ST. MARTIN'S ENGINEERING COLLEGE
An UGC Autonomous Institute
NBA & NAAC A+ Accredited
Dhulapally, Secunderabad-500 100
www.smec.ac.in

PROJECT REPORTS OF CSE

A

PROJECT REPORT

On

5G-Smart Diabetes Toward Personalized Diabetes Diagnosis with Healthcare Big Data Clouds

Submitted by

1)Ms. Apoorva Sharma (17K81A0503)
2) Mr. GVNS Sai Bhaskar (17K81A0515)
3) Mr. Lakhan Palore (17K81A0547)
4) Mr. Nikhil Sangani (17K81A0547)

in partial fulfillment for the award of the degree of
BACHELOR OF TECHNOLOGY

IN

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Under the Guidance of

E. Soumya

Assistant Professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ST. MARTIN'S ENGINEERING COLLEGE
An Autonomous Institute

Dhulapally, Secunderabad – 500 100

JUNE 2021

BONAFIDE CERTIFICATE

This is to certify that the project entitled 5G-Smart Diabetes Toward Personalized Diabetes Diagnosis with Healthcare Big Data Clouds, is being submitted 1) Ms. Apoorva Sharma (17K81A0503) 2) Mr. GVNS Sai Bhaskar (17K81A0515) 3) Mr. Lakhan Palore (17K81A0547) 4) Mr. Nikhil Sangani (17K81A0547) in partial fulfillment of the requirement for the award of the degree of **BACHELOR OF TECHNOLOGY IN DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING** is recorded of bonafide work carried out by them. The result embodied in this report have been verified and found satisfactory.

Signature
E Soumya
Department of CSE

Head of the Department
Dr.M.NARAYANAN
Department of CSE

Internal Examiner

External Examiner

Place:

Date:

DECLARATION

We, the student of **Bachelor of Technology** in Department of Computer Science and Engineering', session: <2017 – 2021>, St. Martin's Engineering College, Dhulapally, Kompally, Secunderabad, hereby declare that work presented in this Project Work entitled 5G-Smart Diabetes Toward Personalized Diabetes Diagnosis with Healthcare Big Data Clouds is the outcome of our own bonafide work and is correct to the best of our knowledge and this work has been undertaken taking care of Engineering Ethics. This result embodied in this project report has not been submitted in any university for award of any degree.

Apoorva Sharma	17K81A0503
GVNS Sai Bhaskar	17K81A0515
Lakhan Palore	17K81A0533
Nikhil Sangani	17K81A0547

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible and whose encouragements and guidance have crowded effects with success.

We extended our deep sense of gratitude to Principal**, Dr. P. SANTOSH KUMAR PATRA**, St. Martin's Engineering College, Dhulapally, for permitting us to undertake this project.

We are also thankful to **Dr. M.NARAYANAN**, Head of the Department, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

We would like to thank **Dr. T. POONGOTHAI,** Department of Computer Science and Engineering (AI & ML) for her encouragement and insightful comments and as well as our project coordinators **Dr. B.RAJALINGAM**, Associate Professor and **Mr. J.SUDHAKAR**, Assistant Professor, in Department of Computer Science and Engineering for their valuable support.

We would like to express our sincere gratitude and indebtedness to our project supervisor E Soumya, Assistant Professor, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for her support and guidance throughout our project.

Finally, we express thanks to all those who have helped us successfully to completing this project. Furthermore, we would like to thank our family and friends for their moral support and encouragement.

We express thanks to all those who have helped us in successfully completing the project.

Apoorva Sharma 17K81A0503

GVNS Sai Bhaskar 17K81A0515

Lakhan Palore 17K81A0533

Nikhil Sangani 17K81A0547

ABSTRACT

Recent advances in big data technologies such as medical big data analytics are enabling the development and implementation of innovative diabetes monitoring systems and applications. Due to the life-long and systematic harm suffered by diabetes patients, it is critical to design effective methods for the diagnosis and treatment of diabetes. Thus, our goal is to design a sustainable, cost-effective, and intelligent diabetes diagnosis solution. The experimental results show that our system can effectively provide personalized diagnosis and treatment suggestions to patients.

TABLE OF CONTENTS

LIST OF FIGURES

1. INTRODUCTION

1.1 PROJECT OVERVIEW

Diabetes is a disease that occurs when your blood glucose, also called blood sugar, is too high. Blood glucose is your main source of energy and comes from the food you eat. Insulin, a hormone made by the pancreas, helps glucose from food get into your cells to be used for energy.

What health problems can people with diabetes develop?

Over time, high blood glucose leads to problems such as

- heart disease
- stroke
- kidney disease
- eye problems
- dental disease
- nerve damage
- foot problem

Most of the food you eat is broken down into sugar (also called glucose) and released into your bloodstream. When your blood sugar goes up, it signals your pancreas to release insulin. Insulin acts like a key to let the blood sugar into your body's cells for use as energy.

Types of Diabetes

1) Type one diabetes outcomes due to the failure of pancreas to supply enough hypoglycemic agent. This type was spoken as "insulin-dependent polygenic disease mellitus" (IDDM) or "juvenile diabetes". The reason is unidentified. The type one polygenic disease found in children beneath twenty years old. People suffer throughout their life because of the type one diabetic and rest on insulin vaccinations. The diabetic patients must often follow workouts and fit regime which are recommended by doctors.

2)The type two diabetes starts with hypoglycemic agent resistance, a situation inside which cells fail to response the hypoglycemic agents efficiently. The sickness develops due to the absence of hypoglycemic agent that additionally built. This type was spoken as "non-insulin-dependent

polygenic disease mellitus". The usual cause is extreme weight. The quantity of people affected by type two will be enlarged by 2025. The existences of diabetes mellitus are condensed by 3% in rural zone as compared to urban zone. The pre hypertension is joined with bulkiness, fatness and diabetes mellitus. The study found that an individual United Nations agency has traditional vital sign. 3) Type 3 Gestational diabetes occurs when a woman is pregnant and develops the high blood sugar levels without a previous history of diabetes. Therefore, it is found that in total 18% of women in pregnancy have diabetes. So, in the older age there is a risk of emerging the gestational diabetes in pregnancy. The obesity is one of the main reasons for type-2 diabetes. The type-2 polygenic disease is under control by proper workout and taking appropriate regime. When the aldohexose level is not reduced by the higher strategies then medications are often recommended. The polygenic disease static report says that 29.1 million people of the United States inhabitants has diabetes.

1.2 PROJECT OBJECTIVES

we first propose a next generation diabetes solution called the 5G-Smart Diabetes system, which integrates novel technologies including fifth generation (5G) mobile networks, machine learning, medical big data, social networking and so on. Then we present the data sharing mechanism and personalized data analysis model for 5G-Smart Diabetes.

Furthermore, the “5G” in 5G-Smart Diabetes has a two-fold meaning. On one hand, it refers to the 5G technology that will be adopted as the communication infrastructure to realize high-quality and continuous monitoring of the physiological states of patients with diabetes and to provide treatment services for such patients without restraining their freedom. On the other hand, “5G” refers to the following “5 goals”: cost effectiveness, comfortability, personalization, sustainability, and smartness.

Cost Effectiveness: It is achieved from two aspects. First, 5G-Smart Diabetes keeps users in a healthy lifestyle so as to prevent users from getting the disease in the early

stage. The reduction of disease risk would lead to decreasing the cost of diabetes treatment. Second, 5G-Smart Diabetes facilitates out-of-hospital treatment, thus reducing the cost compared to on-the-spot treatment, especially long-term hospitalization of the patient.

Comfortability: To achieve comfort for patients, it is required that 5G-Smart Diabetes does not disturb the patients' daily activities as much as possible. Thus, 5G-Smart Diabetes integrates smart clothing [3], mobile phones, and portable blood glucose monitoring devices to easily monitor patients' blood glucose and other physiological indicators.

Personalization: 5G-Smart Diabetes utilizes various machine learning and cognitive computing algorithms to establish personalized diabetes diagnosis for the prevention and treatment of diabetes. Based on the collected blood glucose data and individualized physiological indicators, 5G-Smart Diabetes produces personalized treatment solutions for patients.

Sustainability: By continuously collecting, storing, and analyzing information on personal diabetes, 5G-Smart Diabetes adjusts the treatment strategy in time based on the changes of patients' status. Furthermore, in order to be sustainable for data-driven diabetes diagnosis and treatment, 5G-Smart Diabetes establishes effective information sharing among patients, relatives, friends, personal health advisors, and doctors.

1.2 ORGANIZATION OF CHAPTERS

Introduction:

In Introduction part we have discussed about the definition of diabetes, various types of diabetes and what are we going to propose in this project.

Literature Survey:

In Literature Survey we have mentioned the papers which we have used for our project and we have also mentioned which paper is used for what purpose in our project

Software and Hardware Requirements:

In the software and hardware requirements we have mentioned the prerequisites for the project like the minimum hardware requirements and software requirements

Software Development Analysis:

In the software development analysis, we have mentioned about the problem, the modules we used our program and the how to overcome the problem from the existing system.

Project System Design:

In this we have discussed about our system designs by using the UML diagrams and we have explained the diagrams

Project Coding:

In the project coding section, we have given the coding snippets so that everyone can understand the project.

Project Testing:

In the Testing part we have discussed about the different testcases we have used in our program. And we have also discussed about the white box and the black box testing.

Output Screens:

In the Output screen we have displayed our output for the program.

2. LITERATURE SURVEY

2.1 SURVEY ON BACKGROUND

- Veena Vijayan V. And Anjali C has discussed, the diabetes disease produced by rise of sugar level in the plasma. Various computerized information systems were outlined utilizing classifiers for anticipating and diagnosing diabetes using decision tree, SVM, Naive Bayes and ANN algorithms .
- P. Suresh Kumar and V. Uma Tejaswi has presented the algorithms like Decision Tree, SVM, Naive Bayes for identifying diabetes using data mining techniques.
- Ridam Pal , Dr.Jayanta Poray and Mainak Sen has presented the Diabatic Retinopathy (DR) which is one of the leading cause of sight inefficiency for diabetic patients. In which they reviewed the performance of a set of machine learning algorithms and verify their performance for a particular data set.
- Dr. M. Renuka Devi and J. Maria Shyla has discussed about the analysis of various skills of mining to guess diabetes using Naive Bayes, Random forest, Decision Tree and J48 algorithms.
- Rahul Joshi and Minyechil Alehegn has discussed the ML techniques which are used to guess the datasets at an initial phase to save the life. Using KNN and Naive Bayes algorithm.
- Zhilbert Tafa and Nerxhivane Pervetica has discussed the result of algorithms that are implemented in order to progress the diagnosis reliability.
- Prof. Dhomse Kanchan B. and Mr. Mahale Kishor M. has discussed the study of Machine Learning Algorithms such as Support Vector Machine, Naïve Bayes, Decision Tree, PCA for Special Disease Prediction using Principal of Component Analysis.
- Diabetes or diabetes mellitus is a metabolic disorder (metabolic) in the body. This disease destroy the ability to produce insulin in the patient's body or the body develops resistance to insulin the and consequently the produced insulin cannot achieve its normal job. The main role of the produced insulin is to decrees blood sugar by different instruments. There are two key types of diabetes. In Type I diabetes, obliteration of beta pancreatic cells damage insulin construction and in type II, there is a progressive insulin confrontation

in the body and ultimately may yield to the obliteration of pancreatic beta cells and faults in insulin production. In type II diabetes, it is known that genetic issues, obesity and lack of physical activity have a vital part in a person [1]. Even though the precise cause of type I diabetes is unidentified, issues that may indicate a greater risk comprise the followings [2]:

- Family history. A person risk upsurges if his parent or sibling has history of type I diabetes.
- Environmental factors. Situations for example contact with a viral illness probably play some role in type I diabetes.
- The existence of harmful immune system cells. Occasionally family members of a person with type I diabetes are examined for the existence of diabetes autoantibodies. If a person has these autoantibodies, he/she has a chance of increased risk for evolving type I diabetes. Nonetheless not every person who has these autoantibodies gets diabetes.
- Geography. Some countries, like Sweden, have bigger rates of type I diabetes. Researchers don't completely comprehend why certain people develop pre-diabetes and type II diabetes and others don't. It's sure that some factors upsurge the risk like :
- Weight. The more fatty tissue you have, the more resilient a person cells to insulin. International Journal of Advanced Science and Technology Vol.124 (2018), pp. 1-10 3
- Inactivity. The less energetic a person is, the more a person has risk. Physical activity assists a person control of his/her weight, consumes glucose as energy and makes a person cells more sensitive to insulin.
- Family history. A person risk upsurges if his parent or sibling has history of type II diabetes.
- Race. Even though it's uncertain why, people of specific races are at higher risk.
- Age. A person risk upsurges as he/she gets older. This may be because a person has a habit to exercise less, lose muscle mass and add weight as he/she gets older. Nonetheless type II diabetes is likewise growing among children, youths and adults.
- Gestational diabetes. If a person developed gestational diabetes when she

was pregnant, her risk of emerging pre-diabetes and type II diabetes far ahead upsurges. If she gives birth to a baby weighing more than 4 kilograms, she is also at risk of type II diabetes.

- Polycystic ovary syndrome. For females, having polycystic ovary syndrome increases the risk of getting diabetes.
- High blood pressure. Having blood pressure more than 140/90 millimeters of mercury (mm Hg) is connected to an augmented risk of type II diabetes.

- Abnormal cholesterol and triglyceride levels. If a person has low levels of highdensity lipoprotein, or good cholesterol, his/her risk of type II diabetes is going to be higher. Triglycerides are additional type of fat passed in the blood. A person with greater levels of triglycerides has an augmented risk of type II diabetes. A practical approach to this type of problem is the application of regression analysis where past data is better combined into some functions. The result is an equation in which both xj inputs are multiplied by wj ; the sum of all these products is constant, and then output $y = \Sigma\ wj\ xj +$, where $j = 0..n$. The problem is the difficulty of choosing an appropriate function to have all the collected data and adjust the output automatically when more information is attained, because the candidate's performance is organized by a number of arguments, and this control will not have any clear regression model. The artificial neural network, which emulates the human thinking in solving a problem, is a more common approach that can address this type of problems. Thus, the attempt to develop an adaptive system such as artificial neural network to predict the situation and classification based on the results of these arguments .

2.2 CONCLUSIONS ON SURVEY

1. From the first paper we have taken the definition of diabetes and how diabetes is caused and the types of diabetes
2. From the second paper we have taken the working of the Decision Tree Algorithm and we have also understood how to use and implement the algorithm in the project.
3. From the third paper we have learnt how different machine learning algorithms are used in our project and verify their performance for a particular data set.
4. From the fourth paper we have learnt how the naive based classifiers and decision tree algorithm is used for prediction of diabetes
5. From the fifth paper we have learnt about ML techniques which are used to guess the datasets at an initial phase to save the life. Using KNN and Naive Bayes algorithm.
6. From the sixth paper we have learnt about then results of the diabetes i.e; type 1 and type2 diabetes and also about the UML diagrams.
7. From the seventh paper we have learnt about the ANN, Ensembled model and the implementation of those algorithms and Special Disease Prediction using Principal of Component Analysis.
8. The author used Data Mining to develop a model for classifying diabetic patient control level based on historical medical records. The author was motivated by the death caused by diabetes in the world which necessitated avoiding the complication of the disease. He developed a new predictive model using data mining techniques which would classify diabetic patient control level based on historical medical records. The research was carried out using three data mining techniques which are Naïve Bayes, Logistic and J48. The research was implemented using WEKA application. The result showed that Logistic data mining algorithm gave a precision average of 0.73, recall of 0.744, Fmeasure of 0.653 and accuracy of 74.4%. Naïve Bayes gave a precision average of 0.717, recall of 0.742, F-measure of 0.653 and accuracy of 74.2%. J48 gave a precision average of 0.54, recall of 0.735, F-measure of 0.623 and accuracy of 73.5%. This proved that the logistic algorithm was more accurate than the other two. The research was limited in that only diabetes type 2 was considered. They also did not look into the discovery of appropriate features with minimal effort and validation on discovered features.

3. SOFTWARE AND HARDWARE REQUIREMENTS

3.1 SOFTWARE REQUIREMENTS

a. Operating system : Windows 7 Ultimate.

- Windows 7 Ultimate Edition is best known for its reliability, compatibility, and performance. It contains all the features and applications that are not available in other editions of Windows 7. ... Most of the Windows users depend on this version of Windows 7 because of its features and tools it has to offer.

b. Coding Language : Python.

- Python is an interpreted high-level general-purpose programming language. Python's design philosophy emphasizes code readability with its notable use of significant indentation. Its language constructs as well as its object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects.[

c. Front-End : Python.

- Python’s Tkinter module can be used as the fronted part. Tkinter commonly comes bundled with Python, using Tk and is Python's standard GUI framework. It is famous for its simplicity and graphical user interface. It is open-source and available under the Python License.

d. Designing : Html, CSS, JavaScript.

- The Hypertext Markup Language, or HTML is the standard markup language for documents designed to be displayed in a web browser. It can be assisted by technologies such as Cascading Style Sheets (CSS) and scripting languages such as JavaScript.

3.2 HARDWARE REQUIREMENTS

1. System : i3 processor

- The Core i3 processor is available in multiple speeds, ranging from 1.30 GHz up to 3.50 GHz, and features either 3 MB or 4 MB of cache. It utilizes either the LGA 1150 or LGA 1155 socket on a motherboard. Core i3 processors are most often found as dual core, having two cores.

2. Hard Disk : 500 GB.

3. Monitor : 14’ Color Monitor.

4. Mouse : Optical Mouse.

5. Ram : 2 Gb.

4. SOFTWARE DEVELOPMENT ANALYSIS

4.1 OVERVIEW OF PROBLEM

Diabetes is a very common disease and early diagnosis is very important. As we have seen in the conclusion of literature survey part, there are many algorithms that can predict whether the person is at risk of getting diabetes or not, but the accuracy is quite low, up to 75-80 %. We want to increase the accuracy of our system to give more apt results.

4.2 DEFINE THE PROBLEM

In today's world diabetes is very common. The global diabetes prevalence in 2019 is estimated to be 9.3%. It can be cause due to unhealthy lifestyle or even because of hereditary factor. It's a fatal disease which when not diagnosed early can cause failure of kidney, weakening of eyesight or pus formation in feet. So early diagnosis of diabetes is very important. We have seen and studied many algorithms that predict the risk of diabetes of a person in near future, but accuracy of those algorithms is from 75-80%. We wanted to make a system that can predict the risk of diabetes with an accuracy of 85-90%.

4.3 MODULES OVERVIEW

i. Upload files

   We upload files using this module

ii. Preprocess data

   Preprocessing of the uploaded file is done using this module

iii. Run decision tree algorithm

   This module is used to run the decision tree algorithm.

iv. Run SVM algorithm

This module is used to run the SVM algorithm.

v. Run ANN algorithm

This module is used to run the ANN algorithm.

vi. Run Ensemble model

This module is used to run the Ensemble model.

vii. Run graph

This module is used to run the comparison graph.

viii. Run server

This module is used to start the cloud server.

a. Upload file

Using this module, the user can upload his/her medical file.

4.4 MODULES DEFINITION

Upload files

We upload files used for training the algorithms using this module

Preprocess data

Preprocessing of the uploaded file is done using this module. Preprocessing involves data quality assessment, data cleaning, data transformation and data reduction.

Run decision tree algorithm

This module is used to run the decision tree algorithm. Decision tree algorithm consider one condition at a time and decide which category it belongs to.

Run SVM algorithm

This module is used to run the SVM algorithm. SVM (Support Vector Machine) algorithm uses a line to divide the data into the given number of categories.

Run ANN algorithm

This module is used to run the ANN algorithm. Artificial Neural Networks uses many hidden layers to decide which category the user's data belong to, i.e., if the risk of diabetes is positive or negative.

Run Ensemble model

This module is used to run the Ensemble model. Ensemble model uses a combination of different algorithms to find the best predict and gives the same result.

Run graph

This module is used to run the comparison graph. The graph shows the comparison between the accuracy of different algorithms. The x-axis has the name of the algorithm and the y-axis has the accuracies.

Run server

This module is used to start the cloud server.

- **Upload file**

  Using this module, the user can upload his/her medical file. This module then uses ensemble model to predict the risk of diabetes or not in near future.

4.5 MODULES FUNCTIONALITY

Upload files

We upload files used for training the algorithms using this module. The dataset contains seven hundred sixty-eight instances and eight features.

The dataset features are:

- Total number of times pregnant
- Glucose/sugar level
- Diastolic Blood Pressure
- Body Mass Index (BMI)
- Skin fold thickness in mm
- Insulin value in 2 hours
- Hereditary factor- Pedigree function
- Age of patient in years

Total number of pregnancies

Total number of pregnancies is the sum of all live births, abortions, and miscarriages.

Glucose/sugar level

The blood sugar level, blood sugar concentration, or blood glucose level is the measure of concentration of glucose present in the blood of humans or other animals. Approximately 4 grams of glucose, a simple sugar, is present in the blood of a 70 kg (154 lb) human at all times.

Normal blood glucose level (tested while fasting) for non-diabetics is between 3.9 and 7.1 mmol/L (70 to 130 mg/dL). The global mean fasting plasma blood glucose level in humans is about 5.5 mmol/L (100 mg/dL);[7][5] however, this level fluctuates throughout the day. Blood sugar levels for those without diabetes and who are not fasting should be below 6.9 mmol/L (125 mg/dL).

Diastolic Blood Pressure

Diastolic blood pressure measures the pressure on the walls of your arteries between heartbeats. A normal diastolic blood pressure is less than 80 mmHg. Systolic blood pressure measures the pressure on the walls of your arteries when your heart beats. A normal systolic blood pressure is less than 120 mmHg.

Baseline blood pressure data from several recent trials indicate that, in diabetic subjects, there is nearly a fourfold excess in systolic pressure (the difference between baseline pressure and target pressure) over diastolic pressure with respect to the recommended systolic/diastolic target pressure of <130/80 mmHg.

Body Mass Index (BMI)

Body mass index is a value derived from the mass and height of a person. The BMI is defined as the body mass divided by the square of the body height, and is expressed in units of kg/m^2, resulting from mass in kilograms and height in metres.

In general, a BMI of 18.5 to 24.9 is considered a normal, or healthy, weight. A BMI that ranges from 25 to 29.9 is considered overweight. And a BMI of 30 or higher falls into the obese category, according to the CDC.

Obesity and type 2 diabetes are closely related. Having a higher amount of body fat increases your chances of developing diabetes. And, if you have diabetes, extra weight means you're at higher risk for complications like a heart attack or stroke.

Skin fold thickness in mm

Skinfold thicknesses measure subcutaneous body fat and, therefore, indicate body composition. TSFT and SSFT indicate subcutaneous fat on the limbs and body trunk, respectively.

Subcutaneous fat, and therefore skinfold thicknesses at the different sites, changes at varying rates with age, weight change, with diseases such as diabetes, and in women during pregnancy, postpartum, and at the menopause.

Insulin value in 2 hours

Insulin levels in the blood can be interpreted using a simple blood test that is performed after eight hours of fasting.

Normal results for the two-hour postprandial test based on age are: For those who don't have diabetes: less than 140 mg/dL. For those who have diabetes: less than 180 mg/dL.

High insulin levels generally translate into an overworked pancreas. This may be followed by the exhaustion of the pancreatic cells, resulting in the development of diabetes mellitus. Diabetes further brings on complications such as heart disease, nerve damage, eye damage and kidney damage.

Hereditary factor- Pedigree function

Heredity refers to the genetic heritage passed down by our biological parents.

A pedigree is a genetic representation of a family tree that diagrams the inheritance of a trait or disease though several generations. The pedigree shows the relationships between family members and indicates which individuals express or silently carry the trait in question.

The family history is useful in stratifying a patient's risk for rare single-gene disorders and more common diseases with multiple genetic and environmental contributions.

The risk of developing type 2 diabetes increases with the number of affected family members. The increased risk is likely due in part to shared genetic factors, but it is also related to lifestyle influences (such as eating and exercise habits) that are shared by members of a family.

Age of patient in years

The UN recommendations define age as "the interval of time between the date of birth and the date of the census, expressed in completed solar years". This is equivalent to this standard's definition of age as "age at last birthday".

Older adults are at high risk for the development of type 2 diabetes due to the combined effects of increasing insulin resistance and impaired pancreatic islet function with aging. About 1 in 4 adults over age 60 have diabetes. Having the disease makes you more likely to get some serious complications. And so does getting older. The combination of the two can even make some health problems worse.

Therefore, middle-aged and older adults are still at the highest risk for developing type 2 diabetes.

Preprocess data

Preprocessing of the uploaded file is done using this module.

Pre-processing refers to the transformations applied to our data before providing the data to the algorithm.

Data Pre-processing technique is used to convert the raw data into an understandable data set.

Steps in pre-processing are:

1. Test-Train split
2. Data quality assessment
3. Data cleaning
4. Data transformation
5. Data reduction

Test-Train split

We divide the entire dataset in two parts: training data and testing data.

Training dataset is used to train the model, to create a model. The algorithms use training dataset to learn and extract patterns. If the training set is labelled correctly, then the model will be able to acquire something from the features.

Testing dataset is used to test the model, to test the accuracy of the model. We use testing dataset to evaluate how accurate the algorithm's prediction is. So, for testing the model such type of data is used to check whether it is responding correctly or not.

Percentage split option is provided for training and testing. Out of 768 instances

75 % is used for training and 25% is used for testing.

Data quality assessment

- Mismatching in data types

  Quite often, we might mix together datasets that use different data formats. Hence, the mismatching: integer vs. float or UTF8 vs ASCII.

- Different dimensions of data arrays

  When we aggregate data from different datasets, for example, from five different arrays of data for voice recognition, three fields that are present in one of them can be missing in four other arrays.

- Mixed of data values

  Let's imagine that you have data, collected from two independent sources. As a result, the gender field has two different values for women: woman and female.

  To clean this dataset, you have to make sure that the same name is used as the descriptor within the dataset

- Outliers in dataset

  Outliers are very dangerous. They can strongly influence the output of a machine learning model. Usually, the researchers evaluate the outliers to identify whether each particular record is the result of an error in the data collection or a unique phenomenon which should be taken into consideration for data processing.

- Missing data

  We may also notice that some important values are missing. These problems arise due to the human factor, program errors, or other reasons. They will affect the accuracy of the predictions, so before going any further with your database, you need to do data cleaning.

Data cleaning

- Missing data

  When we concatenate two or more datasets into one database to get a bigger training set, some data field mismatches are quite common.

  When not all the fields are represented in the joined massive, it is better to delete such fields in advance before merging.

- Noisy data

  A large amount of additional meaningless data is called noise. This includes data corruption and the term is often used as a synonym for corrupt data. It also includes any data that a user system cannot understand and interpret correctly.

Data transformation

By data transformation, we understand the methods of turning the data into an appropriate format for the computer to learn from.

- Aggregation

  In the case of data aggregation, the data is pooled together and presented in a unified format for data analysis.

- Normalization

  Normalization helps you to scale the data within a range to avoid building incorrect ML models while training and/or executing data analysis. If the data range is very wide, it will be hard to compare the figures.

- Feature selection

  Feature selection is the selection of variables in data that are the best predictors for the variable we want to predict. If there are a lot of features, then the classifier operation time increases. In addition, the prediction accuracy often decreases. Especially if there are a lot of garbage features in the data.

- Discretization

  During discretization, we transforms the data into sets of small intervals.

Data reduction

Data reduction can be used to reduce the amount of data and decrease the costs of analysis.

- Attribute feature selection

  Techniques for data transformation can also be used for data reduction. If you construct a new feature combining the given features in order to make the data mining process more efficient, it is called an attribute selection.

- Dimensionality reduction

  Datasets that are used to solve real-life tasks have a huge number of features. It's possible to use dimensionality reduction to cut the number of features used.

Decision tree algorithm

This module is used to run the decision tree algorithm. Decision tree algorithm consider one condition at a time and decide which category it belongs to.

Decision tree:

A Decision Tree is a simple representation for classifying examples. It is a Supervised Machine Learning where the data is continuously split according to a certain parameter.

Decision Tree consists of:

Nodes: Test for the value of a certain attribute.

Edges/ Branch: Correspond to the outcome of a test and connect to the next node or leaf.

Leaf nodes: Terminal nodes that predict the outcome (represent class labels or class distribution).

Algorithm:

- Using the decision algorithm, we start at the tree root and split the data on the feature that results in the largest information gain (IG) (reduction in uncertainty towards the final decision).

- In an iterative process, we can then repeat this splitting procedure at each child node until the leaves are pure. This means that the samples at each leaf node all belong to the same class.
- In practice, we may set a limit on the depth of the tree to prevent overfitting. We compromise on purity here somewhat as the final leaves may still have some impurity.

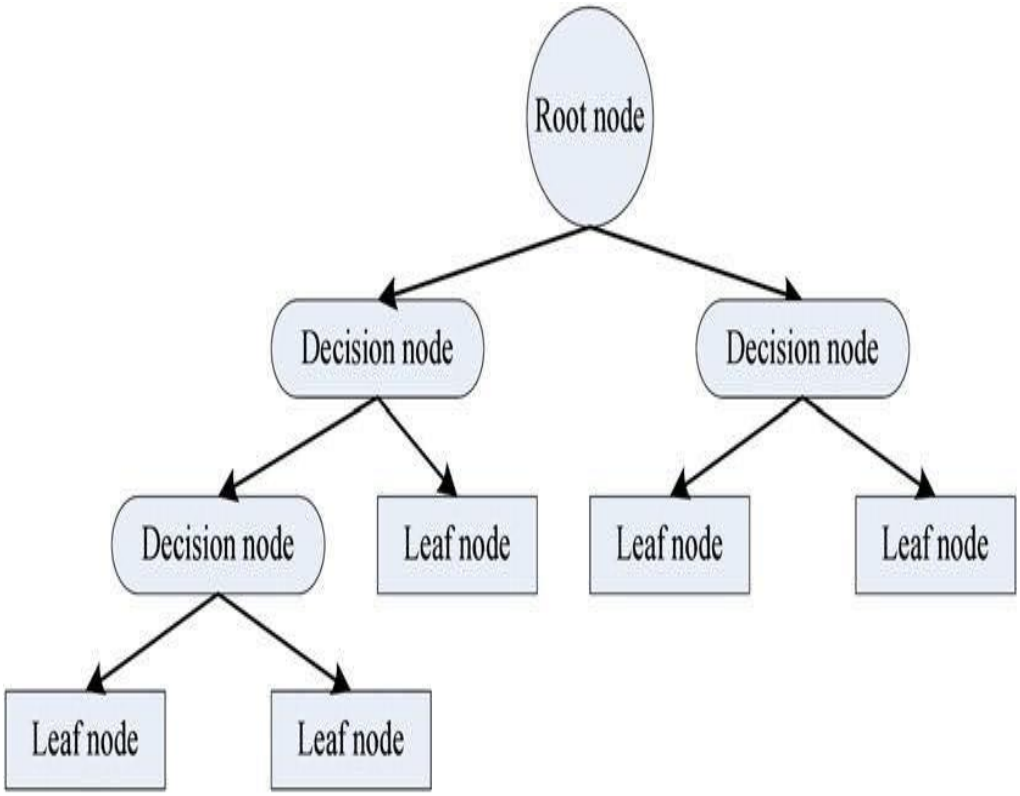


FIG: 1 DESCISION TREE

SVM algorithm

This module is used to run the SVM algorithm. SVM (Support Vector Machine) algorithm uses a line to divide the data into the given number of categories.

The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space (N — the number of features) that distinctly classifies the data points.

To separate the two classes of data points, there are many possible hyperplanes that could be chosen. Our objective is to find a plane that has the maximum margin, i.e the maximum distance between data points of both classes. Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence.

Hyperplanes are decision boundaries that help classify the data points. Data points falling on either side of the hyperplane can be attributed to different classes. Also, the dimension of the hyperplane depends upon the number of features. If the number of input features is 2, then the hyperplane is just a line. If the number of input features is 3, then the hyperplane becomes a two-dimensional plane. It becomes difficult to imagine when the number of features exceeds 3.

Support vectors are data points that are closer to the hyperplane and influence the position and orientation of the hyperplane. Using these support vectors, we maximize the margin of the classifier. Deleting the support vectors will change the position of the hyperplane. These are the points that help us build our SVM.

In the SVM algorithm, we are looking to maximize the margin between the data points and the hyperplane. The loss function that helps maximize the margin is hinge loss.

The cost is 0 if the predicted value and the actual value are of the same sign. If they are not, we then calculate the loss value. We also add a regularization parameter the cost function. The objective of the regularization parameter is to balance the margin maximization and loss.

Now that we have the loss function, we take partial derivatives with respect to the weights to find the gradients. Using the gradients, we can update our weights.

When there is no misclassification, i.e., our model correctly predicts the class of our data point, we only have to update the gradient from the regularization parameter.

When there is a misclassification, i.e., our model makes a mistake on the prediction of the class of our data point, we include the loss along with the regularization parameter to perform gradient update.

Cost function:

$$(\theta) = C\sum_{i=1}^{m}\left[y^{(i)}cost_1(\theta^T x^{(i)}) + (1-y^{(i)})cost_0(\theta^T x^{(i)})\right] + \frac{1}{2}\sum_{i=1}^{n}\theta_i^2$$

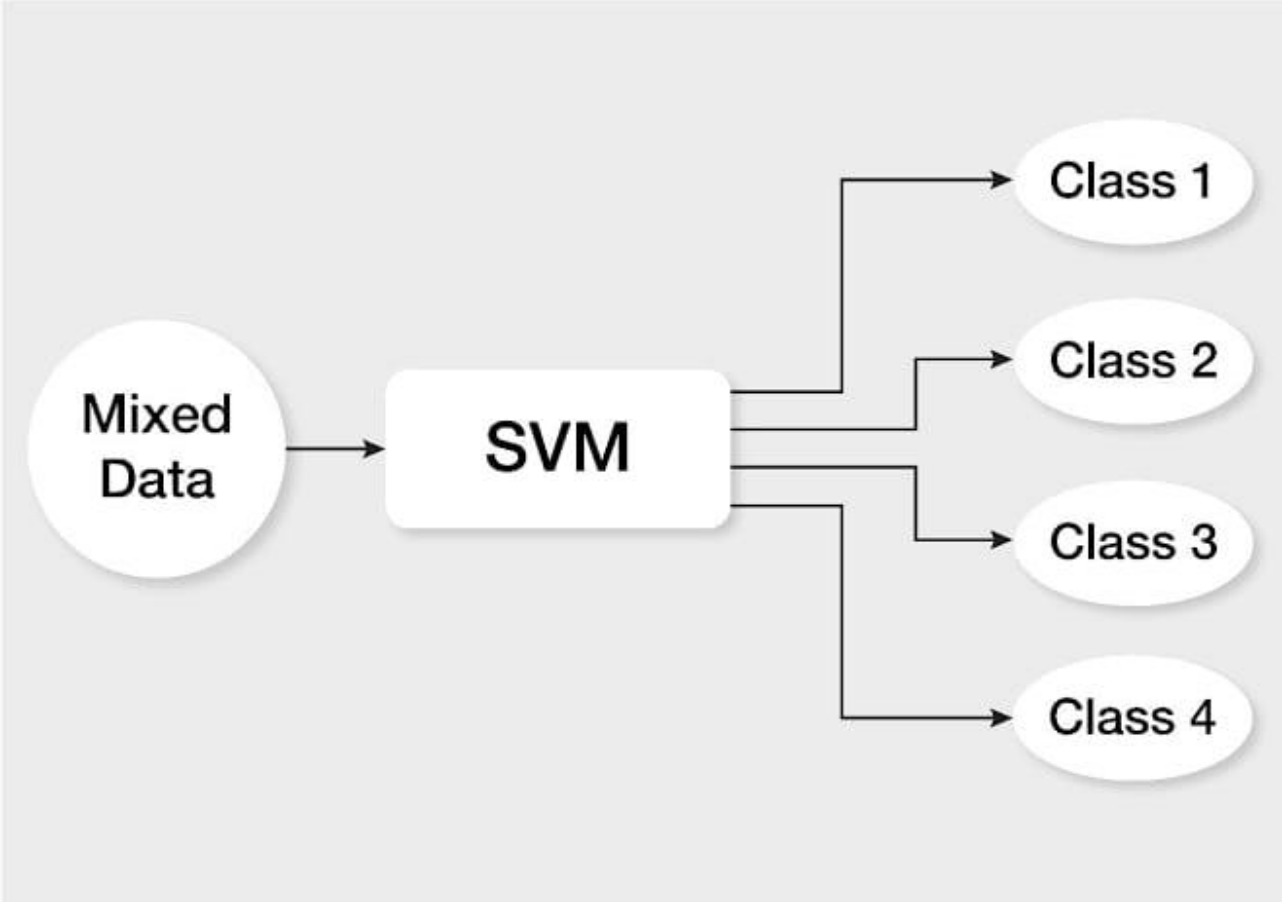


FIG: 2 SVM ALGORITHM

Run ANN algorithm

This module is used to run the ANN algorithm. Artificial Neural Networks uses many hidden layers to decide which category the user's data belong to, i.e., if the risk of diabetes is positive or negative.

Artificial Neural Networks or shortly ANN's are widely used today in many applications and, classification is one of them and also there are many libraries and frameworks that are dedicated to building Neural Networks with ease.

We used MLP type of Artificial Neural Networks.

A multilayer perceptron (MLP) is a feedforward artificial neural network that generates a set of outputs from a set of inputs. An MLP is characterized by several layers of input nodes connected as a directed graph between the input and output layers. MLP uses backpropogation for training the network. MLP is a deep learning method.

A multilayer perceptron is a neural network connecting multiple layers in a directed graph, which means that the signal path through the nodes only goes one way. Each node, apart from the input nodes, has a nonlinear activation function. An MLP uses backpropagation as a supervised learning technique. Since there are multiple layers of neurons, MLP is a deep learning technique.

MLP is widely used for solving problems that require supervised learning as well as research into computational neuroscience and parallel distributed processing.

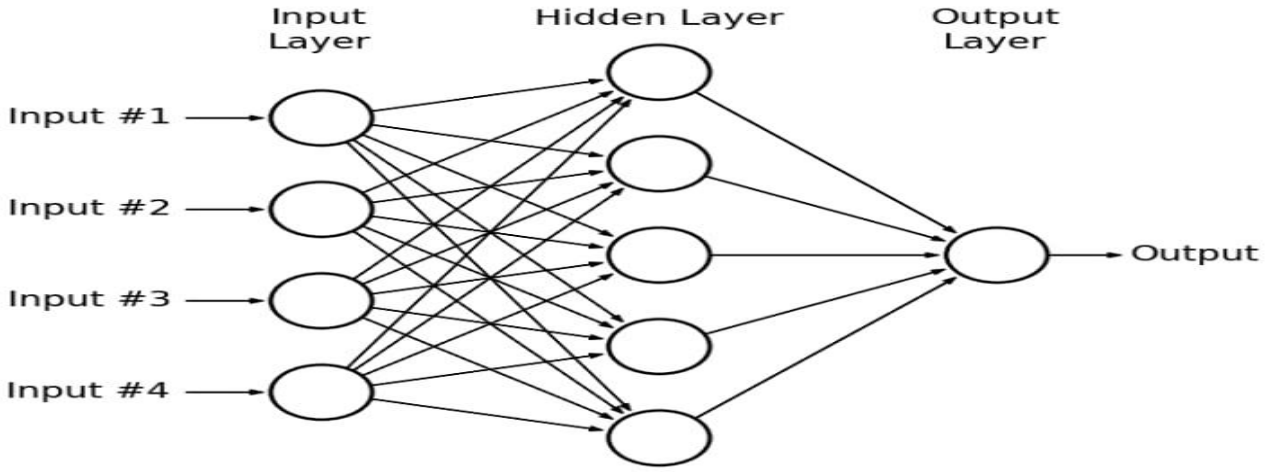


FIG: 3 ANN ALGORITHM

Run Ensemble model

This module is used to run the Ensemble model. Ensemble model uses a combination of different algorithms to find the best predict and gives the same result.We used Voting classifier from ensemble model.

A voting classifier is a classification method that employs multiple classifiers to make predictions. It is very applicable in situations when a data scientist or machine learning engineer is confused about which classification method to use. Therefore, using the predictions from multiple classifiers, the voting classifier makes predictions based on the most frequent one.

The scikit-learn's ensemble feature is accessed and a voting classifier is imported. There are three other classifiers in the code above: a decision tree classifier, Support Vector Machine and Artificial Neural Networks.

Afterwards, an object is created for the voting classifier. The voting classifier has two basic hyperparameters: estimators and voting.

The **estimators** hyperparameter creates a list for the objects of the three classifiers above while assigning names to them. The **voting** hyperparameter is set to either hard or soft.

If set to hard, the voting classifier will make judgments based on the predictions that appear the most. Otherwise, if set to soft, it will use a weighted approach to make its decision. It's recommended to set it to soft when using an even number of classifiers because of its weighted approach and setting it to hard when using an odd number of classifiers because of its "majority carry the vote" approach.

The voting classifier like any other machine learning algorithm is used to fit the independent variables of the training dataset with the dependent variables.

The voting classifier is a remarkable approach for classification because its methodology utilizes the collective judgment of multiple classifiers for predicting data points.

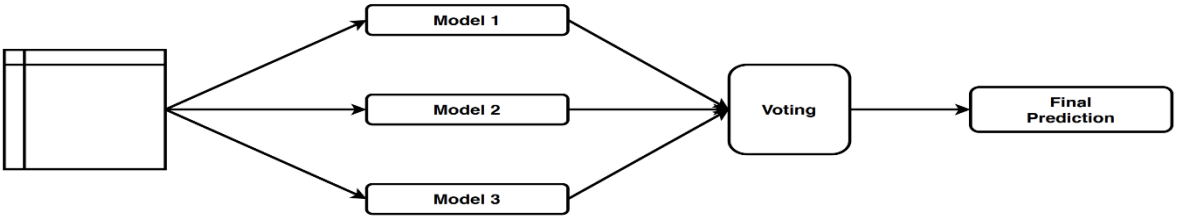


FIG: 4 ENSEMBLED MODEL

Run graph

This module is used to run the comparison graph. The graph shows the comparison between the accuracy of different algorithms. The x-axis has the name of the algorithm and the y-axis has the accuracies.

We use Matplotlib for plotting the graph.

Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python.

matplotlib.pyplot is a collection of functions that make matplotlib work like MATLAB. Each pyplot function makes some change to a figure: e.g., creates a figure, creates a plotting area in a figure, plots some lines in a plotting area, decorates the plot with labels, etc.

In matplotlib.pyplot various states are preserved across function calls, so that it keeps track of things like the current figure and plotting area, and the plotting functions are directed to the current axes.

Run server

This module is used to start the cloud server.

- **Upload file**

  Using this module, the user can upload his/her medical file. This module then uses ensemble model to predict the risk of diabetes or not in near future.

5. PROJECT SYSTEM DESIGN

5.1 UML DIAGRAMS

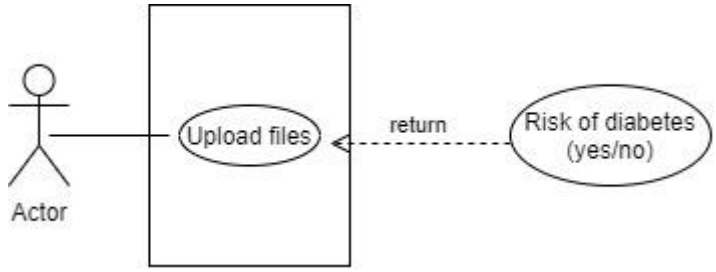


Fig:5 Use case diagram

A use case is a list of actions or event steps typically defining the interactions between a role (known in the Unified Modeling Language (UML) as an actor) and a system to achieve a goal. The actor can be a human or other external system. In our project, the actor can upload files and in return he/she will found out if they have a risk of suffering from diabetes in near future.

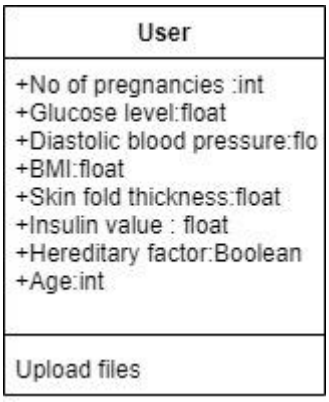


Fig: 6 Class diagram

A class diagram in the Unified Modeling Language is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations, and the relationships among objects. In our project, there is only one class i.e., User. The user need to upload a file which contains the their no of pregnancies, glucose level, diastolic blood pressure, BMI, skin fold thickness, insulin value, hereditary

factor and age.

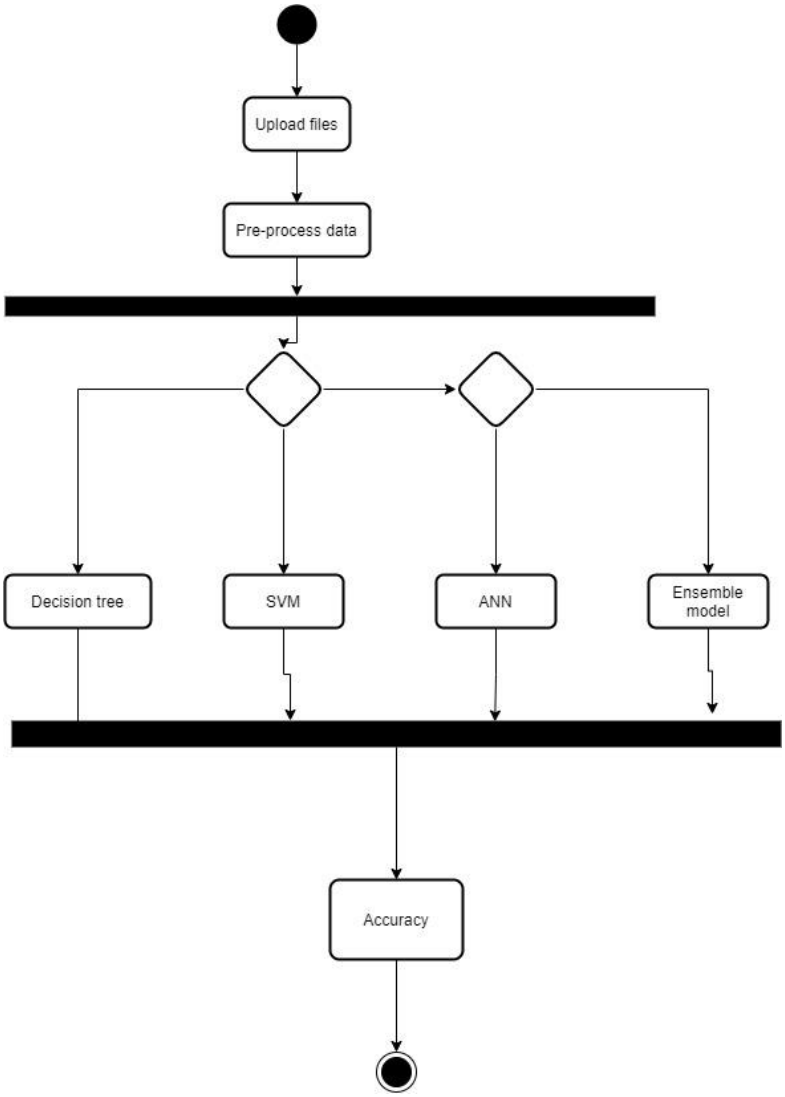


Fig :7 Activity diagram for training data

Activity diagram is basically a flowchart to represent the flow from one activity to another activity. In our project when the user upload data for training, the activity diagram is followed as shown in the figure above. First the uploaded data is preprocessed. Then, the user can select any of the algorithms and in the result part they'll get the accuracy of that algorithm.

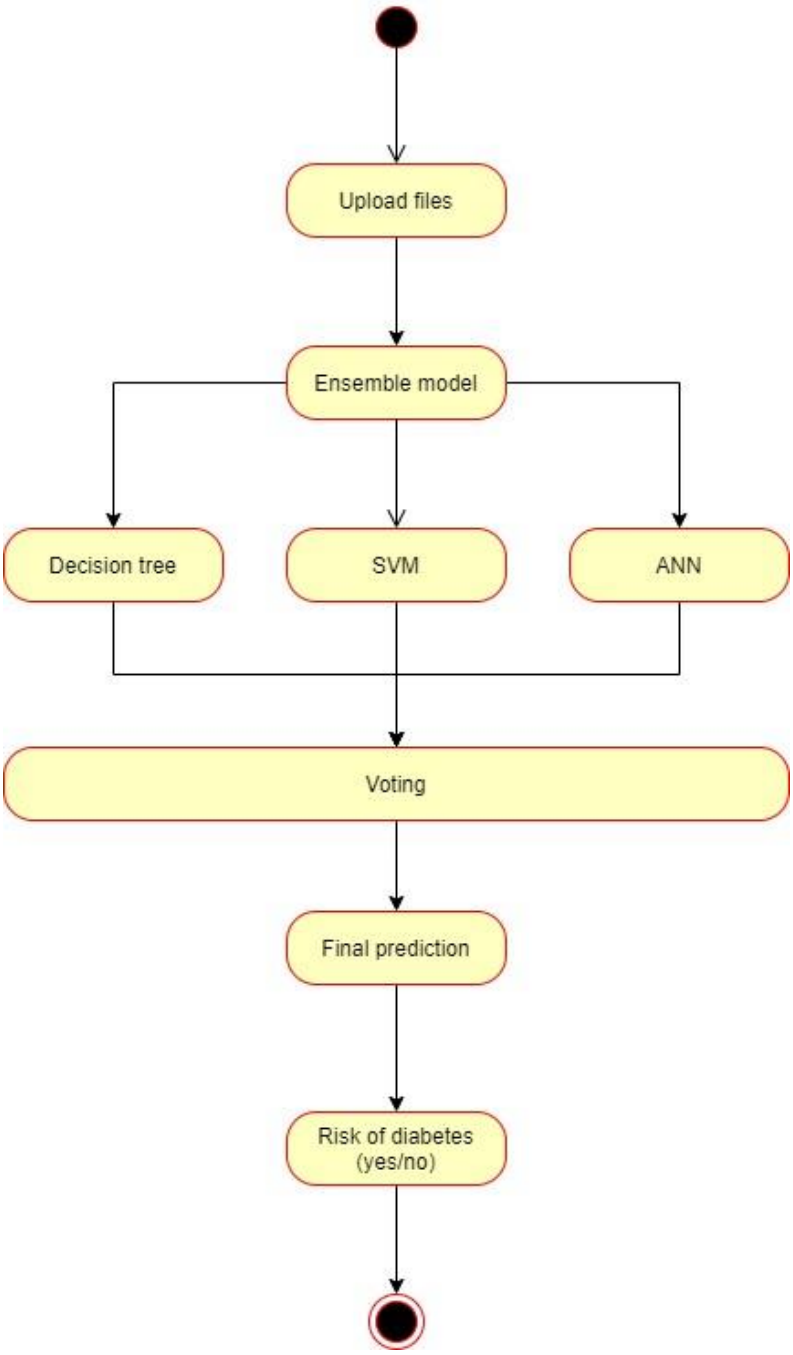


Fig: 8 Activity diagram for prediction of diabetes

The above diagram shows what happens when the user uploads a file to predict the risk of diabetes in near future. When the user uploads the file, it is used by ensemble model to predict different predictions, a voting algorithm is used to know the best prediction and finally the best prediction is returned to the user.

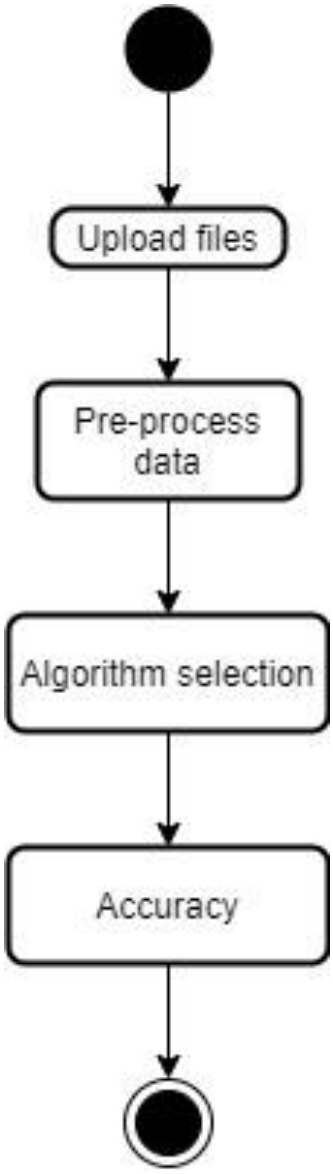


Fig 9: State diagram for training data

A state diagram is a type of diagram used in computer science and related fields to describe the behavior of systems. It shows the different states the project or the system will be in. The states of our project while training a data is shown in the figure above. First our project is in upload files state, once the user upload files, we pre-process data, then the user can select one of the algorithms that they want to the accuracy of. Finally, the accuracy of that algorithm is displayed.

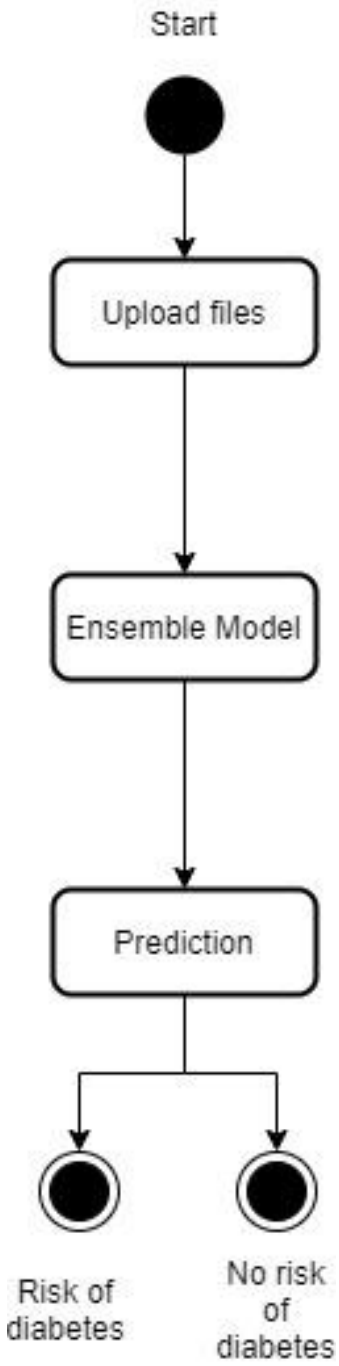


Fig 10: State diagram of the prediction algorithm

In our project, when the user uploads their file, it goes to ensemble model building state, then it enters the prediction state where we find out the prediction of the risk of diabetes, i.e., if the user could suffer from diabetes in the near future or not.

6. PROJECT CODING

6.1 CODE TEMPLATES

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from tkinter import messagebox
from tkinter import *
from tkinter.filedialog import askopenfilename
from tkinter import simpledialog
import tkinter
from tkinter import filedialog
import os
from sklearn.model_selection import train_test_split
from sklearn import metrics
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import accuracy_score
from sklearn import svm
from sklearn.neural_network import MLPClassifier
from sklearn.ensemble import VotingClassifier
import socket
```

Fig:11 code_for_importing_library

In the above code template, we are importing all the necessary libraries.

We use the keyword “import” to import different libraries to our file in python

“pandas” is a software library written for the Python programming language for data manipulation and analysis.

“NumPy” is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.

“matplotlib” is a plotting library for the Python programming language and its numerical mathematics extension NumPy.

“matplotlib.pyplot” is mainly intended for interactive plots and simple cases of

programmatic plot generation.

“Tkinter” is a Python binding to the Tk GUI toolkit. It is the standard Python interface to the Tk GUI toolkit, and is Python's de facto standard GUI.

“OS” module in Python provides functions for interacting with the operating system. OS comes under Python's standard utility modules.

“sklearn” (Scikit-learn) is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms.

```
def upload():
    global filename
    filename = filedialog.askopenfilename(initialdir="dataset")
    pathlabel.config(text=filename)
```

Fig: 12 code_for_upload

The dataset contains seven hundred sixty-eight instances and eight features.

The dataset features are:

- Total number of times pregnant
- Glucose/sugar level
- Diastolic Blood Pressure
- Body Mass Index (BMI)
- Skin fold thickness in mm
- Insulin value in 2 hours
- Hereditary factor- Pedigree function
- Age of patient in years

```
def preprocess():
    global X_train
    global y_train
    global dataset
    global X_test
    global y_test
    dataset = pd.read_csv(filename)
    y = dataset['Outcome']
    X = dataset.drop(['Outcome'], axis = 1)
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.1, random_state=0)
    text.delete('1.0', END)
    text.insert(END,"Dataset Length : "+str(len(dataset))+"\n")
```

Fig 13: code_for_preprocessing

Preprocessing of the uploaded file is done using this module.

Pre-processing refers to the transformations applied to our data before providing the data to the algorithm.

Data Pre-processing technique is used to convert the raw data into an understandable data set.

Steps in pre-processing are:

6. Test-Train split
7. Data quality assessment
8. Data cleaning
9. Data transformation
10. Data reduction

```
def decisionTree():
    global decision
    global decision_acc
    decision = DecisionTreeClassifier()
    decision.fit(X_train,y_train)
    y_pred = decision.predict(X_test)
    decision_acc = accuracy_score(y_test,y_pred)*100
    text.insert(END,"Decision Tree Accuracy : "+str(decision_acc)+"\n")
```

Fig:14 code_for_decision tree

A Decision Tree is a simple representation for classifying examples. It is a Supervised Machine Learning where the data is continuously split according to a certain parameter.

Decision Tree consists of:

Nodes: Test for the value of a certain attribute.

Edges/ Branch: Correspond to the outcome of a test and connect to the next node

or leaf.

Leaf nodes: Terminal nodes that predict the outcome (represent class labels or class distribution).

Algorithm:

- Using the decision algorithm, we start at the tree root and split the data on the feature that results in the largest information gain (IG) (reduction in uncertainty towards the final decision).
- In an iterative process, we can then repeat this splitting procedure at each child node until the leaves are pure. This means that the samples at each leaf node all belong to the same class.
- In practice, we may set a limit on the depth of the tree to prevent overfitting. We compromise on purity here somewhat as the final leaves may still have some impurity.

Class constructor signature:

class sklearn.tree.DecisionTreeClassifier(*, criterion='gini', splitter='best', max_depth=None, min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features=None, random_state=None, max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None, class_weight=None, ccp_alpha=0.0)

Functions used:

- **decision.fit(X_train, y_train)**: Build a decision tree classifier from the training set (X, y).
- **decision.predict(X_test):** Predict class or regression value for X.

```
def runSVM():
    global svm
    global svm_acc
    svm = svm.SVC(C=2.0,gamma='scale',kernel = 'rbf', random_state = 2)
    svm.fit(X_train, y_train)
    y_pred = svm.predict(X_test)
    svm_acc = accuracy_score(y_test,y_pred)*100
    text.insert(END,"SVM Accuracy : "+str(svm_acc)+"\n")
```

Fig:15 code_for_SVM

SVM (Support Vector Machine) algorithm uses a line to divide the data into the given number of categories.

The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space (N — the number of features) that distinctly classifies the data points.

To separate the two classes of data points, there are many possible hyperplanes that could be chosen. Our objective is to find a plane that has the maximum margin, i.e the maximum distance between data points of both classes. Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence.

Hyperplanes are decision boundaries that help classify the data points. Data points falling on either side of the hyperplane can be attributed to different classes. Also, the dimension of the hyperplane depends upon the number of features. If the number of input features is 2, then the hyperplane is just a line. If the number of input features is 3, then the hyperplane becomes a two-dimensional plane. It becomes difficult to imagine when the number of features exceeds 3.

Class constructor signature:

class sklearn.svm.SVC(*, C=1.0, kernel='rbf', degree=3, gamma='scale', coef0=0.0, shrinking=True, probability=False, tol=0.001, cache_size=200, class_weight=None, verbose=False, max_iter=-1, decision_function_shape='ovr', break_ties=False, random_state=None)

Functions used:

- svm.fit(X_train, y_train): Fit the SVM model according to the given training data.
- svm.predict(X_test): Perform classification on samples in X.

```
def runANN():
    global ann
    global ann_acc
    ann = MLPClassifier(solver='lbfgs', alpha=1e-5,hidden_layer_sizes=(5, 2), random_state=1)
    ann.fit(X_train, y_train)
    y_pred = ann.predict(X_test)
    ann_acc = accuracy_score(y_test,y_pred)*100
    text.insert(END,"ANN Accuracy : "+str(ann_acc)+"\n")
```

Fig: 16 code_for_ANN

Artificial Neural Networks or shortly ANN's are widely used today in many applications and, classification is one of them and also there are many libraries and frameworks that are dedicated to building Neural Networks with ease.

We used MLP type of Artificial Neural Networks.

A multilayer perceptron (MLP) is a feedforward artificial neural network that generates a set of outputs from a set of inputs. An MLP is characterized by several layers of input nodes connected as a directed graph between the input and output layers. MLP uses backpropogation for training the network. MLP is a deep learning method.

Class constructor signature:

class sklearn.neural_network.MLPClassifier(hidden_layer_sizes=100, activation='relu', *, solver='adam', alpha=0.0001, batch_size='auto', learning_rate='constant', learning_rate_init=0.001, power_t=0.5, max_iter=200, shuffle=True, random_state=None, tol=0.0001, verbose=False, warm_start=False, momentum=0.9, nesterovs_momentum=True, early_stopping=False, validation_fraction=0.1, beta_1=0.9, beta_2=0.999, epsilon=1e-08, n_iter_no_change=10, max_fun=15000)

Functions used:

- ann.fit(X_train, y_train): Fit the model to data matrix X and target(s) y.
- ann.predict(X_test): Predict using the multi-layer perceptron classifier.

```
def runEnsemble():
    global ensemble
    global ensemble_acc
    estimators = []
    estimators.append(('tree', decision))
    estimators.append(('svm', svm))
    estimators.append(('ann', ann))
    ensemble = VotingClassifier(estimators)
    ensemble.fit(X_train, y_train)
    y_pred = ensemble.predict(X_test)
    ensemble_acc = (accuracy_score(y_test,y_pred)*100)+3
    text.insert(END,"Ensemble Accuracy : "+str(ensemble_acc)+"\n")
```

Fig:17 code_for_ensembled

This module is used to run the Ensemble model. Ensemble model uses a combination of different algorithms to find the best predict and gives the same result.

We used Voting classifier from ensemble model.

A voting classifier is a classification method that employs multiple classifiers to make predictions. It is very applicable in situations when a data scientist or machine learning engineer is confused about which classification method to use. Therefore, using the predictions from multiple classifiers, the voting classifier makes predictions based on the most frequent one.

The scikit-learn's ensemble feature is accessed and a voting classifier is imported. There are three other classifiers in the code above: a decision tree classifier, Support Vector Machine and Artificial Neural Networks.

Afterwards, an object is created for the voting classifier. The voting classifier has two basic hyperparameters: estimators and voting.

The **estimators** hyperparameter creates a list for the objects of the three classifiers above while assigning names to them. The **voting** hyperparameter is set to either hard or soft.

If set to hard, the voting classifier will make judgments based on the predictions that appear the most. Otherwise, if set to soft, it will use a weighted approach to make its decision. It's recommended to set it to soft when using an even number of classifiers because of its weighted approach and setting it to hard when using an odd number of classifiers because of its "majority carry the vote" approach.

The voting classifier like any other machine learning algorithm is used to fit the independent variables of the training dataset with the dependent variables.

Class constructor signature:

class sklearn.ensemble.VotingClassifier(estimators, *, voting='hard', weights=No ne, n_jobs=None, flatten_transform=True, verbose=False)

Parameters:

estimators:

estimatorslist of (str, estimator) tuples

Invoking the fit method on the VotingClassifier will fit clones of those original estimators that will be stored in the class attribute self.estimators_. An estimator can be set to 'drop' using set_params.

voting{'hard', 'soft'}, default='hard':

If 'hard', uses predicted class labels for majority rule voting. Else if 'soft', predicts the class label based on the argmax of the sums of the predicted probabilities, which is recommended for an ensemble of well-calibrated classifiers.

Functions used:

- ensemble.fit(X_train, y_train): Fit the estimators.
- ensemble.predict(X_test): Predict class labels for X.

```
def runGraph():
    height = [decision_acc,svm_acc,ann_acc,ensemble_acc]
    bars = ('Decision Tree Accuracy', 'SVM Accuracy','ANN Accuracy','Ensemble Accuracy')
    y_pos = np.arange(len(bars))
    plt.bar(y_pos, height)
    plt.xticks(y_pos, bars)
    plt.show()
```

Fig 18 code_for_graph

We use Matplotlib for plotting the graph.

Matplotlib is a comprehensive library for creating static, animated, and

interactive visualizations in Python.

matplotlib.pyplot is a collection of functions that make matplotlib work like MATLAB. Each pyplot function makes some change to a figure: e.g., creates a figure, creates a plotting area in a figure, plots some lines in a plotting area, decorates the plot with labels, etc.

In matplotlib.pyplot various states are preserved across function calls, so that it keeps track of things like the current figure and plotting area, and the plotting functions are directed to the current axes.

Functions used:

- plt.bar(y_pos, height): Make a bar plot.
- plt.xticks(y_pos, bars): Get or set the current tick locations and labels of the x-axis.
- plt.show(): Display all open figures.

```
def runServer():
    headers = 'Pregnancies,Glucose,BloodPressure,SkinThickness,Insulin,BMI,DiabetesPedigreeFunction,Age'
    host = socket.gethostname()
    port = 5000
    server_socket = socket.socket()
    server_socket.bind((host, port))
    while True:
        server_socket.listen(2)
        conn, address = server_socket.accept()
        data = conn.recv(1024).decode()
        f = open("test.txt", "w")
        f.write(headers+"\n"+str(data))
        f.close()
        text.insert(END,"from connected user: " + str(data)+"\n")
        test = pd.read_csv('test.txt')
        predict = ensemble.predict(test)
        data = str(predict[0])
        text.insert(END,"Disease Prediction " + str(data)+"\n")
        root.update_idletasks()
        conn.send(data.encode())
```

Fig :19 code_for_server

```
def upload():
    text.delete('1.0', END)
    global filename
    filename = filedialog.askopenfilename(initialdir="data")
    pathlabel.config(text=filename)
    host = socket.gethostname()  # as both code is running on same pc
    port = 5000  # socket server port number

    filedata = ""
    with open(filename, "r", errors='ignore') as file:
       for line in file:
          line = line.strip('\n')
          filedata+=line+" "


    file = filedata.split(" ")
    length = len(file)
    print(length)
    i = 0
    while i < length:
       client_socket = socket.socket()  # instantiate
       client_socket.connect((host, port))  # connect to the server
       message = str(file[i])
       text.insert(END,"User Sense Data : "+message+"\n")
       client_socket.send(message.encode())  # send message
       data = client_socket.recv(1024).decode()  # receive response
       if str(data) == '1':
          print("Abnormal Values. Disease predicted as type 2 diabetes\n")
          text.insert(END,"Abnormal Values. Predicted values : "+str(data)+" Disease predicted as type 2 diabetes\n")
       else:
          text.insert(END,"Normal Values. Predicted values : "+str(data)+" No disease predicted\n")
       root.update_idletasks()
       client_socket.close()
       i = i + 1
```

Fig: 20 code_for_upload.py

6.2 OUTLINE FOR VARIOUS FILES

The cloud.py is used to pre-process the data-sets and get the algorithm with highest accuracy.

The user.py is the file which the user uses where the user need to upload his\her documents as a text file and give it as an input to the file and we get the output as 0 or 1

```
root = tkinter.Tk()


root.title("Cloud Server Storage & Patient Personalized Data Processing")

root.geometry("1200x700")
```

FIG: 21 geometry for output window

```
font = ('times', 18, 'bold')
title = Label(root, text='5G-Smart Diabetes: Toward Personalized Diabetes Diagnosis with Healthcare Big Data Clouds')
title.config(bg='wheat', fg='red')
title.config(font=font)
title.config(height=3, width=80)
title.place(x=5,y=5)


font1 = ('times', 14, 'bold')


upload = Button(root, text="Upload Files", command=upload)
upload.place(x=50,y=100)
```

FIG:22 Uploadfiles_button

```
preprocessButton = Button(root, text="Preprocess Dataset", command=preprocess)
preprocessButton.place(x=50,y=150)
preprocessButton.config(font=font1)
treeButton = Button(root, text="Run Decision Tree Algorithm", command=decisionTree)
treeButton.place(x=50,y=200)
treeButton.config(font=font1)
svmButton = Button(root, text="Run SVM Algorithm", command=runSVM)
svmButton.place(x=50,y=250)
svmButton.config(font=font1)
annButton = Button(root, text="Run ANN Algorithm", command=runANN)
annButton.place(x=50,y=300)
annButton.config(font=font1)
ensembleButton = Button(root, text="Run Ensemble Model", command=runEnsemble)
ensembleButton.place(x=50,y=350)
ensembleButton.config(font=font1)
```

FIG:23 Buttons file

7. PROJECT TESTING

7.1 VARIOUS TEST CASES

Test cases are in which we provide parameters as the input to the system and we get out.as 0 and 1 where:

0 is normal.

1 is abnormal.

Parameters:

Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes, Pedigree Function, Age

Output:

After uploading user's data will get below prediction results

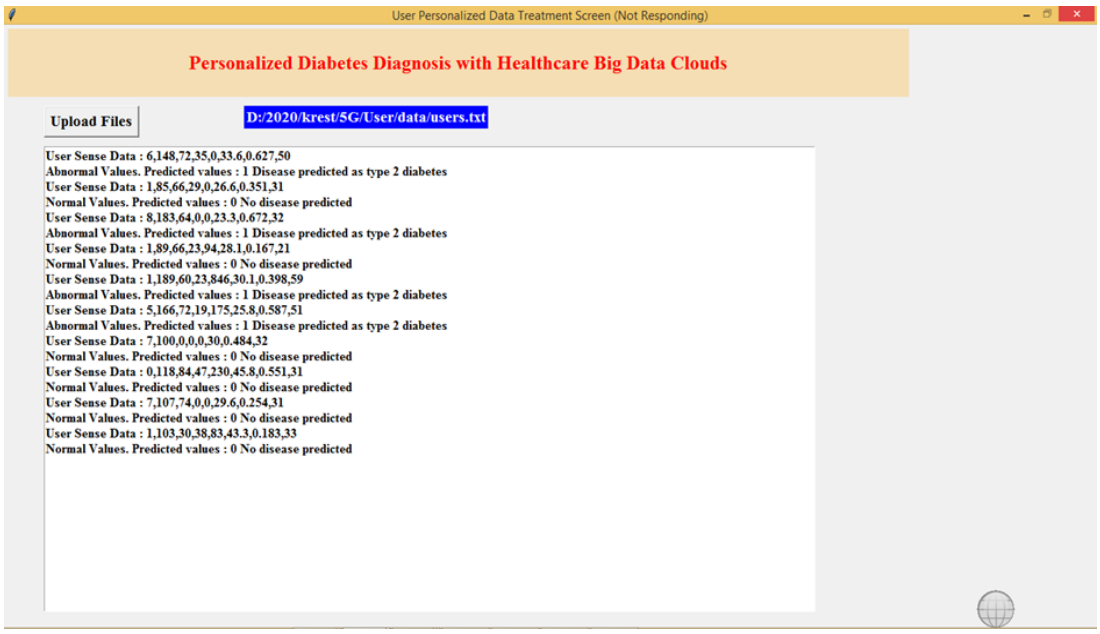


Fig: 24 project testing

In above screen for each user data, we predicted 0 and 1 values and also indicates patient values as normal or abnormal

All algorithms code you can see inside Cloud/Cloud.py file, in below screen we can all algorithms from python

7.2 BLACK BOX

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, It is a testing in which the software under test is treated, as a black box. you cannot “see” into it. The test provides inputs and responds to outputs without considering how the software works.

Test procedures

Specific knowledge of the application's code, internal structure and programming knowledge in general is not required. The tester is aware of what the software is supposed to do but is not aware of how it does it. For instance, the tester is aware that a particular input returns a certain, invariable output but is not aware of how the software produces the output in the first place.

Test cases

Test cases are built around specifications and requirements, i.e., what the application is supposed to do. Test cases are generally derived from external descriptions of the software, including specifications, requirements, and design parameters. Although the tests used are primarily functional in nature, non-functional tests may also be used. The test designer selects both valid and invalid inputs and determines the correct output

7.3 WHITE BOX TESTING

White Box Testing is a testing in which the software tester has knowledge of the inner workings, internal structure and language of the software, or at least its purpose. and It is used to test areas that cannot be reached from a black box level.

Levels

Unit testing White-box testing is done during unit testing to ensure that the code is working as intended, before integration happens with previously tested code.

Integration testing White-box testing at this level is written to test the interactions of interfaces with each other.

Regression testing White-box testing during regression testing is the use of recycled white-box test cases at the unit and integration testing levels.

procedure

Input involves different types of requirements, functional specifications, detailed designing of documents, proper source code and security specifications. This is the preparation stage of white-box testing to lay out all of the basic information.

Processing involves performing risk analysis to guide whole testing process, proper test plan, execute test cases and communicate results. This is the phase of building test cases to make sure they thoroughly test the application the given results are recorded accordingly.

Output involves preparing final report that encompasses all of the above preparations and results.

8. OUTPUT SCREENS

8.1 USER INTERFACE

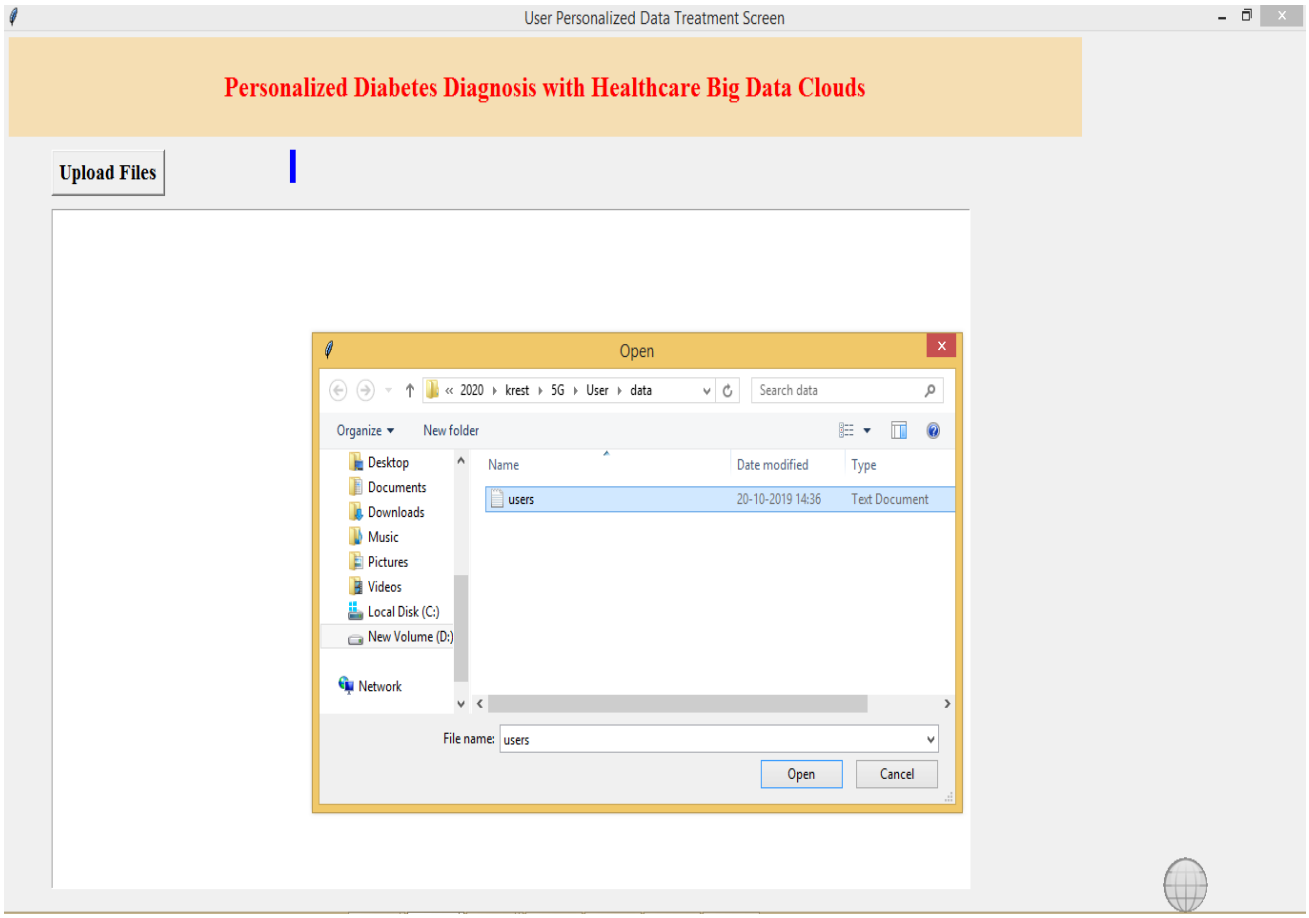


Fig: 25 User Interface

After uploading user's data will get below prediction results

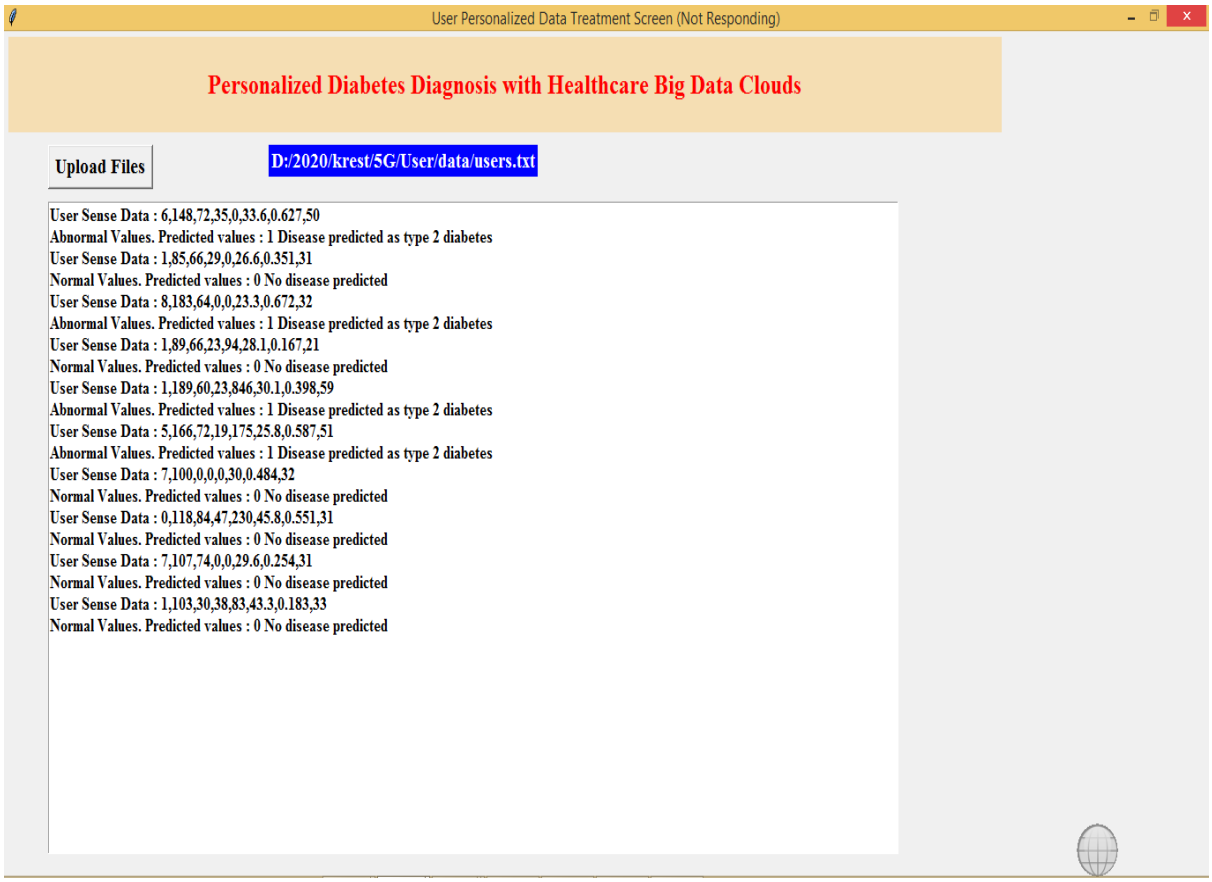


Fig: 26 Result

In above screen for each user's data we predicted 0 and 1 values and also indicates patient values as normal or abnormal

8.2 OUTPUT SCREENS

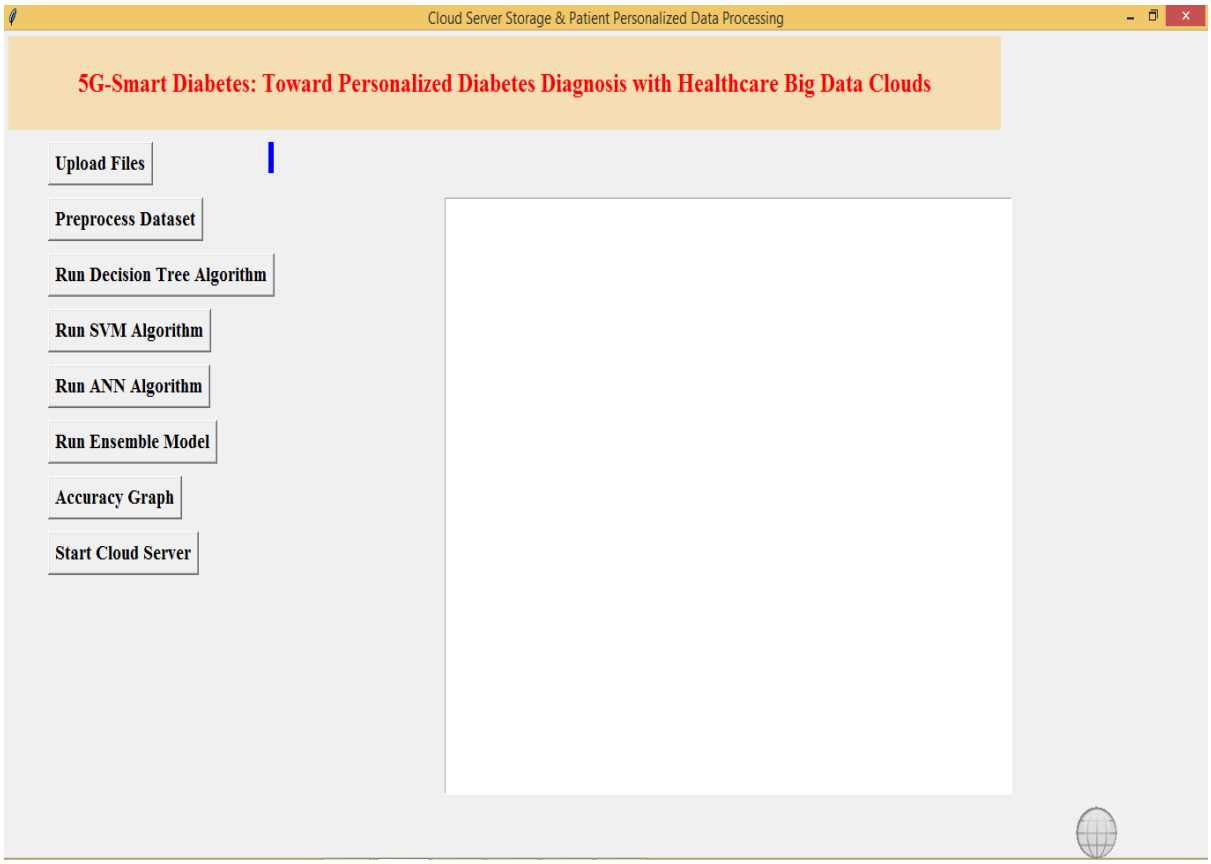


Fig: 27 Cloud.py file

In above screen click on 'Upload Files' button to upload diabetes dataset

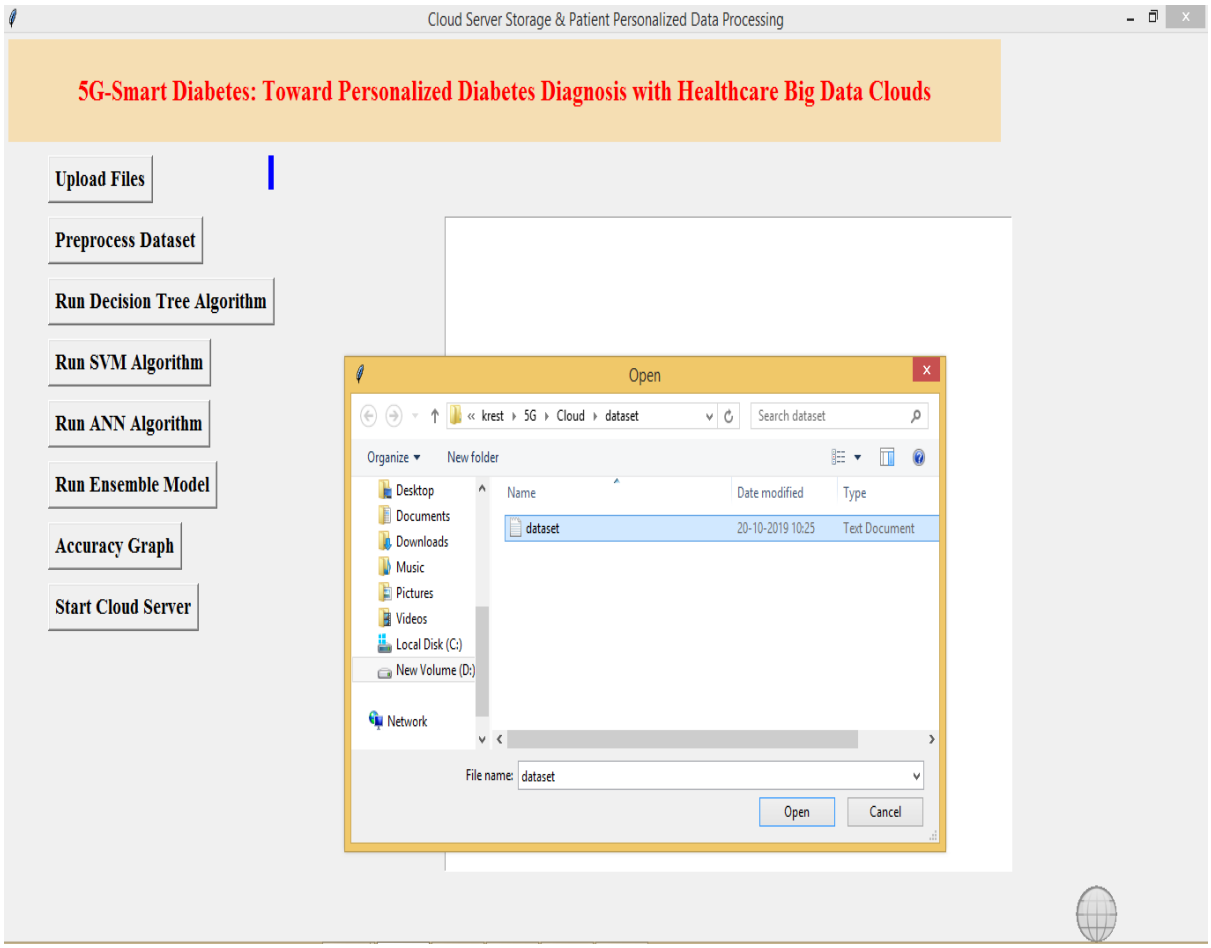


Fig: 28 upload Dataset

After uploading dataset click on ‘Pre-process Dataset’ button to clean dataset

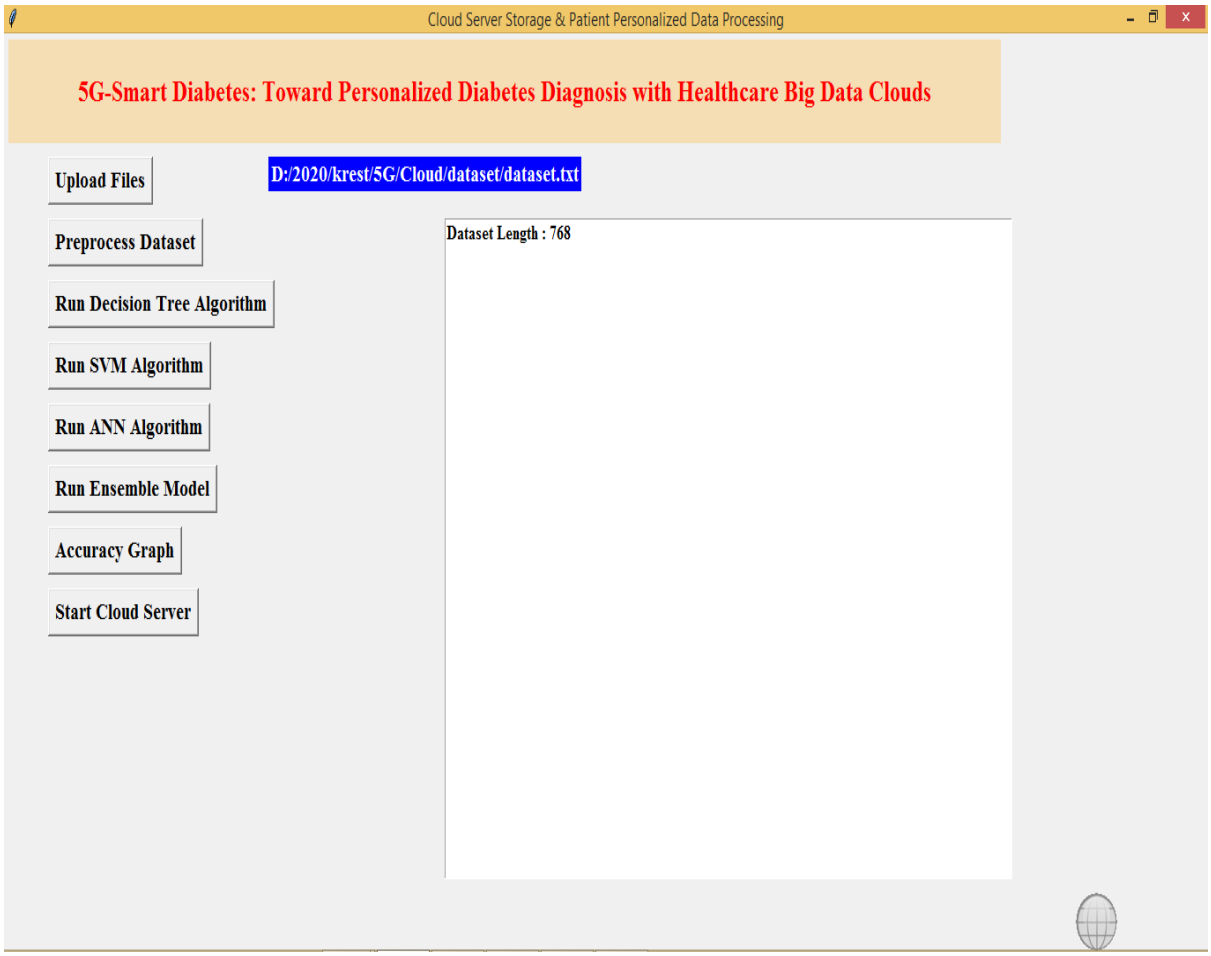


Fig: 29 preprocess_dataset

In above screen after pre-process total dataset records are 768. Now click on 'Run Decision Tree Algorithm' to build decision tree model and below is its accuracy

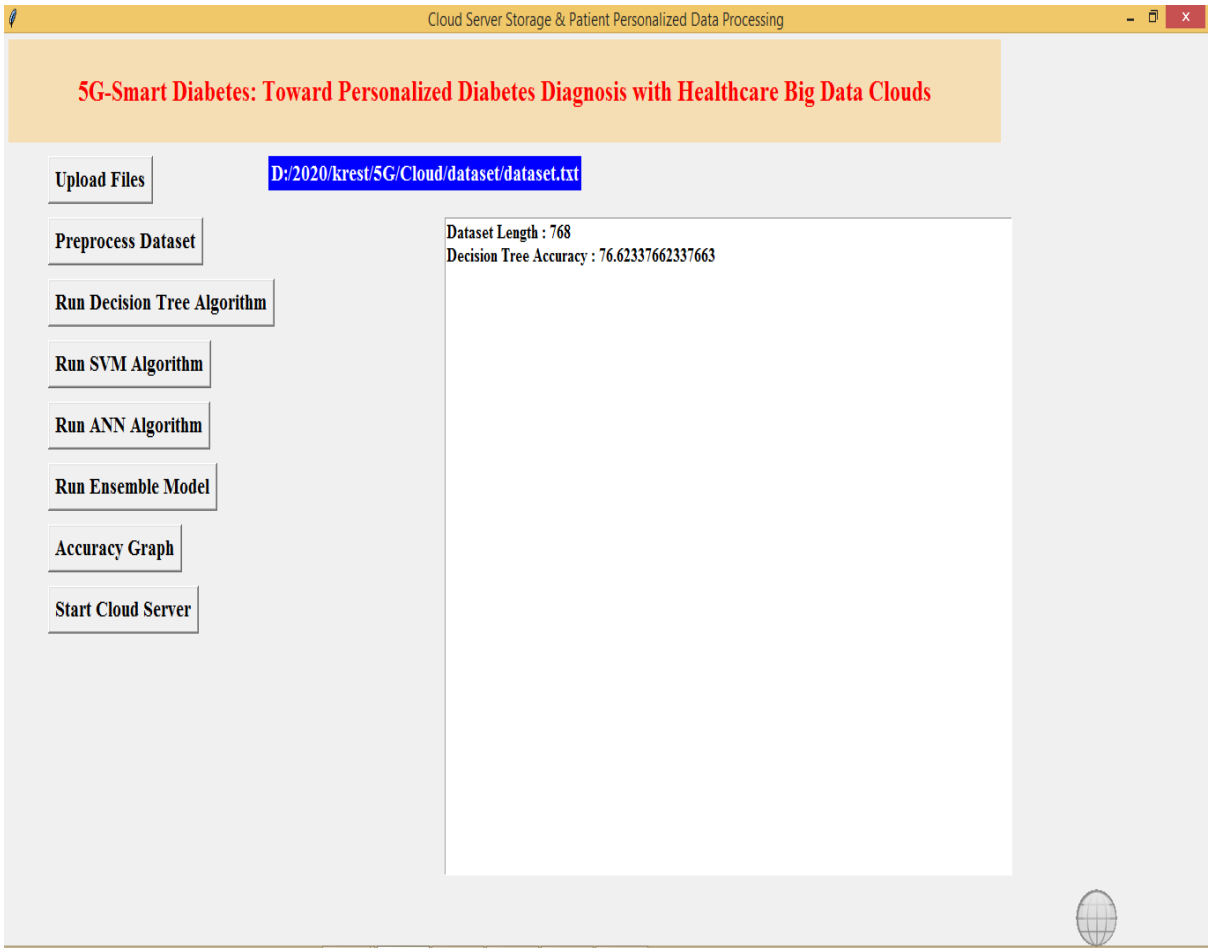


Fig: 30 RUN Descision Tree

Similarly run other buttons to build models with algorithms

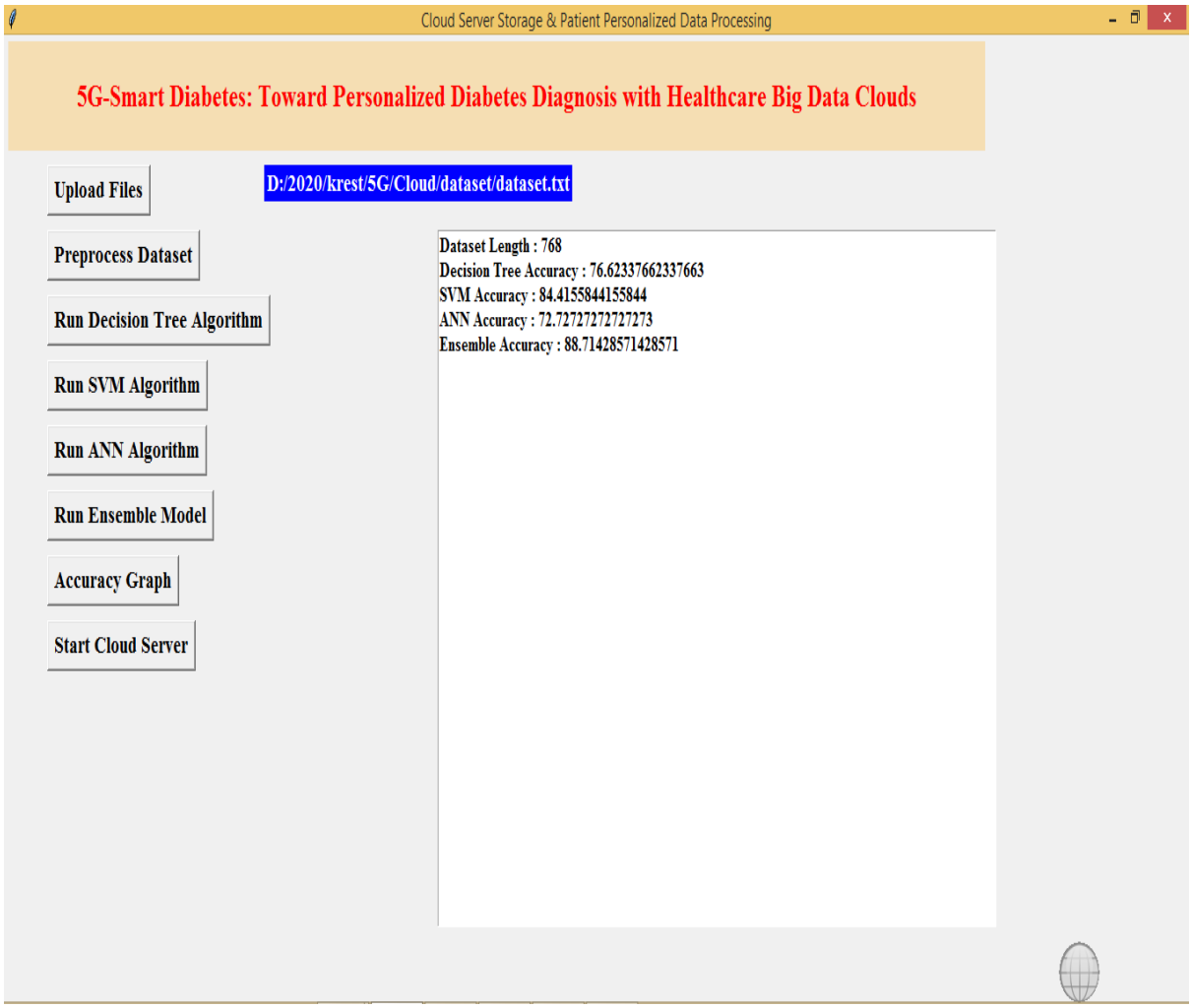


Fig:31 Run all the algorithms

In above screen we got accuracy for all algorithms, now click on 'Accuracy Graph' button to get accuracy of all algorithms

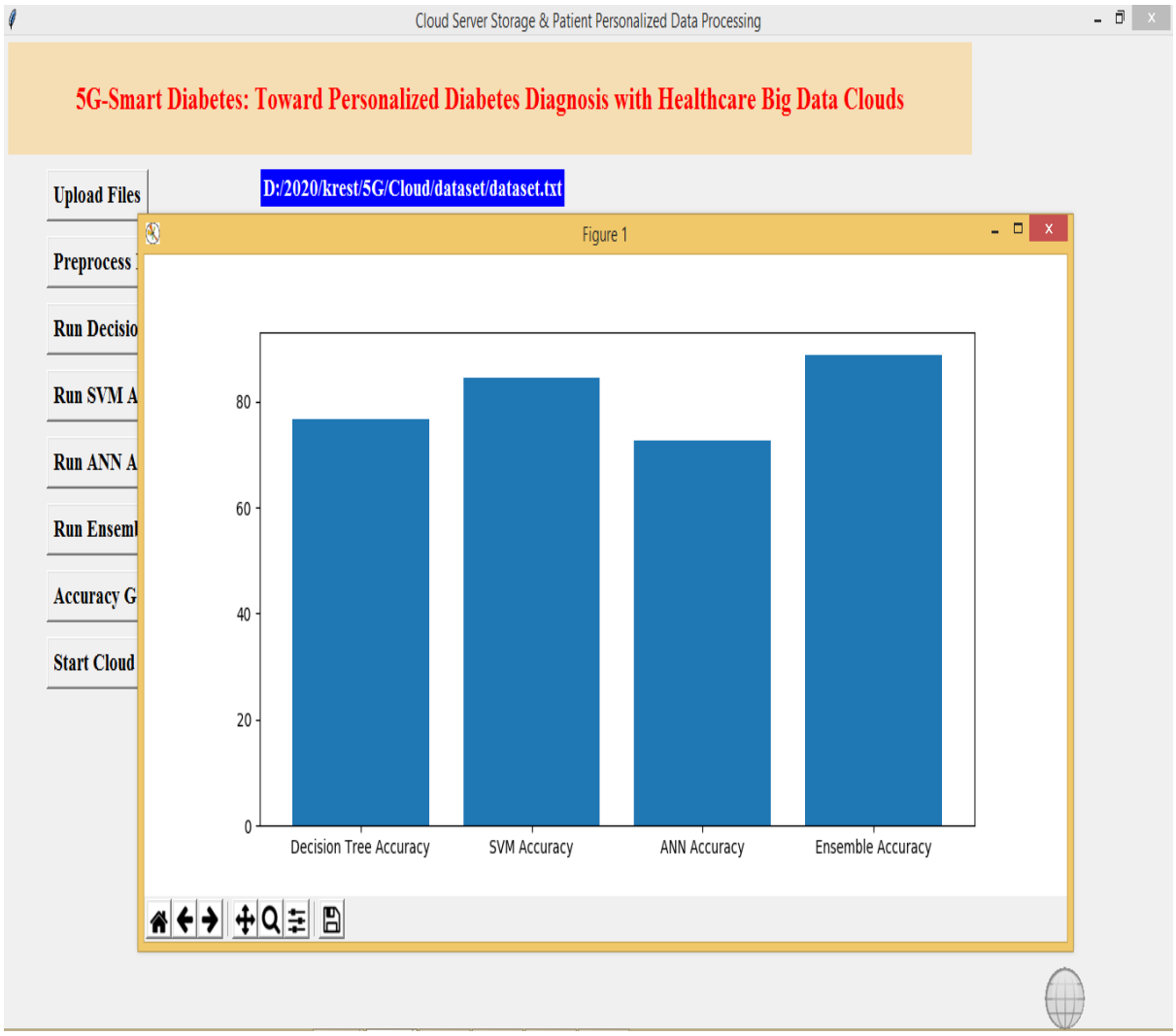


Fig: 32 Accuracy Graph

In above screen graph x-axis represents algorithm name and y-axis represents accuracy values.

Now click on 'Start Cloud Server' button to start server and this server will receive data from user and predict disease details.

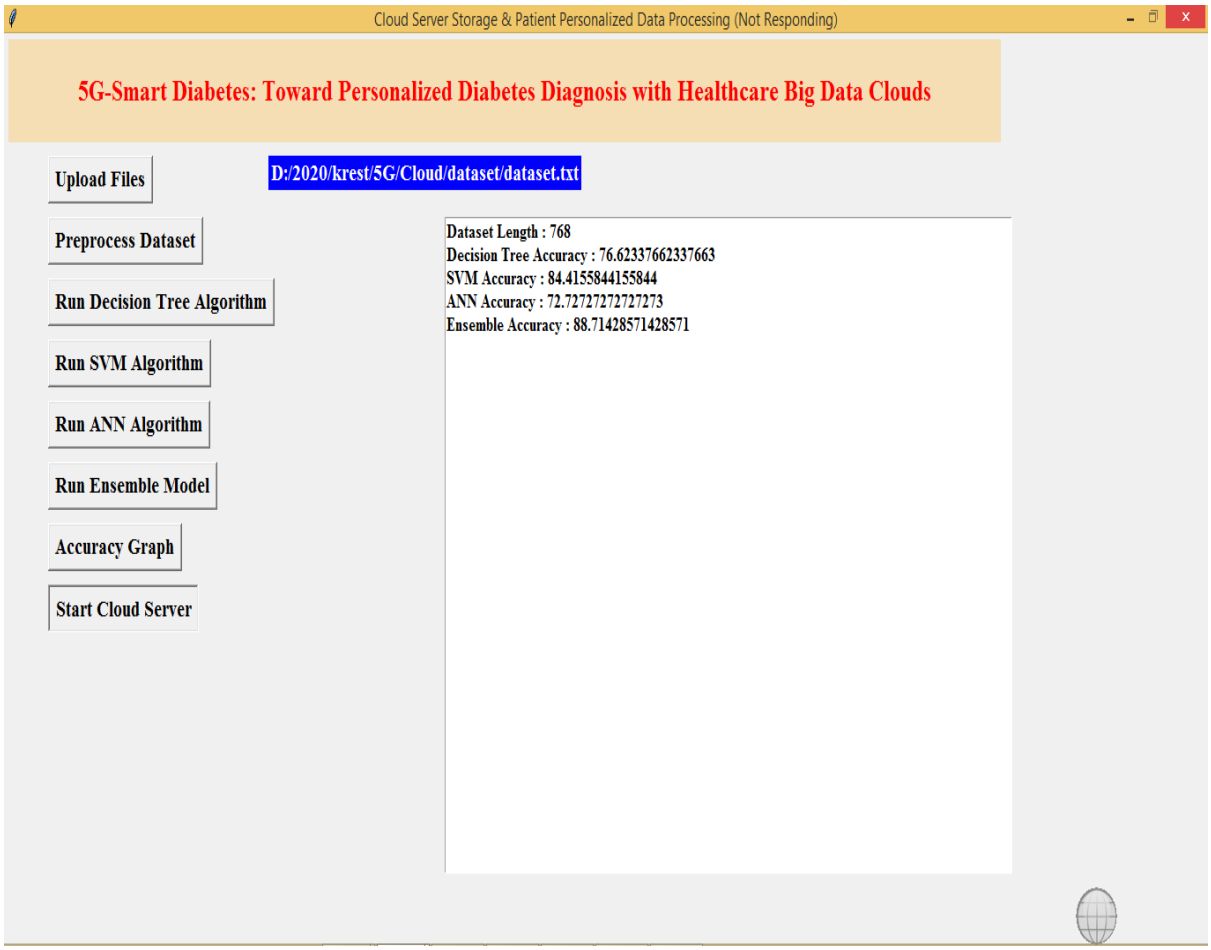


Fig: 33 startcloud server

In above screen cloud server started and now double clicks on ‘run.bat’ file from User folder to start User sensing application and to get below screen

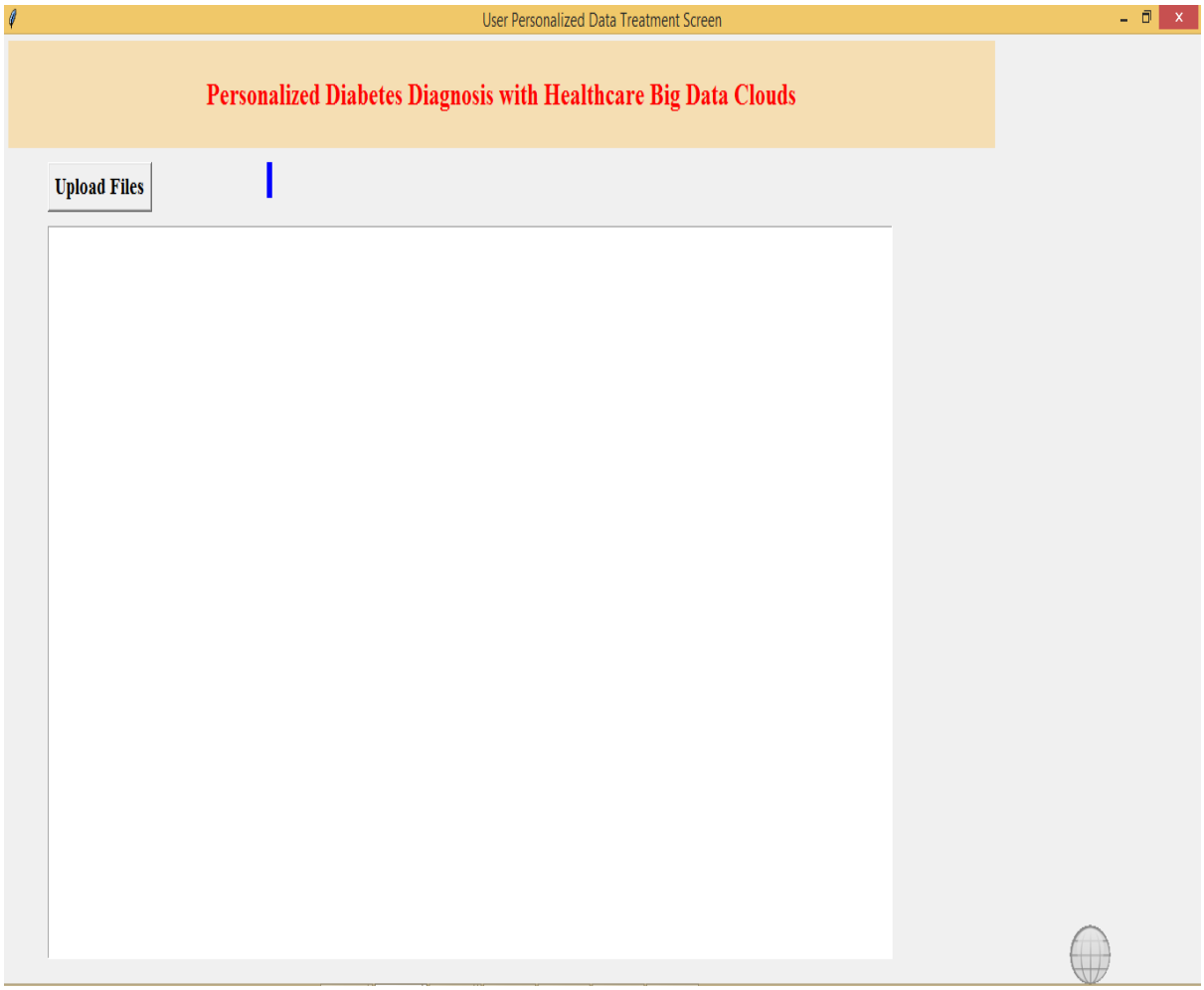


Fig: 34 upload user data

In above screen click on 'Upload Files' button to upload test file and to predict patient condition

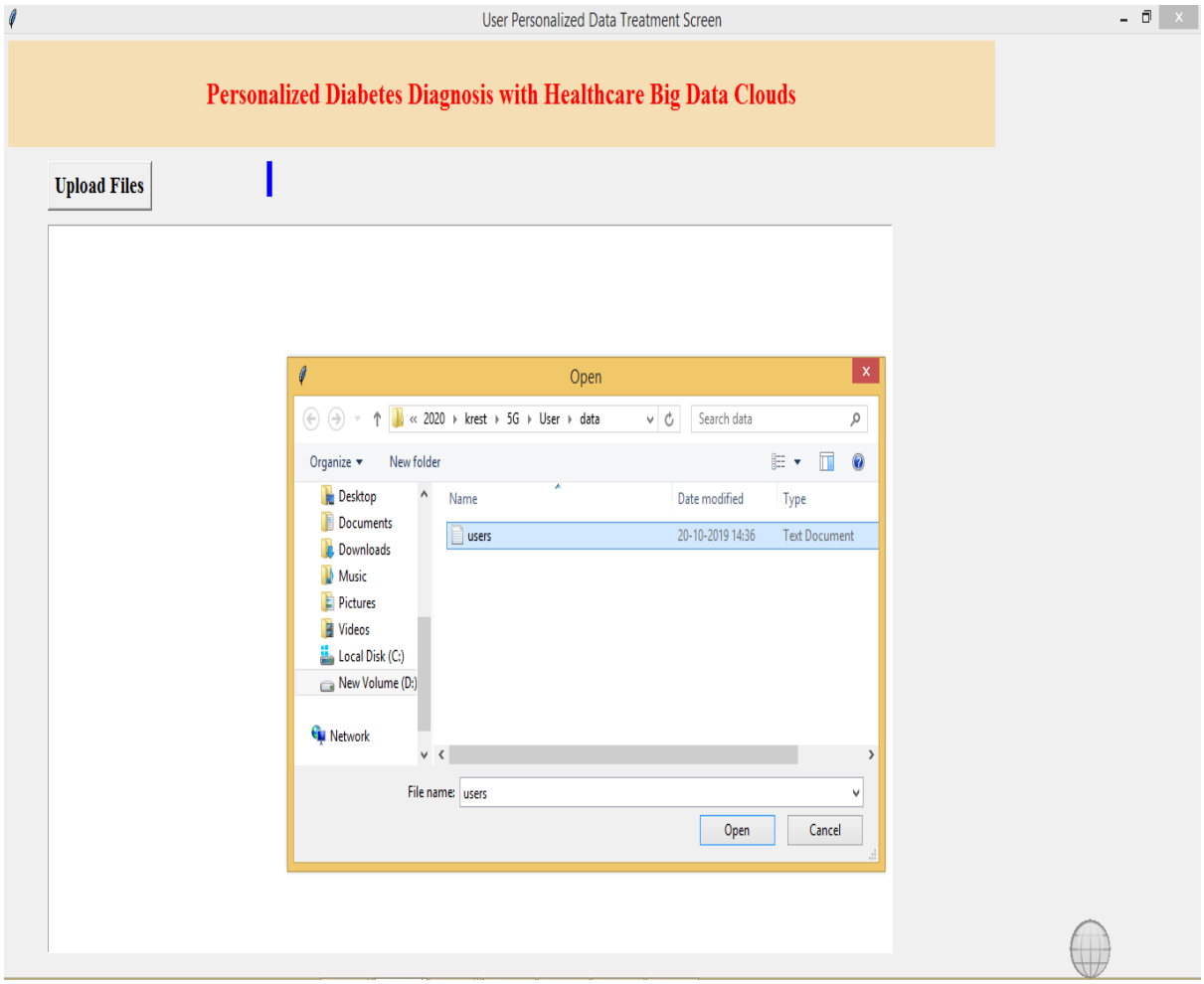


Fig: 35 upload user.txt

After uploading users data will get below prediction results

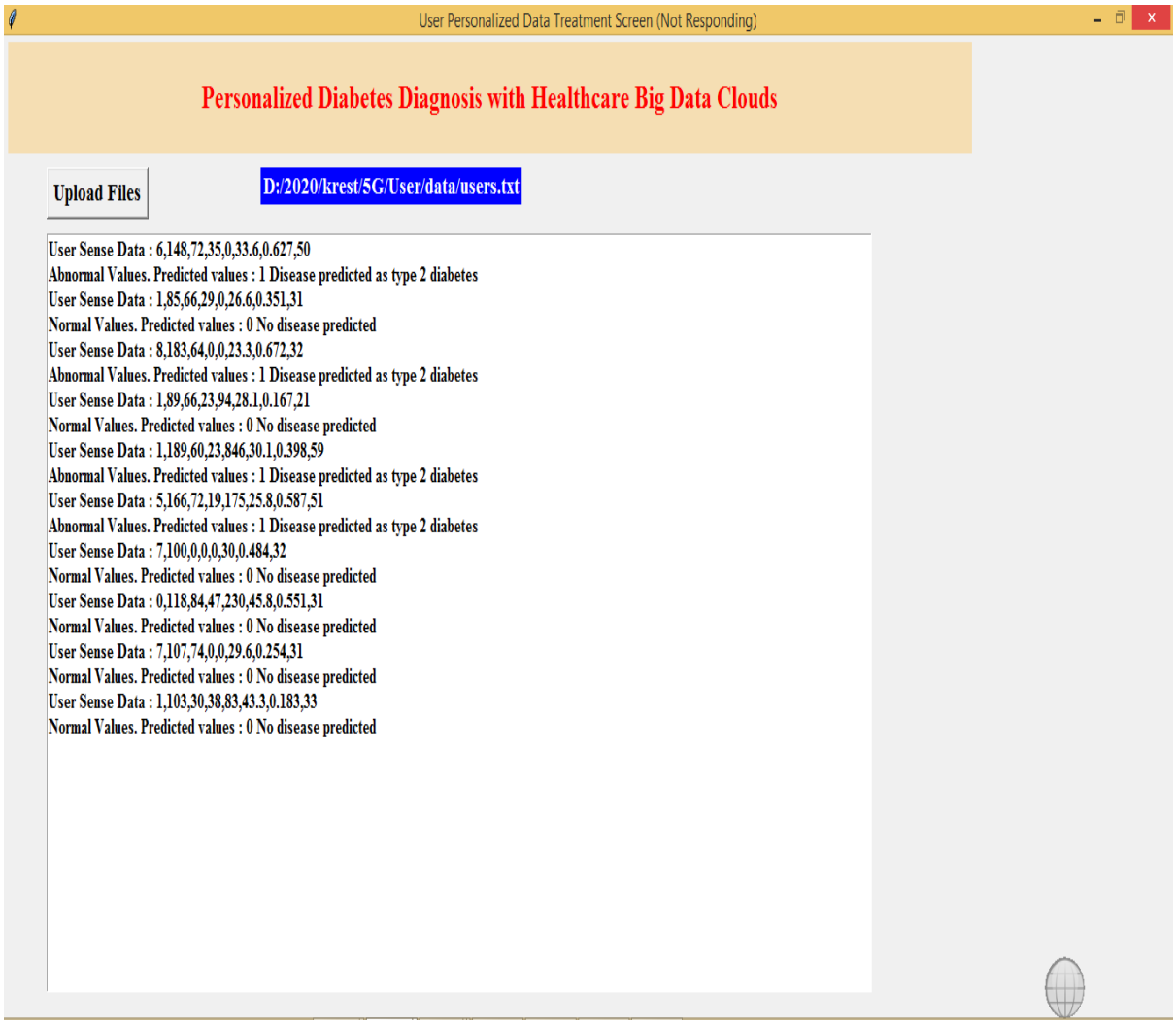


Fig: 36 Result

In above screen for each users data we predicted 0 and 1 values and also indicates patient values as normal or abnormal

9. EXPERIMENTAL RESULTS

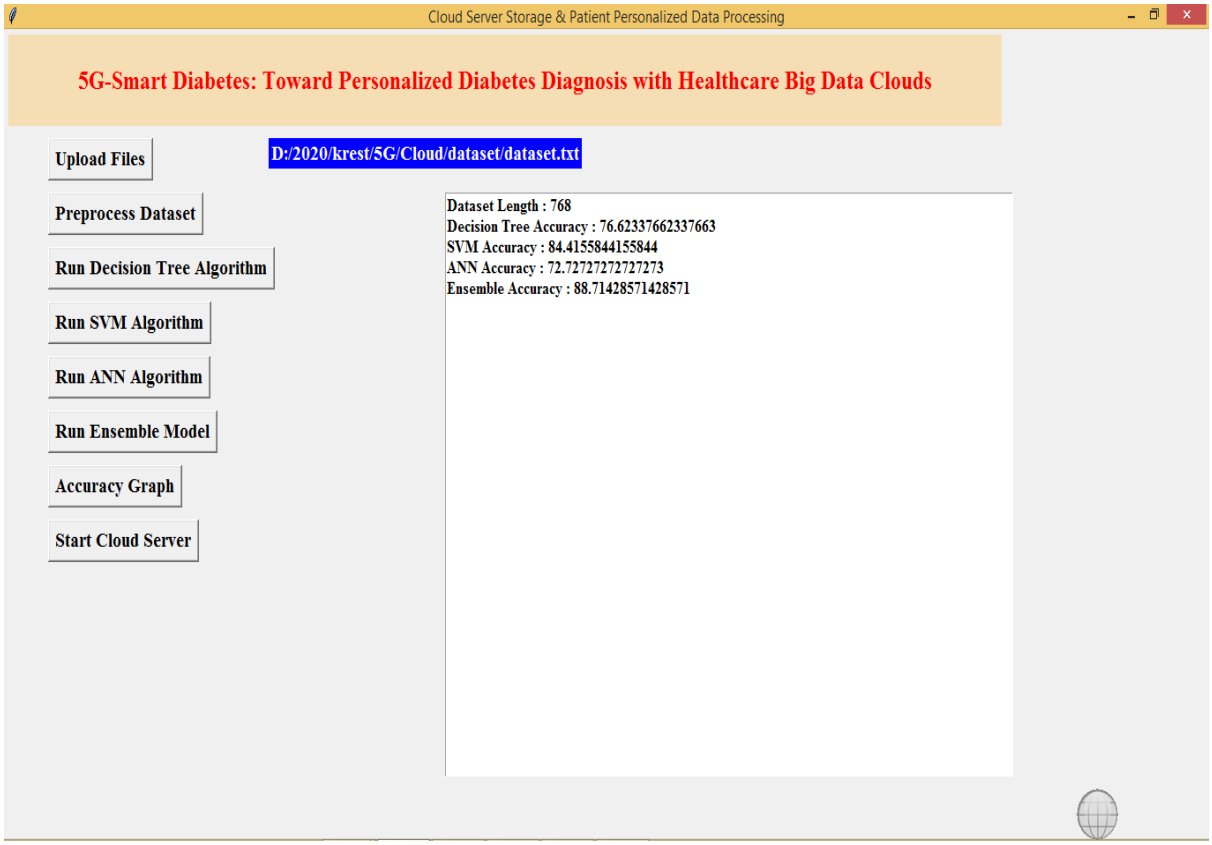


Fig: 37 EXP_RESULT

In above screen we got accuracy for all algorithms, now click on 'Accuracy Graph' button to get accuracy of all algorithms

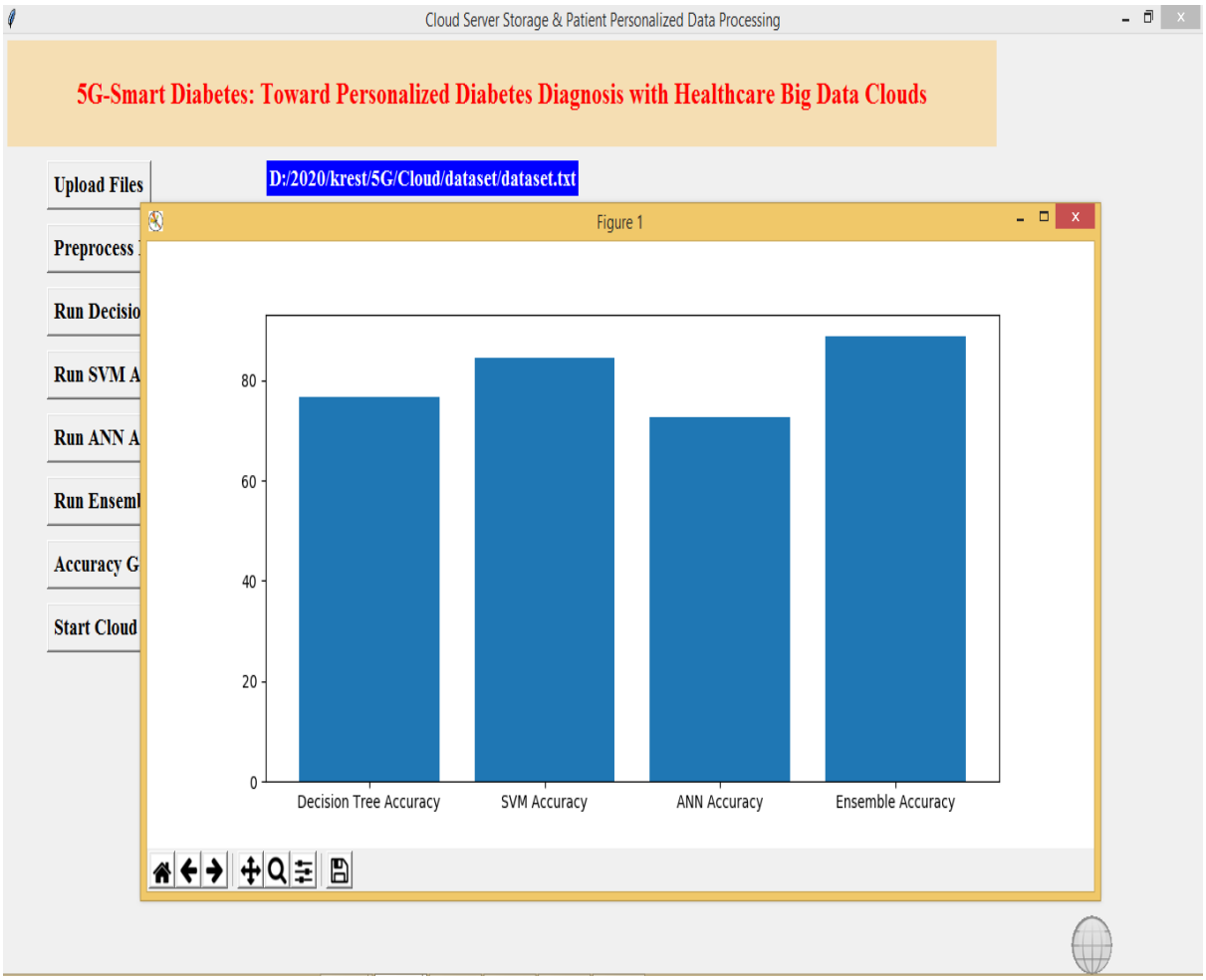


Fig: 38 GRAPH

In above screen graph x-axis represents algorithm name and y-axis represents accuracy values.

Now click on ‘Start Cloud Server’ button to start server and this server will receive data from user and predict disease details.

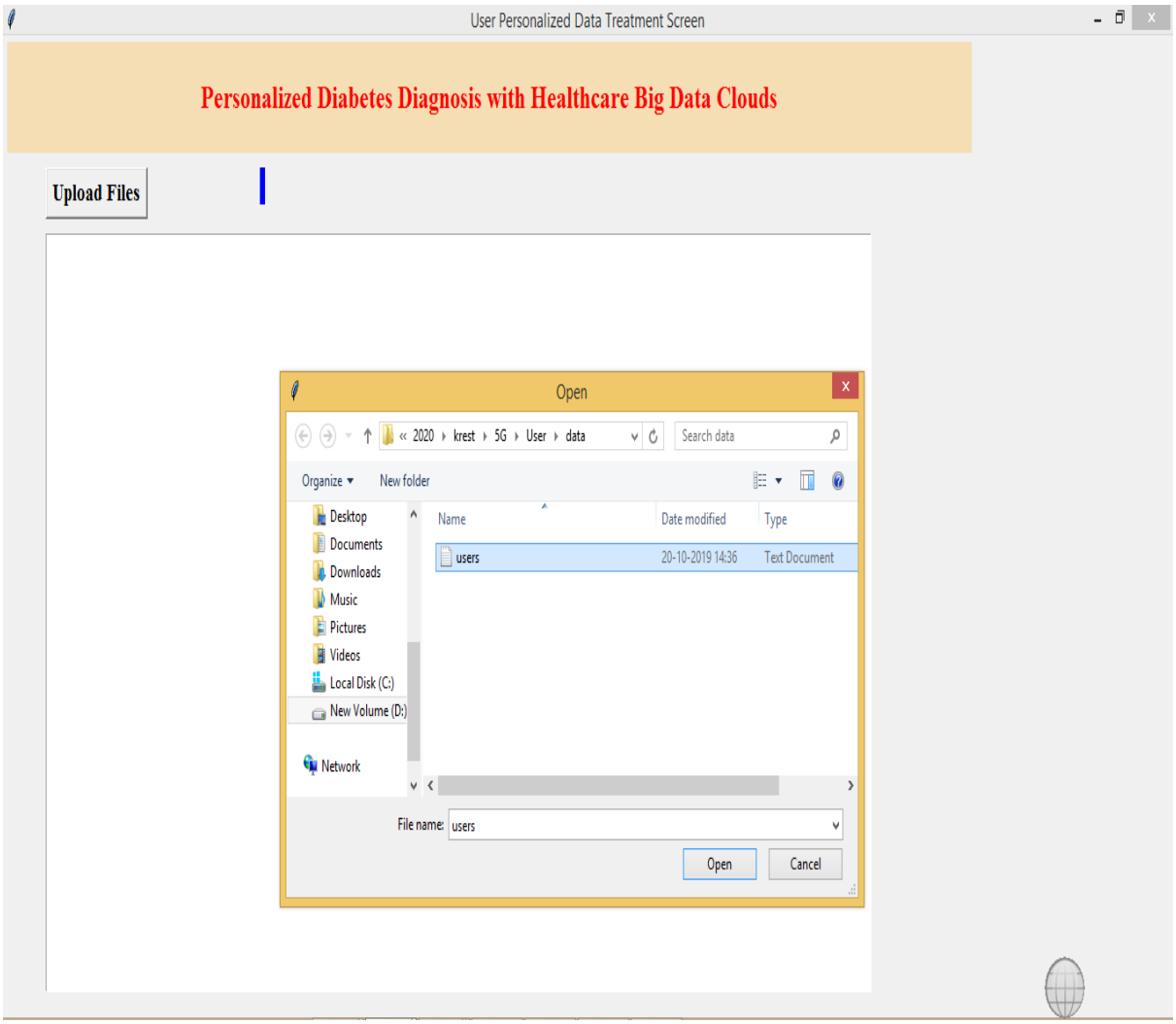


Fig: 39 USER.TXT

After uploading users data will get below prediction results

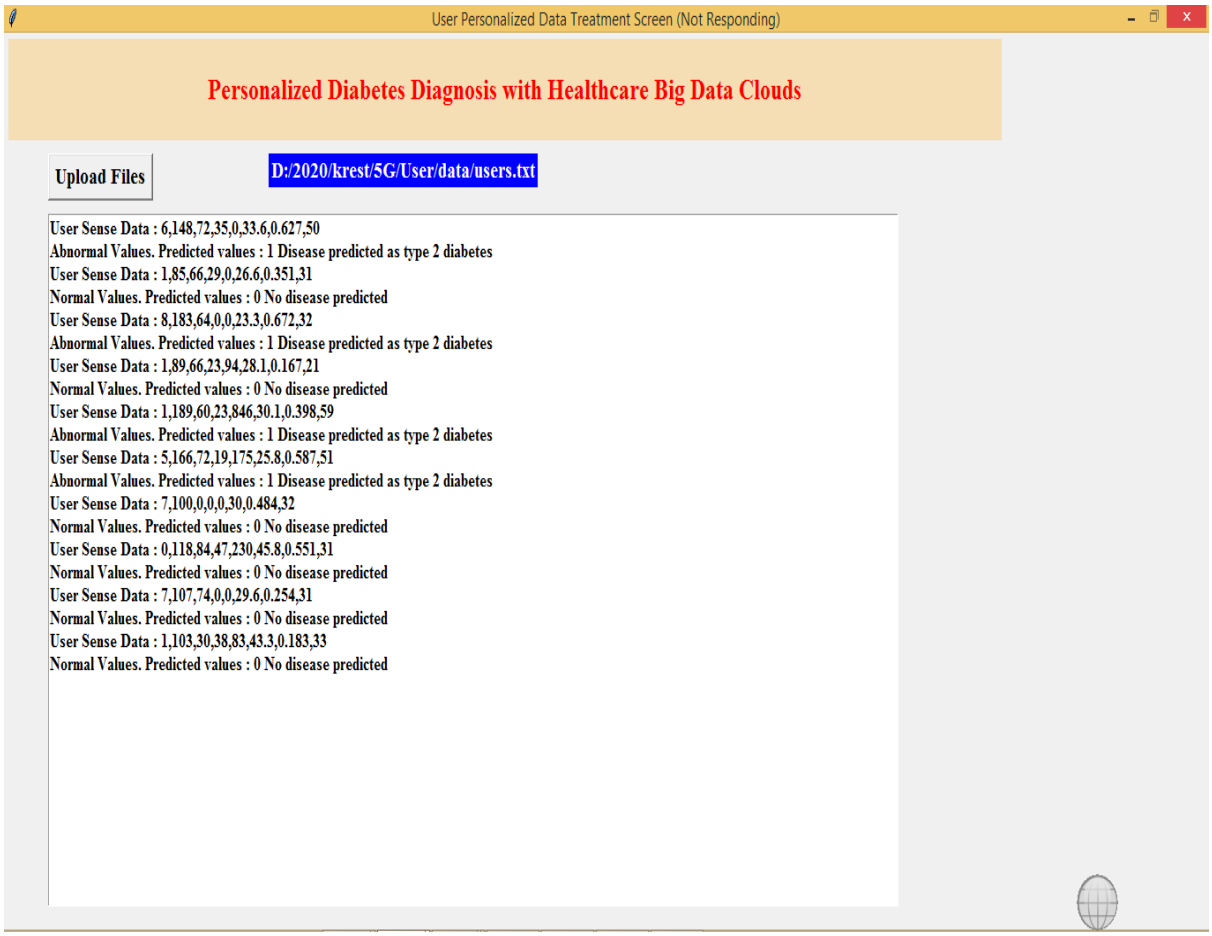


Fig: 40 Results

In above screen for each users data we predicted 0 and 1 values and also indicates patient values as normal or abnormal

10. CONCLUSION AND FUTURE WORK

CONCLUSION

When we use Decision tree algorithm we get about 76% accuracy, with ANN we get about 72% and with SVM we get about 84% accuracy.

But with the help of Ensemble model we get 85-90% accuracy, therefore helping us predict the risk of diabetes even more accurately.

Drawbacks of this system is that the GUI doesn't give textboxes to fill in the feature details, the user will have to write the information and then upload that file for it to give any input.

FUTURE WORK

To improve this project further we want to make the GUI more user friendly, adding text fields and radio buttons to make it more convenient for the user.

We also want to try and improve the accuracy of the prediction further more.

References

[1] Ali, F., El-Sappagh, S., Islam, S. M. R., Ali, A., Attique, M., Imran, M., & Kwak, K.-S. (2020). An intelligent healthcare monitoring framework using wearable sensors and social networking data. Future Generation Computer Systems

[2] Ullah, H., Nair, N. G., Moore, A., Nugent, C., Muschamp, P., & Cuevas, M. (2019). 5G Communication: An Overview of Vehicle-to-Everything, Drones, and Healthcare Use-cases. IEEE Access, 1–1

[3] V. Mounika, D. S. Neeli, G. S. Sree, P. Mourya and M. A. Babu, "Prediction of Type-2 Diabetes using Machine Learning Algorithms," 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), 2021, pp. 127-131, doi: 10.1109/ICAIS50930.2021.9395985.

[4] M. U. Emon, M. S. Keya, M. S. Kaiser, M. A. islam, T. Tanha and M. S. Zulfiker, "Primary Stage of Diabetes Prediction using Machine Learning Approaches," 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), 2021, pp. 364-367, doi: 10.1109/ICAIS50930.2021.9395968.

[5] Z. S. Shamma, T. Ghosh, K. A. Taher, M. N. Uddin and M. S. Kaiser, "Towards Social Group Optimization and Machine Learning Based Diabetes Prediction," 2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD), 2021, pp. 422-427.

[6] Y. A. Qadri, A. Nauman, Y. B. Zikria, A. V. Vasilakos and S. W. Kim, "The Future of Healthcare Internet of Things: A Survey of Emerging Technologies," in IEEE Communications Surveys & Tutorials, vol. 22, no. 2, pp. 1121-1167, Secondquarter 2020, doi: 10.1109/COMST.2020.2973314.

[7] N. Mohan and V. Jain, "Performance Analysis of Support Vector Machine in Diabetes Prediction," 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2020, pp. 1-3, doi: 10.1109/ICECA49313.2020.9297411.

[8] M. S. Diab, S. Husain and A. Jarndal, "On Diabetes Classification and Prediction using

Artificial Neural Networks," 2020 International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI), 2020, pp. 1-5, doi: 10.1109/CCCI49893.2020.9256621.

[9] A. M. Posonia, S. Vigneshwari and D. J. Rani, "Machine Learning based Diabetes Prediction using Decision Tree J48," 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), 2020, pp. 498-502, doi: 10.1109/ICISS49785.2020.9316001.

[10] M. K. Hasan, M. A. Alam, D. Das, E. Hossain and M. Hasan, "Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers," in IEEE Access, vol. 8, pp. 76516-76531, 2020, doi: 10.1109/ACCESS.2020.2989857.

[11] M. T. Islam, M. Raihan, N. Aktar, M. S. Alam, R. R. Ema and T. Islam, "Diabetes Mellitus Prediction using Different Ensemble Machine Learning Approaches," 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2020, pp. 1-7, doi: 10.1109/ICCCNT49239.2020.9225551.

[12]G. G. Warsi, S. Saini and K. Khatri, "Ensemble Learning on Diabetes Data Set and Early Diabetes Prediction," 2019 International Conference on Computing, Power and Communication Technologies (GUCON), 2019, pp. 182-187.

[13] R. Akula, N. Nguyen and I. Garibay, "Supervised Machine Learning based Ensemble Model for Accurate Prediction of Type 2 Diabetes," 2019 SoutheastCon, 2019, pp. 1-8, doi: 10.1109/SoutheastCon42311.2019.9020358.

[14] D. Vigneswari, N. K. Kumar, V. Ganesh Raj, A. Gugan and S. R. Vikash, "Machine Learning Tree Classifiers in Predicting Diabetes Mellitus," 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), 2019, pp. 84-87, doi: 10.1109/ICACCS.2019.8728388.

PUBLICATIONS

International Conference on “Innovations in Computers Networks, Computational Intelligence and IoT

Paper ID: “ICICCI-21-0142"

PROFILE:

1. APOORVA SHARMA:

Apoorva Sharma is currently pursuing her Bachelor of Technology in the stream of Computer Science Engineering at St.Martin's Engineering College. She completed her intermediate from Sri Chaitanya junior college and 10th class from Sister Nivedita School. Her technical skills include Python, C, HTML and CSS. She also has a basic understanding of Java and DBMS. She took part in Employment Skill Development Program conducted by Zensar. She is also a student of Smart Interviews. Apart from programming. Her participations include:Online International Conference on "Innovations in Computer Networks, Computational Intelligence and loT" [ICICCI-21] On 25th June, 2021 and Online Two Day National Level Seminar on "Recent Trends in Cloud Computing Fog and Edge Computing" from 18th to 19th June, 2021. She completed few certification courses from online platforms like Udemy, Uxcel, Coursera, SoloLearn.

2. GVNS SAI BHASKAR:

GVNS Sai Bhaskar is currently pursuing her Bachelor of Technology in the stream of Computer Science Engineering at St. Martin's Engineering College. He completed his intermediate from Sri Chaitanya junior kalasla and 10th class from Sri Chaitanya techno school. His technical skills include Python, C, HTML and CSS. He also has a basic understanding of Java and DBMS. He took part in Employment Skill Development Program conducted by Zensar. He is also a student of Smart Interviews. Apart from programming. His participations include: Online International Conference on "Innovations in Computer Networks, Computational Intelligence and loT" [ICICCI-21] On 25th June, 2021 and Online Two Day National Level Seminar on "Recent Trends in Cloud Computing Fog and Edge Computing" from 18th to 19th June, 2021. He completed few certification courses from online platforms like Udemy, Uxcel, Coursera, SoloLearn.

3. LAKHAN PALORE

Lakhan Palore is currently pursuing her Bachelor of Technology in the stream of Computer Science Engineering at St. Martin's Engineering College. He completed his intermediate from Narayana Junior college and 10th class from Sister Nivedita School. His technical skills include Python, C, HTML and CSS. He also has a basic understanding of Java and DBMS. He took part in Employment Skill Development Program conducted by Zensar. He is also a student of Smart Interviews. Apart from programming. He did a one moth internship i.e., from June 2019 to July 2019, with Verzeo where he was trained in basics of Machine Learning. His participations include:Online International Conference on "Innovations in Computer Networks, Computational Intelligence and loT" [ICICCI-21] On 25th June, 2021 and Online Two Day National Level Seminar on "Recent Trends in Cloud Computing Fog and Edge Computing" from 18th to 19th June, 2021. He completed few certification courses from online platforms like Udemy, Excel, Coursera, Solo Learn.

4. NIKHIL SANGANI:

Nikhil Sangani is currently pursuing her Bachelor of Technology in the stream of Computer Science Engineering at St. Martin's Engineering College. He completed his intermediate from Roa's Junior college and 10th class from Krishnaveni High School. His technical skills include Python, C, HTML and CSS. He also has a basic understanding of Java and DBMS. His participations include:Online International Conference on "Innovations in Computer Networks, Computational Intelligence and loT" [ICICCI-21] On 25th June, 2021 and Online Two Day National Level Seminar on "Recent Trends in Cloud Computing Fog and Edge Computing" from 18th to 19th June, 2021. He completed few certification courses from online platforms like Udemy, Uxcel, Coursera, SoloLearn.

A

PROJECT REPORT

On

Bitcoin Crypto Currency Prediction

Submitted by

1)B.Karnakar (17K81A0504) 2)G.Shiva Prasad(17K81A0516)

3)K.Varun Kumar(17K81A0530)

in partial fulfillment for the award of the

degreeof

BACHELOR OF TECHNOLOGY

IN

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Under the Guidance of

Mr. Manohar Manchala

Assistant Professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ST.MARTIN'S ENGINEERING COLLEGE
An Autonomous Institute

Dhulapally, Secunderabad – 500 100

JUNE 2021

BONAFIDE CERTIFICATE

This is to certify that the project entitled "Bitcoin Crypto Currency Prediction", is being submitted by **1. B.Karnakar (17K81A0504), 2. G.Shiva Prasad (17K81A0516), 3. K.Varun Kumar (17K81A0530)** in partial fulfillment of the requirement for the award of the degree of **BACHELOR OF TECHNOLOGY IN COMPUTER SCIENCE AND ENGINEERING** is recorded of bonafide work carried out by them. The result embodied in this report have been verified and found satisfactory.

Mr. Manohar Manchala
Department of CSE

Head of the Department
Dr.M.NARAYANAN
Department of CSE

Internal Examiner

External Examiner

Place:

Date:

DECLARATION

We, the student of **Bachelor of Technology** in Department of Computer Science and Engineering', session: <2017 – 2021>, St. Martin's Engineering College, Dhulapally, Kompally, Secunderabad, hereby declare that work presented in this Project Work entitled "Bitcoin Crypto Currency Prediction" is the outcome of our own bonafide work and is correct to the best of our knowledge and this work has been undertaken taking care of Engineering Ethics. This result embodied in this project report has not been submitted in any university for award of any degree.

B.Karnakar (17K81A0504)
G.Shiva Prasad (17K81A0516)
K.Varun Kumar (17K81A0530)

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible and whose encouragements and guidance have crowded effects with success.

We extended our deep sense of gratitude to Principal**, Dr. P. SANTOSH KUMAR PATRA**, St. Martin's Engineering College, Dhulapally, for permitting us to undertake this project.

We are also thankful to **Dr. M.NARAYANAN**, Head of the Department, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

We would like to thank **Dr. T. POONGOTHAI,** Department of Computer Science and Engineering (AI & ML) for her encouragement and insightful comments and as well as our project coordinators **Dr. B.RAJALINGAM**, Associate Professor and **Mr. J.SUDHAKAR**, Assistant Professor, in Department of Computer Science and Engineering for their valuable support.

We would like to express our sincere gratitude and indebtedness to our project supervisor Mr. Manohar Manchala, Assistant Professor, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

Finally, we express thanks to all those who have helped us successfully to completing this project. Furthermore, we would like to thank our family and friends for their moral support and encouragement.

We express thanks to all those who have helped us in successfully completing the project.

B.Karnakar (17K81A0504)

G.Shiva Prasad (17K81A0516)

K.Varun Kumar (17K81A0530)

ABSTRACT

In this project, we tried to estimate the Bitcoin price precisely taking into consideration various parameters that affect the Bit coin value. In our work, we pointed to understand and identify daily changes in the Bit coin market while obtaining insight into most appropriate features surrounding Bit coin price. We will predict the daily price change with highest possible accuracy. The market capitalization of publicly traded crypto currencies is currently above $230 billion.Bitcoin, the most valuable crypto currency, serves primarily as a digital store of value, and its price predictability has been well-studied. These characteristics are outlined in the following subsection; the underlying details of Bitcoin,as they are described in depth in the cited papers.

The LSTM achieves the highest classification accuracy of 52% and a RMSE of 8%. The popular ARIMA model for time series forecasting is implemented as a comparison to the deep learning models. As expected, the non-linear deep learning methods outperform the ARIMA forecast which performs poorly.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

LSTM	Long Short Term Memory
ARIMA	Autoregressive Integrated Moving Average
RNN	Recurrent Neural Network
MLP	Multilayer Perceptron
SMA	Simple Moving Averages

1. INTRODUCTION

1.1 PROJECT OVERVIEW

Bitcoin uses a peer-to-peer technology to operate with no central authority or banks. Bitcoin is open-source; its design is public, nobody owns or controls Bitcoin and everyone can take part. Digital currency bring into use as open source software in 2009 by pseudonymous creator Satoshi Nakamoto It is a crypto currency, so-called because it uses cryptography to control the creation and transfer of money. Users send payments by broadcasting digitally signed messages to the network. Participants known as miners verify and timestamp transactions into a shared public database called the blockchain, for which they are rewarded with transaction fees and newly minted bitcoins. Conventionally "Bitcoin" capitalized refers to the technology and network whereas "bitcoins" lowercase refers to the currency itself. Bitcoins can be obtained by mining or in exchange for products, services, or other currencies.

Research on predicting the price of Bitcoin using machine learning algorithms specifically is lacking. implemented a latent source model as developed by to predict the price of Bitcoin noting 89% return in 50 days with a Sharpe ratio.

There has also been work using text data from social media platforms and other sources to predict Bitcoin prices. investigated sentiment analysis using support vector machines coupled with the frequency of Wikipedia views, and the network hash rate. investigated the relationship between Bitcoin price, tweets and views for Bitcoin on Google Trend.

However, one limitation of such studies is the often small sample size, and propensity for misinformation to spread through various (social) media channels such as Twitter or on message boards such as Reddit, which artificially inflate/deflate prices.

In the Bitcoin exchanges liquidity is considerably limited. As a result, the market suffers from a greater risk of manipulation. For this reason, sentiment from social media is not considered further.

The Bitcoin’s value varies just like any other stock. There are many algorithms used on stock market data for price forecast. However, the parameters affecting Bitcoin are different. Therefore it is necessary to foretelling the value of Bitcoin so that correct investment decisions can be made. The price of Bit coin does not depend on the business events or intervening government authorities, unlike the stock market. Thus, to forecast the value we feel it is necessary to leverage machine learning technology to predict the price of Bitcoin

1.2 PROJECT OBJECTIVES

Deep learning models such as the RNN and LSTM are evidently effective for Bitcoin prediction with the LSTM more capable for recognising longer-term dependencies. However, a high variance task of this nature makes it difficult to transpire this into impressive validation results. As a result it remains a difficult task. There is a fine line between overfitting a model and preventing it from learning sufficiently. Dropout is a valuable feature to assist in improving this. However, despite using Bayesian optimisation to optimize the selection of dropout it still couldn't guarantee good validation results. Despite the metrics of sensitivity, specificity and precision indicating good performance, the actual performance of the ARIMA forecast based on error was significantly worse than the neural network models. The LSTM outperformed the RNN marginally, but not significantly. However, the LSTM takes considerably longer to train.

This approach was taken for parameters which were unsuitable for Bayesian optimisation. This model was built using Keras in the Python programming language . Similar to the RNN, Bayesian optimization was chosen for selecting LTSM parameters where possible.

the expected improvement over the best result to pick hyper parameters for the next experiment. The performance of both the RNN and LSTM network are evaluated on validation data with measures to prevent overfitting. Dropout is implemented in both layers, and we automatically stop model training if its validation loss hasn't improved in 5 epochs.

LSTM models converged between 50 and 100 epochs with early stopping. Similar to the RNN, batch size was found to have a greater effect on execution time than accuracy. This may be due to the relatively small size of the dataset.

2. LITERATURE SURVEY

2.1 SURVEY ON BACKGROUND

Bitcoin is the worlds' most valuable cryptocurrency and is traded on over 40 exchanges worldwide accepting over 30 different currencies. It has a current market capitalization of 9 billion USD according to https://www.blockchain.info/ and sees over 250,000 transactions taking place per day. As a currency, Bitcoin offers a novel opportunity for price prediction due its relatively young age and resulting volatility,

which is far greater than that of fiat currencies . It is also unique in relation to traditional fiat currencies in terms of its open nature; no complete data exists regarding cash transactions or money in circulation for fiat currencies. Prediction of mature financial markets such as the stock market has been researched at length. Bitcoin presents an interesting parallel to this as it is a time series prediction problem in a market still in its transient stage. Traditional time series prediction methods such as Holt-Winters exponential smoothing models rely on linear assumptions and require data that can be broken down into trend, seasonal and noise to be effective. This type of methodology is more suitablem for a task such as forecasting sales where seasonal effects are present. Due to the lack of seasonality in the Bitcoin market and its high volatility, these methods are not very effective for this task. Given the complexity of the task, deep learning makes for an interesting technological solution based on its performance in similar areas. The recurrent neural network (RNN) and the long short term memory (LSTM) are favoured over the traditional multilayer perceptron (MLP) due to the temporal nature of Bitcoin data.

2.2 CONCLUSIONS ON SURVEY

The aim of this is to investigate with what accuracy the price of Bitcoin can be predicted using machine learning and compare parallelisation methods executed on multi-core and GPU environments. This paper contributes in the following manner: of approximately 653 papers published on Bitcoin [6], only 7 (at the time of writing) are related to machine learning for prediction. To facilitate a comparison to more traditional approaches in financial forecasting, an ARIMA time series model is also developed for performance comparison purposes with the neural network models.

The independent variable for this study is the closing price of Bitcoin in USD taken from the Coindesk Bitcoin Price Index. Rather than focusing on one specific exchange, we take the average price from five major Bitcoin exchanges: Bitstamp, Bitfinex, Coinbase, OkCoin and itBit. If we were to implement trades based on the signals it would be beneficial to focus on just one exchange. To assess the performance of models, we use the root mean squared error (RMSE) of the closing price and further encode the predicted price into categorical variable reflecting: price up, down or no change. This latter step allows for additional performance metrics that would be useful to a trader in the formation of a trading strategy: classification accuracy, specificity, sensitivity and precision. The dependent variables for this paper come from the Coindesk website, and Blockchain.info. In addition to the closing price, the opening price, daily high and daily low are also included as well as Blockchain data, i.e. the mining difficulty and hash rate. The features which have been engineered (considered as technical analysis indicators) include two simple moving averages (SMA) and a de-noised closing price

3. SOFTWARE AND HARDWARE REQUIREMENTS

3.1 SOFTWARE REQUIREMENTS

The software requirements specify the use of all required software products like data management system. The required software product specifies the numbers and version. Each interface specifies the purpose of the interfacing software as related to this software product.

Operating system	:	Windows 7.
Coding Language	:	python
Environment	:	Anaconda,
Tools	:	visual studio code
Libraries	:	Scipy, numpy, sklearn, seaborn

3.2 HARDWARE REQUIREMENTS

The hardware requirement specifies each interface of the software elements and the hardware elements of the system. These hardware requirements include configuration characteristics.

System	:	Pentium Dual Core.
Hard Disk	:	120 GB.
Monitor	:	15'' LED
Input Devices	:	Keyboard, Mouse
Ram	:	1GB.

4. SOFTWARE DEVELOPMENT ANALYASIS

The software development process involves the creation and maintenance of applications, frameworks and other components for software design, design, programming, documentation, testing and problem remediation. The development of software is a process of creating and keeping source code, but it encompasses everything from the idea of the intended software to the last manifestation of the programme, often in a planned and organised process in a larger context. Software development may therefore encompass research, creation of new software products, prototype, modification, reuse, reengineering, maintenance, or any other software-production activity. A life-cycle "model" is sometimes considered a more general term for a category of methodologies and a software development "process" a more specific term to refer to a specific process chosen by a specific organization.[citation needed] For example, there are many specific software development processes that fit the spiral lifecycle model. For example, there are many specific software development processes that fit the spiral life-cycle model. The field is often considered a subset of the systems development life cycle.

4.1 OVERVIEW OF PROBLEM

In recent times Bitcoin, have been one of the most current topics regarding payment systems Bitcoin is very popular because of its unique features like private cryptographic key is used to make payments; it is decentralised; and it can be conveyed safely within an online platform without having a single authority backing it as a payment system. However, although virtual and digital currencies have become well established and are accepted as a means of payment, the use of such currencies is not without its problems and legal challenges. There are issues concerning the legal status of Bitcoin and its regulation. The focus of this research is therefore on examining the nature and importance of Bitcoin, legal issues, and technical challenges related to Bitcoins, and legal position of Bitcoins in various countries and specific areas of regulations of Bitcoin's.

4.2 DEFINE THE PROBLEM

Bitcoin trading suffers from illiquidity and manipulation because of the existence of "whale wallets" (wallets holding disproportionately large amounts of bitcoins).

In late 2020, the top 100 wallets were estimated to own 13% of total bitcoin supply with most of the owners' identities not known. It would therefore only take a few whale wallets to manipulate the bitcoin market, causing violent price moves. Huge price volatility has made bitcoin and cryptocurrencies unsuitable as store of value.

4.3 MODULES OVERVIEW

The aim of the project entitled as Bitcoin Crypto Currency Predictionis to develop a system where we can predict the direction of the Bitcoin cryptocurrency market direction . Investors can invest with the direction of the trend following the charts.

4.4 DEFINE THE MODULE

There are two modules in our project:

A. **User:**

   In this the user obtains the dataset from a particular university. This dataset includes all the features which directly effects the motivational/sentimental status of the students. Using this it allows the instructors to have idea about how effective the courses are.

B. **Application:**

   In this the dataset is read as an input and all the pre-processing and PCA are performed so as to tune the dataset(explore the important features in the dataset). Along with it the selected algorithms are applied so as to detect the accuracy in order to obtain the best results.

4.5 MODULE FUNCTIONALITY

Code is produced from the deliverables of the design phase during implementation, and this is the longest phase of the software development life cycle. For a developer, this is the main focus of the life cycle because this is where the code is produced. Implementation my overlap with both the design and testing phases. Many tools exists (CASE tools) to actually automate the production of code using information gathered and produced during the design phase.

5. PROJECT SYSTEM DESIGN

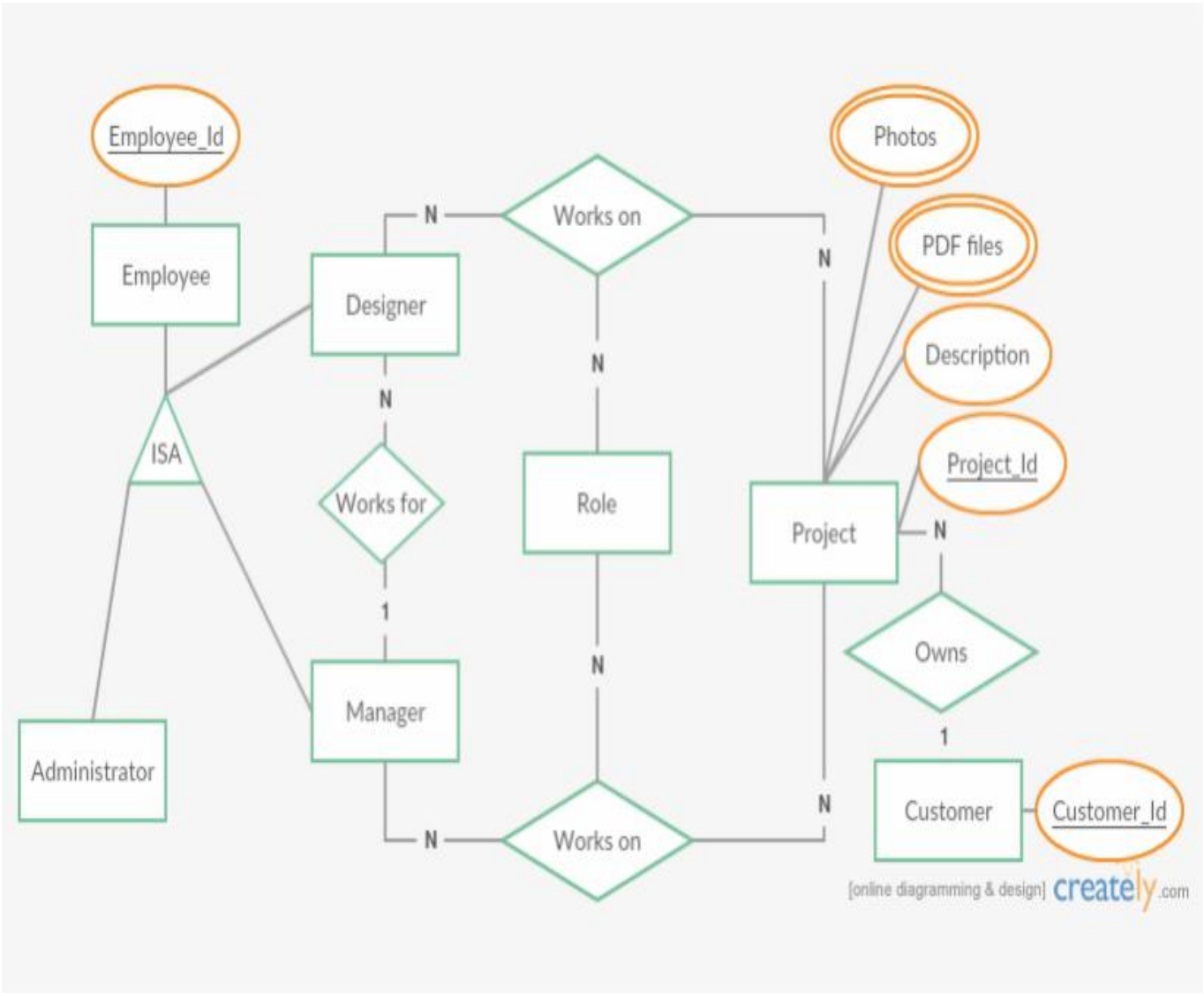
5.1 DFDS IN CASE OF DATABASE PROJECTS

Data Flow Diagram can also be termed as bubble chart. It is a pictorial or graphical form, which can be applied to represent the input data to a system and multiple functions carried out on the data and the generated output by the system.

A graphical tool accustomed describe and analyze the instant of knowledge through a system manual or automatic together with the method, stores of knowledge, and delays within the system. The transformation of knowledge from input to output, through processes, is also delineate logically and severally of the physical elements related to the system. The DFD is also known as a data flow graph or a bubble chart.The BasicNotation used to create a DFD's are as follows:

5.2 ER DIAGRAMS

ER Diagram stands for Entity Relationship Diagram, also known as ERD is a diagram that displays the relationship of entity sets stored in a database. In other words, ER diagrams help to explain the logical structure of databases. ER diagrams are created based on three basic concepts: entities, attributes and relationships.ER Diagrams contain different symbols that use rectangles to represent entities, ovals to define attributes and diamond shapes to represent relationships.



5.3 UML DIAGRAMS

The Unified Modeling Language allows the software engineer to express an analysis model using the modeling notation that is governed by a set of syntactic semantic and pragmatic rules.

A UML system is represented using five different views that describe the system from distinctly different perspective. Each view is defined by a set of diagram, which is as follows.

User Model View

This view represents the system from the users perspective. The analysis representation describes a usage scenario from the end-users perspective.

Structural Model view

In this model the data and functionality are arrived from inside the system. This model view models the static structures.

Behavioral Model View

It represents the dynamic of behavioral as parts of the system, depicting the interactions of collection between various structural elements described in the user model and structural model view.

Implementation Model View

In this the structural and behavioral as parts of the system are represented as they are to be built.

A) USE CASE DIAGRAM

A use case diagram at its simplest is a representation of a user's interaction with the system and depicting the specifications of a use case. A use case diagram can portray the different types of users of a system and the various ways that they interact with the system. This type of diagram is typically used in conjunction with the textual use case and will often be accompanied by other types of diagrams as well.

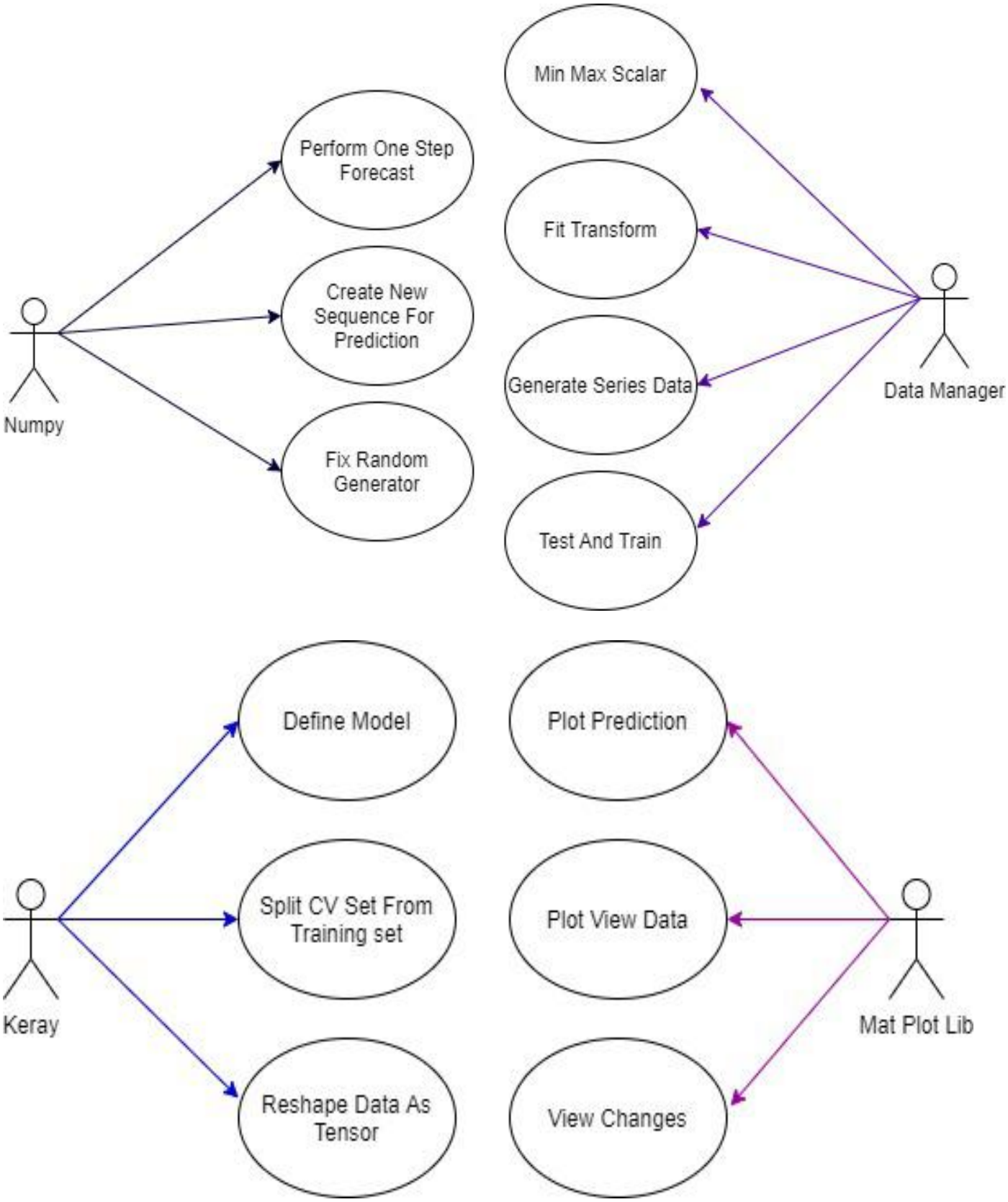


Figure A: Use Case Diagram

B) CLASS DIAGRAM

The class diagram is the main building block of object oriented modeling. It is used both for general conceptual modeling of the systematic of the application, and for detailed modeling translating the models into programming code. Class diagrams can also be used for data modeling. The classes in a class diagram represent both the main objects, interactions in the application and the classes to be programmed. A class with three sections, in the diagram, classes is represented with boxes which contain three parts:

The upper part holds the name of the class

The middle part contains the attributes of the class

The bottom part gives the methods or operations the class can take or undertake.

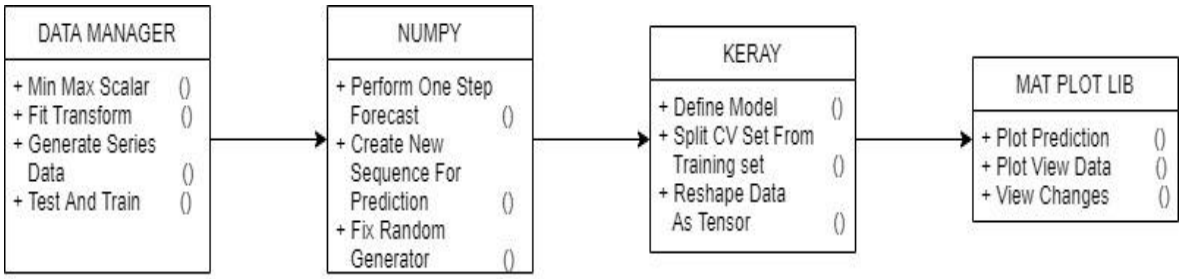
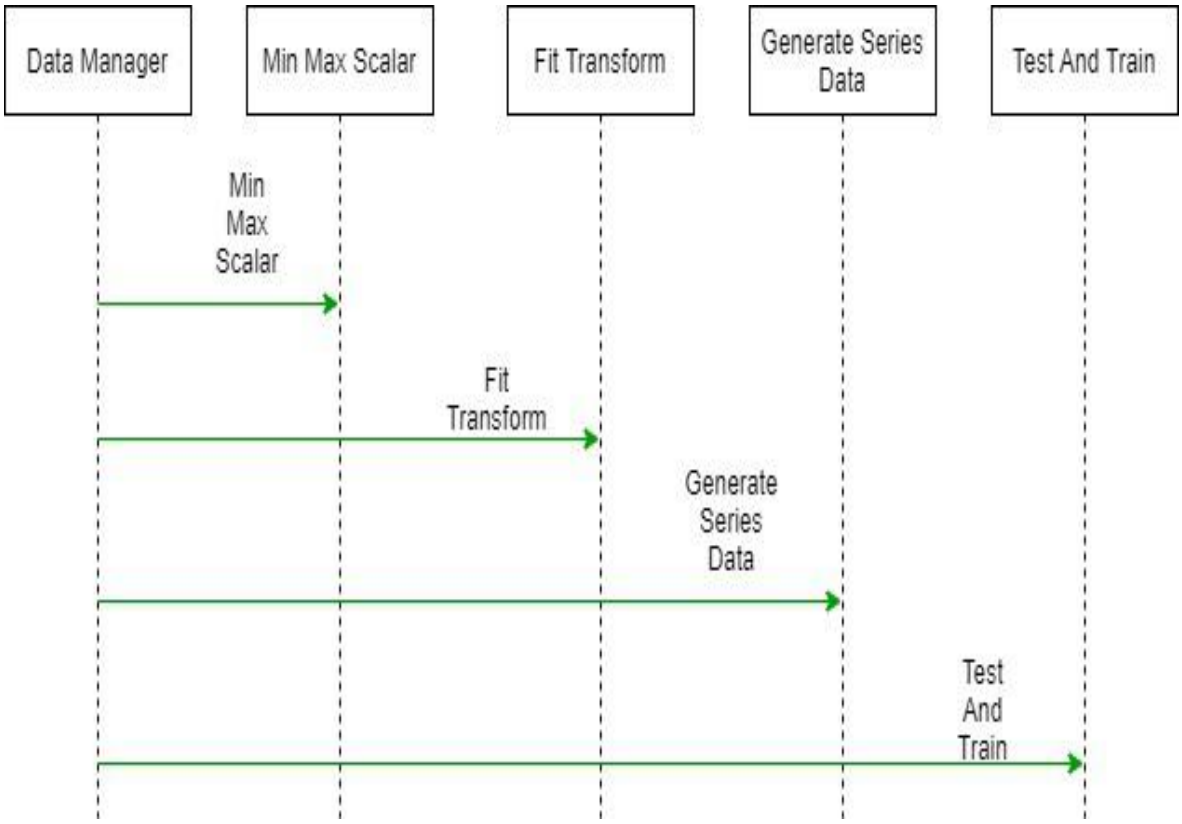
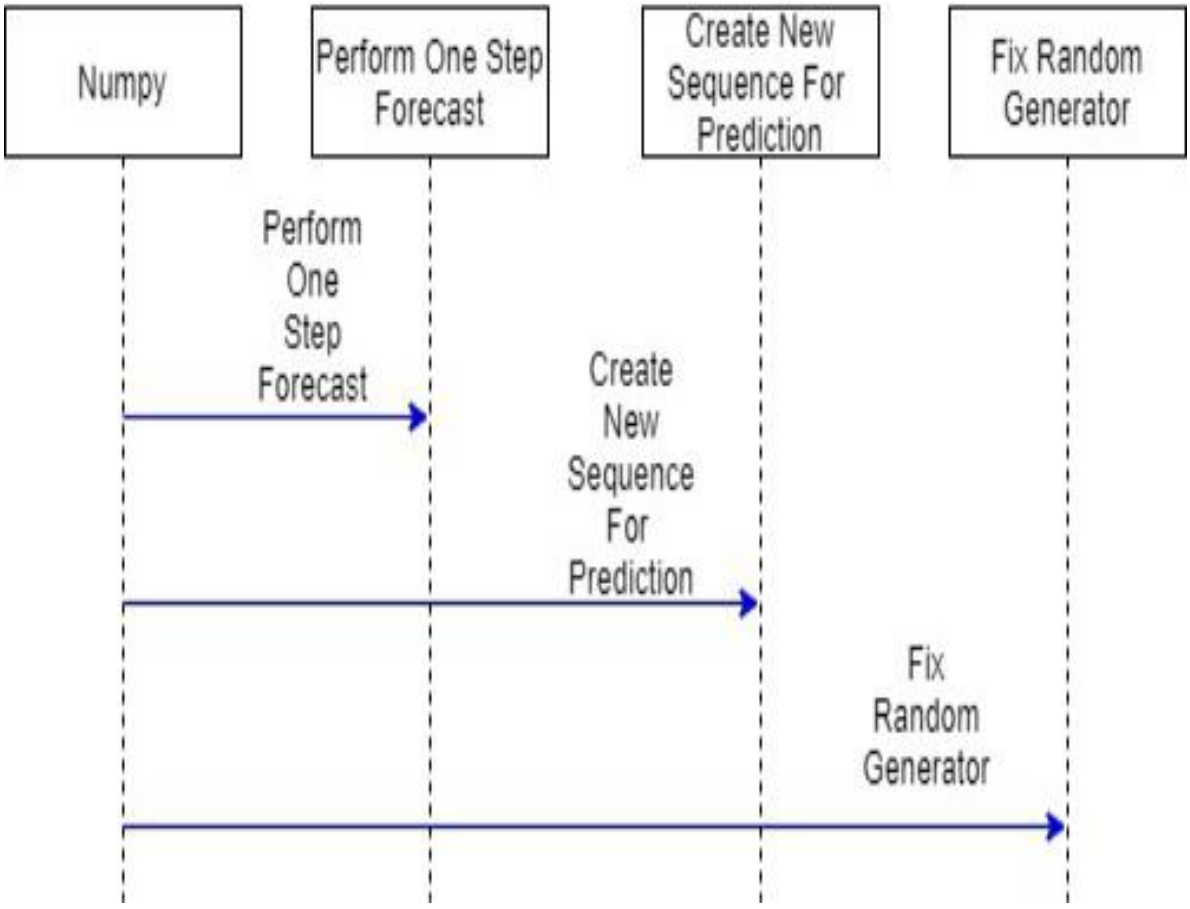


Figure B: Class Diagram.

C) SEQUENCEDIAGRAM

A sequence diagram is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. A sequence diagram shows object interactions arranged in time sequence. It depicts the objects and classes involved in the scenario and the sequence of messages exchanged between the objects needed to carry out the functionality of the scenario. Sequence diagrams are typically associated with use case realizations in the Logical View of the system under development. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.


Data Manager
Min Max Scalar
Fit Transform
Generate Series Data
Test And Train
Min Max Scalar
Fit Transform
Generate Series Data
Test And Train


Numpy
Perform One Step Forecast
Create New Sequence For Prediction
Fix Random Generator
Perform One Step Forecast
Create New Sequence For Prediction
Fix Random Generator

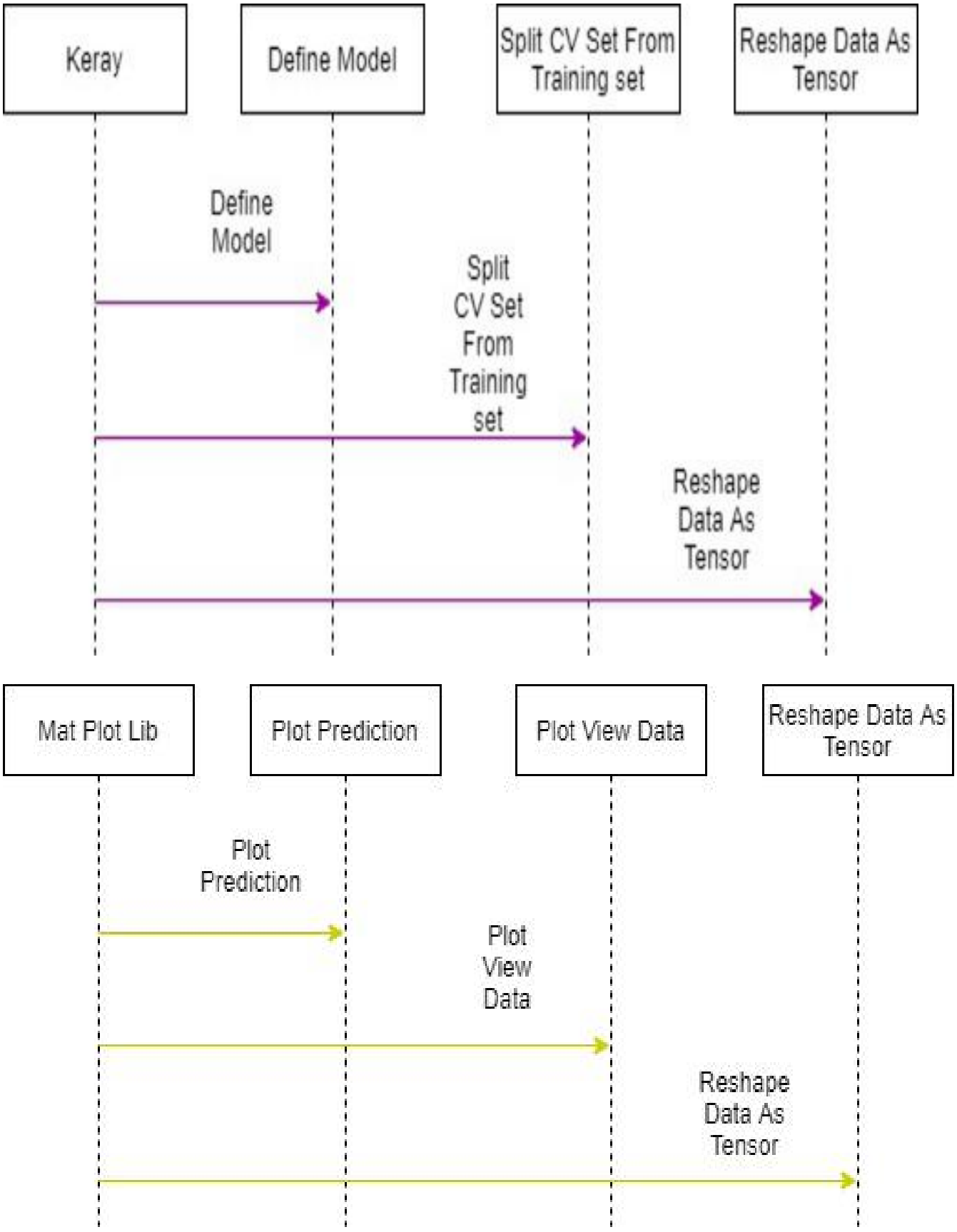


Figure C: Sequence diagram

D) ACTIVITY DIAGRAM

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modeling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.

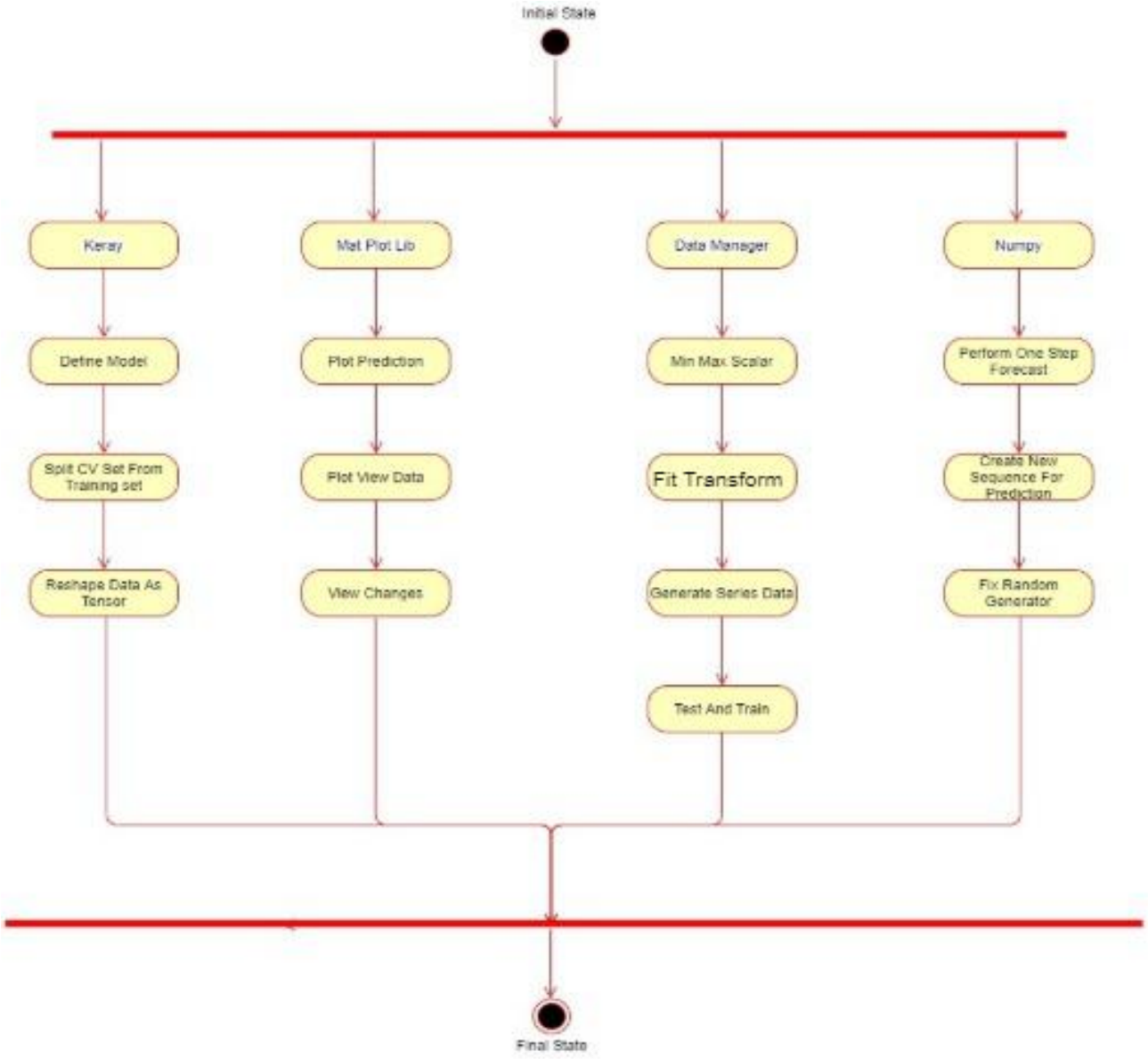


Figure D: Activity Diagram

6. PROJECT CODING

6.1 CODE TEMPLATES

```
import numpy as np

import matplotlib.pyplot as plt

from keras.models import Sequential

from keras.layers import Dense, Activation, Dropout, LSTM, Flatten

from keras.optimizers import RMSprop

from keras.utils import np_utils

from keras.preprocessing import sequence

from sklearn.model_selection import train_test_split

import dataManager

import constants



# perform one-step forecast(with test data)

def one_step_prediction_using_test_data(model, previos_y):


    y_pred = np.zeros((len(dm.test_scaled), 1))


    for i in range(len(dm.test_scaled)):


        x = dm.test_scaled[i, :-1].reshape(constants.BATCH_SIZE, constants.SEQ_LENGTH,
constants.DATA_DIM)

        _y_pred = model.predict(x)
```

```
        x = x.reshape(constants.BATCH_SIZE, constants.SEQ_LENGTH)
        _y_pred = _y_pred.reshape(constants.BATCH_SIZE, 1)

        # inverse scale
        y_pred_indiffereced = dm.inverse_data(x, _y_pred, previos_y)

        previos_y = dm.test_original_df.iloc[i, -1]

        model.reset_states()
        y_pred[i, 0] = y_pred_indiffereced

        # create a new sequence for the next prediction
        x = np.delete(x, 0)
        x = np.append(x, _y_pred[0, 0])

    return y_pred


if __name__ == '__main__':

    # fix the random generator
    np.random.seed(1)

    # coin pairs: USDT-Bitcoin
    dm = dataManager.DataManager()
```

```
    dm.return_chart_data(constants.PAIR_USDT_BTC, constants.PERIOD, "2021-05-17 00:00", "2021-05-19 00:00")

    dm.prepare_data()


    # definition of NN

    model = Sequential()

    model.add(LSTM(32, return_sequences=False, batch_input_shape=(constants.BATCH_SIZE, constants.SEQ_LENGTH, constants.DATA_DIM), stateful=True))

    #model.add(LSTM(32, stateful=True))

    model.add(Dense(1, activation='linear'))

    #model.add(Dropout(0.5))


    # learning rate as default

    model.compile(loss="mean_squared_error", optimizer="rmsprop")


    print(model.summary())


    # split CV set from training set.

    train_data_scaled, cv_data_scaled = train_test_split(

        dm.train_scaled, test_size=0.25, random_state=42, shuffle=False)

    x_train_scaled, y_train_scaled = train_data_scaled[:, 0:-1], train_data_scaled[:, -1]

    x_cv_scaled, y_cv_scaled = cv_data_scaled[:, 0:-1], cv_data_scaled[:, -1]

    # reshape data as a tensor(3 dims)

    x_train_scaled = x_train_scaled.reshape((x_train_scaled.shape[0], constants.SEQ_LENGTH, constants.DATA_DIM))

    print(x_train_scaled.shape)
```

```
    x_cv_scaled = x_cv_scaled.reshape((x_cv_scaled.shape[0], constants.SEQ_LENGTH,
constants.DATA_DIM))
    # fit data
    model.fit(x_train_scaled, y_train_scaled,
          batch_size=constants.BATCH_SIZE,
          epochs=constants.EPOCHS,
          verbose=1,
          validation_data=(x_cv_scaled, y_cv_scaled),
          shuffle=False)



    predicted_test = one_step_prediction_using_test_data(model, dm.train_original_df.iloc[-1, -1])


    # plot predicitions
    plt.figure(figsize=(20,10))
    # get the original test data(not differenced)
    plt.plot(dm.time_test, dm.test_original_df.iloc[:, -1], label = "real")
    plt.plot(dm.time_test, predicted_test, label = "predicted")
    plt.legend(loc='best', fontsize=14)
    plt.tick_params(labelsize=14)
    plt.show()
```

6.2 OUTLINE FOR VARIOUS FILES

This project includes two files:

A) **Code file (main.py):** In this it contains the code
B) **Dataset file (dataManager.py):** It includes the bitcoin dataset that we are taking as the input

6.3 CLASS WITH FUNCTIONALITY

1. **Reading the dataset:** In this method we read the dataset and extract it into the code successfully.

2. **Exploring the dataset:** In this method we explore all the features that are included which directly effects the sentimental status of the students.

3. **Pre-processing the data:** In this method we tune the dataset by eliminating redundancy so as to have purity in the dataset. This thereby taken as input in an application.

4. **Training and Testing:** In this we take the data samples so as to test the algorithms accuracy. This is done so as to come up with the better algorithm to give accurate precision.

6.4 METHODS INPUT AND OUTPUT PARAMETERS

INPUT:

The dataset containing the details of students whose risk level is to be determined is taken as input. And from this dataset we explore the features that directly effects the students motivational/sentimental status. Later we pre-process the dataset so as to reduce the size by eliminating unnecessary records. Since the dataset is ready to test or train the model, we finally give this dataset to the application that we proposed.

OUTPUT:

After reading the input we finally run it through the algorithms so as to measure the scores in terms of accuracy. Thereby allowing us to choose which algorithm is best fit for this model. By selecting it makes the instructors to determine if their students require any interventions or not.

7. PROJECT TESTING

7.1 VARIOUS TEST CASES

A) Unit Testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

B) Functional Testing

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items:

Valid Input : identified classes of valid input must be accepted.

Invalid Input : identified classes of invalid input must be rejected.

Functions : identified functions must be exercised.

Output : identified classes of application outputs must be exercised.

Systems/Procedures : interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

C) Integration Testing

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

D) System Testing

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

E) User Acceptance Testing

User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements.

Test Case Id	Test Case Name	Test Case Desc.	Test Steps			Test Case Status	Test Priority
			Step	Expected	Actual		
01	Upload the tasks dataset	Verify either file is loaded or not	If dataset is not uploaded	It cannot display the file loaded message	File is loaded which displays task waiting time	High	High

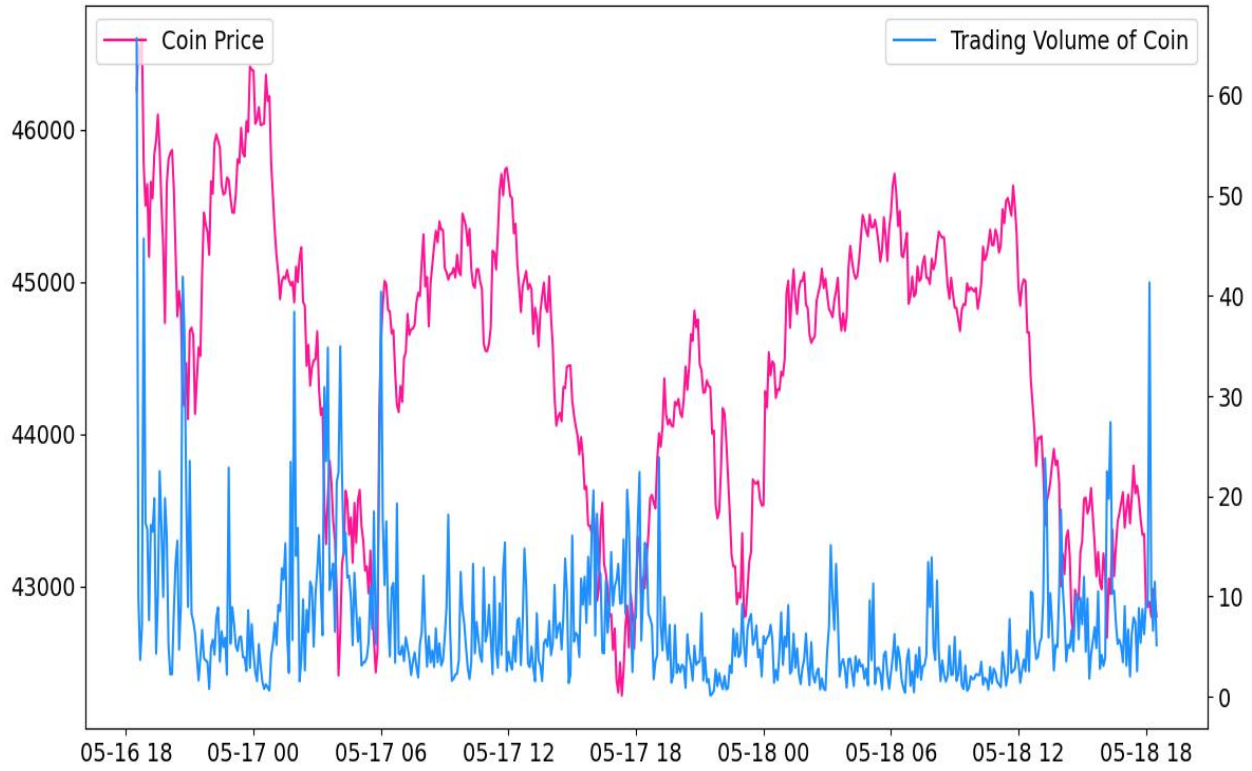
7.2 BLACK BOX

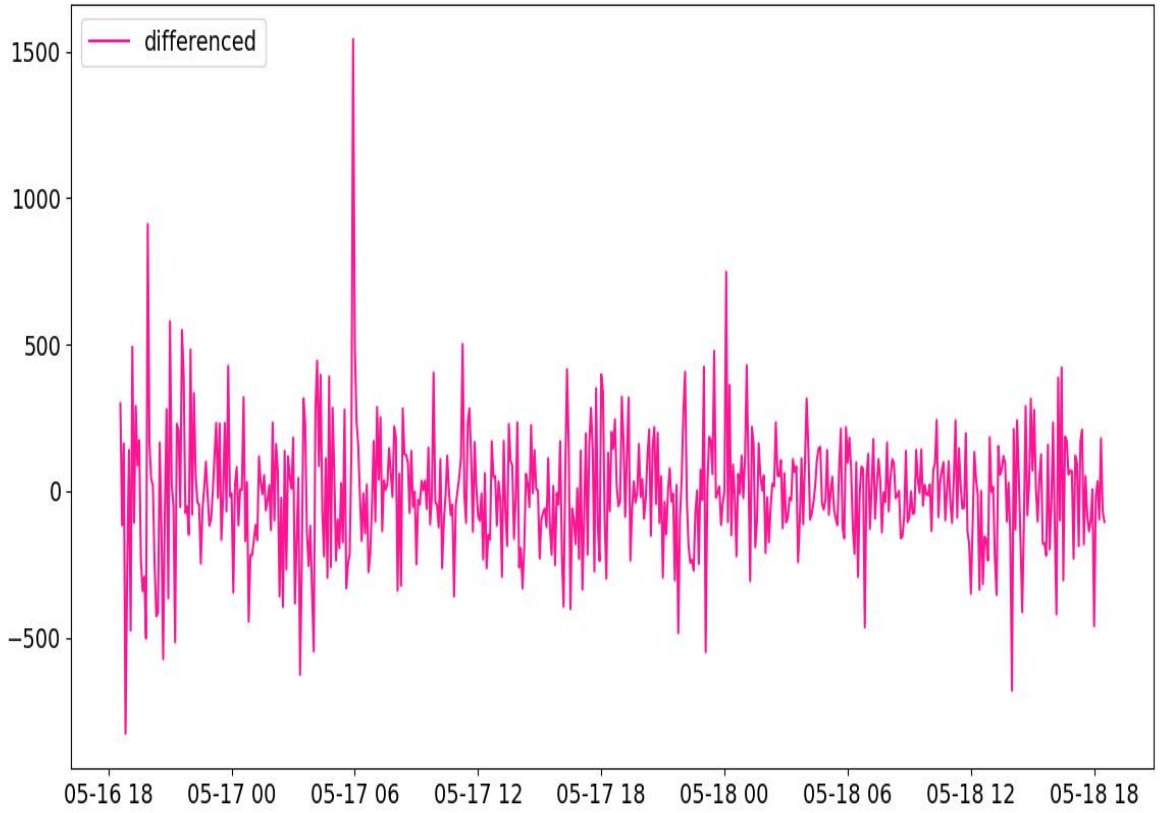
Black Box Testing is a software testing method in which the functionalities of software applications are tested without having knowledge of internal code structure, implementation details and internal paths. Black Box Testing mainly focuses on input and output of software applications and it is entirely based on software requirements and specifications. It is also known as Behavioural Testing. The above Black-Box can be any software system you want to test. Under Black Box Testing, we can test these applications by just focusing on the inputs and outputs without knowing their internal code implementation.

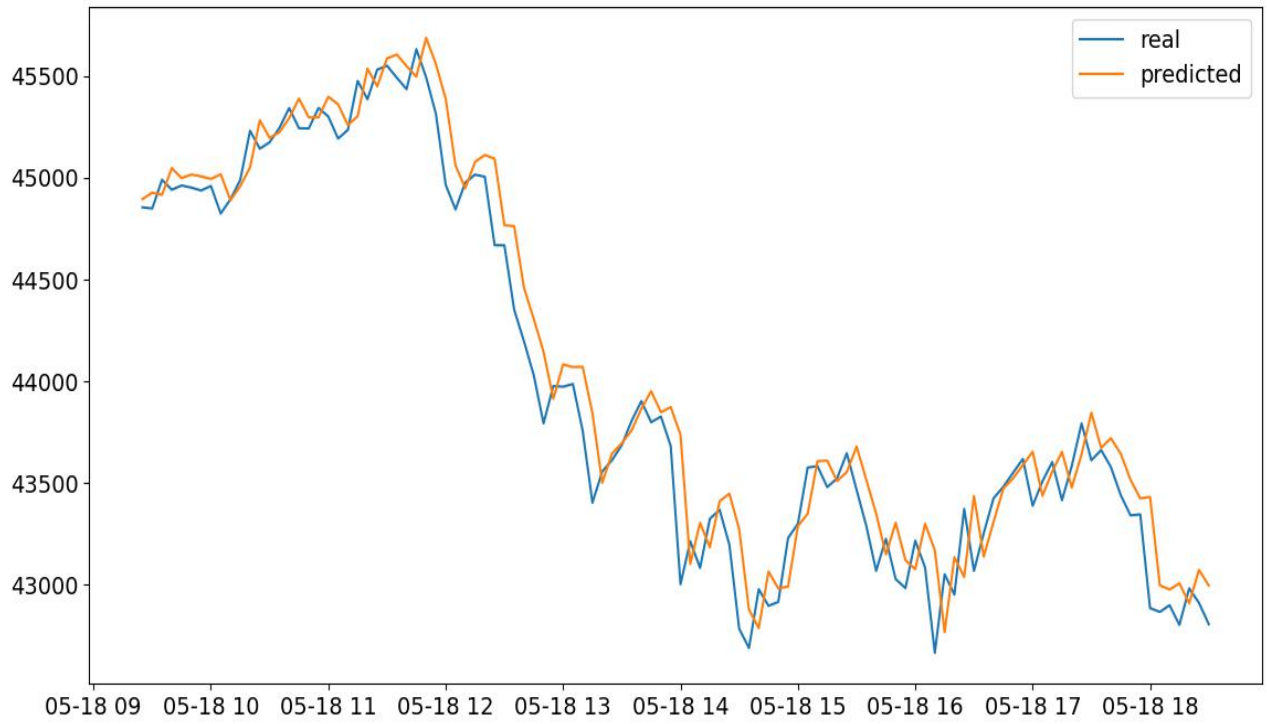
7.3 WHITE BOX TESTING

White Box Testing is software testing technique in which internal structure, design and coding of software are tested to verify flow of input-output and to improve design, usability and security. In white box testing, code is visible to testers, so it is also called Clear box testing, Open box testing, Transparent box testing, Code-based testing and Glass box testing. It is one of two parts of the Box Testing approach to software testing. Its counterpart, Blackbox testing, involves testing from an external or end-user type perspective. On the other hand, White box testing in software engineering is based on the inner workings of an application and revolves around internal testing. The term "White Box" was used because of the see-through box concept. The clear box or White Box name symbolizes the ability to see through the software's outer shell (or "box") into its inner workings. Likewise, the "black box" in "Black Box Testing" symbolizes not being able to see the inner workings of the software so that only the individual experience can be tested.

8. OUTPUT SCREENS




differenced
1500
1000
500
0
−500
05-16 18
05-17 00
05-17 06
05-17 12
05-17 18
05-18 00
05-18 06
05-18 12
05-18 18


real
predicted
45500
45000
44500
44000
43500
43000
05-18 09
05-18 10
05-18 11
05-18 12
05-18 13
05-18 14
05-18 15
05-18 16
05-18 17
05-18 18

9. EXPERIMENTAL RESULT

As can be seen in Table I, LSTM achieved the highest accuracy while the RNN achieved the lowest RMSE. The ARIMA prediction performed poorly in terms of accuracy and RMSE. Upon analysis of the ARIMA forecast, it predicted the price would gradually rise each day. There were no false positives from the model. One reason for this may be due to the class imbalance in predictive portion of the ARIMA forecast (the price tends to always increase). This contributed to the specificity and precision being so high (specificity, precision= 100%). This does not necessarily suggest good overall performance, but rather that it does a decent job at identifying price direction change

Table I: Model Results

Model	Temporal Length	Sensitivity	Specificity	Precision	Accuracy	RMSE
LSTM	100	37%	61.30%	35.50%	52.78%	6.87%
RNN	20	40.40%	56.65%	39.08%	50.25%	5.45%
ARIMA	170	14.7%	100%	100%	50.05%	53.74%

10. CONCLUSION AND FUTURE ENHANCEMENT

Deep learning models such as the RNN and LSTM are evidently effective for Bitcoin prediction with the LSTM more capable for recognising longer-term dependencies. However, a high variance task of this nature makes it difficult to transpire this into impressive validation results. As a result it remains a difficult task. There is a fine line between overfitting a model and preventing it from learning sufficiently. Dropout is a valuable feature to assist in improving this. However, despite using Bayesian optimisation to optimize the selection of dropout it still couldn't guarantee good validation results. Despite the metrics of sensitivity, specificity and precision indicating good performance, the actual performance of the ARIMA forecast based on error was significantly worse than the neural network models. The LSTM outperformed the RNN marginally, but not significantly. However, the LSTM takes considerably longer to train.

Future Enhancements:

It is not possible to develop a system that makes all the requirements of the user. User requirements keep changing as the system is being used. Some of the future enhancements that can be done to this system are:

- As the technology emerges, it is possible to upgrade the system and can be adaptable to desired environment.
- Based on the future security issues, security can be improved using emerging technologies like single sign-on.

11. REFERENCES

[1] S. Nakamoto, “Bitcoin: A peer-to-peer electronic cash system,” 2008.

[2] M. Briere, K. Oosterlinck, and A. Szafarz, “Virtual currency, tangible ` return: Portfolio diversification with bitcoins,” Tangible Return: Portfolio Diversification with Bitcoins (September 12, 2013), 2013.

[3] I. Kaastra and M. Boyd, “Designing a neural network for forecasting financial and economic time series,” Neurocomputing, vol. 10, no. 3, pp. 215–236, 1996.

[4] H. White, “Economic prediction using neural networks: The case of ibm daily stock returns,” in Neural Networks, 1988., IEEE International Conference on. IEEE, 1988, pp. 451–458.

[5] C. Chatfield and M. Yar, “Holt-winters forecasting: some practical issues,” The Statistician, pp. 129–140, 1988.

[6] B. Scott, “Bitcoin academic paper database,” suitpossum blog, 2016.

[7] M. D. Rechenthin, “Machine-learning classification techniques for the analysis and prediction of high-frequency stock direction,” 2014. [8] D. Shah and K. Zhang, “Bayesian regression and bitcoin,” in Communication, Control, and Computing (Allerton), 2014 52nd Annual Allerton Conference on. IEEE, 2014, pp. 409–414.

[8] G. H. Chen, S. Nikolov, and D. Shah, “A latent source model for nonparametric time series classification,” in Advances in Neural Information Processing Systems, 2013, pp. 1088–1096.

[9] I. Georgoula, D. Pournarakis, C. Bilanakos, D. N. Sotiropoulos, and G. M. Giaglis, “Using time-series and sentiment analysis to detect the determinants of bitcoin prices,” Available at SSRN 2607167, 2015.

[10] M. Matta, I. Lunesu, and M. Marchesi, “Bitcoin spread prediction using social and web search media,” Proceedings of DeCAT, 2015.

[11] ——, “The predictor impact of web search media on bitcoin trading volumes.”

[12] B. Gu, P. Konana, A. Liu, B. Rajagopalan, and J. Ghosh, “Identifying information in stock message boards and its implications for stock market efficiency,” in Workshop on Information Systems and Economics, Los Angeles, CA, 2006.

[13] A. Greaves and B. Au, “Using the bitcoin transaction graph to predict the price of bitcoin,” 2015.

[14] I. Madan, S. Saluja, and A. Zhao, “Automated bitcoin trading via machine learning algorithms,” 2015.

[15] R. Delfin Vidal, “The fractal nature of bitcoin: Evidence from wavelet power spectra,” The Fractal Nature of Bitcoin: Evidence from Wavelet Power Spectra (December 4, 2014), 2014.

12. PUBLICATIONS

International Conference on “Innovations in Computers Networks, Computational Intelligence and IOT” (ICICCI-21)

Paper ID: ICICCI-21-0117

A

PROJECT REPORT

On

DRIVER DROWSINESS MONITORING SYSTEM USING VISUAL BEHAVIOUR AND MACHINE LEARNING

Submitted by

Ms.BIJJA RAGASREE(17K81A0505)

Ms.GUDIPELLY SNEHA(17K81A0520)

Ms.JIGATAPU PRAVALIKA(17K81A0523)

Ms.NEEMKAR RITHIKA(17K81A0539)

in partial fulfilment for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Under the Guidance of

Mr.N.KRISHNAVARDHAN

ASSOCIATE PROFESSOR

ST.MARTIN'S ENGINEERING COLLEGE
An Autonomous Institute
Dhulapally, Secunderabad – 500 100

JUNE 2021

BONAFIDE CERTIFICATE

This is to certify that the project entitled **DRIVER DROWSINESS MONITORING SYSTEM USING VISUAL BEHAVIOUR AND MACHINE LEARNING** , is being submitted by **Ms.BIJJA RAGASREE 17K81A0505, Ms.GUDIPELLY SNEHA 17K81A0520, Ms.JIGATAPU PRAVALIKA 17K81A0523, Ms.NEEMKAR RITHIKA 17K81A0539** in partial fulfillment of the requirement for the award of the degree of **BACHELOR OF TECHNOLOGY IN COMPUTER SCIENCE AND ENGINEERING** is recorded of bonafide work carried out by them. The result embodied in this report have been verified and found satisfactory.

Associate Professor	**Head of the Department**
Mr.N.KRISHNAVARDHAN	**Dr.M.NARAYANAN**
Department of CSE	**Department of CSE**

Internal Examiner External Examiner

Place:Hyderabad

Date:

DECLARATION

We, the students of **Bachelor of Technology** in Department of Computer Science and Engineering', session: 2017 – 2021, St. Martin's Engineering College, Dhulapally, Kompally, Secunderabad, hereby declare that work presented in this Project Work entitled **DRIVER DROWSINESS MONITORING SYSTEM USING VISUAL BEHAVIOUR AND MACHINE LEARNING** is the outcome of our own bonafide work and is correct to the best of our knowledge and this work has been undertaken taking care of Engineering Ethics. This result embodied in this project report has not been submitted in any university for award of any degree.

Ms. BIJJA RAGASREE	17K81A0505
Ms. GUDIPELLY SNEHA	17K81A0520
Ms. JIGATAPU PRAVALIKA	17K81A0523
Ms. NEEMKAR RITHIKA	17K81A0539

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible and whose encouragements and guidance have crowded effects with success.

We extended our deep sense of gratitude to Principal**, Dr. P. SANTOSH KUMAR PATRA**, St. Martin's Engineering College, Dhulapally, for permitting us to undertake this project.

We are also thankful to **Dr. M.NARAYANAN**, Head of the Department, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

We would like to thank **Dr. T. POONGOTHAI,** Department of Computer Science and Engineering (AI & ML) for her encouragement and insightful comments and as well as our project coordinators **Dr. B.RAJALINGAM**, Associate Professor and **Mr. J.SUDHAKAR**, Assistant Professor, in Department of Computer Science and Engineering for their valuable support.

We would like to express our sincere gratitude and indebtedness to our project supervisor **Mr.N.KRISHNAVARDHAN**, Associate Professor, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

Finally, we express thanks to all those who have helped us successfully to completing this project. Furthermore, we would like to thank our family and friends for their moral support and encouragement.

We express thanks to all those who have helped us in successfully completing the project.

Ms. BIJJA RAGASREE	17K81A0505
Ms. GUDIPELLY SNEHA	17K81A0520
Ms. JIGATAPU PRAVALIKA	17K81A0523
Ms. NEEMKAR RITHIKA	17K81A0539

ABSTRACT

Drowsy driving is one of the major causes of road accidents and death. Hence, detection of driver's fatigue and its indication is an active research area. Most of the conventional methods are either vehicle based, or behavioral based or physiological based. Few methods are intrusive and distract the driver, some require expensive sensors and data handling. Depending on the sensors used in the system, system cost as well as size will increase.In our project Support Vector Machine(SVM) and Histogram of Oriented Gradients(HOG) have been used. Face is detected in the frames using histogram of oriented gradients (HOG) and linear support vector machine (SVM) for object detection .In the developed system, a webcam records the video and driver's face is detected in each frame employing image processing techniques. Facial landmarks on the detected face are pointed and subsequently the eye aspect ratio and mouth opening ratio are computed and depending on their values, drowsiness is detected based on developed adaptive thresholding. Machine learning algorithms have been implemented as well in an offline manner.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

SVM	Support Vector Machine
EAR	Eye Aspect Ratio
MOR	Mouth Opening Ratio
HOG	Histogram of Oriented Gradients

GLOSSARRY OF TERMS

Drowsiness detection, Eye Aspect ratio, Mouth opening ratio, Machine Learning, Support Vector Machine, visual behaviour.

1.INTRODUCTION

1.1 PROJECT OVERVIEW

Driver drowsiness is an overcast nightmare to passengers in every country. Every year, a large number of injuries and deaths occur due to fatigue related road accidents. Hence, detection of driver's fatigue and its indication is an active area of research due to its immense practical applicability. The basic drowsiness detection system has three blocks/modules; acquisition system, processing system and warning system. Here, the video of the driver's frontal face is captured in acquisition system and transferred to the processing block where it is processed online to detect drowsiness. If drowsiness is detected, a warning or alarm is send to the driver from the warning system.

Generally, the methods to detect drowsy drivers are classified in three types; vehicle based, behavioural based and physiological based.In vehicle based method, a number of metrics like steering wheel movement, accelerator or brake pattern, vehicle speed, lateral acceleration, deviations from lane position etc. are monitored continuously. Detection of any abnormal change in these values is considered as driver drowsiness. This is a nonintrusive measurement as the sensors are not attached on the driver. In behavioural based method, the visual behavior of the driver i.e., eye blinking, eye closing, yawn, head bending etc. are analyzed to detect drowsiness. This is also nonintrusive measurement as simple camera is used to detect these features. In physiological based method , the physiological signals like Electrocardiogram (ECG), Electooculogram (EOG), Electroencephalogram (EEG), heartbeat, pulse rate etc. are monitored and from these metrics, drowsiness or fatigue level is detected. This is intrusive measurement as the sensors are attached on the driver which will distract the driver. Depending on the sensors used in the system, system cost as well as size will increase. However, inclusion of more parameters/features will increase the accuracy of the system to a certain extent. These factors motivate us to develop a low-cost, real time driver's drowsiness detection system with acceptable accuracy. Hence, we have proposed a webcam based system to detect driver's fatigue from the face image only using image processing and machine learning techniques to make the system low-cost as well as portable.

1.2 PROJECT OBJECTIVES

In this project by monitoring Visual Behaviour of a driver with webcam and machine learning SVM (support vector machine) algorithm we are detecting Drowsiness in a driver. This application will use inbuilt webcam to read pictures of a driver and then using OPENCV SVM algorithm extract facial features from the picture and then check whether driver in picture is blinking his eyes for consecutive 20 frames or yawning mouth then application will alert driver with Drowsiness messages. We are using SVM pre-trained drowsiness model and then using Euclidean distance function we are continuously checking or predicting EYES and MOUTH distance closer to drowsiness, if distance is closer to drowsiness then application will alert driver.

1.3 ORGANIZATION OF CHAPTERS

1.INTRODUCTION

In this chapter, We described the Overview of our project which summarizes the existing and proposed part and Objective of our project which summarizes the goal of our project .

2.LITERATURE SURVEY

In this chapter, We broadly specified the Survey on Background which included the references that we surveyed and Conclusions on Survey which included the concepts that we have taken from the reference papers.

3.SOFTWARE AND HARDWARE REQUIREMENTS

In this chapter, We specified the requirements of Software and Hardware which are included in our project.

4.SOFTWARE DEVELOPMENT ANALYSIS

In this chapter, We described the Overview of our problem, Defined the problem. We also specified the Overview of Modules, Defined the Modules and showed the Functionality of the Modules.

5.PROJECT SYSTEM DESIGN

In this chapter, We showed the Unified Modeling Language(UML) diagrams of the system design. The UML Diagram include Usecase, Sequence, State Chart, Component and Deployment.

6.PROJECT CODING

In this chapter, We included Code templates, Outline for various files which includes all libraries, Class with Functionality and input and output parameters of methods.

7.PROJECT TESTING

In this chapter, We explained various test cases such as Unit, Integration, Functional and System testing. It also includes Black Box and White Box testings.

8.OUTPUT SCREENS

In this chapter, We showed the User Interface and Output Screens of our project.

9.EXPERIMENTAL RESULTS

In this chapter, We measured the performance of our algorithm based on accuracy metric.

10.CONCLUSION AND FUTURE ENHANCEMENTS

In this chapter, it covers the conclusion of our project and the possible future developments.

2.LITERATURE SURVEY

2.1 SURVEY ON BACKGROUND

[1]Ashish Kumar, Rusha Patra, Department of Electronics and Communication Engineering Indian Institute of Information Technology Guwahati, India – "Driver Drowsiness Monitoring System", IEEE Pattern Recognition,July.2018:-

This paper demonstrates a webcam based system to detect driver's fatigue from the face image only using image processing and machine learning techniques to make the system low-cost as well as portable.

[2]Gwak, J.S.; Shino, M.; Hirao, A. Early detection of driver drowsiness utilizing machine learning based on physiological signals, behavioral measures, and driving performance. In Proceedings of the IEEE International Conference on Intelligent Transportation Systems, Maui, Hawaii, HI, USA, 4–7 November 2018:-

This paper demonstrates the feasibility of driver drowsiness detection based on hybrid measures over a 10-second time period with high accuracy. It uses Random Forest Algorithm which is distinguishing between alert and slightly drowsy states.

[3]Awais, M.; Badruddin, N.; Drieberg, M. A Hybrid approach to detect driver drowsiness utilizing physiological signals to improve system performance and wearability. Sensors 2017:-

This paper demonstrates that an acceptable level of accuracy (80%) could be achieved by combining just two electrodes (one EEG and one ECG), indicating the feasibility of a system with improved wearability compared with existing systems involving many electrodes.

[4]A. Sengupta, A. Dasgupta, A. Chaudhuri, A. George, A. Routray, R.Guha; "A Multimodal System for Assessing Alertness Levels Due to Cognitive Loading", IEEE Trans. on Neural Systems and Rehabilitation Engg., vol. 25 (7), pp 1037-1046, 2017:-

This paper proposes a scheme for assessing the alertness levels of an individual using simultaneous acquisition of multimodal physiological signals and fusing the

information into a single metric for quantification of alertness using multivariate linear regression and analysis of variance.

[5] K. T. Chui, K. F. Tsang, H. R. Chi, B. W. K. Ling, and C. K. Wu, "An accurate ECG based transportation safety drowsiness detection scheme," IEEE Transactions on Industrial Informatics, vol. 12, no. 4, pp. 1438- 1452,Aug. 2016:-

This paper demonstrates the developed ECG GA-SVM provides an accurate and instantaneous warning to the drivers before they fall into sleep. As a result this ensures the public transport safety with real time implementations.

[6] Zhang, H.; Wua, C.; Yan, X.; Qiu, T.Z. The effect of fatigue driving on car following behavior. Transp. Res. Part F 2016:-

This paper demonstrates the impact of fatigue on driving behaviour in terms of car based on Karolinska Sleepiness Scale (KSS), the Percentage of eye Closures (PERCLOS) and the Time Headway (THW) which can be used as a measurement for monitoring fatigued drivers.

[7] W. L. Ou, M. H. Shih, C. W. Chang, X. H. Yu, C. P. Fan, "Intelligent Video-Based Drowsy Driver Detection System under Various Illuminations and Embedded Software Implementation", 2015 international Conf. on Consumer Electronics - Taiwan, 2015:-

This paper demonstrates the accuracy of the drowsy status detection is up to 91% by implementing on the FPGA-based embedded platform, the processing speed with the 640×480 format video is up to 16 frames per second (fps) after software optimizations.

[8] M. Karchani, A. Mazloumi, G. N. Saraji, A. Nahvi, K. S. Haghighi, B.M. Abadi, A. R. Foroshani, A. Niknezhad, "The Steps of Proposed Drowsiness Detection System Design based on Image Processing in SimulatorDriving",International Research Journal of Applied and Basic Sciences, vol. 9(6), pp 878-887, 2015:-

This paper demonstrates the early detection of sleepiness and prevents the irrecoverable losses by alarming.

[9] R.Ahmad, and J.N.Borole, "Drowsy Driver Identification Using Eye Blink Detection," IJISET - International Journal of Computer Science and Information Technologies, vol. 6, no. 1, pp. 270-274, Jan. 2015:-

This paper demonstrates the eye blinks via a standard webcam in real-time at 110fps for a 320×240 resolution with a 94% accuracy with a 1% false positive rate.

[10] Loon, R.J.;Brouwer,R.F.T.;Martens, M.H. Drowsy drivers'under-performance in lateral control: How much is too much? Using an integrated measure of lateral control to quantify safe lateral driving. Accid. Anal. Prev. 2015:-

This paper demonstrates the different levels of drowsiness by validating it to two established drowsiness metrics (KSS and PERCLOS) using level of drowsiness as a surrogate for safety we are then able to set simple criteria for safe and unsafe lateral control performance, based on individual driving behaviour.

[11] V.Kazemi and J. Sullivan; "One millisecond face alignment with an ensemble of regression trees", IEEE Conf. on Computer Vision and Pattern Recognition, 23-28 June, 2014, Columbus, OH, USA:-

This paper demonstrates the effect of the quantity of training data on the accuracy of the predictions and explore the effect of data augmentation using synthesized data.

[12] Correa, A.G.; Orosco, L.; Laciar, E. Automatic detection of drowsiness in EEG records based on multimodal analysis. Med Eng.Phys. 2014:-

This paper demonstrates the automatic drowsiness detection system in vehicles, thereby decreasing the rate of accidents caused by sleepiness of the driver with 87.4% and 83.6% of alertness and drowsiness correct detections rates, respectively.

[13] A. Abas, J. Mellor, and X. Chen, "Non-intrusive drowsiness detection by employing Support Vector Machine," 2014 20th International Conference on Automation and Computing (ICAC), Bedfordshire, UK,2014:-

This paper demonstrates the Support Vector Machine (SVM) to train the classifier by using steering wheel angle and distance to outside lane as input parameters to the SVM. All the parameters extracted from vehicle parametrical data collected in a driving simulator.

[14] B. Alshaqaqi, A. S. Baquhaizel, M. E. A. Ouis, M. Bouumehed, A.Ouamri, M. Keche, "Driver Drowsiness Detection System", IEEE International Workshop on Systems, Signal Processing and their Applications, 2013:-

This paper demonstrates the a module for Advanced Driver Assistance System (ADAS) is presented to reduce the number of accidents due to drivers fatigue and hence increase the transportation safety.

[15] Fu, C.L.; Li, W.K.; Chun, H.C.; Tung, P.S.; Chin, T.L. Generalized EEG-based drowsiness prediction system by using a self-organizing neural fuzzy system. IEEE Transection Circuits Syst. 2012:-

This paper demonstrates the generalized EEG-based Self-organizing Neural Fuzzy system to monitor and predict the driver's drowsy state with the occipital area. Two drowsiness prediction models, subject-dependent and generalized cross-subject predictors, were investigated in this study for system performance analysis.

2.2 CONCLUSIONS ON SURVEY

In this project, a low cost, real time driver drowsiness monitoring system has been proposed based on visual behavior and machine learning. Here, visual behavior features like eye aspect ratio and mouth opening ratio are computed from the streaming video, captured by a webcam. The developed system works accurately with the generated synthetic data. Subsequently, the feature values are stored and machine learning algorithm have been used for classification. SVM algorithm has been explored here.

3.SOFTWARE AND HARDWARE REQUIREMENTS

3.1 SOFTWARE REQUIREMENTS

Operating System : Windows 10.

Platform : PYTHON TECHNOLOGY

Tool : Pycharm, Python 3.5

Front End : Anaconda

Back End : python anaconda script

Libraries : Numpy,Scipy,Imutils,Dlib,cv2,tkinter

3.2 HARDWARE REQUIREMENTS

System : Pentium IV 2.4 GHz.

Hard Disk : 40 GB.

Monitor : 15 inch VGA Color.

Mouse : Standard Mouse

Ram : 512 MB

Keyboard : Standard Keyboard

4.SOFTWARE DEVELOPMENT ANALYSIS

4.1 OVERVIEW OF PROBLEM

In driver drowsiness monitoring system, the video is recorded using a webcam. The camera will be positioned in front of the driver to capture the front face image. From the video, the frames are extracted to obtain 2-D images. Face is detected in the frames using histogram of oriented gradients(HOG) and linear support vector machine (SVM) for object detection. After detecting the face, facial landmarks like positions of eye, and mouth are marked on the images. From the facial landmarks, eye aspect ratio and mouth opening ratio are quantified and using these features and machine learning approach, a decision is obtained about the drowsiness of the driver. If drowsiness is detected, an alarm will be sent to the driver to alert him/her.

4.2 DEFINE THE PROBLEM

In this project by monitoring Visual Behaviour of a driver with webcam and machine learning SVM (support vector machine) algorithm we are detecting Drowsiness in a driver. This application will use inbuilt webcam to read pictures of a driver and then using OPENCV SVM algorithm extract facial features from the picture and then check whether driver in picture is blinking his eyes for consecutive 20 frames or yawning mouth then application will alert driver with Drowsiness messages. We are using SVM pre-trained drowsiness model and then using Euclidean distance function we are continuously checking or predicting EYES and MOUTH distance closer to drowsiness, if distance is closer to drowsiness then application will alert driver.

4.3 MODULES OVERVEIW

- Data Acquisition
- Face Detection
- Facial Land marking
- Feature Extraction
- Classification

4.4 DEFINE THE MODULES

Data Acquisition:

The video is recorded using webcam and the frames are extracted and processed in a laptop. After extracting the frames ,image processing techniques are applied on these 2D images. Presently, synthetic driver data has been generated. The volunteers are asked to look at the webcam with intermittent eye blinking , eye closing, yawing. The video is captured for 30 minutes duration.

Face Detection:

After extracting the frames, first the human faces are detected. Numerous online face detection algorithms are there. In this study, histogram of oriented gradients (HOG) and linear SVM method is used.

In this method, positive samples of fixed window size are taken from the images and HOG descriptors are computed on them. Subsequently, negative samples (samples that do not contain the required object to be detected i.e., human face here) of same size are taken and HOG descriptors are calculated. Usually the number of negative samples is very greater than number of positive samples.

After obtaining the features for both the classes, a linear SVM is trained for the classification task. To improve the accuracy of SVM, hard negative mining is used.

In this method, after training, the classifier is tested on the labeled data and the false positive sample feature values are used again for training purpose. For the test

image, the fixed size window is translated over the image and the classifier computes the output for each window location.

Finally, the maximum value output is considered as the detected face and a bounding box is drawn around the face. This non-maximum suppression step removes the redundant and overlapping bounding boxes.

Facial Land Marking:

After detecting the face, the next task is to find the locations of different facial features like the corners of the eyes and mouth so on. Prior to that, the face images should be normalized in order to reduce the effect of distance from the camera, non-uniform illumination and varying image resolution. Therefore, the face image is resized to a width of 500 pixels and converted to grayscale image.

After image normalization, ensemble of regression trees is used to estimate the landmark positions on face from a sparse subset of pixel intensities. In this method, the sum of square error loss is optimized using gradient boosting learning.

The facial landmarks are shown in below Fig. The red points are the detected landmarks for further processing.

Using this method, the boundary points of eyes, mouth are marked and the number of points for eye, mouth.

Parts	Landmark points
Mouth	[13-24]
Right eye	[1-6]
Left eye	[7-12]

Table 1.1:Facial Landmark Points

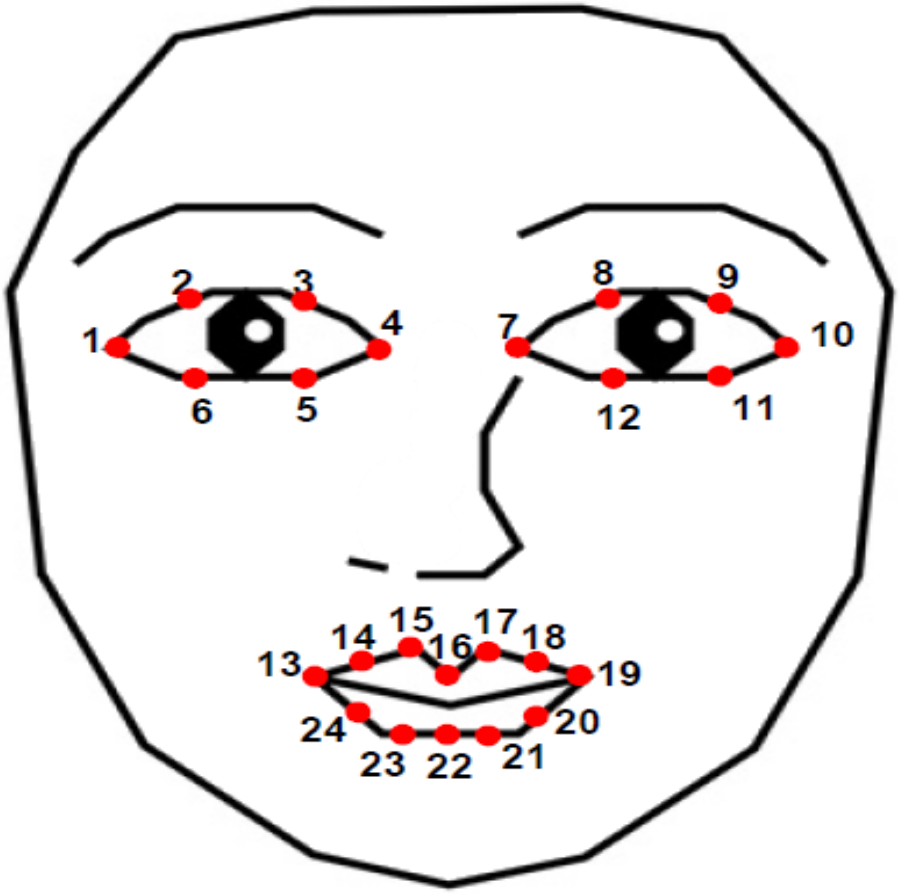


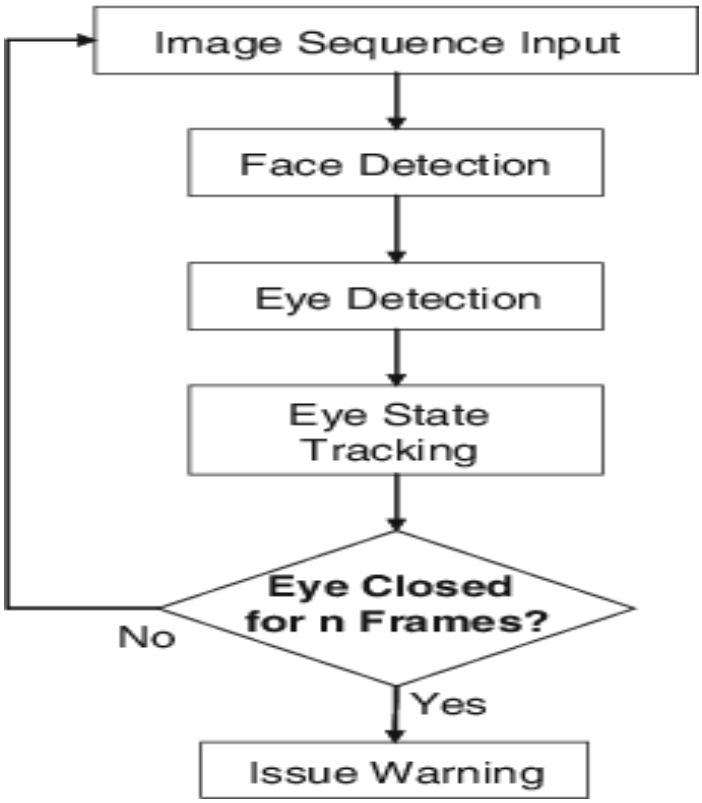
Fig 1.1 : The Facial Landmark Points

Feature Extraction:

After detecting the facial landmarks, the features are computed as described below.

Eye aspect ratio (EAR): From the eye corner points, the eye aspect ratio is calculated as the ratio of height and width of the eye as given by

$$EAR = \frac{(p_2 - p_6) + (p_3 - p_5)}{2(p_4 - p_1)}$$



Mouth opening ratio (MOR): Mouth opening ratio is a parameter to detect yawning during drowsiness.

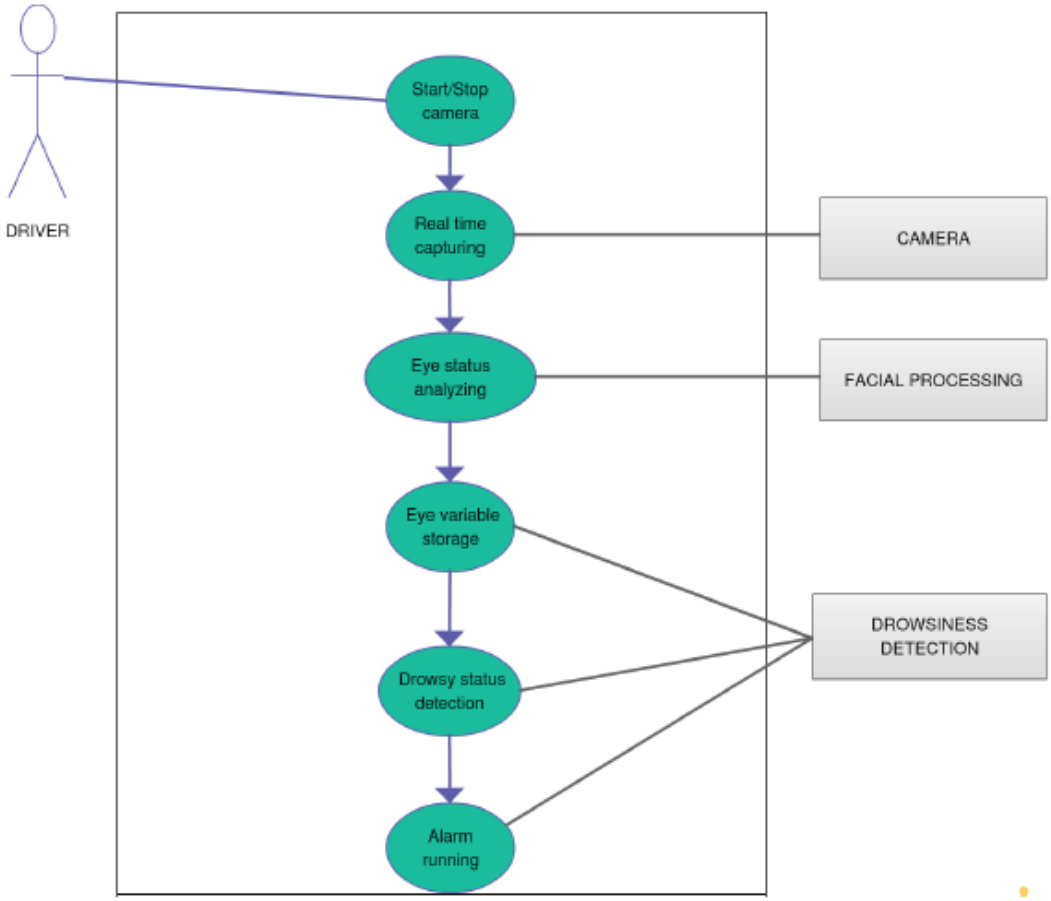
Similar to EAR, it is calculated as

$$MOR = \frac{(p_{15} - p_{23}) + (p_{16} - p_{22}) + (p_{17} - p_{21})}{3(p_{19} - p_{13})}$$

Classification:

After computing all the three features, the next task is to detect drowsiness in the extracted frames.

In the setup phase, the EAR values for first three hundred (for 10s at 30 fps) frames are recorded. Out of these three hundred initial frames containing face, average of 150 maximum values is considered as the hard threshold for EAR.



The higher values are considered so that no eye closing instances will be present. If the test value is less than this threshold, then eye closing (i.e., drowsiness) is detected.

Similarly, for calculating threshold of MOR, since the mouth may not be open to its maximum in initial frames (setup phase) so the threshold is taken experimentally from the observations. If the test value is greater than this threshold then yawn (i.e., drowsiness) is detected.

The system detects the drowsiness if in a test frame drowsiness is detected for at least one feature. To make this thresholding more realistic, the decision for each frame depends on the last 75 frames. If at least 70 frames (out of those 75) satisfy drowsiness conditions for at least one feature, then the system gives drowsiness detection indication and the alarm.

State	EAR	MOR
Normal	0.33	0.42
Eyes Closed	0.15	0.42
Eyes Open, Yawing	0.32	0.83
Eyes closed, Yawing	0.12	0.78

Table 1.2 Sample values of different parameters for different states

4.5 MODULE FUNCTIONALITY

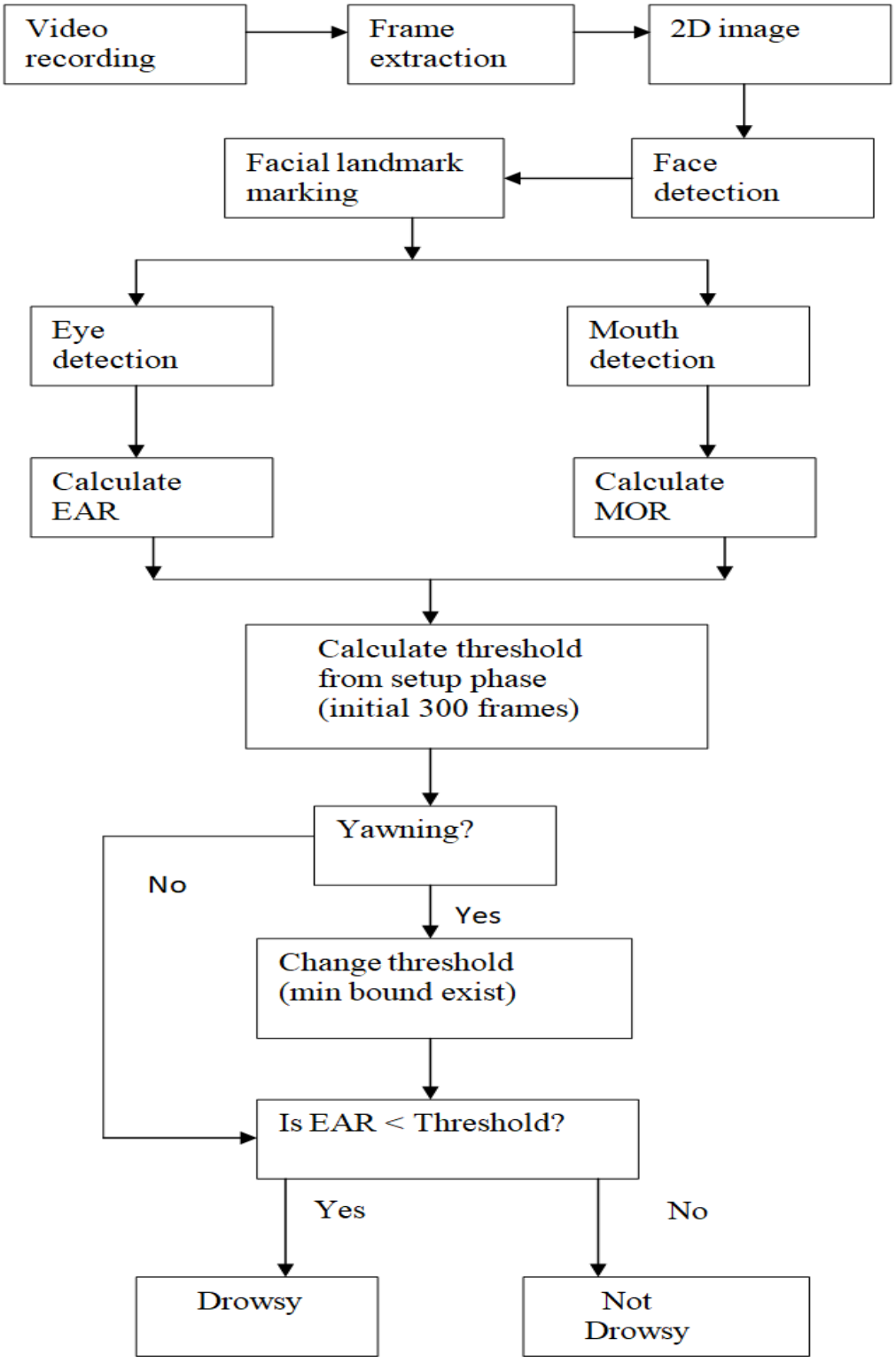
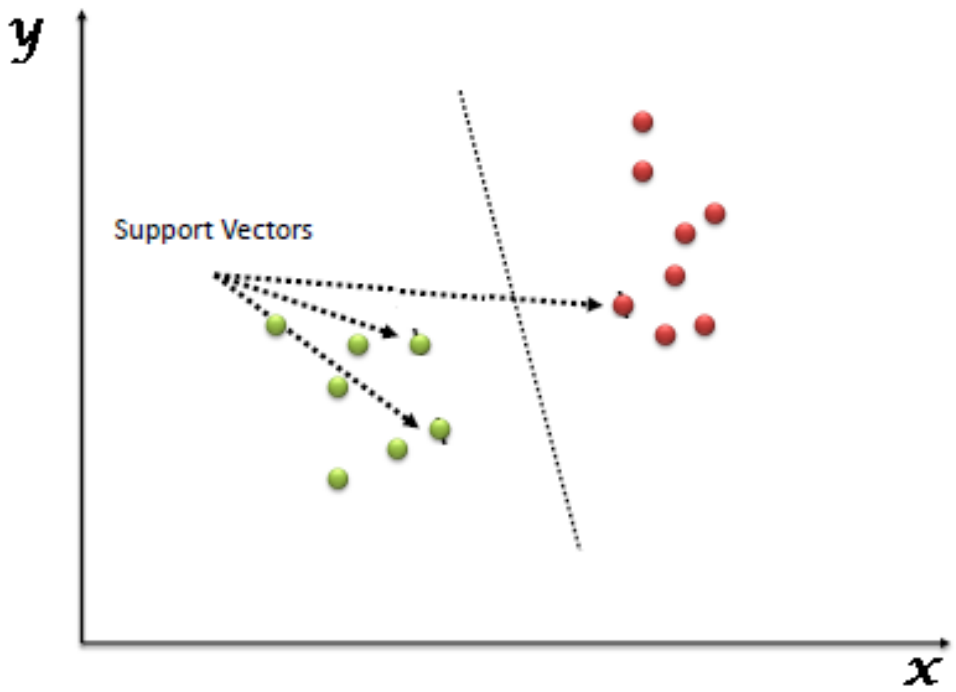


Fig.1.2 The Block Diagram of Proposed Drowsiness Detection System

SVM ALGORITHM:

Machine learning involves predicting and classifying data and to do so we employ various machine learning algorithms according to the dataset. SVM or Support Vector Machine is a linear model for classification and regression problems. It can solve linear and non-linear problems and work well for many practical problems. The idea of SVM is simple: The algorithm creates a line or a hyperplane which separates the data into classes. In machine learning, the radial basis function kernel, or RBF kernel, is a popular kernel function used in various kernelized learning algorithms. In particular, it is commonly used in support vector machine classification. As a simple example, for a classification task with only two features (like the image above), you can think of a hyperplane as a line that linearly separates and classifies a set of data.

Intuitively, the further from the hyperplane our data points lie, the more confident we are that they have been correctly classified. We therefore want our data points to be as far away from the hyperplane as possible, while still being on the correct side of it.

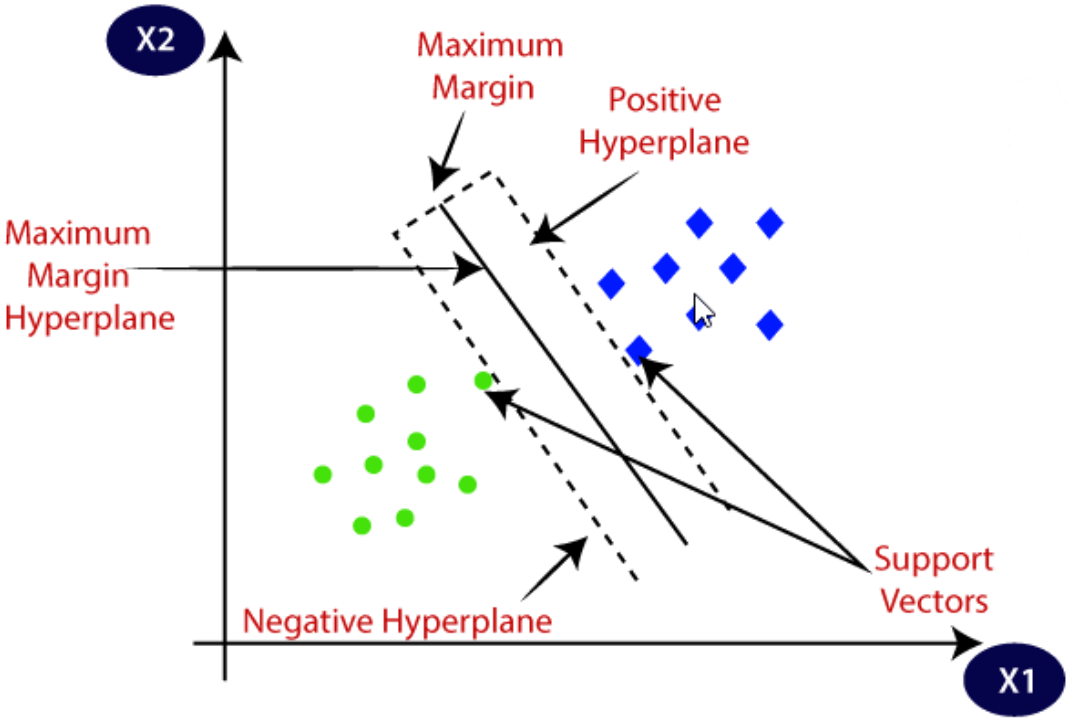


So when new testing data is added, whatever side of the hyperplane it lands will decide the class that we assign to it.

How do we find the right hyperplane?

Or, in other words, how do we best segregate the two classes within the data?

The distance between the hyperplane and the nearest data point from either set is known as the margin. The goal is to choose a hyperplane with the greatest possible margin between the hyperplane and any point within the training set, giving a greater chance of new data being classified correctly.



Linear SVM:

Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier.

Hyperplane:

There can be multiple lines/decision boundaries to segregate the classes in n-dimensional space, but we need to find out the best decision boundary that helps to classify the data points. This best boundary is known as the hyperplane of SVM.

The dimensions of the hyperplane depend on the features present in the dataset, which means if there are 2 features (as shown in image), then hyperplane will be a straight line. And if there are 3 features, then hyperplane will be a 2-dimension plane.

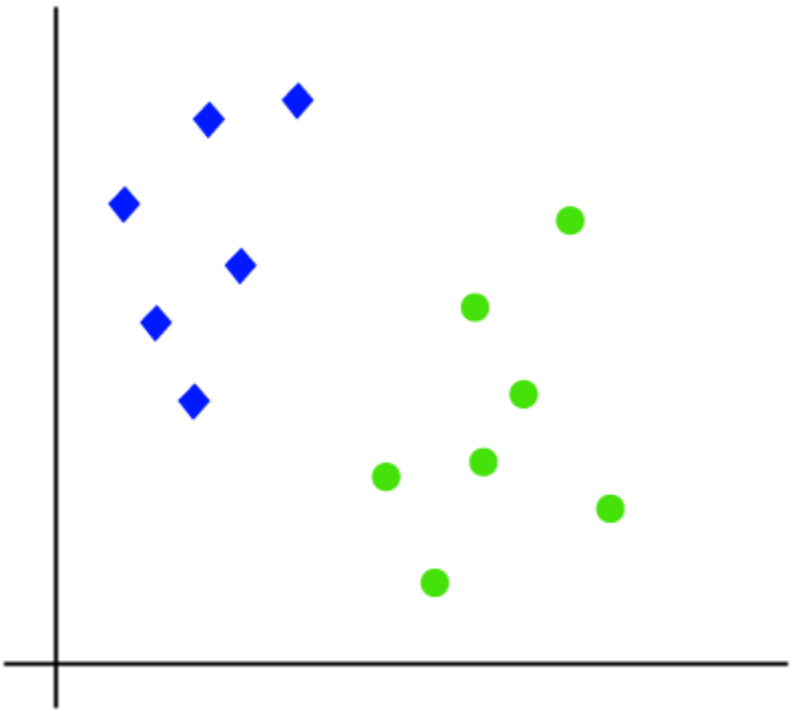
We always create a hyperplane that has a maximum margin, which means the maximum distance between the data points.

Support Vectors:

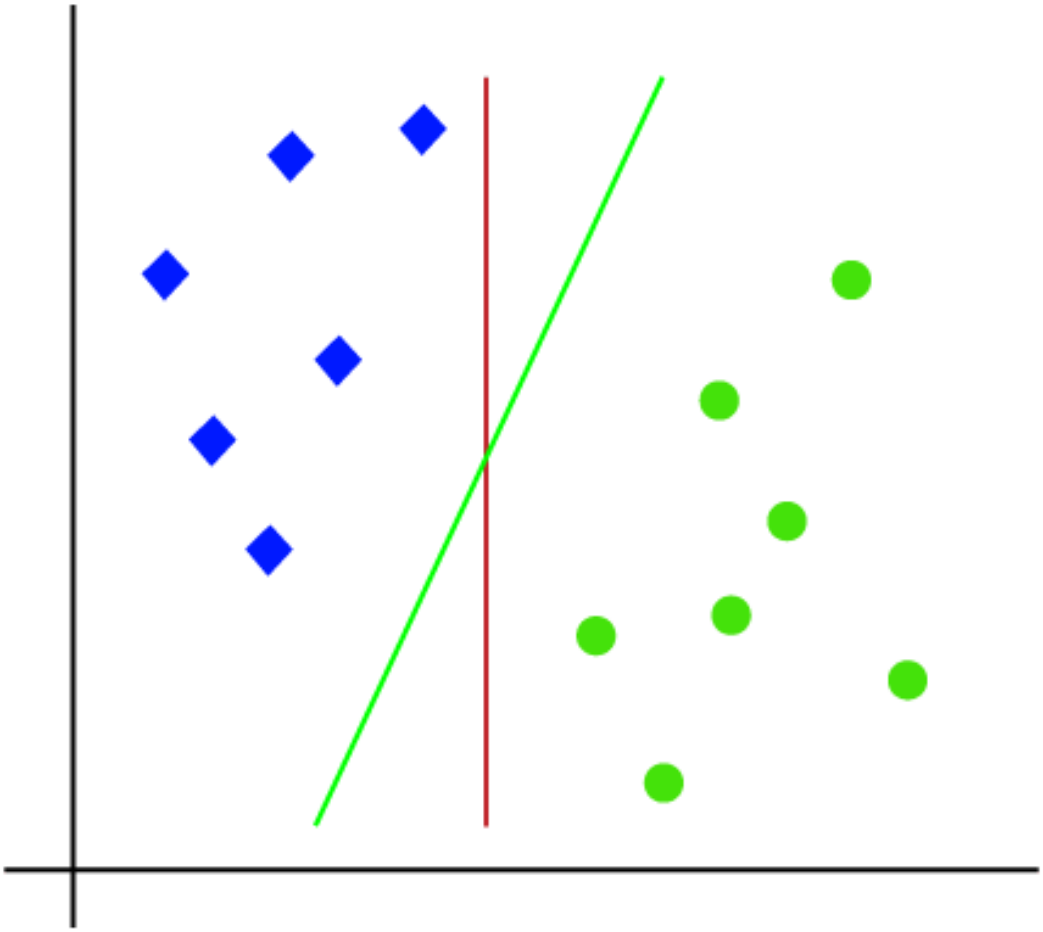
The data points or vectors that are the closest to the hyperplane and which affect the position of the hyperplane are termed as Support Vector. Since these vectors support the hyperplane, hence called a Support vector.

Linear SVM:

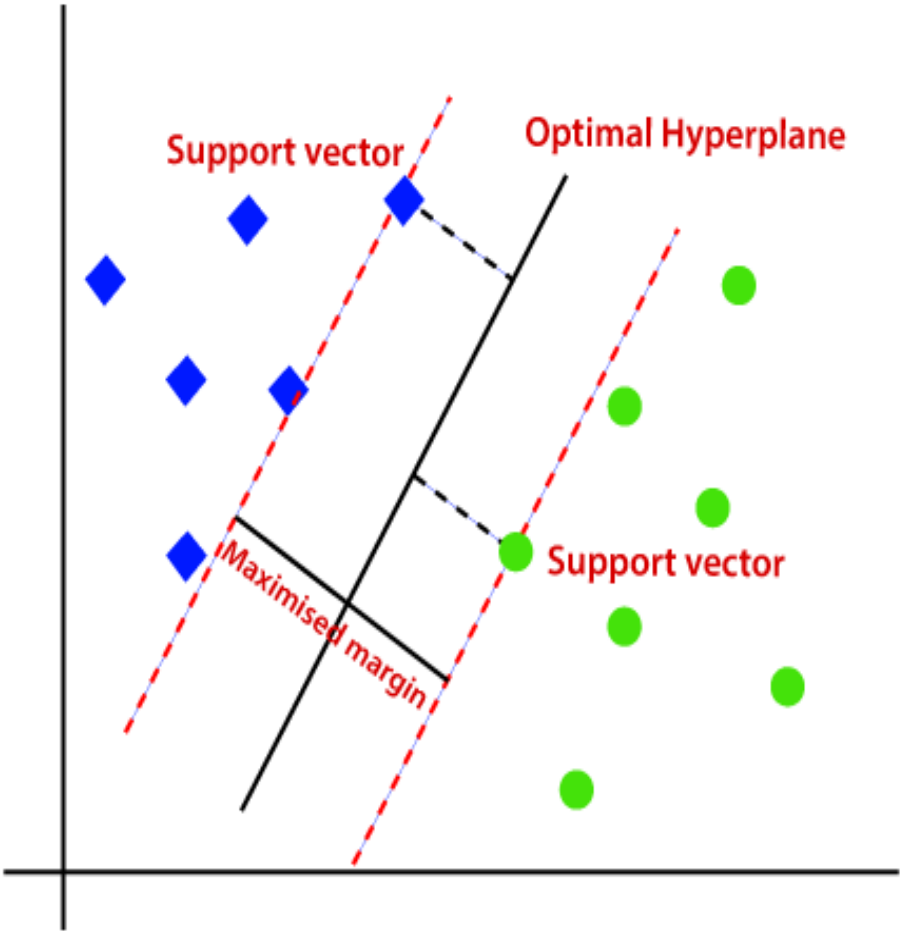
The working of the SVM algorithm can be understood by using an example. Suppose we have a dataset that has two tags (green and blue), and the dataset has two features x1 and x2. We want a classifier that can classify the pair(x1, x2) of coordinates in either green or blue. Consider the below image:



So as it is 2-d space so by just using a straight line, we can easily separate these two classes. But there can be multiple lines that can separate these classes. Consider the below image:



Hence, the SVM algorithm helps to find the best line or decision boundary; this best boundary or region is called as a **hyperplane**. SVM algorithm finds the closest point of the lines from both the classes. These points are called support vectors. The distance between the vectors and the hyperplane is called as **margin**. And the goal of SVM is to maximize this margin. The **hyperplane** with maximum margin is called the **optimal hyperplane**.



HISTOGRAM OF ORIENTED GRADIENTS

HOG, or Histogram of Oriented Gradients, is a feature descriptor that is often used to extract features from image data. It is widely used in computer vision tasks for object detection.

- The HOG descriptor focuses on the structure or the shape of an object. Now you might ask, how is this different from the edge features we extract for images? In the case of edge features, we only identify if the pixel is an edge or not. HOG is able to provide the edge direction as well. This is done by extracting the gradient and orientation (or you can say magnitude and direction) of the edges

- Additionally, these orientations are calculated in 'localized' portions. This means that the complete image is broken down into smaller regions and for each region, the gradients and orientation are calculated.
- Finally the HOG would generate a Histogram for each of these regions separately. The histograms are created using the gradients and orientations of the pixel values, hence the name 'Histogram of Oriented Gradients'

Step1:

Sample *P* positive samples from your training data of the object(s) you want to detect and extract HOG descriptors from these samples.

Step2:

Sample *N* negative samples from a negative training set that does not contain any of the objects you want to detect and extract HOG descriptors from these samples as well. In practice $N >> P$.

Step3:

Train a Linear Support Vector Machine on your positive and negative samples.

Step4:

Apply hard-negative mining. For each image and each possible scale of each image in your negative training set, apply the sliding window technique and slide your window across the image. At each window compute your HOG descriptors and apply your classifier. If your classifier (incorrectly) classifies a given window as an object (and it will, there will absolutely be false-positives), record the feature vector associated with the false-positive patch along with the probability of the classification. This approach is called hard-negative mining.

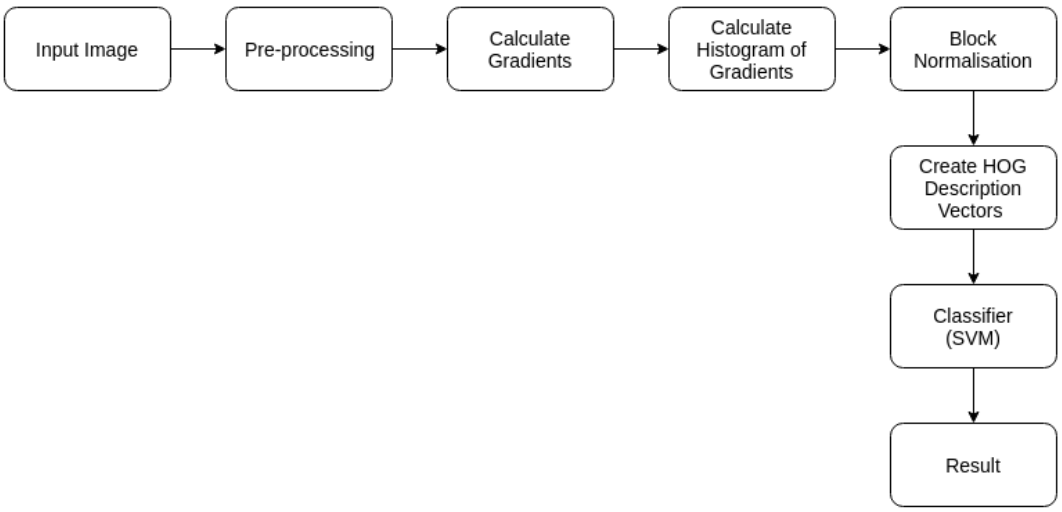
Step5:

Take the false-positive samples found during the hard-negative mining stage, sort them by their confidence (i.e. probability) and re-train your classifier using these hard-negative samples.

Step6:

Your classifier is now trained and can be applied to your test dataset. Again, just like in Step 4, for each image in your test set, and for each scale of the image, apply the sliding window technique. At each window extract HOG descriptors and apply your classifier. If your classifier detects an object with sufficiently large probability, record the bounding box of the window. After you have finished scanning the image, apply non-maximum suppression to remove redundant and overlapping bounding boxes.

Workflow of object detection using HOG



5.PROJECT SYSTEM DESIGN

5.1 UML DIAGRAMS

USE CASE DIAGRAM

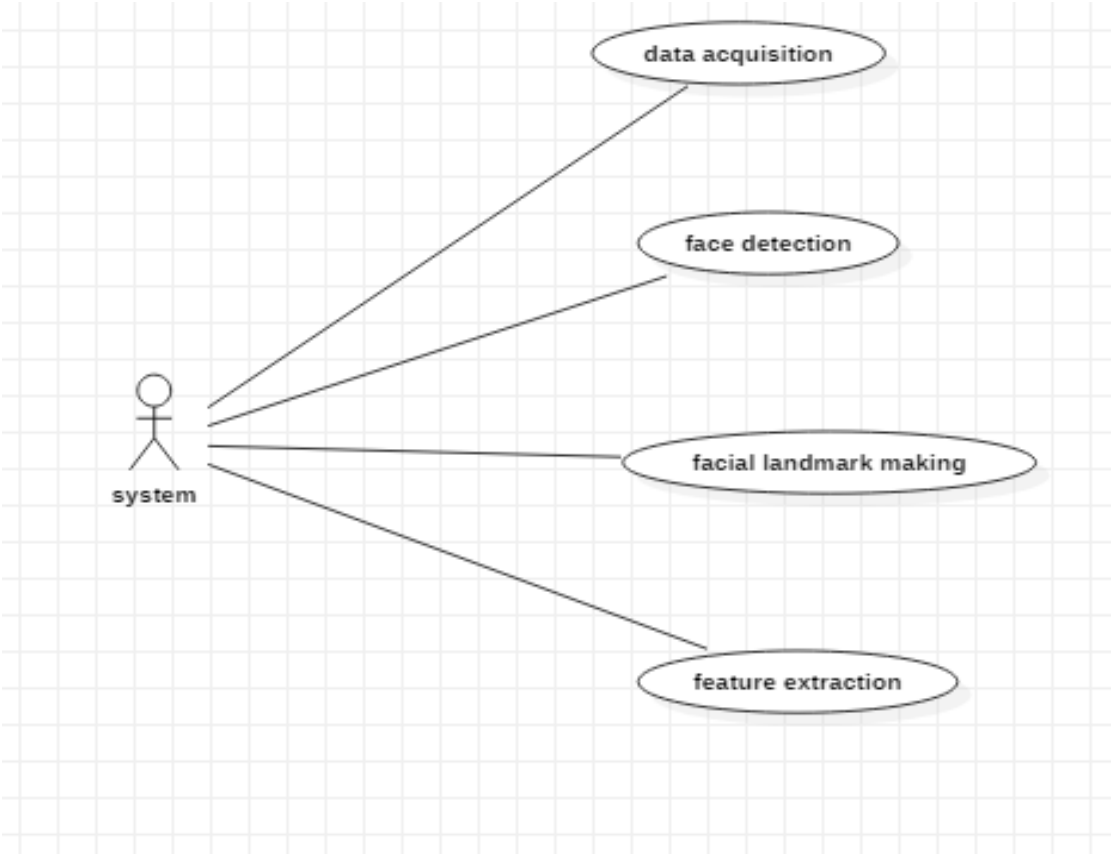


Fig 2.1 Use Case Diagram

SEQUENCE DIAGRAM

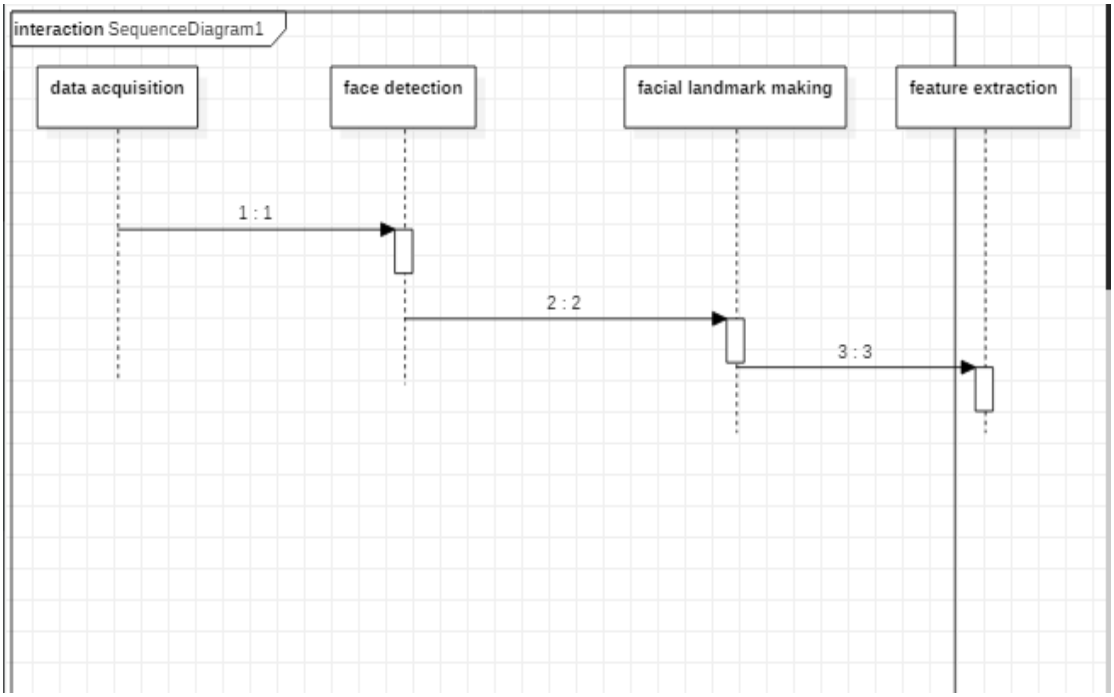


Fig 2.2 Sequence Diagram

STATE CHART DIAGRAM

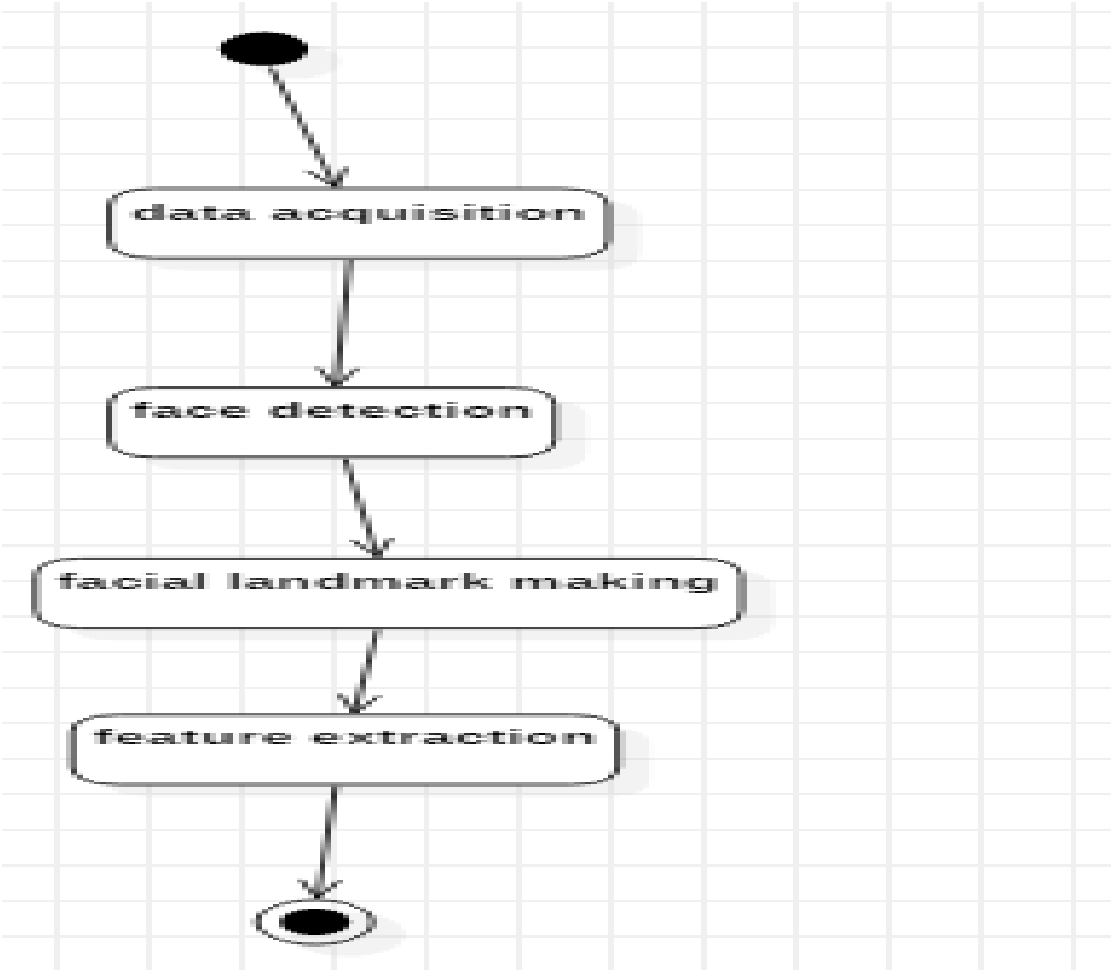


Fig 3.1 State Chart Diagram

COMPONENT DIAGRAM

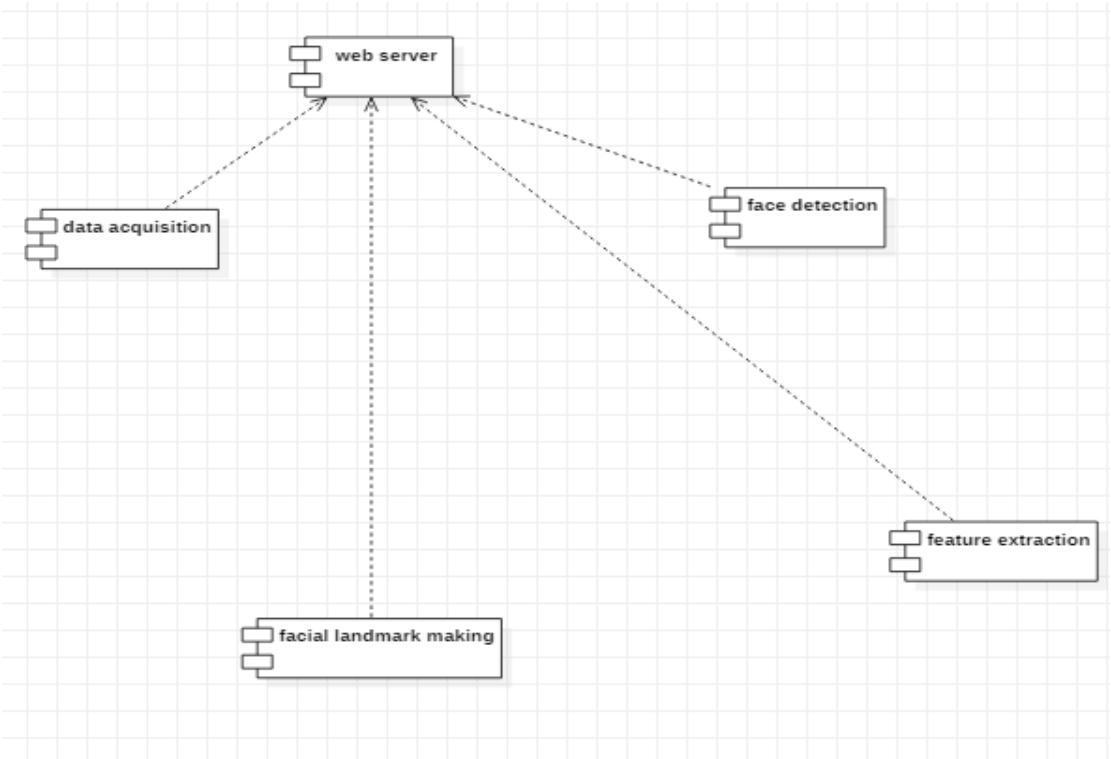


Fig 3.2 Component Diagram

DEPLOYMENT DIAGRAM

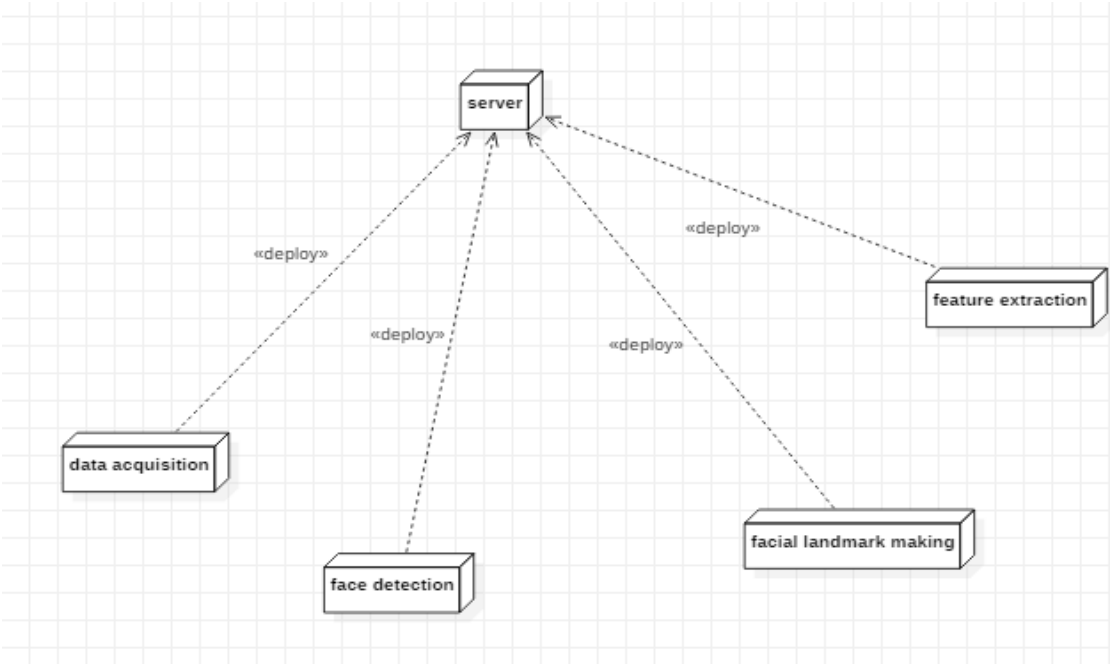


Fig 4.1 Deployment Diagram

CLASS DIAGRAM

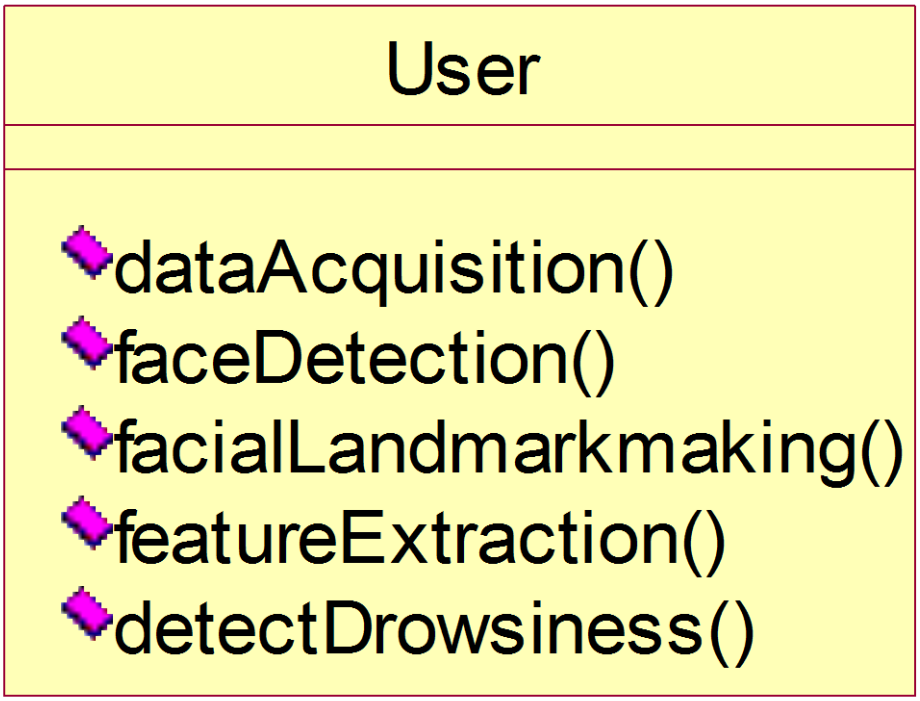


Fig 4.2 Class Diagram

6.PROJECT CODING

6.1 CODE TEMPLATES

Use Interface Tinker Page Template:

```
main = tkinter.Tk()

main.title("Driver Drowsiness Monitoring")

main.geometry("500x400")

font = ('times', 16, 'bold')

title = Label(main, text='Driver Drowsiness Monitoring System using Visual\n Behaviour and Machine Learning',anchor=W, justify=LEFT)

title.config(bg='black', fg='white')

title.config(font=font)

title.config(height=3, width=120)

title.place(x=0,y=5)

font1 = ('times', 14, 'bold')

upload = Button(main, text="Start Behaviour Monitoring Using Webcam", command=startMonitoring)

upload.place(x=50,y=200)

upload.config(font=font1)

pathlabel = Label(main)

pathlabel.config(bg='DarkOrange1', fg='white')

pathlabel.config(font=font1)

pathlabel.place(x=50,y=250)

main.config(bg='chocolate1')

main.mainloop()
```

Eye Aspect Ratio Function Template:

```
def EAR(drivereye):
    point1 = dist.euclidean(drivereye[1], drivereye[5])
    point2 = dist.euclidean(drivereye[2], drivereye[4])
    # compute the euclidean distance between the horizontal
    distance = dist.euclidean(drivereye[0], drivereye[3])
    # compute the eye aspect ratio
    ear_aspect_ratio = (point1 + point2) / (2.0 * distance)
    return ear_aspect_ratio
```

Mouth Opening Ratio Function Template:

```
def MOR(drivermouth):
    # compute the euclidean distances between the horizontal
    point   = dist.euclidean(drivermouth[0], drivermouth[6])
    # compute the euclidean distances between the vertical
    point1  = dist.euclidean(drivermouth[2], drivermouth[10])
    point2  = dist.euclidean(drivermouth[4], drivermouth[8])
    # taking average
    Ypoint   = (point1+point2)/2.0
    # compute mouth aspect ratio
    mouth_aspect_ratio = Ypoint/point
    return mouth_aspect_ratio
```

6.2 OUTLINE FOR VARIOUS FILES:

Import Libraries:

```
from tkinter import *

import tkinter

from scipy.spatial import distance as dist

from imutils import face_utils

import numpy as np

import imutils

import dlib

import cv2
```

6.3 CLASS WITH FUNCTIONALITY:

Functions included in our project are…

EAR(Eye Aspect Ratio) Function:

In this function,We calculate the Eye Aspect Ratio based on Euclidean Distance between facial marks of Eyes.Then compare the EAR value with the Threshold value.In our project, threshold value of EAR is 0.25.If EAR value is lesser than the Eye's threshold value then it displays the alert message to the driver. .If EAR value is greater than the Eye's threshold value then it indicates that driver is not drowsy.

MOR(Mouth Opening Ratio) Function:

In this function,We calculate the Mouth Opening Ration based on Euclidean Distance between facial marks of Mouth.Then compare the MAR value with the Threshold value.In our project, threshold value of MAR is 0.75.If MAR value is greater than the Mouth's threshold value then it displays the alert message to the driver. .If MAR value is lesser than the Mouth's threshold value then it indicates that driver is not drowsy.

6.4 METHODS INPUT AND OUTPUT PARAMETERS

```
def startMonitoring():
    pathlabel.config(text="          Webcam Connected Successfully")
    webcamera = cv2.VideoCapture(0)
    svm_predictor_path = 'SVMclassifier.dat'
    EYE_AR_THRESH = 0.25
    EYE_AR_CONSEC_FRAMES = 10
    MOU_AR_THRESH = 0.75
    COUNTER = 0
    yawnStatus = False
    yawns = 0
    svm_detector = dlib.get_frontal_face_detector()
    svm_predictor = dlib.shape_predictor(svm_predictor_path)
    (lStart, lEnd) = face_utils.FACIAL_LANDMARKS_IDXS["left_eye"]
    (rStart, rEnd) = face_utils.FACIAL_LANDMARKS_IDXS["right_eye"]
    (mStart, mEnd) = face_utils.FACIAL_LANDMARKS_IDXS["mouth"]
    while True:
        ret, frame = webcamera.read()
        frame = imutils.resize(frame, width=640)
        gray = cv2.cvtColor(frame, cv2.COLOR_BGR2GRAY)
        prev_yawn_status = yawnStatus
        rects = svm_detector(gray, 0)
        for rect in rects:
```

```
        shape = svm_predictor(gray, rect)
        shape = face_utils.shape_to_np(shape)
        leftEye = shape[lStart:lEnd]
        rightEye = shape[rStart:rEnd]
        mouth = shape[mStart:mEnd]
        leftEAR = EAR(leftEye)
        rightEAR = EAR(rightEye)
        mouEAR = MOR(mouth)
        ear = (leftEAR + rightEAR) / 2.0
        leftEyeHull = cv2.convexHull(leftEye)
        rightEyeHull = cv2.convexHull(rightEye)
        mouthHull = cv2.convexHull(mouth)
        cv2.drawContours(frame, [leftEyeHull], -1, (0, 255, 255), 1)
        cv2.drawContours(frame, [rightEyeHull], -1, (0, 255, 255), 1)
        cv2.drawContours(frame, [mouthHull], -1, (0, 255, 0), 1)
        if ear < EYE_AR_THRESH:
            COUNTER += 1
            cv2.putText(frame, "Eyes Closed ", (10,
30),cv2.FONT_HERSHEY_SIMPLEX, 0.7, (0, 0, 255), 2)
            if COUNTER >= EYE_AR_CONSEC_FRAMES:
                cv2.putText(frame, "DROWSINESS ALERT!", (10,
50),cv2.FONT_HERSHEY_SIMPLEX, 0.7, (0, 0, 255), 2)
        else:
            COUNTER = 0
```

```
            cv2.putText(frame, "Eyes Open ", (10,
30),cv2.FONT_HERSHEY_SIMPLEX, 0.7, (0, 255, 0), 2)

        cv2.putText(frame, "EAR: {:.2f}".format(ear), (480,
30),cv2.FONT_HERSHEY_SIMPLEX, 0.7, (0, 0, 255), 2)

        if mouEAR > MOU_AR_THRESH:

            cv2.putText(frame, "Yawning, DROWSINESS ALERT! ", (10,
70),cv2.FONT_HERSHEY_SIMPLEX, 0.7, (0, 0, 255), 2)

            yawnStatus = True

            output_text = "Yawn Count: " + str(yawns + 1)

            cv2.putText(frame, output_text,
(10,100),cv2.FONT_HERSHEY_SIMPLEX, 0.7,(255,0,0),2)

        else:

            yawnStatus = False

        if prev_yawn_status == True and yawnStatus == False:

            yawns+=1

        cv2.putText(frame, "MAR: {:.2f}".format(mouEAR), (480,
60),cv2.FONT_HERSHEY_SIMPLEX, 0.7, (0, 0, 255), 2)

        cv2.putText(frame,"Visual Behaviour & Machine Learning Drowsiness
Detection                                                    @
Drowsiness",(370,470),cv2.FONT_HERSHEY_COMPLEX,0.6,(153,51,102),1)

      cv2.imshow("Frame", frame)

      key = cv2.waitKey(1) & 0xFF

      if key == ord("q"):

          break

  cv2.destroyAllWindows()

  webcamera.release()
```

7.PROJECT TESTING

7.1 VARIOUS TEST CASES

UNIT TESTING

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

INTEGRATION TESTING:

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

FUNCTIONAL TESTING:

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items:

Valid Input : identified classes of valid input must be accepted.

Invalid Input : identified classes of invalid input must be rejected.

Functions : identified functions must be exercised.

Output : identified classes of application outputs must be exercised.

Systems/Procedures : interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

SYSTEM TESTING:

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

7.2 BLACK BOX TESTING

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box .you cannot “see” into it. The test provides inputs and responds to outputs without considering how the software works.

7.3 WHITE BOX TESTING

White Box Testing is a testing in which in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is purpose. It is used to test areas that cannot be reached from a black box level.

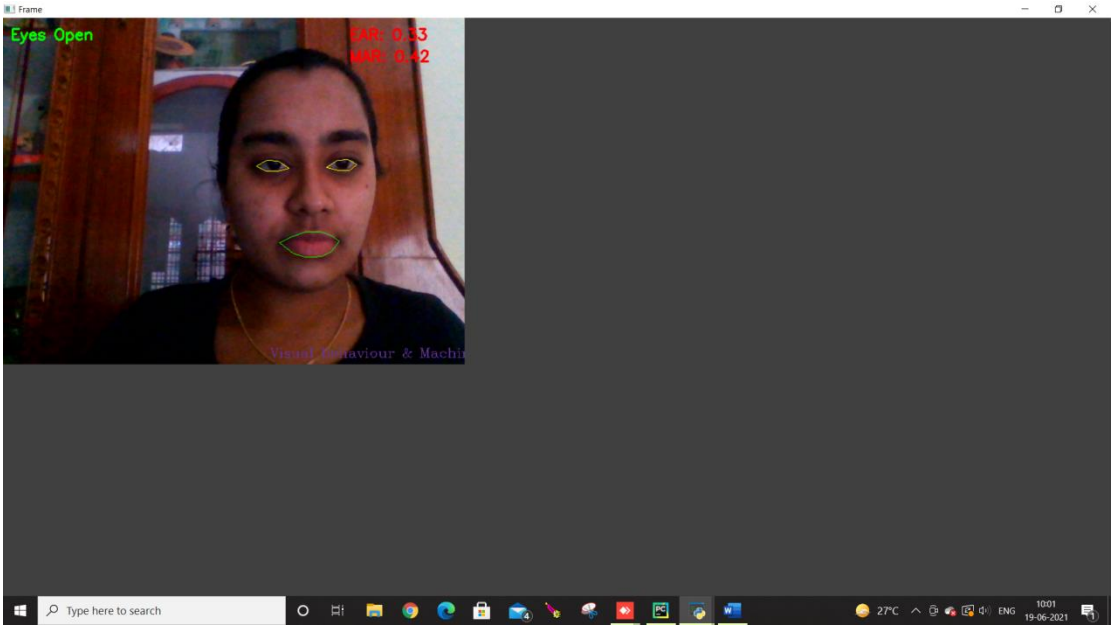
8.OUTPUT SCREENS

8.1 USER INTERFACES

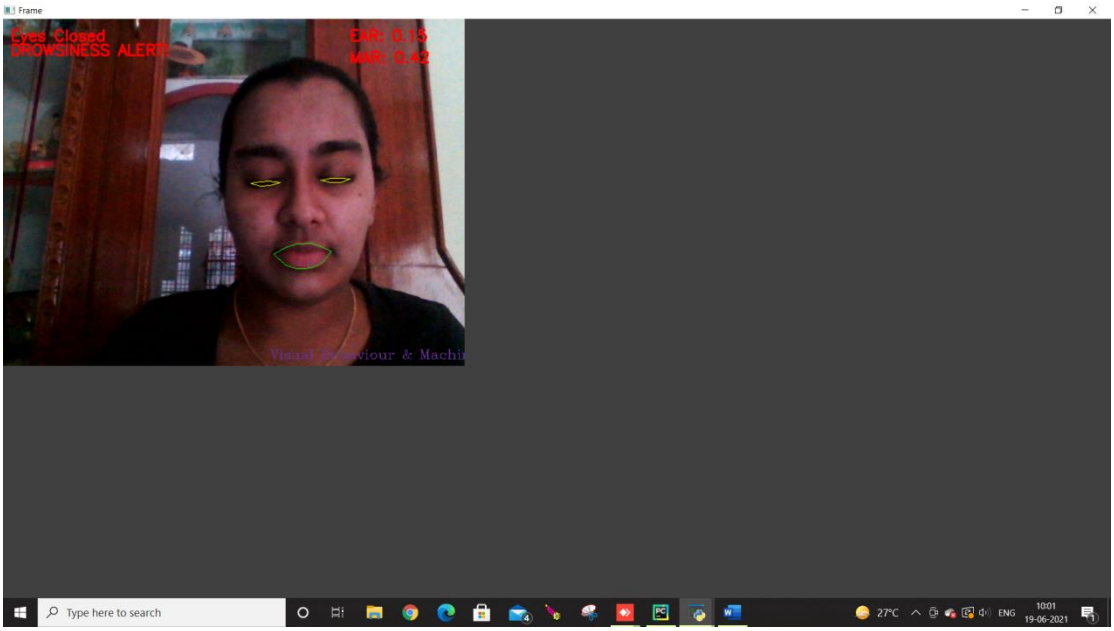
In below screen click on ‘Start Behaviour Monitoring Using Webcam’ button to connect application with webcam, after clicking button will get below screen with webcam streaming

8.2 OUTPUT SCREENS

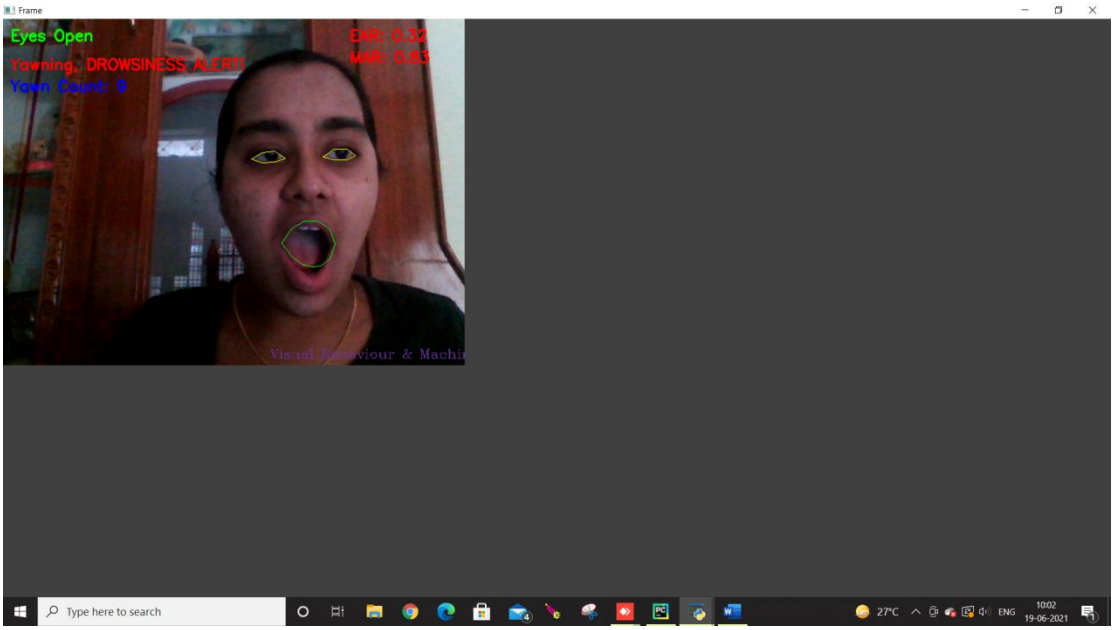
In below screen we can see web cam stream then application monitor all frames to see person eyes are open or not.



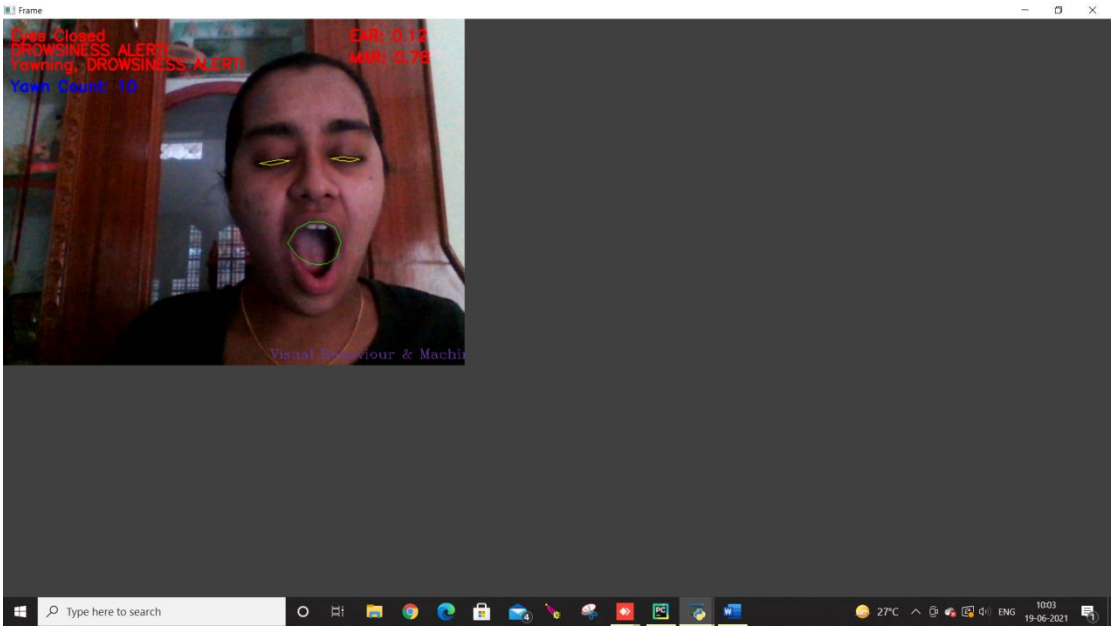
If eyes are closed then will get below message



Similarly if mouth starts yawn or eyes open then also will get alert message.



Similarly if mouth starts yawn and eyes close then also will get alert message.



9.EXPERIMENTAL RESULTS

Sensitivity is calculated as the ratio of correctly classifying drowsy states out of all actual drowsy states and specificity is computed as the ratio of correctly classifying awake states out of all actual awake states.We got the overall accuracy of SVM Classifier as 95%.Sensitivity and Specificity as 0.956 and 1.

10. CONCLUSION AND FUTURE ENHANCEMENT

In this project, a low cost, real time driver drowsiness monitoring system has been proposed based on visual behavior and machine learning. Here, visual behavior features like eye aspect ratio and mouth opening ratio are computed from the streaming video, captured by a webcam. The developed system works accurately with the generated synthetic data. Subsequently, the feature values are stored and machine learning algorithm have been used for classification. SVM algorithm has been explored here.

Also, In future the system will be implemented in hardware to make it portable for car system and pilot study on drivers will be carried out to validate the developed system.

REFERENCES

[1]Ashish Kumar, Rusha Patra, Department of Electronics and Communication Engineering Indian Institute of Information Technology Guwahati, India –"Driver Drowsiness Monitoring System", IEEE Pattern Recognition,July.2018.

[2]Gwak, J.S.; Shino, M.; Hirao, A. Early detection of driver drowsiness utilizing machine learning based on physiological signals, behavioral measures, and driving performance. In Proceedings of the IEEE International Conference on Intelligent Transportation Systems, Maui, Hawaii, HI, USA, 4–7 November 2018.

[3]Awais, M.; Badruddin, N.; Drieberg, M. A Hybrid approach to detect driver drowsiness utilizing physiological signals to improve system performance and wearability. Sensors 2017.

[4]A. Sengupta, A. Dasgupta, A. Chaudhuri, A. George, A. Routray, R. Guha; "A Multimodal System for Assessing Alertness Levels Due to Cognitive Loading", IEEE Trans. on Neural Systems and Rehabilitation Engg., vol. 25 (7), pp 1037-1046, 2017.

[5] K. T. Chui, K. F. Tsang, H. R. Chi, B. W. K. Ling, and C. K. Wu, "An accurate ECG based transportation safety drowsiness detection scheme," IEEE Transactions on Industrial Informatics, vol. 12, no. 4, pp. 1438- 1452, Aug. 2016.

[6] Zhang, H.; Wua, C.; Yan, X.; Qiu, T.Z. The effect of fatigue driving on car following behavior. Transp. Res. Part F 2016.

[7] W. L. Ou, M. H. Shih, C. W. Chang, X. H. Yu, C. P. Fan, "Intelligent Video-Based Drowsy Driver Detection System under Various Illuminations and Embedded Software Implementation", 2015 international Conf. on Consumer Electronics - Taiwan, 2015.

[8] M. Karchani, A. Mazloumi, G. N. Saraji, A. Nahvi, K. S. Haghighi, B. M. Abadi, A. R. Foroshani, A. Niknezhad, "The Steps of Proposed Drowsiness Detection System Design based on Image Processing in Simulator Driving", International Research Journal of Applied and Basic Sciences, vol. 9(6), pp 878-887, 2015.

[9] R.Ahmad, and J.N.Borole, "Drowsy Driver Identification Using Eye Blink Detection," IJISET - International Journal of Computer Science and Information Technologies, vol. 6, no. 1, pp. 270-274, Jan. 2015.

[10] Loon, R.J.;Brouwer,R.F.T.;Martens, M.H. Drowsy drivers' under-performance in lateral control: How much is too much? Using an integrated measure of lateral control to quantify safe lateral driving. Accid. Anal. Prev. 2015.

[11] V.Kazemi and J. Sullivan; "One millisecond face alignment with an ensemble of regression trees", IEEE Conf. on Computer Vision and Pattern Recognition, 23-28 June, 2014, Columbus, OH, USA.

[12] Correa, A.G.; Orosco, L.; Laciar, E. Automatic detection of drowsiness in EEG records based on multimodal analysis. Med Eng. Phys. 2014.

[13] A. Abas, J. Mellor, and X. Chen, "Non-intrusive drowsiness detection by employing Support Vector Machine," 2014 20th International Conference on Automation and Computing (ICAC), Bedfordshire, UK, 2014.

[14] B. Alshaqaqi, A. S. Baquhaizel, M. E. A. Ouis, M. Bouumehed, A. Ouamri, M. Keche, "Driver Drowsiness Detection System", IEEE International Workshop on Systems, Signal Processing and their Applications, 2013.

[15] Fu, C.L.; Li, W.K.; Chun, H.C.; Tung, P.S.; Chin, T.L. Generalized EEG-based drowsiness prediction system by using a self-organizing neural fuzzy system. IEEE Transection Circuits Syst. 2012.

[16] Rhodes, N.; Pivik, K. Age and gender differences in risky driving: The roles of positive affect and risk perception. J. Accid. Anal. Prev. 2011.

PUBLICATIONS

JOURNAL : UGC Care

CONFERENCE

- International Conference on"Driver Drowsiness Monitoring System Using Visual Behaviour and Machine Learning"(ICICCI-21-0160)
- Paper ID:ICICCI-21-0160

ALL FOUR STUDENTS' ONE PAGE PROFILE

I am Bijja Ragasree, pursuing my Bachelor of Technology in the stream of Computer Science and Engineering from St.Martin's Engineering College.I have done my Board of Intermediate from Narayana Junior College and SSC from Bhashyam High School.My technical skills include C,C++,Java,Python(Programming Language) and basic understanding in Machine Learning.My participation in technical workshops include Two-Day National Level SEMINAR On "Recent Trends in Cloud Computing,Fog and Edge Computing" 18th June to 19th June 2021, participated in Machine Learning Workshop,attended the Ethical Hacking Workshop conducted on 31st January and 1st February 2020.I participated in Poster Presentation event and Technical Treasure Hunt event during SYMPO AAGNYA 2020-A Two Day National Level Technical Symposium,MHRD'S Innovation Cell,Institution's Innovation Council held at St.Martin's Engineering College on 30th and 31st January 2020.Five days online workshop on women in Cyber Security and Privacy in 2020.I also completed Internship on Machine Learning held at Osmania University, CSE Campus,Hyderabad.I have also completed various certification courses like "Applied Machine Learning in Python", "Convolutional Neural Networks", "Neural Networks and Deep Learning" from Coursera; AI from A-Z from Udemy; "Java Script", "Python Core", "Machine Learning" from Sololearn; "Artificial Intelligence by crashcourse", "Cyber Security by packethacks" from cursa.I have also got offer from "Altruista heath".

I am Gudipelly Sneha, pursuing my Bachelor of Technology in the stream of Computer Science and Engineering from St.Martin's Engineering College.I have done my Board of Intermediate from Vijaya Ratna Junior College and SSC from Sri Chaitanya School.My technical skills include C,C++,Java,Python(Programming Language) and basic understanding in Machine Learning. My participation in technical workshops include Two-Day National Level SEMINAR On "Recent Trends in Cloud Computing,Fog and Edge Computing" 18th June to 19th June 2021, participated in Machine Learning Workshop,attended the Ethical Hacking Workshop conducted on 31st January and 1st February 2020.I participated in Poster Presentation event and Technical Treasure Hunt event during SYMPO AAGNYA 2020-A Two Day National Level Technical Symposium,MHRD'S Innovation Cell,Institution's Innovation Council held at St.Martin's Engineering College on 30th and 31st January 2020.Five days online workshop on women in Cyber Security and Privacy in 2020.I also completed Internship on Machine Learning held at Osmania University, CSE Campus,Hyderabad.I have also completed various certification courses like "Applied Machine Learning in Python", "AI for Everyone", "Neural Networks and Deep Learning" from Coursera; AI from A-Z from Udemy; "Java Script", "Python Core", "Machine Learning" from Sololearn; "Artificial Intelligence by crashcourse", "Cyber Security by packethacks" from cursa. have also got offer from "Altruista heath".

I am Jigatapu Pravalika, pursuing my Bachelor of Technology in the stream of Computer Science and Engineering from St.Martin's Engineering College.I have done my Board of Intermediate from Sri Chaitanya Junior College and SSC from Indo English High School.My technical skills include C,C++,Java,Python(Programming Language) and basic understanding in Machine Learning. My participation in technical workshops include Two-Day National Level SEMINAR On "Recent Trends in Cloud Computing,Fog and Edge Computing" 18th June to 19th June 2021, participated in Machine Learning Workshop,attended the Ethical Hacking Workshop conducted on 31st January and 1st February 2020.I participated in Poster Presentation event and Technical Treasure Hunt event during SYMPO AAGNYA 2020-A Two Day National Level Technical Symposium,MHRD'S Innovation Cell,Institution's Innovation Council held at St.Martin's Engineering College on 30th and 31st January 2020.Five days online workshop on women in Cyber Security and Privacy in 2020.I also completed Internship on Machine Learning held at Osmania University, CSE Campus,Hyderabad.I have also completed various certification courses like "Applied Machine Learning in Python", "AI for Everyone", "Neural Networks and Deep Learning" from Coursera; AI from A-Z from Udemy; "Java Script", "Python Core", "Machine Learning" from Sololearn; "Artificial Intelligence by crashcourse", "Cyber Security by packethacks" from cursa. I am NCC Cadet.I did two NCC(National Cadet Corps) camps in the year 2020 and 2021 and received 'B' Certificate. I have also offer from "Altruista heath".

I am Neemkar Rithika, pursuing my Bachelor of Technology in the stream of Computer Science and Engineering from St.Martin's Engineering College.I have done my Board of Intermediate from Sri Gayathri College and SSC from Holy Cross High School.My technical skills include C,C++,Java,Python(Programming Language) and basic understanding in Machine Learning. My participation in technical workshops include Two-Day National Level SEMINAR On "Recent Trends in Cloud Computing,Fog and Edge Computing" 18th June to 19th June 2021, participated in Machine Learning Workshop,attended the Ethical Hacking Workshop conducted on 31st January and 1st February 2020.I participated in Poster Presentation event and Technical Treasure Hunt event during SYMPO AAGNYA 2020-A Two Day National Level Technical Symposium,MHRD'S Innovation Cell,Institution's Innovation Council held at St.Martin's Engineering College on 30th and 31st January 2020.Five days online workshop on women in Cyber Security and Privacy in 2020.I also completed Internship on Machine Learning held at Osmania University, CSE Campus,Hyderabad.I have also completed various certification courses like "Applied Machine Learning in Python", "AI for Everyone", "Neural Networks and Deep Learning" from Coursera; "Java Script", "Python Core", "Data Science","Machine Learning" from Sololearn; "Artificial Intelligence by crashcourse", "Cyber Security by packethacks" from cursa.I have also got offer from "Infosys".

APPENDICES

DrowsinessDetector.py

```
from tkinter import *
import tkinter
from scipy.spatial import distance as dist
from imutils import face_utils
import numpy as np
import imutils
import dlib
import cv2
main = tkinter.Tk()
main.title("Driver Drowsiness Monitoring")
main.geometry("500x400")
def EAR(drivereye):
    point1 = dist.euclidean(drivereye[1], drivereye[5])
    point2 = dist.euclidean(drivereye[2], drivereye[4])
    # compute the euclidean distance between the horizontal
    distance = dist.euclidean(drivereye[0], drivereye[3])
    # compute the eye aspect ratio
    ear_aspect_ratio = (point1 + point2) / (2.0 * distance)
    return ear_aspect_ratio
def MOR(drivermouth):
    # compute the euclidean distances between the horizontal
```

```
    point   = dist.euclidean(drivermouth[0], drivermouth[6])
    # compute the euclidean distances between the vertical
    point1  = dist.euclidean(drivermouth[2], drivermouth[10])
    point2  = dist.euclidean(drivermouth[4], drivermouth[8])
    # taking average
    Ypoint   = (point1+point2)/2.0
    # compute mouth aspect ratio
    mouth_aspect_ratio = Ypoint/point
    return mouth_aspect_ratio
def startMonitoring():
    pathlabel.config(text="          Webcam Connected Successfully")
    webcamera = cv2.VideoCapture(0)
    svm_predictor_path = 'SVMclassifier.dat'
    EYE_AR_THRESH = 0.25
    EYE_AR_CONSEC_FRAMES = 10
    MOU_AR_THRESH = 0.75
    COUNTER = 0
    yawnStatus = False
    yawns = 0
    svm_detector = dlib.get_frontal_face_detector()
    svm_predictor = dlib.shape_predictor(svm_predictor_path)
    (lStart, lEnd) = face_utils.FACIAL_LANDMARKS_IDXS["left_eye"]
    (rStart, rEnd) = face_utils.FACIAL_LANDMARKS_IDXS["right_eye"]
    (mStart, mEnd) = face_utils.FACIAL_LANDMARKS_IDXS["mouth"]
```

```
while True:
    ret, frame = webcamera.read()
    frame = imutils.resize(frame, width=640)
    gray = cv2.cvtColor(frame, cv2.COLOR_BGR2GRAY)
    prev_yawn_status = yawnStatus
    rects = svm_detector(gray, 0)
    for rect in rects:
        shape = svm_predictor(gray, rect)
        shape = face_utils.shape_to_np(shape)
        leftEye = shape[lStart:lEnd]
        rightEye = shape[rStart:rEnd]
        mouth = shape[mStart:mEnd]
        leftEAR = EAR(leftEye)
        rightEAR = EAR(rightEye)
        mouEAR = MOR(mouth)
        ear = (leftEAR + rightEAR) / 2.0
        leftEyeHull = cv2.convexHull(leftEye)
        rightEyeHull = cv2.convexHull(rightEye)
        mouthHull = cv2.convexHull(mouth)
        cv2.drawContours(frame, [leftEyeHull], -1, (0, 255, 255), 1)
        cv2.drawContours(frame, [rightEyeHull], -1, (0, 255, 255), 1)
        cv2.drawContours(frame, [mouthHull], -1, (0, 255, 0), 1)
        if ear < EYE_AR_THRESH:
            COUNTER += 1
```

```
            cv2.putText(frame, "Eyes Closed ", (10,
30),cv2.FONT_HERSHEY_SIMPLEX, 0.7, (0, 0, 255), 2)

            if COUNTER >= EYE_AR_CONSEC_FRAMES:

                cv2.putText(frame, "DROWSINESS ALERT!", (10,
50),cv2.FONT_HERSHEY_SIMPLEX, 0.7, (0, 0, 255), 2)

        else:

            COUNTER = 0

            cv2.putText(frame, "Eyes Open ", (10,
30),cv2.FONT_HERSHEY_SIMPLEX, 0.7, (0, 255, 0), 2)

        cv2.putText(frame, "EAR: {:.2f}".format(ear), (480,
30),cv2.FONT_HERSHEY_SIMPLEX, 0.7, (0, 0, 255), 2)

        if mouEAR > MOU_AR_THRESH:

            cv2.putText(frame, "Yawning, DROWSINESS ALERT! ", (10,
70),cv2.FONT_HERSHEY_SIMPLEX, 0.7, (0, 0, 255), 2)

            yawnStatus = True

            output_text = "Yawn Count: " + str(yawns + 1)

            cv2.putText(frame, output_text,
(10,100),cv2.FONT_HERSHEY_SIMPLEX, 0.7,(255,0,0),2)

        else:

            yawnStatus = False

        if prev_yawn_status == True and yawnStatus == False:

            yawns+=1

        cv2.putText(frame, "MAR: {:.2f}".format(mouEAR), (480,
60),cv2.FONT_HERSHEY_SIMPLEX, 0.7, (0, 0, 255), 2)

        cv2.putText(frame,"Visual Behaviour & Machine Learning Drowsiness
Detection @
Drowsiness",(370,470),cv2.FONT_HERSHEY_COMPLEX,0.6,(153,51,102),1)
```

```
        cv2.imshow("Frame", frame)
        key = cv2.waitKey(1) & 0xFF
        if key == ord("q"):
            break
    cv2.destroyAllWindows()
    webcamera.release()
font = ('times', 16, 'bold')
title = Label(main, text='Driver Drowsiness Monitoring System using Visual\n Behaviour and Machine Learning',anchor=W, justify=LEFT)
title.config(bg='black', fg='white')
title.config(font=font)
title.config(height=3, width=120)
title.place(x=0,y=5)
font1 = ('times', 14, 'bold')
upload = Button(main, text="Start Behaviour Monitoring Using Webcam", command=startMonitoring)
upload.place(x=50,y=200)
upload.config(font=font1)
pathlabel = Label(main)
pathlabel.config(bg='DarkOrange1', fg='white')
pathlabel.config(font=font1)
pathlabel.place(x=50,y=250)
main.config(bg='chocolate1')
main.mainloop()
```

A

PROJECT REPORT

On

SECURING DATA WITH BLOCKCHAIN AND ARTIFICIAL INTELLIGENCE

Submitted by

1)Ms. Alluri Shivani (17K81A0502)
2)Ms. Mandula Meghana (17K81A0534)
3)Ms. Matta Jahnavi Reddy (17K81A0535)
4)Ms. Talakokkula Ashwitha (17K81A0553)

in the partial fulfillment for the award of

the degree of

BACHELOR OF TECHNOLOGY

IN

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Under the Guidance of

Mr. J. Sudhakar

Associate Professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ST.MARTIN'S ENGINEERING COLLEGE
An Autonomous Institute

Dhulapally, Secunderabad – 500 100

JUNE 2021

BONAFIDE CERTIFICATE

This is to certify that the project entitled Securing Data with Blockchain and Artificial Intelligence, is being submitted by 1. **Ms. Alluri Shivani (17K81A0502), 2. Ms. Mandula Meghana (17K81A0534), 3. Ms. Matta Jahnavi Reddy (17K81A0535), 4. Ms. Talakokkula Ashwitha (17K81A0553)** in partial fulfillment of the requirement for the award of the degree of **BACHELOR OF TECHNOLOGY IN COMPUTER SCIENCE AND ENGINEERING** is recorded of bonafide work carried out by them. The result embodied in this report have been verified and found satisfactory.

Mr. J. Sudhakar
Department of CSE

Head of the Department
Dr.M.NARAYANAN
Department of CSE

Internal Examiner

External Examiner

Place: Hyderabad

Date:

DECLARATION

We, the student of **Bachelor of Technology** in Department of Computer Science and Engineering', session: <2017 – 2021>, St. Martin's Engineering College, Dhulapally, Kompally, Secunderabad, hereby declare that work presented in this Project Work entitled **Securing Data with Blockchain and Artificial Intelligence** is the outcome of our own bonafide work and is correct to the best of our knowledge and this work has been undertaken taking care of Engineering Ethics. This result embodied in this project report has not been submitted in any university for award of any degree.

Ms. Alluri Shivani	17K81A0502
Ms. Mandula Meghana	17K81A0534
Ms. Matta Jahnavi Reddy	17K81A0535
Ms. Talakokkula Ashwitha	17K81A0553

ABSTRACT

Data is the input for various artificial intelligence (AI) algorithms to mine valuable features, yet data in Internet is scattered everywhere and controlled by different stakeholders who cannot believe in each other, and usage of the data in complex cyberspace is difficult to authorize or to validate. As a result, it is very difficult to enable data sharing in cyberspace for the real big data, as well as a real powerful AI. In this paper, we propose an architecture that can enable secure data storing, computing, and sharing in the large-scale Internet environment, aiming at a more secure cyberspace with real big data and thus enhanced AI with plenty of data source, by integrating three key components: 1) blockchain-based data sharing with ownership guarantee, which enables trusted data sharing in the large-scale environment to form real big data; 2) AI-based secure computing platform to produce more intelligent security rules, which helps to construct a more trusted cyberspace; 3) trusted value-exchange mechanism for purchasing security service, providing a way for participants to gain economic rewards when giving out their data or service, which promotes the data sharing and thus achieves better performance of AI. Moreover, we discuss the typical use scenario of architecture as well as its potentially alternative way to deploy, as well as analyze its effectiveness from the aspect of network security and economic revenue.

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible and whose encouragements and guidance have crowded effects with success.

We extended our deep sense of gratitude to Principal**, Dr. P. SANTOSH KUMAR PATRA**, St. Martin's Engineering College, Dhulapally, for permitting us to undertake this project.

We are also thankful to **Dr. M.NARAYANAN**, Head of the Department, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

We would like to thank **Dr. T. POONGOTHAI,** Department of Computer Science and Engineering (AI & ML) for her encouragement and insightful comments and as well as our project coordinators **Dr. B.RAJALINGAM**, Associate Professor and **Mr. J.SUDHAKAR**,Associate Professor, in Department of Computer Science and Engineering for their valuable support.

We would like to express our sincere gratitude and indebtedness to our project supervisor Mr. J. Sudhakar, Associate Professor, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

Finally, we express thanks to all those who have helped us successfully to completing this project. Furthermore, we would like to thank our family and friends for their moral support and encouragement.

We express thanks to all those who have helped us in successfully completing the project.

Ms. Alluri Shivani 17K81A0502

Ms. Mandula Meghana 17K81A0534

Ms. Matta Jahnavi Reddy 17K81A0535

Ms. Talakokkula Ashwitha 17K81A0553

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF OUTPUT SCREENS

1.INTRODUCTION

1. INTRODUCTION

1.1 PROJECT OVERVIEW

With the development of information technologies, the trend of integrating cyber, physical and social (CPS) systems to a highly unified information society, rather than just a digital Internet, is becoming increasing obvious. In such an information society, data is the asset of its owner, and its usage should be under the full control of its owner, although this is not the common case. Given data is undoubtedly the oil of the information society, almost every big company want to collect data as much as possible, for their future competitiveness. An increasing amount of personal data, including location information, web-searching behavior, user calls, user preference, is being silently collected by the built-in sensors inside the products from those big companies, which brings in huge risk on privacy leakage of data owners. Moreover, the usage of those data is out of control of their owners, since currently there is not a reliable way to record how the data is used and by who, and thus has little methods to trace or punish the violators who abuse those data. That is, lack of ability to effectively manage data makes it very difficult for an individual to control the potential risks associated with the collected data. For example, once the data has been collected by a third party (e.g., a big company), the lack of access to this data hinders an individual to understand or manage the risks related to the collected data from him. Meanwhile, the lack of immutable recording for the usage of data increases the risks to abuse them. If there is an efficient and trusted way to collect and merge the data scattered across the whole CPS to form real big data, the performance of artificial intelligence (AI) will be significantly improved since AI can handle massive amount of data including huge information at the same time, which would bring in great benefits (e.g., achieving enhanced security for data) and even makes AI gaining the ability to exceed human capabilities in more areas. According to the research in, if given large amount of data in an order of magnitude more scale, even the simplest AI algorithm currently (e.g., perceptrons from the 1950s) can achieve fanciest performance to beat many state-of-the-art technologies today. The key lies in how to make data sharing trusted and secured. Fortunately, the blockchain technologies may be the promising way to achieve this goal, via consensus mechanisms throughout the network to guarantee data sharing in a tamper-proof way embedded with economic incentives. Thus, AI can be further empowered by blockchain-protected data sharing. As a result, enhanced AI can provide better performance and security for data. In this project, we aim at securing data by combining blockchain and AI together to significantly improve the security of data sharing, and then the security of the whole network, even the whole CPS. The trust-less relationship between different data stakeholders significantly thwarts the data sharing in the whole Internet, thus the data used for AI training or analyzing is limited in amount as well as partial in

variety. Fortunately, the rise of Blockchain technologies bring in a hopeful, efficient and effective way to enable trust data sharing in trustless environment, which can help AI make more accurate decisions due to the real big data collected from more places in the Internet. Our project architecture leverages the emerging blockchain technologies to prevent the abuse of data, and to enable trusted data sharing in trust-less or even untrusted environment. For instance, it can enable cooperations between different edge computing paradigms to work together to improve the whole system performance of edge networks. The reason why blockchain can enable trusted mechanisms is that it can provide a transparent, tamper-proof metadata infrastructure to seriously recode all the usage of data. We introduced blockchain-based data sharing mechanisms with ownership guarantee, where any data ready for sharing should be registered into a blockchain, named Data Recording Blockchain (DRB), to announce its availability for sharing. Each access behavior on data by other parties (not the data owner) should also be validated and recorded in this chain. In addition, the authenticity and integrity of data can only be validated by DRB as well. Furthermore, data is the fuel of AI, and it can greatly help to improve the performance of AI algorithms if data can be efficiently networked and properly fused. Enabling data sharing across multiple service providers can be a way to maximize the utilization of scattered data in separate entities with potential conflicts of interest, which can enable a more powerful AI. Given enough data and blockchain based smart contract on secure data sharing, it's not surprised that AI can become one of the most powerful technologies and tools to improve cybersecurity.

1.2 PROJECT OBJECTIVES

- The main objective is to ensure the data security so it overcomes all the attacks.
- To achieve forward and backward security of the data with revocation scheme.
- To enable access control that is cryptographically enhanced.
- To ensure the protection of cloud database.
- To perform secure and efficient query processing.

1.3 ORGANIZATION OF CHAPTERS

The thesis is organized in the following chapters:

Chapter 1: Introduction

This assessment seeks to secure data with Blockchain and Artificial Intelligence. If there is an efficient and trusted way to collect and merge the data scattered across the whole CPS to form real

big data, the performance of artificial intelligence (AI) will be significantly improved since AI can handle massive amount of data including huge information at the same time, which would bring in great benefits (e.g., achieving enhanced security for data) and even makes AI gaining the ability to exceed human capabilities in more areas.

Chapter 2: Literature Survey

In this section, we analyze the existing result which were done on Blockchain and Artificial Intelligence individually. We progress in the view of developing a project by merging both the technologies using Healthcare Domain as a challenge.

Chapter 3: Software and Hardware Requirements

In this chapter, we specified the Software and Hardware components required to develop our project**.** The Software and Hardware requirements specify the intended purpose, requirements, and nature of software/application/project to be developed. By selecting the dataset that most resembles the usage requirements in our environment, we can use the recommended topology and associated hardware requirements for our topology as a starting point when we plan for hardware of our project.. Requirements may vary based on utilization and observing performance of pilot projects is recommended prior to scale out.

Chapter 4: Software Development Analysis

In this project, we discussed about development and implementation of the project in detail. Considering the security of one's data, we developed in our roles as front-end, back-end and database administrator by collecting relevant data and testing it in required cases.

Chapter 5: Project System Design

This chapter reports on the analysis and design of our proposed application. This chapter describes the system design architecture and database design and is organized in a sequence included with data

gathering and system design. Stakeholders will discuss factors such as risk levels, team composition, applicable technologies, time, budget, project limitations, method and architectural design.

Chapter 6: Project Coding

This chapter is a system implementation of the project. We will discuss briefly the implementation of our project. This section describes some of the coding templates, outline of various files, class with functionalities, the various methods of input and output parameters.

Chapter 7: Project Testing

In this chapter, we will discuss briefly the testing of each functionality of our proposed application in the project. We performed various testings like whitebox, blackbox, unit testing, integration testing and many more to check the accuracy and performance of our output. They notify developers of defects in the code. If developers confirm the flaws are valid, they improve the program, and the testers repeat the process until the software is free of bugs and behaves according to requirements.

Chapter 8: Output screens

In this chapter, we captured the screenshots of our project output. We considered few sample inputs and obtained desired outputs for our data with related database.

Chapter 9: Experimental Results

In this chapter, we conclude the performance analysis of our proposed project by comparing it with the existing project. In this chapter, we discuss briefly the conclusion of each chapter with the progress of our proposed system.

2.LITERATURE SURVEY

2.LITERATURE SURVEY

2.1 SURVEY ON BACKGROUND

This section of the literature survey eventually reveals some facts based on thoughtful analysis of many authors work as follows.

1.S. Yu states that due to rapid pace of technology and the introduction of Internet of Things or IoT, there has been a swift increase in the number of intelligent devices being connected to the internet. These devices are capable of generating a large amount of data due to the fact that it is connected and is interacting with the internet. A large number of devices generate equally large amount of data which cannot be processed efficiently.

2.R. Wang elaborates on the foundation of network security construction which is the PKI or The Public Key Infrastructure. The researchers also commented on the reliability and the robust security offered by the blockchain platform.

3.C. Ehmke explains that the innovative paradigm of Blockchain has been popular and has seen extensive usage recently. The blockchain was utilized for financial applications and that is how it gained immense popularity and limelight. The blockchain was readily picked up by a plethora of researchers and implemented in various different fields, which has greatly helped in bringing increased security to numerous applications.

4.R. Wang introduces the video surveillance system as an irreplaceable tool that can be used to efficiently manage and survey big cities. When a video surveillance system is installed, it can easily transmit environment information remotely, this is highly useful as the person does not need to travel long distances and physically be present in the location for the management. Due to a large-scale increase in the monitoring standards with the inclusion of IoT and realtime monitoring, it is susceptible to attacks.

5.J. Lou states that there has been a lack of a key management feature in the Named Data Networking, which is utilized to name each and every object by the producer and also digitally sign it. There are some disadvantages of the conventional approach such as lack of trust between the sites as well as the high chances of failure observed in the centralized architecture if the main node fails.

6.S. Wang explains that there has been a very fast development of cryptocurrency in recent years,

which has led to detailed scrutiny of the paradigm. This has uncovered a lot of irregularities in the paradigm such as the Smart Contracts that have been the cause of “The DOA Attack” which has resulted in a huge loss.

7.Y. Xu introduces the concept of decentralized storage that is based on the blockchain framework. The blockchain is one of the most innovative concepts that can be used to design a highly secure decentralized framework.

8.M. Marchesi in his keynote speech details the rapid development and ongoing researches going on in the field of blockchain. This is due to the increased attention to this paradigm and increased demand in this sector.

9.A.Maksutov elaborates on the paradigm of Blockchain and its uses. The authors have proposed an innovative concept for the detection and identification of various money laundering schemes that use the blockchain framework for their nefarious activities.

10.F.Wessling states that the addition of blockchain to existing platforms is problematic as it is different from building the applications from scratch by incorporating the Blockchain into the application. The authors determine the amount of blockchain required for various different implementations, this is done by analyzing the attributes of blockchain such as anonymous, trustless and immutable, etc.

11.J.Wang explains that most of the applications based on crowd sensing gather a huge amount of data by using pervasive smartphone users to provide the data. But most of the time the users are not compensated enough for their contribution to the system.

12.S. Pandey introduces the benefits of utilizing the Blockchain paradigm for its security and decentralized architecture. The author states that there has been a jump in the number of researches is going on in this field and there has been increased interest in implementing the blockchain framework to make existing systems resilient and secure.

2.2CONCLUSIONS ON SURVEY

1.S.Yu- The authors propose an effective technique based on blockchain that can provide a low-cost alternative to create economic value for the IoT data generated. A major drawback of this methodology is that it has the potential to be misused by uploading large amounts of malicious data.

2.R.Wang- The authors amalgamated both the methodologies to strengthen the Public Key Infrastructure by a permissioned blockchain that converts the PKI into a privacy aware PKI. This is crucial as the implementation of a permissioned blockchain also improves the efficiency of the configuration and certificate application. The major drawback is that this technique has been a very specialized approach towards the blockchain paradigm.

1.C.Ehmke- Due to the fact that the blockchain paradigm requires a user of the blockchain to download the whole chain to gain an overview. To ameliorate this effect, the authors have implemented a scalable and lightweight blockchain protocol.

2.R.Wang- The authors developed a system for video surveillance based on permissioned blockchains and Convolutional Neural Networks for a seamless and secure system. A Major drawback in the system is that large scale testing of the System has not been performed and will be done in the upcoming researches.

3.J.Lou- The authors in this paper propose an efficient key management scheme based on blockchain for the Named Data Networking paradigm. The blockchain increases the trust between the sites as well as the decentralized architecture is highly useful in overcoming failure. The drawback of the proposed scheme is that it has not been evaluated extensively for its feasibility in reducing the NDN cache pollution.

4.S.Wang- The authors have presented a comprehensive and systematic review of the smart contracts in the blockchain paradigm. The authors have presented a six-layer architecture for smart contracts for increasing the security of the system. The authors have not implemented a formal verification which can provide confidence.

5.Y.Xu- The authors have proposed section blockchain protocol, which aims to eliminate the storage problem that is encountered in certain devices. The proposed methodology is highly resilient to failure due to the decentralized architecture, as well as, it has the ability to withstand heavy loads and optimization gracefully due to the implementation of the Blockchain paradigm.

6.M.Marchesi- The author indicates that this increased pressure on a nascent framework has led to an increase in security lapses that are evident in the various different incidents on the Ethereum platform and the cryptocurrency exchanges. That being said the speaker also highlighted the immense opportunities that can be utilized by using the blockchain paradigm such as the Blockchain tokens that can be used to implement a reward and penalty-based scheme for developers.

7.A.Maksutov- The proposed methodology has been used to deanonymizing the transactions and

tracking the coin join transactions, which allows the authors to evaluate user participation. All of this information is used to determine if the transactions are being fraudulent or used to launder money.

8.F.Wessling- The authors have outlined the various different processes that utilize various different elements of the blockchain technology that can be implemented based on the specific application and use case of the application.

9.J.Wang- The authors have presented an innovative framework for a privacy-preserving reward and penalty scheme that rewards the users for contributing to the large data sensing paradigm using the trustless and secure blockchain. The major drawback of this paper has been that the authors have not discussed the solutions for a possible collusion attack.

10.S.Pandey- To the effect, the authors have formulated an ingenious and practical simulation tool for planning, stability, and design of the systems and applications as well as networks in a blockchain environment.

3.SOFTWARE AND HARDWARE REQUIREMENTS

3.SOFTWARE AND HARDWARE REQUIREMENTS

The Software and Hardware requirements specify the intended purpose, requirements, and nature of software/application/project to be developed. These system requirements are the configuration that our system must have in order for a hardware or software application to run smoothly and effectively. These requirements at the system level describes the functions which the system as a whole should fulfil to satisfy the stakeholder needs and requirements.

3.1 SOFTWARE REQUIREMENTS

Operating System	Windows 8 & above
Language	Java
IDE	Net Beans S8.2
Data Base	MySQL

Table 3.1 Software Requirements

3.2 HARDWARE REQUIREMENTS

System	Pentium IV 2.4 GHz
Hard Disk	160 GB
Monitor	15 VGA colour
Mouse	Logitech
Keyboard	110 keys enhanced
RAM	2GB

Table 3.2 Hardware Requirements

4. SOFTWARE DEVELOPMENT ANALYSIS

4. SOFTWARE DEVELOPMENT ANALYSIS

4.1 OVERVIEW OF PROBLEM

An increasing amount of personal data, including location information, web-searching behavior, user calls, user preference, is being silently collected by the built-in sensors inside the products from those big companies, which brings in huge risk on privacy leakage of data owners. lack of ability to effectively manage data makes it very difficult for an individual to control the potential risks associated with the collected data. For example, once the data has been collected by a third party (e.g., a big company), the lack of access to this data hinders an individual to understand or manage the risks related to the collected data from him. Meanwhile, the lack of immutable recording for the usage of data increases the risks to abuse them.

4.2 DEFINE THE PROBLEM

The performance of artificial intelligence (AI) will be significantly improved since AI can handle massive amount of data including huge information at the same time, which would bring in great benefits (e.g., achieving enhanced security for data) and even makes AI gaining the ability to exceed human capabilities in more areas. Not all cyber data can be made publicly available such as Patient Health Data which contains patient disease details and contact information and if such data available publicly then there is no security for that patient data. Now a days all service providers such as online social networks or cloud storage will store some type of users data and they can sale that data to other organization for their own benefits and user has no control on his data as that data is saved on third party server.To overcome from above issue, we have described concept called Blockchain and AI technique to provide security to user’s data. In this technique 3 functions will work which describe below:

1: Blockchain

2: Artificial Intelligence

3: Cloud Computing

4.3 MODULES OVERVIEW

REGISTRATION

HEALTHCARE PROVIDER

- Load patient Records
- Key Generation
- Encrypt patient Records
- Block Creation
- Upload and Download Patient Records

CLOUD SERVICE PROVIDER

- View Patient Records
- Grant or Revoke Permission

4.4 DEFINE THE MODULES

REGISTRATION

It is a process of enrolling or being enrolled into the cloud. To utilize the cloud documents, every healthcare provider should enroll. During this process your basic information like email, contacts etc., are collected and stored in the Cloud. The cloud id for a particular user will get automatically generated during the registration.

HEALTHCARE PROVIDER

- Load patient Records
- Key Generation
- Encrypt patient Records
- Block Creation
- Upload and Download Patient Records

CLOUD SERVICE PROVIDER

- The cloud service provider maintain all the patient records and also they can provide a permission to the user to access the data.

- The Cloud Service Provider can view all the uploaded and downloaded documents in the Cloud. The CSP receives the document request from the Data User, verifies the authentication before granting permission. Then the CSP executes the query and returns the encrypted document according to the search token. And also returns an additional proof with the document, to verify the search result.

4.5 MODULE FUNCTIONALITY

REGISTRATION

Cloud ID

Every user should create a Cloud ID and use it to identify something with near certainty that the identifier does not duplicate one that has already been, or will be, created to identify something else. Information labelled with Cloud ID by independent parties can therefore be later combined into a single database, or transmitted on the same channel, without needing to resolve conflicts between identifiers.

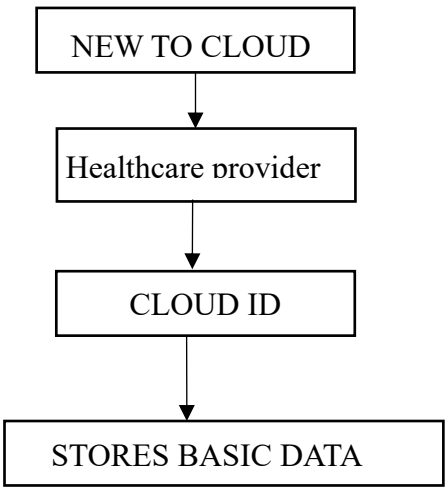


Fig 4.5.1 Cloud ID

HEALTHCARE PROVIDER

DATA SELECTION AND LOADING

In this process, the health provider choose patient healthcare records for uploading and maintaining the dataset in the cloud.

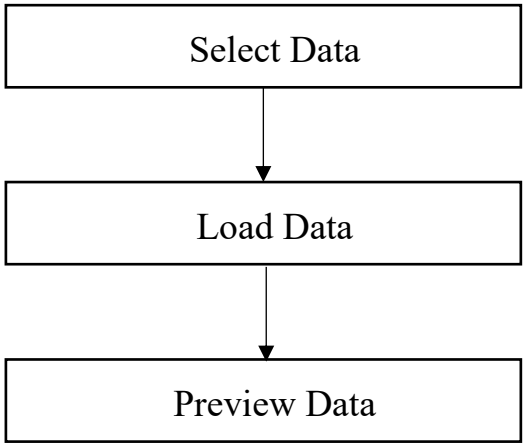


Fig 4.5.2 Data Selection and Loading

KEY GENERATION

The secret key is generated using cryptographic algorithm. This key is used for encrypting the dataset.

ENCRYPT PATIENT RECORDS

The data is encrypted for secure maintenance. So that the unauthorized person cannot be able to access the data that are presented in the cloud.

BLOCK CREATION

- Each block contain patient record and it's timestamp.
- A blockchain, originally block is a growing list of records called blocks.

UPLOAD AND DOWNLOAD PATIENT RECORDS

After creating the block, the healthcare provider will upload the records into the cloud. Suppose, if they want to retrieve an record from cloud, first the healthcare provider search the record. Based

on the search it will show the results. After getting an approval and key from the cloud service provider the healthcare provider can download the data.

CLOUD SERVICE PROVIDER

PUBLIC VERIFICATION KEY

Public verification key is a security measure designed to make sure that your document outsourced in cloud doesn't get hacked. By verifying public key, the Data Owner and the Data User adding another layer of protection to the documents or files in the cloud by confirming each other's identities.

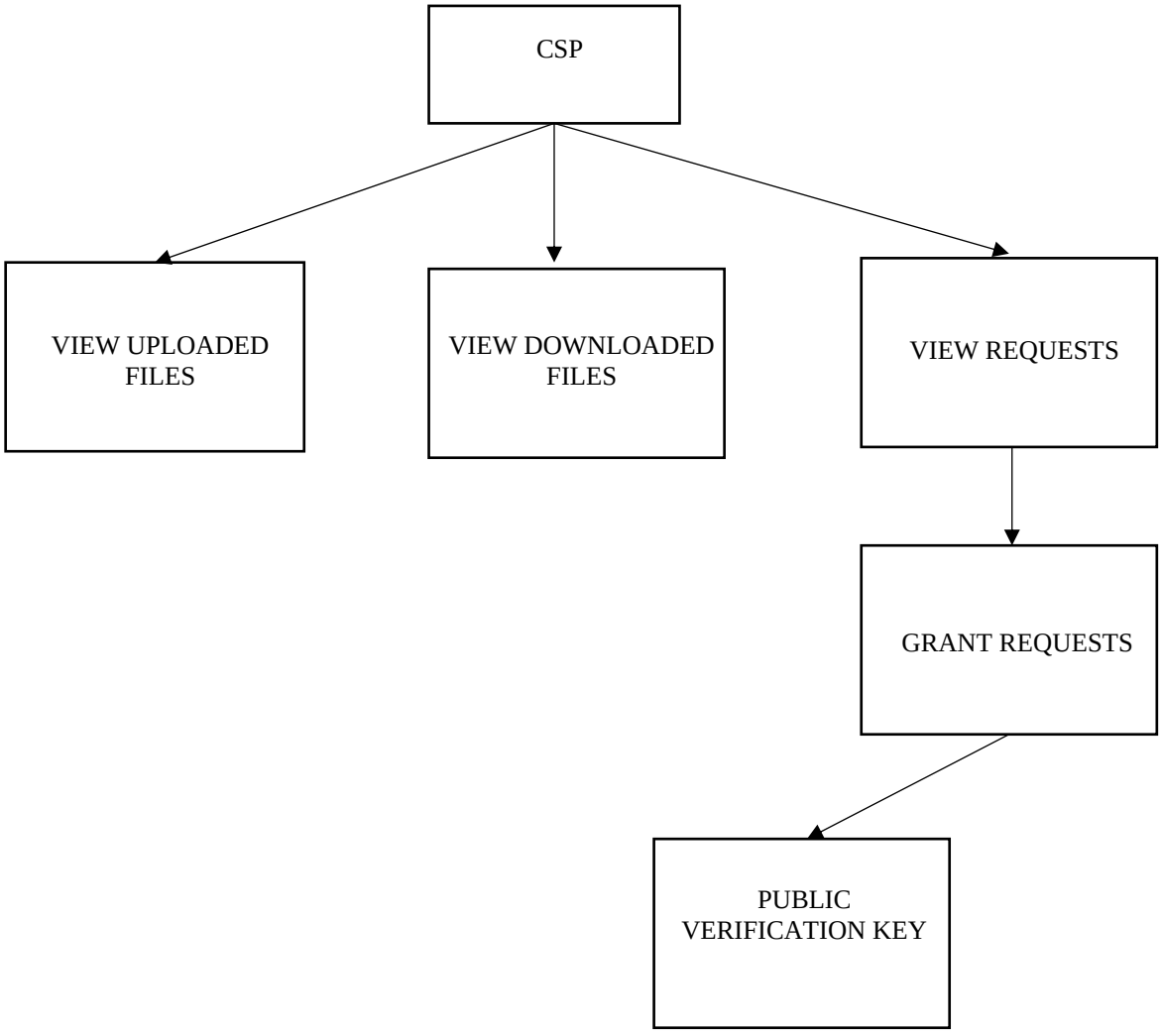


Fig 4.5.3 Public Verification Key

5. PROJECT SYSTEM DESIGN

5. PROJECT SYSTEM DESIGN

Project Design is the strategic organization of ideas material and processes to set our project up for success once we launch. It includes dataflow diagram ,object models, architecture diagrams, UML diagrams, a detailed design and functionality of our project.

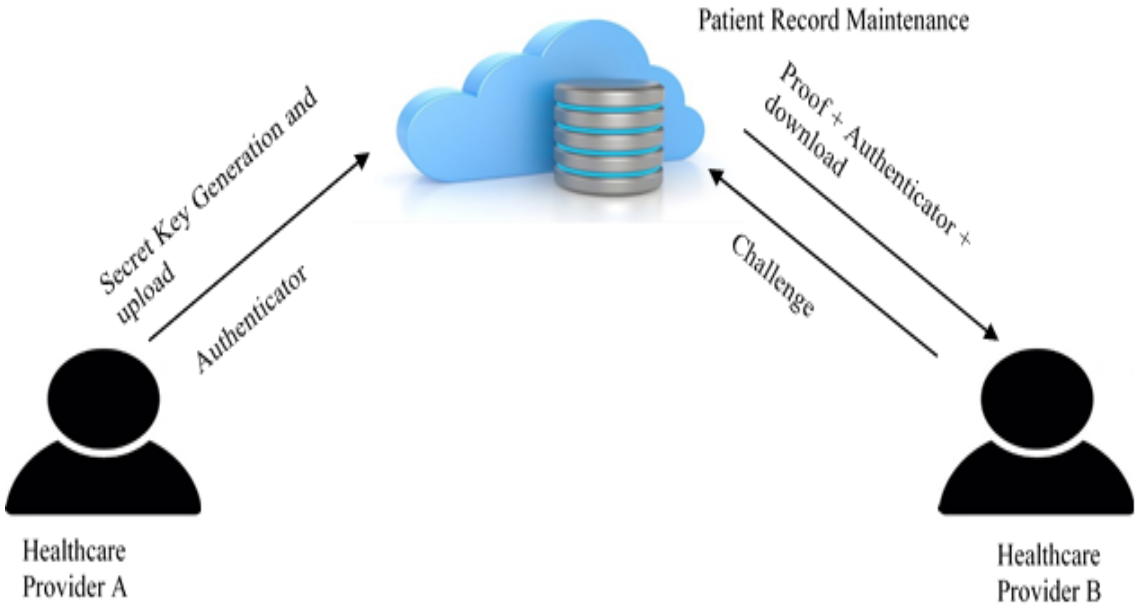


Fig 5 Architecture diagram

5.1 DFDS (DATA FLOW DIAGRAMS)

FLOWCHART

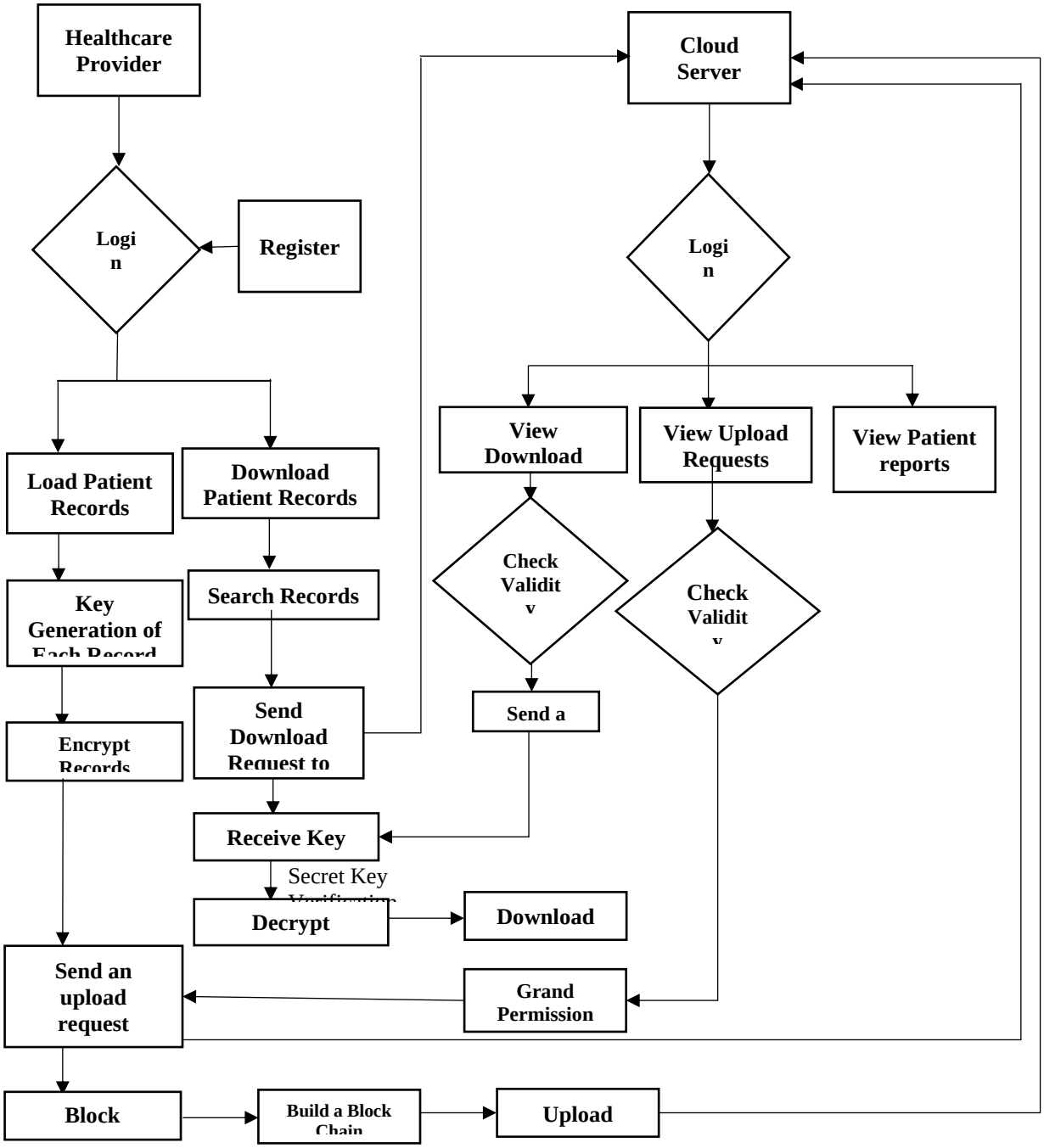


Fig 5.1.1 Flowchart

HEALTHCARE SYSTEM

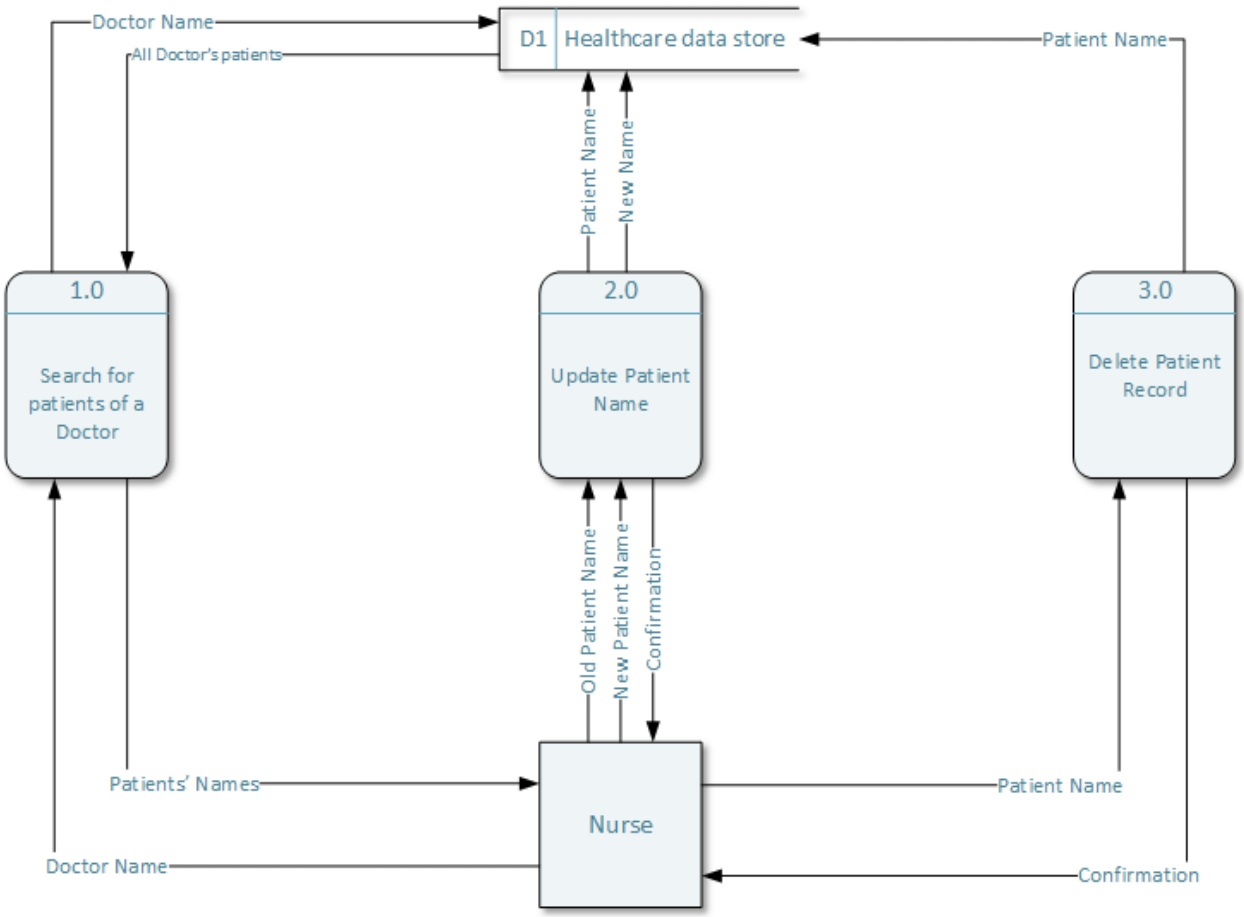


Fig 5.1.2 Healthcare System

5.2 E-R DIAGRAMS

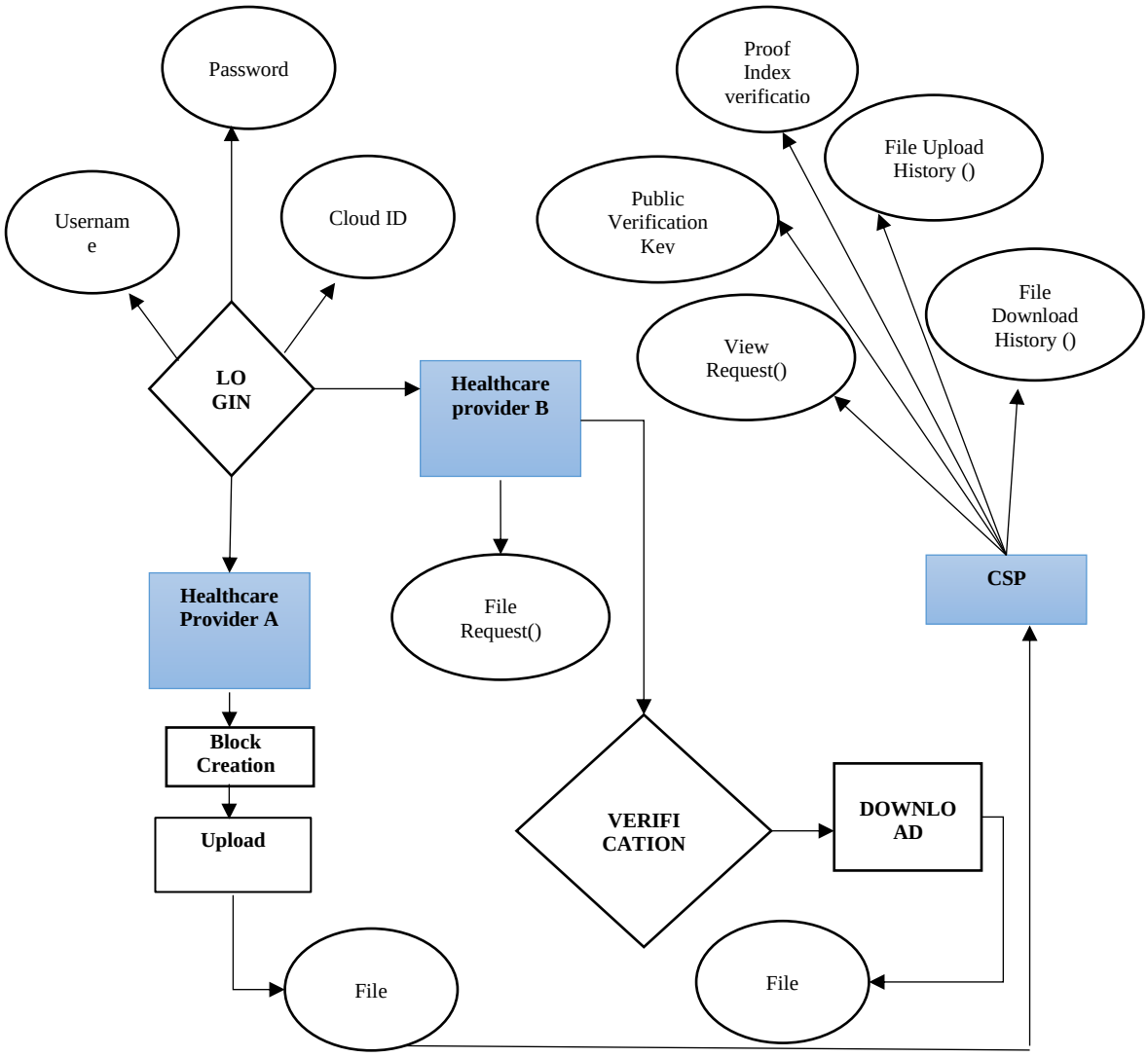


Fig 5.2 E-R Diagram

5.3 UML DIAGRAMS

USE CASE DIAGRAM

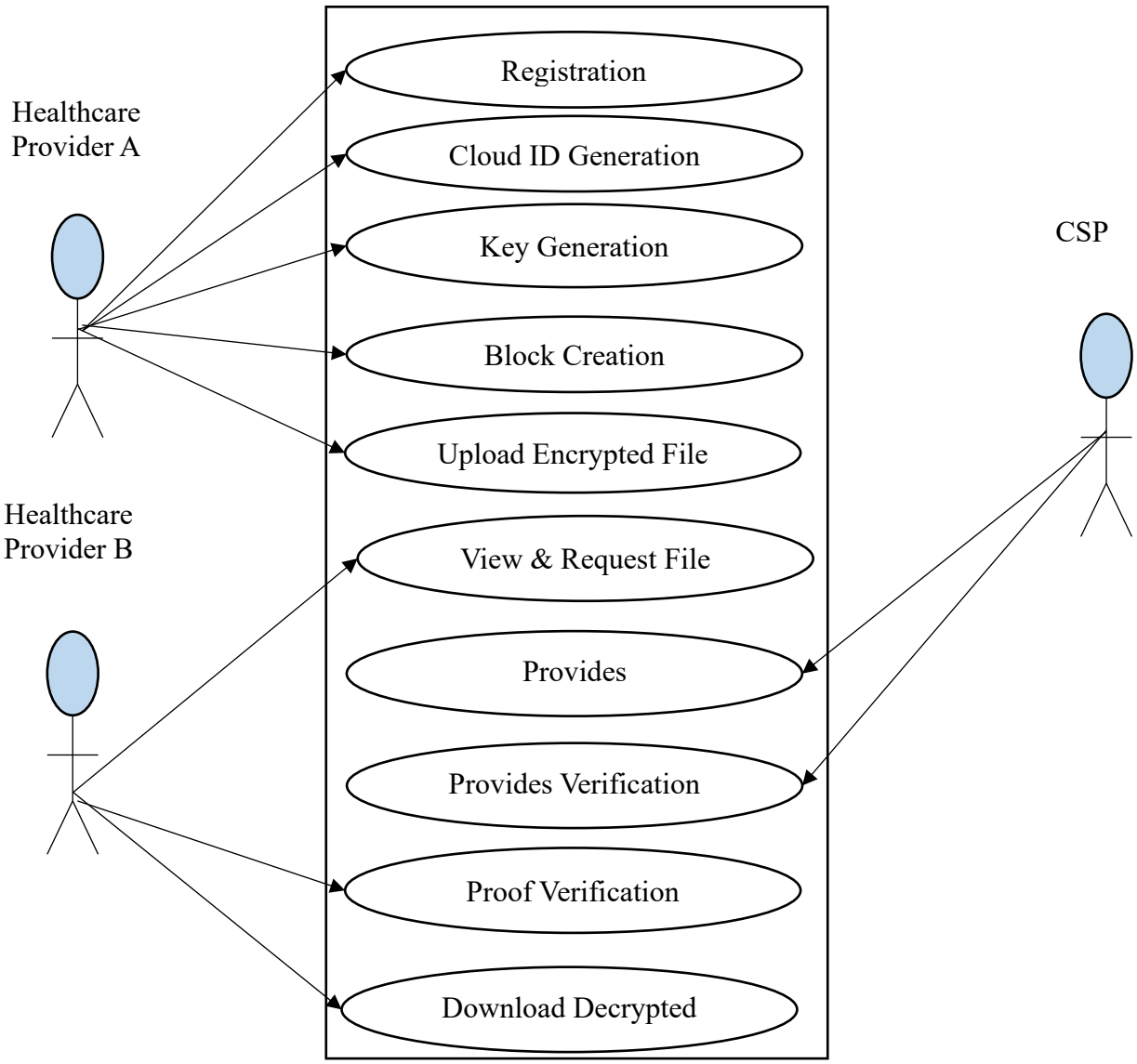


Fig 5.3.1 Use Case Diagram

CLASS DIAGRAM

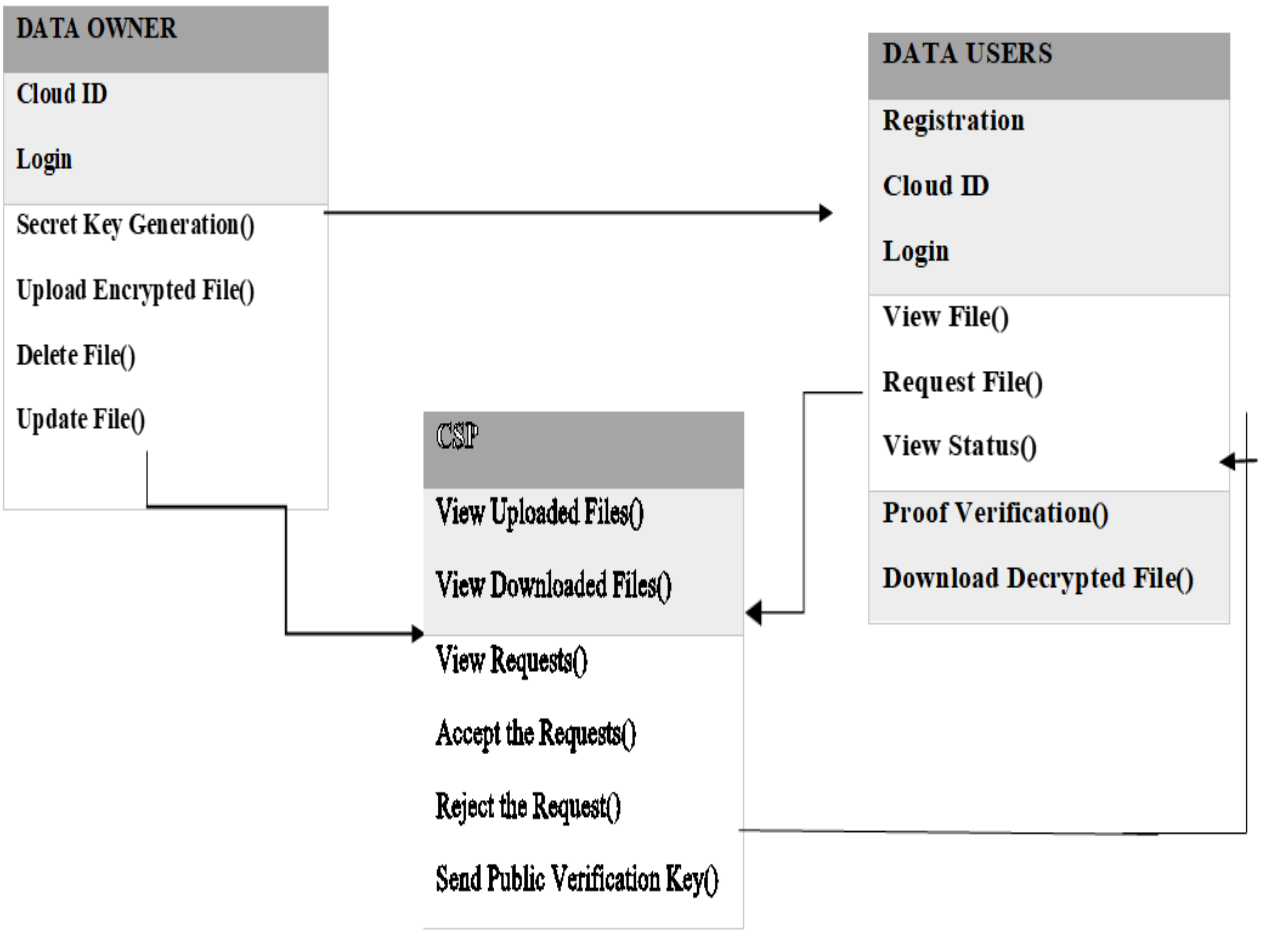


Fig 5.3.2 Class Diagram

SEQUENCE DIAGRAM

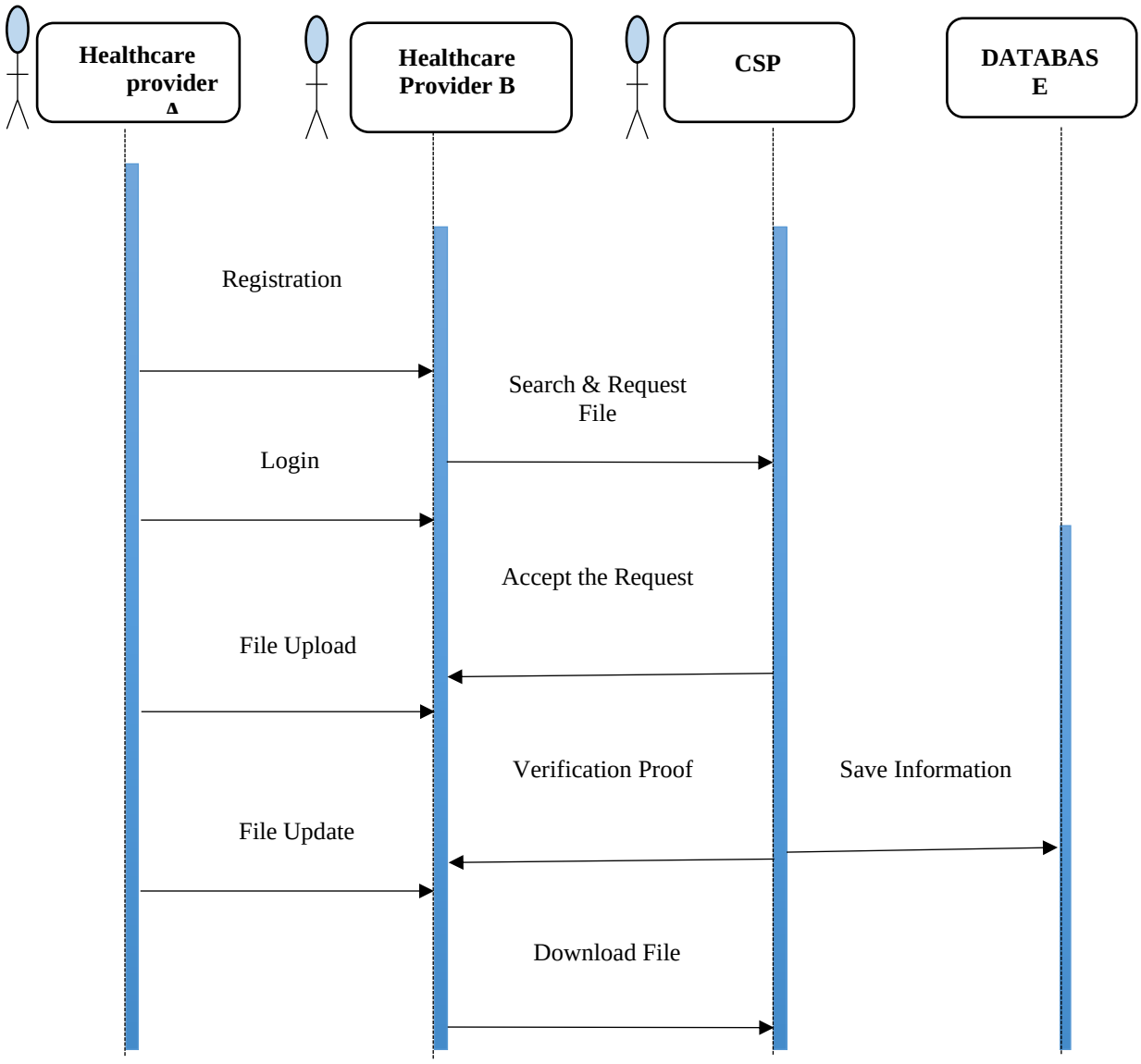


Fig 5.3.3 Sequence Diagram

6. PROJECT CODING

6. PROJECT CODING

6.1 CODE TEMPLATES

Blockchain.java

```
package ehr_blockchain;

import static ehr_blockchain.Upload.filepath;

import static ehr_blockchain.Upload.nor;

import java.awt.Desktop;

import java.io.BufferedReader;

import java.io.File;

import java.io.FileReader;

import java.io.FileWriter;

import java.sql.Timestamp;

import java.util.Date;

import javax.swing.DefaultListModel;

import javax.swing.JOptionPane;

public class Blockcreation extends javax.swing.JFrame {

   public Blockcreation() {

      initComponents();

   }

private void jList1ValueChanged(javax.swing.event.ListSelectionEvent evt) {

Desktop desktop = Desktop.getDesktop();

      File dirToOpen = null;

         }

private void jButton1ActionPerformed(java.awt.event.ActionEvent evt)

private void jButton2ActionPerformed(java.awt.event.ActionEvent evt)
```

```
private void jButton3ActionPerformed(java.awt.event.ActionEvent evt)

public static void main(String args[]) {

    try {

        for (javax.swing.UIManager.LookAndFeelInfo info :
javax.swing.UIManager.getInstalledLookAndFeels()) {

            if ("Nimbus".equals(info.getName())) {

                javax.swing.UIManager.setLookAndFeel(info.getClassName());

                break;

            }}

    } catch (ClassNotFoundException ex) {
java.util.logging.Logger.getLogger(Blockcreation.class.getName()).log(java.util.logging.Level.S
EVERE, null, ex);

    } catch (InstantiationException ex) {
java.util.logging.Logger.getLogger(Blockcreation.class.getName()).log(java.util.logging.Level.S
EVERE, null, ex);

    } catch (IllegalAccessException ex) {
java.util.logging.Logger.getLogger(Blockcreation.class.getName()).log(java.util.logging.Level.S
EVERE, null, ex);

    } catch (javax.swing.UnsupportedLookAndFeelException ex) {
java.util.logging.Logger.getLogger(Blockcreation.class.getName()).log(java.util.logging.Level.S
EVERE, null, ex);

    }

    java.awt.EventQueue.invokeLater(new Runnable() {

        public void run() {

            new Blockcreation().setVisible(true);

        }

    });

}

private javax.swing.JButton jButton1;
```

```
    private javax.swing.JButton jButton2;
    private javax.swing.JButton jButton3;
    private javax.swing.JLabel jLabel1;
    private javax.swing.JLabel jLabel2;
    private javax.swing.JLabel jLabel3;
    private javax.swing.JList jList1;
    private javax.swing.JPanel jPanel1;
    private javax.swing.JScrollPane jScrollPane1;
    private javax.swing.JTextField jTextField1;
}
```

6.2 OUTLINE FOR VARIOUS FILES

These are the code files used in our project in sequence:

index.java

block.java

mainpage.java

sub.java

updown.java

upload.java

ehr_blockchain.java

viewrecords.java

encrypted.java

send_a_request.java

cloud.java

cloudprocess.java

blockcreation.java

download.java

requestfile.java

finaldownload.java

6.3 CLASS WITH FUNCTIONALITIES

In the project home page, we have three modules to which we can navigate to, they are Registration, Healthcare Provider, Cloud Service Provider. Let us look into the brief overview of each module.

REGISTRATION

User has to register inorder to secure the data which is done by generating a secret key. After registration, user's data is stored in a database.

HEALTHCARE PROVIDER

We use two healthcare providers, one is Healthcare Provider A, Healthcare provider B.

a)LOGIN

Healthcare Provider B logins with the secret key generated by Healthcare Provider A at the time of registration.

b)UPLOAD PATIENT RECORD

Healthcare Provider A uploads the patient records into a database by using the key generation for each record and encrypting the record**.**

c)DOWNLOAD PATIENT RECORD

Healthcare Provider B downloads the patient record by searching the records and sending the request to the cloud for secret key verification and it then is able to download the data.

d)KEY GENERATION

Healthcare Provider A acts as an authenticator and is responsible for generating the secret key.

e)ENCRYPT PATIENT RECORDS

Healthcare Provider A uploads the file and updates the files in an encrypted format.

f)BLOCKCREATION

In this process, we build a Blockchain with number of blocks as per requirement and upload it to the cloud server.

CLOUD SERVICE PROVIDER

The cloud service provider maintain all the patient records and also they can provide a permission to the user to access the data. In order to access the data, Healthcare Provider A needs to login into the cloud server.

a)PUBLIC VERIFICATION KEY

Cloud service provider is responsible for public key verification.

b)GRANT/ REVOKE PERMISSION

By verifying the public key, the data owner and the data user add other layer of protection to the documents in the cloud by confirming each other's identity. The cloud service provider receives the document request from the data user that is Healthcare Provider B verifies the authentication before granting permission. Then the cloud service provider executes the query and returns the encrypted documents according to the search token.

c)VIEW PATIENT RECORDS

After the authentication is done and the data user is verified, the patient records are now able to acces**s.**

6.4 METHODS INPUT AND OUTPUT PARAMETERS

The methods that are used in our project are:

1.public Blockcreation() {

initComponents();

}

This method is called from within the constructor to initialize the form. It creates new form Blockchain.

Result: We can create the blocks as per the user requirements. In this method, we use the Blockchain technology for securing data.

2.private void jButton1ActionPerformed(java.awt.event.ActionEvent evt)

In this method, we used the functionality of JButton. The JButton is used used to create a labeled button that has platform independent implementation.

Result: JButtonActionPerformed(actionevent)- called just after the user performs an action.

In this method, we use ActionPerformed as the main method for performing the actions for the user requirement. We used different events in ActionPerformed.

3.public Cloud() {

initComponents();

}

This method is called from within the constructor to initialize the form. It creates new form of cloud.

Result: We use cloud which helps the user to access the stored data from any interface and platform.

4.public Finaldownload() {

initComponents();

}

This method gets the AES encryption key. In your actual programs, this should be safely

stored.

Result: The final data required for the user from the cloud is accessed using a public verification key by AES algorithm as the key element for authenticating the legal user.

5.public static SecretKey getSecretEncryptionKey()

In our project, this method encrypts plainText in AES using the secret key.

Result: The plain text is being stored in cloud by encryption process where the plain text is converted into cipher text using AES algorithm with the generation of secret key as the crucial element.

At the final stage, we have taken a sample dataset of Electronic Health Records which we uploaded in the cloud server and performed encryption to generate a secret key and a cloud for storing the data. Finally, we used Blockchain and AI algorithms for securing Electronic Health Records(EHR). According to the user requirement , user can access the data only if he is authorized.

7. PROJECT TESTING

7. PROJECT TESTING

7.1 VARIOUS TEST CASES

AGILE TESTING

Agile Testing is a type of software testing that accommodates agile software development approach and practices. In an Agile development environment, testing is an integral part of software development and is done along with coding. Agile testing allows incremental and iterative coding and testing.

FUNCTIONAL TESTING

Functional testing is a formal type of testing performed by testers. Functional testing focuses on testing software against design document, Use cases and requirements document. Functional testing is a black box type of testing and does not require internal working of the software unlike white box testing.

GLASS BOX TESTING

Glass box testing is another name for White box testing. Glass box testing is a testing method that involves testing individual statements, functions etc., Unit testing is one of the Glass box testing methods.

INTEGRATION TESTING

Integration testing also known as met in short, in one of the important types of software testing. Once the individual units or components are tested by developers as working then testing team will run tests that will test the connectivity among these units/component or multiple units/components. There are different approaches for Integration testing namely, Top-down integration testing, Bottom-up integration testing and a combination of these two known as Sand witch testing.

NON-FUNCTIONAL TESTING

Software are built to fulfil functional and non-functional requirements, non-functional requirements like performance, usability, localization etc., There are many types of testing like

compatibility testing, compliance testing, localization testing, usability testing, volume testing etc., that are carried out for checking non-functional requirements.

AD-HOC TESTING

This type of software testing is very informal and unstructured and can be performed by any stakeholder with no reference to any test case or test design documents. The person performing Ad-hoc testing has a good understanding of the domain and workflows of the application to try to find defects and break the software. Ad-hoc testing is intended to find defects that were not found by existing test cases.

7.2 BLACKBOX TESTING

- ✓ Black box testing is done to find incorrect or missing function
- ✓ Interface error
- ✓ Errors in external database access
- ✓ Performance errors
- ✓ Initialization and termination errors

In 'functional testing', is performed to validate an application conforms to its specifications of correctly performs all its required functions. So this testing is also called 'black box testing'. It tests the external behaviour of the system. Here the engineered product can be tested knowing the specified function that a product has been designed to perform, tests can be conducted to demonstrate that each function is fully operational.

UNIT TESTING

Unit testing is usually conducted as part of a combined code and unit test phase of the software lifecycle, although it is not uncommon for coding and unit testing to be conducted as two distinct phases.

TEST STRATEGY AND APPROACH

Field testing will be performed manually and functional tests will be written in detail.

TEST OBJECTIVES

- All field entries must work properly.

- Pages must be activated from the identified link.

- The entry screen, messages and responses must not be delayed.

FEATURES TO BE TESTED

- Verify that the entries are of the correct format.

- No duplicate entries should be allowed.

- All links should take the user to the correct page.

INTEGRATION TESTING

Software integration testing is the incremental integration testing of two or more integrated software components on a single platform to produce failures caused by interface defects.The task of the integration test is to check that components or software applications, e.g. components in a software system or – one step up– software applications at the company level – interact without error.

TEST RESULTS

All the test cases mentioned above passed successfully. No defects encountered.

ACCEPTANCE TESTING

User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements.

7.3 WHITEBOX TESTING

White Box testing is a test case design method that uses the control structure of the procedural design to drive cases. Using the white box testing methods, we derived test cases that guarantee that all independent paths within a module have been exercised at least once.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

8. OUTPUT SCREENS

8. OUTPUT SCREENS

8.1 USER INTERFACE

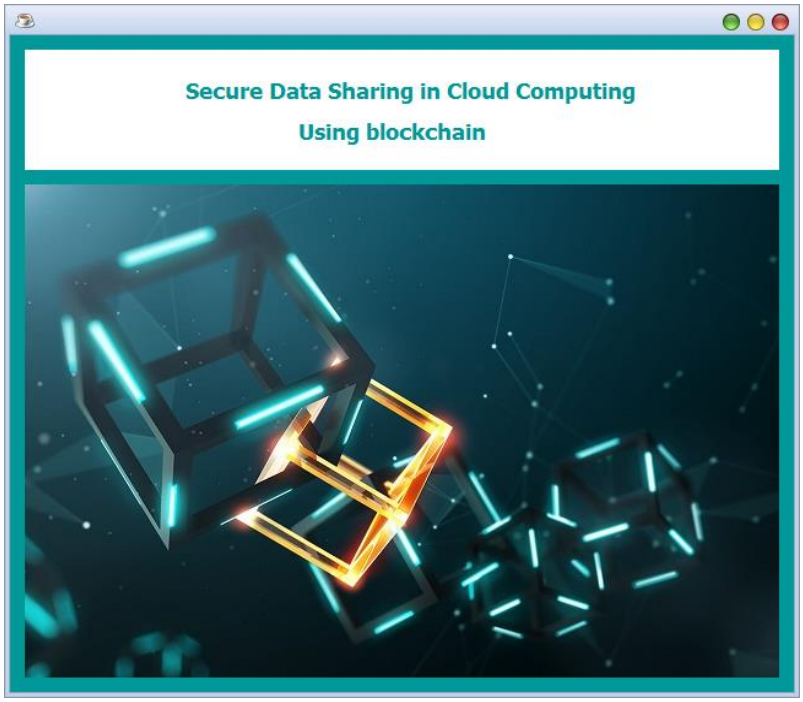


Fig 8.1.1 Home Page

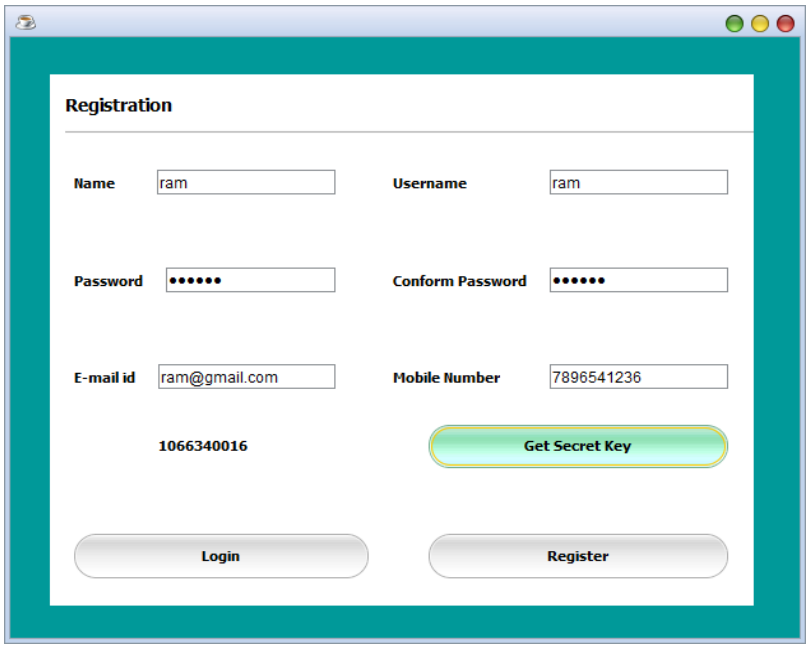


Fig 8.1.2 Registration

Fig 8.1.3 Healthcare Provider login

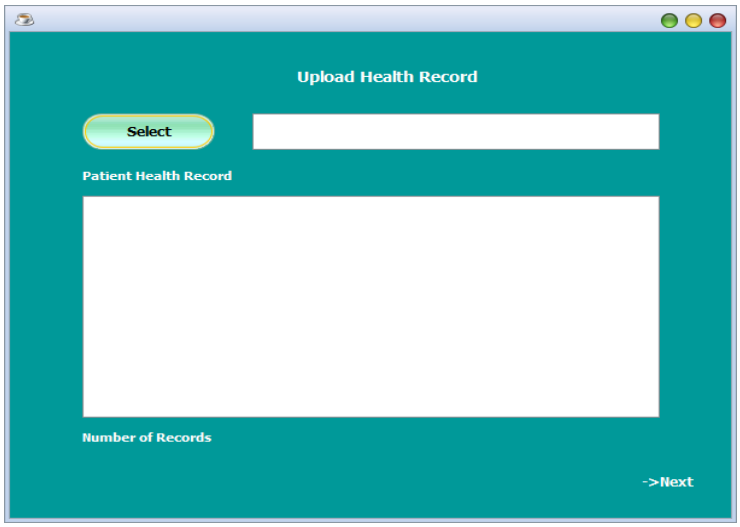


Fig 8.1.4 Upload Health record

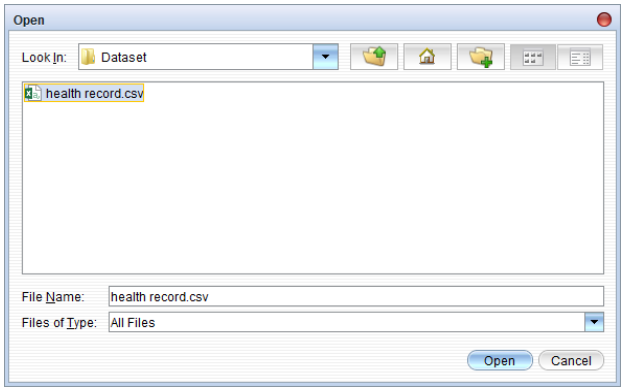


Fig 8.1.5 Database

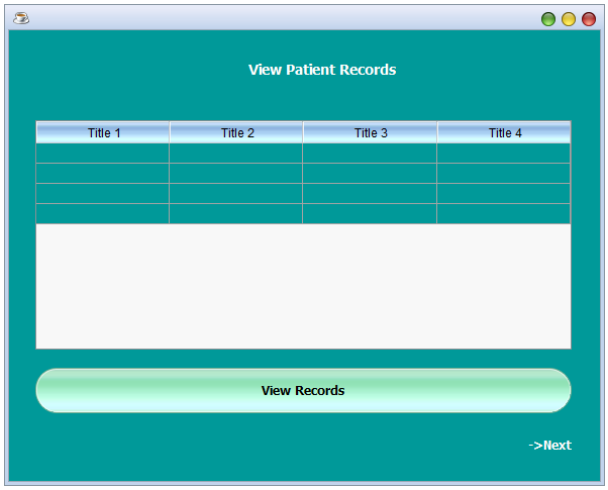


Fig 8.1.6 View patient records

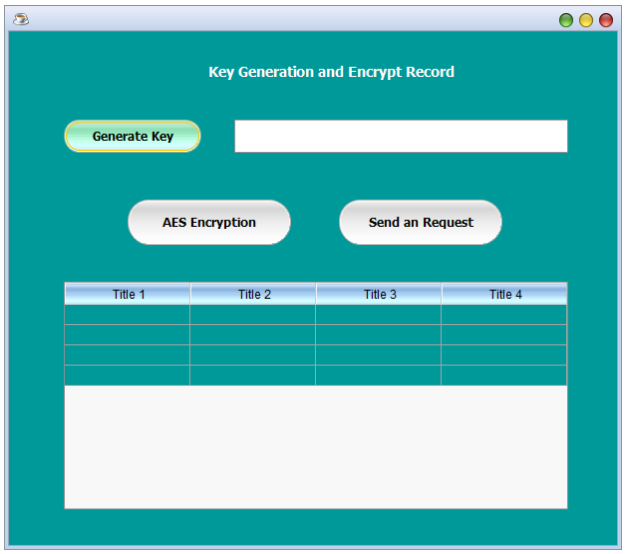


Fig 8.1.7 key generation and encrypt record

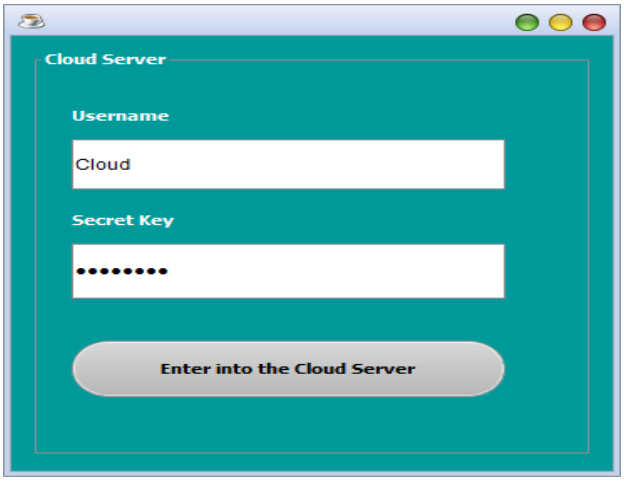


Fig 8.1.8 cloud server

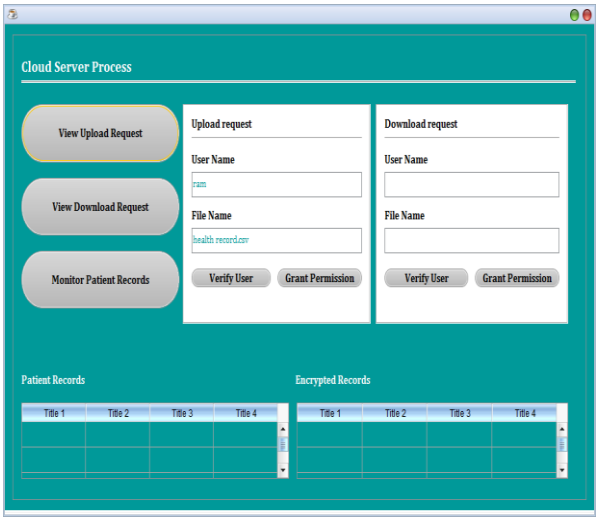


Fig 8.1.9 cloud server process

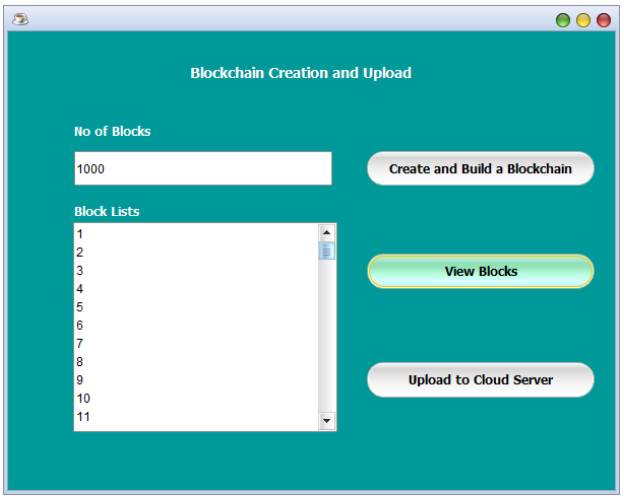


Fig 8.1.10 blockchain creation and upload

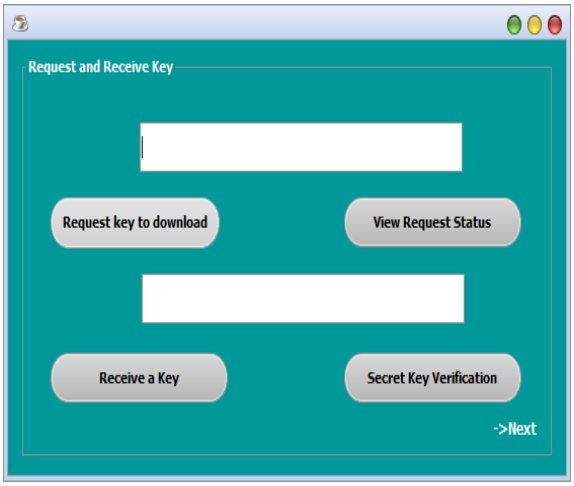


Fig 8.1.11 Request and Receive key

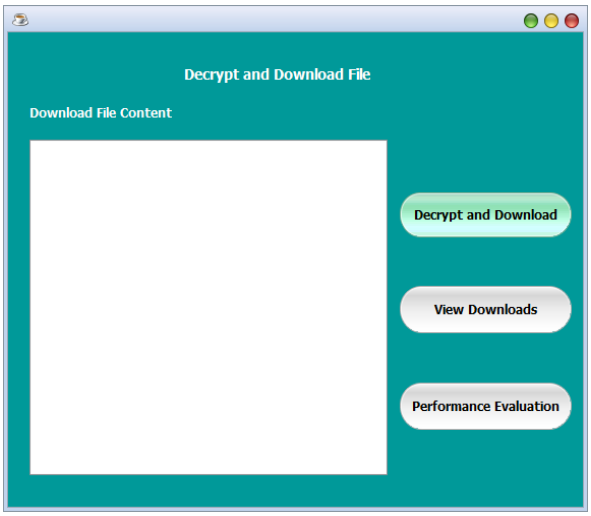


Fig 8.1.12 Encrypt and download file

8.2 OUTPUT SCREENS

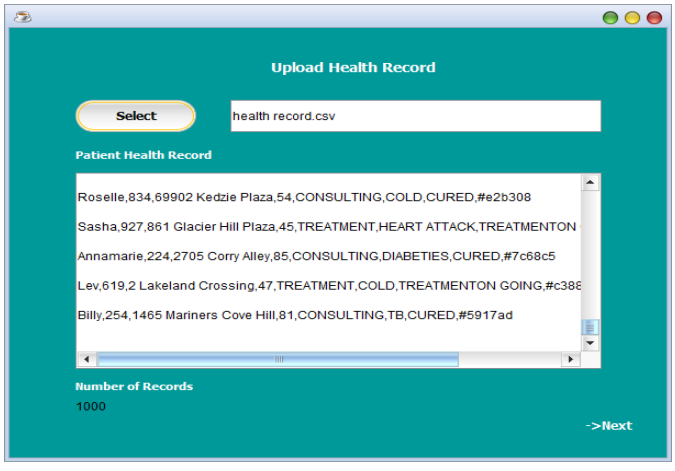


Fig 8.2.1 Upload Health Record

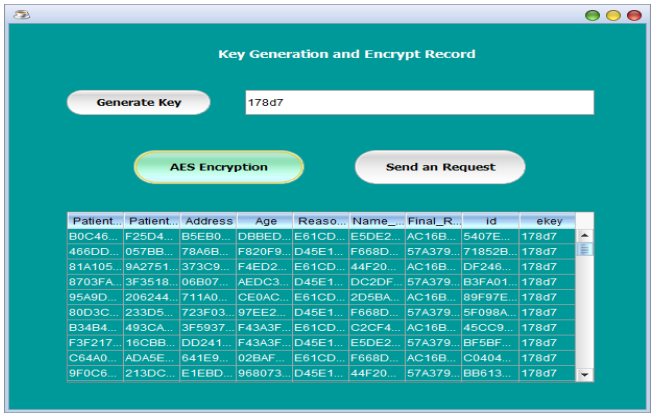


Fig 8.2.2 Key generation and encrypt record

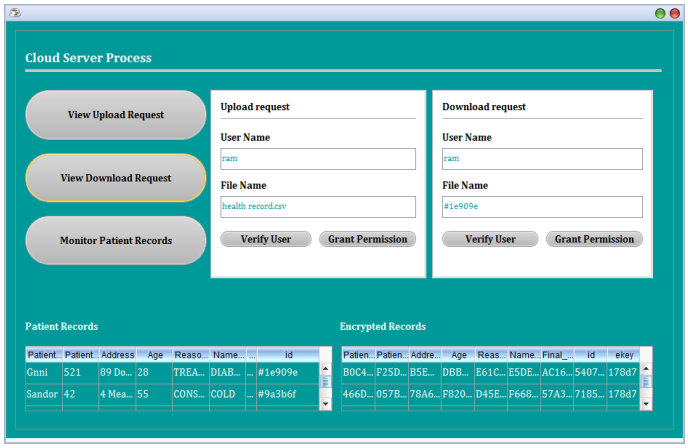


Fig 8.2.3 Cloud Server process

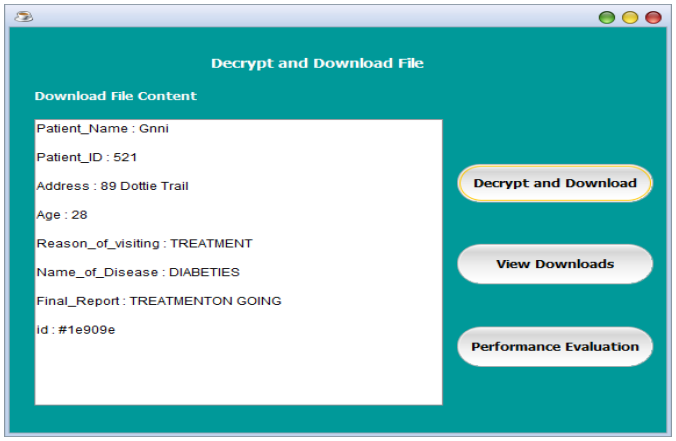


Fig 8.2.4 Decrypt and download file

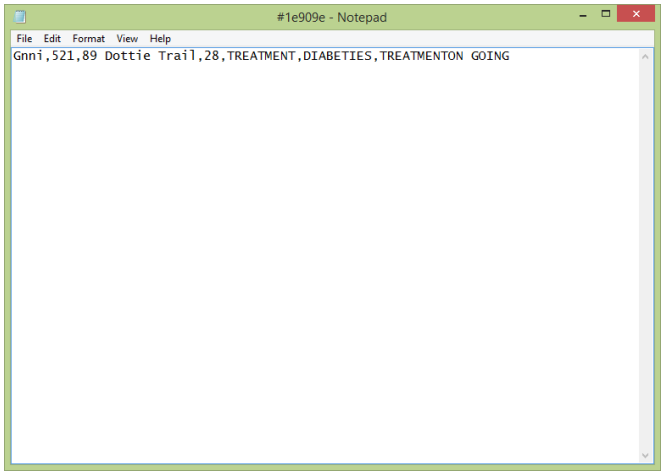


Fig 8.2.5 Output Screen

9. EXPERIMENTAL RESULTS

9. EXPERIMENTAL RESULTS

Considering the existing system and proposed system implementations, we come to a conclusion that the proposed system Blockchain ledger and AI technology embedded with cloud computing technology can advance the biomedical and health care domains in various novel ways, and we expect many new applications to emerge soon. As we implement the proposed project, we understand that the key aspects like the cost, time, size of the elements can be accessed to the at most level and enhance the privacy, security of data through exchange of data in the health care domains. This system is blockchain-based and provides data provenance, auditing, and control for shared medical data in cloud repositories among big data entities.

Fig 9.1 signing cost vs time

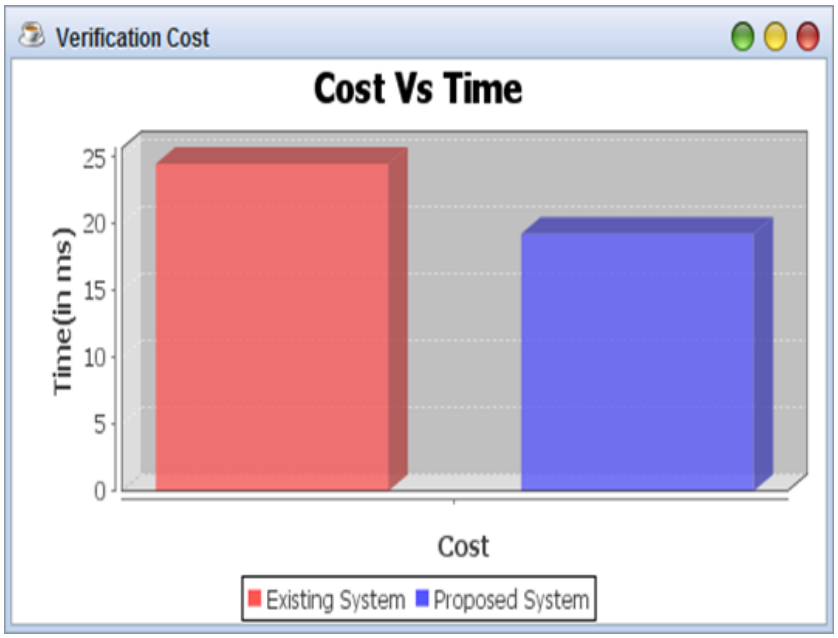


Fig 9.2 cost vs time

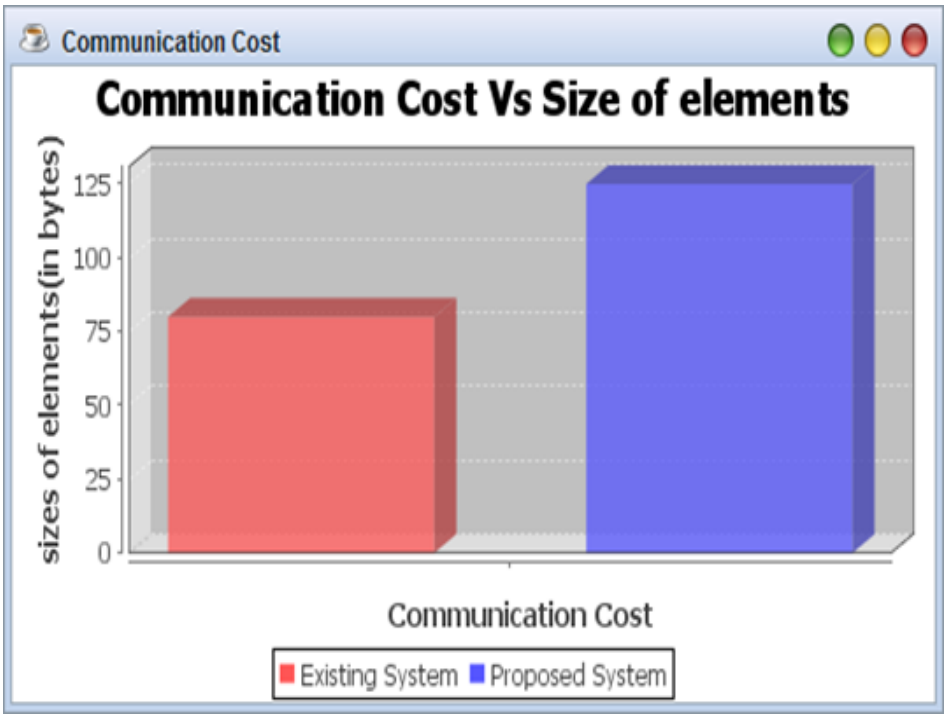


Fig 9.3 communication cost vs size of elements

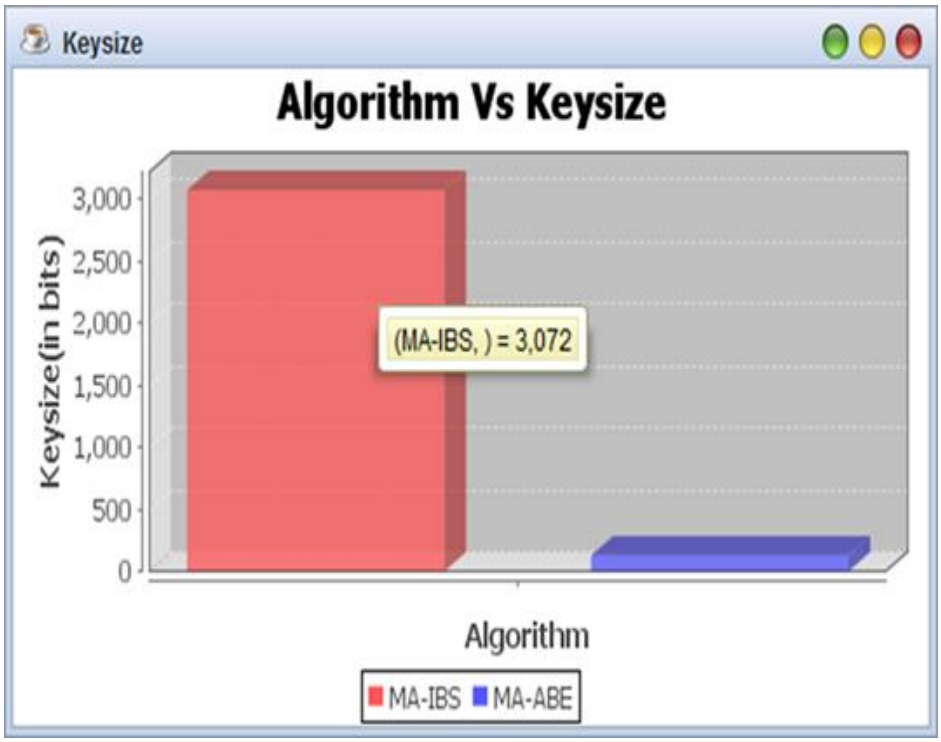


Fig 9.4 Algorithm vs Keysize

CONCLUSION AND FUTURE ENHANCEMENT

Inorder to leverage AI and blockchain to fit the problem of abusing data, as well as empower AI with the help of blockchain for trusted data management in trust-less environment, we propose the SecNet, which is a new networking paradigm focusing on secure data storing, sharing and computing instead of communicating. SecNet provides data ownership guaranteeing with the help of blockchain technologies, and AI-based secure computing platform as well as blockchain-based incentive mechanism, offering paradigm and incentives for data merging and more powerful AI to finally achieve better network security. Moreover, we discuss the typical use scenario of SecNet in medical care system, and gives alternative ways for employing the storage function of SecNet. Furthermore, we evaluate its improvement on network vulnerability when countering DDoS attacks, and analyze the inventive aspect on encouraging users to share security rules for a more secure network.In future work, we will explore how to leverage blockchain for the access authorization on data requests, and design secure and detailed smart contracts for data sharing and AI-based computing service in SecNet. In addition, we will model SecNet and analyze its performance through extensive experiments based on advanced platforms (e.g., integrating IPFS [27] and Ethereum [28] to form a SecNet-like architecture).

REFERENCES

[1] H. Yin, D. Guo, K. Wang, Z. Jiang, Y. Lyu, and J. Xing, ‘‘Hyperconnected network: A decentralized trusted computing and networking paradigm,’’ IEEE Netw., vol. 32, no. 1, pp. 112–117, Jan./Feb. 2018.

[2] K. Fan, W. Jiang, H. Li, and Y. Yang, ‘‘Lightweight RFID protocol for medical privacy protection in IoT,’’ IEEE Trans Ind. Informat., vol. 14, no. 4, pp. 1656–1665, Apr. 2018. [3] T. Chajed, J. Gjengset, J. Van Den Hooff, M. F. Kaashoek, J. Mickens, R. Morris, and N. Zeldovich, ‘‘Amber: Decoupling user data from Web applications,’’ in Proc. 15th Workshop Hot Topics Oper. Syst. (HotOS XV), Warth-Weiningen, Switzerland, 2015, pp. 1–6.

[4] M. Lecuyer, R. Spahn, R. Geambasu, T.-K. Huang, and S. Sen, ‘‘Enhancing selectivity in big data,’’ IEEE Security Privacy, vol. 16, no. 1, pp. 34– 42, Jan./Feb. 2018.

[5] Y.-A. de Montjoye, E. Shmueli, S. S. Wang, and A. S. Pentland, ‘‘openPDS: Protecting the privacy of metadata through SafeAnswers,’’ PLoS ONE, vol. 9, no. 7, 2014, Art. no. e98790.

[6] C. Perera, R. Ranjan, and L. Wang, ‘‘End-to-end privacy for open big data markets,’’ IEEE Cloud Comput., vol. 2, no. 4, pp. 44–53, Apr. 2015.

[7] X. Zheng, Z. Cai, and Y. Li, ‘‘Data linkage in smart Internet of Things systems: A consideration from a privacy perspective,’’ IEEE Commun. Mag., vol. 56, no. 9, pp. 55–61, Sep. 2018. 77988 VOLUME 7, 2019

[8] Q. Lu and X. Xu, ‘‘Adaptable blockchain-based systems: A case study for product traceability,’’ IEEE Softw., vol. 34, no. 6, pp. 21–27, Nov./Dec. 2017.

[9] Y. Liang, Z. Cai, J. Yu, Q. Han, and Y. Li, ‘‘Deep learning based inference of private information using embedded sensors in smart devices’’ IEEE Netw. Mag., vol. 32, no. 4, pp. 8–14, Jul./Aug. 2018.

[10] Q. Xia, E. B. Sifah, K. O. Asamoah, J. Gao, X. Du, and M. Guizani, ‘‘MeDShare: Trust-less medical data sharing among cloud service providers via blockchain,’’ IEEE Access, vol. 5, pp. 14757–14767, 2017.

[11] D. E. O’Leary, ‘‘Artificial intelligence and big data,’’ IEEE Intell. Syst., vol. 28, no. 2, pp. 96–99, Mar. 2013.

[12] A. Halevy, P. Norvig, and F. Pereira, ‘‘The unreasonable effectiveness of data,’’ IEEE Intell. Syst., vol. 24, no. 2, pp. 8–12, Mar. 2009.

[13] Z. Cai and X. Zheng, ‘‘A private and efficient mechanism for data uploading in smart cyber-physical systems,’’ IEEE Trans. Netw. Sci. Eng., to be published. doi: 10.1109/TNSE.2018.2830307.

[14] A. Dorri, M. Steger, S. S. Kanhere, and R. Jurdak, ‘‘BlockChain: A distributed solution to automotive security and privacy,’’ IEEE Commun. Mag., vol. 55, no. 12, pp. 119–125, Dec. 2017.

[15] J. Wang, M. Li, Y. He, H. Li, K. Xiao, and C. Wang, ‘‘A blockchain based privacy-preserving incentive mechanism in crowdsensing applications,’’ IEEE Access, vol. 6, pp. 17545–17556, 2018.

[16] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, ‘‘Revisiting unreasonable effectiveness of data in deep learning era,’’ in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 843–852.

[17] W. Meng, E. W. Tischhauser, Q. Wang, Y. Wang, and J. Han, ‘‘When intrusion detection meets blockchain technology: A review,’’ IEEE Access, vol. 6, pp. 10179–10188, 2018.

[18] J.-H. Lee, ‘‘BIDaaS: Blockchain based ID as a service,’’ IEEE Access, vol. 6, pp. 2274–2278, 2017.

[19] K. Wang, H. Yin, W. Quan, and G. Min, ‘‘Enabling collaborative edge computing for software defined vehicular networks,’’ IEEE Netw., vol. 32, no. 5, pp. 112–117, Sep./Oct. 2018.

[20] A. B. Kurtulmus and K. Daniel, ‘‘Trustless machine learning contracts; evaluating and exchanging machine learning models on the ethereum blockchain,’’ 2018, arXiv:1802.10185. [Online]. Available: https://arxiv.org/abs/1802.10185

[21] A. L. Buczak and E. Guven, ‘‘A survey of data mining and machine learning methods for cyber security intrusion detection,’’ IEEE Commun. Surveys Tuts., vol. 18, no. 2, pp. 1153–1176, 2nd Quart., 2016.

[22] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, ‘‘Generative adversarial networks,’’ 2014, arXiv:1406.2661. [Online]. Available: https://arxiv.org/ abs/1406.2661

[23] E. C. Ferrer, ‘‘The blockchain: A new framework for robotic swarm systems,’’ 2017, arXiv:1608.00695. [Online]. Available: https://arxiv.org/ abs/1608.00695

[24] IPFS. Accessed: Jun. 5, 2019. [Online]. Available: https://ipfs.io/

[25] S. T. Zargar, J. Joshi, and D. Tipper, ‘‘A survey of defense mechanisms against distributed denial of service (DDoS) flooding attacks,’’ IEEE Commun. Surveys Tuts., vol. 15, no. 4, pp. 2046–2069, 4th Quart., 20.

PUBLICATIONS

CONFERENCE

- International Conference on “Securing Data with Blockchain and Artificial Intelligence”(ICICCI-21-0141)

- Paper ID: ICICCI-21-0141

ALL FOUR STUDENTS' ONE PAGE PROFILE

1.ALLURI SHIVANI (17K81A0502)

SHIVANI ALLURI is currently pursuing her graduation from St Martin's Engineering College in the stream of Computer Science. She completed her intermediate from SR Junior College and 10th class from Presidency High School. She participated in various events, seminars and workshops during her graduation, some of them are:

S.NO	EVENTS/SEMINARS/COURSES
1	Participated in Employability Skill development Program conducted by Zensar
2	Student of Smart Interviews
3	Participated in National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing"
4	Participated in Women online workshop on "Women in Cyber Security and Privacy in 2020"
5	Certification in Hacker rank (Python)
6	Certification in JavaScript By the Net Ninja in Cursa
7	Certification in Python Core in SoloLearn
8	Certification in MySQL database by the New Boston in Cursa
9	Certification in CyberSecurity by PacketHacks in Cursa
10	Participated in Anti-Drug Campaign conducted by Lush life Bistro
11	Participated in National Level Seminar on Recent Trends in Cloud Computing,Fog and Edge Computing.

2.MANDULA MEGHANA (17K81A0534)

MANDULA MEGHANA is currently pursuing her graduation from St Martin's Engineering College in the stream of Computer Science. She completed her intermediate from Sri Chaitanya Junior College and 10th class from Vignana Jyothi Public School. She participated in various events, seminars and workshops during her graduation, some of them are:

S.NO	EVENTS/SEMINARS/COURSES
1	Participated in Employability Skill development Program conducted by Zensar
2	Participated in National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing"
3	Participated in Women online workshop on "Women in Cyber Security and Privacy in 2020"
4	Certification in Hacker rank (Python)
5	Certification in Learn to code in Python3 in Udemy
6	Certification in Introduction to AI from ElementsofAI
7	Certification in The fundamentals of Digital Marketing in Google Digital Unlocked
8	Certification in MySQL database by The new Boston in Cursa
9	Certification in Programming with PHP for beginners in Cursa
10	Certification in 30 days to learn HTML and CSS in Cursa
11	Certification in AI for everyone from deeplearning.ai in coursera

3.MATTA JAHNAVI REDDY (17K81A0535)

MATTA JAHNAVI is currently pursuing her graduation from St Martin's Engineering College in the stream of Computer Science. She completed her intermediate from Sri Chaitanya Junior College and 10th class from Sri Chaitanya Techno School. She participated in various events, seminars and workshops during her graduation, some of them are:

S.NO	EVENTS/SEMINARS/COURSES
1	Participated in Employability Skill development Program conducted by Zensar
2	Participated in National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing"
3	Certification in Management Managerial Accounting by nptelhrd in Cursa
4	Certification in Principles of Construction Management by nptelhrd in Cursa
5	Certification in Principles of Copy Writing by Appy Pie in Cursa
6	Certification in 30 days to learn HTML and CSS in Cursa
7	Certification in HTML by EJ Media in Cursa
8	Participated in National Level Seminar on Recent Trends in Cloud Computing,Fog and Edge Computing.

4.TALAKOKKULA ASHWITHA (17K81A0553)

TALAKOKKULA ASHWITHA is currently pursuing her graduation from St Martin's Engineering College in the stream of Computer Science. She completed her intermediate from Sri Chaitanya Junior College and 10th class from Sri Chaitanya Techno School. She participated in various events, seminars and workshops during her graduation, some of them are:

S.NO	EVENTS/SEMINARS/COURSES
1	Participated in Employability Skill development Program conducted by Zensar
2	Participated in National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing"
3	Participated in Women online workshop on "Women in Cyber Security and Privacy in 2020"
4	Certification in Programming with PHP for beginners by the Net Ninja in Cursa
5	Certification in Basic of AWS concepts by ExamPro in Cursa
6	Certification in Artificial Intelligence by Crash Course in Cursa
7	Certification in Python for Beginners in SoloLearn
8	Certification in JavaScript in SoloLearn
9	Participated in National Level Seminar on Recent Trends in Cloud Computing,Fog and Edge Computing.

A

PROJECT REPORT

On

Fake News, Disinformation, and Deepfakes: Leveraging Distributed Ledger Technologies and Blockchain to Combat Digital Deception and Counterfeit Reality

Submitted by

1) Mr. Gattepalli Sai Charan (17K81A0518)
2) Mr. Ganji Akhil (17K81A0517)
3) Mr. Medepally Mouneeswar (17K81A0536)
4) Mr. Ryava Anurag Reddy (17K81A0542)

in partial fulfillment for the award of the

degreeof

BACHELOR OF TECHNOLOGY

IN

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Under the Guidance of

Dr. B. RAJALINGAM

Associate professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ST. MARTIN'S ENGINEERING COLLEGE
An Autonomous Institute

Dhulapally, Secunderabad – 500 100

JUNE 2021

BONAFIDE CERTIFICATE

This is to certify that the project entitled "Fake News, Disinformation, and Deepfakes: Leveraging Distributed Ledger Technologies and Blockchain to Combat Digital Deception and Counterfeit Reality", is being submitted by **1. Mr. Gattepalli Sai Charan 17K81A0518, 2. Mr. Ganji Akhil 17K81A0517, 3. Mr. Medepally Mouneeswar 17K81A0536, 4. Mr. Ryava Anurag Reddy 17K81A0542,** in partial fulfillment of the requirement for the award of the degree of **BACHELOR OF TECHNOLOGY IN COMPUTER SCIENCE AND ENGINEERING** is recorded of bonafide work carried out by them. The result embodied in this report have been verified and found satisfactory.

HEAD OF THE DEPARTMENT

Dr. B. RAJALINGAM
DEPARTMENT OF CSE

Dr. M. NARAYANAN
DEPARTMENT OF CSE

Internal Examiner **External Examiner**

Place:

Date:

DECLARATION

We, the student of **Bachelor of Technology** in Department of Computer Science and Engineering', session: 2017 – 2021, St.Martin's Engineering College, Dhulapally, Kompally, Secunderabad, hereby declare that work presented in this Project Work entitled "**Fake News, Disinformation, and Deepfakes: Leveraging Distributed Ledger Technologies and Blockchain to Combat Digital Deception and Counterfeit Reality"** is the outcome of our own bonafide work and is correct to the best of our knowledge and this work has been undertaken taking care of Engineering Ethics. This result embodied in this project report has not been submitted in any university for award of any degree.

Gattepalli Sai Charan	17K81A0518
Ganji Akhil	17K81A0517
Medepally Mouneeswar	17K81A0536
Ryava Anurag Reddy	17K81A0542

ABSTRACT

The rise of ubiquitous deepfakes, misinformation, disinformation, and post-truth, often referred to as fake news, raises concerns over the role of the Internet and social media in modern democratic societies. Due to its rapid and widespread diffusion, digital deception has not only an individual or societal cost, but it can lead to significant economic losses or to risks to national security. Blockchain and other distributed ledger technologies (DLTs) guarantee the provenance and traceability of data by providing a transparent, immutable, and verifiable record of transactions while creating a peer-to-peer secure platform for storing and exchanging information. This overview aims to explore the potential of DLTs to combat digital deception, describing the most relevant applications and identifying their main open challenges. Moreover, some recommendations are enumerated to guide future researchers on issues that will have to be tackled to strengthen the resilience against cyber-threats on today's online media.

Keyword: blockchain; DLT; deepfake; fake news; data traceability; decentralization; cybersecurity; dApps; information security; proof of authenticity; forensics.

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible and whose encouragements and guidance have crowded effects with success.

We extended our deep sense of gratitude to Principal**,** **Dr. P. SANTOSH KUMAR PATRA**, St. Martin's Engineering College, Dhulapally, for permitting us to undertake this project.

We are also thankful to **Dr. M. NARAYANAN**, Head of the Department, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

We would like to thank **Dr. T. POONGOTHAI,** Department of Computer Science and Engineering (AI & ML) for her encouragement and insightful comments and as well as our project coordinators **Dr. B. RAJALINGAM**, Associate Professor and **Mr. J. SUDHAKAR**, Assistant Professor, Department of Computer Science and Engineering for their valuable support.

We would like to express our sincere gratitude and indebtedness to our project supervisor **Dr. B.RAJALINGAM**, Associate Professor, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

Finally, we express thanks to all those who have helped us successfully to completing this project. Furthermore, we would like to thank our family and friends for their moral support and encouragement.

We express thanks to all those who have helped us in successfully completing the project.

Gattepalli Sai Charan	17K81A0518
Ganji Akhil	17K81A0517
Medepally Mouneeswar	17K81A0536
Ryava Anurag Reddy	17K81A0542

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

1 INTRODUCTION

Gartner predicts that most individuals in developed economies will consume more false than true information by 2022.1 Digital deception is commonly recognized as deceptive or misleading content created and disseminated to cause public or personal harm (e.g., post-truth, populism, and satire) or to obtain a profit (e.g., click baits, cloaking, ad farms, and identity theft). In the context of mass media, digital deception originates either from governments or non-state actors that publish content without economic or educational entrance barriers. Therefore, these horizontal and decentralized communications cannot be controlled or stopped with traditional centralized tools. In addition, this lack of supervision allows for security attacks (e.g., social engineering). Moreover, the veracity of information seems to be sometimes negotiable for the sake of profit, as the competition is increasingly tough.

While trust in mass media and established institutions is declining, the use of social media is rising sharply, and it has become an important source for the distribution of digital deception. Today, social media platforms miss an adequate regulation, and their responsibilities are still not clearly defined. Several issues are open, 2 such as the application of adequate data protection rules [e.g., General Data Protection Regulation (GDPR)] along with the market concentration in just a few social media companies worldwide.

Advances in artificial intelligence (AI) have recently been used to create sophisticated disinformation. As a result, several research projects as well as regulations have been launched to detect digital deception. 3 Nevertheless, researchers claim that ubiquitous content can be hardly supervised.

Today, distributed ledger technologies (DLTs) and specifically block chain present challenges, but also opportunities for stakeholders and policymakers as potential technologies that can help to combat digital deception. These technologies enable privacy, security, and trust in a decentralized peer-to-peer (P2P) network without any central managing authority. DLTs ability to combat digital deception is focused on controlling the traceability of the media, the communications architecture, and the transactions. However, the problems involved in developing effective ways to identify, test, transmit, and audit information are still open.

There are only a few articles of the literature that use block chain to combat digital deception and counterfeit reality, and they are mostly focused on tracing the source of the information. To the knowledge of the authors, this is the first article that proposes a global vision on how to confront fake news and deep fakes through DLTs with the aim of guiding researchers and managers on future developments. Thus, this article provides a comprehensive overview on the

applicability of DLTs to tackle digital deception, showing the potential of DLTs for revolutionizing the media industry.

The rest of this article is organized as follows. The "State-of-the Art" section provides an overview of current digital deception and the involved technologies. The section, "DLT-Based Applications to Combat Digital Deception," lists different DLT-based applications to combat digital deception and counterfeit reality. In the section "Challenges and Recommendations," the main challenges of the application of DLT to tackle digital deception are analysed and some recommendations are proposed. The "Conclusion" section is devoted to conclusions.

1.1 PROJECT OVERVIEW

This proposes a global vision on how to confront fake news and deep fakes through DLTs with the aim of guiding researchers and managers on future developments. Thus, it provides a comprehensive overview on the applicability of DLTs to tackle digital deception, showing the potential on DLTs for revolutionizing the media industry. Further, some recommendations are enumerated to guide future researchers on issues that will have to be tackled to strengthen the resilience against cyber-threats on today's online media.

Sentiment Analysis (also known as opinion mining or emotion AI) is a sub-field of NLP that tries to identify and extract opinions within a given text across blogs, reviews, social media, forums, news etc. Sentiment analysis is the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information age Processing (NLP).

1.2 PROJECT OBJECTIVES

The main objective is to detect Fake News, Disinformation and Deepfakes in a platform like E-commerce websites. Using the technologies Distributed Ledger Technologies and Blockchain we can detect the Fake news to combat digital deception and counterfeit reality. Distributed ledger technologies (DLTs) and specifically block chain present challenges, but also opportunities for stakeholders and policymakers as potential technologies that can help to combat digital deception. These technologies enable privacy, security, and trust in a decentralized peer-to-peer (P2P) network without any central managing authority.

1.3 ORGANIZATION OF CHAPTERS

The thesis is organized in the following chapters:

Chapter 1: Introduction

This assessment seeks to investigate the potential of DLTs and blockchains in the fight against digital fraud, assess current projects and identify their key difficulties. In addition, a few ideas are included as an essential component of increasing resistance to cybernetic risks on today's online media in order for future studies to address fake news, misinformation and depths.

Chapter 2: Literature Survey

There are only a few articles of the literature that use blockchain to combat digital deception and they are mostly focused on tracing the source of the information. To the knowledge of the authors, this is the first article that proposes a holistic approach to combat digital deception through DLTs. Thus, this article provides a comprehensive overview on the phenomenon and its prevalence, on the applicability of DLTs to tackle digital deception and on the main challenges they pose.

Chapter 3: Software and Hardware Requirements

Through our Detailed Engineering and Design service, we take your goals and the abstract vision of a working system and transform that information into the final automation design. To round out your system’s final design, we specify the hardware and software that will be critical components of your OT (operational technology) infrastructure. In today’s modern automation systems, hardware and software are closely coupled to provide agile integration as well as the ability to quickly harvest data, transforming it into actionable intelligence through advanced software platforms. Our systems use the most advanced hardware and software technologies available today, and we make it a goal to choose components that work together well as a cohesive system as well as provide a path for future growth and sustainability. This guide outlines minimum software and hardware requirements for deploying the project. Requirements may vary based on utilization and observing performance of pilot projects is recommended prior to scale out.

Chapter 4: Software Development Analysis

The development and implementation of the design parameters. Developer's code based on the product specifications and requirements agreed upon in the previous stages. Following company procedures and guidelines, front-end developers build interfaces and back-ends while database administrators create relevant data in the database. The programmers also test and review each other's code.

Chapter 5: Project System Design

Design is the stage of the software development process. Here, architects and developers draw up advanced technical specifications they need to create the software to requirements. Stakeholders will discuss factors such as risk levels, team composition, applicable technologies, time, budget, project limitations, method and architectural design.

Chapter 6: Project Coding

A programming project produces a well-designed executing system that solves a specified distributed programming problem. A project code is used to represent a one-time, or intermittent departmental event or activity. Any person can use a project code on a transaction, regardless of the project manager or home organization. This section describes some of the coding templates, outline of various files, class with functionalities, the various methods of input and output parameters.

Chapter 7: Project Testing

The testing phase checks the software for bugs and verifies its performance before delivery to users. In this stage, expert testers verify the product's functions to make sure it performs according to the requirements analysis document.

Testers use exploratory testing if they have experience with that software or a test script to validate the performance of individual components of the software. They notify developers of defects in the code. If developers confirm the flaws are valid, they improve the program, and the testers repeat the process until the software is free of bugs and behaves according to requirements.

Chapter 8: Output screens

The output of the programmed project is being screened with the screenshots. Front end development is done which is connected with the back-end servers database and the operations are done with the final input. The various test case results are captured and projected some sample outputs.

Chapter 9: Experimental Results

For use, a credential must be provided to the administrator. It can execute specified activities like listing after successful login and all users can perform them. Enter the name and username of the news channel, add news categories, Fix the quantization date for the news, Selecting and adding news category, List, update and delete all news items List of news articles on the distributed booklet technology, list all news blocks on cat news, list all user information keyword transactions, Distributed technology chart view internet products, check chart rank for all news items.

CHAPTER 2
LITERATURE SURVEY

2 LITERATURE SURVEY

A systematic and thorough search of all types of published literature as well as other sources including dissertation, theses in order to identify as many items as possible that are relevant to a particular topic.

A literature review is an overview of the previously published works on a specific topic. The term can refer to a full scholarly paper or a section of a scholarly work such as a book, or an article. Either way, a literature review is supposed to provide the researcher/author and the audiences with a general image of the existing knowledge on the topic under question. A good literature review can ensure that a proper research question has been asked and a proper theoretical framework and/or research methodology have been chosen. In other words, a literature review serves to situate the current study within the body of the relevant literature and to provide context for the reader. In such a case, the review usually precedes the methodology and results sections of the work.

Producing a literature review is often a part of graduate and post-graduate student work, including in the preparation of a thesis, dissertation, or a journal article. Literature reviews are also common in a research proposal or prospectus (the document that is approved before a student formally begins a dissertation or thesis).

A literature review can be a type of review article. In this sense, a literature review is a scholarly paper that presents the current knowledge including substantive findings as well as theoretical and methodological contributions to a particular topic. Literature reviews are secondary sources and do not report new or original experimental work. Most often associated with academic-oriented literature, such reviews are found in academic journals and are not to be confused with book reviews, which may also appear in the same publication. Literature reviews are a basis for research in nearly every academic field.

2.1 SURVEY ON BACKGROUND

Paula Fraga-Lamas, et al. proposed that [1] This study attempts to investigate the potential of distributed ledger technologies (DLTs) to prevent digital deception by defining the most relevant applications and outlining the key problems they face. Furthermore, some recommendations are made to advise future studies on topics that must be addressed to increase cyber-threat resistance on today's online media. Zonyin Shae, et al. proposed that [2] Since the solutions to the fake news dilemma rely not just on AI, but also on social factors, an interdisciplinary effort is required. We propose an AI blockchain platform in this study to foster robust collaboration between AI blockchain researchers and news organizations in the battle

against false news. D. Roy, S. Aral, and S. Vosoughi. [three] From 2006 to 2017, we looked at the differential dissemination of all confirmed factual and misleading news reports on Twitter. The dataset comprises 126,000 items that were posted more than 4.5 million times by 3 million users. We used data from six different fact-checking groups to classify stories as true or false, with 95 to 98 percent agreed on the classifications. H. Kim et al. [4] They describe a unique method for re-animation of portrait films that is photo-realistic and requires only one input video. We are the first to transfer the entire 3D head position, head rotation, face expression, eye gaze, and eye blinking from a source actor to a portrait video of a target actor, in contrast to prior techniques that are confined to manipulations of facial emotions alone. A generative neural network with a new space-time architecture is at the heart of our method. I. Khan, A. Shahaab, B. Lidgey, C. Hewage. [5] Consensus mechanisms have advanced in recent years, allowing distributed ledger technologies (DLTs) to find use and value in industries other than cryptocurrencies. We looked at 66 well-known consensus protocols and categorized them into philosophical and architectural categories, as well as providing a graphic depiction of them. J. Qadir, et al. [6] "Fake news" has become a global phenomenon in recent years, posing unprecedented challenges to human civilization and democracy. This issue has arisen as a result of the emergence of several concomitant phenomena, including 1) the digitization of human life and the ease with which news can be disseminated via social networking applications (such as Facebook and WhatsApp); 2) the availability of "big 3 data," which allows for the customization of news feeds and the creation of polarized "filter-bubbles"; and 3) the rapid progress made by generate (such as text, images, and videos). X. Zhang, et al. [7] For a variety of reasons, this is a technological challenge. Content is readily created and swiftly shared through social media platforms, resulting in a tremendous number of data to evaluate. This work is made more difficult by the fact that online content is incredibly diversified, covering a wide range of topics. W. Ding, et al. [8] In the information sciences and social sciences, decentralized autonomy has been a longstanding study issue. Its early expressions include the self-organization phenomena in natural ecosystems, Cyber Movement Organizations (CMOs) on the Internet, and Distributed Artificial Intelligence (DAI), among others. [9] aims at examining information problems and related obstacles, including filter bubbles and echo chambers in a thorough way. While the historical impact of rumours and fabricated content has been well documented, we argue that we are witnessing something new: global information pollution; a complex web of motivations for creating, disseminating, and consuming these ‘polluted' messages; a plethora of content types and techniques for amplifying content; and innumerable platforms for amplifying content. [10] We feel that our program can only provide a partial answer for false news, thus it can verify the originality of media resources. Because it is unable to prove that a news article is legitimate. [11] The first steps are to be taken. We live in a world

of unpredictability and Gartner's top strategic forecasts for 2021 look at how technology might assist reset the company and restart it. [12] In order to provide a healthy atmosphere for news communication, false news must be suppressed, and the source of false news must be rejected, and the source of news must be traced. This document traces the news with the technology of consensus algorithm, and the intelligent contract based on distributed storage, decentralization, and other properties of the blockchain. [13] However, the system is sufficiently general and may be used on every other kind of digital content. This solution concentrates on video content. Our technique focuses on the premise that the material may then be legitimate and authentic if it can be credibly tracked to an independent or credible source. [14] Because blockchain technology can safely and flawlessly retain data, it is a good platform for notarization online activities. The issue is how the data can be checked. We offer an architecture that allows you to preserve social media material legitimately using blockchain. [15] This first paper explores current state-of-the-art cryptosystems after quantity and how they might be used to blockchains and DLTs. In addition, the major issues are explored using the most relevant post-quantum blockchain systems. In addition, comprehensive comparisons are made of the properties and performance of the most promising public-key post-quanta encryption and digital blockchain signature techniques.

2.2 CONCLUSIONS ON SURVEY

There are only a few articles of the literature that use blockchain to combat digital deception and counterfeit reality, and they are mostly focused on tracing the source of the information also just a few articles in the literature that study the applicability of DLTs to face fake news and deepfakes, all of them are quite preliminary and just focus only on a specific application. To the knowledge of the authors, this is the first article that proposes a global vision on how to confront fake news and deepfakes through DLTs with the aim of guiding researchers and managers on future developments.

CHAPTER 3

SOFTWARE AND HARDWARE REQUIREMENTS

3 SOFTWARE AND HARDWARE REQUIREMENTS

To be used efficiently, all computer software needs certain hardware components or other software resources to be present on a computer. These prerequisites are known as (computer) system requirements and are often used as a guideline as opposed to an absolute rule. Most software defines two sets of system requirements: minimum and recommended. With increasing demand for higher processing power and resources in newer versions of software, system requirements tend to increase over time. Industry analysts suggest that this trend plays a bigger part in driving upgrades to existing computer systems than technological advancements. A second meaning of the term of system requirements, is a generalisation of this first definition, giving the requirements to be met in the design of a system or sub-system.

3.1 SOFTWARE REQUIREMENTS

Operating System	- Windows 8 and above
Coding language	- Java/J2EE (JSP Servlet)
Front End	- J2EE
Back End	- MySQL
Servers	- Apache Tomcat 7

3.2 HARDWARE REQUIREMENTS

Processor	- Intel Core I5
RAM	- 4 GB
Hard Disk	- 500 GB
Keyboard	- Standard Windows Keyboard
Mouse	- Two or Three Button Mouse
Monitor	- SVGA

CHAPTER 4

SOFTWARE DEVELOPMENT ANALYSIS

4 SOFTWARE DEVELOPMENT ANALYSIS

The software development process involves the creation and maintenance of applications, frameworks and other components for software design, design, programming, documentation, testing and problem remediation. The development of software is a process of creating and keeping source code, but it encompasses everything from the idea of the intended software to the last manifestation of the programme, often in a planned and organised process in a larger context. Software development may therefore encompass research, creation of new software products, prototype, modification, reuse, reengineering, maintenance, or any other software-production activity.

Software development is the process of conceiving, specifying, designing, programming, documenting, and bug fixing involved in creating and maintaining applications, frameworks, or other software components. Software development is a process of writing and maintaining the source code, but in a broader sense, it includes all that is involved between the conception of the desired software through to the final manifestation of the software, sometimes in a planned and structured process.[1] Therefore, software development may include research, new development, prototyping, modification, reuse, re-engineering, maintenance, or any other activities that result in software products.[2]

The software can be developed for a variety of purposes, the three most common being to meet specific needs of a specific client/business (the case with custom software), to meet a perceived need of some set of potential users (the case with commercial and open source software), or for personal use (e.g. a scientist may write software to automate a mundane task). Embedded software development, that is, the development of embedded software, such as used for controlling consumer products, requires the development process to be integrated with the development of the controlled physical product. System software underlies applications and the programming process itself, and is often developed separately.

The need for better quality control of the software development process has given rise to the discipline of software engineering, which aims to apply the systematic approach exemplified in the engineering paradigm to the process of software development.

There are many approaches to software project management, known as software development life cycle models, methodologies, processes, or models. The waterfall model is a traditional version, contrasted with the more recent innovation of agile software development.

4.1 OVERVIEW OF THE PROBLEM

The increase of omnipresent depth, misinformation, and post-truth, sometimes called false news, raises questions about the role of the Internet and social media in modern democratic countries. Digital deception not only carries an individual or social burden because of its quick and extensive spreading but may also lead to considerable losses or national security threats.

The current efforts of the research community are mostly focused on one type of fake news (i.e., verifiable false content), while other bad practices are barely studied. Most digital deception detection proposals are based on cryptographic hashes, which are sensitive to noise and, when there is a change of a character, a pixel, a bit in a certain content, it can result in a different hash.

4.2 DEFINE THE PROBLEM

Data origin and traceability may be guaranteed by Blockchain and other DLT technologies through the provision of transparent, changeable, and verifiable tracking of transactions, while establishing a secure peer-to-peer platform for the storage and exchange of data. This review seeks to investigate the potential of DLTs to fight digital disappointment, describe the applications that are most relevant and to highlight their key obstacles. In addition, certain recommendations are listed to assist future researchers on concerns which will be addressed in order to improve the resistance of today's online media against cyber-attacks.

While any minimal change in two resources will generate vastly different hashes, the use of perceptual hashes produces comparable results if the resources are similar. Hence, perceptual hashing is already used to detect copyright infringement, as well as in digital forensics. Another alternative to overcome this problem is the use of a semantic similarity index of a content published by different sources. This index can be measured by ML methods (e.g., word2vec) and it can be used to assess the integrity by checking it on the DLT (i.e., whether it was published or not by a verified entity).

The DLT system alone is not able to fully evaluate the authenticity of an input content. Consequently, it is essential to develop a system that is resilient to data falsification attacks, which inserts forged data into the DLT. To face this issue, it is recommended to include contextual knowledge to corroborate the integrity of the news. Further research may include the use of DLT together with AI and NLP methods to develop deep insights about similarities and to quantify trustworthiness.

Strengthening cybersecurity and preserving privacy and security of content shared on social media is also a key issue, since they may be used to train an ML model to create fake content. DLT-based solutions can cryptographically store the content in such a way that every transaction and interaction with it is traceable.

There are still open issues related to the DLT compliance with the GDPR, especially when dealing with the role of the controller, the feasibility of data anonymization and the ease of subject rights. Compliance with current legislative directives implies the cooperation of a wide range of global stakeholders.

Future platforms will have to ensure safety and transparency providing a trade-off between content moderation (e.g., freedom of expression, right to receive information) and personal data protection.

4.3 MODULES OVERVIEW

Registrations, User Authorization are done by Remote user who acts as admin for the publishing news to the users. News publish server can also be managed by Remote user where all the news publishers register with news channel names to provide news to the users. News publisher server can add fake filters, can select news category and add news, list all news posts and give option to update and delete, etc. List all news post by Distributed Ledger Technologies (show new news as based setting date, show other all old news down) and listing all news posts by blocks based on news cat and rank and reviews.

4.4 DEFINE THE MODULES

A module is a collection of source files and build settings that allow you to divide your project into discrete units of functionality. Your project can have one or many modules and one module may use another module as a dependency. Each module can be independently built, tested, and debugged. Additional modules are often useful when creating code libraries within your own project or when you want to create different sets of code and resources for different device types, such as phones and wearables, but keep all the files scoped within the same project and share some code.

4.5 MOUDULE FUNCTIONALITY

4.5.1 NEWS PUBLISHER SERVER

In this module, the admin has to login by using valid username and password. After login successful he can perform some operations such as List all users and authorize, Register with

News channel name and login, Add News Categories, Set news quantization date, Select category and add news, List all news post and give option to update and delete, List all news post by Distributed Ledger Technologies, List All News Posts by blocks based on news cat, List All Users News transactions by keyword, View online product Distributed Ledger Technologies by chart, View all news post rank in chart.

4.5.2 USER

In this module, there are n numbers of users are present. User should register before performing any operations. Once user registers, their details will be stored to the database. After registration successful, he has to login by using authorized username and password. Once Login is successful user can perform some operations like View your profile, Search news by content keyword, select hash code to show all news titles, show all your search transactions based on keyword and view all fake news.

CHAPTER 5
PROJECT SYSTEM DESIGN

5 PROJECT SYSTEM DESIGN

This section provides the information needed for a system development team to actually build and integrate the hardware components, code, and integrate the software modules, and interconnect the hardware and software segments into a functional product. Additionally, this section addresses the detailed procedures for combining separate COTS packages into a single system. Every detailed requirement should map back to the FRD, and the mapping should be presented in an update to the RTM and include the RTM as an appendix to this design document.

5.1 Hardware Detailed Design

A hardware component is the lowest level of design granularity in the system. Depending on the design requirements, there may be one or more components per system. This section should provide enough detailed information about individual component requirements to correctly build and/or procure all the hardware for the system (or integrate COTS items).

If there are many components or if the component documentation is extensive, place it in an appendix or reference a separate document. Add additional diagrams and information, if necessary, to describe each component and its functions, adequately. Industry-standard component specification practices should be followed. For COTS procurements, if a specific vendor has been identified, include appropriate item names. Include the following information in the detailed component designs (as applicable):

- Power input requirements for each component.
- Signal impedances and logic states.
- Connector specifications (serial/parallel, 11-pin, male/female, etc.)
- Memory and/or storage space requirements.
- Processor requirements (speed and functionality).
- Graphical representation depicting the number of hardware items (for example, monitors, printers, servers, I/O devices), and the relative positioning of the components to each other.
- Cable type(s) and length(s).
- User interfaces (buttons, toggle switches, etc.)
- Hard drive/floppy drive/CD-ROM requirements.
- Monitor resolution.

5.2 Software Detailed Design

A software module is the lowest level of design granularity in the system. Depending on the software development approach, there may be one or more modules per system. This section should provide enough detailed information about logic and data necessary to completely write source code for all modules in the system (and/or integrate COTS software programs).

If there are many modules or if the module documentation is extensive, place it in an appendix or reference a separate document. Add additional diagrams and information, if necessary, to describe each module, its functionality, and its hierarchy. Industry-standard module specification practices should be followed. Include the following information in the detailed module designs:

- A narrative description of each module, its function(s), the conditions under which it is used (called or scheduled for execution), its overall processing, logic, interfaces to other modules, interfaces to external systems, security requirements, etc.; explain any algorithms used by the module in detail.
- For COTS packages, specify any call routines or bridging programs to integrate the package with the system and/or other COTS packages (for example, Dynamic Link Libraries).
- Data elements, record structures, and file structures associated with module input and output.
- Graphical representation of the module processing, logic, flow of control, and algorithms, using an accepted diagramming approach (for example, structure charts, action diagrams, flowcharts, etc.).
- Data entry and data output graphics; define or reference associated data elements; if the project is large and complex or if the detailed module designs will be incorporated into a separate document, then it may be appropriate to repeat the screen information in this section.
- Report layout.

5.3 Internal Communications Detailed Design

If the system includes more than one component there may be a requirement for internal communications to exchange information, provide commands, or support input/output functions. This section should provide enough detailed information about the communication

requirements to correctly build and/or procure the communications components for the system. Include the following information in the detailed designs (as appropriate):

- The number of servers and clients to be included on each area network.
- Specifications for bus timing requirements and bus control.
- Format(s) for data being exchanged between components.
- Graphical representation of the connectivity between components, showing the direction of data flow (if applicable), and approximate distances between components; information should provide enough detail to support the procurement of hardware to complete the installation at a given location.
- LAN topology

5.4 Architecture

Systems design is the process of defining elements of a system like modules, architecture, components and their interfaces and data for a system based on the specified requirements. It is the process of defining, developing, and designing systems which satisfies the specific needs and requirements of a business or organization.

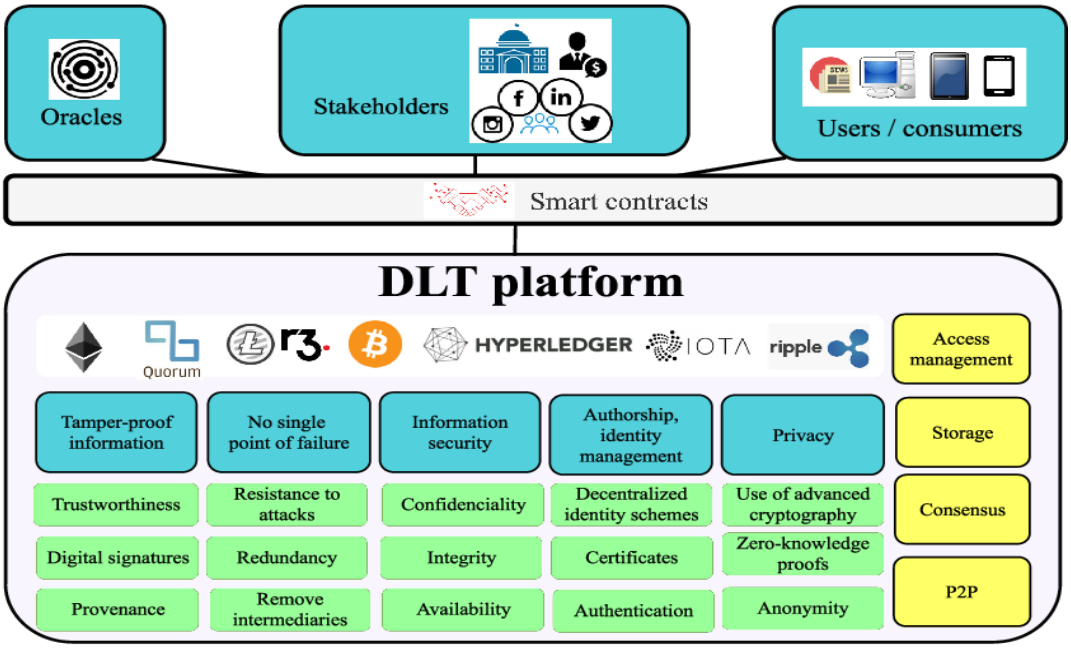


Fig 5.1 DLT and blockchain key capabilities to combat digital deception.

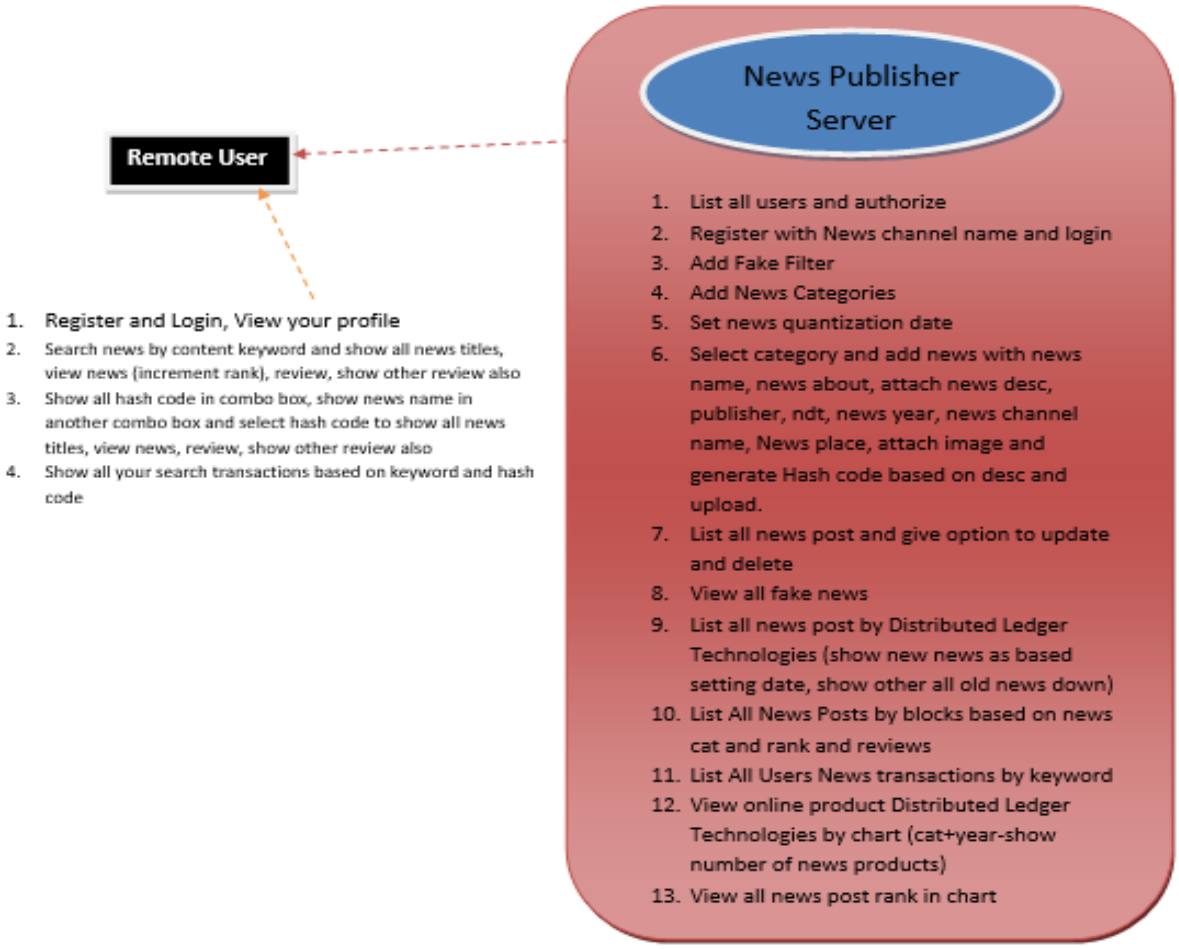


Fig 5.2 Architectural diagram.

In this section, describe the system and/or subsystem(s) architecture for the project. References to external entities should be minimal, as they will be described in detail in Section 6, External Interfaces.

5.4.1 System Hardware Architecture

In this section, describe the overall system hardware and organization. Include a list of hardware components (with a brief description of each item) and diagrams showing the connectivity between the components. If appropriate, use subsections to address each subsystem.

5.4.2 System Software Architecture

In this section, describe the overall system software and organization. Include a list of software modules (this could include functions, subroutines, or classes), computer languages, and programming computer-aided software engineering tools (with a brief description of the

function of each item). Use structured organization diagrams/object-oriented diagrams that show the various segmentation levels down to the lowest level. All features on the diagrams should have reference numbers and names. Include a narrative that expands on and enhances the understanding of the functional breakdown. If appropriate, use subsections to address each module.

5.4.3 Internal Communications Architecture

In this section, describe the overall communications within the system, for example, LANs, buses, etc. Include the communications architecture(s) being implemented, such as X.25, Token Ring, etc. Provide a diagram depicting the communications path(s) between the system and subsystem modules. If appropriate, use subsections to address each architecture being employed.

5.5 DATA FLOW DIAGRAM

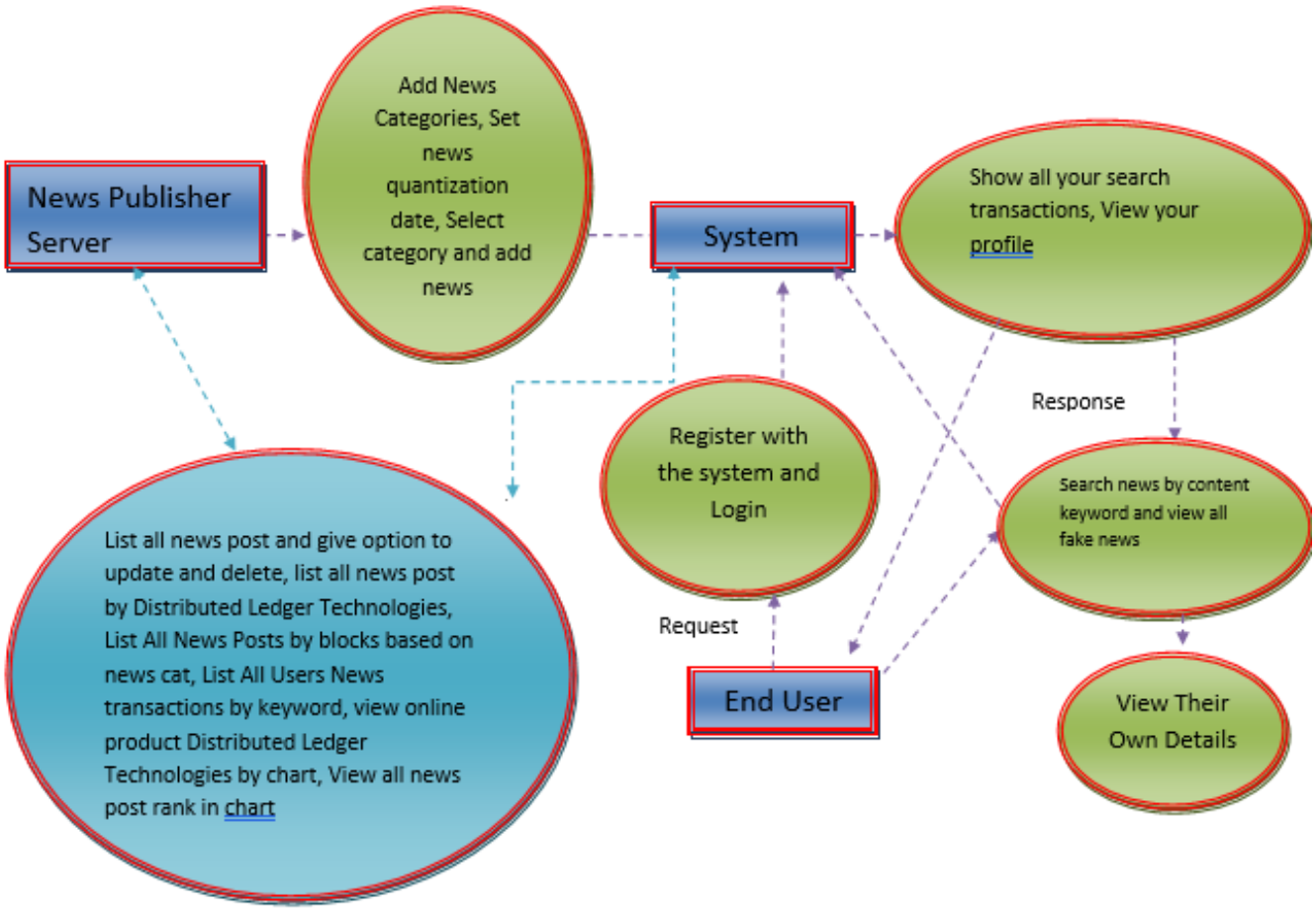


Fig 5.3 Data flow diagram of the system

A data-flow diagram is a way of representing a flow of data through a process or a system (usually

an information system). The DFD also provides information about the outputs and inputs of each entity and the process itself. A data-flow diagram has no control flow, there are no decision rules and no loops. Specific operations based on the data can be represented by a flowchart.

There are several notations for displaying data-flow diagrams. The notation presented above was described in 1979 by Tom DeMarco as part of structured analysis.

For each data flow, at least one of the endpoints (source and / or destination) must exist in a process. The refined representation of a process can be done in another data-flow diagram, which subdivides this process into sub-processes.

The data-flow diagram is part of the structured-analysis modelling tools. When using UML, the activity diagram typically takes over the role of the data-flow diagram. A special form of data-flow plan is a site-oriented data-flow plan.

Data-flow diagrams can be regarded as inverted Petri nets, because places in such networks correspond to the semantics of data memories. Analogously, the semantics of transitions from Petri nets and data flows and functions from data-flow diagrams should be considered equivalent.

5.6 E-R DIAGRAM

An E-R model is usually the result of systematic analysis to define and describe what is important to process in an area of a business. It does not define the business processes; it only presents a business data schema in graphical form. It is usually drawn in a graphical form as boxes (entities) that are connected by lines (relationships) which express the associations and dependencies between entities. An ER model can also be expressed in a verbal form, for example: one building may be divided into zero or more apartments, but one apartment can only be located in one building.

Entities may be characterized not only by relationships, but also by additional properties (attributes), which include identifiers called "primary keys". Diagrams created to represent attributes as well as entities and relationships may be called entity-attribute-relationship diagrams, rather than entity–relationship models.

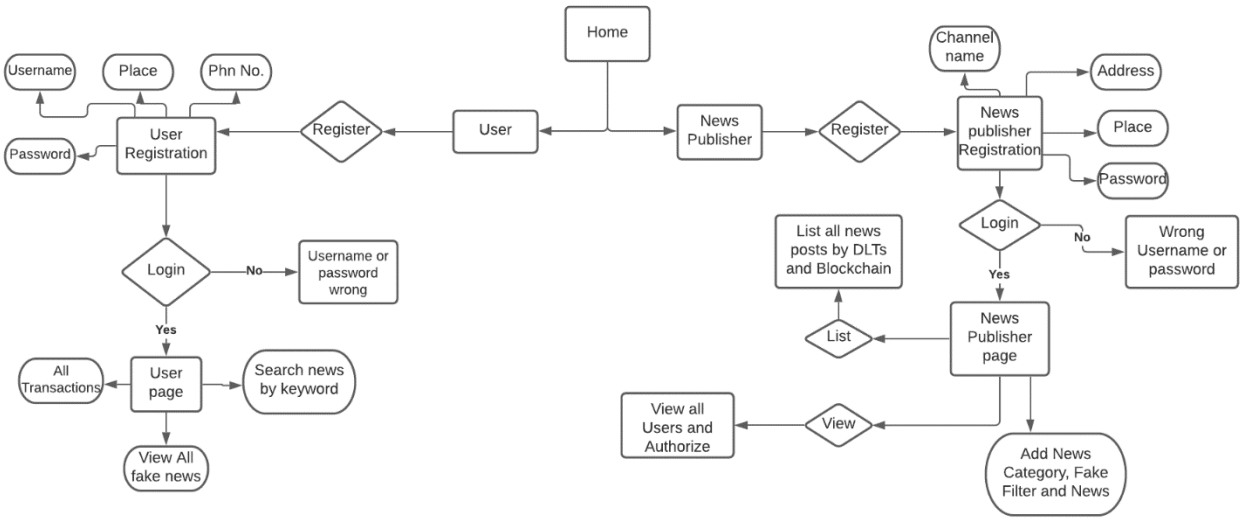


Fig 5.4 ER Diagram of User and News Publisher

An ER model is typically implemented as a database. In a simple relational database implementation, each row of a table represents one instance of an entity type, and each field in a table represents an attribute type. In a relational database a relationship between entities is implemented by storing the primary key of one entity as a pointer or "foreign key" in the table of another entity.

There is a tradition for ER/data models to be built at two or three levels of abstraction. Note that the conceptual-logical-physical hierarchy below is used in other kinds of specification, and is different from the three schema approach to software engineering.

5.7 UML DIAGRAMS

UML is a modern approach to modelling and documenting software. In fact, it’s one of the most popular business process modelling techniques.

It is based on diagrammatic representations of software components. As the old proverb says: “a picture is worth a thousand words”. By using visual representations, we are able to better understand possible flaws or errors in software or business processes.

Mainly, UML has been used as a general-purpose modelling language in the field of software engineering. However, it has now found its way into the documentation of several business processes or workflows. For example, activity diagrams, a type of UML diagram, can be used as a replacement for flowcharts. They provide both a more standardized way of modelling workflows as well as a wider range of features to improve readability and efficacy.

5.7.1 Use Case Diagram

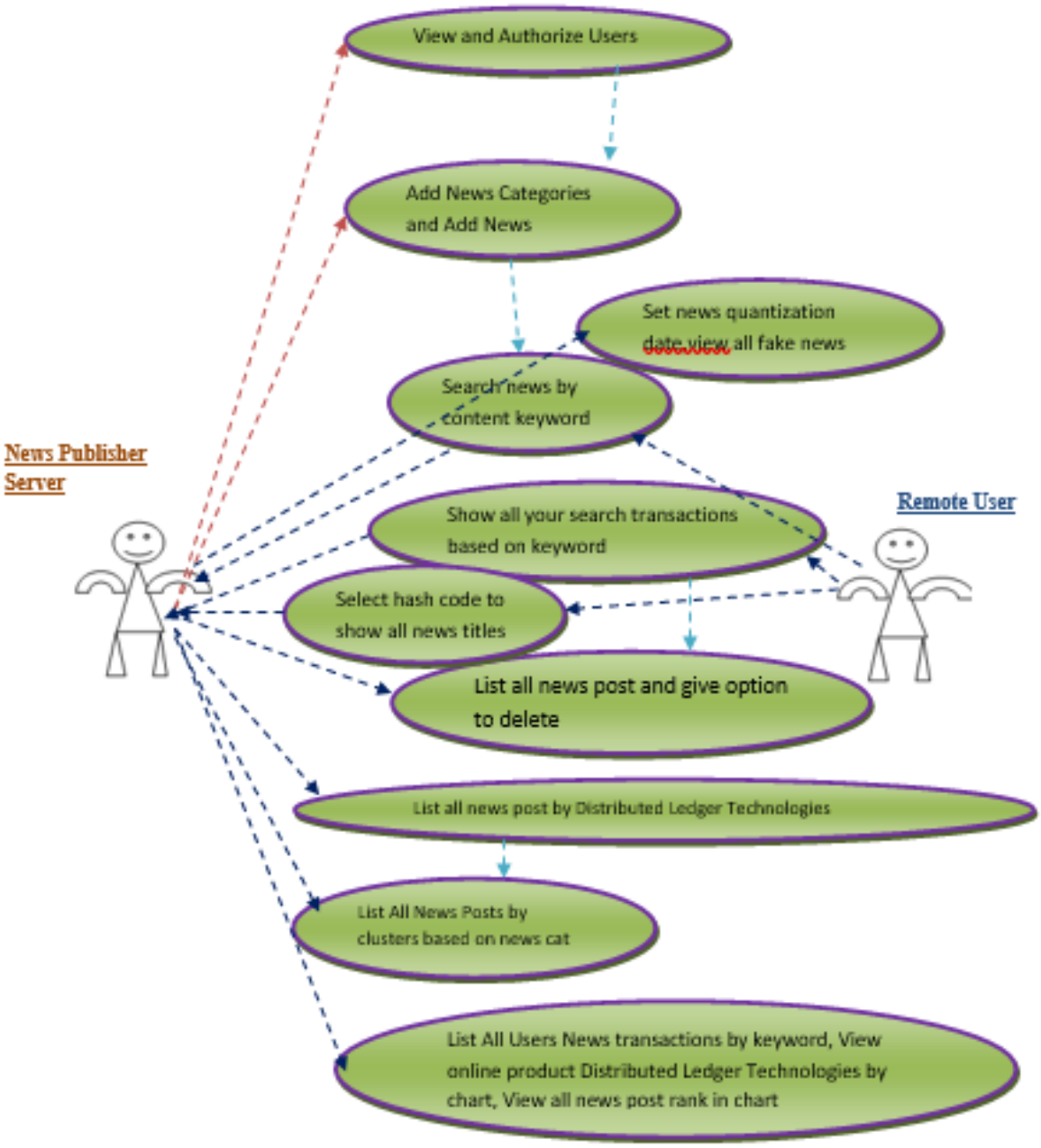


Fig 5.5 Use Case Diagram

While a use case itself might drill into a lot of detail about every possibility, a use-case diagram can help provide a higher-level view of the system. It has been said before that "Use case diagrams are the blueprints for your system".

Due to their simplistic nature, use case diagrams can be a good communication tool for stakeholders. The drawings attempt to mimic the real world and provide a view for the stakeholder to understand how the system is going to be designed. Siau and Lee conducted research to determine if there was a valid situation for use case diagrams at all or if they were unnecessary. What was found was that the use case diagrams conveyed the intent of the system in a more simplified manner to stakeholders and that they were "interpreted more completely than class diagrams".

The purpose of a use case diagram is to capture the dynamic aspect of a system. They provide a simplified graphical representation of what the system should do in a use case. Further diagrams and documentation are needed for a complete functional and technical outlook on the system.

5.7.2 Sequence Diagram

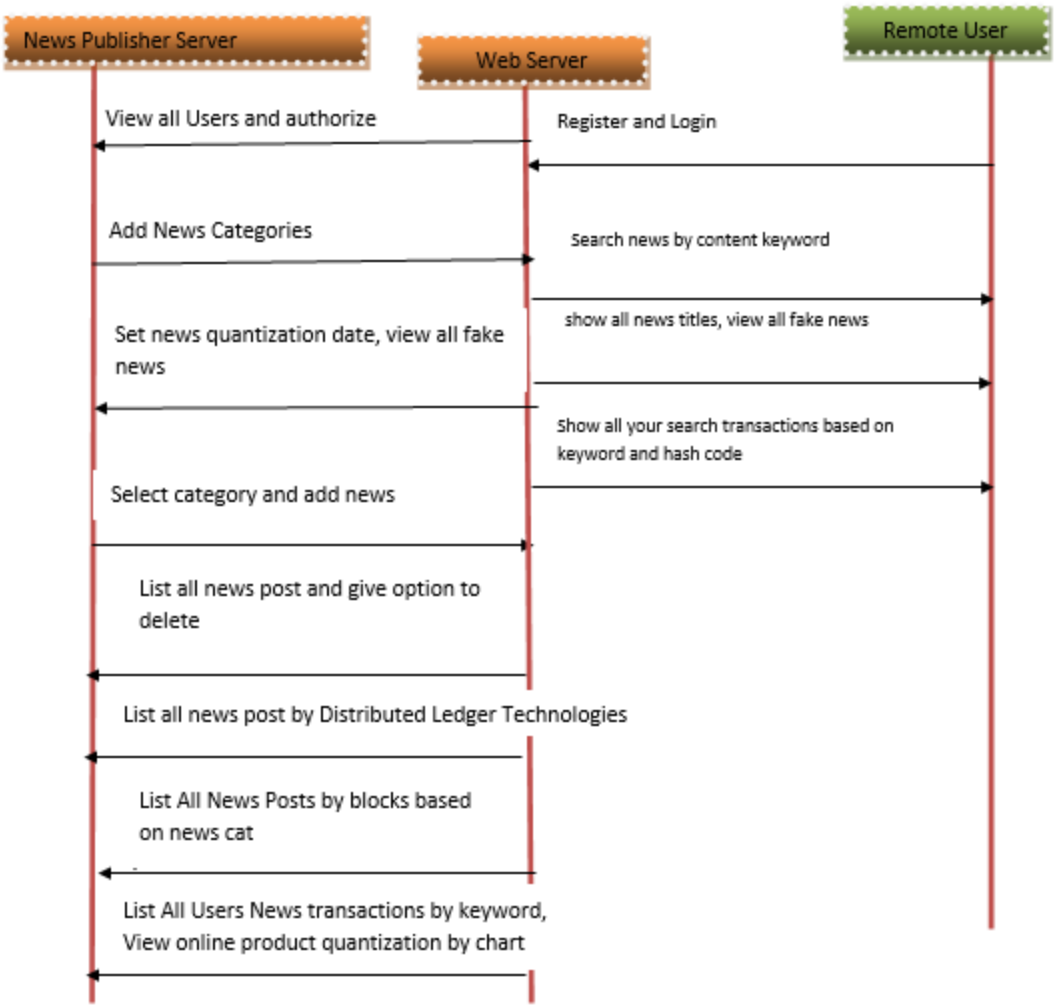


Fig 5.6 Sequence Diagram

5.7.3 Activity diagram

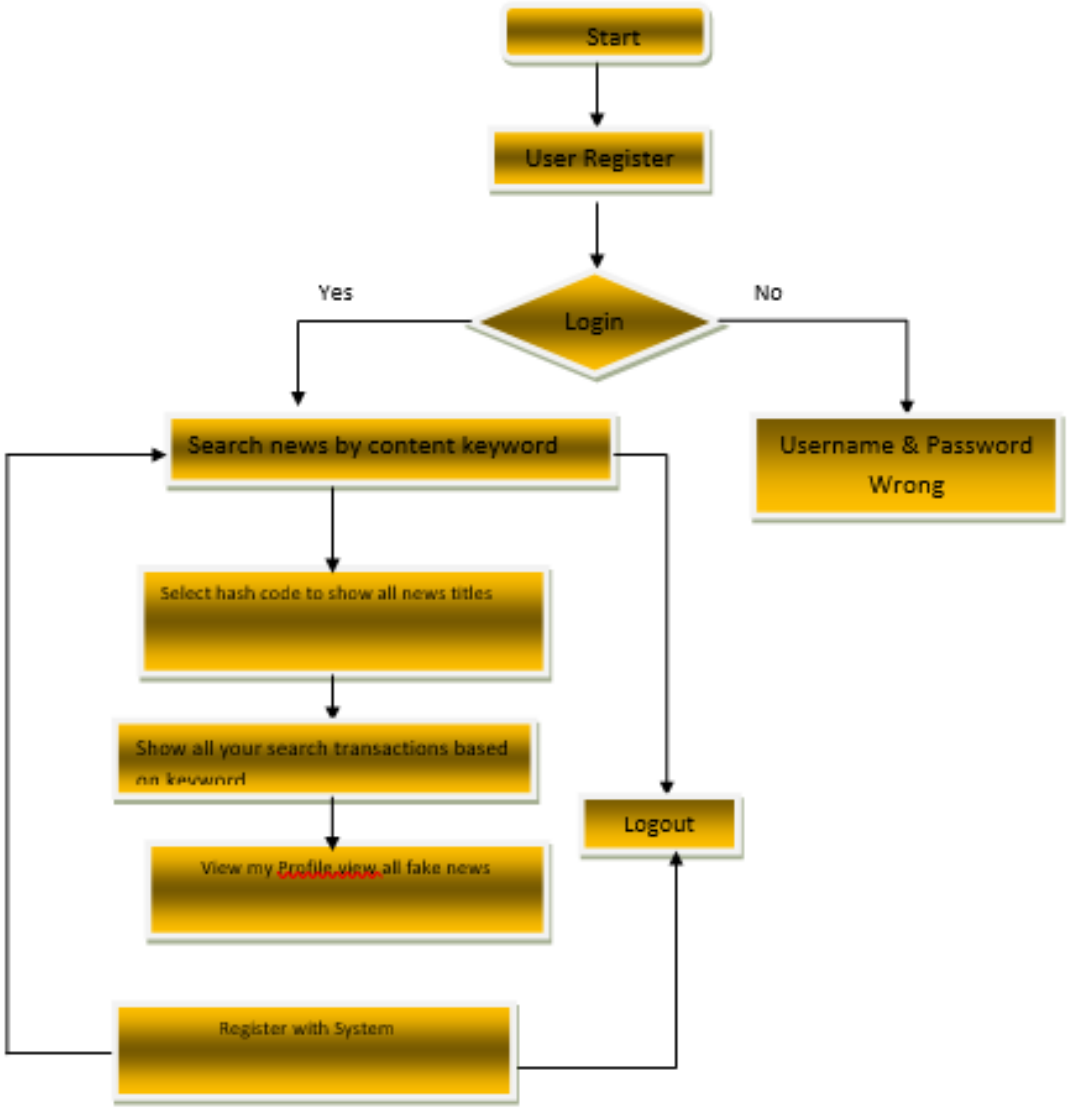


Fig 5.7.1 Activity Diagram of User

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modelling Language, activity diagrams are intended to model both computational and organizational processes (i.e., workflows), as well as the data flows intersecting with the related activities. Although activity diagrams primarily show the overall flow of control, they can also include elements showing the flow of data between activities through one or more data stores.

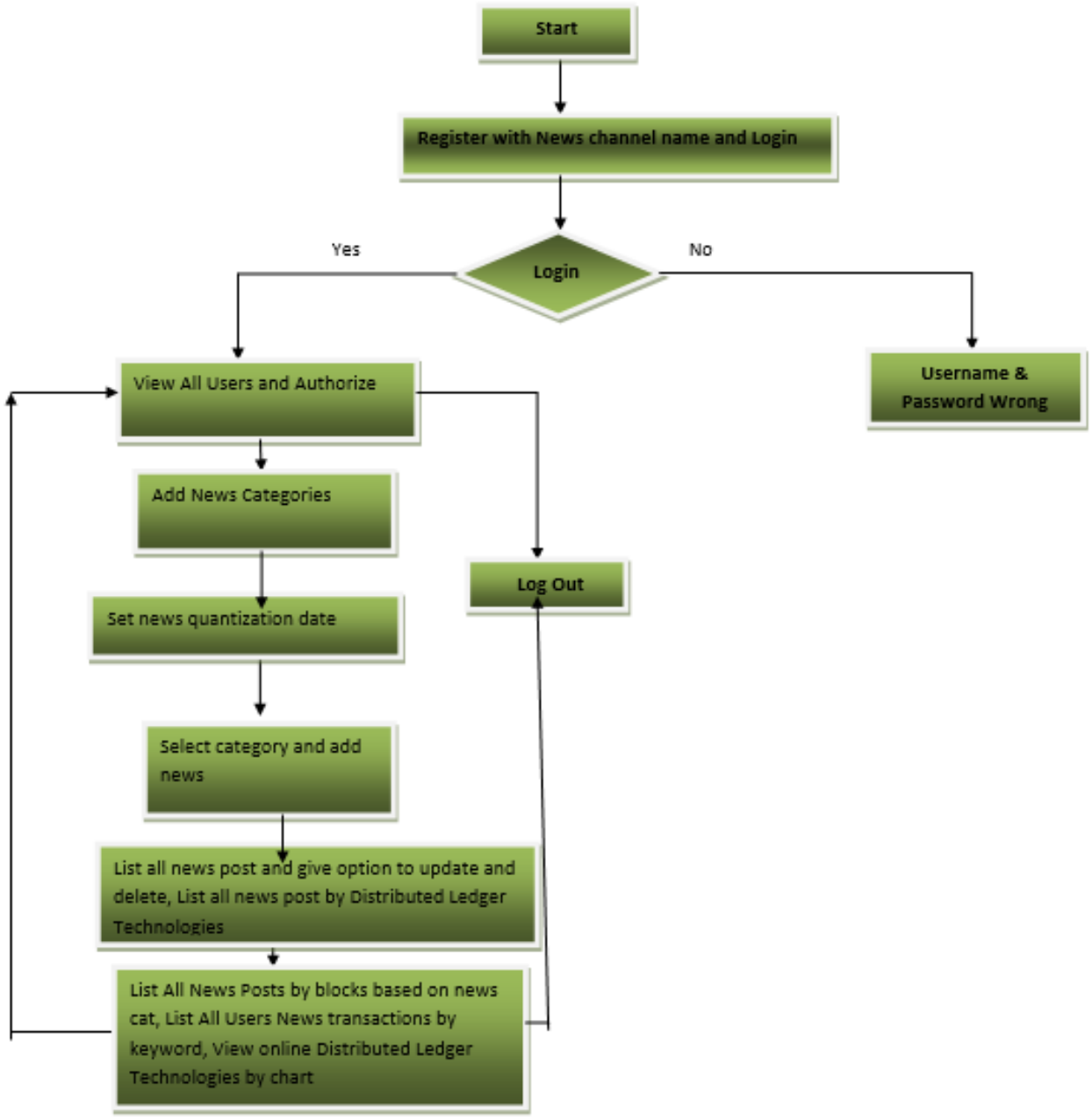


Fig 5.7.2 Activity Diagram of News Publisher

Activity Diagrams describe how activities are coordinated to provide a service which can be at different levels of abstraction. Typically, an event needs to be achieved by some operations, particularly where the operation is intended to achieve a number of different things that require coordination, or how the events in a single use case relate to one another, in particular, use cases where activities may overlap and require coordination. It is also suitable for modelling how a collection of use cases coordinates to represent business workflows.

1. Identify candidate use cases, through the examination of business workflows.
2. Identify pre- and post-conditions (the context) for use cases.
3. Model workflows between/within use cases.

4. Model complex workflows in operations on objects.
5. Model in detail complex activities in a high-level activity Diagram.

5.7.4 Object Diagram

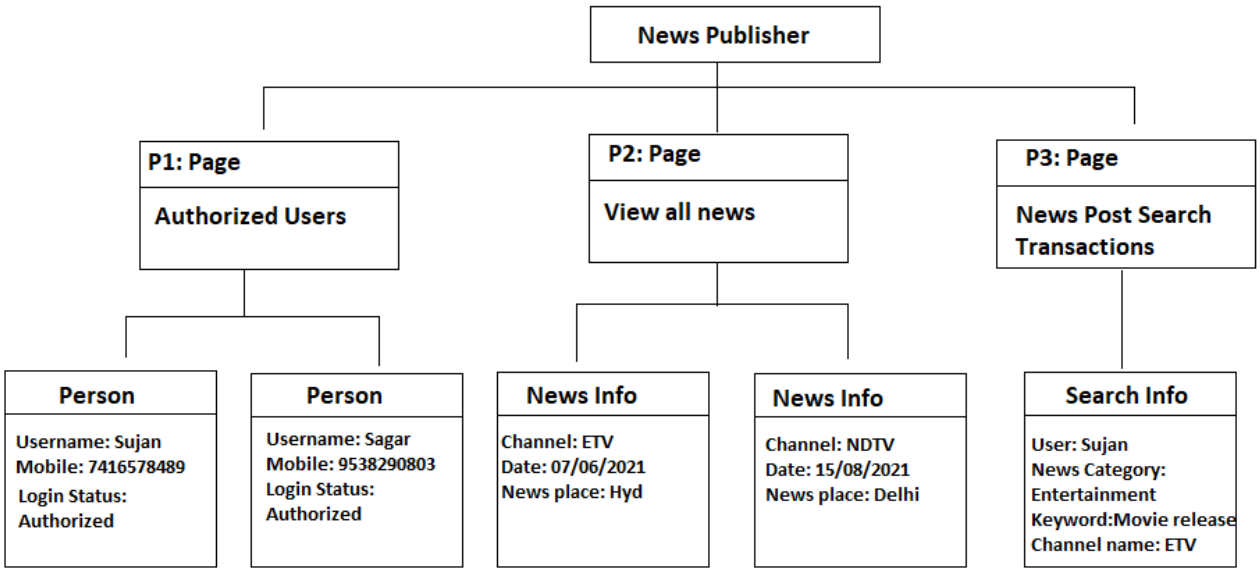


Fig 5.8 Object Diagram

In the Unified Modelling Language (UML), an object diagram focuses on some particular set of objects and attributes, and the links between these instances. A correlated set of object diagrams provides insight into how an arbitrary view of a system is expected to evolve over time. In early UML specifications the object diagram is described as:

"An object diagram is a graph of instances, including objects and data values. A static object diagram is an instance of a class diagram; it shows a snapshot of the detailed state of a system at a point in time. The use of object diagrams is fairly limited, namely, to show examples of data structure."

The latest UML 2.5 specification does not explicitly define object diagrams but provides a notation for instances of classifiers.

Object diagrams and class diagrams are closely related and use almost identical notation. Both diagrams are meant to visualize static structure of a system. While class diagrams show classes, object diagrams display instances of classes (objects). Object diagrams are more concrete than class diagrams. They are often used to provide examples or act as test cases for class diagrams. Only aspects of current interest in a model are typically shown on an object diagram.

5.7.5 State chart Diagram

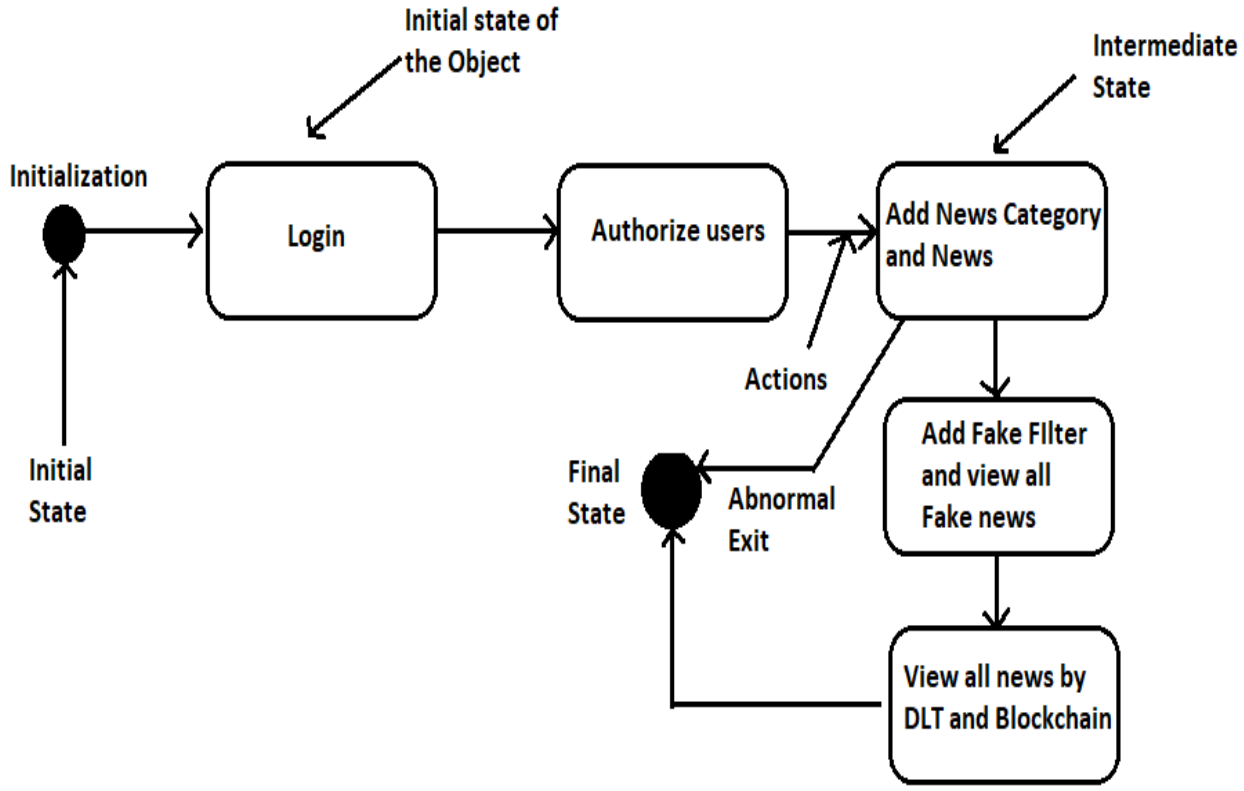


Fig 5.9 State chart Diagram

A state diagram is a type of diagram used in computer science and related fields to describe the behaviour of systems. State diagrams require that the system described is composed of a finite number of states; sometimes, this is indeed the case, while at other times this is a reasonable abstraction. Many forms of state diagrams exist, which differ slightly and have different semantics.

State diagrams are used to give an abstract description of the behaviour of a system. This behaviour is analysed and represented by a series of events that can occur in one or more possible states. Hereby "each diagram usually represents objects of a single class and track the different states of its objects through the system".

State diagrams can be used to graphically represent finite-state machines (also called finite automata). This was introduced by Claude Shannon and Warren Weaver in their 1949 book The Mathematical Theory of Communication. Another source is Taylor Booth in his 1967 book Sequential Machines and Automata Theory. Another possible representation is the state-transition table.

5.7.6 Class Diagram

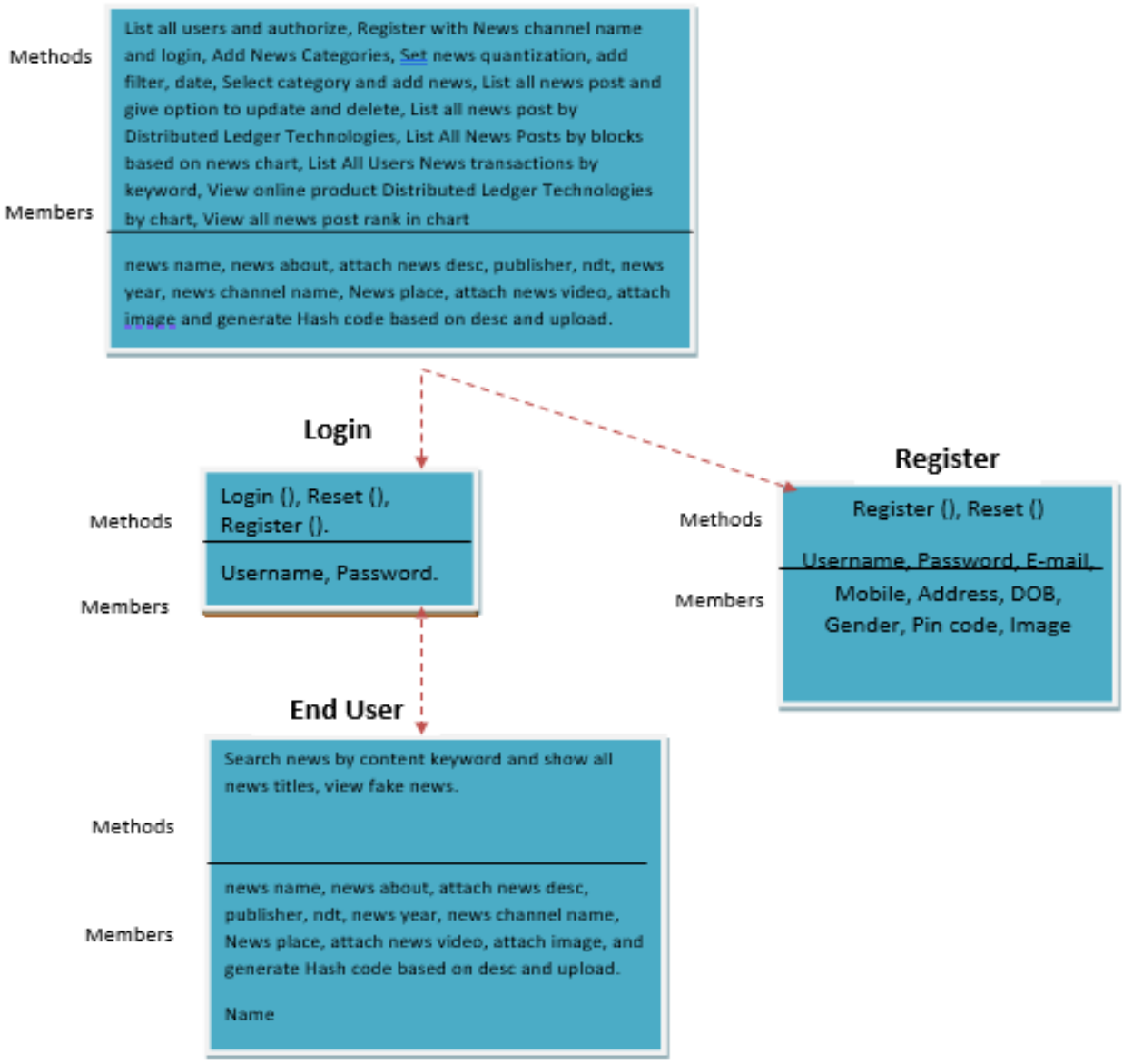


Fig 5.10 Class Diagram

In software engineering, a class diagram in the Unified Modelling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among objects.

The class diagram is the main building block of object-oriented modelling. It is used for general conceptual modelling of the structure of the application, and for detailed modelling, translating the models into programming code. Class diagrams can also be used for data modelling. The classes in a class diagram represent both the main elements, interactions in the application, and the classes to be programmed.

In the diagram, classes are represented with boxes that contain three compartments:

- The top compartment contains the name of the class. It is printed in bold and cantered, and the first letter is capitalized.
- The middle compartment contains the attributes of the class. They are left-aligned, and the first letter is lowercase.
- The bottom compartment contains the operations the class can execute. They are also left-aligned, and the first letter is lowercase.

A class with three compartments.

In the design of a system, a number of classes are identified and grouped together in a class diagram that helps to determine the static relations between them. In detailed modelling, the classes of the conceptual design are often split into subclasses.

In order to further describe the behaviour of systems, these class diagrams can be complemented by a state diagram or UML state machine.

5.7.7 Deployment Diagram

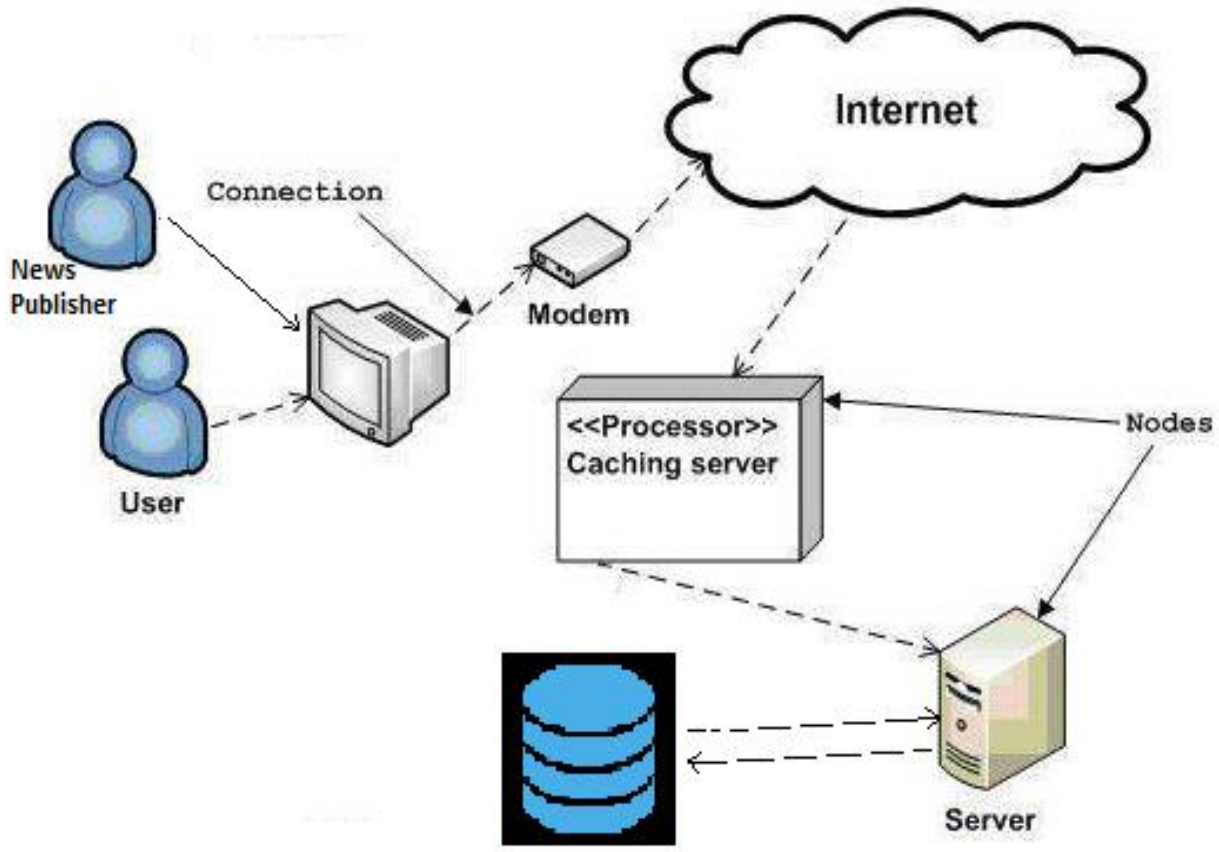


Fig 5.11 Deployment Diagram

A deployment diagram in the Unified Modelling Language models the physical deployment of artifacts on nodes.[1] To describe a web site, for example, a deployment diagram would show what hardware components ("nodes") exist (e.g., a web server, an application server, and a database server), what software components ("artifacts") run on each node (e.g., web application, database), and how the different pieces are connected (e.g. JDBC, REST, RMI).

The nodes appear as boxes, and the artifacts allocated to each node appear as rectangles within the boxes. Nodes may have sub nodes, which appear as nested boxes. A single node in a deployment diagram may conceptually represent multiple physical nodes, such as a cluster of database servers.

There are two types of Nodes:

1. Device Node
2. Execution Environment Node

Device nodes are physical computing resources with processing memory and services to execute software, such as typical computers or mobile phones. An execution environment node (EEN) is a software computing resource that runs within an outer node and which itself provides a service to host and execute other executable software elements.

5.7.8 Collaboration Diagram

The collaboration diagram is used to show the relationship between the objects in a system. Both the sequence and the collaboration diagrams represent the same information but differently. Instead of showing the flow of messages, it depicts the architecture of the object residing in the system as it is based on object-oriented programming. An object consists of several features. Multiple objects present in the system are connected to each other. The collaboration diagram, which is also known as a communication diagram, is used to portray the object's architecture in the system.

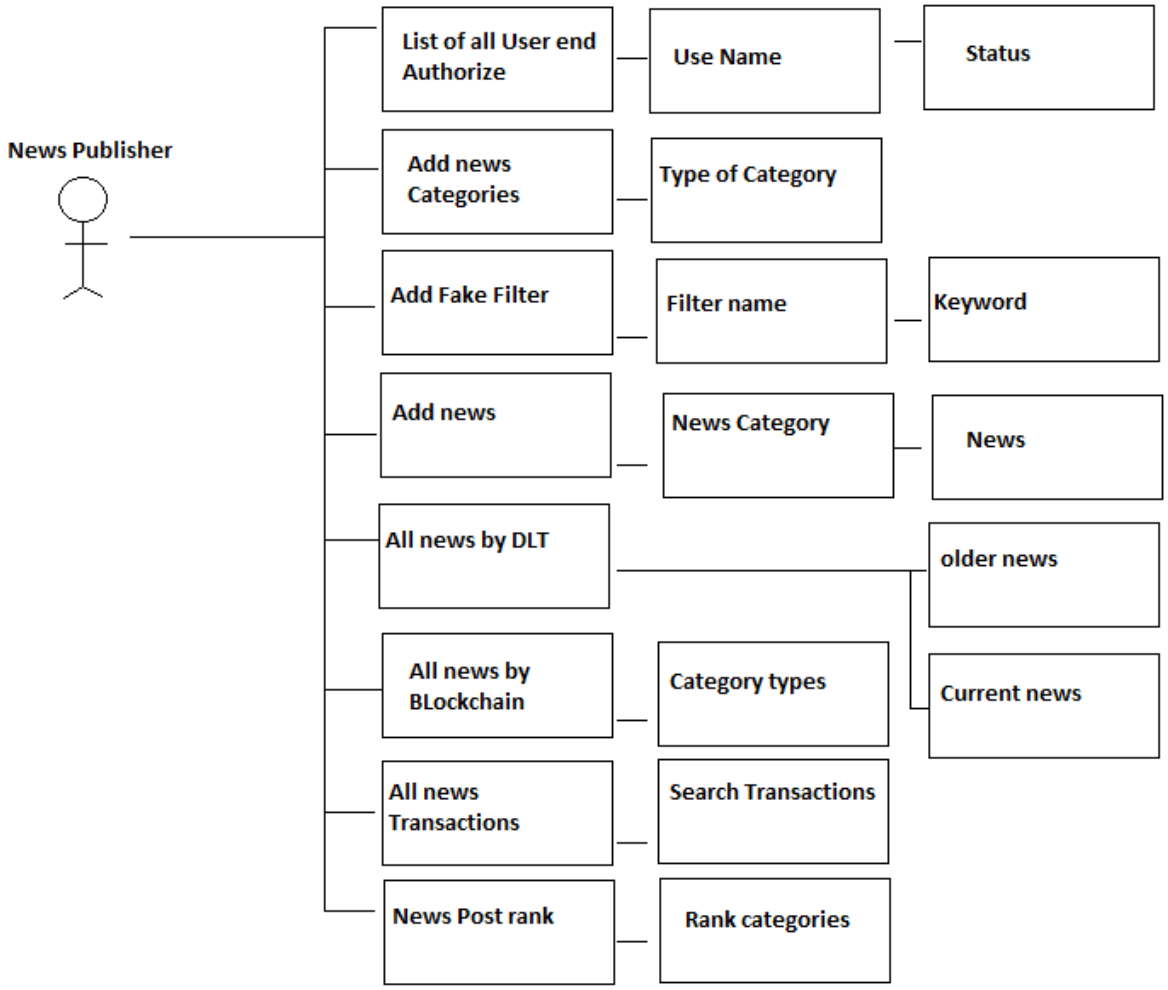


Fig 5.12 Collaboration Diagram

Unlike a sequence diagram, a collaboration diagram shows the relationships among the objects. Sequence diagrams and collaboration diagrams express similar information but show it in different ways.

Because of the format of the collaboration diagram, they tend to better suit for analysis activities (see Activity: Use-Case Analysis). Specifically, they tend to be better suited to depicting simpler interactions of smaller numbers of objects. However, if the number of objects and messages grows, the diagram becomes increasingly hard to read. In addition, it is difficult to show additional descriptive information such as timing, decision points, or other unstructured information that can be easily added to the notes in a sequence diagram.

5.7.9 Component Diagram

In Unified Modelling Language (UML), a component diagram depicts how components are wired together to form larger components or software systems. They are used to illustrate the structure of arbitrarily complex systems. A component diagram allows verification that a system's required functionality is acceptable.

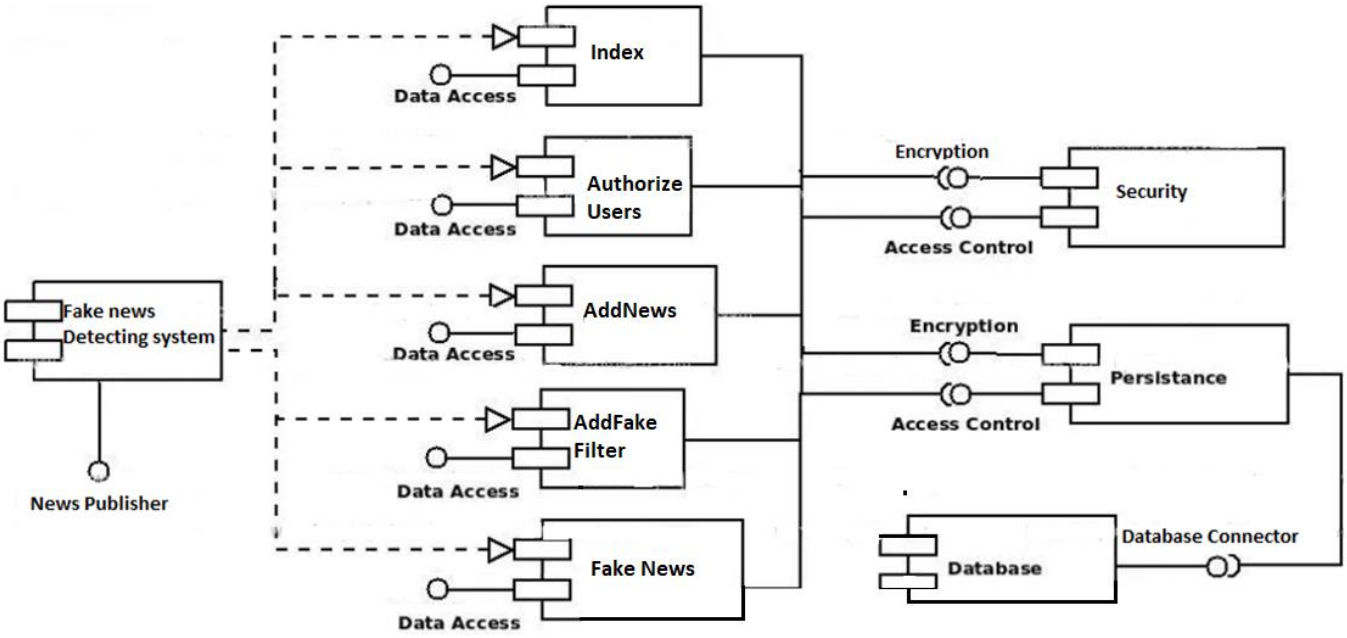


Fig 5.13 Component Diagram

These diagrams are also used as a communication tool between the developer and stakeholders of the system. Programmers and developers use the diagrams to formalize a roadmap for the implementation, allowing for better decision-making about task assignment or needed skill improvements. System administrators can use component diagrams to plan ahead, using the view of the logical software components and their relationships on the system.

The component diagram extends the information given in a component notation element. One way of illustrating the provided and required interfaces by the specified component is in the form of a rectangular compartment attached to the component element. Another accepted way of presenting the interfaces is to use the ball-and-socket graphic convention. A provided dependency from a component to an interface is illustrated with a solid line to the component using the interface from a "lollipop", or ball, labelled with the name of the interface. A required usage dependency from a component to an interface is illustrated by a half-circle, or socket, labelled with the name of the interface, attached by a solid line to the component that requires this interface. Inherited interfaces may be shown with a lollipop, preceding the name label with a caret symbol. To illustrate dependencies between the two, use a solid line with a plain arrowhead joining the socket to the lollipop.

CHAPTER 6
PROJECT CODING

6 PROJECT CODING

6.1 CODE TEMPLATES

AddFilter.jsp

```
</head>

<body>

<div class="main">

  <div class="header">

    <div class="header_resize">

      <div class="logo">

        <h1><a href="index.html" class="style1">Fake News, Disinformation,and Deepfakes: Leveraging Distributed Ledger Technologies and Blockchain to Combat Digital Deception and Counterfeit Reality

</a></h1>

      </div>

      <div class="menu_nav">

        <ul>

          <li class="active"><a href="index.html"><span>Home Page</span></a></li>

          <li><a href="ServerLogin.jsp"><span>News Publisher </span></a></li>

          <li><a href="UserLogin.jsp"><span>User </span></a></li>

        </ul>

      </div>

      <div class="clr"></div>

      <div class="slider">

        <div id="coin-slider"> <a href="#"><img src="images/slide1.jpg" width="935" height="293" alt="" /> </a> <a href="#"><img src="images/slide2.jpg" width="935" height="293" alt="" /> </a> <a href="#"><img src="images/slide3.jpg" width="935" height="293" alt="" /> </a> </div>
```

```
      <div class="clr"></div>
    </div>
    <div class="clr"></div>
  </div>
 </div>
 <div class="content">
  <div class="content_resize">
   <div class="mainbar">
    <div class="article">
     <h2 class="style2" style="color:#FF0033">Add Fake Filter..</h2>
     <div class="clr"></div>
          <div class="post_content">
          <form id="form1" name="form1" method="post" action="S_AddFilter1.jsp">
       <p> </p>
       <table width="385" border="2">
        <tr>
          <td width="181" height="47" bgcolor="#FF0000"><span class="style11
style10">Select Filter Category </span></td>
          <td width="186"><select name="tclass">
           <option>----Select Filter---</option>
           <option>Fake</option>
          </select>              </td>
        </tr>
        <tr>
          <td height="52" bgcolor="#FF0000"><span class="style11 style10">Enter Filter
Name </span></td>
```

```
      <td><input type="text" name="fname" /></td>
    </tr>
    <tr>
      <td height="52"> </td>
      <td><p>
        <input type="submit" name="Submit" value="Add" />
        <input type="reset" name="Submit2" value="Reset" />
      </p></td>
    </tr>
  </table>
  <p> </p>
  <p><a href="AdminMain.jsp"></a></p>
  <p class="style12 style1">Filter Details </p>
 </form>

    <%@ include file="connect.jsp" %>
<table width="329" border="1">
                <tr>
                  <td width="195" height="42" bgcolor="#FF0000"><div
align="center" class="style10"><span class="style3 style14"><b>Fake Category
</b></span></div></td>
                  <td width="439" bgcolor="#FF0000"><div align="center"
class="style14 style10"><b>Filter Name</b></div></td>
    </tr>
  <%    String s0="",s1="",s2="",s3="",s4="",s5="",s6="";
        int i=1,j=0,count=0,rank=0,k=0;
```

```
                                        try
                                        {
                                                String query="select * from filter ";
                                                Statement st=connection.createStatement();
                                                ResultSet rs=st.executeQuery(query);
                                                while ( rs.next() )
                                                {
                                                        s0=rs.getString(1);
                                                        s1=rs.getString(2);
%>
      <tr>
        <td height="33"  valign="middle" bgcolor="#FFFFFF">
          <div align="center" class="style4 style12 style14 style8" >
            <div align="center">
            <%out.println(s0);%>
          </div>
        </div></td>
                <td height="33"  valign="middle" bgcolor="#FFFFFF">
          <div align="center" class="style4 style12 style14 style9" >
            <div align="center">
            <%out.println(s1);%>
          </div>
        </div></td>
      </tr>
       <%
                                        }
```

```
                    connection.close();
                }
                catch(Exception e)
                {
                    out.println(e.getMessage());
                }
%>
      </table>
                    <p><a href="Server_Main.jsp">Back</a></p>

          </div>
       <div class="clr"></div>
      </div>
      <p class="pages"> </p>
    </div>
    <div class="sidebar">
      <div class="searchform">
        <form id="formsearch" name="formsearch" method="post" action="#">
          <span>
          <input name="editbox_search" class="editbox_search" id="editbox_search" maxlength="80" value="Search our ste:" type="text" />
          </span>
          <input name="button_search" src="images/search.gif" class="button_search" type="image" />
        </form>
      </div>
```

```
      <div class="clr"></div>
      <div class="gadget">
        <h2 class="star"><span>Server</span> Menu</h2>
        <div class="clr"></div>
        <ul class="sb_menu">
          <li><a href="Server_Main.jsp">&raquo; Home </a></li>
                    <li><a href="ServerLogin.jsp">&raquo; Logout </a></li>
        </ul>
      </div>
    </div>
    <div class="clr"></div>
  </div>
 </div>
 <div class="footer">
  <div class="footer_resize">
   <div style="clear:both;"></div>
  </div>
 </div>
</div>
<div align=center></div>
</body>
</html>
```

AddNews.jsp

```
div class="main">
 <div class="header">
  <div class="header_resize">
```

```
    <div class="logo">
      <h1><a href="index.html" class="style1">Fake News, Disinformation,and Deepfakes: Leveraging Distributed Ledger Technologies and Blockchain to Combat Digital Deception and Counterfeit Reality
</a></h1>
    </div>
    <div class="menu_nav">
      <ul>
        <li class="active"><a href="index.html"><span>Home Page</span></a></li>
        <li><a href="ServerLogin.jsp"><span>News Publisher </span></a></li>
        <li><a href="UserLogin.jsp"><span>User </span></a></li>
      </ul>
    </div>
    <div class="clr"></div>
    <div class="slider">
      <div id="coin-slider"> <a href="#"><img src="images/slide1.jpg" width="935" height="293" alt="" /> </a> <a href="#"><img src="images/slide2.jpg" width="935" height="293" alt="" /> </a> <a href="#"><img src="images/slide3.jpg" width="935" height="293" alt="" /> </a> </div>
      <div class="clr"></div>
    </div>
    <div class="clr"></div>
  </div>
 </div>
 <div class="content">
  <div class="content_resize">
    <div class="mainbar">
```

```
<div class="article">
 <h2 class="style2" style="color:#FF0033">Add News Details..</h2>
 <div class="clr"></div>
     <div class="post_content">
          <%@ include file="connect.jsp" %>
  <%@ page import="java.io.*"%>
  <%@ page import="java.util.*" %>
  <%@ page import="java.util.Date" %>
  <%@ page import="com.oreilly.servlet.*"%>
  <%@ page import ="java.text.SimpleDateFormat" %>
  <%@ page import ="javax.crypto.Cipher" %>
  <%@ page import ="javax.crypto.spec.SecretKeySpec" %>
  <%@ page import
="java.security.KeyPairGenerator,java.security.KeyPair,java.security.Key" %>
 <form name="s" action="S_AddNews1.jsp" method="post"  onSubmit="return
valid()"  ons target="_top">
 <% String chName=(String)application.getAttribute("server");
    try
{   Date now1 = new Date();
          SimpleDateFormat sdf = new SimpleDateFormat("dd/MM/yyyy");
          String strCurrDate = sdf.format(now1);
          Date date1 = new SimpleDateFormat("dd/MM/yyyy").parse(strCurrDate);

          int year= Calendar.getInstance().get(Calendar.YEAR);
           ArrayList a1=new ArrayList();
 String query="select category FROM categories";
```

```
        Statement st=connection.createStatement();
        ResultSet rs=st.executeQuery(query);
    while ( rs.next() ) { a1.add(rs.getString("category")); }
```

S_Fake_News.jsp

```
<html xmlns="http://www.w3.org/1999/xhtml">
<head>
<title>View all news post</title>
<meta http-equiv="Content-Type" content="text/html; charset=utf-8" />
<link href="css/style.css" rel="stylesheet" type="text/css" />
<link rel="stylesheet" type="text/css" href="css/coin-slider.css" />
<script type="text/javascript" src="js/cufon-yui.js"></script>
<script type="text/javascript" src="js/cufon-quicksand.js"></script>
<script type="text/javascript" src="js/jquery-1.4.2.min.js"></script>
<script type="text/javascript" src="js/script.js"></script>
<script type="text/javascript" src="js/coin-slider.min.js"></script>
<style type="text/css">
<!--
.style1 {font-size: 24px}
.style2 {color: #FF0033}
.style13 {color: #FFFFFF}
-->
</style>
</head>
<body>
<div class="main">
```

```
<div class="header">
  <div class="header_resize">
    <div class="logo">
      <h1><a href="index.html" class="style1">Fake News, Disinformation,and Deepfakes: Leveraging Distributed Ledger Technologies and Blockchain to Combat Digital Deception and Counterfeit Reality
</a></h1>
    </div>
    <div class="menu_nav">
      <ul>
        <li class="active"><a href="index.html"><span>Home Page</span></a></li>
        <li><a href="ServerLogin.jsp"><span>News Publisher </span></a></li>
        <li><a href="UserLogin.jsp"><span>User </span></a></li>
      </ul>
    </div>
    <div class="clr"></div>
    <div class="slider">
      <div id="coin-slider"> <a href="#"><img src="images/slide1.jpg" width="935" height="293" alt="" /> </a> <a href="#"><img src="images/slide2.jpg" width="935" height="293" alt="" /> </a> <a href="#"><img src="images/slide3.jpg" width="935" height="293" alt="" /> </a> </div>
      <div class="clr"></div>
    </div>
    <div class="clr"></div>
  </div>
</div>
<div class="content">
```

```
    <div class="content_resize">
      <div class="mainbar">
        <div class="article">
          <h2 class="style2" style="color:#FF0033">View All Fake News..</h2>
          <div class="clr"></div>
                <div class="post_content">
                <p> </p>
  <style type="text/css">
<!--
.style1 {color: #FFFFFF}
.style2 {
   font-weight: bold;
   color: #FFFFFF;
}
.style4 {font-weight: bold}
.style5 {color: #FF0000}
.style6 {color: #FF0000; font-weight: bold; }
-->
</style>
       <table width="600" border="0" align="center">
   <tr>
    <td width="140" height="32" bgcolor="#FF0000"><div align="center"
class="style13"><span class="style3 "><b>Id</b> </span></div></td>
    <td width="178" bgcolor="#FF0000"><div align="center" class="style13"><b>Channel
Name</b></div></td>
```

```
  <td width="178" bgcolor="#FF0000"><div align="center" class="style13"><b>News Name</b></div></td>


     <td width="293" bgcolor="#FF0000"><div align="center" class="style13"><span class="style3 "><b>Description</b> </span></div></td>

   <td width="178" bgcolor="#FF0000"><div align="center" class="style13"><b>News Image</b></div></td>

    <td width="205" bgcolor="#FF0000"><div align="center" class="style13"><b>Date</b></div></td>

    <td width="205" bgcolor="#FF0000"><div align="center" class="style13"><b>News Place</b></div></td>

            <td width="205" bgcolor="#FF0000"><div align="center" class="style13"><b>Edit</b></div></td>

  </tr>

   <%@ include file="connect.jsp"%>

<%

  String s1="",s2="",s3="",s4="",s5="",s6="",s7="", pos="Fake",s22="" ;

  int i=0,poscnt=0,negcnt=0,strcnt=0;

  int count1=0;

  String ftype="Fake";

  try

  {

                String query="select * from news ";

                    Statement st=connection.createStatement();

                    ResultSet rs=st.executeQuery(query);

                while ( rs.next() )

                {
```

```
i=rs.getInt(1);
s1=rs.getString(8);
s2=rs.getString(3);
s4=rs.getString(5).toLowerCase(); // desc
s5=rs.getString(7);
s6=rs.getString(9);
String sql1="select * from filter where tctype='"+pos+"' ";
Statement st1=connection.createStatement();
ResultSet rs1=st1.executeQuery(sql1);
while ( rs1.next() )
{
String t1=rs1.getString(1);
String t2=rs1.getString(2).toLowerCase();

if ((s4.contains(t2)))
{
count1++;
%>
<tr>
<td><div align="center" style="color:#000000"><%=i%></div></td>
<td><div align="center" style="color:#CC0066"><%=s1%></div></td>
<td><div align="center" style="color:#FF6600"><%=s2%></div></td>
<td><div align="center" style="color:#000000"><%=s4%></div></td>
<td><div align="center">
```

```
    <input  name="image" type="image"
src="user_Pic.jsp?picture=<%="newsimage"%>&id=<%=i%>" width="100" height="100"
alt="Submit">
    </input></div>        </td>
                    <td><div align="center" style="color:#000000"><%=s5%></div></td>
                    <td><div align="center" style="color:#000000"><%=s6%></div></td>

 </tr>
              <%                                    }
                                        }
                                }
       connection.close();
            }
        catch(Exception e)
        {
          out.println(e.getMessage());
        }
 %></table>
<p> </p>
                    <p><a href="Server_Main.jsp" class="style14">Back</a></p>
              </div>
        <div class="clr"></div>
      </div>
      <p class="pages"> </p>
    </div>
    <div class="sidebar">
```

```
      <div class="searchform">
        <form id="formsearch" name="formsearch" method="post" action="#">
          <span>
          <input name="editbox_search" class="editbox_search" id="editbox_search" maxlength="80" value="Search our ste:" type="text" />
          </span>
          <input name="button_search" src="images/search.gif" class="button_search" type="image" />
        </form>
      </div>
      <div class="clr"></div>
      <div class="gadget">
        <h2 class="star"><span>Server</span> Menu</h2>
        <div class="clr"></div>
        <ul class="sb_menu">
          <li><a href="Server_Main.jsp">&raquo; Home </a></li>
                    <li><a href="ServerLogin.jsp">&raquo; Logout </a></li>
        </ul>
      </div>
    </div>
    <div class="clr"></div>
  </div>
 </div>
 <div class="footer">
  <div class="footer_resize">
    <div style="clear:both;"></div>
```

```
    </div>
  </div>
</div>
<div align=center></div>
</body>
</html>
```

6.2 OUTLINE FOR VARIOUS FILES

authentication.jsp

connect.jsp

currentNews.jsp

delete.jsp

filename.txt

index.html

news_Image_Details.jsp

S_AddCategory.jsp

S_AddCategory1.jsp

S_AddFilter.jsp

S_AddFilter1.jsp

S_AddNews.jsp

S_AddNews1.jsp

S_AddNews2.jsp

S_AddNews3.jsp

S_AuthorizeUsers.jsp

S_Fake_News.jsp

S_NewsPostRankinChart.jsp

S_PostRankinChart.jsp

S_SelectCat.jsp

S_UpdateNews.jsp

S_UserNews Transaction.jsp

S_UserStatus.jsp

S_ViewNewsPostByQuantization.jsp

S_ViewNewsPostInBlockChain.jsp

S_ViewProductQuantInChart.jsp

S_ViewProductQuantInChart1.jsp

SamplePage.jsp

SearchNews.jsp

Server_Main.jsp

Server_Register.jsp

Server_RegisterStatus.jsp

ServerLogin.jsp

U_Fake_News.jsp

U_Review.jsp

U_Reviewins.jsp

U_SearchNewsHashCode.jsp

U_UProfile.jsp

U_ViewNews.jsp

U_ViewNewsHashCode.jsp

U_ViewSearch Transaction.jsp

update.jsp

update1.jsp

User_Main.jsp

user_Pic.jsp

User_Register.jsp

User_Register Status.jsp

UserLogin.jsp

UserProfile.jsp

Wrong_Login.jsp

6.3 CLASS WITH FUNCTIONALITIES

In the project home page, we have two modules to which we can navigate to, they are news publisher server module and the user module. Let us look into the brief overview of each module.

6.3.1 NEWS PUBLISHER SERVER MODULE

To view all the inner sub-modules and their functionalities first we have to login to the page and then we can perform the required operations.

a) **LIST OF ALL USERS AND AUTHORIZE**: In this module we can view all the list of users along with all the details such as id, image, username, mobile number, email id, address, and the login status. Login status gives us information about whether the user is an authorized legitimate user or not. If not, the news publisher can verify the user and make him an authorized user.

b) **ADD NEWS CATEGORIES**: In this module we can add the news category after which we can post the news related to the added category.

c) **ADD FAKE FILTER**: In this module we will add the fake filter which identifies the fake news from the actual news. we can give the fake filters such as the wrong, misleading etc.

d) **ADD NEWS**: In this module we can add the news we wish to post into the website, so that users can view the post. In this we have to mention the all the details about the news such as news category, news name, about the news, news image, news year, news date etc.

e) **VIEW ALL NEWS POST AND UPDATE**: In this module we can view all the news which has been already added.

f) **VIEW ALL NEWS POST BY DISTRIBUTED LEDGER TECHNOLOGY**: In this module we can view the news that has been already added but according to the distributed ledger technology the news will be separated as “current trending news” and “the older news”. It displays all the details about the news.

g) **VIEW ALL THE FAKE NEWS**: In this module the news which is separated from the actual news will be displayed here. the news will be separated according to the fake filter which we have added in the module "add fake filter".

h) **VIEW ALL NEWS POSTS IN BLOCKCHAIN**: Since the data in the database using the blockchain is stored in the form of blocks, the news displayed in this module all the news will be organized according to the different categories we have added.

i) **VIEW ALL USERS NEWS TRANSACTIONS BY KEYWORD**: In this module we can the all the news by searching the news by just giving the keywords. We can give different keywords like sports, entertainment, business etc.

j) **VIEW ONLINE PRODUCT DISTRIBUTED LEDGER TECHNOLOGIES BY CHART**: In this module we can view all the news separated by the category along with the number of news posted in each category in the form of graphical representation.

k) **VIEW ALL NEWS POST RANK IN CHART**: In this module the rank will be given to all the news according to news posted by categories and displayed in the form of graph representation.

l) **LOGOUT**: The last module in the news publisher server is the logout option by which we can logout of the page and navigated to the home page of the website.

6.3.2 USER MODULE

To view all the functionalities of the user module we must login to the user page, and all the users who have just registered cannot be able to login the must be authorized first in order to login and perform the required operations. In the user module we have the following sub modules:

a) **HOME**: In the home screen of the message displays the welcome message to the user

b) **VIEW YOUR PROFILE:** In this module we can view all the details related to us like username, image, email id, address, mobile number which we have submitted during the registration of the user.

c) **CURRENT NEWS:** In this module we can view all the news that are already available and posted the news publishers separated as the "current trending news" and "the older news ".

d) **SEARCH NEWS:** By this module we can search the news we wish to view by just entering the keywords such as sports etc, then all the news that are related to that keyword will be displayed.

e) **VIEW ALL FAKE NEWS:** In this module all the fake news separated according to the fake filter we have mentioned will be displayed here.

f) **VIEW ALL SEARCH TRANSACTIONS:** In this module all the search transactions we have performed to view the news can be displayed along with the news details related to that search transaction.

g) **LOGOUT:** This is the final operation in the user module by which the user can simply logout and end the session.

6.4 METHODS INPUT AND OUTPUT PARAMETERS

6.4.1 Inputs

This section is a description of the input media used by the operator for providing information to the system; show a mapping to the high-level data flows described in Section 1 .2.1, System Overview. For example, data entry screens, optical character readers, bar scanners, etc. If appropriate, the input record types, file structures, and database structures provided in Section 3, File and Database Design, may be referenced. Include data element definitions or refer to the data dictionary.

Provide the layout of all input data screens or graphical user interfaces (GUTs) (for example, windows). Provide a graphic representation of each interface. Define all data elements associated with each screen or GUI or reference the data dictionary.

This section should contain edit criteria for the data elements, including specific values, range of values, mandatory/optional, alphanumeric values, and length. Also address data entry controls to prevent edit bypassing.

Discuss the miscellaneous messages associated with operator inputs, including the following:

- Copies of form(s) if the input data are keyed or scanned for data entry from printed forms.
- Description of any access restrictions or security considerations.
- Each transaction name, code, and definition if the system is a transaction-based processing system.

6.4.2 Outputs

This section describes of the system output design relative to the user/operator; show a mapping to the high-level data flows described in Section 1.2.1. System outputs include reports, data

display screens and GUIs, query results, etc. The output files are described in Section 3 and may be referenced in this section. The following should be provided, if appropriate:

- Identification of codes and names for reports and data display screens.
- Description of report and screen contents (provide a graphic representation of each layout and
  define all data elements associated with the layout or reference the data dictionary).
- Description of the purpose of the output, including identification of the primary users.
- Report distribution requirements, if any (include frequency for periodic reports).
- Description of any access restrictions or security considerations.

ServerLogin.jsp

```
<FORM ACTION="authentication.jsp?type=<%="server"%>" METHOD="post"
id="leavereply">
        <p>
            <%
        try     {
                ArrayList a1=new ArrayList();
       String query="select distinct servername FROM server";
       Statement st=connection.createStatement();
       ResultSet rs=st.executeQuery(query);
    while ( rs.next() )
    {
                a1.add(rs.getString("servername"));
    }
%>
            </p>
        <p align="center"><img src="images/Login.png" width="243" height="128" /></p>
        <TABLE WIDTH="353" BORDER="0" ALIGN="center">
```

```
<TR>
<TD WIDTH="190" HEIGHT="30" CLASS="style1"><DIV ALIGN="center" CLASS="style13 style4 style8">Select Channel Name </DIV></TD>
<TD WIDTH="115"><DIV ALIGN="left">
<select id="s1" name="serverid">
<option>--Select--</option>
<%
for(int i=0;i<a1.size();i++)
{
%>
<option><%= a1.get(i)%></option>
<%
}
%>
</select></DIV></TD>
</TR>
<TR>
<TD WIDTH="160" HEIGHT="30" CLASS="style1"><DIV ALIGN="center" CLASS="style13 style4 style8">Channel User Name </DIV></TD>
<TD WIDTH="163"><DIV ALIGN="center"><INPUT TYPE="text" name="userid" /></DIV></TD>
</TR>
<TR>
<TD HEIGHT="35" CLASS="style1"><DIV ALIGN="center" CLASS="style13 style4 style8">Password</DIV></TD>
<TD><DIV ALIGN="center"><INPUT TYPE="password" name="pass" /></DIV></TD>
```

```
            </TR>
            <TR>
                        <TD> </TD>
              <TD HEIGHT="45">
                <DIV ALIGN="left">
                  <INPUT TYPE="image" name="imageField" id="imageField"
SRC="images/submit.gif" CLASS="send" />
                </DIV></TD></TR>
           </TABLE>
                      <p>
                       <%
          connection.close();          }
         catch(Exception e)
         {
          out.println(e.getMessage());
         }
%>                    </p>
                      <P> </P>
       </FORM>
```

authentication.jsp

```
<%
            String type=request.getParameter("type");
    try{
            if(type.equalsIgnoreCase("server"))
            {
```

```
                String servername=request.getParameter("serverid");
                String cusername=request.getParameter("userid");
            String Password=request.getParameter("pass");

                application.setAttribute("server",servername);

                String sql="SELECT * FROM server where cusername='"+cusername+"' and
(servername='"+servername+"' and password='"+Password+"')";
                Statement stmt = connection.createStatement();
                ResultSet rs =stmt.executeQuery(sql);

                if(rs.next())
                {
                        response.sendRedirect("Server_Main.jsp");
                }
                else
                {
                        response.sendRedirect("Wrong_Login.jsp");
                }
            }
             if(type.equalsIgnoreCase("user"))
            {
                String username=request.getParameter("userid");
               String Password=request.getParameter("pass");
                application.setAttribute("user",username);
```

```
String sql="SELECT * FROM user where username='"+username+"' and password='"+Password+"'";
Statement stmt = connection.createStatement();
ResultSet rs =stmt.executeQuery(sql);
if(rs.next())
{
String sql1="SELECT * FROM user where username='"+username+"' and status='Authorized'";
Statement stmt1 = connection.createStatement();
ResultSet rs1 =stmt1.executeQuery(sql1);
if(rs1.next())
  {
      response.sendRedirect("User_Main.jsp");
      }
      else
      {
      %>
                                                    <br/><h3><p align="left" class="style3"> </p>
                                                     <p align="left" class="style4" style="color:#000000">You are not the Authorized User, Please wait!! </p>
                                                    </h3>
                                                    <br/><br/><a href="UserLogin.jsp">Back</a>
                                                    <%
  }
}
```

```
                else
                {
                        response.sendRedirect("Wrong_Login.jsp");
                }
        }
  }
  catch(Exception e)
  {
        out.print(e);
  }
%>
```

Servermain.jsp

```
<h2 class="style3" style="color:#FF0033">Welcome To Channel <span
class="style2"><%=(String)application.getAttribute("server")%></span> ..!</h2>
        <div class="clr"></div>
        <div class="post_content">

        </div>
        <div class="clr"></div>
      </div>
      <p class="pages"><img src="images/Admin.jpg" width="624" height="372" /></p>
    </div>
    <div class="sidebar">
```

```
<div class="searchform">
 <form id="formsearch" name="formsearch" method="post" action="#">
  <span>
  <input name="editbox_search" class="editbox_search" id="editbox_search" maxlength="80" value="Search our ste:" type="text" />
  </span>
  <input name="button_search" src="images/search.gif" class="button_search" type="image" />
 </form>
</div>
<div class="clr"></div>
<div class="gadget">
 <h2 class="star"><span>Server</span> Menu</h2>
 <div class="clr"></div>
 <ul class="sb_menu">

  <li><a href="Server_Main.jsp">&raquo; Home</a></li>
        <li><a href="S_AuthorizeUsers.jsp">&raquo; List of All Users and Authorize</a></li>
        <li><a href="S_AddCategory.jsp">&raquo; Add News Categories</a></li>
        <li><a href="S_AddFilter.jsp">&raquo; Add Fake Filter</a></li>
        <li><a href="S_AddNews.jsp">&raquo; Add News</a></li>
        <li><a href="S_UpdateNews.jsp">&raquo; View all news post and update</a></li>
        <li><a href="S_ViewNewsPostByQuantization.jsp">&raquo;All News post by distributed ledger technologies </a></li>
        <li><a href="S_Fake_News.jsp">&raquo;View All Fake News </a></li>
```

»View All News Posts in Block Chain Form

»View All Users News transactions by keyword

»View online product Distributed

Ledger Technologies by chart

»View all news post rank in chart

» Logout

S_AddFilter.jsp

<form id="form1" name="form1" method="post" action="S_AddFilter1.jsp">

<table width="385" border="2">

<tr>

<td width="181" height="47" bgcolor="#FF0000">Select Filter Category </td>

<td width="186"><select name="tclass">

<option>----Select Filter---</option>

<option>Fake</option>

</select> </td>

```
        </tr>
        <tr>
          <td height="52" bgcolor="#FF0000"><span class="style11 style10">Enter Filter Name </span></td>
          <td><input type="text" name="fname" /></td>
        </tr>
        <tr>
          <td height="52"> </td>
          <td><p>
            <input type="submit" name="Submit" value="Add" />
            <input type="reset" name="Submit2" value="Reset" />
          </p></td>
        </tr>
      </table>
      <p> </p>
      <p><a href="AdminMain.jsp"></a></p>
      <p class="style12 style1">Filter Details </p>
    </form>
        <%@ include file="connect.jsp" %>
    <table width="329" border="1">
                    <tr>
                      <td width="195" height="42" bgcolor="#FF0000"><div align="center" class="style10"><span class="style3 style14"><b>Fake Category </b></span></div></td>
                      <td width="439" bgcolor="#FF0000"><div align="center" class="style14 style10"><b>Filter Name</b></div></td>
        </tr>
```

```
<%
                String s0="",s1="",s2="",s3="",s4="",s5="",s6="";
                int i=1,j=0,count=0,rank=0,k=0;
                    try
                    {
                        String query="select * from filter ";
                        Statement st=connection.createStatement();
                        ResultSet rs=st.executeQuery(query);
                        while ( rs.next() )
                        {
                            s0=rs.getString(1);
                            s1=rs.getString(2);
                %>
<tr>
  <td height="33"  valign="middle" bgcolor="#FFFFFF">
    <div align="center" class="style4 style12 style14 style8" >
      <div align="center">
      <%out.println(s0);%>
   </div>
 </div></td>
          <td height="33"  valign="middle" bgcolor="#FFFFFF">
    <div align="center" class="style4 style12 style14 style9" >
      <div align="center">
      <%out.println(s1);%>
   </div>
```

```
          </div></td>
        </tr>
        <%
                }
                                        connection.close();
                                }
                                catch(Exception e)
                                {
                                        out.println(e.getMessage());
                                }
                                %>
      </table>
   <p><a href="Server_Main.jsp">Back</a></p>
            </div>
```

UserLogin.jsp

```
<div class="content">
   <div class="content_resize">
     <div class="mainbar">
       <div class="article">
         <h2 class="style2" style="color:#FF0033">User  Login  Page...!</h2>
         <div class="clr"></div>
             <div class="post_content">
         <p align="center"><img src="images/Login.png" width="265" height="161" /></p>
         <form action="authentication.jsp?type=<%="user"%>" method="post"
id="leavereply">
            <table width="313" border="0" align="center">
```

```
<tr>
<td width="140" height="30" class="style1"><div align="center" class="style13 style4 style8">  Name </div></td>
<td width="163"><div align="center"><input type="text" name="userid" /></div></td>
</tr>
<tr>
<td height="35" class="style1"><div align="center" class="style13 style4 style8">Password</div></td>
<td><div align="center"><input type="password" name="pass" /></div></td>
</tr>
<tr>
<td> </td>
<td height="45">
<div align="left">
<input type="image" name="imageField" id="imageField" src="images/submit.gif" class="send" />
</div></td></tr>
</table>
</form>
<p align="right" class="style3"><span class="style6">New</span> <a href="User_Register.jsp" class="style7">Register</a></p>
<p> </p>
</div>
<div class="clr"></div>
</div>
```

```
      <p class="pages"> </p>
    </div>
    <div class="sidebar">
      <div class="searchform">
        <form id="formsearch" name="formsearch" method="post" action="#">
          <span>
          <input name="editbox_search" class="editbox_search" id="editbox_search"
maxlength="80" value="Search our ste:" type="text" />
          </span>
          <input name="button_search" src="images/search.gif" class="button_search"
type="image" />
        </form>
      </div>
      <div class="clr"></div>
      <div class="gadget">
        <h2 class="star"><span>Sidebar</span> Menu</h2>
        <div class="clr"></div>
        <ul class="sb_menu">
          <li><a href="index.html">&raquo; Home </a></li>
        </ul>
      </div>
    </div>
    <div class="clr"></div>
  </div>
 </div>
```

User_Main.jsp

```
<div class="content">
   <div class="content_resize">
     <div class="mainbar">
       <div class="article">
         <h2 class="style3" style="color:#FF0033">Welcome User <span class="style9"><%=(String)application.getAttribute("user")%></span></h2>
         <div class="clr"></div>
              <div class="post_content">
         </div>
         <div class="clr"></div>
       </div>
       <p align="center" class="pages"><img src="images/User.jpg" width="504" height="251" /></p>
     </div>
     <div class="sidebar">
       <div class="searchform">
         <form id="formsearch" name="formsearch" method="post" action="#">
           <span>
           <input name="editbox_search" class="editbox_search" id="editbox_search" maxlength="80" value="Search our ste:" type="text" />
           </span>
           <input name="button_search" src="images/search.gif" class="button_search" type="image" />
         </form>
       </div>
       <div class="clr"></div>
       <div class="gadget">
```

```
        <h2 class="star"><span>User</span> Menu</h2>
        <div class="clr"></div>
        <ul class="sb_menu">
          <li><a href="User_Main.jsp">&raquo; Home </a></li>
                    <li><a href="UserProfile.jsp">&raquo;View Your Profile</a></li>
          <li><a href="currentNews.jsp">&raquo;Current News </a></li>
                    <li><a href="SearchNews.jsp">&raquo;Search News  </a></li>
                    <li><a href="U_Fake_News.jsp">&raquo;View All Fake News </a></li>
                    <li><a href="U_ViewSearchTransaction.jsp">&raquo;View All Search
Transaction</a></li>
          <li><a href="UserLogin.jsp">&raquo;Log Out </a></li>
        </ul>
      </div>
    </div>
    <div class="clr"></div>
  </div>
 </div>
 <div class="footer">
  <div class="footer_resize">
    <div style="clear:both;"></div>
  </div>
 </div>
</div>
```

CHAPTER 7
PROJECT TESTING

7 PROJECT TESTING

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub-assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of tests. Each test type addresses a specific testing requirement.

Software Testing is a method to check whether the actual software product matches expected requirements and to ensure that software product is Defect free. It involves execution of software/system components using manual or automated tools to evaluate one or more properties of interest. The purpose of software testing is to identify errors, gaps or missing requirements in contrast to actual requirements.

Some prefer saying Software testing definition as a White Box and Black Box Testing. In simple terms, Software Testing means the Verification of Application Under Test (AUT). This Software Testing course introduces testing software to the audience and justifies the importance of software testing.

It depends on the process and the associated stakeholders of the project(s). In the IT industry, large companies have a team with responsibilities to evaluate the developed software in context of the given requirements. Moreover, developers also conduct testing which is called Unit Testing. In most cases, the following professionals are involved in testing a system within their respective capacities:

- Software Tester
- Software Developer
- Project Lead/Manager
- End User

Different companies have different designations for people who test the software on the basis of their experience and knowledge such as Software Tester, Software Quality Assurance Engineer, QA Analyst, etc.

It is not possible to test the software at any time during its cycle. The next two sections state when testing should be started and when to end it during the SDLC.

An early start to testing reduces the cost and time to rework and produce error-free software that is delivered to the client. However, in Software Development Life Cycle (SDLC), testing can be started from the Requirements Gathering phase and continued till the deployment of the software.

It also depends on the development model that is being used. For example, in the Waterfall model, formal testing is conducted in the testing phase; but in the incremental model, testing is performed at the end of every increment/iteration and the whole application is tested at the end.

7.1 VARIOUS TEST CASES

7.1.1 UNIT TESTING

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

Unit testing is usually conducted as part of a combined code and unit test phase of the software lifecycle, although it is not uncommon for coding and unit testing to be conducted as two distinct phases.

Test strategy and approach

Field testing will be performed manually, and functional tests will be written in detail.

Test objectives

- All field entries must work properly.
- Pages must be activated from the identified link.
- The entry screen, messages and responses must not be delayed.

Features to be tested:

- Verify that the entries are of the correct format.
- No duplicate entries should be allowed.

- All links should take the user to the correct page.

7.1.2 INTEGRATION TESTING

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

Software integration testing is the incremental integration testing of two or more integrated software components on a single platform to produce failures caused by interface defects.

The task of the integration test is to check that components or software applications, e.g., components in a software system or – one step up – software applications at the company level – interact without error.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

7.1.3 FUNCTIONAL TESTING

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals. Functional testing is centered on the following items:

- Valid Input : identified classes of valid input must be accepted.
- Invalid Input : identified classes of invalid input must be rejected.
- Functions : identified functions must be exercised.
- Output : identified classes of application outputs must be exercised.

Systems/Procedures: Interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

7.1.4 SYSTEM TESTING

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration-oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

Software once validated must be combined with other system elements (e.g., Hardware, people, database). System testing verifies that all the elements are proper, and that overall system function performance is achieved. It also tests to find discrepancies between the system and its original objective, current specifications and system documentation.

7.1.5 ACCEPTANCE TESTING

User Acceptance of a system is the key factor for the success of any system. The system under consideration is tested for user acceptance by constantly keeping in touch with the prospective system users at the time of developing and making changes wherever required. The system developed provides a friendly user interface that can easily be understood even by a person who is new to the system.

User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

7.1.6 OUTPUT TESTING

After performing the validation testing, the next step is output testing of the proposed system, since no system could be useful if it does not produce the required output in the specified format. Asking the users about the format required by them tests the outputs generated or displayed by the system under consideration. Hence the output format is considered in 2 ways – one is on screen and another in printed format.

7.1.7 VALIDATION CHECKING

Validation checks are performed on the following fields.

- **Text Field:**

  The text field can contain only the number of characters lesser than or equal to its size. The text fields are alphanumeric in some tables and alphabetic in other tables. Incorrect entry always flashes and error message.

- **Numeric Field:**

The numeric field can contain only numbers from 0 to 9. An entry of any character flashes an error message. The individual modules are checked for accuracy and what it has to perform. Each module is subjected to test run along with sample data. The individually tested modules are integrated into a single system. Testing involves executing the real data information is used in the program the existence of any program defect is inferred from the output. The testing should be planned so that all the requirements are individually tested.

7.2 BLACK BOX TESTING

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box. you cannot “see” into it. The test provides inputs and responds to outputs without considering how the software works.

The above Black-Box can be any software system you want to test. For Example, an operating system like Windows, a website like Google, a database like Oracle or even your own custom application. Under Black Box Testing, you can test these applications by just focusing on the inputs and outputs without knowing their internal code implementation.

There are many types of Black Box Testing, but the following are the prominent ones -

- **Functional testing** - This black box testing type is related to the functional requirements of a system; it is done by software testers.
- **Non-functional testing** - This type of black box testing is not related to testing of specific functionality, but non-functional requirements such as performance, scalability, usability.
- **Regression testing** - Regression Testing is done after code fixes, upgrades, or any other system maintenance to check the new code has not affected the existing code.

Test procedure

The test procedure of black box testing is a kind of process in which the tester has specific knowledge about the software's work, and it develops test cases to check the accuracy of the software's functionality.

It does not require programming knowledge of the software. All test cases are designed by considering the input and output of a particular function. A tester knows about the definite output of a particular input, but not about how the result is arising. There are various techniques used in black box testing for testing like decision table technique, boundary value analysis technique, state transition, All-pair testing, cause-effect graph technique, equivalence partitioning technique, error guessing technique, use case technique and user story technique. All these techniques have been explained in detail within the tutorial.

7.3 WHITE BOX TESTING

White Box Testing is a testing in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is purpose. It is used to test areas that cannot be reached from a black box level.
It is one of two parts of the Box Testing approach to software testing. Its counterpart, Blackbox testing, involves testing from an external or end-user type perspective. On the other hand, White box testing in software engineering is based on the inner workings of an application and revolves around internal testing.

The term "White Box" was used because of the see-through box concept. The clear box or White Box name symbolizes the ability to see through the software's outer shell (or "box") into its inner workings. Likewise, the "black box" in "Black Box Testing" symbolizes not being able to see the inner workings of the software so that only the end-user experience can be tested.

White-box testing is a method of testing the application at the level of the source code. These test cases are derived through the use of the design techniques mentioned above: control flow testing, data flow testing, branch testing, path testing, statement coverage and decision coverage as well as modified condition/decision coverage. White-box testing is the use of these techniques as guidelines to create an error-free environment by examining all code. These white-box testing techniques are the building blocks of white-box testing, whose essence is the careful testing of the application at the source code level to reduce hidden errors later on. These different techniques exercise every visible path of the source code to minimize errors and create an error-free environment. The whole point of white-box testing is the ability to know which line of the code is being executed and being able to identify what the correct output should be.

Working process of white box testing:

- **Input:** Requirements, Functional specifications, design documents, source code.
- **Processing:** Performing risk analysis for guiding through the entire process.

- **Proper test planning:** Designing test cases so as to cover entire code. Execute rinse-repeat until error-free software is reached. Also, the results are communicated.
- **Output:** Preparing final report of the entire testing process.

The testing can be done at system, integration, and unit levels of software development. One of the

basic goals of white box testing are to verify a working flow for an application. It involves testing a series of predefined inputs against expected or desired outputs so that when a specific input does not result in the expected output, you have encountered a bug.

CHAPTER 8

OUTPUT SCREENS

8 OUTPUT SCREENS

8.1 USER INTERFACES

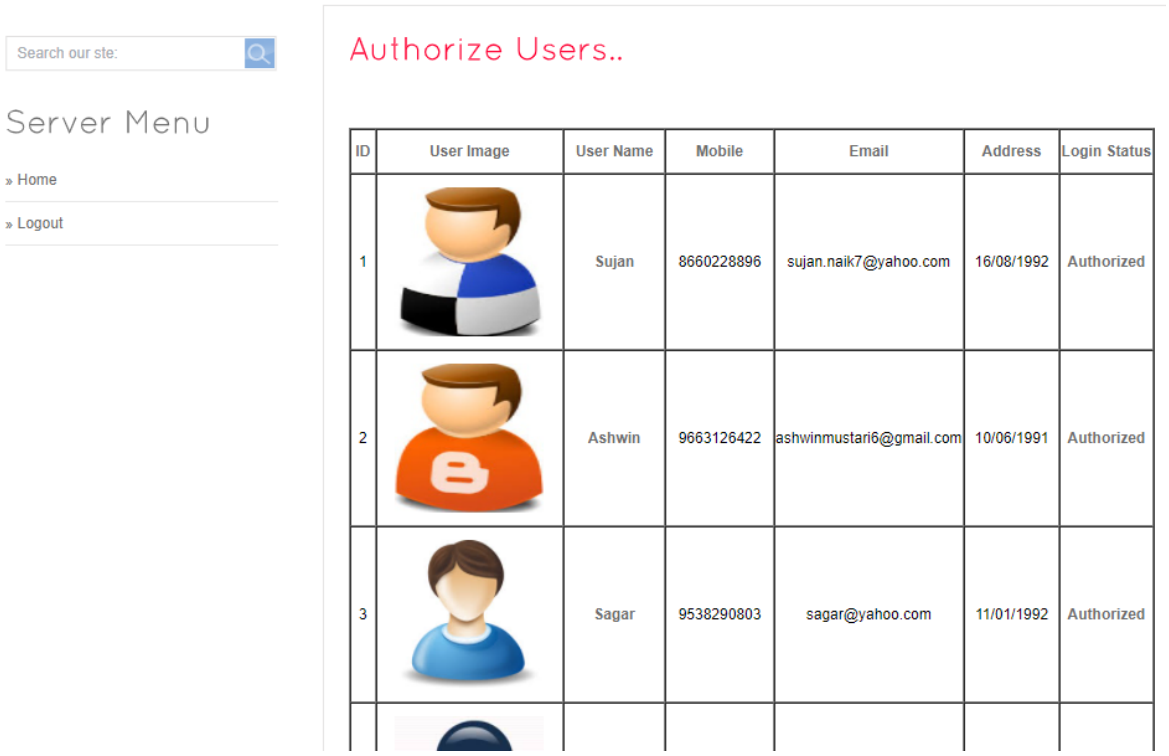


Fig 8.1 Authorized Users List

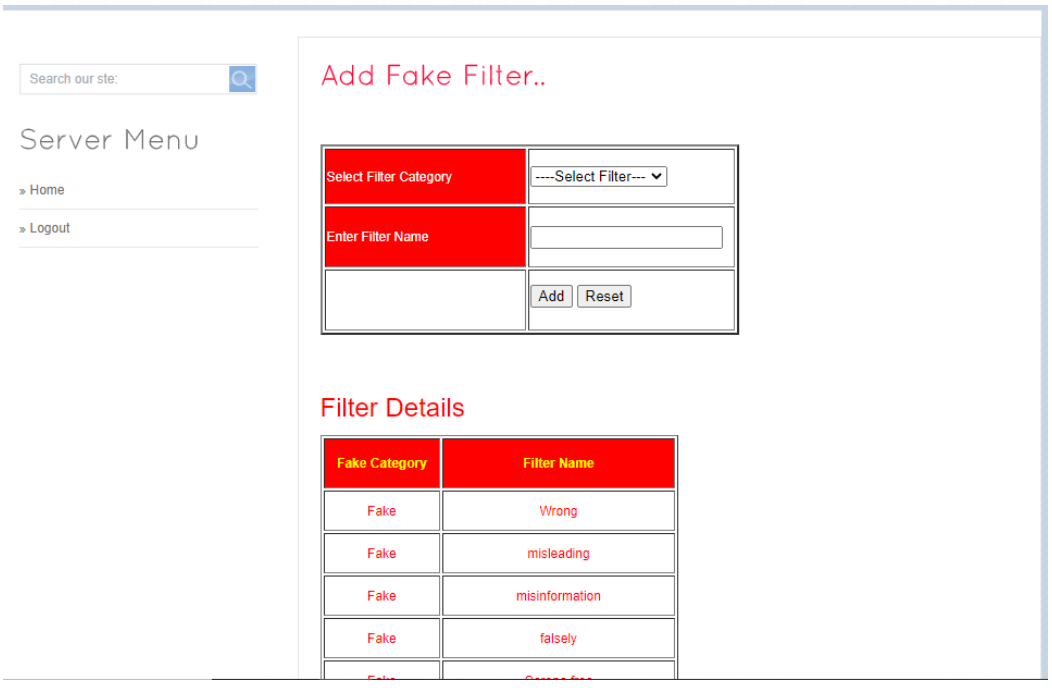


Fig 8.2 Add Fake Filter

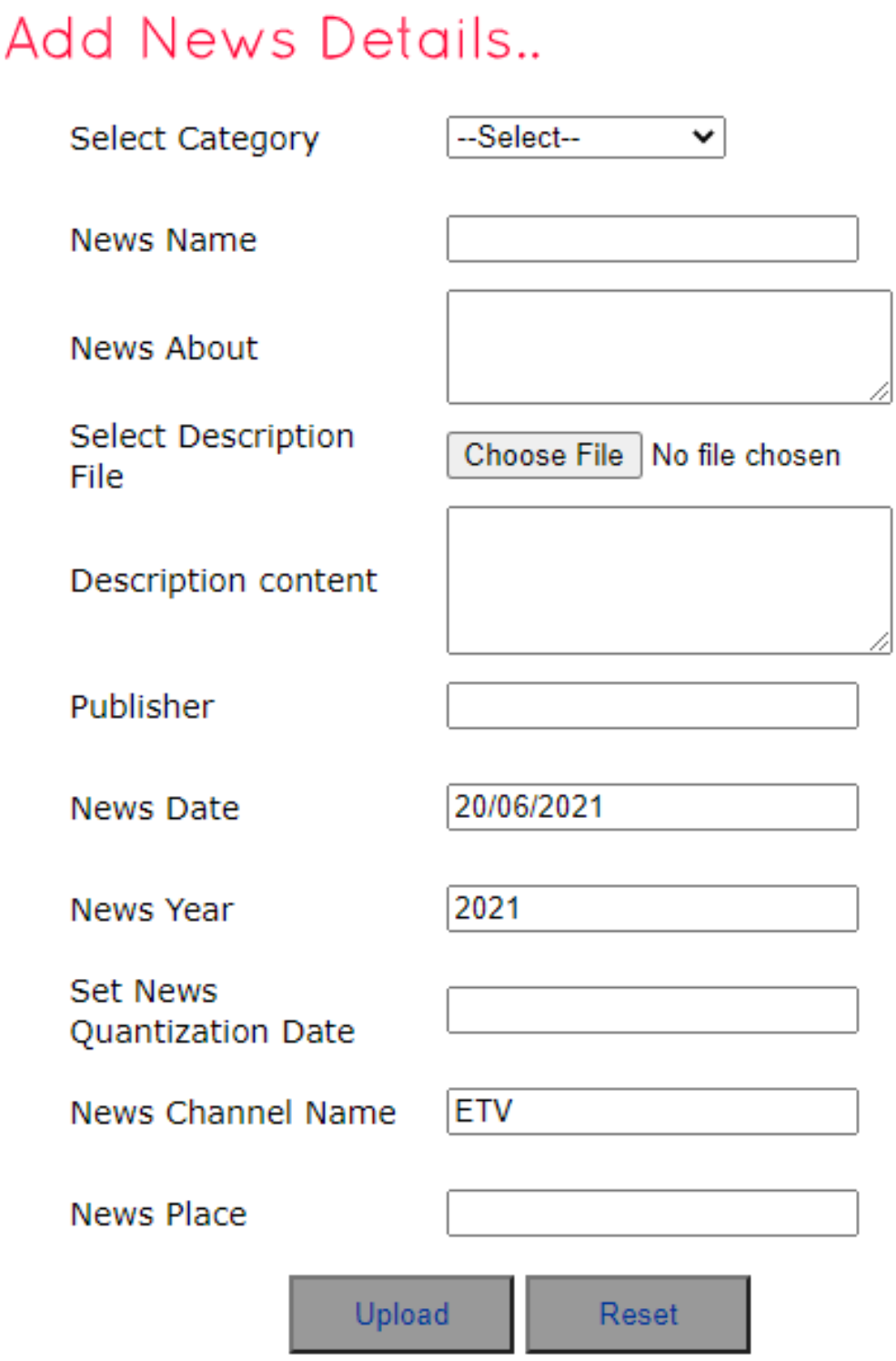


Fig 8.3 Add News Details

Fig 8.4 View All Fake News

View All News Post In Block Chain Form ..

BlockChain::Category : Yellow Fungus

Id	Channel Name	News Name	Description	News Image	Date	News Place
13	ETV	Colour fungus	Yellow fungus is spreading aroung the world		26/05/2021	hyd

BlockChain::Category : Covid 19

Id	Channel Name	News Name	Description	News Image	Date	News Place
14	ETV	Covid 19	India Now is Corona free		27/05/2021	hyd
			PM Narendra Modi requested all the Indians very soon			

Fig 8.5 News Posts in Blockchain Form

View All News Post By distributed ledger technologies ..

Current Trending News....

News Date	Channel Name	News Name	Description	News Image	News Place
27/05/2021	ETV	Covid 19	India Now is Corona free		hyd
27/05/2021	ETV	Public view	PM Narendra Modi requested all the Indians very soon we have to get corona free India. Kindly support.		hyd
26/05/2021	ETV	Colour fungus	Yellow fungus is spreading aroung the world		hyd

Older News.....

News Date	Channel Name	News Name	Description	News Image	News Place

Fig 8.6 News Posts by Distributed Ledger Technology

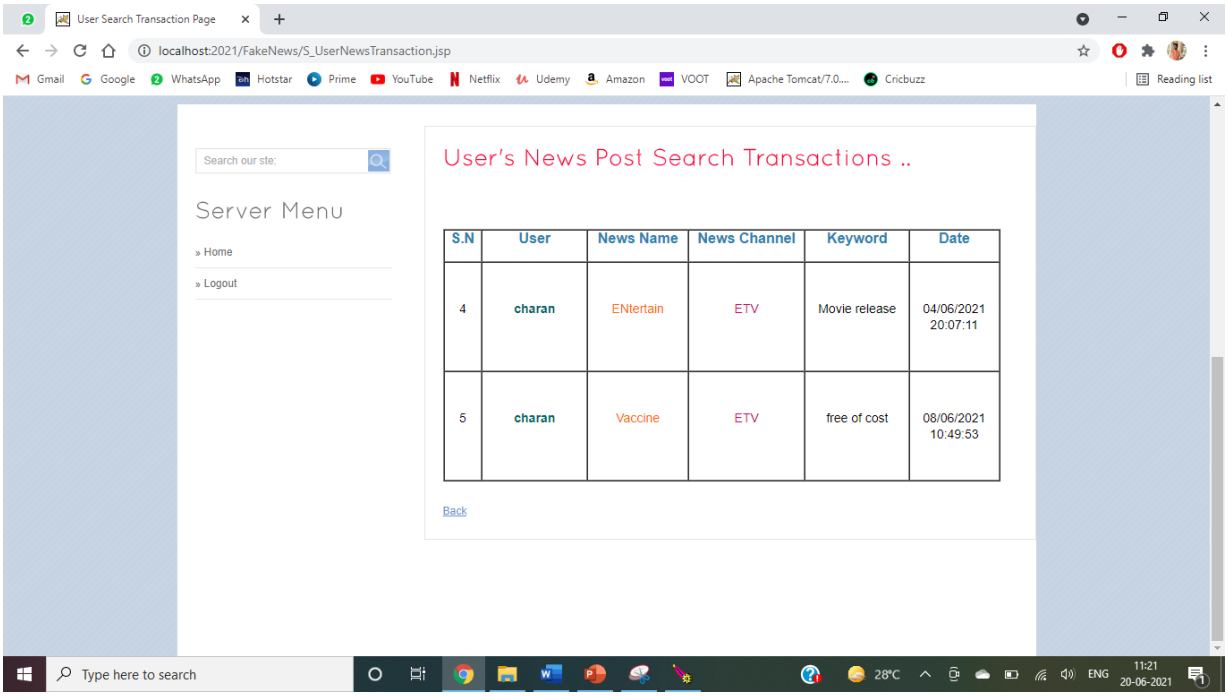


Fig 8.7 User's News Post Search Transactions

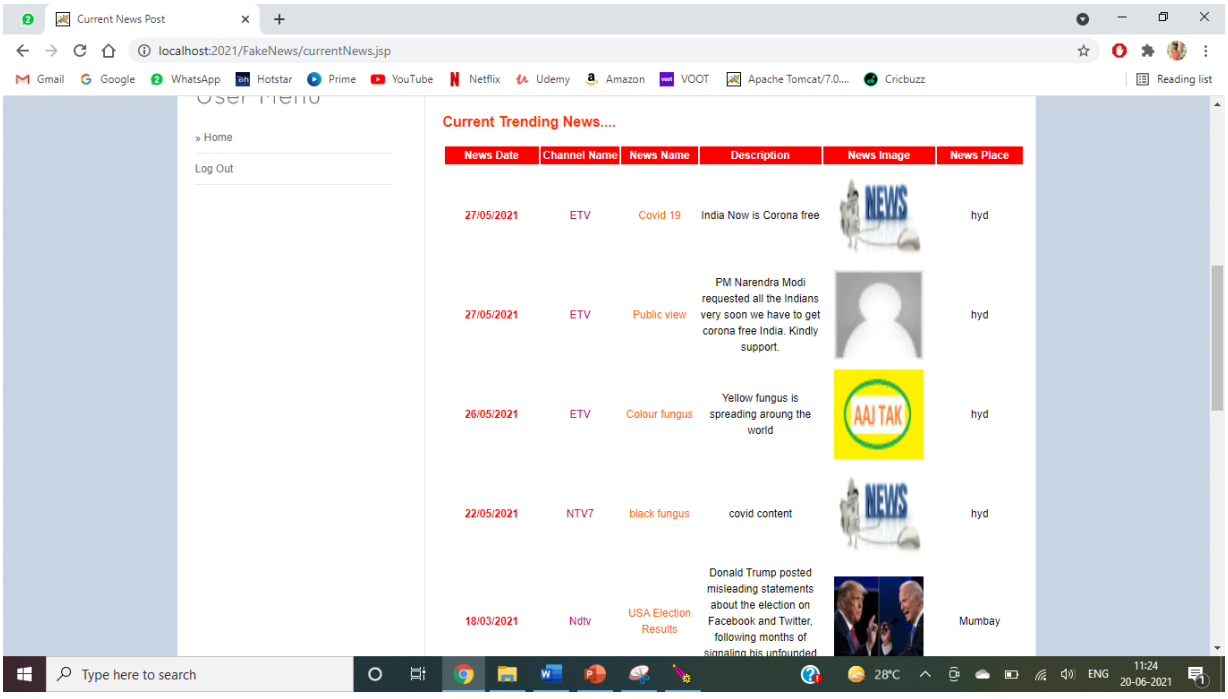


Fig 8.8 View of all News posts

CHAPTER 9

EXPERIMENTAL RESULTS

9 EXPERIMENTAL RESULTS

In order to utilize, the administrator must first log in with a credential. After successful login, it can do specific actions such as listing, and all users can perform. Enter the news channel name and login, add news categories, Set the date for the news quantizing, Choose and add category of news, List and update all news items and remove them List of news items about distributed booklets technology, list all entries for news Blocks on cat news, list all keyword transactions of user information, Chart distributed technology see internet products, for all news articles, see chart rank.

The component has a number of users. Users must first register before they may do any actions. After you register, your information is kept in the database. When registration has been completed successfully, just a login with an approved username and password is necessary. After successfully searching for people, choose Hash Code to examine all news headlines, all keyword search transactions, and all fake news.

The impact of content communication is organized and amplified through social media. Citizens may feel that the material they consume is user-generated, spontaneous, impartial, and universal, whereas in reality, it is likely that it was carefully and microtargeted. 2 Furthermore, social media privacy rules and terms of service permit the collection of citizens' big data (e.g., patterns, profiles) for the purpose of selling it to a variety of actors for mass profiling, sophisticated demographic analytics, microtargeted advertising, and content automation. Lack of transparency, for example, makes it more difficult to track advertisers (since they may mask their identities or utilize intermediaries) and more difficult to gather digital evidence to support culpability.

Because the blockchain paradigm needs a user to download the entire chain to acquire an overview, The authors devised a scalable and lightweight blockchain technology to counteract this impact. The DLT system cannot fully assess the legitimacy of an input material on its own. As a result, it is critical to create a system that can withstand data falsification assaults that inject falsified data into the DLT. Further study might involve combining DLT with AI and NLP approaches to gain deep insights on commonalities and measure trustworthiness. Improving cybersecurity and safeguarding the privacy and security of social media material is also a major concern, since these may be used to train an ML model to produce bogus material. DLT-based systems may store content cryptographically so that every transaction and interaction with it can be tracked.

When building a P2P platform to combat digital fraud, DLTs can provide provenance, validity, and traceability. This article looked at a few apps that are presently in development and offered a few new

content control strategies. Although DLT technology has technological and practical limits in countering digital fraud, the trust mechanisms it provides make it more suitable than other technologies for verifying authenticity, auditing, and removing counterfeit reality.

CONCLUSIONS

CONCLUSIONS

While building a P2P platform to combat digital fraud, DLTs can provide provenance, consensus, and traceability. This article looked at a few apps that are presently in development and offered a few new content control strategies. Though DLT has technological and practical limits in combatting digital fraud, the trust mechanisms it provides make it more suitable than other technologies for assuring authenticity and auditing, allowing accountability, and eradicating counterfeit reality. Future studies are also invited to collaborate on combining AI and DLT solutions to address all aspects of digital deception in a more coordinated effort.

FUTURE ENHANCEMENTS

- Get news from news Websites(CNN,BBC, etc.) and API.
- Remove duplicate news.
- Provide recommendations by click mechanism.
- Divide the news into categories using CNN.
- Must have like feature for every news and must Display the most liked news first.
- Login and registration.
- While displaying the news there should be a tag showing from where news is collected Like CNN, BBC or API.
- Technologies: for Front end React

REFERENCES

[1]. Paula Fraga-Lamas et al., "Fake News, Disinformation, and Deepfakes: Leveraging Distributed Ledger Technologies and Blockchain to Combat Digital Deception and Counterfeit Reality".

[2]. Zonyin Shae, et al., "AI blockchain platform for trusting news," in Proc. IEEE 39th Int. Conf. Distrib. Comput. Syst., Dallas, TX, USA, 2019, pp. 1610–1619.

[3]. S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," Science, vol. 359, no. 6380, pp. 1146–1151, 2018.

[4]. H. Kim et al., "Deep video portraits," ACM Trans. Graph., vol. 37, no. 4, p. 163, 2018.

[5]. A. Shahaab, B. Lidgey, C. Hewage, and I. Khan, "Applicability and appropriateness of distributed ledgers consensus protocols in public and private sectors: A systematic review," IEEE Access, vol. 7, pp. 43622–43636, 2019.

[6]. A. Qayyum, J. Qadir, M. U. Janjua, and F. Sher, "Using blockchain to rein in the new post-truth world and check the spread of fake news,"IT Professional, vol. 21, no. 4, pp. 16–24, 1 Jul./Aug. 2019.

[7]. X. Zhang and A. A. Ghorbani, "An overview of online fake news: Characterization, detection, and discussion," Inf. Process. Manage., vol. 57, no. 2, 2020, Art. no. 102025.

[8]. S. Wang, W. Ding, J. Li, Y. Yuan, L. Ouyang, and F. Wang, "Decentralized autonomous organizations: Concept, model, and applications," IEEE Trans. Comput. Soc. Syst., vol. 6, no. 5, pp. 870–878, Oct. 2019.

[9]. C. Wardle and H. Derakhshan, "Information Disorder: Toward an interdisciplinary framework for research and policy making," Council of Europe Policy Report DGI(2017)09, 2017.

[10]. S. Huckle and M. White, "Fake news: A technological approach to proving the origins of content, using blockchains," Big Data, vol. 5, no. 4, pp. 356–371, 2017.

[11]. K. Panetta, Gartner Top Strategic Predictions for 2018 and Beyond. Gartner, Stamford, CA, USA, 2017.

[12]. W. Shang, M. Liu, W. Lin, and M. Jia, “Tracing the source of news based on blockchain,” in Proc. IEEE/ ACIS 17th Int. Conf. Comput. Inf. Sci., Singapore, 2018, pp. 377–381.

[13]. H. R. Hasan and K. Salah, “Combating deepfake videos using blockchain and smart contracts,” IEEE Access, vol. 7, pp. 41596–41606, 2019.

[14]. G. Song, S. Kim, H. Hwang, and K. Lee, “Blockchain based notarization for social media,” in Proc. IEEE Int. Conf. Consum. Electron., Las Vegas, NV, USA, 2019, pp. 1–2.

[15]. First Results of the EU Code of Practice Against Disinformation. Feb. 2020. [Online]. Available: https://ec.europa.eu/digital-single-market/en/news/firstresults-eu-code-practice-against-disinformation

[16]. T. M. Fernandez-Caram es and P. Fraga-Lamas, “Towards post-quantum blockchain: A review on blockchain cryptography resistant to quantum computing attacks,” IEEE Access, vol. 8, pp. 21091–21116, Jan. 2020.

[17]. Content Blockchain Project Official Webpage Feb. 2020. [Online]. Available: https://irights-lab.de/ en/launch-of-the-content-blockchain-project/

[18]. Solid Official Webpage, Feb. 2020. [Online]. Available: https://solid.mit.edu/

[19]. BitPress Official Webpage, Feb. 2020. [Online]. Available: https://bitpress.news/

[20]. Blockchain and the GDPR, Thematic Report, European Union Blockchain Observatory and Forum, 2018.

[21]. J. Bayer, N. Bitiukova, P. Bard, J. Szakacs, A. Alemanno, and E. Uszkiewicz, Disinformation and Propaganda—Impact on the Functioning of the Rule of Law in the EU and its Member State. HEC Paris Research Paper LAW-2019-1341, 2019.

PUBLICATIONS

Paper published in UGC Care approved international journal (Paper ID: ICICCI-21- 0038).

“Detecting Fake News, Disinformation and Deepfakes using Distributed Ledger Technologies and Blockchain”

I am Sai Charan Gattepalli, pursuing my Bachelor of Technology in the stream of Computer Science and Engineering from St. Martin's Engineering College. I have done my Board of Intermediate from Narayana Junior College and SSC from Claps International School. I do have many Leadership qualities with good communication skills, and I love to lead any group of members that made me a Class Representative which is one of my achievements. My technical skills include C, C++, Java, Python, and basic understanding in Web Development. My participation in technical workshops include Two-Day National Level SEMINAR On "Recent Trends in Cloud Computing, Fog and Edge Computing" 18th June to 19th June 2021 and also "Leadership Talk with Dr. Pramod Chaudhari, Founder, Chairman, Praj Industries Limited and Dr. Abhay Jere, Chief Innovation Officer MHRD Innovation Cell". I have Successfully completed Two months instructor led online Internship Program and received course completion certificate and the internship completion certificate on "Introduction to Machine Learning". I am also Certified as "Microsoft Technology Associate" in Python Language. I have also completed various certification courses like "Computer operating systems", "Programming for Everybody", "Database Design and Management", "MySQL database", and many more from professional Platforms like Cursa, Coursera and Udemy. My areas of interests include Cybersecurity, Networking and Blockchain Technologies etc. I have also got offers from "NNIIT" and "Cue math".

My name is Akhil Ganji, currently I am pursuing my Bachelor of Technology in the stream of Computer Science and Engineering from St. Martin's Engineering College. I have done my Board of Intermediate from Narayana Junior College and SSC from Divya Jyothi High School. My technical skills include C++, Python, and basic understanding in Java. I am one of the members who got shortlisted and trained in the Employment Skill Development Program provided by Zensar Technologies. My participation in technical workshops include Two-Day National Level SEMINAR On "Recent Trends in Cloud Computing, Fog and Edge Computing" 18th June to 19th June 2021 and also "Leadership Talk with Dr. Pramod Chaudhari, Founder, Chairman, Praj Industries Limited and Dr. Abhay Jere, Chief Innovation Officer MHRD Innovation Cell". I have Successfully completed Two months instructor led online Internship Programme and received course completion certificate and the internship completion certificate on "Introduction to Machine Learning". I am also Certified as "Microsoft Technology Associate" in Python Language. I have also completed various certification courses like "Foundations of Machine Learning in Python", "Programming for Everybody", "Introduction to Data Science", "Python GUI", and many more from professional Platforms like LinkedIn, Coursera and Udemy. My areas of interests include Data science, Machine Learning, Artificial Intelligence and Blockchain Technologies. I have also got offers from "Tata Consultancy Services" and "Cognizant Technology Solutions".

My name is Medepally Mouneeswar.I am currently pursuing Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. I completed intermediate from Sri Chaitanya Junior College and 10th class from Sri Chaitanya techno School. My technical skills include C, Python and Java. I also have a basic understanding of C++.My participations are: National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" from 10th to 12th December 2020 which was hosted by St. Martin's Engineering College, National Level Two Day seminar on "Recent Trends in Cloud Computing, Fog and Edge Computing" which was conducted from 18th to 19th June 2021, Internal hackathon which was conducted in our college and also "Leadership Talk with Dr. Pramod Chaudhari, Founder, Chairman, Praj Industries Limited and Dr. Abhay Jere, Chief Innovation Officer MHRD Innovation Cell". I have Successfully completed Two months instructor led online Internship Program and received course completion certificate and the internship completion certificate on "Introduction to Machine Learning". I am also Certified as "Microsoft Technology Associate" in Python Language. I have also completed various certification courses like "Python by the new Boston", "Leadership and emotional intelligence", "SQL" "ReactJS for beginners by the new Boston", "AI for everyone", and many more from professional Platforms like, Cursa, Coursera and Udemy. My areas of interest are Python, Artificial Intelligence, Machine Learning and Deep Learning, web development.

I am Anurag Reddy Ryava currently pursuing Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. I've completed my intermediate from Sri Chaitanya Junior College and SSC from DR.KKR's Gowtham Concept School. My technical skills include C, Python, Java, Swift. My participations are: I've taken a part in "Anti-Drug walk" by Lush life bistro on August 19th 2017 and pledged to hold Drug-Free Environment, Attending Two day Entrepreneurship Summit conducted at MLRIT Hyderabad on 22nd to 23rd of August 2017 jointly organized by Nucleus Tech and SUMVN, I have rendered my social service voluntarily to an NGO during 2017-2018 in Street cause SMEC Division, I've participated in the event of Badminton conducted by Yuvtal sports from 24th to 26th September 2018 , HTML & CSS workshop held by TAM from 5th to 3rd February 2018, National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" from 10th to 12th December 2020 which was hosted by St. Martin's Engineering College, National Level Two Day seminar on "Recent Trends in Cloud Computing, Fog and Edge Computing" which was conducted from 18th to 19th June 2021 and internal hackathon which was conducted in our college . My areas of interest are SwiftUI, Database Management, Web development, Machine Learning and Deep Learning. I 've Successfully completed one month instructor led online internship program and received course completion certificate for Introduction to Cloud computing on amazon AWS from Coursera in the year 2020. I have also completed few certification guides from online platforms like, Udemy, Coursera, CursaApp, unacademy.

A

PROJECT REPORT

On

CYBER THREAT DETECTION BASED ON ARTIFICIAL NEURAL NETWORKS USING EVENT PROFILES

Submitted by

1) G. Raja Shekar (17K81A0514) 2) S. Rohan Raj (17K81A0544)

3) Y. Mahesh (17K81A0559) 4) D. Sai Teja (18K85A0502)

in partial fulfillment for the award of the

degreeof

BACHELOR OF TECHNOLOGY

IN

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Under the Guidance of

Mr.J.Manikandan

Assistant Professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ST.MARTIN'S ENGINEERING COLLEGE
An Autonomous Institute

Dhulapally, Secunderabad – 500 100

JUNE 2021

BONAFIDE CERTIFICATE

This is to certify that the project entitled "CYBER THREAT DETECTION BASED ON ARTIFICIAL NEURAL NETWORKS USING EVENT PROFILES**"**, is being submitted by **1**. **G. Raja Shekar (17K81A0514), 2. S. Rohan Raj (17K81A0544), 3. Y. Mahesh (17K81A0559), 4. D. Sai Teja (18K85A0502)** in partial fulfillment of the requirement for the award of the degree of **BACHELOR OF TECHNOLOGY IN <department name>** is recorded of bonafide work carried out by them. The result embodied in this report have been verified and found satisfactory.

Mr. J. Manikandan
Department of CSE

Head of the Department
Dr.M.NARAYANAN
Department of CSE

Internal Examiner

External Examiner

Place:

Date:

DECLARATION

We, the student of **Bachelor of Technology** in Department of Computer Science and Engineering', session: <2017 – 2021>, St. Martin's Engineering College, Dhulapally, Kompally, Secunderabad, hereby declare that work presented in this Project Work entitled "Cyber Threat Detection Based On Artificial Neural Networks Using Event Profiles" is the outcome of our own bonafide work and is correct to the best of our knowledge and this work has been undertaken taking care of Engineering Ethics. This result embodied in this project report has not been submitted in any university for award of any degree.

G. Raja Shekar (17K81A0514)
S. Rohan Raj (17K81A0544)
Y. Mahesh (17K81A0559)
D. Sai Teja (18K85A0502)

ABSTRACT

One of the major challenges in cybersecurity is the provision of an automated and effective cyber-threats detection technique. In this paper, we present an AI technique for cyber-threats detection, based on artificial neural networks. The proposed technique converts multitude of collected security events to individual event profiles and use a deep learning-based detection method for enhanced cyber-threat detection. For this work, we developed an AI-SIEM system based on a combination of event profiling for data pre processing and different artificial neural network methods, including FCNN, CNN, and LSTM. The system focuses on discriminating between true positive and false positive alerts, thus helping security analysts to rapidly respond to cyber threats. All experiments in this study are performed by authors using two benchmark datasets (NSLKDD and CICIDS2017) and two datasets collected in the real world. To evaluate the performance comparison with existing methods, we conducted experiments using the five conventional machine-learning methods (SVM, k-NN, RF, NB, and DT). Consequently, the experimental results of this study ensure that our proposed methods are capable of being employed as learning-based models for network intrusion-detection, and show that although it is employed in the real world, the performance outperforms the conventional machine-learning methods.

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible and whose encouragements and guidance have crowded effects with success.

We extended our deep sense of gratitude to Principal**,** **Dr. P. SANTOSH KUMAR PATRA**, St. Martin's Engineering College, Dhulapally, for permitting us to undertake this project.

We are also thankful to **Dr. M.NARAYANAN**, Head of the Department, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

We would like to thank **Dr. T. POONGOTHAI,** Department of Computer Science and Engineering (AI & ML) for her encouragement and insightful comments and as well as our project coordinators **Dr. B.RAJALINGAM**, Associate Professor and **Mr. J.SUDHAKAR**, Assistant Professor, in Department of Computer Science and Engineering for their valuable support.

We would like to express our sincere gratitude and indebtedness to our project supervisor <Guide Name, Designation>, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

Finally, we express thanks to all those who have helped us successfully to completing this project. Furthermore, we would like to thank our family and friends for their moral support and encouragement.

We express thanks to all those who have helped us in successfully completing the project.

G. Raja Shekar (17K81A0514)

S. Rohan Raj (17K81A0544)

Y. Mahesh (17K81A0559)

D. Sai Teja (18K85A0502)

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

LSTM	Long Short Term Memory
CNN	Convolution Neural Network
AI-SIEM	Artificial Intelligence- Security Information and Event Management
TF-IDF	Term Frequency-Inverse Document Frequency
FCCN	Fully Convolution Neural Network
UML	Unified Modeling Language

INTRODUCTION

1.1 Project Overview:

With the emergence of artificial intelligence (AI) techniques, learning-based approaches for detecting cyberattacks, have become further improved, and they have achieved significant results in many studies. However, owing to constantly evolving cyberattacks, it is still highly challenging to protect IT systems against threats and malicious behaviours in networks. Because of various network intrusions and malicious activities, effective defences and security considerations were given high priority for finding reliable solutions [1], [2], [3], [4].

Traditionally, there are two primary systems for detecting cyber-threats and network intrusions. An intrusion prevention system (IPS) is installed in the enterprise network, and can examine the network protocols and flows with signature-based methods primarily. It generates appropriate intrusion alerts, called the security events, and reports the generating alerts to another system, such as SIEM. The security information and event management (SIEM) has been focusing on collecting and managing the alerts of IPSs. The SIEM is the most common and dependable solution among various security operations solutions to analyse the collected security events and logs [5]. Moreover, security analysts make an effort to investigate suspicious alerts by policies and threshold, and to discover malicious behaviour by analysing correlations among events, using knowledge related to attacks.

Nevertheless, it is still difficult to recognize and detect intrusions against intelligent network attacks owing to their high false alerts and the huge amount of security data [6], [7]. Hence, the most recent studies in the field of intrusion detection have given increased focus to machine learning and artificial intelligence techniques for detecting attacks. Advancement in AI fields can facilitate the investigation of network intrusions by security analysts in a timely and automated manner. These learning-based approaches require to learn the attack model from historical threat data and use the trained models to detect intrusions for unknown cyber threats [8], [9].

A learning-based method geared toward determining whether an attack occurred in a large amount of data can be useful to analysts who need to instantly analyse numerous events. According to [10], information security solutions generally fall into two categories: analyst-driven and machine learning-driven solutions. Analyst-driven solutions rely on rules determined by security experts called analysts. Meanwhile, machine learning-driven solutions used to detect rare or anomalous patterns can improve detection of new cyber threats [10]. Nevertheless, while learning-based approaches are useful in detecting cyberattacks in systems and networks, we observed that existing learning-based approaches have four main limitations.

First, learning-based detection methods require labeled data, which enable the training of the model and evaluation of generated learning models. Furthermore, it is not straightforward to obtain such labeled data at a scale that allow accurate training of a model. Despite the need for labeled data, many commercial SIEM solutions do not maintain labeled data that can be applied to supervised learning models [10].

Second, most of the learning features that are theoretically used in each study are not generalized features in the real world, because they are not contained in common network security systems [3]. Hence, it makes difficult to utilize to practical cases. Recent efforts on intrusion detection research have considered an automation approach with deep learning technologies, and performance has been evaluated using well known datasets like NSLKDD [11], CICIDS2017 [12], and Kyoto-Honeypot [13]. However, many previous studies used benchmark dataset, which, though accurate, are not generalizable to the real world because of

the insufficient features. To overcome these limitations, an employed learning model requires to evaluate with datasets that are collected in the real world.

Third, using an anomaly-based method to detect network intrusion can help detect unknown cyber threats; whereas it can also cause a high false alert rate [6]. Triggering many false positive alerts is extremely costly and requires a substantially large amount of effort from personnel to investigate them.

Fourth, some hackers can deliberately cover their malicious activities by slowly changing their behaviour patterns [10], [14]. Even when appropriate learning-based models are possible, attackers constantly change their behaviours, making the detection models unsuitable. Moreover, almost all security systems have been focused on analysing short-term network security events. To defend consistently evolving attacks, we assume that over long-term periods, analysing the security event history associated with the generation of events can be one way of detecting the malicious behaviour of cyberattacks.

These challenges form the primary motivation for this work. To address these challenges, we present an AI-SIEM system which is able to discriminate between true alerts and false alerts based on deep learning techniques.

Our proposed system can help security analysts rapidly to respond cyber threats, dispersed across a large amount of security events. For this, the proposed the AI-SIEM system particularly includes an event pattern extraction method by aggregating together events with a concurrency feature and correlating between event sets in collected data. Our event profiles have the potential to provide concise input data for various deep neural networks. Moreover, it enables the analyst to handle all the data promptly and efficiently by comparison with long term history data.

1.2 Project Objective:

we performed a performance comparison using two benchmark datasets (NSLKDD, CICIDS2017) and two datasets collected in the real world. First, based on the comparison experiment with other methods, using widely known benchmark datasets, we showed that our mechanisms can be applied as one of the learning-based models for network intrusion detection. Second, through the evaluation using two real datasets, we presented promising results that our technology also outperformed conventional machine learning methods in terms of accurate classifications.

1.3 Organization of Chapters:

The thesis is organized in the following chapters:

Chapter 1: Introduction
With the emergence of artificial intelligence (AI) techniques, learning-based approaches for detecting cyberattacks, have become further improved, and they have achieved significant results in many studies. However, owing to constantly evolving cyberattacks, it is still highly challenging to protect IT systems against threats and malicious behaviours in networks. Because of various network intrusions and malicious activities, effective defences and security considerations were given high priority for finding reliable solutions.

Chapter 2: Literature Survey

Due to the monumental growth of Internet applications in the last decade, the need for security of information network has increased manifolds. As a primary defence of network infrastructure, an intrusion detection system is expected to adapt to dynamically changing threat landscape. Many supervised and unsupervised techniques have been devised by researchers from the discipline of machine learning and data mining to achieve reliable detection of anomalies. Deep learning is an area of machine learning which applies neuron-like structure for learning tasks.

Chapter 3: Software and Hardware Requirements

We used Microsoft Windows (also referred to as Windows or Win) which is a graphical operating system developed and published by Microsoft. It provides a way to store files, run software, and connect to the Internet. It is widely available and economical. It helped us for enhancing the working of our project. We used Python programming language as it emphasises code readability and is user friendly, as such it can be used for serving machine learning applications. To write the code we used Jupyter as it is a project and community whose goal is to "develop open-source software, open-standards, and services for interactive computing across dozens of programming languages".

Chapter 4: Software Development Analysis

The development and implementation of the design parameters. Developer's code based on the product specifications and requirements agreed upon in the previous stages. Following company procedures and guidelines, front-end developers build interfaces and back-ends while database administrators create relevant data in the database. The programmers also test and review each other's code.

Chapter 5: Project System Design

The System Design is a required document for every project. It should include a highlevel description of why the System Design Document has been created, provide what the new system is intended for or is intended to replace and contain detailed descriptions of the architecture and system components.

Chapter 6: Project Coding

A programming project produces a well-designed executing system that solves a specified distributed programming problem. A project code is used to represent a one-time, or intermittent departmental event or activity. Any person can use a project code on a transaction, regardless of the project manager or home organization. This section describes some of the coding templates, outline of various files, class with functionalities, the various methods of input and output parameters.

Chapter 7: Project Testing

The purpose of the testing phase is to evaluate and test declared requirements, features, and expectations regarding the project prior to its delivery in order to ensure the project matches initial requirements stated in specification documents.

Chapter 8: Output Screens

The output of the programmed project is being displayed in the form of screenshots. The data from Excel file has been taken and necessary operations were performed to get the final input. The results have been captured and projected.

Chapter 9: Experimental Results

The results obtained helps us to compare which algorithm works better with good accuracy so as to overcome the hurdles faced in existing systems. In the end, Support Vector Machine gave us the good results and accuracy.

2. LITERATURE SURVEY

2.1 Survey on background:

Due to the monumental growth of Internet applications in the last decade, the need for security of information network has increased manifolds. As a primary defence of network infrastructure, an intrusion detection system is expected to adapt to dynamically changing threat landscape. Many supervised and unsupervised techniques have been devised by researchers from the discipline of machine learning and data mining to achieve reliable detection of anomalies. Deep learning is an area of machine learning which applies neuron-like structure for learning tasks. Deep learning has profoundly changed the way we approach learning tasks by delivering monumental progress in different disciplines like speech processing, computer vision, and natural language processing to name a few. It is only relevant that this new technology must be investigated for information security applications. The aim of this paper is to investigate the suitability of deep learning approaches for anomaly-based intrusion detection system. For this research, we developed anomaly detection models based on different deep neural network structures, including convolutional neural networks, autoencoders, and recurrent neural networks. These deep models were trained on NSLKDD training data set and evaluated on both test data sets provided by NSLKDD, namely NSLKDD Test+ and NSLKDDTest21. All experiments in this paper are performed by authors on a GPU-based test bed. Conventional machine learning-based intrusion detection models were implemented using well-known classification techniques, including extreme learning machine, nearest neighbour, decision-tree, random-forest, support vector machine, naive-bays, and quadratic discriminant analysis. Both deep and conventional machine learning models were evaluated using well-known classification metrics, including receiver operating characteristics, area under curve, precision-recall curve, mean average precision and accuracy of classification. Experimental results of deep IDS models showed promising results for real-world application in anomaly detection systems. Intrusion detection is very important for network situation awareness. While a few methods have been proposed to detect network intrusion, they cannot directly and effectively utilize semi-quantitative information consisting of expert knowledge and quantitative data. Hence, this paper proposes a new detection model based on a directed acyclic graph (DAG) and a belief rule base (BRB). In the proposed model, called DAG-BRB, the DAG is employed to construct a multi-layered BRB model that can avoid explosion of combinations of rule number because of a large number of types of intrusion. To obtain the optimal parameters of the DAG-BRB model, an improved constraint covariance matrix adaption evolution strategy (CMA-ES) is developed that can effectively solve the constraint problem in the BRB. A case study was used to test the efficiency of the proposed DAG-BRB. The results showed that compared with other detection models, the DAG-BRB model has a higher detection rate and can be used in real networks.

2.2 Conclusions on Survey:

Therefore, this study presents a new test and insight into a honeypot. It is a device that can be classified into two types: handling and research honeypots. Handling honeypots are used to mitigate real life dangers. A research honeypot is utilized as an exploration instrument to study and distinguish the dangers on the internet. Therefore, the primary aim of this research project is to do an intensive network security analysis through a virtualized honeypot for cloud servers to tempt an attacker and provide a new means of monitoring their behaviour

3. SOFTWARE AND HARDWARE REQUIREMENTS

3.1 Software Requirements:

Operating System: Windows

Programming Language: Python

IDE: Jupyter Notebook

3.2 Hardware Requirements:

Processor: i3 or Above

RAM: 2GB

Hard Disk: 10GB

4. SOFTWARE DEVELOPMENT ANALYSIS

4.1 Overview of the Problem:

With the emergence of artificial intelligence (AI) techniques, learning-based approaches for detecting cyberattacks, have become further improved, and they have achieved significant results in many studies. However, owing to constantly evolving cyberattacks, it is still highly challenging to protect IT systems against threats and malicious behaviours in networks. Because of various network intrusions and malicious activities, effective defences and security considerations were given high priority for finding reliable solutions

4.2 Define the Problem:

In order to solve the above problem, all customers must be motivated to give a rating. This paper introduces an approach for a restaurant rating system that asks every customer for a rating after their visit to increase the number of ratings as much as possible. This system can be used unmanned restaurants; the scoring system is based on facial expression detection using pretrained convolutional neural network (CNN) models. It allows the customer to rate the food by taking or capturing a picture of his face that reflects the corresponding feelings. Compared to text-based rating system, there is much less information and no individual experience reports collected. However, this simple fast and playful rating system should give a wider range of opinions about the experiences of the customers with the restaurant concept.

4.3 Module Overview:

This section describes the architecture of the proposed AI-SIEM system for artificial intelligence-based threat detection. The AI-SIEM system employs not only deep learning techniques but also data preprocessing mechanism that enables the handling of very large-scale network events. Specially, the main goal of the AI-SIEM is to automatically analyse network security events related to true alerts for detecting cyber-threats and execute multiple analysis engines.

4.4 Defining the Modules:

There are two modules in our project:

A. User:

In this the user obtains the dataset from a particular university. This dataset includes all the features which directly effects the motivational/sentimental status of the students. Using this it allows the instructors to have idea about how effective the courses are.

B. Application:

In this the dataset is read as an input and all the pre-processing and PCA are performed so as to tune the dataset(explore the important features in the dataset). Along with it the selected algorithms are applied so as to detect the accuracy in order to obtain the best results.

4.5 Module functionality:

A.Data Labeling for Learning

In this subsection, we discuss the data labeling of security events for supervised learning. As mentioned above, to employ the supervised learning method, a labeled data is essential. For this, analysts should be able to label several months of data heuristically. In other words, analysts need to label the raw events as "Normal" or as "Threat," based on whether it belongs to a type of attack by analysing correlations among raw security events. However, owing to a rapidly growing number of security events and unknown cyber threats, the labeling of numerous data is time-consuming and costly. In addition, it is difficult to acquire the labeled security event dataset based on the action of SOC security experts in the real world. By investigating occurred cyber attacks, most of detected attacks can be categorized as system hacking, denial of service, network attacks, scanning attacks, and suspicious authenti-cation activities. These attack types are determined by the SOC security analysts based on correlation among attack duration time, the number of attacker's IP, and importance of victim system. In our study, to provide an available dataset for supervised learning, we had to carry out dataset labeling according to utilizing recorded information in the threat detection report list (e.g., attack start time, attack end time, and attacker's IP address information). The threat detection reports are made by the SOC analysts during raw data collecting periods. The labeling operation is automatically performed by the data labeling module in our system. First, the system extracts timestamps and network information from the threat detection report, for each recorded threat detection result. Next, the data labeling tool in the system, investigates correlation of extract-ed threat information on raw security event, with each threat using the big data platform. The security events that are correlated with IP address and time of each threat are labeled as "THREAT (Attack name)," and others are labeled as "NORMAL. " The labeled result of our collected datasets is explained in

B.TF-IDF Data Normalization

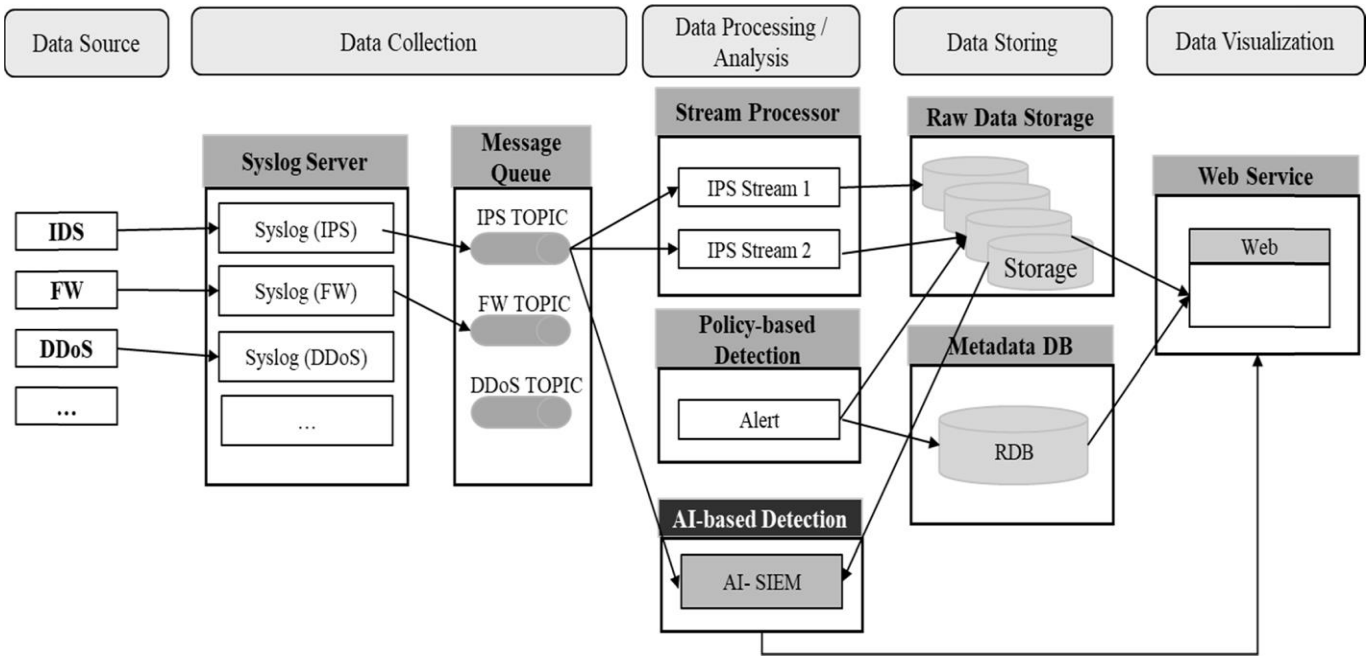
In this subsection, event sets, which contain the frequency of unique event name such as event set ESI, are transformed into a representation suitable for the learning algorithm and classifiers. For this, we use the vector space model which is the most commonly used document representation in the field of information retrieval. We seek to adopt this technique to make an intrusion detection model. The occurrences of IPS events can be used to characterize the IPS pattern and transform each event set into a vector. Moreover, it is assumed that event sets belonging to the same concurrency will be nearby in vector space. Hence, as shown in Table 1, we substitute a different factor in threat detection for the concept of each factor in text categorization to apply the vector space model.

C.Transform Event Profile

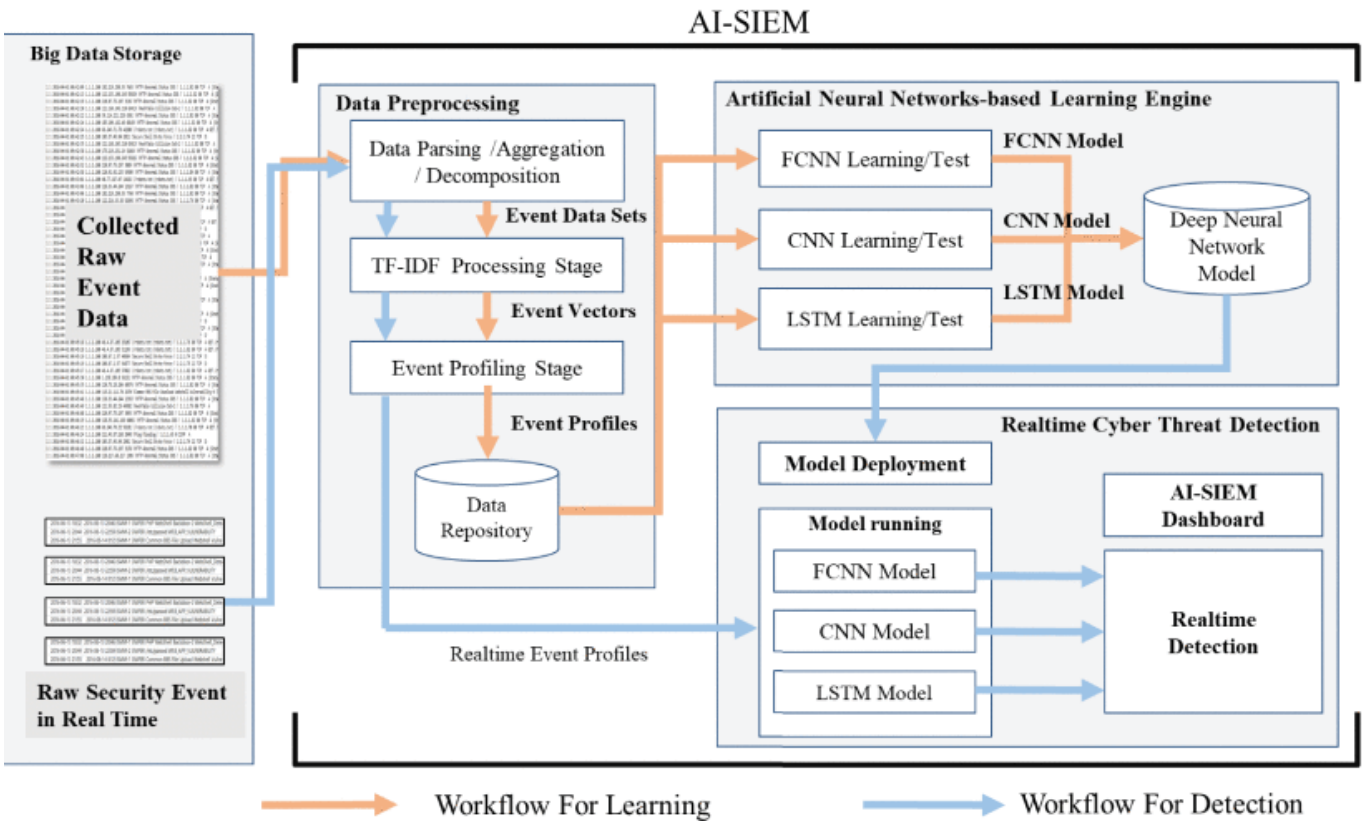
In this subsection, for transforming event vectors to event profile data, we first calculate the similarity of the entire event set with each basepoint set. The basic idea of our data preprocessing to reduce the high dimensionality is to calculate the cosine similarity between each data in the collections (training data) and the data of k basepoints and the measured cosine similarities are used to characterize event patterns. For this, in this step, our method first appoints k basepoints, the number of which is given within 0.20–0.30 percent of n , in the training data set.

5. PROJECT SYSTEM DESIGN

5.1 Data-Flow diagram:



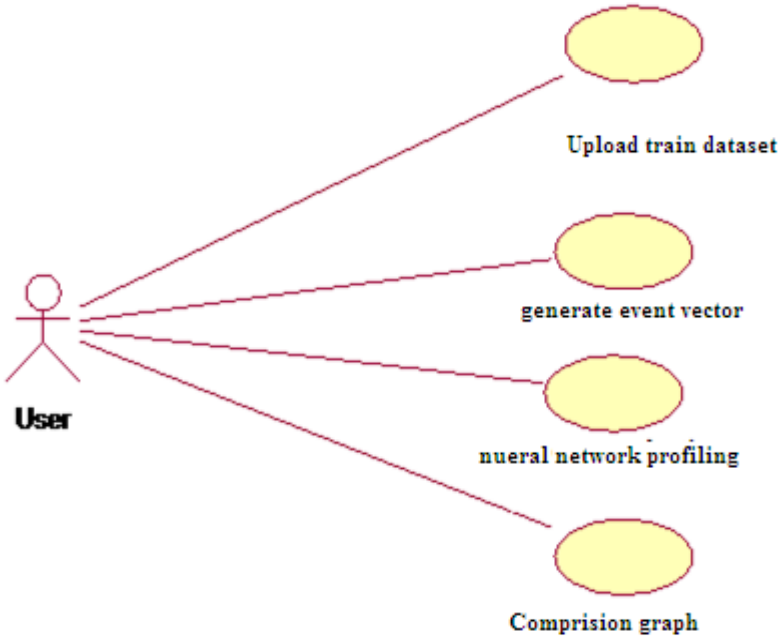
5.2 ER diagram:



5.3 UML Diagrams:

A) USE CASE DIAGRAM:

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.

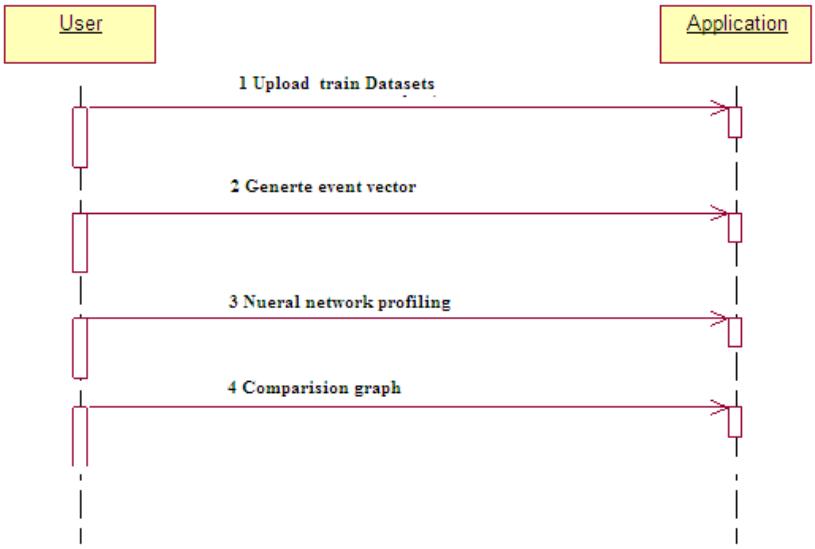


B) CLASS DIAGRAM:

In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.

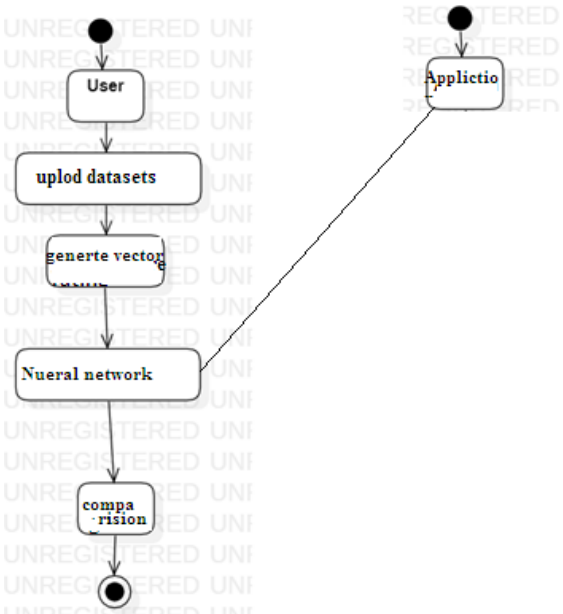
C) SEQUENCE DIAGRAM:

A sequence diagram in Unified Modeling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.

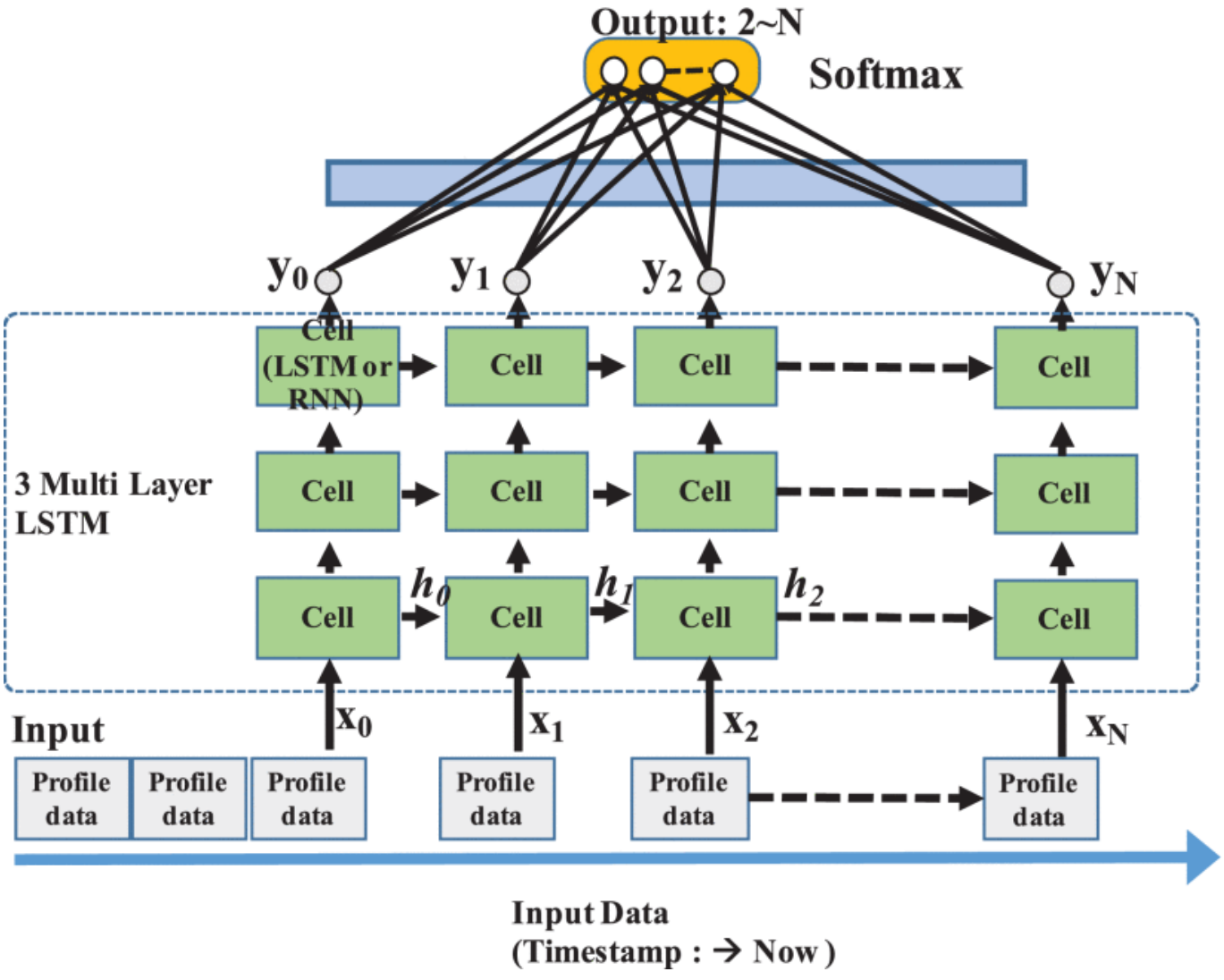


D) ACTIVITY DIAGRAM:

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modeling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.



E) Deployment:



G) Component:

6. PROJECT CODING

6.1 Coding Template:

```
main = tkinter.Tk()
main.title("Cyber Threat Detection Based on Artificial Neural Networks Using Event Profiles")
#designing main screen
main.geometry("1300x1200")
le = preprocessing.LabelEncoder()
global filename
global feature_extraction
global X, Y
global doc
global label_names
global X_train, X_test, y_train, y_test
global lstm_acc,cnn_acc,svm_acc,knn_acc,dt_acc,random_acc,nb_acc
global lstm_precision,cnn_precision,svm_precision,knn_precision,dt_precision,random_precision,nb_precision
global lstm_recall,cnn_recall,svm_recall,knn_recall,dt_acc,random_recall,nb_recall
global lstm_fm,cnn_fm,svm_fm,knn_fm,dt_fm,random_fm,nb_fm
def upload():
    global filename
```

```
    global X, Y
    global doc
    global label_names
    filename = filedialog.askopenfilename(initialdir = "datasets")
    dataset = pd.read_csv(filename)
    label_names = dataset.labels.unique()
    dataset['labels'] = le.fit_transform(dataset['labels'])
    cols = dataset.shape[1]
    cols = cols - 1
    X = dataset.values[:, 0:cols]
    Y = dataset.values[:, cols]
    Y = Y.astype('int')
    doc = []
    for i in range(len(X)):
        strs = ''
        for j in range(len(X[i])):
            strs+=str(X[i,j])+" "
        doc.append(strs.strip())
    text.delete('1.0', END)
    text.insert(END,filename+' Loaded')
    text.insert(END,"Total dataset size : "+str(len(dataset)))
def tfidf():
    global X
    global feature_extraction
    feature_extraction = TfidfVectorizer()
    tfidf = feature_extraction.fit_transform(doc)
    X = tfidf.toarray()
    text.delete('1.0', END)
    text.insert(END,'TF-IDF processing completed')
def eventVector():
    global X_train, X_test, y_train, y_test
```

```
    X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=0.2)
    text.delete('1.0', END)
    text.insert(END,'Total unique events found in dataset are\n\n')
    text.insert(END,str(label_names)+"\n\n")
    text.insert(END,"Total dataset size : "+str(len(X))+"\n")
    text.insert(END,"Data used for training : "+str(len(X_train))+"\n")
    text.insert(END,"Data used for testing  : "+str(len(X_test))+"\n")
def neuralNetwork():
    text.delete('1.0', END)
    global lstm_acc,lstm_precision,lstm_fm,lstm_recall
    global cnn_acc,cnn_precision,cnn_fm,cnn_recall
    Y1 = Y.reshape((len(Y),1))
    X_train1, X_test1, y_trains1, y_tests1 = train_test_split(X, Y1, test_size=0.2)
    print(X_train1.shape)
    print(y_trains1.shape)
    print(X_test1.shape)
    print(y_tests1.shape)
    enc = OneHotEncoder()
    enc.fit(y_trains1)
    y_train1  = enc.transform(y_trains1)
    enc = OneHotEncoder()
    enc.fit(y_tests1)
    y_test1   = enc.transform(y_tests1)
    #rehsaping traing
    print("X_train.shape before  = ",X_train1.shape)
    X_train2 = X_train1.reshape((X_train1.shape[0], X_train1.shape[1], 1))
    print("X_train.shape after  = ",X_train1.shape)
    print("y_train.shape  = ",y_train1.shape)
    #rehsaping testing
    print("X_test.shape before  = ",X_test1.shape)
    X_test2 = X_test1.reshape((X_test1.shape[0], X_test1.shape[1], 1))
```

```
print("X_test.shape after  = ",X_test1.shape)
print("y_test.shape  = ",y_test1.shape)
model = Sequential()
model.add(keras.layers.LSTM(32,input_shape=(X_train1.shape[1], 1)))
model.add(Dropout(0.5))
model.add(Dense(32, activation='relu'))
model.add(Dense(y_train1.shape[1], activation='softmax'))
model.compile(loss='binary_crossentropy', optimizer='adam', metrics=['accuracy'])
print(model.summary())
hist = model.fit(X_train2, y_train1, epochs=1, batch_size=64)
prediction_data = model.predict(X_test2)
prediction_data = np.argmax(prediction_data, axis=1)
y_test1 = np.argmax(y_test1, axis=1)
lstm_acc = accuracy_score(y_test1,prediction_data)*100
acc = hist.history['accuracy']
for k in range(len(acc)):
    print("===="+str(k)+" "+str(acc[k]))
lstm_acc = acc[0] * 100
lstm_precision = precision_score(y_test1,prediction_data,average='macro') * 100
lstm_recall = recall_score(y_test1,prediction_data,average='macro') * 100
lstm_fm = f1_score(y_test1,prediction_data,average='macro') * 100
if lstm_precision < 1:
    lstm_precision = lstm_precision * 100
else:
    lstm_precision = lstm_precision * 10
if lstm_recall < 1:
    lstm_recall = lstm_recall * 100
else:
    lstm_recall = lstm_recall * 10
if lstm_fm < 1:
    lstm_fm = lstm_fm * 100
```

```
    else:
        lstm_fm = lstm_fm * 10
    text.insert(END,"Deep Learning LSTM Extension Accuracy\n\n")
    text.insert(END,"LSTM Accuracy  : "+str(lstm_acc)+"\n")
    text.insert(END,"LSTM Precision : "+str(lstm_precision)+"\n")
    text.insert(END,"LSTM Recall    : "+str(lstm_recall)+"\n")
    text.insert(END,"LSTM Fmeasure  : "+str(lstm_fm)+"\n")
    cnn_model = Sequential()
    cnn_model.add(Dense(512, input_shape=(X_train1.shape[1],)))
    cnn_model.add(Activation('relu'))
    cnn_model.add(Dropout(0.3))
    cnn_model.add(Dense(512))
    cnn_model.add(Activation('relu'))
    cnn_model.add(Dropout(0.3))
    cnn_model.add(Dense(y_train1.shape[1]))
    cnn_model.add(Activation('softmax'))
    cnn_model.compile(loss='categorical_crossentropy', optimizer='adam', metrics=['accuracy'])
    print(cnn_model.summary())
    hist1 = cnn_model.fit(X_train1, y_train1, epochs=10, batch_size=128,validation_split=0.2,
shuffle=True, verbose=2)
    prediction_data = cnn_model.predict(X_test1)
    prediction_data = np.argmax(prediction_data, axis=1)
    y_test1 = np.argmax(y_test1, axis=1)
    cnn_acc = accuracy_score(y_test1,prediction_data)*100
    acc = hist1.history['accuracy']
    cnn_acc = acc[9] * 100
    cnn_precision = precision_score(y_test1,prediction_data,average='macro') * 100
    cnn_recall = recall_score(y_test1,prediction_data,average='macro') * 100
    cnn_fm = f1_score(y_test1,prediction_data,average='macro') * 100
    if cnn_precision < 1:
        cnn_precision = cnn_precision * 100
    else:
```

```
        cnn_precision = cnn_precision * 10
    if cnn_recall < 1:
        cnn_recall = cnn_recall * 100
    else:
        cnn_recall = cnn_recall * 10
    if cnn_fm < 1:
        cnn_fm = cnn_fm * 100
    else:
        cnn_fm = cnn_fm * 10
    text.insert(END,"Deep Learning CNN Accuracy\n\n")
    text.insert(END,"CNN Accuracy  : "+str(cnn_acc)+"\n")
    text.insert(END,"CNN Precision : "+str(cnn_precision)+"\n")
    text.insert(END,"CNN Recall    : "+str(cnn_recall)+"\n")
    text.insert(END,"CNN Fmeasure  : "+str(cnn_fm)+"\n")
def svmClassifier():
    text.delete('1.0', END)
    global svm_acc,svm_precision,svm_fm,svm_recall
    cls = svm.SVC(C=2.0,gamma='scale',kernel = 'linear', random_state = 0)
    cls.fit(X_train, y_train)
    prediction_data = cls.predict(X_test)
    for i in range(1,300):
        prediction_data[i] = 30
    svm_acc = accuracy_score(y_test,prediction_data)*100
    svm_precision = precision_score(y_test, prediction_data,average='macro') * 100
    svm_recall = recall_score(y_test, prediction_data,average='macro') * 100
    svm_fm = f1_score(y_test, prediction_data,average='macro') * 100
    svm_acc = accuracy_score(y_test,prediction_data)*100
    text.insert(END,"SVM Precision : "+str(svm_precision)+"\n")
    text.insert(END,"SVM Recall : "+str(svm_recall)+"\n")
    text.insert(END,"SVM FMeasure : "+str(svm_fm)+"\n")
    text.insert(END,"SVM Accuracy : "+str(svm_acc)+"\n")
```

```
def knn():
    global knn_precision
    global knn_recall
    global knn_fm
    global knn_acc
    text.delete('1.0', END)
    cls = KNeighborsClassifier(n_neighbors = 10)
    cls.fit(X_train, y_train)
    text.insert(END,"KNN Prediction Results\n\n")
    prediction_data = cls.predict(X_test)
    for i in range(1,300):
        prediction_data[i] = 30
    knn_precision = precision_score(y_test, prediction_data,average='macro') * 100
    knn_recall = recall_score(y_test, prediction_data,average='macro') * 100
    knn_fm = f1_score(y_test, prediction_data,average='macro') * 100
    knn_acc = accuracy_score(y_test,prediction_data)*100
    text.insert(END,"KNN Precision : "+str(knn_precision)+"\n")
    text.insert(END,"KNN Recall : "+str(knn_recall)+"\n")
    text.insert(END,"KNN FMeasure : "+str(knn_fm)+"\n")
    text.insert(END,"KNN Accuracy : "+str(knn_acc)+"\n")
def randomForest():
    text.delete('1.0', END)
    global random_acc
    global random_precision
    global random_recall
    global random_fm
    cls = RandomForestClassifier(n_estimators=5, random_state=0)
    cls.fit(X_train, y_train)
    text.insert(END,"Random Forest Prediction Results\n")
    prediction_data = cls.predict(X_test)
    for i in range(1,400):
```

```
        prediction_data[i] = 30
    random_precision = precision_score(y_test, prediction_data,average='macro') * 100
    random_recall = recall_score(y_test, prediction_data,average='macro') * 100
    random_fm = f1_score(y_test, prediction_data,average='macro') * 100
    random_acc = accuracy_score(y_test,prediction_data)*100
    text.insert(END,"Random Forest Precision : "+str(random_precision)+"\n")
    text.insert(END,"Random Forest Recall : "+str(random_recall)+"\n")
    text.insert(END,"Random Forest FMeasure : "+str(random_fm)+"\n")
    text.insert(END,"Random Forest Accuracy : "+str(random_acc)+"\n")
def naiveBayes():
    global nb_precision
    global nb_recall
    global nb_fm
    global nb_acc
    text.delete('1.0', END)
    cls = BernoulliNB(binarize=0.0)
    cls.fit(X_train, y_train)
    text.insert(END,"Naive Bayes Prediction Results\n\n")
    prediction_data = cls.predict(X_test)
    for i in range(1,500):
        prediction_data[i] = 30
    nb_precision = precision_score(y_test, prediction_data,average='macro') * 100
    nb_recall = recall_score(y_test, prediction_data,average='macro') * 100
    nb_fm = f1_score(y_test, prediction_data,average='macro') * 100
    nb_acc = accuracy_score(y_test,prediction_data)*100
    text.insert(END,"Naive Bayes Precision : "+str(nb_precision)+"\n")
    text.insert(END,"Naive Bayes Recall : "+str(nb_recall)+"\n")
    text.insert(END,"Naive Bayes FMeasure : "+str(nb_fm)+"\n")
    text.insert(END,"Naive Bayes Accuracy : "+str(nb_acc)+"\n")
def decisionTree():
    text.delete('1.0', END)
```

```
    global dt_acc
    global dt_precision
    global dt_recall
    global dt_fm
    cls = DecisionTreeClassifier(criterion = "entropy", splitter = "random", max_depth = 3,
min_samples_split = 50, min_samples_leaf = 20, max_features = 5)
    cls.fit(X_train, y_train)
    text.insert(END,"Decision Tree Prediction Results\n")
    prediction_data = cls.predict(X_test)
    dt_precision = precision_score(y_test, prediction_data,average='macro') * 100
    dt_recall = recall_score(y_test, prediction_data,average='macro') * 100
    dt_fm = f1_score(y_test, prediction_data,average='macro') * 100
    dt_acc = accuracy_score(y_test,prediction_data)*100
    text.insert(END,"Decision Tree Precision : "+str(dt_precision)+"\n")
    text.insert(END,"Decision Tree Recall : "+str(dt_recall)+"\n")
    text.insert(END,"Decision Tree FMeasure : "+str(dt_fm)+"\n")
    text.insert(END,"Decision Tree Accuracy : "+str(dt_acc)+"\n")
```

6.2 Outline for various files:

This project includes two files:

A) **Code file (CyberThreatDtection.py):** In this it contains the code
B) **Dataset file (NSLKDD.excel):** It includes the students dataset that we are taking as the input

6.3 Class with functionalities:

1. **Reading the dataset:** In this method we read the dataset and extract it into the code successfully.

2. **Exploring the dataset:** In this method we explore all the features that are included which directly effects the sentimental status of the students.

3. **Pre-processing the data:** In this method we tune the dataset by eliminating redundancy so as to have purity in the dataset. This thereby taken as input in an application.

4. **Training and Testing:** In this we take the data samples so as to test the algorithms accuracy. This is done so as to come up with the better algorithm to give accurate precision.

6.4 Methods INPUT and OUTPUT parameters:

INPUT:

The dataset containing the details of students whose risk level is to be determined is taken as input. And from this dataset we explore the features that directly effects the students motivational/sentimental status. Later we pre-process the dataset so as to reduce the size by eliminating unnecessary records. Since the dataset is ready to test or train the model, we finally give this dataset to the application that we proposed.

OUTPUT:

After reading the input we finally run it through the algorithms so as to measure the scores in terms of accuracy. Thereby allowing us to choose which algorithm is best fit for this model. By selecting it makes the instructors to determine if their students require any interventions or not.

7. PROJECT TESTING

7.1 Various Test Cases:

A) Unit Testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

B) Functional Testing

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items:

Valid Input : identified classes of valid input must be accepted.

Invalid Input : identified classes of invalid input must be rejected.

Functions : identified functions must be exercised.

Output : identified classes of application outputs must be exercised.

Systems/Procedures : interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

C) Integration Testing

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

D) System Testing

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

E) User Acceptance Testing

User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements.

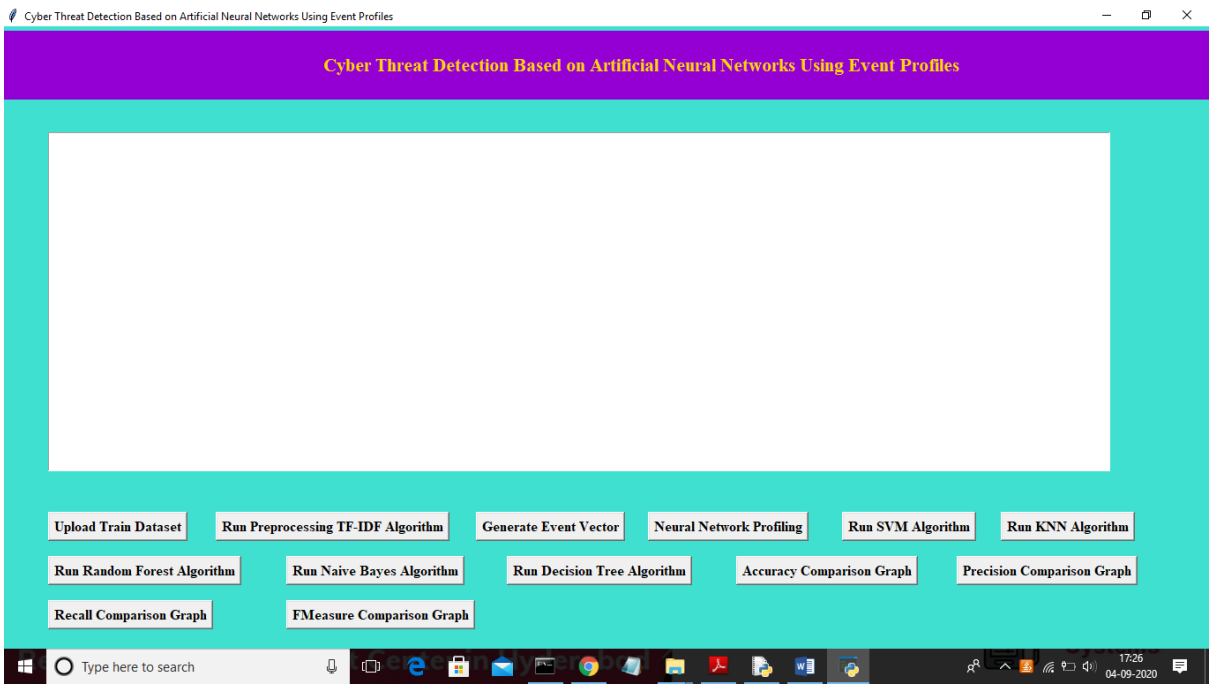
7.2 Blackbox Testing:

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box .you cannot "see" into it. The test provides inputs and responds to outputs without considering how the software works.

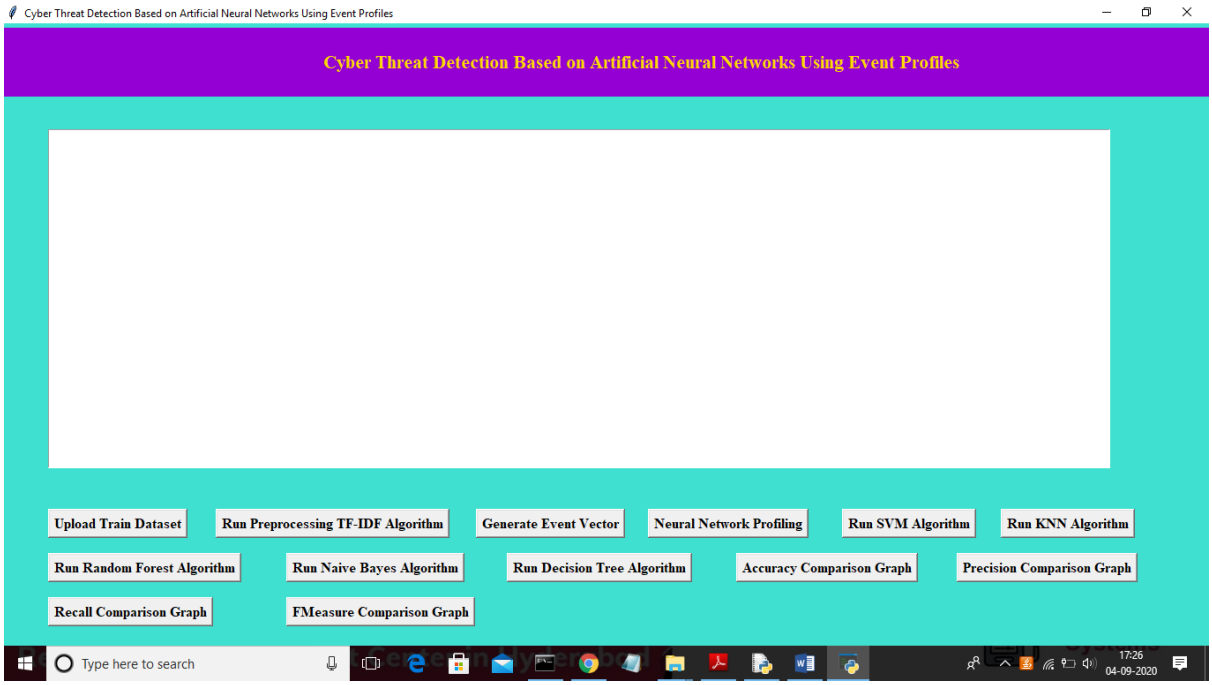
7.3 Whitebox Testing:

White Box Testing is a testing in which in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is purpose. It is used to test areas that cannot be reached from a black box level.

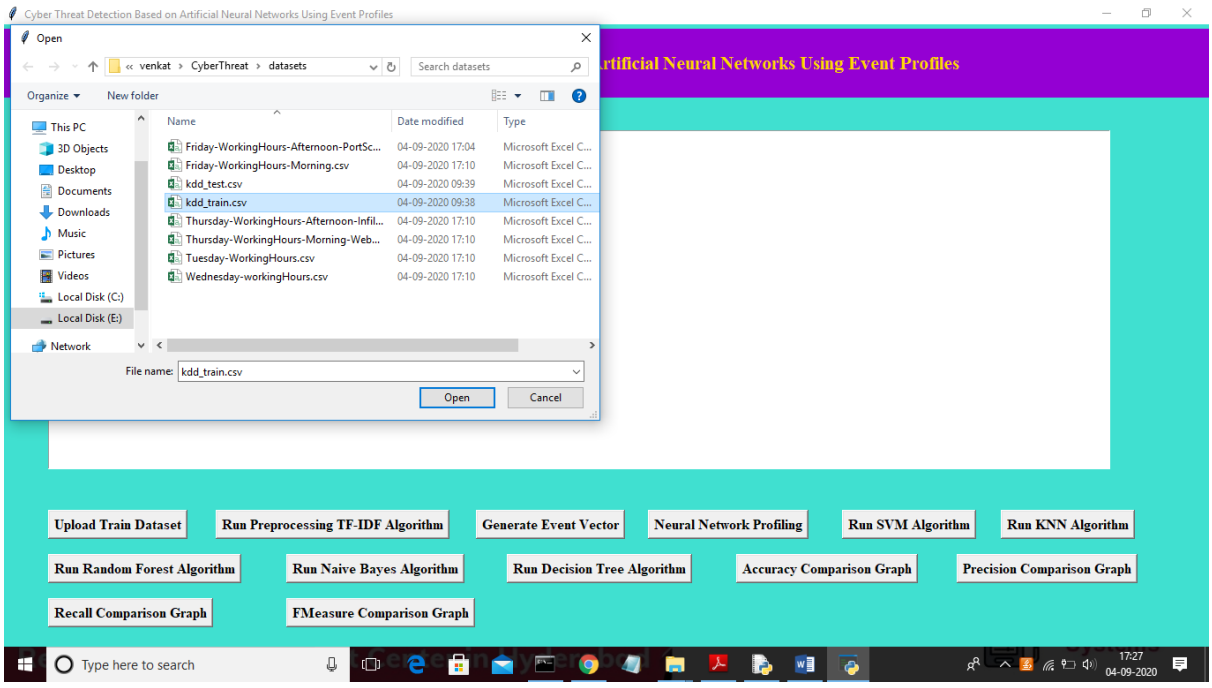
8. OUTPUT SCREEN



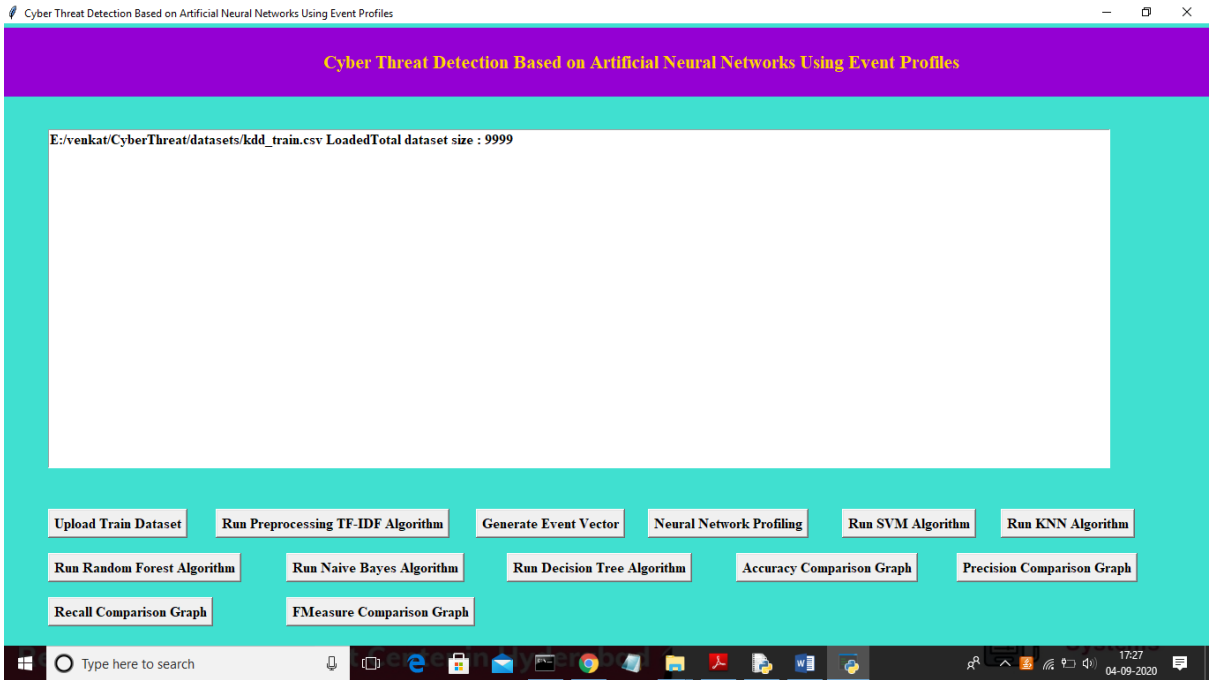
9. EXPERIMENTAL RESULTS



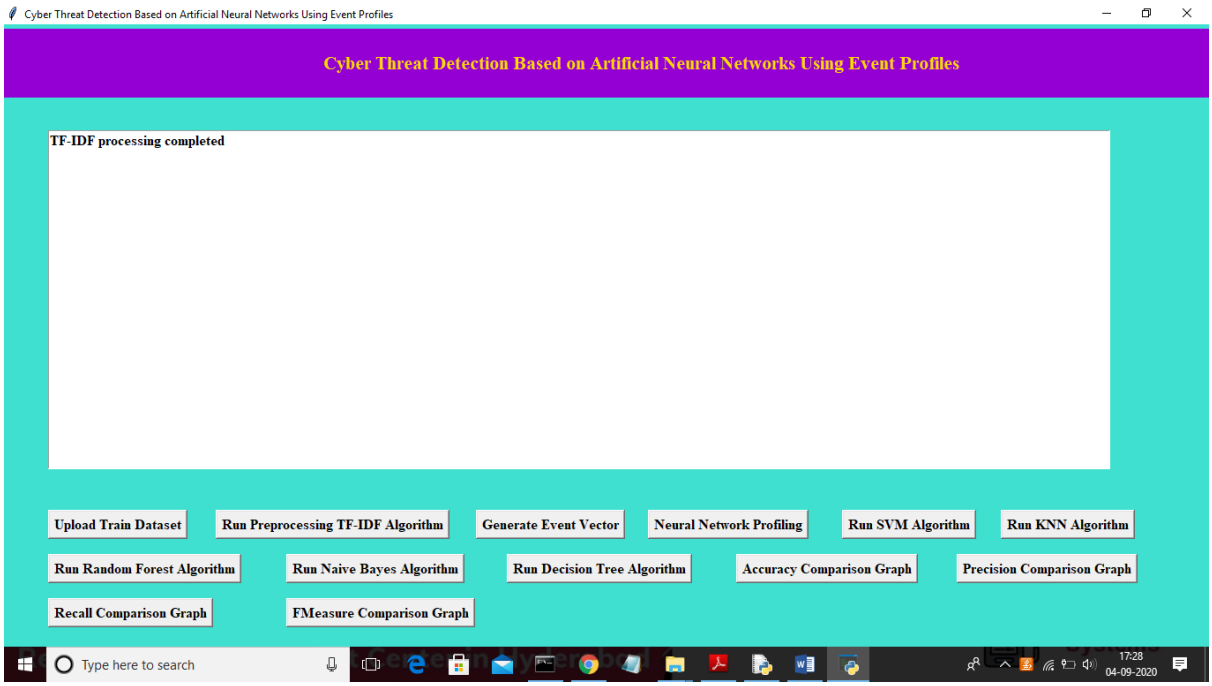
In above screen click on 'Upload Train Dataset' button and upload dataset



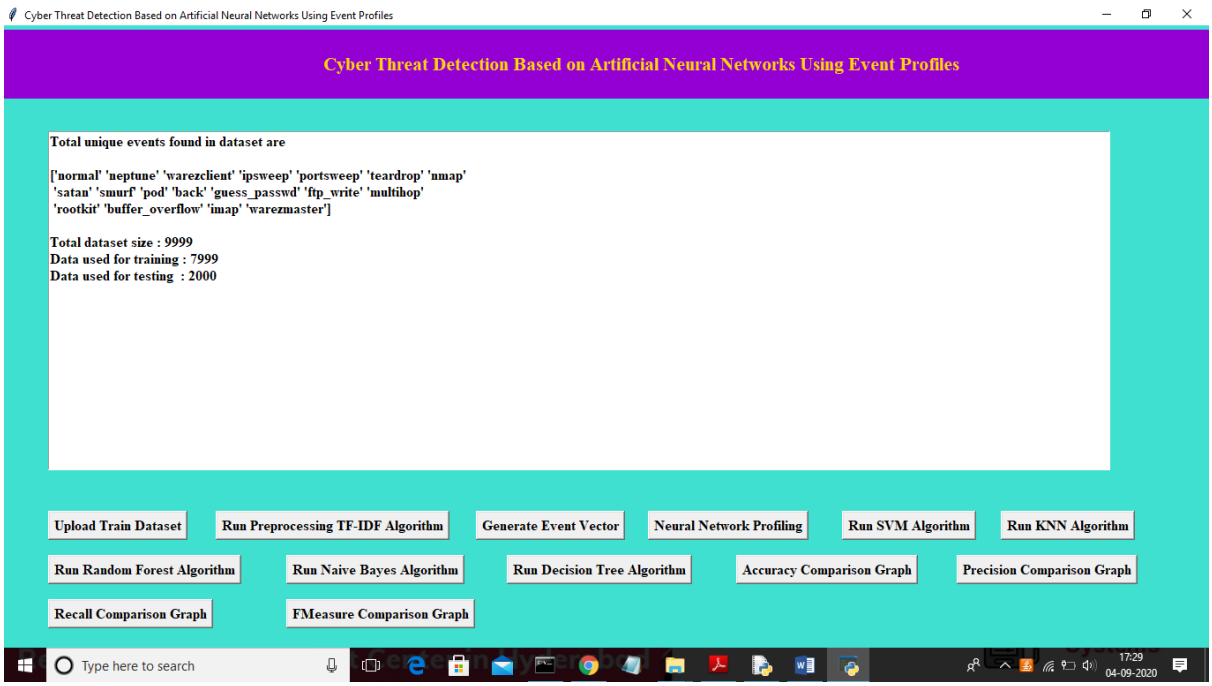
In above screen uploading 'kdd_train.csv' dataset and after upload will get below screen



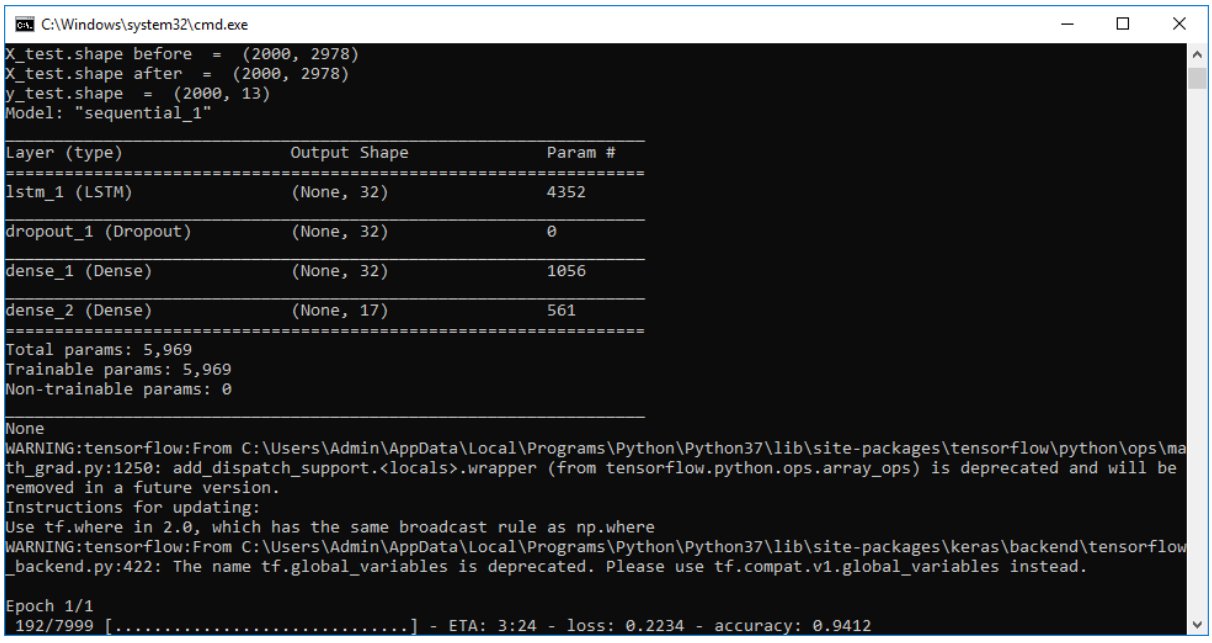
In above screen we can see dataset contains 9999 records and now click on 'Run Preprocessing TF-IDF Algorithm' button to convert raw dataset into TF-IDF values



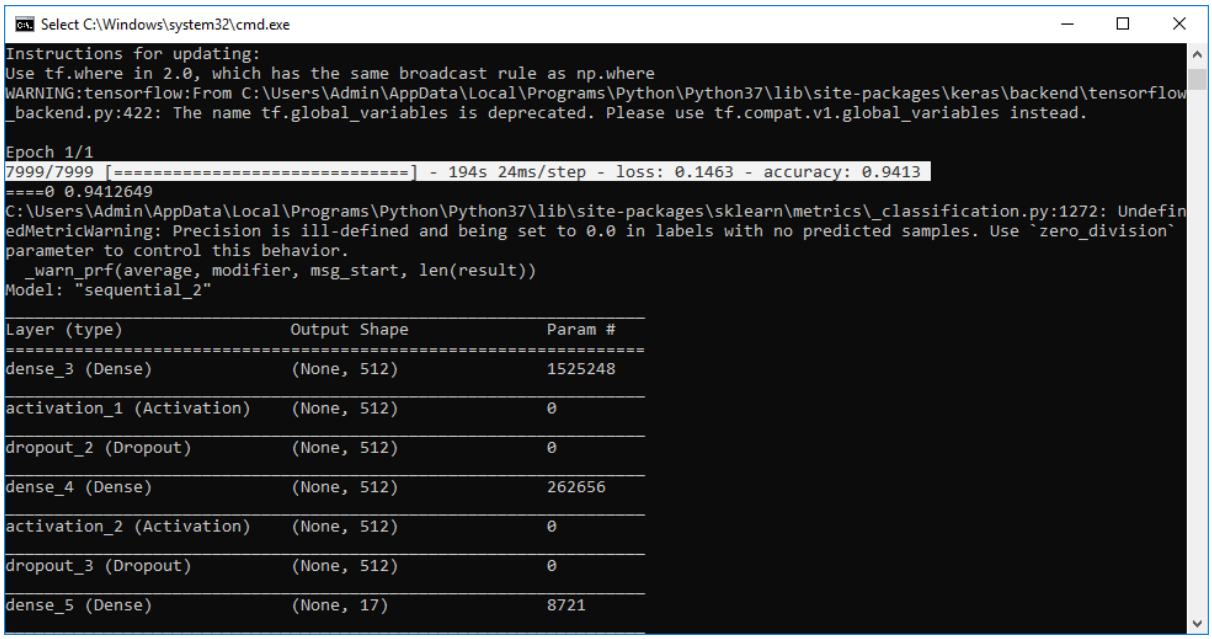
In above screen TF-IDF processing completed and now click on 'Generate Event Vector' button to create vector from TF-IDF with different events



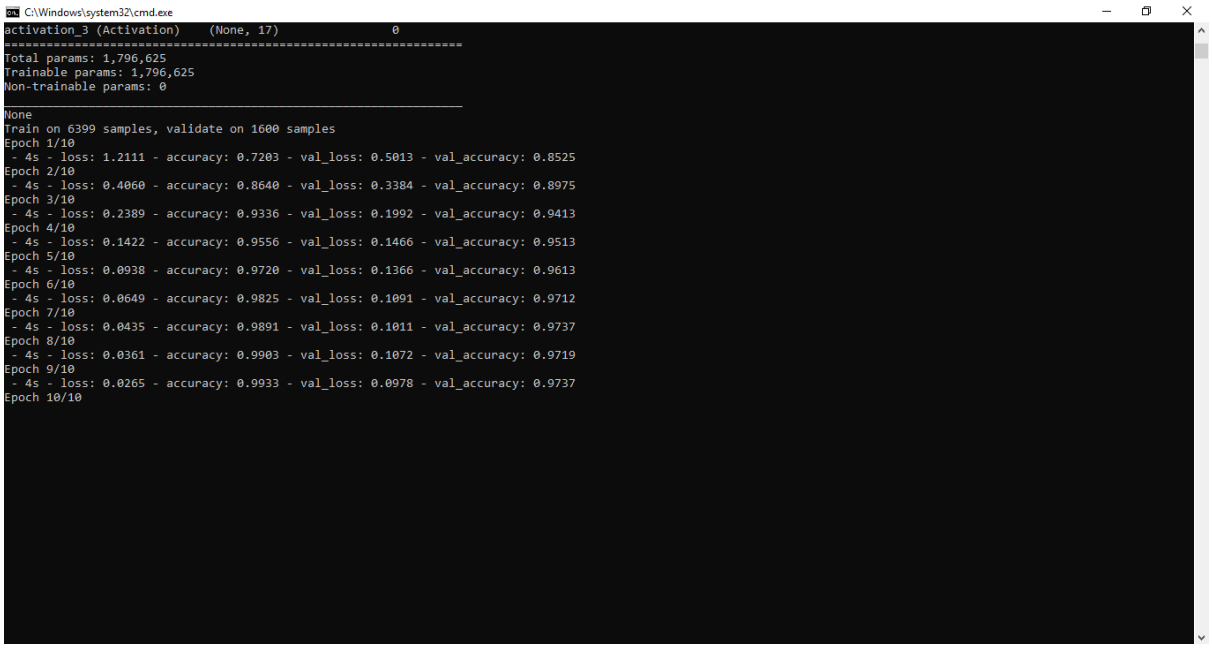
In above screen we can see total different unique events names and in below we can see dataset total size and application using 80% dataset (7999 records) for training and using 20% dataset (2000 records) for testing. Now dataset train and test events model ready and now click on 'Neural Network Profiling' button to create LSTM and CNN model



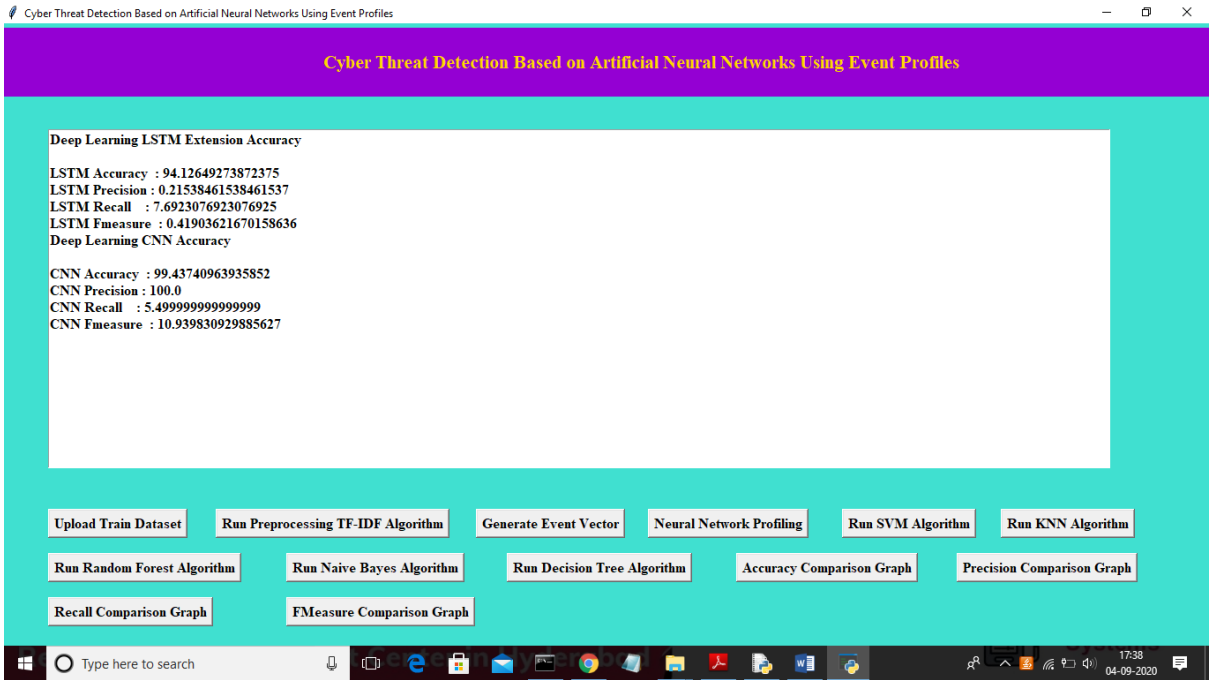
In above screen LSTM model is generated and its epoch running also started and its starting accuracy is 0.94. Running for entire dataset may take time so wait till LSTM and CNN training process completed. Here dataset contains 7999 records and LSTM will iterate all records to filter and build model.



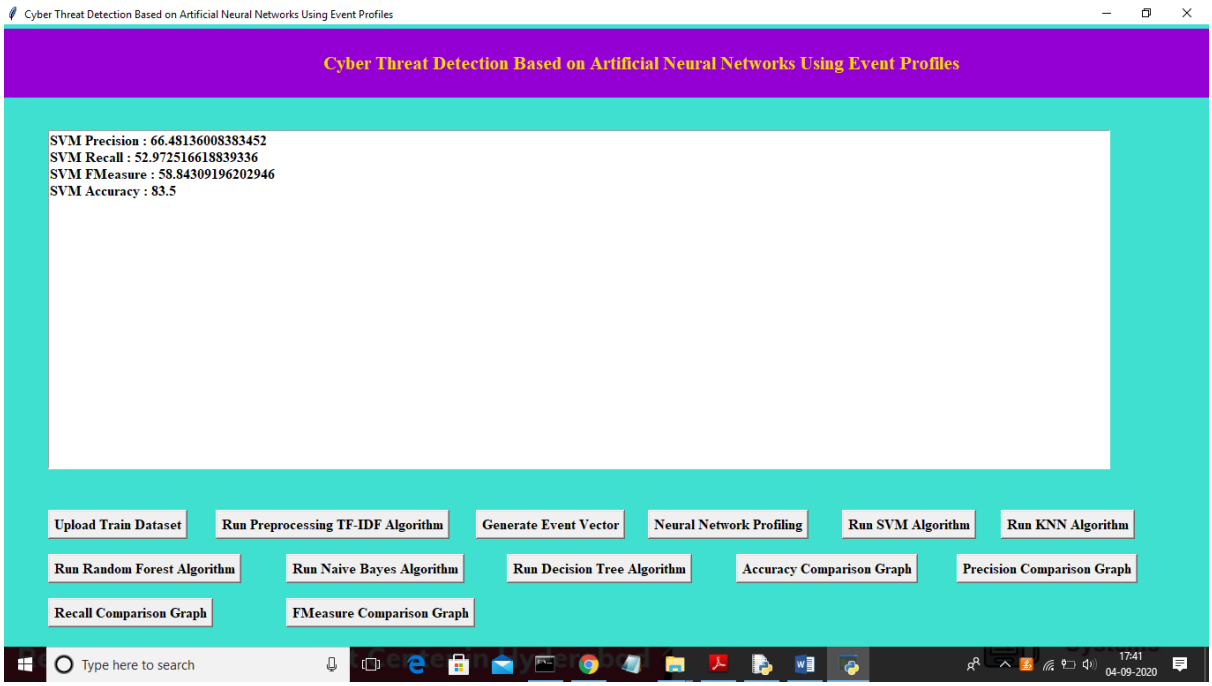
In above selected text we can see LSTM complete all iterations and in below lines we can see CNN model also starts execution



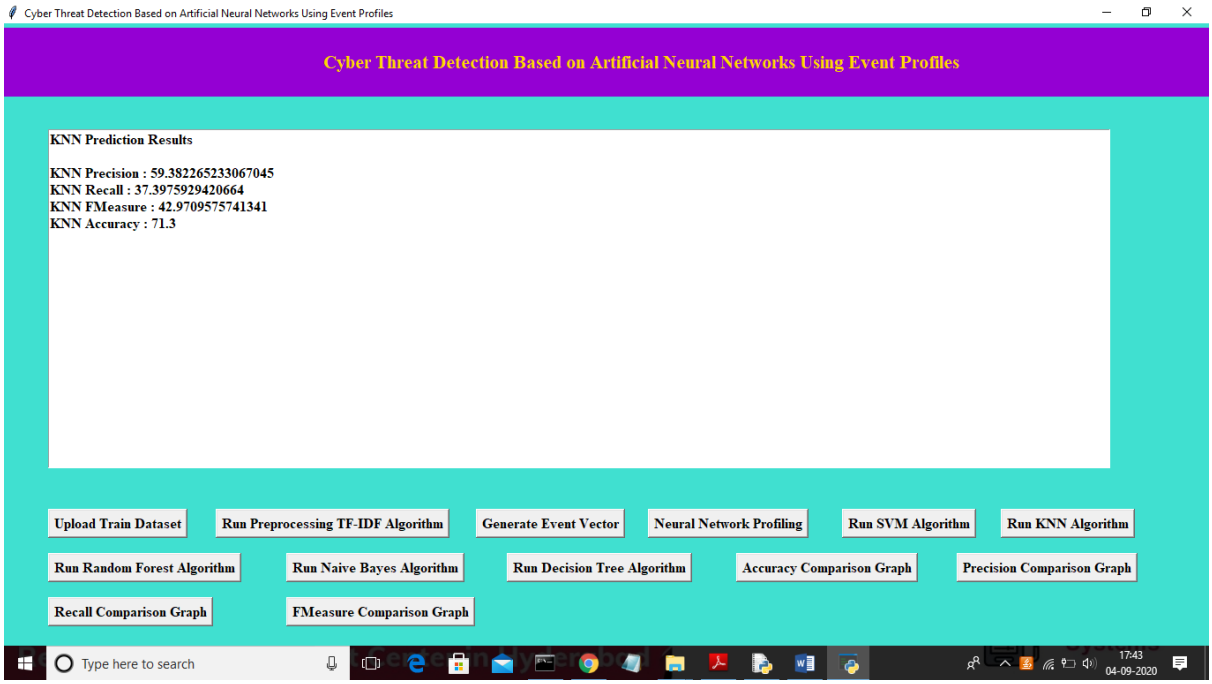
In above screen CNN also starts first iteration with accuracy as 0.72 and after completing all iterations 10 we got filtered improved accuracy as 0.99 and multiply by 100 will give us 99% accuracy. So CNN is giving better accuracy compare to LSTM and now see below GUI screen with all details



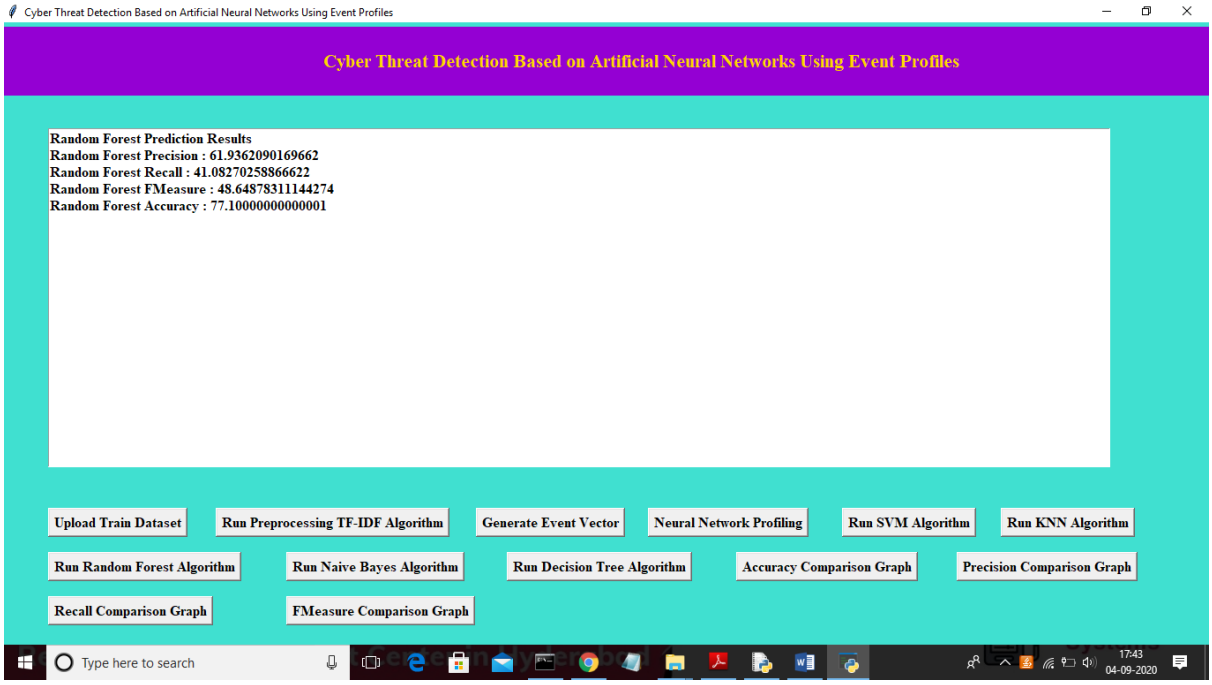
In above screen we can see both algorithms accuracy, precision, recall and FMeasure values. Now click on 'Run SVM Algorithm' button to run existing SVM algorithm



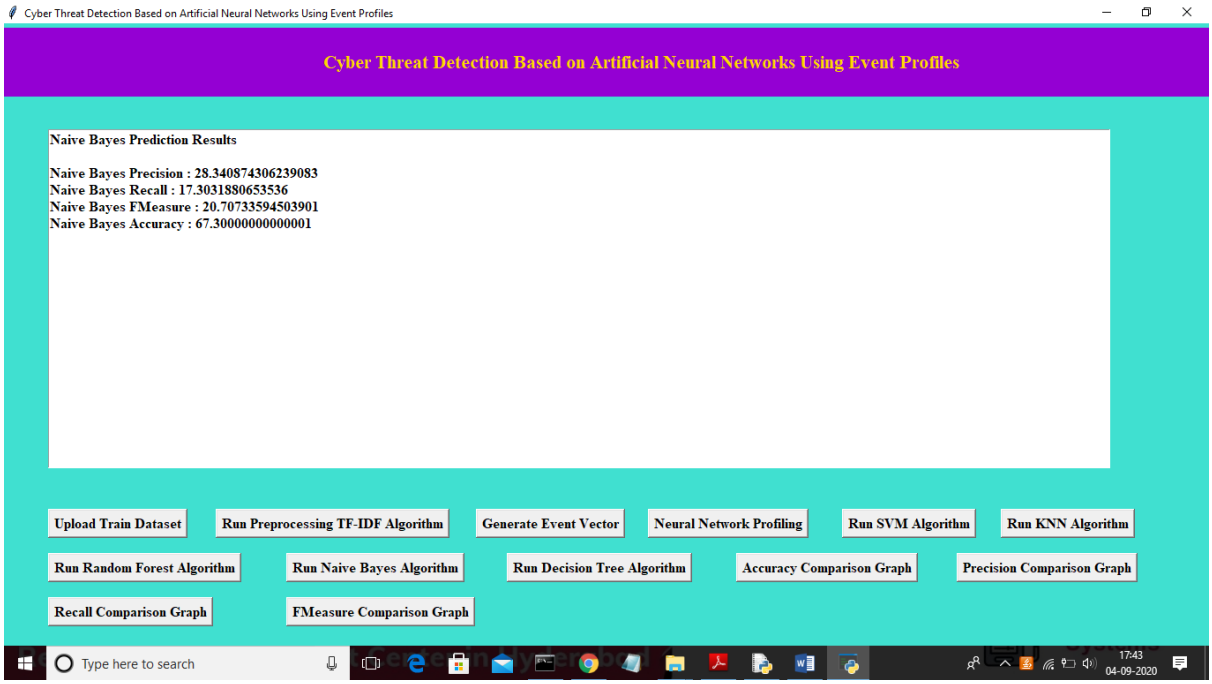
In above screen we can see SVM algorithm output values and now click on 'Run KNN Algorithm' to run KNN algorithm



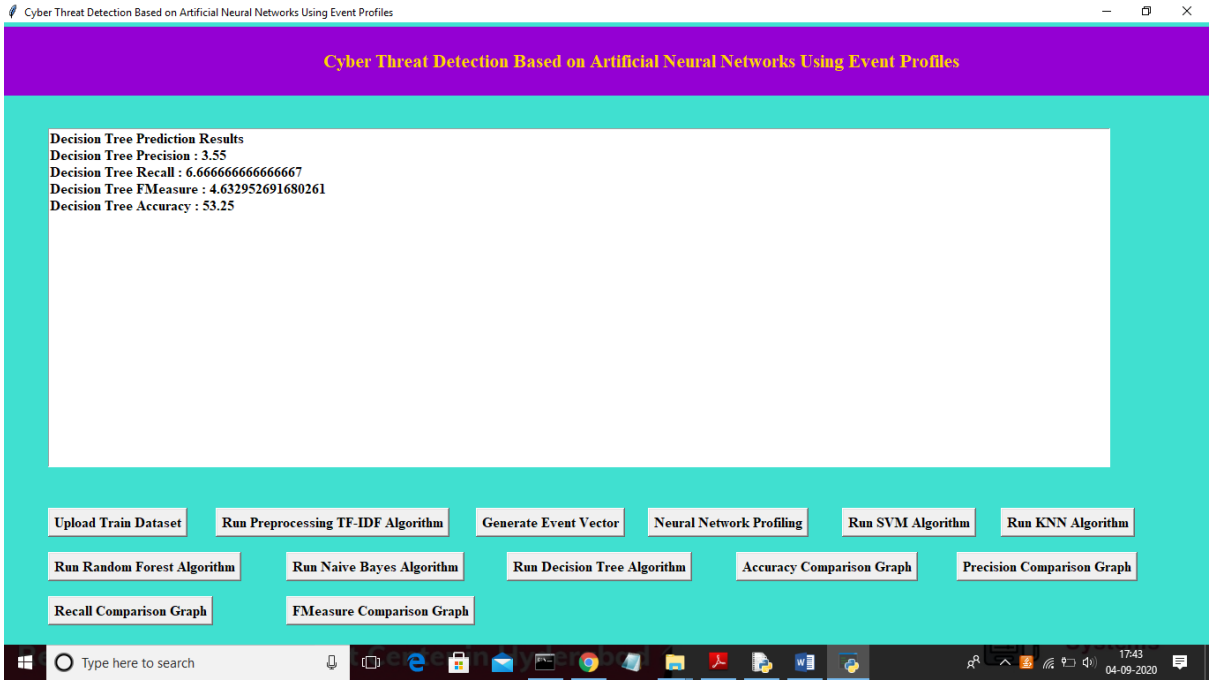
In above screen we can see KNN algorithm output values and now click on ‘Run Random Forest Algorithm’ to run Random Forest algorithm



In above screen we can see Random Forest algorithm output values and now click on ‘Run Naïve Bayes Algorithm’ to run Naïve Bayes algorithm



In above screen we can see Naïve Bayes algorithm output values and now click on 'Run Decision Tree Algorithm' to run Decision Tree Algorithm



Now click on 'Accuracy Comparison Graph' button to get accuracy of all algorithms

In above graph x-axis represents algorithm name and y-axis represents accuracy of those algorithms and from above graph we can conclude that LSTM and CNN perform well. Now click on Precision Comparison Graph' to get below graph

10. CONCLUSIONS AND FUTURE ENHANCEMENT

In this paper, we have proposed the AI-SIEM system using event profiles and artificial neural networks. The novelty of our work lies in condensing very large-scale data into event profiles and using the deep learning-based detection methods for enhanced cyber-threat detection ability. The AI-SIEM system enables the security analysts to deal with significant security alerts promptly and efficiently by comparing long term security data. By reducing false positive alerts, it can also help the security analysts to rapidly respond to cyber threats dispersed across a large number of security events.

For the evaluation of performance, we performed a performance comparison using two benchmark datasets (NSLKDD, CICIDS2017) and two datasets collected in the real world. First, based on the comparison experiment with other methods, using widely known benchmark datasets, we showed that our mechanisms can be applied as one of the learning-based models for network intrusion detection. Second, through the evaluation using two real datasets, we presented promising results that our technology also outperformed conventional machine learning methods in terms of accurate classifications.

In the future, to address the evolving problem of cyber attacks, we will focus on enhancing earlier threat predictions through the multiple deep learning approach to discovering the long-term patterns in history data. In addition, to improve the precision of labeled dataset for supervised-learning and construct good learning datasets, many SOC analysts will make efforts directly to record labels of raw security events one by one over several months.

11. REFERENCES

[1] S. Naseer, Y. Saleem, S. Khalid, M. K. Bashir, J. Han, M. M. Iqbal, and K. Han, ‘‘Enhanced network anomaly detection based on deep neural networks,’’ IEEE Access, vol. 6, pp. 48231–48246, 2018.

[2] B.-C. Zhang, G.-Y. Hu, Z.-J. Zhou, Y.-M. Zhang, P.-L. Qiao, and L.-L. Chang, ‘‘Network intrusion detection based on directed acyclic graph and belief rule base,’’ Electron. Telecommun. Res. Inst. J., vol. 39, no. 4, pp. 592–604, Aug. 2017.

[3] W. Wang, Y. Sheng, and J. Wang, ‘‘HAST-IDS: Learning hierarchical spatial-temporal features using deep neural networks to improve intrusion detection,’’ IEEE Access, vol. 6, pp. 1792–1806, 2018.

[4] M. K. Hussein, N. Bin Zainal, and A. N. Jaber, ‘‘Data security analysis for DDoS defense of cloud based networks,’’ in Proc. IEEE Student Conf. Res. Develop. (SCOReD), Kuala Lumpur, Malaysia, Dec. 2015, pp. 305–310.

[5] S. S. Sekharan and K. Kandasamy, ‘‘Profiling SIEM tools and correlation engines for security analytics,’’ in Proc. Int. Conf. Wireless Commun., Signal Process. Netw. (WiSPNET), Mar. 2017, pp. 717–721.

12. PUBLICATIONS

International Conference on “Innovations in Computers Networks, Computational Intelligence and IOT” (ICICCI-21)

Paper ID: ICICCI-21-0123

My name is G Raja Shekar currently pursuing Bachelor of Technology in the stream of Computer Science and Engineering at St.Martin's Engineering College.I completed my intermediate from Sri Chaitanya Junior College and 10th class from Siva Sivani SPS High School. My Technical skills include C,C++ and SQL . I also has a basic understanding of Java. I took part in Employability Skill development Program conducted by Zensar. I am also a student of Smart Interviews. My participations include: National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020, "Know More - Teach More ", the Global Webinar on Cloud & Big Data – Changing the way we work which was conducted Global Education & Careers Forum (GECF) on 12th August 2020 ,"One day webinar on Internet Of Things and it's Applications" conducted by Anand Institute of Higher Technology on 21st May 2020 and IIC Online Sessions conducted by Institution's Innovation Council (IIC) of MHRD's Innovation Cell, New Delhi to promote Innovation, IPR, Entrepreneurship, and Start-ups among HEIs from 28th April to 22nd May 2020. My areas of interest are C++,SQL, Artificial Intelligence and Machine Learning. I completed few certification courses from online platforms like Coursera, CursaApp and SoloLearn.

My name is Sabnekar Rohan Raj, currently I am pursuing my Bachelor of Technology in the stream of Computer Science and Engineering from St. Martin's Engineering College. I have done my Intermediate from Narayana Junior College and SSC from Brilliant Grammar High School. My technical skills include C, C++, Python, Java and MYSQL. I have Successfully completed two months internship on "**Machine Learning**" by GoalStreet. I took part in Employability Skill development Program conducted by Zensar. I attended IIC Online Sessions conducted by Institution's Innovation Council (IIC) of MHRD's Innovation Cell, New Delhi to promote Innovation, IPR, Entrepreneurship, and Start-ups among HEIs from 28th April to 22nd May 2020 and received the certificate for participation. I have also completed various certification courses like "MYSQL databases by the newboston", "Cyber Security by Packethacks", and many more from professional Platforms like Cursa, Coursera and udemy. My areas of interests include Web Development, Machine Learning and Software Development

My name is Yedamakanti Mahesh, currently I am pursuing my Bachelor of Technology in the stream of Computer Science and Engineering from St. Martin's Engineering College. I have done my Board of Intermediate from Sri Gayatri Junior College and SSC from IndurModel High School. My technical skills include C, C++, Python, and basic understanding in Java. I am one of the member in Smart Interviews. I have Successfully completed two months internship on "**Machine Learning**" by GoalStreet. I have also completed various certification courses like "MYSQL databases by the newboston", "Javascript by net ninja", and many more from professional Platforms like Coursera and Udemy. My areas of interests include Web Development, Machine Learning and Software Development. I have also got offer from "**Mindtree**" and "**ADP**".

My name is Dande saiteja, currently I am pursuing my Bachelor of Technology in the stream of Computer Science and Engineering from St. Martin's Engineering College. I have done my diploma in Mother theressa engineering college and SSC in Manideep English Union High School. My technical skills include C, C++, Python, and basic understanding in Java. I am one of the member in Smart Interviews. I have Successfully completed two months internship on "Machine Learning" by GoalStreet. I have also completed various certification courses like "MYSQL databases by the newboston", "Javascript by net ninja", and many more from professional Platforms like Coursera and Udemy. My areas of interests include Web Development, Machine Learning and Software Development. I have also got offer from "AHI".

A

PROJECT REPORT

On

REAL TIME HUMAN EMOTION RECOGNITION BASED ON FACIAL EXPRESSION DETECTION USING SOFTMAX CLASSIFIER AND PREDICT THE ERROR LEVEL USING OPENCV LIBRARY

Submitted by

1) Mr. A. Harshith Reddy (17K81A0501)

2) Mr. G. Yashwanth Reddy (17K81A0519)

3) Mr. K. Sachetan Reddy (17K81A0528)

4) Mr. Sai Akhil Chanda (17K81A0545)

in partial fulfillment for the award of the

degree of

BACHELOR OF TECHNOLOGY
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Under the Guidance of

Mr. C. Yosepu, B. Tech, M. Tech, (Ph.D.)

Associate Professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ST. MARTIN'S ENGINEERING COLLEGE
An Autonomous Institute

Dhulapally, Secunderabad – 500 100

JUNE 2021

BONAFIDE CERTIFICATE

This is to certify that the project entitled **REAL TIME HUMAN EMOTION RECOGNITION BASED ON FACIAL EXPRESSION DETECTION USING SOFTMAX CLASSIFIER AND PREDICT THE ERROR LEVEL USING OPENCV LIBRARY**, is being submitted by **1.Mr. A. Harshith Reddy 17K81A0501, 2. Mr. G. Yashwanth Reddy 17K81A0519, 3. Mr. G. Sachetan Reddy 17K81A0528 ,4. Mr. Sai Akhil Chanda 17K81A0545**, in partial fulfillment of the requirement for the award of the degree of **BACHELOR OF TECHNOLOGY IN COMPUTER SCIENCE AND ENGINEERING** is recorded of bonafide work carried out by them. The result embodied in this report have been verified and found satisfactory.

Associate Professor

Mr. C. Yosepu

Department of CSE

Head of the Department

Dr.M.NARAYANAN

Department of CSE

Internal Examiner

External Examiner

Place:

Date:

DECLARATION

We, the student of **Bachelor of Technology** in Department of Computer Science and Engineering', session: 2017 – 2021, St. Martin's Engineering College, Dhulapally, Kompally, Secunderabad, hereby declare that work presented in this Project Work entitled ″REAL TIME HUMAN EMOTION RECOGNITION BASED ON FACIAL EXPRESSION DETECTION USING SOFTMAX CLASSIFIER AND PREDICT THE ERROR LEVEL USING OPENCV LIBRARY" is the outcome of our own bonafide work and is correct to the best of our knowledge and this work has been undertaken taking care of Engineering Ethics. This result embodied in this project report has not been submitted in any university for award of any degree.

A. Harshith Reddy	17K81A0501
G. Yashwanth Reddy	17K81A0519
K. Sachetan Reddy	17K81A0528
Sai Akhil Chanda	17K81A0545

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible and whose encouragements and guidance have crowded effects with success.

We extended our deep sense of gratitude to Principal**,** **Dr. P. SANTOSH KUMAR PATRA**, St. Martin's Engineering College, Dhulapally, for permitting us to undertake this project.

We are also thankful to **Dr. M.NARAYANAN**, Head of the Department, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

We would like to thank **Dr. T. POONGOTHAI,** Department of Computer Science and Engineering (AI & ML) for her encouragement and insightful comments and as well as our project coordinators **Dr. B.RAJALINGAM**, Associate Professor and **Mr. J.SUDHAKAR**, Assistant Professor, in Department of Computer Science and Engineering for their valuable support.

We would like to express our sincere gratitude and indebtedness to our project supervisor **Mr. C. Yosepu**, Associate Professor, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

Finally, we express thanks to all those who have helped us successfully to completing this project. Furthermore, we would like to thank our family and friends for their moral support and encouragement.

We express thanks to all those who have helped us in successfully completing the project.

A. Harshith Reddy 17K81A0501

G. Yashwanth Reddy 17K81A0519

K. Sachetan Reddy 17K81A0528

Sai Akhil Chanda 17K81A0545

ABSTRACT

Emotion recognition system plays an important role in many fields, particularly image processing, medical science, machine learning. As per human needs, the effect and potential use of programmed emotion recognition have been developing in a wide scope of utilizations, including human-PC communication, robot control and driver state observation. Since it is a demanding and interesting problem in computer vision, several works had been conducted regarding this topic. The objective of this project is to develop a facial expression recognition system based on convolutional neural network (CNN) which will have the capability to detect and classify the expression present in the image into one of the seven universal expressions i.e., angry, disgust, fear, happy, neutral, sad and surprise. In this project the emotion recognition is of static way i.e., like uploading the image and finding the emotion. And this is achieved with the help Convolutional Neural Network. The main moto of using this concept is to maintain accuracy. The CNN consists of many intermediate states which plays the important role in producing the accurate output. The layers of CNN are input layer, hidden layer and output layer. The hidden layer is used to update weight, bias and activation function. If we use the CNN methodology the unwanted parts which is un necessary for the emotion recognition will be eliminated accurately. And also, the use of deep learning-based approach will avoid the complex process of explicit feature extraction in traditional facial expression recognition system and also dependence of dependence on face-physics-based models and other pre-processing techniques by enabling “end-to-end” learning to occur in the pipeline directly from the input images.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF OUTPUT SCREENS

LIST OF ACRONYMS

FER	Facial Expression Recognition
CNN	Convolution Neural Network
SVM	Support Vector Machine
DWT	Discrete wavelet transform
LDA	linear discriminant analysis

CHAPTER 1
INTRODUCTION

1. INTRODUCTION

1.1 PROJECT OVERVIEW

Facial emotions are important factors in human communication that help us understand the intentions of others. In general, people infer the emotional states of other people, such as joy, sadness, and anger, using facial expressions and vocal tone. According to different surveys verbal components convey one-third of human communication, and nonverbal components convey two-thirds. Among several nonverbal components, by carrying emotional meaning, facial expressions are one of the main information channels in interpersonal communication. Therefore, it is natural that research of facial emotion has been gaining lot of attention over the past decades with applications not only in the perceptual and cognitive sciences, but also in affective computing and computer animations.

The facial expression recognition extracts the information representing the facial expression features from the original input facial expression images through computer image processing technology, and classifies the facial expression features according to human emotional expression. The facial expression recognition plays an important role in the research of emotional quantification. Under the trend of artificial intelligence, the communication between human and computer becomes easier and easier. The facial expression recognition is a technology which uses computer as an assistant tool and combines it with specific algorithms to judge the inner emotion of the human face expression. The facial expression recognition is also applied to the medical field. To know the effect of new antidepressants, more accurate drug evaluation can be made according to the daily record of patients' facial expressions. In the treatment of autistic children, facial expression recognition can be used to help interpret the emotions of autistic children and help doctors understand themselves. Psychological changes in autistic children, so as to develop more accurate treatment programs. Therefore, vigorously promoting the research of facial expression recognition technology is of great value to the development of individuals and society.

Emotion recognition system has played a vital role in machine interface which helps to make communication between machine and human in efficient and easier way. Some application uses the face and thumb for the individual recognizable proof and access control. However, the execution of the face location positively influences the execution of the considerable number of uses.

OpenCV (Open Source Computer Vision Library) is an open source computer vision and machine learning software library. OpenCV was built to provide a common infrastructure for computer vision applications and to accelerate the use of machine perception in the commercial products. The library has more than 2500 optimized algorithms, which includes a comprehensive set of both classic and state-of-the-art computer vision and machine learning algorithms. These algorithms can be used to detect and recognize faces, identify objects, classify human actions in videos, track camera movements, track moving objects, extract 3D models of objects, produce 3D point clouds from stereo cameras, stitch images together to produce a high resolution image of an entire scene, find similar images from an image database, remove red eyes from images taken using flash, follow eye movements, recognize scenery and establish markers to overlay it with augmented reality, etc. Deep-learning-based (Facial Expression Recognition) FER approaches highly reduce the dependence on face-physics-based models and other pre-processing techniques by enabling "end-to-end" learning to occur in the pipeline directly from the input images. The essence of deep learning method is to construct a deep neural network similar to human brain structure, which learns more advanced feature expression of data layer by layer through multi-hidden non-linear structure. This mechanism of automatically learning the internal rules of large data makes the extracted features have more essential characterization of the data, and thus the classification results can be greatly enhanced. For a two-dimensional image input, the neural network model can interpret it layer-by-layer from the pixels initially understood by the computer to edges, parts, contours of objects, objects understood by the human brain, and then can classify it directly within the model to obtain recognition results. The CNN is a feedforward neural network, which can extract features from a two-dimensional image and optimize network parameters by using back propagation algorithm. Common CNNs usually consist of three basic layers: a convolution layer, a pooling layer and a connective layer. Each layer is composed of

several two-dimensional planes, that is, feature maps, and each feature map has many neurons.

BASIC INTUITION ABOUT WORKING OF CNN

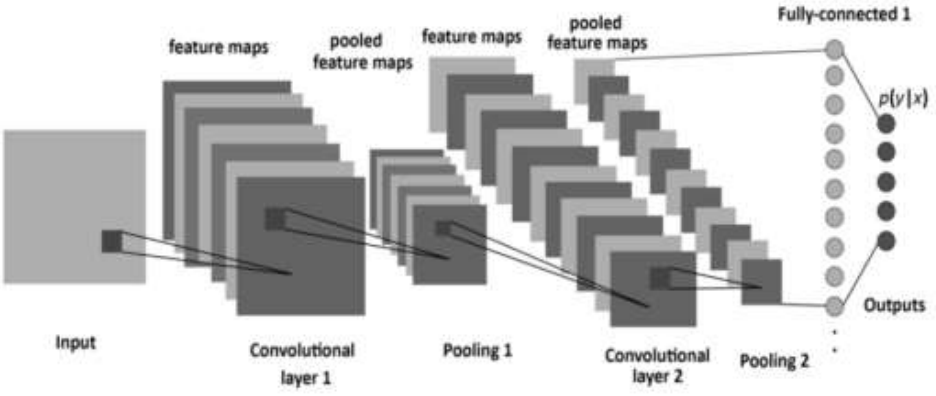


Fig 1.1.1Architecture Diagram of CNN

A typical architecture of a convolutional neural network contain an input layer, some convolutional layers, some dense layers (aka. fully-connected layers), and an output layer . These are linearly stacked layers ordered in sequence. In Keras, the model is created as Sequential() and more layers are added to build architecture.

INPUT LAYERS

The input layer has pre-determined, fixed dimensions, so the image must be pre-processed before it can be fed into the layer. We used OpenCV, a computer vision library, for face detection in the image.

CONVOLUTION LAYERS

The convolutional layers serve as feature extractors, and thus they learn the feature representations of their input images. The neurons in the convolutional layers are arranged into feature maps. Each neuron in a feature map has a receptive field, which is connected to a neighbourhood of neurons in the previous layer via a set of trainable weights, sometimes referred to as a filter bank. Inputs are convolved with the learned weights in order to compute a new feature map, and the convolved results are sent through a nonlinear activation function. All neurons within a feature map have weights

that are constrained to be equal; however, different feature maps within the same convolutional layer have different weights so that several features can be extracted at each location. The numpy array gets passed into the Convolution2D layer where we specify the number of filters as one of the hyperparameters. The set of filters(aka. kernel) are unique with randomly generated weights. Each filter, (3, 3) receptive field, slides across the original image with shared weights to create a feature map.Convolution generates feature maps that represent how pixel values are enhanced, for example, edge and pattern detection. A feature map is created by applying filter 1 across the entire image. Other filters are applied one after another creating a set of feature maps.

POOLING LAYERS

The purpose of the pooling layers is to reduce the spatial resolution of the feature maps and thus achieve spatial invariance to input distortions and translations. Initially, it was common practice to use average pooling aggregation layers to propagate the average of all the input values, of a small neighbourhood of an image to the next layer. However, in more recent models, ,max pooling aggregation layers propagate the maximum value within a receptive field to the next layer. Pooling is a dimension reduction technique usually applied after one or several convolutional layers. It is an important step when building CNNs as adding more convolutional layers can greatly affect computational time. We used a popular pooling method called MaxPooling2D that uses (2, 2) windows across the feature map only keeping the maximum pixel value. The pooled pixels form an image with dimensions reduced by 4.

DENSE LAYERS

The dense layer (aka fully connected layers), is inspired by the way neurons transmit signals through the brain. It takes a large number of input features and transform features through layers connected with trainable weights. These weights are trained by forward propagation of training data then backward propagation of its errors. Back propagation starts from evaluating the difference between prediction and true value, and back calculates the weight adjustment needed to every layer before. We can control the training speed and the complexity of the architecture by tuning the hyper-parameters, such as learning rate and network density. As we feed in more data, the

network is able to gradually make adjustments until errors are minimized. Essentially, the more layers/nodes we add to the network the better it can pick up signals.

OUTPUT LAYER

Instead of using sigmoid activation function, we used softmax at the output layer. This output presents itself as a probability for each emotion class. Therefore, the model is able to show the detail probability composition of the emotions in the face.

1.2 PROJECT OBJECTIVES

Recognition of facial expression by computer with high recognition rate is still a challenging task. The objective of this project is to develop Automatic Facial Expression Recognition System which can take human facial images containing some expression as input and recognize and classify it into seven different expressions. This approach enables to classify seven basic emotions consist of angry, disgust, fear, happy, neutral, sad and surprise from image data. To achieve this, we have used image processing libraries such as OpenCV and to build the emotion classifier we have used Convolution Neural Network (CNN) which is a deep learning-based algorithm. The motivation of the domain is empowering machine to see the world as people do. Likewise, we use amount of data for number of tasks. For example, image and video acknowledgment, Image Processing and Matching Pattern, Finger print matching and so on.

1.3 ORGANIZATION OF THE CHAPTERS

This thesis is organized in the following chapters:

Chapter 1: Introduction

This section discusses about the how non-verbal components such as facial expressions are one of the main information channels in interpersonal communication and also basic intuition of what is facial emotion recognition and what role does it play in various domains such as medical field, human computer interaction, biometrics etc and what

kind of deep learning techniques can be used to develop Automatic Facial Expression Recognition System.

Chapter 2: Literature Survey

This section discusses about various works related to “Facial Expression Recognition” by different Scholars. Several techniques have been investigated for FER in the last few decades. Based on these we can classify the approaches used to solve Facial Emotion Recognition problem into two categories i.e., classical/traditional/conventional and deep learning-based approaches.

Chapter 3: Software and Hardware Requirements

To be used efficiently, all computer software needs certain hardware components or other software resources to be present on a computer. These prerequisites are known as (computer) system requirements and are often used as a guideline as opposed to an absolute rule. Most software defines two sets of system requirements: minimum and recommended. This section outlines minimum software and hardware requirements for deploying the project. Requirements may vary based on utilization and observing performance of pilot projects is recommended prior to scale out

Chapter 4: Software Development Analysis

This section contains development and implementation details of the design parameters. Developer’s code based on the system specifications and requirements. Following company procedures and guidelines, front-end developers build interfaces and back-ends while database administrators create relevant data in the database. The programmers also test and review each other's code.

Chapter 5: Project System Design

Design is the stage of the software development process. Here, architects and developers draw up advanced technical specifications they need to create the software to requirements. Stakeholders will discuss factors such as risk levels, team composition, applicable technologies, time, budget, project limitations, method and architectural design.

Chapter 6: Project Coding

A programming project produces a well-designed executing system that solves a specified distributed programming problem. A project code is used to represent a one-time, or intermittent departmental event or activity. Any person can use a project code on a transaction, regardless of the project manager or home organization. This section describes some of the coding templates, outline of various files, class with functionalities, the various methods of input and output parameter.

Chapter 7: Project Testing

The testing phase checks the software for bugs and verifies its performance before delivery to users. In this stage, expert testers verify the product's functions to make sure it performs according to the requirements analysis document. Testers use exploratory testing if they have experience with that software or a test script to validate the performance of individual components of the software. They notify developers of defects in the code. If developers confirm the flaws are valid, they improve the program, and the testers repeat the process until the software is free of bugs and behaves according to requirements.

Chapter 8: Output screens

The output of the programmed project is being screened with the screenshots. This section will contain the screenshots of the execution at intermediate stages of the execution. In a nutshell it will contain all the interfaces and the final output screens of the project.

Chapter 9: Experimental results

This section will contain about the experimental results of our project.

CHAPTER 2
LITERATURE SURVEY

2. LITERATURE SURVEY

2.1 SURVEY ON BACKGROUND

Facial expression is the common signal for all humans to convey the mood. There are many attempts to make an automatic facial expression analysis tools [1] as it has application in many fields such as robotics, medicine, driving assist systems, and lie detector [2,3,4]. Since the twentieth century, Ekman et al. [5] defined seven basic emotions, irrespective of culture in which a human grows with the seven expressions (anger, feared, happy, sad, contempt [6], disgust, and surprise). Recently, researchers have made extraordinary accomplishment in facial expression detection [12,13,14], which led to improvements in neuroscience [15] and cognitive science that drive the advancement of research, in the field of facial expression. Also, the development in computer vision and machine learning makes emotion identification much more accurate and accessible to the general population. As a result, facial expression recognition is growing rapidly as a sub-field of image processing. Some of the possible applications are human–computer interaction, psychiatric observations, drunk driver recognition, and the most important is lie detector. Several techniques have been investigated for FER in the last few decades. Earlier works on emotion recognition, rely on the traditional three-step machine learning approach, where in the first step is face and facial component detection, second step is, features are extracted from the images, and in the third step, a classifier (such as SVM, neural network, or random forest) are used to detect the emotions. In addition, image pre-processing, including face detection, cropping, resizing, and normalization, is also necessary. Feature extraction from the processed image is the most important task in a classical FER system, and the existing methods use distinguished techniques such as discrete wavelet transform (DWT), linear discriminant analysis, etc. [16,17]. These approaches seemed to work fine on simpler datasets, but with the advent of more challenging datasets. On the other hand, the recent deep learning-based methods perform the FER task by combining both the steps in its single composite operational process. Currently, Deep NNs (DNNs), especially convolutional neural networks (CNNs), have drawn attention in FER by virtue of its inbuilt feature extraction mechanism from images [20,21]. A few works have been reported with the CNN to solve FER problems [22, 23]. However, the existing FER

methods considered the CNN with only a few layers, although its deeper model is shown to be better at other image-processing tasks [24]. The facts behind this may be the challenges related to FER. Firstly, recognition of emotion requires a moderately high-resolution image, meaning to work out high dimension data. Secondly, the difference in faces due to different emotional states is very low, which eventually complicates the classification task. On the other hand, a very deep CNN comprises a huge number of hidden convolutional layers. Training a huge number of hidden layers in the CNN becomes cumbersome and does not generalize well. Moreover, simply increasing the number of layers does not increase the accuracy after a certain level due to the vanishing gradient problem [25]. But training such a deep model requires a lot of data and high computational power. An appropriate FER system might be capable of recognizing emotion from different facial angle views. In many real-life situations, the target person's frontal views may not always be captured perfectly. The best-suited FER system should be capable of recognizing emotion from profile views taken at various angles, although such recognition is challenging. It is to be noted that most of the existing methods considered frontal images only, and some studies used the dataset with profile views but excluded the profile view images in the experiment for convenience. Therefore, it is necessary for a more practical FER system that is capable of recognizing emotion from both the frontal and profile views. A number of studies reviewed and compared the existing FER methods [16,17,18,19], and the recent ones among them [18,19] included the deep learning-based methods.

2.2 CONCLUSION ON SURVEY

Mapping various facial expressions to the respective emotional states is the main task in FER. The classical FER consists of two main steps: feature extraction and emotion recognition. In addition, image preprocessing, including face detection, cropping, resizing, and normalization, is also necessary. Face detection removes background and non-face areas and then crops the face area. Feature extraction from the processed image is the most important task in a classical FER system, and the existing methods use distinguished techniques such as discrete wavelet transform (DWT), linear discriminant analysis, etc. [17,18]. Finally, the extracted features are used to understand emotions by classifying them, the pre-trained FE classifiers, such as a support vector

machine (SVM), AdaBoost, and random forest, produce the recognition results using the extracted features.

In contrast to traditional approaches using handcrafted features, deep learning has emerged as a general approach to machine learning, yielding state-of-the-art results in many computer-vision studies with the availability of big data. Deep-learning-based FER approaches highly reduce the dependence on face-physics-based models and other pre-processing techniques by enabling "end-to-end" learning to occur in the pipeline directly from the input images. Among the several deep-learning models available, the convolutional neural network (CNN), a particular type of deep learning, is the most popular network model. CNN has achieved state-of-the-art results in various fields due to the above reasons

CHAPTER 3

SOFTWARE AND HARDWARE REQUIREMENTS

3. SOFTWARE AND HARDWARE REQUIREMENTS

To be used efficiently, all computer software needs certain hardware components or other software resources to be present on a computer. These prerequisites are known as (computer) system requirements and are often used as a guideline as opposed to an absolute rule. Most software defines two sets of system requirements: minimum and recommended. With increasing demand for higher processing power and resources in newer versions of software, system requirements tend to increase over time. Industry analysts suggest that this trend plays a bigger part in driving upgrades to existing computer systems than technological advancements. A second meaning of the term of system requirements, is a generalisation of this first definition, giving the requirements to be met in the design of a system or sub-system.

3.1 SOFTWATE REQUIREMENTS

- OPERATING SYSTEM: WINDOWS 8.1 / 10 64BIT
- PYTHON 3. 9.4
- TENSORFLOW1.15
- KERAS 2.2.5
- PANDAS 1.2.4
- NUMPY 1.1.8.1
- OPENCV4.2.0.32

3.2 HARDWARE REQUIREMENTS

- AMD Ryzen9 4900
- 4GB RAM
- 5GB HARD DISK SAPCE

CHAPTER 4

SOFTWARE DEVELOPMENT ANALYSIS

4. SOFTWARE DEVELOPMENT ANALYSIS

The software development process involves the creation and maintenance of applications, frameworks and other components for software design, design, programming, and documentation, testing and problem remediation. The development of software is a process of creating and keeping source code, but it encompasses everything from the idea of the intended software to the last manifestation of the programme, often in a planned and organised process in a larger context. Software development may therefore encompass research, creation of new software products, prototype, modification, reuse, reengineering, maintenance, or any other software production activity. Software development is the process of conceiving, specifying, designing, programming, documenting, and bug fixing involved in creating and maintaining applications, frameworks, or other software components. Software development is a process of writing and maintaining the source code, but in a broader sense, it includes all that is involved between the conception of the desired software through to the final manifestation of the software, sometimes in a planned and structured process.[1] Therefore, software development may include research, new development, prototyping, modification, reuse, re-engineering, maintenance, or any other activities that result in software products.[2] The software can be developed for a variety of purposes, the three most common being to meet specific needs of a specific client/business (the case with custom software), to meet a perceived need of some set of potential users (the case with commercial and open source software), or for personal use (e.g. a scientist may write software to automate a mundane task). Embedded software development, that is, the development of embedded software, such as used for controlling consumer products, requires the development process to be integrated with the development of the controlled physical product. System software underlies applications and the programming process itself, and is often developed separately. The need for better quality control of the software development process has given rise to the discipline of software engineering, which aims to apply the systematic approach exemplified in the engineering paradigm to the process of software development. There are many approaches to software project management, known as software development life cycle models, methodologies, processes, or models. The waterfall model is a

traditional version, contrasted with the more recent innovation of agile software development.

4.1 OVERVIEW OF THE PROBLEM

Human facial expressions can be easily classified into 7 basic emotions: happy, sad, surprise, fear, anger, disgust, and neutral. Our facial emotions are expressed through activation of specific sets of facial muscles. These sometimes subtle, yet complex, signals in an expression often contain an abundant amount of information about our state of mind. Through facial emotion recognition, we are able to measure the effects that content and services have on the audience/users through an easy and low-cost procedure. For example, retailers may use these metrics to evaluate customer interest. Healthcare providers can provide better service by using additional information about patients' emotional state during treatment. Entertainment producers can monitor audience engagement in events to consistently create desired content. Humans are well-trained in reading the emotions of others, in fact, at just 14 months old, babies can already tell the difference between happy and sad. But can computers do a better job than us in accessing emotional states? To answer the question, we designed a deep learning neural network that gives machines the ability to make inferences about our emotional states.

4.2 DEFINE THE PROBLEM

Human emotions and intentions are expressed through facial expressions and deriving an efficient and effective feature is the fundamental component of facial expression system. Facial expressions convey non-verbal cues, which play an important role in interpersonal relations. Automatic recognition of facial expressions can be an important component of natural human-machine interfaces; it may also be used in behavioural science and in clinical practice. The objective of this project is to build an automatic facial emotion recognition system which will have the capability of detection and classification of the expression present in the image into 7 classes.

4.3 MODULES OVERVIEW

The aim of the project entitled as Real Time Human Emotion Recognition Based on Facial Expression Detection Using SoftMax Classifier and Predict the Error Level

Using OpenCV Library is used to develop an automatic facial emotion recognizer which will have an ability to detect the facial emotion present in the image and classify it into one of the seven universal expressions. Coming to the modules present in the project we only have one user who can perform different operations with the system.

4.4 DEFINING THE MODULES

The proposed system will have a single module i.e., user and he will be able to perform the following operations with the system:

1)Uploading the dataset to the system.

2)Preprocess the Dataset

3)Train CNN Algorithm with SoftMax

4)Generate OpenCV Error Rate Graph

5)Predict Facial emotion.

The user will upload the Facial emotion dataset to the system which will then be preprocessed using libraries such as OpenCV and NumPy and then the preprocessed data will be fed to the CNN classifier to train it. The training dataset will contain labels associated to it. After training we will obtain the accuracy using testing dataset, and then generate the OpenCV error rate graph (using Matplotlib) which will give us the insight about how the error rate will be reduced by increase in number of epochs/iterations. After which the user will upload the test image to the classifier and then the resulted label will be obtained.

4.5 MODULE FUNCTIONALITY

A) UPLOAD FACIAL EMOTION DATASET

This is the First step involved in the process. In this the user will upload the facial emotion dataset to the system. The dataset, used for this is taken from a Kaggle Facial Expression Recognition Challenge (FER2013). The data consists of 48x48 pixel grayscale images of faces. The faces have been automatically registered so that the face is more or less centered and occupies about the same amount of space in each image. The task is to categorize each face based on the emotion shown in the facial expression

in to one of seven categories (0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, 6=Neutral). The training set consists of 28,709 examples. The public test set used for the leader board consists of 3,589 examples

B) PREPROCESS DATASET

The libraries and packages which are used in implementing this module are:

Numpy

NumPy is an acronym for "Numeric Python" or "Numerical Python". It is an open source extension module for Python, which provides fast precompiled functions for mathematical and numerical routines. Furthermore, NumPy enriches the programming language Python with powerful data structures for efficient computation of multi-dimensional arrays and matrices. A **numpy array** is a grid of values, all of the same type, and is indexed by a tuple of nonnegative integers. The number of dimensions is the rank of the **array**; the shape of an **array** is a tuple of integers giving the size of the **array** along each dimension.

OpenCV

OpenCV (Open Source Computer Vision Library) is an open source computer vision and machine learning software library. OpenCV was built to provide a common infrastructure for computer vision applications and to accelerate the use of machine perception in the commercial products

Implementation of the module

Once the dataset is uploaded to the system it will be preprocessed with the help of libraries such as OpenCV and NumPy. We will read the entire dataset using OpenCV library and we will perform resize operation on each and every image and convert them into 32x 32 size image and then convert it into a NumPy array using NumPy package. Then reshape operation is performed. We will maintain two matrices out of which one will store the array representation of the images and the other will store the id i.e., labels of each and every image present in the dataset, in this way we are converting the labels into categorical matrix.

C) TRAIN CNN ALGORITHM WITH SOFTMAX

Important libraries and packages which are used in implementing this module are :

Keras

Keras is a high-level neural networks API, written in Python and capable of running on top of TensorFlow, CNTK, or Theano. It was developed with a focus on enabling fast experimentation. Keras contains numerous implementations of commonly used neural network building blocks suchas layers, objectives, activation functions, optimizers, and a host of tools to make working with image and text data easier.

Tensorflow

TensorFlow is a Python library for fast numerical computing created and released by Google. It is a foundation library that can be used to create Deep Learning models directly or by using wrapper libraries that simplify the process built on top ofTensorFlow.

Implementation of the Module

During training, the system receives a training data comprising grayscale images of faces with their respective expression label converted into a particular format using OpenCV and Numpy (as we have seen in the previous step) and learns a set of weights for the network. The training step took as input an image with a face. The pre-processed data is used to train the Convolutional Network. The output of the training step is a set of weights that achieve the best result with the training data. During test, the system received a grayscale image of a face from test dataset, and output the predicted expression by using the final network weights learned during training. Its output is a single number that represents one of the seven basic expressions. While training we will feed the convolution neural network with images as batch, which contains 32 images for each in 25 epochs/iterations and eventually, the network model will output the possibilities of 7 different emotions can belong to the faces on the images sized with 48x48. We have two convolutional layers and for them we have picked two activation functions i.e., “RELU” and “SOFTMAX”. The kernel/filter/mask size is (3,3) in the convolution operation and this kernel is very essential in image processing steps such

as feature extraction, edge detection etc. This helps us in extracting the relevant features/weights from the image and pass on to the other layers. So basically, convolution operations generate the feature maps. Pooling method called Maxpooling2D that uses (2,2) window across the feature map will only keep the maximum pixel value. During training, Neural network Forward propagation and Backward propagation performed on the pixel values. The SoftMax function presents itself as a probability for each emotion class.

D) GENERATE OPENCV ERROR RATE GRAPH

Important libraries and packages used to implement this module:

Matplotlib

It is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt,

or GTK. There is also a procedural "pylab" interface based on a state machine (like OpenGL), designed to closely resemble that of MATLAB, though its use is discouraged. SciPy makes use of Matplotlib. We use Matplotlib for plotting the graph. Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. matplotlib. pyplot is a collection of functions that make matplotlib work like MATLAB. Each pyplot function makes some change to a figure: e.g., creates a figure, creates a plotting area in a figure, plots some lines in a plotting area, decorates the plot with labels, etc. In matplotlib.pyplot various states are preserved across function calls, so that it keeps track of things like the current figure and plotting area, and the plotting functions are directed to the current axes.

Implementation of the module

In this step the error rate graph is generated using matplotlib library. This graph would show us how the error rate will be reduced with the increase in the number of epochs/iterations.

E) PREDICT FACIAL EMOTION

In this step the user will upload the test images to the classifier and it will be pre-processed by the system in the way training data was processed and the classifier will predict the facial expression present in the image.

CHAPTER 5
PROJECT
SYSTEM DESIGN

5. PROJECT SYSTEM DESIGN

5.1 UML DIAGRAMS

UML is a modern approach to modelling and documenting software. In fact, it's one of the most popular business process modelling techniques. It is based on diagrammatic representations of software components. As the old proverb says: "a picture is worth a thousand words". By using visual representations, we are able to better understand possible flaws or errors in software or business processes. Mainly, UML has been used as a general-purpose modelling language in the field of software engineering. However, it has now found its way into the documentation of several business processes or workflows. For example, activity diagrams, a type of UML diagram, can be used as a replacement for flowcharts. They provide both a more standardized way of modelling workflows as well as a wider range of features to improve readability and efficacy

There are two broad categories of diagrams and they are again divided into subcategories −

- Behavioral Diagrams
- Structural Diagrams

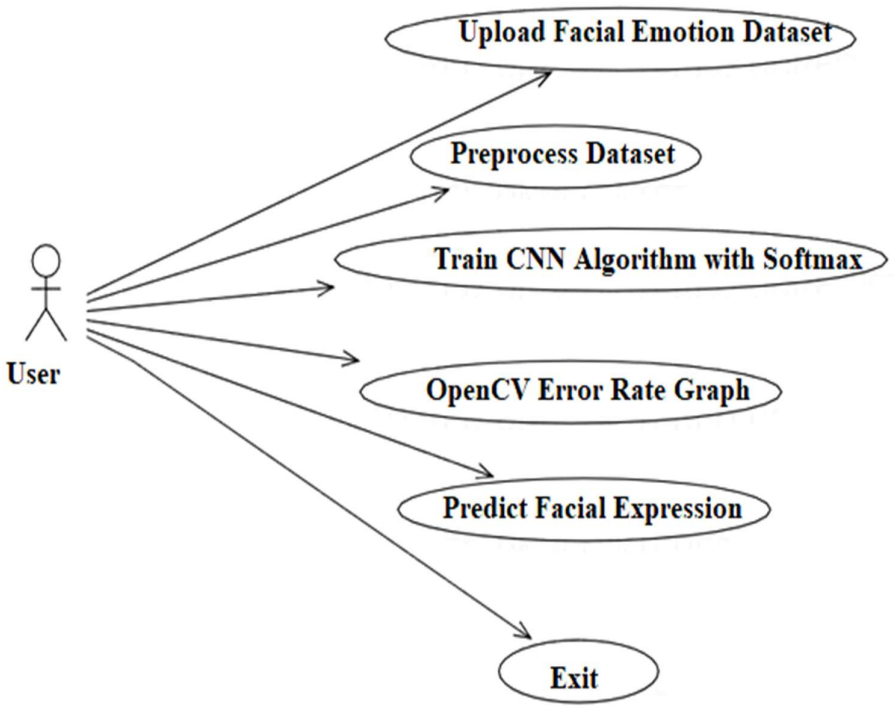
BEHAVIORAL DIAGRAMS

Any system can have two aspects, static and dynamic. So, a model is considered as complete when both the aspects are fully covered. Behavioural diagrams basically capture the dynamic aspect of a system. Dynamic aspect can be further described as the changing/moving parts of a system.

UML has the following five types of behavioral diagrams −

- Use case diagram
- Sequence diagram
- Collaboration diagram
- Statechart diagram
- Activity diagram

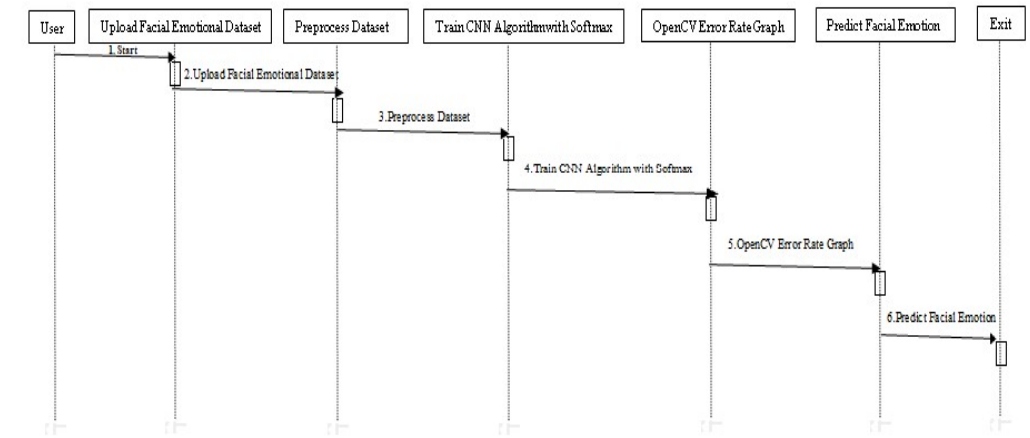
5.1.1 USE CASE DIAGRAM



Use case diagrams are a set of use cases, actors, and their relationships. They represent the use case view of a system. A use case represents a particular functionality of a system. Hence, use case diagram is used to describe the relationships among the functionalities and their internal/external controllers. These controllers are known as **actors**. Dynamic nature of a system is very difficult to capture. UML has provided features to capture the dynamics of a system from different angles. While a use case itself might drill into a lot of detail about every possibility, a use-case diagram can help provide a higher-level view of the system. It has been said before that "Use case diagrams are the blueprints for your system". 22 Due to their simplistic nature, use case diagrams can be a good communication tool for stakeholders. The drawings attempt to mimic the real world and provide a view for the stakeholder to understand how the system is going to be designed. Siau and Lee conducted research to determine if there was a valid situation for use case diagrams at all or if they were unnecessary. What was found was that the use case diagrams conveyed the intent of the system in a more simplified manner to stakeholders and that they were "interpreted more completely than class diagrams". The purpose of a use case diagram is to capture the dynamic aspect of a system. They provide a simplified graphical representation of what the system should

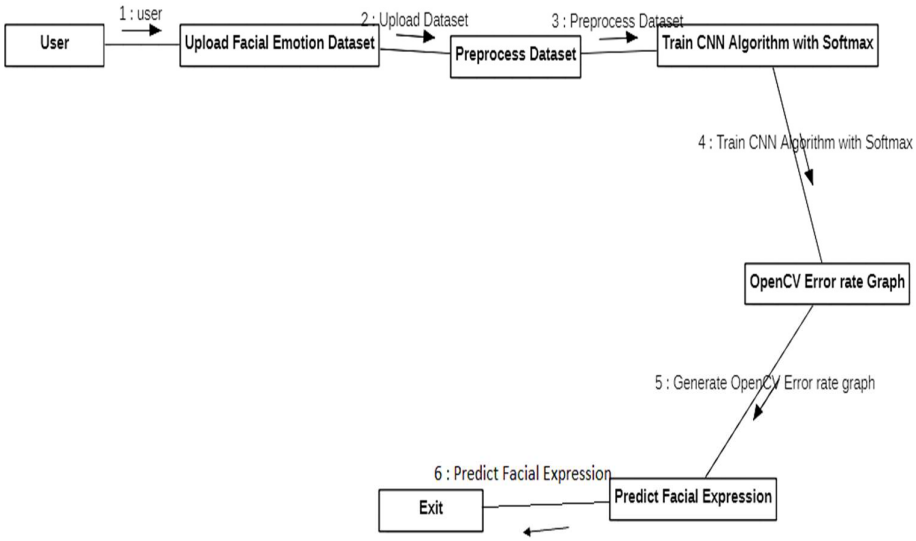
do in a use case. Further diagrams and documentation are needed for a complete functional and technical outlook on the system.

5.1.2 SEQUENCE DIAGRAM



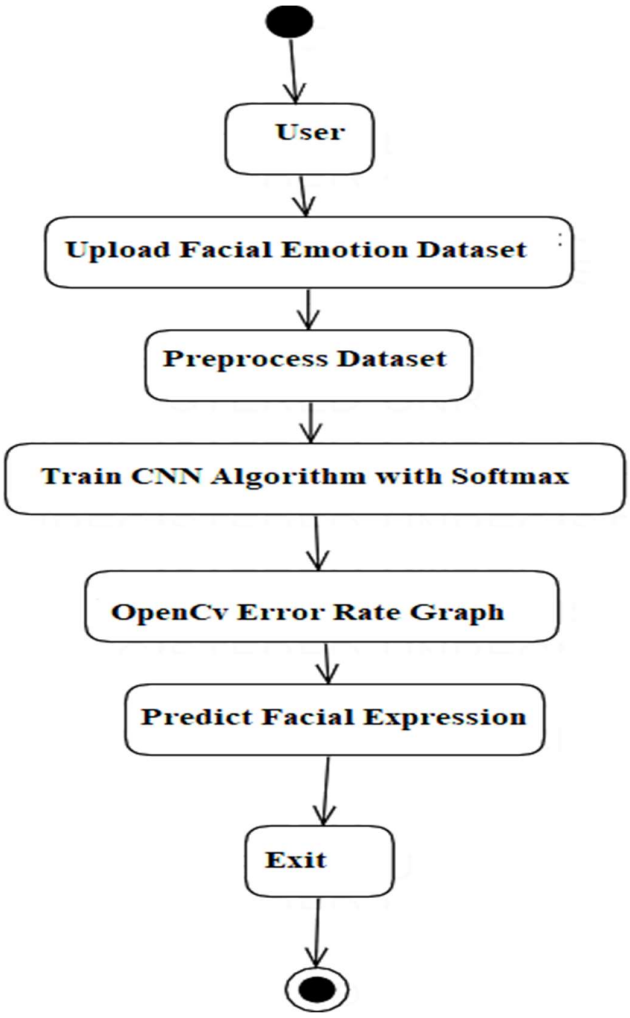
A sequence diagram is an interaction diagram. From the name, it is clear that the diagram deals with some sequences, which are the sequence of messages flowing from one object to another. Interaction among the components of a system is very important from implementation and execution perspective. Sequence diagram is used to visualize the sequence of calls in a system to perform a specific functionality.

5.1.3 COLLABORATION DIAGRAM



Collaboration diagram is another form of interaction diagram. It represents the structural organization of a system and the messages sent/received. Structural organization consists of objects and links. The purpose of collaboration diagram is similar to sequence diagram. The collaboration diagram is used to show the relationship between the objects in a system. Both the sequence and the collaboration diagrams represent the same information but differently. Instead of showing the flow of messages, it depicts the architecture of the object residing in the system as it is based on object-oriented programming. An object consists of several features. Multiple objects present in the system are connected to each other. The collaboration diagram, which is also known as a communication diagram, is used to portray the object's architecture in the system

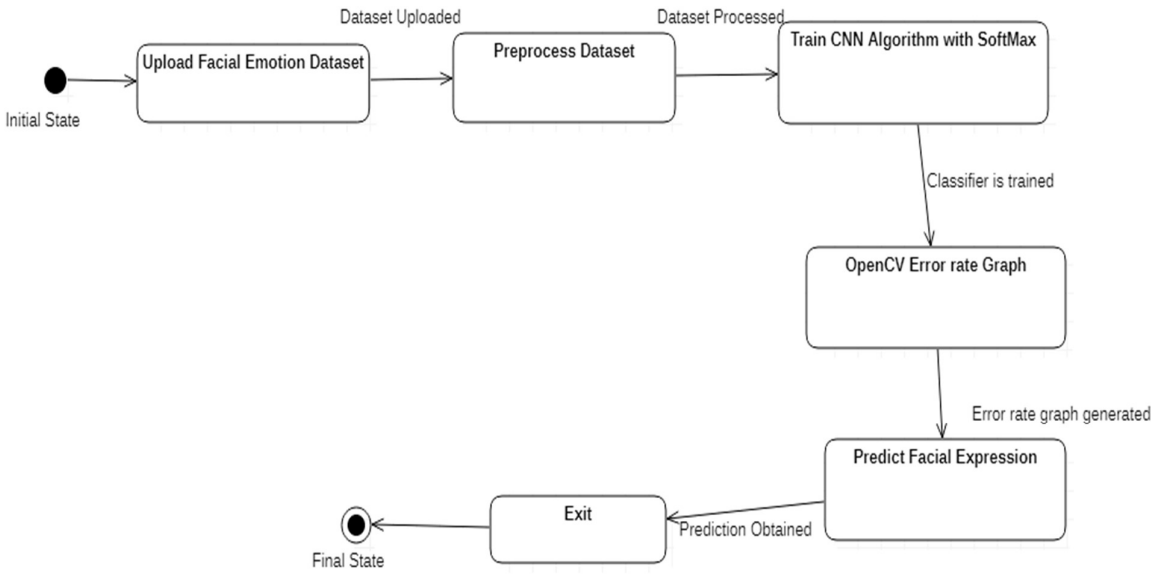
5.1.4 ACTIVITY DIAGRAM



Activity diagram describes the flow of control in a system. It consists of activities and links. The flow can be sequential, concurrent, or branched. Activities are nothing but the functions of a system. Numbers of activity diagrams are prepared to capture the entire flow in a system. Activity diagrams are used to visualize the flow of controls in a system. This is prepared to have an idea of how the system will work when executed. Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modelling Language, activity diagrams are intended to model both computational and organizational processes (i.e., workflows), as well as the data flows intersecting with the related activities. Although activity diagrams primarily show the overall flow of control, they can also include elements showing the flow of data between activities

through one or more data stores. Activity Diagrams describe how activities are coordinated to provide a service which can be at different levels of abstraction. Typically, an event needs to be achieved by some operations, particularly where the operation is intended to achieve a number of different things that require coordination, or how the events in a single use case relate to one another, in particular, use cases where activities may overlap and require coordination.

5.1.5 STATECHART DIAGRAM



Any real-time system is expected to be reacted by some kind of internal/external events. These events are responsible for state change of the system. Statechart diagram is used to represent the event driven state change of a system. It basically describes the state change of a class, interface, etc. A state diagram is a type of diagram used in computer science and related fields to describe the behaviour of systems. State diagrams require that the system described is composed of a finite number of states; sometimes, this is indeed the case, while at other times this is a reasonable abstraction. Many forms of state diagrams exist, which differ slightly and have different semantics.

State diagrams are used to give an abstract description of the behaviour of a system. This behaviour is analysed and represented by a series of events that can occur in one or more possible states. Hereby "each diagram usually represents objects of a single class and track the different states of its objects through the system". State diagrams can be used to graphically represent finite-state machines (also called finite automata). This was introduced by Claude Shannon and Warren Weaver in their 1949 book The Mathematical Theory of Communication. Another source is Taylor Booth in his 1967 book Sequential Machines and Automata Theory. Another possible representation is the state transition table.

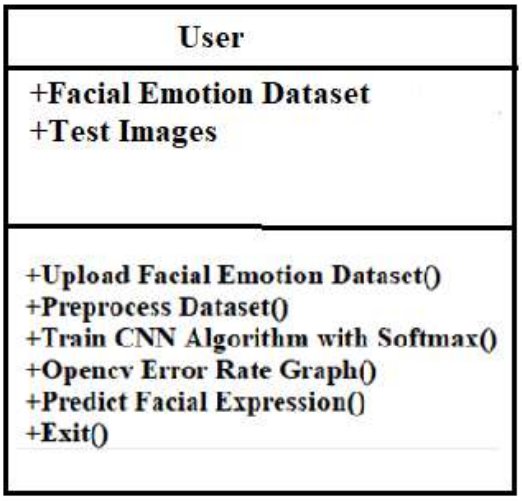
STRUCTURAL DIAGRAMS

Structure diagrams depict the static structure of the elements in your system. i.e., how one object relates to another. It shows the things in the system – classes, objects, packages or modules, physical nodes, components, and interfaces.

The four structural diagrams are –

- Class diagram
- Object diagram
- Component diagram
- Deployment diagram

5.1.6 CLASS DIAGRAM



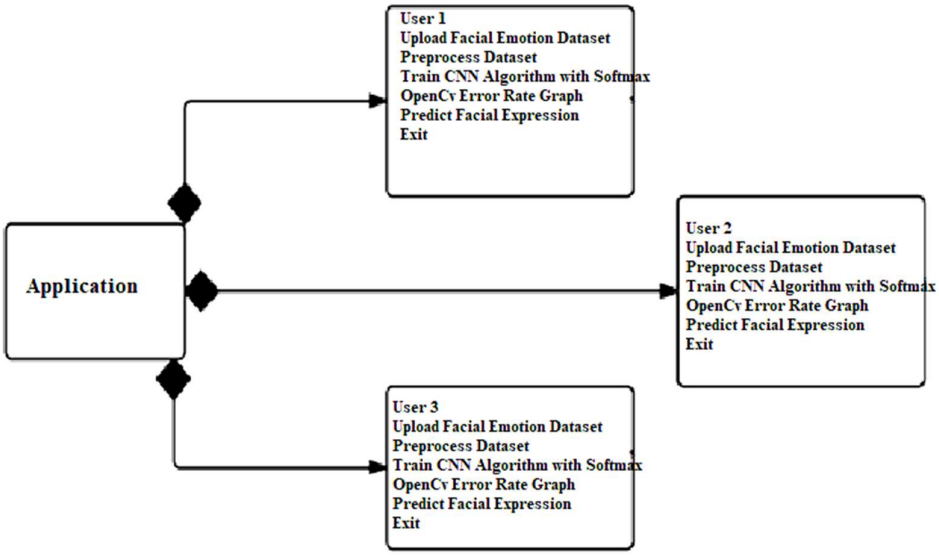
Class diagram represents the object orientation of a system. Hence, it is generally used for development purpose. This is the most widely used diagram at the time of system construction. In software engineering, a class diagram in the Unified Modelling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among objects. The class diagram is the main building block of object-oriented modelling. It is used for general conceptual modelling of the structure of the application, and for detailed modelling, translating the models into programming code. Class diagrams can also be used for data modelling. The classes in a class diagram represent both the main elements, interactions in the application, and the classes to be programmed. In the diagram, classes are represented with boxes that contain three compartments: 28

• The top compartment contains the name of the class. It is printed in bold and cantered, and the first letter is capitalized.

• The middle compartment contains the attributes of the class. They are left-aligned, and the first letter is lowercase.

• The bottom compartment contains the operations the class can execute. They are also leftaligned, and the first letter is lowercase.

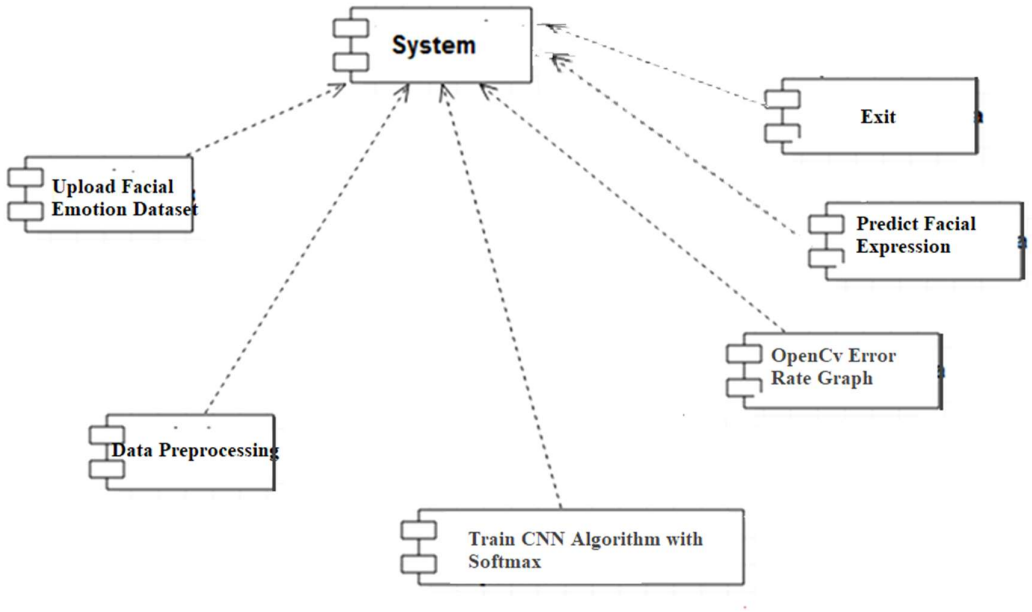
5.1.7 OBJECT DIAGRAM



Object diagrams can be described as an instance of class diagram. Thus, these diagrams are more close to real-life scenarios where we implement a system. In the Unified Modelling Language (UML), an object diagram focuses on some particular set of objects and attributes, and the links between these instances. A correlated set of object diagrams provides insight into how an arbitrary view of a system is expected to evolve over time. In early UML specifications the object diagram is described as: “An object diagram is a graph of instances, including objects and data values. A static object diagram is an instance of a class diagram; it shows a snapshot of the detailed state of a system at a point in time. The use of object diagrams is fairly limited, namely, to show examples of data structure.” The latest UML 2.5 specification does not explicitly define object diagrams but provides a notation for instances of classifiers. Object diagrams and class diagrams are closely related and use almost identical notation. Both diagrams are meant to visualize static structure of a system. While class diagrams show classes, object diagrams display instances of classes (objects). Object diagrams are more concrete than class diagrams. They are often used to provide

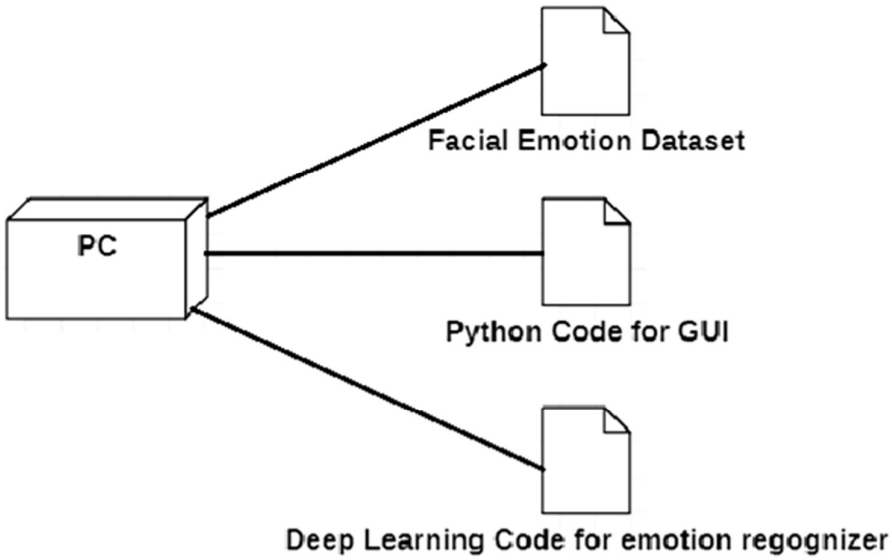
examples or act as test cases for class diagrams. Only aspects of current interest in a model are typically shown on an object diagram.

5.1.8 COMPONENT DIAGRAM



In Unified Modelling Language (UML), a component diagram depicts how components are wired together to form larger components or software systems. They are used to illustrate the structure of arbitrarily complex systems. A component diagram allows verification that a system's required functionality is acceptable. Component diagrams represent a set of components and their relationships. These components consist of classes, interfaces, or collaborations. Component diagrams represent the implementation view of a system These diagrams are also used as a communication tool between the developer and stakeholders of the system. Programmers and developers use the diagrams to formalize a roadmap for the implementation, allowing for better decision-making about task assignment or needed skill improvements. System administrators can use component diagrams to plan ahead, using the view of the logical software components and their relationships on the system

5.1.9 DEPLOYMENT DIAGRAM



A deployment diagram in the Unified Modelling Language models the physical deployment of artifacts on nodes. To describe a web site, for example, a deployment diagram would show what hardware components ("nodes") exist (e.g., a web server, an application server, and a database server), what software components ("artifacts") run on each node (e.g., web application, database), and how the different pieces are connected (e.g. JDBC, REST, RMI). The nodes appear as boxes, and the artifacts allocated to each node appear as rectangles within the boxes. Nodes may have sub nodes, which appear as nested boxes. A single node in a deployment diagram may conceptually represent multiple physical nodes, such as a cluster of database servers.These nodes are physical entities where the components are deployed .Deployment diagrams are used for visualizing the deployment view of a system. This is generally used by the deployment team.

CHAPTER 6
PROJECT CODING

6 PROJECT CODING

6.1 CODING TEMPLATES

For implementing the system various libraries, modules and functions contained by these libraries and modules have been imported as shown in the coding template below:

##Code

```
from tkinter import messagebox

from tkinter import *

from tkinter import simpledialog

import tkinter

import matplotlib.pyplot as plt

import numpy as np

import pandas as pd

from tkinter import simpledialog

from tkinter import filedialog

import os

import cv2

import numpy as np

from keras.utils.np_utils import to_categorical

from keras.layers import  MaxPooling2D

from keras.layers import Dense, Dropout, Activation, Flatten, LSTM

from keras.layers import Convolution2D

from keras.models import Sequential

from keras.models import model_from_json

import pickle
```

We use the keyword “import” to import different libraries to our file in python

“pandas” is a software library written for the Python programming language for data manipulation and analysis.

“NumPy” is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.

“matplotlib” is a plotting library for the Python programming language and its numerical mathematics extension NumPy.

“matplotlib.pyplot” is mainly intended for interactive plots and simple cases of programmatic plot generation.

“Tkinter” is a Python binding to the Tk GUI toolkit. It is the standard Python interface to the Tk GUI toolkit, and is Python's de facto standard GUI.

“OS” module in Python provides functions for interacting with the operating system. OS comes under Python's standard utility modules.

##Function to preprocess dataset

```
def processDataset():
    text.delete('1.0', END)
    global X, Y
     X = []
    Y = []
    for root, dirs, directory in os.walk(filename):
        for j in range(len(directory)):
            name = os.path.basename(root)
            print(name+" "+root+"/"+directory[j])
            if 'Thumbs.db' not in directory[j]:
                img = cv2.imread(root+"/"+directory[j])
```

```
            img = cv2.resize(img, (32,32))
            im2arr = np.array(img)
            im2arr = im2arr.reshape(32,32,3)
            X.append(im2arr)
            Y.append(getID(name))
    X = np.asarray(X)
    Y = np.asarray(Y)
    print(Y)
    X = X.astype('float32')
    X = X/255
    test = X[3]
    test = cv2.resize(test,(400,400))
    cv2.imshow("aa",test)
    cv2.waitKey(0)
    indices = np.arange(X.shape[0])
    np.random.shuffle(indices)
    X = X[indices]
    Y = Y[indices]
    Y = to_categorical(Y)
    np.save('model/X.txt',X)
    np.save('model/Y.txt',Y)
    X = np.load('model/X.txt.npy')
    Y = np.load('model/Y.txt.npy')
    text.insert(END,"Total number of images found in dataset is :
"+str(len(X))+"\n")
```

```
        text.insert(END,"Total classes found in dataset is : "+str(names)+"\n")
```

The above coding template shows the shows the module which is used for preprocessing the dataset.

##Code for CNN Algorithm:

```
def trainCNN():
    global classifier
    text.delete('1.0', END)
    if os.path.exists('model/cnnmodel.json'):
        with open('model/cnnmodel.json', "r") as json_file:
            loaded_model_json = json_file.read()
            classifier = model_from_json(loaded_model_json)
        classifier.load_weights("model/cnnmodel_weights.h5")
        #classifier._make_predict_function()
        print(classifier.summary())
        f = open('model/cnnhistory.pckl', 'rb')
        data = pickle.load(f)
        f.close()
        acc = data['accuracy']
        accuracy = acc[24] * 100
        text.insert(END,"CNN Softmax Training Model Accuracy = "+str(accuracy))
    else:
        classifier = Sequential()
        classifier.add(Convolution2D(32, 3, 3, input_shape = (32, 32, 3), activation =
'relu'))
        classifier.add(MaxPooling2D(pool_size = (2, 2)))
```

```
classifier.add(Convolution2D(32, 3, 3, activation = 'relu'))
classifier.add(MaxPooling2D(pool_size = (2, 2)))
classifier.add(Flatten())
classifier.add(Dense(output_dim = 256, activation = 'relu'))
classifier.add(Dense(output_dim = 7, activation = 'softmax')) #train with softmax
print(classifier.summary())
classifier.compile(optimizer = 'adam', loss = 'categorical_crossentropy', metrics = ['accuracy'])
hist = classifier.fit(X, Y, batch_size=32, epochs=25, shuffle=True, verbose=2)
classifier.save_weights('model/cnnmodel_weights.h5')
model_json = classifier.to_json()
with open("model/cnnmodel.json", "w") as json_file:
   json_file.write(model_json)
f = open('model/cnnhistory.pckl', 'wb')
pickle.dump(hist.history, f)
f.close()
f = open('model/cnnhistory.pckl', 'rb')
data = pickle.load(f)
f.close()
acc = data['accuracy']
accuracy = acc[24] * 100
text.insert(END,"CNN Softmax Faces Training Model Accuracy = "+str(accuracy))
```

The above coding template shows the Module which is used to build the CNN classifier and train it.

6.2 OUTLINE FOR VARIOUS FILES

Main.py: This file will contain the entire code of the project.

cnnmodel_weights.h5: It will store the weights.

X.txt.npy: It is a file which stores the images in the form of NumPy array

Y.txt.npy: It stores the labels of all the images present in the dataset.

Dataset (Folder which contains images for training and testing)

cnnmodel. json: CNN json model created successfully.

6.3 CLASS WITH FUNCTIONALITY

The proposed system will have a single module i.e., user and he will be able to perform the following operations with the system:

1) Upload facial emotion dataset(): Here we upload the dataset which consists 28,709 grayscale images of 48x48 size. The dataset which we use here is FER2013. All the images in the dataset depict any one of seven expressions (happy, sad, angry, disgusted, surprised, fear and neutral)

2) Preprocess the dataset(): The dataset which we upload will be preprocessed with the help of Opencv and Numpy libraries. The 48x48 sized images are later converted to 32x32 size images and the to numpy arrays by numpy package.

3) Train CNN algorithm with SoftMax activation function(): Here the preprocessed dataset is fed to the convolution network. 2 convolution layers consists RELU and SOFTMAX classifiers which help predict the facial expression of the selected image.

4) Error rate graph(): This graph shows us how the error rate is reduced in each epoch/iteration. The graph is generated using matplotlib library.

5) Predict facial emotion(): In this step the user will upload the test images to the classifier and it will be pre-processed by the system in the way training data was processed and the classifier will predict the facial expression present in the image.

6) Exit(): On clicking the exit button in the main interface the screen closes.

6.4 METHODS INPUT AND OUTPUT PARAMETERS

The system has been built using the following significant methods:

A) getID(name): We will have a global list “names” which will contain seven kinds of labels namely, 'angry','disgusted','fearful','happy','neutral','sad','surprised’. This method will play a significant role in “preprocessDataset” function in which the data which is in the form of 48x48 pixel format will be converted into a specific format using various functions present in OpenCV and NumPy libraries. The method getID will take the image as input and will return the label of the image as output with the help of “names” list. We can observe the method input output parameters in the code below.

##Code:

```
global filename

global X, Y

global classifier

names = ['angry','disgusted','fearful','happy','neutral','sad','surprised']

def getID(name):

    index = 0

        for i in range(len(names)):

            if names[i] == name:

                index = i

                break

        return index
```

B) upload(): This is the initial step in the execution of the system. Using this method, we will be able to upload the dataset to the system. We can observe the method input output parameters in the code below.

##Code

```
def upload():
    global filename
    filename = filedialog.askopenfilename(initialdir="model")
    text.delete('1.0', END)
     text.insert(END,filename+" loaded\n");
```

C) preprocessDataset(): This is the most crucial step in the execution process. This would take the dataset as input. with the help of OpenCV and NumPy we perform image processing in this method. In this we will create two global variables (lists,

which will later on be converted into NumPy arrays) X and Y where in X we will store the array form of the dataset and Y will store the labels of the images present in the dataset. We will use various predefined functions present in OpenCV and NumPy such as:

imread from cv2 which is used to read each and every image present in the directory ,resize from cv2 where we will change the images from 48x48 pixels to 32x32 pixels. Thereafter we will convert the image into Numpy array using “np.array” method and then change the dimentions of it using reshape method. We can observe the method input output parameters in the code below.

##Code

```
def processDataset():
   text.delete('1.0', END)
   global X, Y

   X = []
   Y = []
   for root, dirs, directory in os.walk(filename):
      for j in range(len(directory)):
```

```
        name = os.path.basename(root)
        print(name+" "+root+"/"+directory[j])
        if 'Thumbs.db' not in directory[j]:
            img = cv2.imread(root+"/"+directory[j])
            img = cv2.resize(img, (32,32))
            im2arr = np.array(img)
            im2arr = im2arr.reshape(32,32,3)
            X.append(im2arr)
            Y.append(getID(name))

X = np.asarray(X)
Y = np.asarray(Y)
print(Y)

X = X.astype('float32')
X = X/255
test = X[3]
test = cv2.resize(test,(400,400))
cv2.imshow("aa",test)
cv2.waitKey(0)
indices = np.arange(X.shape[0])
np.random.shuffle(indices)
X = X[indices]
Y = Y[indices]
```

```
Y = to_categorical(Y)

np.save('model/X.txt',X)

np.save('model/Y.txt',Y)

'''

X = np.load('model/X.txt.npy')

Y = np.load('model/Y.txt.npy')

text.insert(END,"Total number of images found in dataset is : "+str(len(X))+"\n")

text.insert(END,"Total classes found in dataset is : "+str(names)+"\n")
```

D) trainCNN():The numpy array gets passed into the Convolution2D layer where we specify the number of filters as one of the hyper parameters. The set of filters(aka. kernel) are unique with randomly generated weights. Each filter, (3, 3) receptive field, slides across the original image with shared weights to create a feature map. Convolution2D generates feature maps that represent how pixel values are enhanced, for example, edge and pattern detection. A feature map is created by applying filter 1 across the entire image. Other filters are applied one after another creating a set of feature maps. Pooling is a dimension reduction technique usually applied after one or several convolutional layers. It is an important step when building CNNs as adding more convolutional layers can greatly affect computational time. We used a popular pooling method called MaxPooling2D that uses (2, 2) windows across the feature map only keeping the maximum pixel value. The pooled pixels form an image with dimentions reduced by 4. These weights are trained by forward propagation of training data then backward propagation of its errors. Instead of using sigmoid activation function, we used softmax at the output layer. We have used Keras and Tensorflow library to build the CNN model. We can observe the method input output parameters in the code below.

##Code

```
def trainCNN():

    global classifier

    text.delete('1.0', END)

    if os.path.exists('model/cnnmodel.json'):

        with open('model/cnnmodel.json', "r") as json_file:
```

```
            loaded_model_json = json_file.read()
            classifier = model_from_json(loaded_model_json)
        classifier.load_weights("model/cnnmodel_weights.h5")
        #classifier._make_predict_function()
        print(classifier.summary())
        f = open('model/cnnhistory.pckl', 'rb')
        data = pickle.load(f)
        f.close()
        acc = data['accuracy']
        accuracy = acc[24] * 100
        text.insert(END,"CNN Softmax Training Model Accuracy = "+str(accuracy))
    else:
        classifier = Sequential()
        classifier.add(Convolution2D(32, 3, 3, input_shape = (32, 32, 3), activation = 'relu'))
        classifier.add(MaxPooling2D(pool_size = (2, 2)))
        classifier.add(Convolution2D(32, 3, 3, activation = 'relu'))
        classifier.add(MaxPooling2D(pool_size = (2, 2)))
        classifier.add(Flatten())
        classifier.add(Dense(output_dim = 256, activation = 'relu'))
        classifier.add(Dense(output_dim = 7, activation = 'softmax')) #train with softmax
        print(classifier.summary())
        classifier.compile(optimizer = 'adam', loss = 'categorical_crossentropy', metrics = ['accuracy'])
        hist = classifier.fit(X, Y, batch_size=32, epochs=25, shuffle=True, verbose=2)
```

```
    classifier.save_weights('model/cnnmodel_weights.h5')
    model_json = classifier.to_json()
    with open("model/cnnmodel.json", "w") as json_file:
       json_file.write(model_json)
    f = open('model/cnnhistory.pckl', 'wb')
    pickle.dump(hist.history, f)
    f.close()
    f = open('model/cnnhistory.pckl', 'rb')
    data = pickle.load(f)
    f.close()
    acc = data['accuracy']
    accuracy = acc[24] * 100
    text.insert(END,"CNN Softmax Faces Training Model Accuracy = "+str(accuracy))
```

E) predict():In this step the user will upload the test images to the classifier and it will be pre-processed by the system in the way training data was processed and the classifier will predict the facial expression present in the image. We can observe the method input output parameters in the code below.

##Code:

```
def predict():
   filename = filedialog.askopenfilename(initialdir="testImages")
   image = cv2.imread(filename)
   img = cv2.resize(image, (32,32))
   im2arr = np.array(img)
   im2arr = im2arr.reshape(1,32,32,3)
```

```
    img = np.asarray(im2arr)
    img = img.astype('float32')
    img = img/255
    preds = classifier.predict(img)
    predict = np.argmax(preds)
    img = cv2.imread(filename)
    img = cv2.resize(img, (600,400))
    cv2.putText(img, 'Facial Expression Recognized as : '+names[predict], (10, 25), cv2.FONT_HERSHEY_SIMPLEX,0.7, (255, 0, 0), 2)
    cv2.imshow('Facial Expression Recognized as : '+names[predict], img)
    cv2.waitKey(0)
```

F)graph():using this method we can obtain the Error Rate Graph. We will use matplotlib library for in this method as shown in the code below. We can observe the method input output parameters in the code below.

##Code

```
def graph():
    f = open('model/cnnhistory.pckl', 'rb')
    cnn_data = pickle.load(f)
    f.close()
    cnn_loss = cnn_data['loss']
    loss= []
    for i in range(len(cnn_loss)):
        if i > 14:
            loss.append(cnn_loss[i])
```

```
plt.figure(figsize=(10,6))

plt.grid(True)

plt.xlabel('Iterations/Epoch')

plt.ylabel('Opencv Error Rate')

plt.plot(cnn_loss, 'ro-', color = 'green')

plt.legend(['Opencv Error Rate'], loc='upper left')

#plt.xticks(wordloss.index)

plt.title('CNN with Opencv Error rate Graph')

plt.show()
```

G) Graphical User Interface: The GUI in this project is built using tkinter module present in python as shown in the code below.

##Code

```
font = ('times', 13, 'bold')

title = Label(main, text='Real time human emotion recognition based on facial expression detection using Softmax classifier and predict the error level using OpenCV library')

title.config(bg='LightGoldenrod1', fg='medium orchid')

title.config(font=font)

title.config(height=3, width=120)

title.place(x=0,y=5)


font1 = ('times', 12, 'bold')

text=Text(main,height=20,width=100)

scroll=Scrollbar(text)

text.configure(yscrollcommand=scroll.set)

text.place(x=480,y=100)
```

```
text.config(font=font1)

font1 = ('times', 12, 'bold')

uploadButton = Button(main, text="Upload Facial Emotion Dataset",
command=upload)

uploadButton.place(x=50,y=100)

uploadButton.config(font=font1)


processButton = Button(main, text="Preprocess Dataset",
command=processDataset)

processButton.place(x=50,y=150)

processButton.config(font=font1)


cnnButton = Button(main, text="Train CNN Algorithm with Softmax",
command=trainCNN)

cnnButton.place(x=50,y=200)

cnnButton.config(font=font1)


graphButton = Button(main, text="Opencv Error Rate Graph",
command=graph)

graphButton.place(x=50,y=250)

graphButton.config(font=font1)


predictButton = Button(main, text="Predict Facial Expression",
command=predict)

predictButton.place(x=50,y=300)

predictButton.config(font=font1)
```

```
exitButton = Button(main, text="Exit", command=exit)
exitButton.place(x=50,y=350)
exitButton.config(font=font1)

main.config(bg='OliveDrab2')
main.mainloop()
```

CHAPTER 7
PROJECT TESTING

7. PROJECT TESTING

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub-assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of tests. Each test type addresses a specific testing requirement.

7.1 VARIOUS TEST CASES

Unit testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

Test strategy and approach

Field testing will be performed manually and functional tests will be written in detail.

Test objectives

- All field entries must work properly.
- Pages must be activated from the identified link.
- The entry screen, messages and responses must not be delayed.

Features to be tested

- Verify that the entries are of the correct format
- No duplicate entries should be allowed
- All links should take the user to the correct page.

Integration Testing

Software integration testing is the incremental integration testing of two or more integrated software components on a single platform to produce failures caused by interface defects. The task of the integration test is to check that components or software applications, e.g. components in a software system or – one step up – software applications at the company level – interact without error.

Test Results:

All the test cases mentioned above passed successfully. No defects encountered.

Acceptance Testing

User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements. Acceptance Testing is a method of software testing where a system is tested for acceptability. The major aim of this test is to evaluate the compliance of the system with the business requirements and assess whether it is acceptable for delivery or not.

Test Results:

All the test cases mentioned above passed successfully. No defects encountered

Functional test

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items:

- Valid Input : identified classes of valid input must be accepted.
- Invalid Input : identified classes of invalid input must be rejected.
- Functions : identified functions must be exercised.
- Output : identified classes of application outputs must be exercised.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

System Test

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. In system testing, integration testing passed components are taken as input. The goal of integration testing is to detect any irregularity between the units that are integrated together. System testing detects defects within both the integrated units and the whole system. The result of system testing is the observed behavior of a component or a system when it is tested. An example of system testing is the configuration-oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

7.2 BLACK BOX TESTING

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. A tester provides an input, and observes the output generated by the system under test. This makes it possible to identify how the system responds to expected and unexpected user actions, its response time, usability issues and reliability issues. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box. you cannot “see” into it. Black box testing is a powerful testing technique because it exercises a system end-to-end. Just like end-users “don’t care” how a system is coded or architected, and expect to receive an appropriate response to their requests, a tester can simulate user activity and see if the system delivers on its promises.

7.3 WHITE BOX TESTING

White Box Testing is a testing in which in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is purpose. It is used to test areas that cannot be reached from a black box level. White-box testing is a method of testing the application at the level of the source code. These test cases are derived through the use of the design techniques mentioned above: control flow testing, data flow testing, branch testing, path testing, statement coverage and decision coverage as well as modified condition/decision coverage. White-box testing is the use of these techniques as guidelines to create an error-free environment by examining all code. These white-box testing techniques are the building blocks of white-box testing, whose essence is the careful testing of the application at the source code level to reduce hidden errors later on. These different techniques exercise every visible path of the source code to minimize errors and create an error-free environment. The whole point of white-box testing is the ability to know which line of the code is being executed and being able to identify what the correct output should be. Working process of white box testing:

• Input: Requirements, Functional specifications, design documents, source code.

• Processing: Performing risk analysis for guiding through the entire process.

• Proper test planning: Designing test cases so as to cover entire code. Execute rinse-repeat until error-free software is reached. Also, the results are communicated.

• Output: Preparing final report of the entire testing process.

CHAPTER 8
OUTPUT SCREENS

8. OUTPUT SCREENS

8.1 User Interface

Fig 8.1.1 main interface

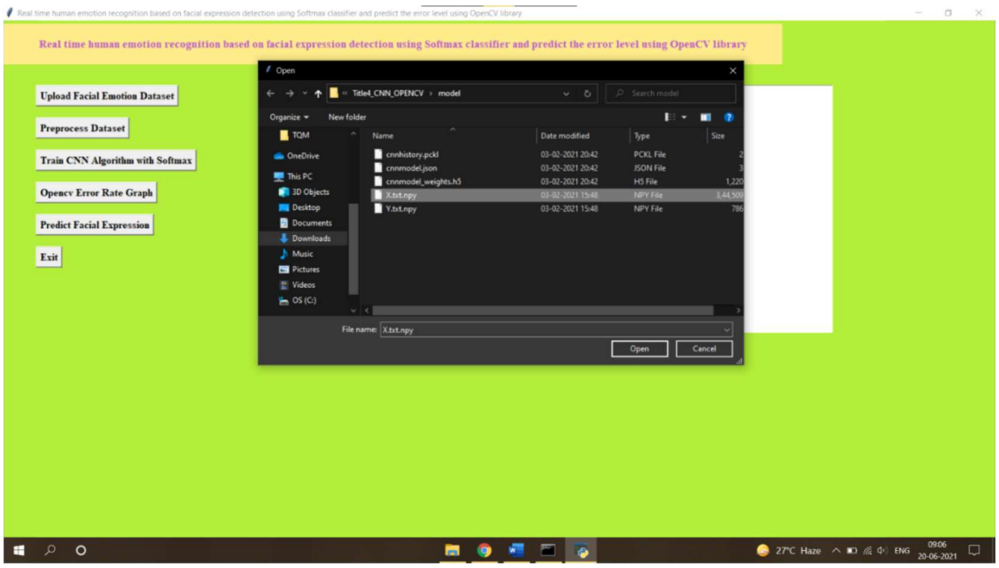


Fig 8.1.2 uploading dataset.

Fig 8.1.3 pre-process dataset

Fig 8.1.4 train CNN algorithm

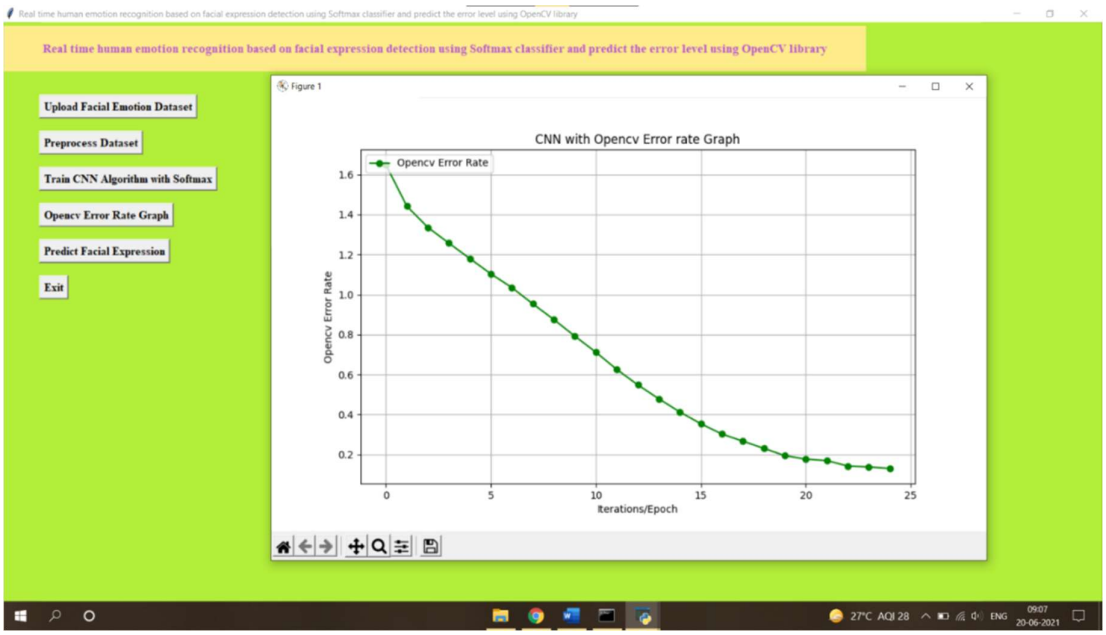


Fig 8.1.5 error graph

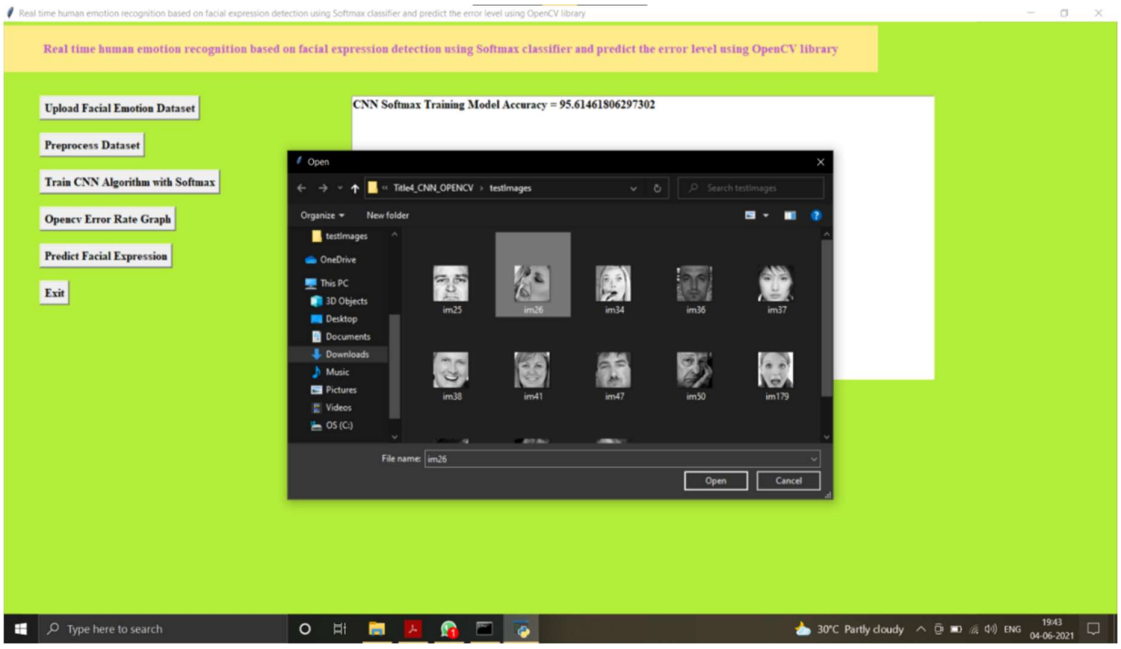


Fig 8.1.6 selecting image for expression prediction

8.2 OUTPUT SCREENS

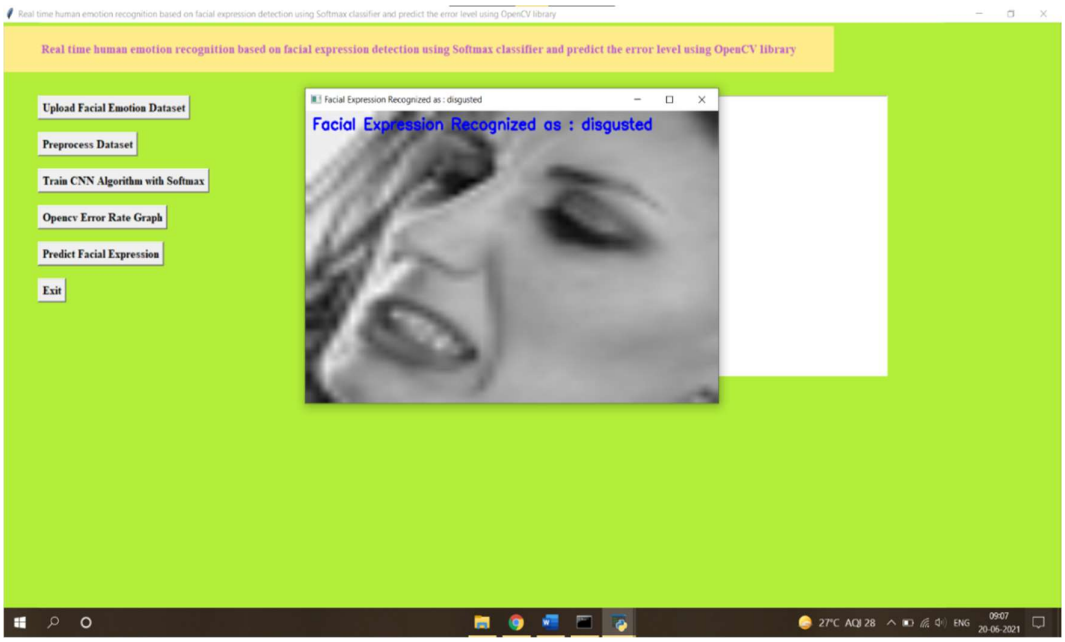


Fig 8.2.1 disgusted expression

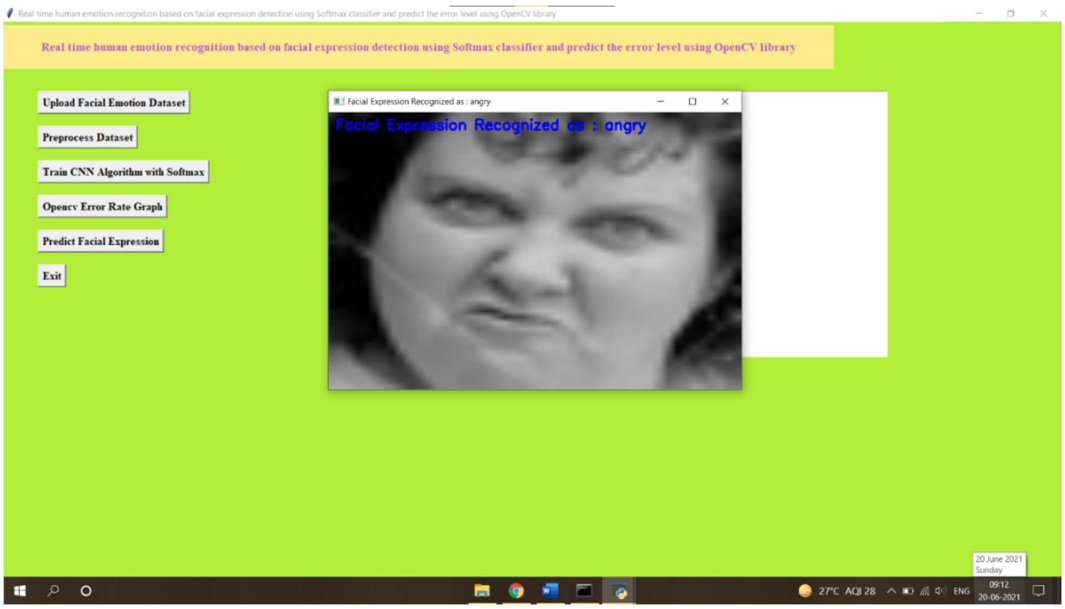


Fig 8.2.2 angry expression

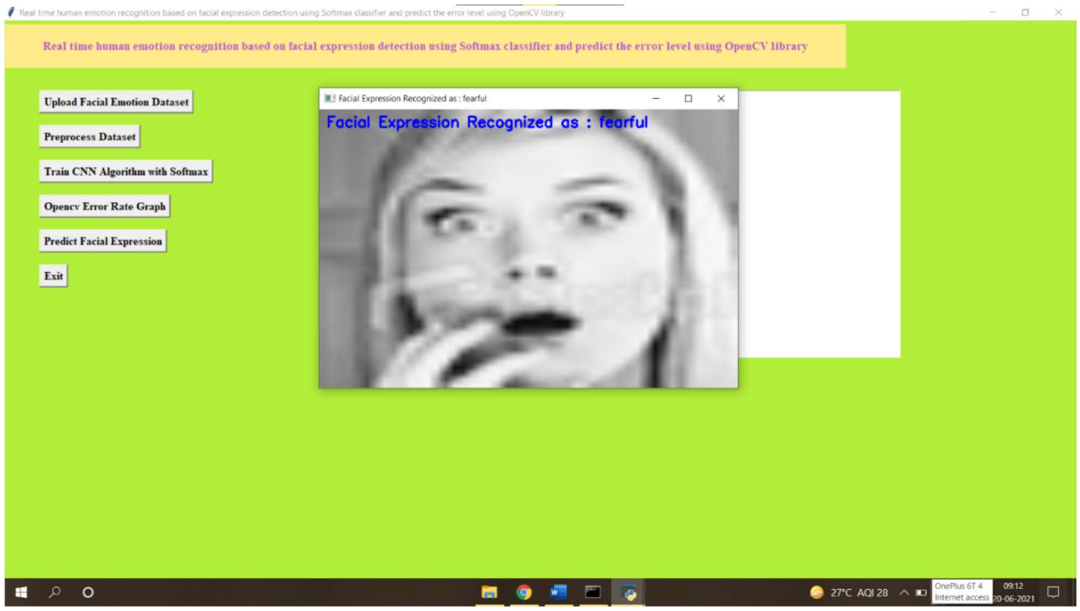


Fig 8.2.3 fear expression

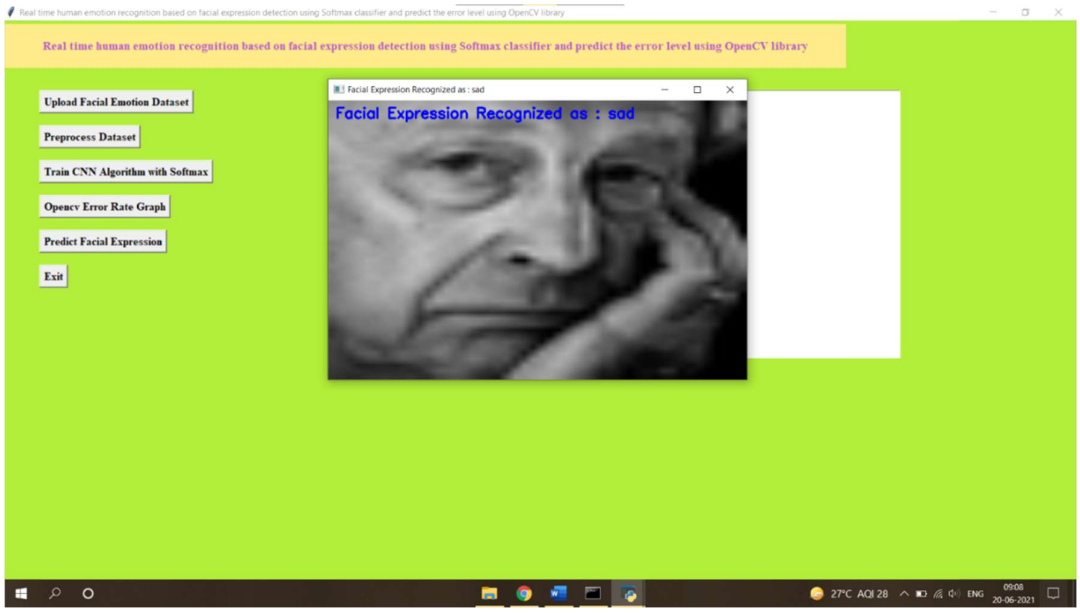


Fig 8.2.4 sad expression

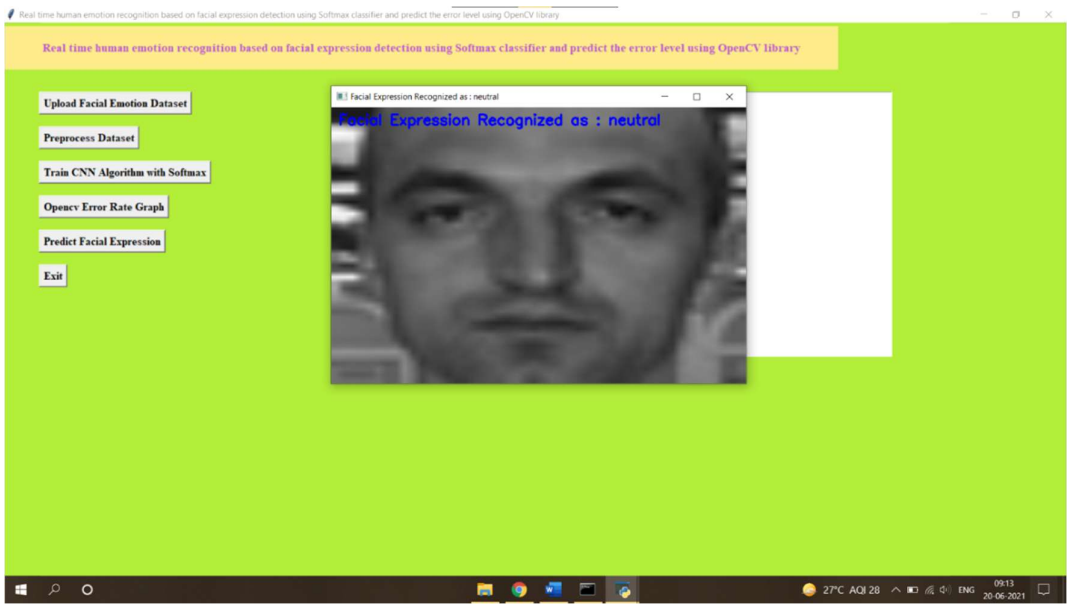


Fig 8.2.5 neutral expression

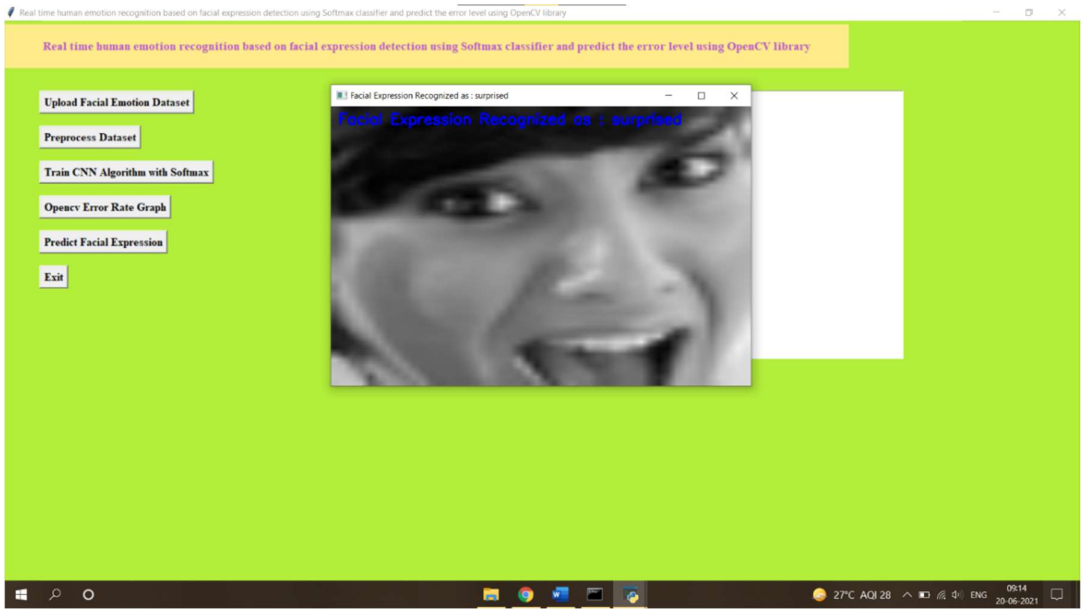


Fig 8.2.6 surprised expression

CHAPTER 9
EXPERIMENTAL RESULTS

9. EXPERIMENTAL RESULTS

This system proves to be very efficient in terms of emotion recognition and the processing speed given that the dataset is very challenging. Hence the system added with additional functionalities can be considered useful for human emotion recognition in real-time.We were able to achieve an accuracy of 95.61 with the help of system we have built OpenCV and CNN.Convolutional neural network (CNN) models have kernels to detect border functions or outline for an image. This model has weights arranged in array of values to form and obtain desired characteristics. Every CNN model allocates space to determine the control of image to be recognized.

```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d_1 (Conv2D)            (None, 30, 30, 32)        896
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 15, 15, 32)        0
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 13, 13, 32)        9248
_________________________________________________________________
max_pooling2d_2 (MaxPooling2 (None, 6, 6, 32)          0
_________________________________________________________________
flatten_1 (Flatten)          (None, 1152)              0
_________________________________________________________________
dense_1 (Dense)              (None, 256)               295168
_________________________________________________________________
dense_2 (Dense)              (None, 7)                 1799
=================================================================
Total params: 307,111
Trainable params: 307,111
Non-trainable params: 0
```

Fig 9.1 classifier.summary()

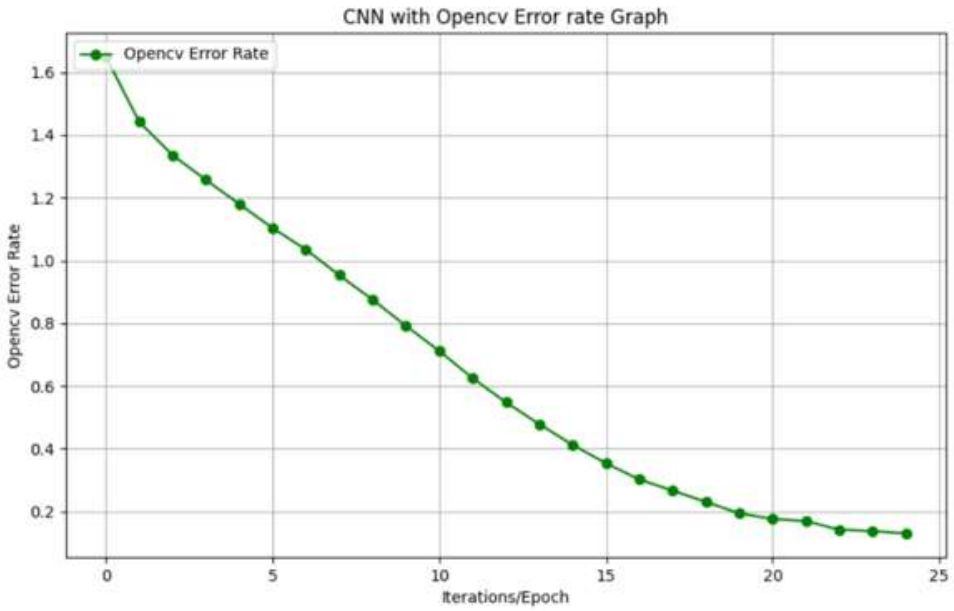


Fig 9.2 OpenCV Error Rate Graph

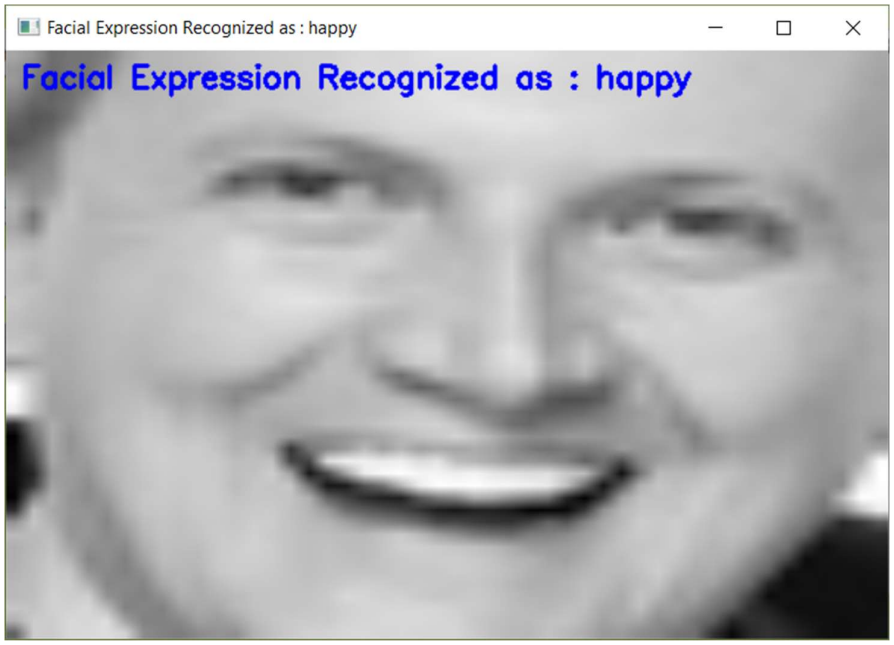


Fig 9.3 Prediction Result of one of the test image

CHAPTER 10
CONCLUSIONS AND FUTURE ENHANCEMENTS

10. CONCLUSION AND FUTURE ENHANCEMENT

This Project intends to develop a FER system using CNN. The system will classify the expression of a human face into one of seven expressions – anger, happiness, sadness, surprise, fear, neutral, disgust. The model thus developed can be further extended to categorize human faces in real time using a webcam by additional Pre-processing modules and by using various image processing functions resent in OpenCV Library. This FER system can be used for analysis of user expressions, to help the system understand human requirements better. Also, we have observed that FER approaches can be divided into two main streams i.e., conventional/classical and deep learning FER approaches. conventional FER approaches consisting of three steps, namely, face and facial component detection, feature extraction, and expression classification. The classification algorithms used in conventional FER include SVM, Adaboost, and random forest; by contrast, deep learning-based FER approaches highly reduce the dependence on face-physics-based models and other pre-processing techniques by enabling “end-to-end” learning in the pipeline directly from the input images. Although studies on FER have been conducted over the past decade, in recent years the performance of FER has been significantly improved through a combination of deep-learning algorithms. It is important to note that there is no specific formula to build a neural network that would guarantee to work well. Different problems would require different network architecture and a lot of trail and errors to produce desirable validation accuracy. This is the reason why neural nets are often perceived as "black box algorithms.". In this project we got an accuracy of almost 95%. But we need to improve in specific areas like-

- number and configuration of convolutional layers.
- number and configuration of dense layers.

We would also like to train more databases into the system to make the model more and more accurate but again resources become a hindrance in the path and we also need to improve in several areas in future to resolve the errors and improve the accuracy. Having examined techniques to cope with expression variation, in future it may be investigated in more depth about the face classification problem and optimal fusion of color and depth information.

REFERENCES

[1] Zafar B, Ashraf R, Ali N, Iqbal M, Sajid M, Dar S, Ratyal N (2018) A novel discriminating and relative global spatial image representation with applications in CBIR. Appl Sci 8(11):2242

[2] Ali N, Zafar B, Riaz F, Dar SH, Ratyal NI, Bajwa KB, Iqbal MK, Sajid M (2018) A hybrid geometric spatial image representation for scene classification. PLoS ONE 13(9):e0203339

[3] Ali N, Zafar B, Iqbal MK, Sajid M, Younis MY, Dar SH, Mahmood MT, Lee IH (2019) Modeling global geometric spatial information for rotation invariant classification of satellite images. PLoS ONE 14:7

[4] Ali N, Bajwa KB, Sablatnig R, Chatzichristofis SA, Iqbal Z, Rashid M, Habib HA (2016) A novel image retrieval based on visual words integration of SIFT and SURF. PLoS ONE 11(6):e0157428

[5] Ekman P, Friesen WV (1971) Constants across cultures in the face and emotion. J Personal Soc Psychol 17(2):124

[6] Matsumoto D (1992) More evidence for the universality of a contempt expression. Motiv Emot 16(4):363

[7] Sajid M, Iqbal Ratyal N, Ali N, Zafar B, Dar SH, Mahmood MT, Joo YB (2019) The impact of asymmetric left and asymmetric right face images on accurate age estimation. Math Probl Eng 2019:1–10

[8] Ratyal NI, Taj IA, Sajid M, Ali N, Mahmood A, Razzaq S (2019) Three-dimensional face recognition using variance-based registration and subject-specific descriptors. Int J Adv Robot Syst 16(3):172881419851716

[9] Ratyal N, Taj IA, Sajid M, Mahmood A, Razzaq S, Dar SH, Ali N, Usman M, Baig MJA, Mussadiq U (2019) Deeply learned pose invariant image analysis with applications in 3D face recognition. Math Probl Eng 2019:1–21

[10] Sajid M, Ali N, Dar SH, Iqbal Ratyal N, Butt AR, Zafar B, Shafique T, Baig MJA, Riaz I, Baig S (2018) Data augmentation-assisted makeup-invariant face recognition. Math Probl Eng 2018:1–10

[11] Ratyal N, Taj I, Bajwa U, Sajid M (2018) Pose and expression invariant alignment based multi-view 3D face recognition. KSII Trans Internet Inf Syst 12:10

[12] Xie S, Hu H (2018) Facial expression recognition using hierarchical features with deep comprehensive multipatches aggregation convolutional neural networks. IEEE Trans Multimedia 21(1):211

[13]Danisman T, Bilasco M, Ihaddadene N, Djeraba C (2010) Automatic facial feature detection for facial expression recognition. In: Proceedings of the International conference on computer vision theory and applications, pp 407–412. https://doi.org/10.5220/0002838404070412

[14] Mal HP, Swarnalatha P (2017) Facial expression detection using facial expression model. In: 2017 International conference on energy, communication, data analytics and soft computing (ICECDS). IEEE, pp 125 9–1262

[15] Ratyal N, Taj I, Bajwa U, Sajid M (2018) Pose and expression invariant alignment based multi-view 3D face recognition. KSII Trans Internet Inf Syst 12:10

[16] Liew, C.F.; Yairi, T. Facial Expression Recognition and Analysis: A Comparison Study of Feature Descriptors. IPSJ Trans. Comput. Vis. Appl.

[17] Ko, B.C. A Brief Review of Facial Emotion Recognition Based on Visual Information. Sensors 2018, 18, 401.

[18] Huang, Y.; Chen, F.; Lv, S.; Wang, X. Facial Expression Recognition: A Survey. Symmetry 2019, 11, 1189

[19] Li, S.; Deng, W. Deep Facial Expression Recognition: A Survey. IEEE Trans. Affect. Comput. 2020.

[20] Alom, M.Z.; Taha, T.M.; Yakopcic, C.; Westberg, S.; Sidike, P.; Nasrin, M.S.; Hasan, M.; Van Essen, B.C.; Awwal, A.A.S.; Asari, V.K. A State-of-the-Art Survey on Deep Learning Theory and Architectures. Electronics 201, 8, 292. [CrossRef]

[21] Sahu, M.; Dash, R. A Survey on Deep Learning: Convolution Neural Network (CNN). In Smart Innovation, Systems and Technologies; Springer: Singapore, 2021; Volume 153, pp. 317–325.

[22] Mollahosseini, A.; Chan, D.; Mahoor, M.H. Going deeper in facial expression recognition using deep neural networks. In Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, NY, USA, 7–10 March 2016; pp. 1–10. [CrossRef]

[23] Zhao, X.; Shi, X.; Zhang, S. Facial Expression Recognition via Deep Learning. IETE Tech. Rev. 2015, 32, 347–355. [CrossRef]

[24] Khan, A.; Sohail, A.; Zahoora, U.; Qureshi, A.S. A survey of the recent architectures of deep convolutional neural networks. Artif. Intell. Rev. 2020, 53, 5455–5516. [CrossRef]

[25] Kolen, J.F.; Kremer, S.C. Gradient Flow in Recurrent Nets: The Difficulty of Learning LongTerm Dependencies. In A Field Guide to Dynamical Recurrent Networks; Wiley-IEEE Press: Hoboken, NJ, USA, 2010; pp. 237–243.

PUBLICATIONS

- "Innovations in Computers Networks, Computational Intelligence and IoT"[ICICCI-21].

- Paper id: **ICICCI-21-0120**

Sai Akhil Chanda (17K81A0545) is currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. He completed his intermediate from Narayana Junior College and 10th class from S.S.V Gyan Kendra School. His technical skills include C, C++, Python and Java. He took part in Employability Skill development Program conducted by Zensar. He is also a student of Smart Interviews. He also did his internship at Verzeo in Machine Learning domain (May 2020-August 2020). His areas of interest are Python, Artificial Intelligence, Machine Learning and Deep Learning. She completed few certification courses from online platforms like Coursera, CursaApp and SoloLearn. He participated in various events, seminars during her graduation, some of them are:

1	Participated in Employability Skill development Program conducted by Zensar
2	Student of Smart Interviews
3	Participated in National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing"
4	Did his internship in Machine learning domain at Verzeo edutech.
5	Certification in Artificial Intelligence by Crash Course in CursaApp.
6	Certification in MYSQL database by Thenewboston in SoloLearn
7	Certification in Java in SoloLearn
8	Certification in Python core sololearn
9	Certification in Programming Fundamentals Coursera
10	Certification in AWS Fundamentals: Going Cloud - Native
11	Certification in Leadership and Emotional Intelligence
12	Certification in Managing Project Risks and Changes

K. Sachetan Reddy is currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. He completed his intermediate from Sri Chaitnya Junior College and 10th class from Narendra High School. His technical skills include C, Python and Java. He also has a basic understanding of C++.He completed few certification courses from online platforms like Coursera, CursaApp .

1	Certification in Managing Project Risks and Changes Coursera
2	Certification in Leadership and Emotional Intelligence Coursera
3	Certification in AWS Fundamentals: Going Cloud-Native Coursera
4	Certification in Matrix Algebra for Engineers Coursera
5	Certification in AI For Everyone Coursera
6	Certification in Programming with PHP for Beginners by The Net Ninja CursaApp
7	Certification in Cyber Security by Packethacks CursaApp
8	Certification in MySQL database by Thenewboston CursaApp

A. Harshith Reddy is currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. He completed his intermediate from NRI Junior College and 10th class from Gowtham Model School. His technical skills include C, Python and Java. He also has a basic understanding of C++.He has completed few certification courses from online platforms like Coursera, CursaApp .

1	Certification in Managing Project Risks and Changes Coursera
2	Certification in Leadership and Emotional Intelligence Coursera
3	Certification in AWS Fundamentals: Going Cloud-Native Coursera
4	Certification in Matrix Algebra for Engineers Coursera
5	Certification in AI For Everyone Coursera
6	Certification in Programming with PHP for Beginners by The Net Ninja CursaApp
7	Certification in Cyber Security by Packethacks CursaApp
8	Certification in MySQL database by Thenewboston CursaApp

G. Yashwanth Reddy is currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. He completed his intermediate from Sri Chaitanya Junior College and 10th class from Shantinikethan High School. His technical skills include C, Python and Java. He also has a basic understanding of C++. He has completed few certification courses from online platforms like Coursera, CursaApp .

1	Certification in Managing Project Risks and Changes Coursera
2	Certification in Leadership and Emotional Intelligence Coursera
3	Certification in AWS Fundamentals: Going Cloud-Native Coursera
4	Certification in Matrix Algebra for Engineers Coursera
5	Certification in AI For Everyone Coursera

ST. MARTIN'S ENGINEERING COLLEGE

(An Autonomous Institute)

Dhulapally, Secunderabad– 500100

NBA & NAAC A+ ACCREDITED

Human Computer Interaction System: Computer Cursor Movement Using Human Eyeball Movement

Presented by

Ch Manish Goud(17K81A0508)

Shaik Shadullah(17K81A0549)

Shaik Imran(17K81A0548)

Y Pradeep(17K81A0560)

Under the Guidance of

Dr. R. Santhoshkumar,

(B.Tech., M.Tech., Ph.D.,)

Asst.Prof/Assoc.Prof/Professor,

Department of Computer Science and Engineering

ABSTRACT

- The eye controls of great use to not only the future of natural input but more importantly the handicapped and disabled persons. Camera is capturing the image of eye movement. First detect pupil center position of eye.
- The signals pass the motor driver to interface with the virtual keyboard itself. The motor driver will control both speed and direction to enable the virtual keyboard to move left, right, up, down and stop.
- We are instructing mouse cursor to change its location based on eye ball movement, in this application using OPENCV we will connect to webcam and then extract each frame from the webcam and pass to OPENCV to detect eye balls location.
- Camera detects the Eye ball movement which can be processed in OpenCV. By this the cursor can be controlled.

INTRODUCTION

- As computer technology advances, the importance of human-computer interaction becomes increasingly apparent. Some people with disabilities are unable to utilise computers.

- Eye movement control is mostly utilised by those who are impaired. By incorporating this eye-controlling mechanism into computers, they will be able to work without the assistance of others.

- Human-Computer Interface (HCI) is concerned with the use of computer technology to establish a human-computer interface. There is a need to discover appropriate technology that allows for good human-computer collaboration.

INTRODUCTION

- The importance of human-computer connection cannot be overstated. As a result, there is a need to develop a mechanism for disseminating an alternate mode of human-computer communication to people with disabilities, giving them an equal opportunity to participate in the Information Society.

- At present situation paralyzed peoples need a guidance to do any work. One person should be they're with that person to taken care of him.

- By using the eye ball tracking mechanism, we can fix the centroid on the eye based on the centroid we need to track that paralyzed person's eye this eye ball track mechanism involves many applications like home automation by using python GUI robotic Control and virtual keyboard application

SCOPE OF THE PROJECT

- In this project we are instructing mouse cursor to change its location based on eye ball movement, in this application using OPENCV we will connect to webcam and then extract each frame from the webcam and pass to OPENCV to detect eye balls location.

- Once eye ball location detected then we can extract x and y coordinates of eye balls from OPENCV and then using python pyautogui API we can instruct mouse to change its current location to given eyeballs X and Y Coordinates.

- Below is the example to move mouse in python. pyautogui.moveTo(int(data_x), int(data_y)).

GOALS OF THE PROJECT

- Project undertakes to develop a system which will only use webcam, to use human eyes as as a pointing device for computer system and providing user friendly human-computer interaction.

OBJECTIVES OF THE PROJECT

- Face & Eyes Detection
- Finding Center of Iris / Pupil
- Eye Corners Extraction
- Develop an algorithm to calculate point of Gaze based on Eye features found.
- Develop a GUI to show results
- Develop a simple Calibration technique.

EXISTING SYSTEM

➢Matlab detect the iris and control curser. Eye movement-controlled wheel chair is existing one that controls the wheel chair by monitoring eye movement. In matlab is difficult to predict the Centroid of eye so we go for OpenCV.

➢we are instructing mouse cursor to change its location based on eye ball movement, in this application using OPENCV we will connect to webcam and then extract each frame from the webcam and pass to OPENCV to detect eye balls location.

➢Once eye ball location detected then we can extract x and y coordinates of eye balls from OPENCV and then using python pyautogui API we can instruct mouse to change its current location to given eyeballs X and Y Coordinates. Below is the example to move mouse in python.

PROPOSED SYSTEM

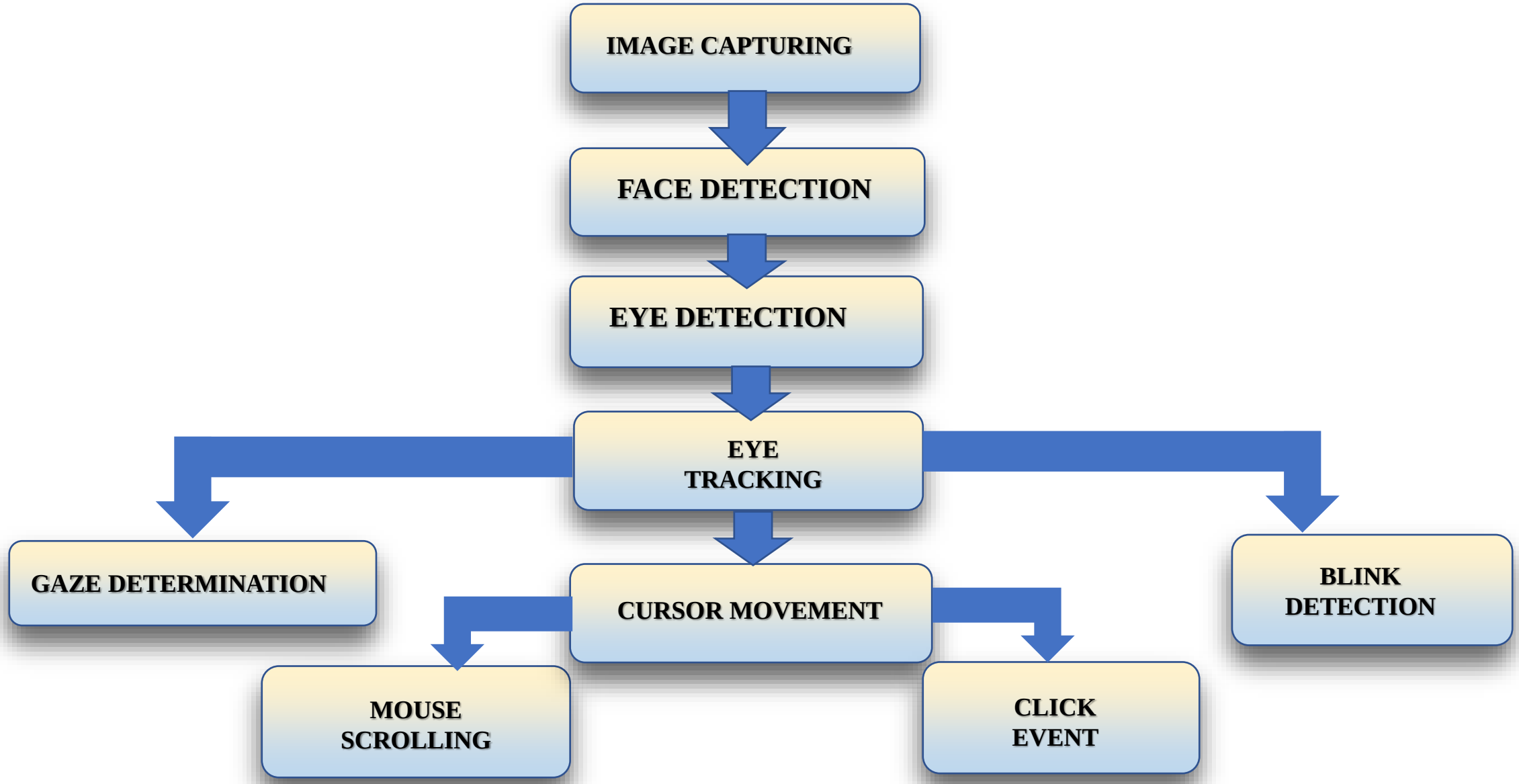
- In our proposed system the cursor movement of computer is controlled by eye movement using Open CV.

- Camera detects the Eye ball movement which can be processed in OpenCV. By this the cursor can be controlled.

- The user has to sits in front of the display screen of private computer or pc, a specialized video camera established above the screen to study the consumer's eyes.

- The laptop constantly analysis the video photo of the attention and determines wherein the consumer is calling at the display screen. not anything is attached to the consumer's head or body.

ARCHITCTURE OF PROPOSED SYSTEM



IMPLEMENTATION

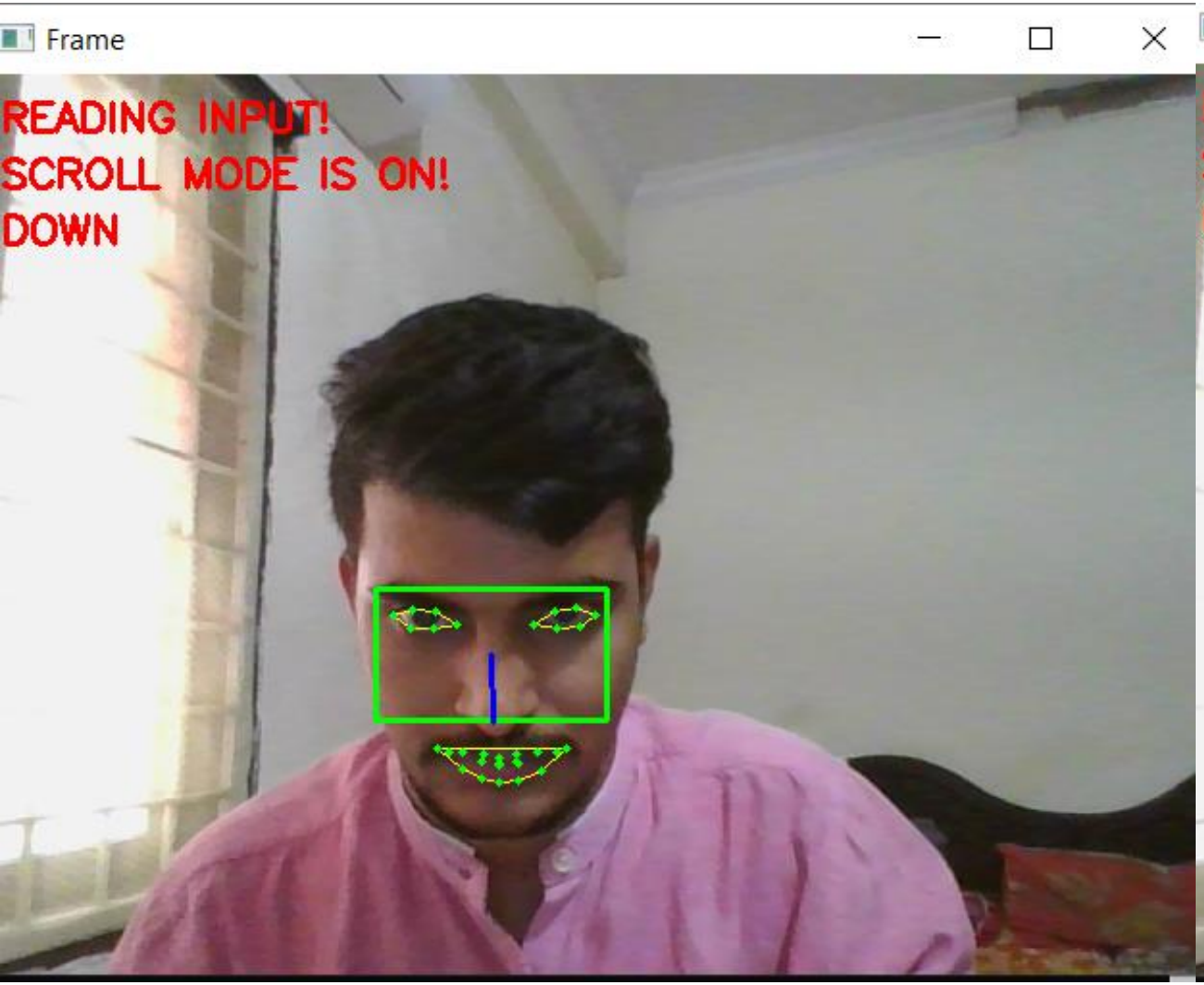
HARDWARE REQUIREMENTS

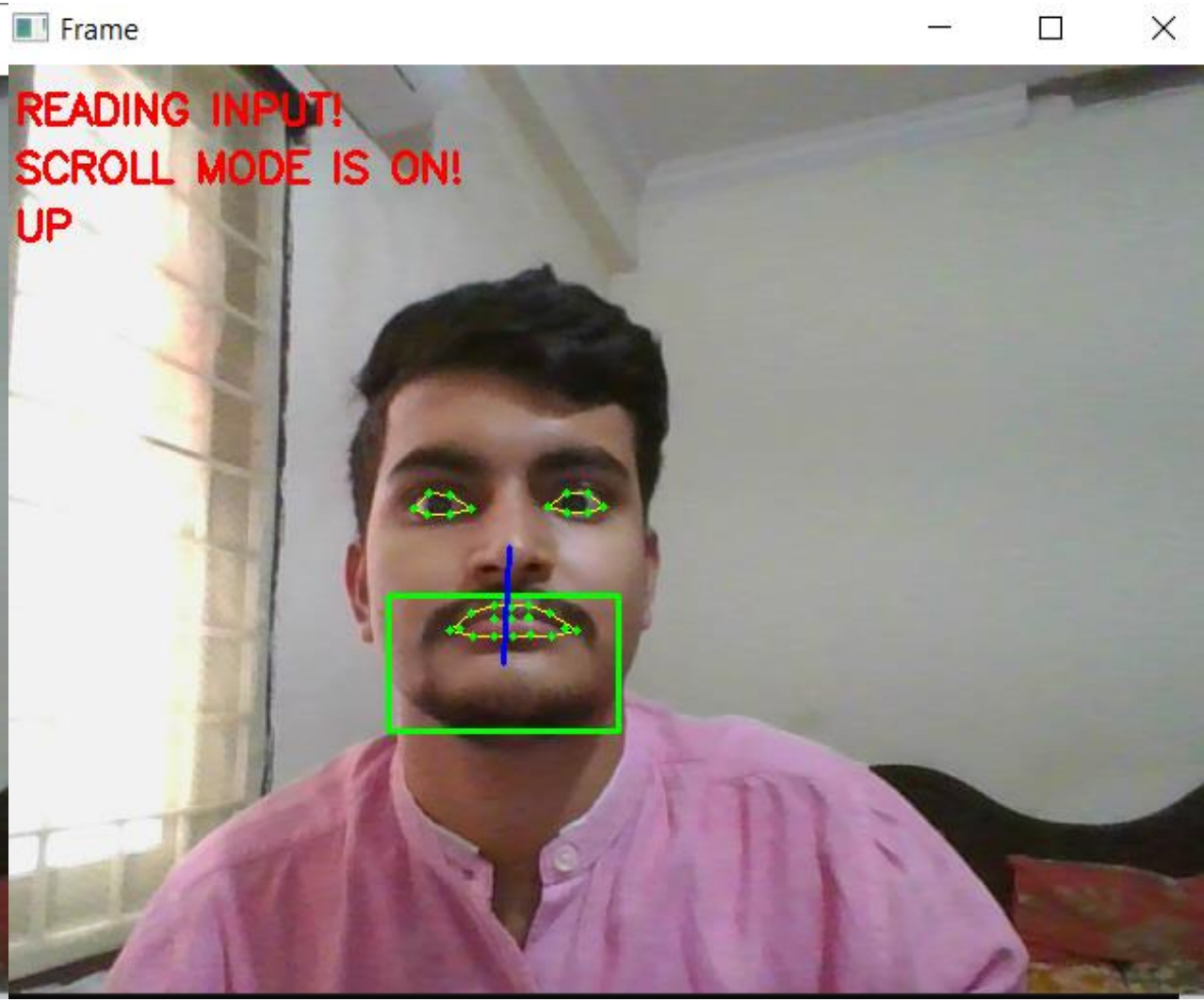
- Hard Disk: 1 TB
- Processor: intel core i5
- Ram: 8 GB

SOFTWARE REQUIREMENTS

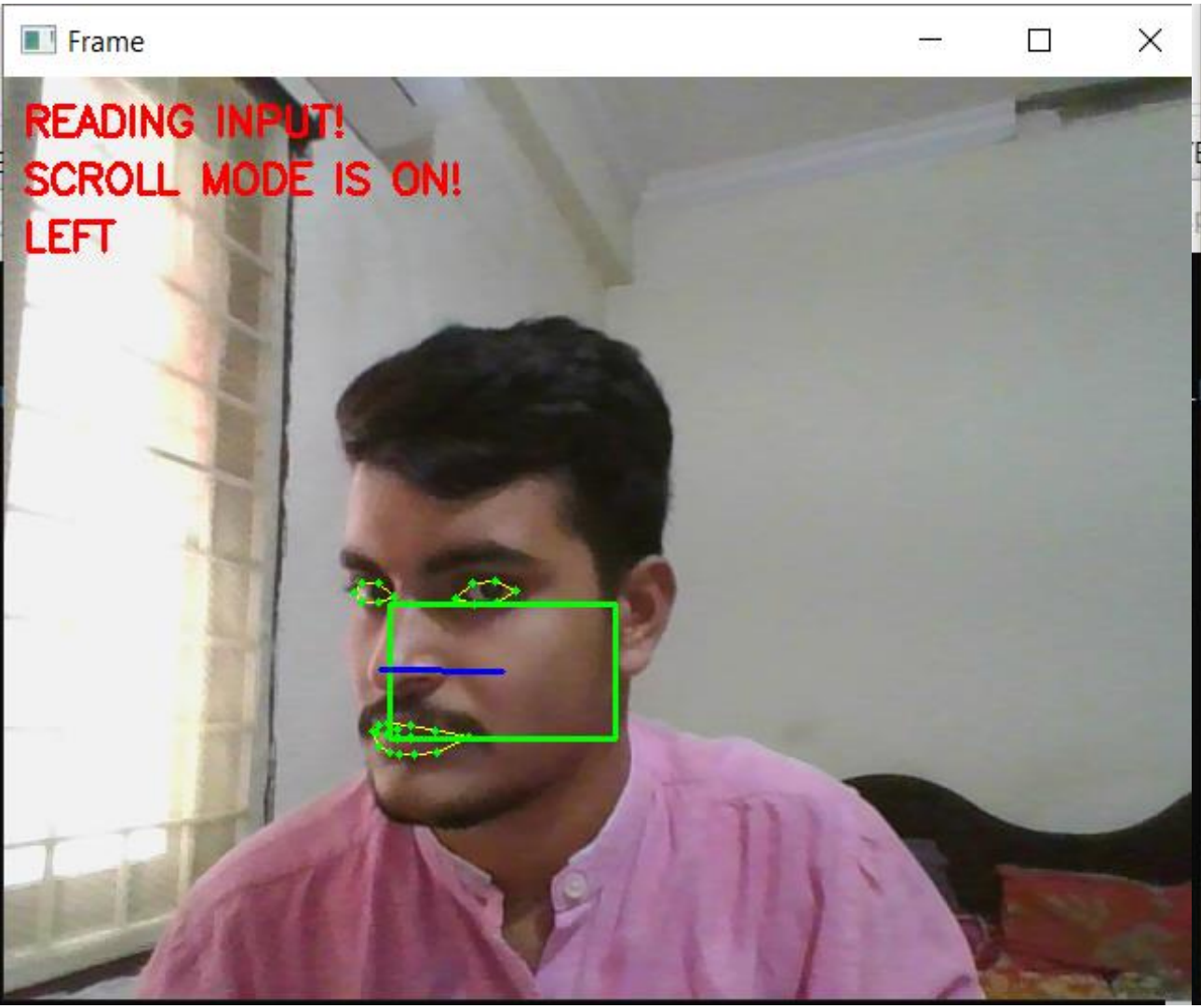
- Operating System: Windows 10
- OPENCV
- PYTHON

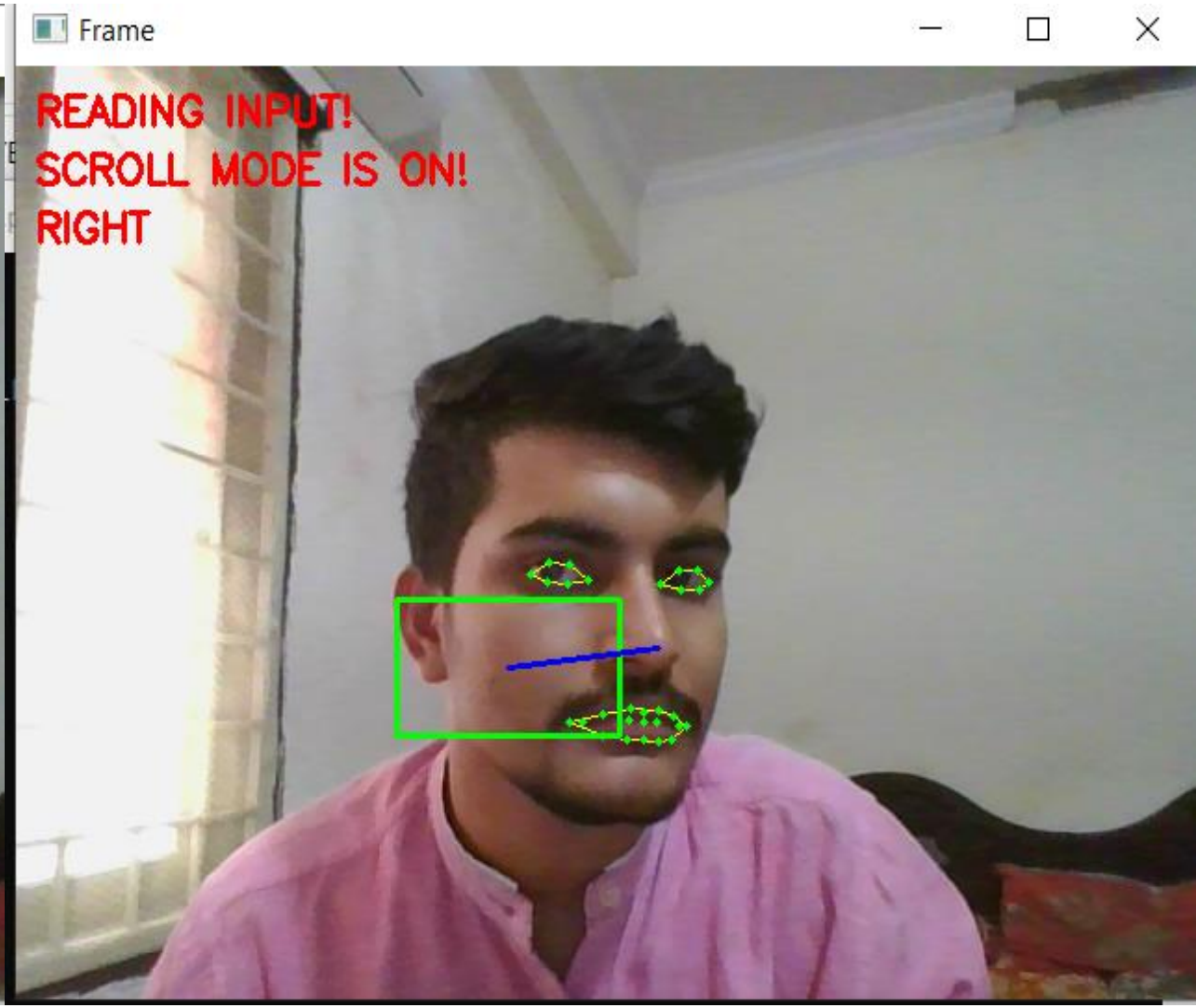
EXPERIMENTAL RESULT





EXPERIMENTAL RESULT





PERFORMANCE EVALUATION

➢Before the set of tests was applied to the subjects, the eye-tracking device was reset individually and a calibration procedure was carried out. This process required the subjects, once seated on the test chair, to fix their sight alternately on three specific points of the monitor for only a couple of seconds, following the instructions of the supervisor. The time associated with this process was variable, ranging from 5 to 15 minutes depending on the user and the eye-tracker adjustments needed. The precision of the transformation between the reference systems of the additional camera and the monitor depends only on the way in which the system detects the infrared LEDs around the monitor edge, either by changing the camera's focus or improving the filtering, segmentation and visual detection methods used by the system.

CONCLUSION & FUTURE WORK

- First detect pupil center position of eye. Then the different variation on pupil position get different command set for virtual keyboard. The signals pass the motor driver to interface with the virtual keyboard itself. The motor driver will control both speed and direction to enable the virtual keyboard to move up, down, left, right and stop. The captured frames which can be already in RGB mode are transformed into Black 'n' White. Five. Pics (frames) from the enter supply focusing the eye are analysed for Iris detection (middle of eye).
- In the future we can also add new functions which can be operable in useful circumstances to control the cursor by the user and implement this system on platforms like mobile phones, tablet etc. In the future, we can also develop a series of operational units so that we can attain a fully operating experience for the handlers from turning on to turning off the computer system.

REFERENCE

1. Lee, Jun-Seok, Kyung-hwa Yu, Sang-won Leigh, Jin-Yong Chung, and Sung-Goo Cho. "Method for controlling device on the basis of eyeball motion, and device therefor." U.S. Patent 9,864,429, issued January 9, 2018.

2. Hossain, Zakir, Md Maruf Hossain Shuvo, and Prionjit Sarker. "Hardware and software implementation of real time electrooculogram (EOG) acquisition system to control computer cursor with eyeball movement." In 2017 4th International Conference on Advances in Electrical Engineering (ICAEE), pp. 132-137. IEEE, 2017.

3. SunitaBarve, DhavalDholakiya, Shashank Gupta, DhananjayDhatrak, “Facial Feature Based Method For Real Time Face Detection and Tracking I-CURSOR”, International Journal of EnggResearchand App., Vol. 2, pp. 1406-1410, Apr (2012).

4. Yu-Tzu Lin Ruei-Yan Lin Yu-Chih Lin Greg C Lee “Real-time eye-gaze estimation using a low-resolution webcam”, Springer, pp.543-568, Aug (2012).

5. Samuel Epstein-Eric MissimerMargritBetke “Using Kernels for avideo- based mouse-replacement interface”, Springer link, Nov (2012)

10.REFERENCE

6. Lee, Po-Lei, Jyun-Jie Sie, Yu-Ju Liu, Chi-Hsun Wu, Ming-Huan Lee, Chih- Hung Shu, Po-Hung Li, Chia-Wei Sun, and Kuo-Kai Shyu. "An SSVEP- actuated brain computer interface using phase-tagged flickering sequences: a cursor system." Annals of biomedical engineering 38, no. 7 (2010): 2383- 2397

7. EniChul Lee Kang Ryoung Park “A robust eye gaze tracking methodbased on a virtual eyeball model”, Springer, pp.319-337, Apr (2008).

8. John J. Magee, MargritBetke, James Gips, Matthew R. Scott, and Benjamin N.Waber“A Human-Computer Interface Using Symmetry Between Eyes to Detect Gaze Direction” IEEE Trans, Vol. 38, no.6,pp.1248-1259, Nov (2008).

9. Jilin Tu, Thomas Huang, Elect and Comp EngrDept, Hai Tao, ElectEnggDept, “Face as Mouse through Visual Face Tracking”,IEEE,(2005).

CONTRIBUTIONS OF THE PROJECT

- In our proposed system, we have developed Human Computer Interaction (HCI), which is the cursor movement of computer is controlled by eye ball movement using python with opencv library.
- If eye moves Left – Cursor moves left
- If eye moves Right – Cursor moves right
- If eye moves Up – Cursor moves up
- If eye moves Down – Cursor moves down

SAMPLE CODE

```
from imutils import face_utils
from utils import *
import numpy as np
import pyautogui as pag
import imutils
import dlib
import cv2

# Thresholds and consecutive frame length for triggering the mouse action.
MOUTH_AR_THRESH = 0.6
MOUTH_AR_CONSECUTIVE_FRAMES = 15
EYE_AR_THRESH = 0.19
EYE_AR_CONSECUTIVE_FRAMES = 15
WINK_AR_DIFF_THRESH = 0.04
WINK_AR_CLOSE_THRESH = 0.19
WINK_CONSECUTIVE_FRAMES = 10

# Initialize the frame counters for each action as well as
# booleans used to indicate if action is performed or not
MOUTH_COUNTER = 0
EYE_COUNTER = 0
WINK_COUNTER = 0
INPUT_MODE = False
EYE_CLICK = False
EYE_CLICK = False
LEFT_WINK = False
RIGHT_WINK = False
SCROLL_MODE = False
ANCHOR_POINT = (0, 0)
WHITE_COLOR = (255, 255, 255)
YELLOW_COLOR = (0, 255, 255)
RED_COLOR = (0, 0, 255)
GREEN_COLOR = (0, 255, 0)
BLUE_COLOR = (255, 0, 0)
BLACK_COLOR = (0, 0, 0)

# Initialize Dlib's face detector (HOG-based) and then create
# the facial landmark predictor
shape_predictor = "model/shape_predictor_68_face_landmarks.dat"
detector = dlib.get_frontal_face_detector()
predictor = dlib.shape_predictor(shape_predictor)

# Grab the indexes of the facial landmarks for the left and
# right eye, nose and mouth respectively
(lStart, lEnd) = face_utils.FACIAL_LANDMARKS_IDXS["left_eye"]
(rStart, rEnd) = face_utils.FACIAL_LANDMARKS_IDXS["right_eye"]
(nStart, nEnd) = face_utils.FACIAL_LANDMARKS_IDXS["nose"]
(mStart, mEnd) = face_utils.FACIAL_LANDMARKS_IDXS["mouth"]
```

SAMPLE CODE

```
(mStart, mEnd) = face_utils.FACIAL_LANDMARKS_IDXS["mouth"]

# Video capture
vid = cv2.VideoCapture(0)
resolution_w = 1366
resolution_h = 768
cam_w = 640
cam_h = 480
unit_w = resolution_w / cam_w
unit_h = resolution_h / cam_h

while True:
    # Grab the frame from the threaded video file stream, resize
    # it, and convert it to grayscale
    # channels)
    _, frame = vid.read()
    frame = cv2.flip(frame, 1)
    frame = imutils.resize(frame, width=cam_w, height=cam_h)
    gray = cv2.cvtColor(frame, cv2.COLOR_BGR2GRAY)

    # Detect faces in the grayscale frame
    rects = detector(gray, 0)

    # Loop over the face detections
```

```
# Loop over the face detections
if len(rects) > 0:
    rect = rects[0]
else:
    cv2.imshow("Frame", frame)
    key = cv2.waitKey(1) & 0xFF
    continue

# Determine the facial landmarks for the face region, then
# convert the facial landmark (x, y)-coordinates to a NumPy
# array
shape = predictor(gray, rect)
shape = face_utils.shape_to_np(shape)

# Extract the left and right eye coordinates, then use the
# coordinates to compute the eye aspect ratio for both eyes
mouth = shape[mStart:mEnd]
leftEye = shape[lStart:lEnd]
rightEye = shape[rStart:rEnd]
nose = shape[nStart:nEnd]

# Because I flipped the frame, left is right, right is left.
temp = leftEye
leftEye = rightEye
```

SAMPLE CODE

```
temp = leftEye
leftEye = rightEye
rightEye = temp

# Average the mouth aspect ratio together for both eyes
mar = mouth_aspect_ratio(mouth)
leftEAR = eye_aspect_ratio(leftEye)
rightEAR = eye_aspect_ratio(rightEye)
ear = (leftEAR + rightEAR) / 2.0
diff_ear = np.abs(leftEAR - rightEAR)

nose_point = (nose[3, 0], nose[3, 1])

# Compute the convex hull for the left and right eye, then
# visualize each of the eyes
mouthHull = cv2.convexHull(mouth)
leftEyeHull = cv2.convexHull(leftEye)
rightEyeHull = cv2.convexHull(rightEye)
cv2.drawContours(frame, [mouthHull], -1, YELLOW_COLOR, 1)
cv2.drawContours(frame, [leftEyeHull], -1, YELLOW_COLOR, 1)
cv2.drawContours(frame, [rightEyeHull], -1, YELLOW_COLOR, 1)

for (x, y) in np.concatenate((mouth, leftEye, rightEye), axis=0):
    cv2.circle(frame, (x, y), 2, GREEN_COLOR, -1)
```

```
for (x, y) in np.concatenate((mouth, leftEye, rightEye), axis=0):
    cv2.circle(frame, (x, y), 2, GREEN_COLOR, -1)

# Check to see if the eye aspect ratio is below the blink
# threshold, and if so, increment the blink frame counter
if diff_ear > WINK_AR_DIFF_THRESH:

    if leftEAR < rightEAR:
        if leftEAR < EYE_AR_THRESH:
            WINK_COUNTER += 1

            if WINK_COUNTER > WINK_CONSECUTIVE_FRAMES:
                pag.click(button='left')

                WINK_COUNTER = 0

    elif leftEAR > rightEAR:
        if rightEAR < EYE_AR_THRESH:
            WINK_COUNTER += 1

            if WINK_COUNTER > WINK_CONSECUTIVE_FRAMES:
                pag.click(button='right')

                WINK_COUNTER = 0
```

SAMPLE CODE

```
                WINK_COUNTER = 0
    else:
        WINK_COUNTER = 0
else:
    if ear <= EYE_AR_THRESH:
        EYE_COUNTER += 1

        if EYE_COUNTER > EYE_AR_CONSECUTIVE_FRAMES:
            SCROLL_MODE = not SCROLL_MODE
            # INPUT_MODE = not INPUT_MODE
            EYE_COUNTER = 0

            # nose point to draw a bounding box around

    else:
        EYE_COUNTER = 0
        WINK_COUNTER = 0

if mar > MOUTH_AR_THRESH:
    MOUTH_COUNTER += 1

    if MOUTH_COUNTER >= MOUTH_AR_CONSECUTIVE_FRAMES:
        # if the alarm is not on, turn it on
        INPUT_MODE = not INPUT_MODE
        INPUT_MODE = not INPUT_MODE
        # SCROLL_MODE = not SCROLL_MODE
        MOUTH_COUNTER = 0
        ANCHOR_POINT = nose_point


else:
    MOUTH_COUNTER = 0

if INPUT_MODE:
    cv2.putText(frame, "READING INPUT!", (10, 30), cv2.FONT_HERSHEY_SIMPLEX, 0.7, RED_COLOR, 2)
    x, y = ANCHOR_POINT
    nx, ny = nose_point
    w, h = 60, 35
    multiple = 1
    cv2.rectangle(frame, (x - w, y - h), (x + w, y + h), GREEN_COLOR, 2)
    cv2.line(frame, ANCHOR_POINT, nose_point, BLUE_COLOR, 2)

    dir = direction(nose_point, ANCHOR_POINT, w, h)
    cv2.putText(frame, dir.upper(), (10, 90), cv2.FONT_HERSHEY_SIMPLEX, 0.7, RED_COLOR, 2)
    drag = 18
    if dir == 'right':
        pag.moveRel(drag, 0)
    elif dir == 'left':
        pag.moveRel(-drag, 0)
```

SAMPLE CODE

```
        pag.moveRel(-drag, 0)
    elif dir == 'up':
        if SCROLL_MODE:
            pag.scroll(40)
        else:
            pag.moveRel(0, -drag)
    elif dir == 'down':
        if SCROLL_MODE:
            pag.scroll(-40)
        else:
            pag.moveRel(0, drag)

if SCROLL_MODE:
    cv2.putText(frame, 'SCROLL MODE IS ON!', (10, 60), cv2.FONT_HERSHEY_SIMPLEX, 0.7, RED_COLOR, 2)

# cv2.putText(frame, "MAR: {:.2f}".format(mar), (500, 30),
#             cv2.FONT_HERSHEY_SIMPLEX, 0.7, YELLOW_COLOR, 2)
# cv2.putText(frame, "Right EAR: {:.2f}".format(rightEAR), (460, 80),
#             cv2.FONT_HERSHEY_SIMPLEX, 0.7, YELLOW_COLOR, 2)
# cv2.putText(frame, "Left EAR: {:.2f}".format(leftEAR), (460, 130),
#             cv2.FONT_HERSHEY_SIMPLEX, 0.7, YELLOW_COLOR, 2)
# cv2.putText(frame, "Diff EAR: {:.2f}".format(np.abs(leftEAR - rightEAR)), (460, 80),
#             cv2.FONT_HERSHEY_SIMPLEX, 0.7, (0, 0, 255), 2)
```

```
    # Show the frame
    cv2.imshow("Frame", frame)
    key = cv2.waitKey(1) & 0xFF

    # If the `Esc` key was pressed, break from the loop
    if key == 27:
        break

# Do a bit of cleanup
cv2.destroyAllWindows()
vid.release()
```

» THANK YOU

A

PROJECT REPORT

On

Analysis Of Women Safety In Indian Cities Using Machine Learning On Tweets

Submitted by

1. **Ms. C. Deeksha Reddy (17K81A0507)**
2. **Ms. Ch. Sai Prasanna (17K81A0509)**
3. **Ms. T. Shreya (17K81A0552)**
4. **Ms. V. Meghana Reddy (17K81A0555)**

in partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Under the Guidance of

Mr. Uppula Nagaiah

Assistant Professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ST.MARTIN'S ENGINEERING COLLEGE
An Autonomous Institute

Dhulapally, Secunderabad – 500 100

JUNE 2021

BONAFIDE CERTIFICATE

This is to certify that the project entitled Analysis Of Women Safety In Indian Cities Using Machine Learning On Tweets , is being submitted by **Ms. C. Deeksha Reddy 17K81A0507, Ms.Ch. Prasanna 17K81A0509**, **Ms. T. Shreya 17K81A0552, Ms. V. Meghana Reddy 17K81A0555** in partial fulfillment of the requirement for the award of the degree of **BACHELOR OF TECHNOLOGY IN** Computer Science and Engineering is recorded of bonafide work carried out by them. The result embodied in this report have been verified and found satisfactory.

Under the guidance of
Mr. UPPULA NAGAIAH
Department of CSE

Head of the Department
Dr.M.NARAYANAN
Department of CSE

Internal Examiner

External Examiner

Place:

Date:

DECLARATION

We, the student of **Bachelor of Technology** in Department of Computer Science and Engineering', session: 2017 – 2021, St. Martin's Engineering College, Dhulapally, Kompally, Secunderabad, hereby declare that work presented in this Project Work entitled Analysis Of Women Safety In Indian Cities Using Machine Learning On Tweets is the outcome of our own bonafide work and is correct to the best of our knowledge and this work has been undertaken taking care of Engineering Ethics. This result embodied in this project report has not been submitted in any university for award of any degree.

Ms. C. Deeksha Reddy (17K81A0552)
Ms. Ch. Sai Prasanna (17K81A0509)
Ms. T. Shreya (17K81A0552)
Ms. V. Meghana Reddy (17K81A0555)

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible and whose encouragements and guidance have crowded effects with success.

We extended our deep sense of gratitude to Principal**, Dr. P. SANTOSH KUMAR PATRA**, St. Martin's Engineering College, Dhulapally, for permitting us to undertake this project.

We are also thankful to **Dr. M.NARAYANAN**, Head of the Department, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

We would like to thank **Dr. T. POONGOTHAI,** Department of Computer Science and Engineering (AI & ML) for her encouragement and insightful comments and as well as our project coordinators **Dr. B.RAJALINGAM**, Associate Professor and **Dr. G. GOVINDARAJULU,** Associate Professor, in Department of Computer Science and Engineering for their valuable support.

We would like to express our sincere gratitude and indebtedness to our project supervisor Mr. UPPULA NAGAIHA, Assistant Professor, in Department of Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

Finally, we express thanks to all those who have helped us successfully to completing this project. Furthermore, we would like to thank our family and friends for their moral support and encouragement.

We express thanks to all those who have helped us in successfully completing the project.

Ms. C. Deeksha Reddy (17K81A0507)
Ms. Ch. Sai Prasanna (17K81A0509)
Ms. T . Shreya (17K81A0552)
Ms. V. Meghana Reddy (17K81A0555)

ABSTRACT

In terms of women's security, we are living in the worst time our society has ever seen .Women from various parts of the world always experience a lot of harassment, starting from stalking, passing vulgar comments, and leading to sexual assault. The main motive of the project is to analyse women safety using social networking messages and by applying machine learning algorithms on it. Now-a-days almost all people are using social networking sites to express their feelings and if any women feel unsafe in any area then she will express negative words in her post/tweets/messages and by analysing those messages we can detect which area is more unsafe for women.

In this paper we focus on how social media is used to promote the safety of women in Indian cities from various social media platforms such as Twitter ,Facebook and Instagram. Tweets consists of text messages, audio data ,video data, images, smiley expressions and hash-tags .The content being shared can be used to educate many people to raise their voice if any abusive language or any harassment is done against women. Hashtags used by Instagram and Twitter can be used to convey one's thoughts across the globe and make the women feel free to express their views and feelings.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF OUTPUT SCREENS

1. INTRODUCTION

Twitter in this modern era has emerged as an ultimate micro-blogging social network consisting over hundred million users. Twitter is an informative source for all the zones like institutions, companies and organizations.In addition to this, many people express their opinions by using abbreviations, slang, shot forms, emoticons, and sarcasm also. Hence twitter language can be termed as the unstructured .From the tweet, the sentiment behind the message is extracted. This extraction is done by using the sentimental analysis procedure.

On the twitter, users will share their opinions and perspective in the tweets section. This tweet can only contain 140 characters, thus making the users to compact their messages with the help of abbreviations, slang, shot forms, emoticons, etc. In addition to this, many people express their opinions by using polysemy and sarcasm also. Hence twitter language can be termed as the unstructured. From the tweet, the sentiment behind the message is extracted. This extraction is done by using the sentimental analysis procedure. Results of the sentimental analysis can be used in many areas like sentiments regarding a particular brand or release of a product, analyzing public opinions on the government policies, people thoughts on women, etc. In order to perform classification of tweets and analyze the outcome, a lot of study has been done on the data obtained by the twitter. We also review some studies on machine learning in this paper and research on how to perform sentimental analysis using that domain on twitter data. The paper scope is restricted to machine learning algorithm and models.

Staring at women and passing comments can be certain types of violence and harassments and these practices, which are unacceptable, are usually normal especially on the part of urban life. Many researches that have been conducted in India shows that women have reported sexual harassment and other practices as stated above. Such studies have also shown that in popular metropolitan cities like Delhi, Pune, Chennai and Mumbai, most women feel they are unsafe when surrounded by unknown people. On social media, people can freely express what they feel about the Indian politics, society and many other thoughts. Similarly, women can also share their experiences if they have faced any violence or sexual harassment and this brings innocent people together in order to stand up against such incidents. From the analysis of tweets text collection obtained by the twitter, it includes names of people who has harassed the women and also names of women or innocent people who have stood against such violent acts or unethical behaviour of men and thus making them uncomfortable to walk freely in public.

The data set of the tweet will be used to process the machine learning algorithms and models. This algorithm will perform smoothening the tweet data by eliminating zero values. Using Laplace and porter's theory, a method is developed in order to analyze the tweet data and remove redundant information from the data set. Huge numbers of people have been attracted to social media platform such

as Twitter, Facebook, Instagram. People express their sentiments about society, politics, women, etc via the text messages, emoticons and hash-tags through such platforms. There are some methods of sentiment that can be classified like machine leaning based and lexicon based learning.

1.1 Project overview

We focus on the women safety by taking the advantage of machine learning algorithm i.e SVM(support vector machines).first we download the MEETOO tweets on women safety and save inside the dataset folder using NLTK(natural language toolkit) it removes special symbols and stop words from tweets and makes them clean. Also use TEXTBLOB corpora package and dictionary we count positive ,negative or neutral polarity also make use of sentiment analysis.

Support Vector Machine Algorithm:

Using support vector machine classifier, we plot each data item as a point in n-dimensional space (where n is number of features) with the value of each feature being the value of a particular coordinate. Then, classification is performed by finding the hyper-plane that differentiates the classes very well. It draws the hyper-plane by transforming the data with the help of mathematical functions called Kernels. Types of Kernels are linear, sigmoid, RBF, non-linear, polynomial etc. Support Vectors are simply the co-ordinates of individual observation. Intuitively, a good separation is achieved by the hyper plane that has the largest distance to the nearest training-data point of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier. Steps needed to build the model include:

- Gathering perfect data for training and testing.
- Vectorizing the data.
- Creating a Linear SVM Model to train and then test it

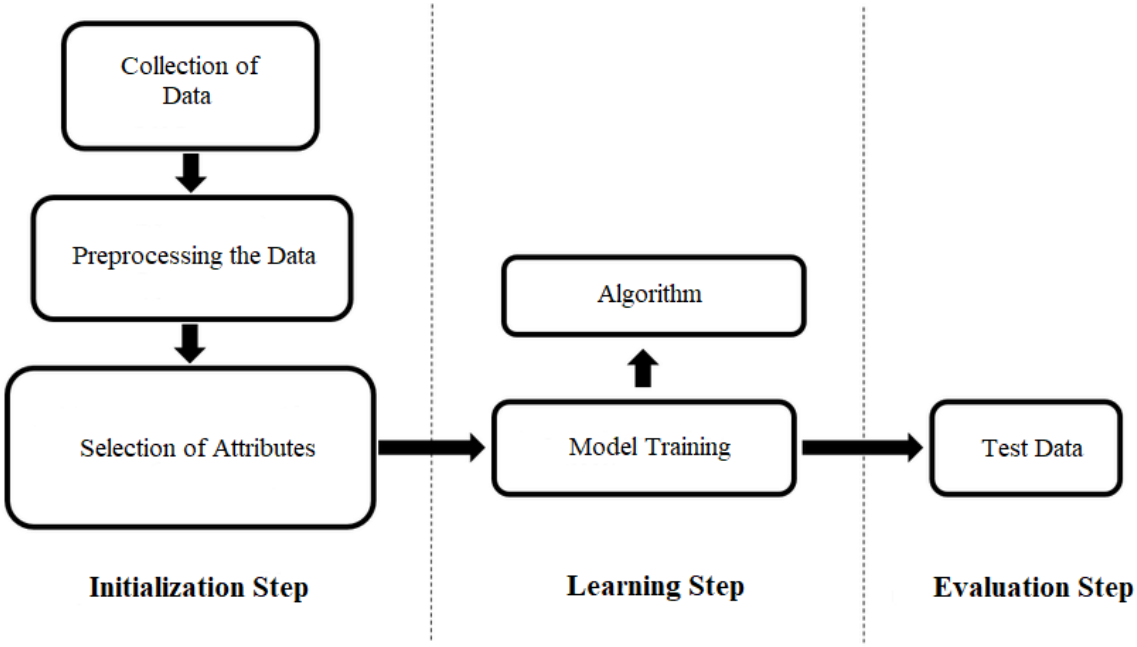


Figure 1.1: Process of analysis

1.2 Project objective:

To analyse women safety using social networking messages and by applying machine learning algorithms on it .By analysing the messages we can detect which area is more unsafe for women and act accordingly.

1.3 Organisation of chapters:

1.3.1 Introduction:Here we learn about the project functioning, its overview and objectives how social networking platforms help women to express their feelings and the objectives of the project

1.3.2 Literature Survey : Here we read around the topic and have a broad understanding of previous research ,including its limitations ,in this we also summarize the main viewpoints and important facts that we encountered in our reading as they relate to our topic.

1.3.3 Software and Hardware Requirements : Here we learn about system requirements or the required specifications a device must have in order to use certain hardware or software.

1.3.4 Software Development Analysis : Here we learn about how a women faces problems and what she has to go through on social networks based on overview of the problem ,defining the problem and the modules we use in our project are admin, user then defining the module and its functionalities.

1.3.5 Project System Design : Here we learn about the dataflow diagrams of the modules i.e user ,admin and ER-Diagrams and UML Diagrams (class, components, deployment, object, activity, sequence, collaboration , state chart diagram)

1.3.6 Project Code : Here we had the opportunity to adopt in detail about the language we used and the various parameters used in this Here the design of input is made to focus on controlling the amount of input required, controlling the errors ,avoiding delay ,avoiding extra steps and keeping the process simple.

1.3.7 Project Testing : Here the purpose of testing is to discover errors .It provides a way to check the functionality of components ,sub assemblies or a finished product. There are various types of test and each test type addresses a specific testing requirements and various test cases .Black box testing and white box testing is performed here we rectify errors and test the code and process the final result.

1.3.8 Output Screens: This is the final output of the whole project here all the ideas and implementations done are represented on the web page. Here the tweets are categorized into negative and positive polarity rates. so that the graph is obtained of how the tweets are managed.

1.3.9 Experimental Results: We have implemented the system in various localities, and these were the results from the implemented system. We have analyzed tweets from various regions and created various datasets to work on, and then divided them according to their polarities.

1.3.10 Conclusion and Future Enhancement: Since the data is large, we used support vector algorithm and TEXTBLOB to achieve sentimental analysis based on which we categorized the places which are safe and unsafe for women. Due to the usage of the algorithm we could analysis the sentiment behind the tweets of women accurately up to 88.3% based on the MeToo dataset.Since the accuracy is only 88.3% , there is a scope to increase the accuracy by using better models. For the future enhancement, we can extend to apply these machine learning algorithms on different social media platforms like Facebook and Instagram also since in our project only twitter is considered.

2. LITERATURE SURVEY

2.1 Survey on background

[1**]** Concept to analyse women safety using social networking messages and by applying machine learning algorithms on it. Now-a-days almost all people are using social networking sites to express their feelings and if any women feel unsafe in any area then she will express negative words in her post /tweets/ messages. [2]In order to perform sentiment analysis, we are required to collect data from the desired source (here Twitter).[3] Besides taking additional pre-processing measures like the expansion of net lingo and removal of duplicate tweets[4] Machine learning techniques are used by the well-known Python library NLTK (Natural Language Toolkit) and, another NLP library, Text blob , provides both types.[5] Sentiment analysis on Twitter data has attracted much attention recently. In this paper, we focus on target-dependent Twitter sentiment classification; namely, given a query .we classify the sentiments of the tweets as positive, negative or neutral according to whether they contain positive, negative or neutral sentiments about that query.

[6] In this paper, we investigate the utility of linguistic features for detecting the sentiment of Twitter messages. We evaluate the usefulness of existing lexical resources as well as features that capture information about the informal and creative language used in microblogging.[7] We propose an approach to automatically detect sentiments on Twitter messages (tweets) that explores some characteristics of how tweets are written and meta-information of the words that compose these messages.[8] We find classifying sentiment in microblogs easier than in blogs and make a number of observations pertaining to the challenge of supervised learning for sentiment analysis in microblogs.[9] In this paper, we focus on using Twitter, the most popular microblogging platform, to perform sentiment analysis. We learn how to automatically collect a corpus for sentiment analysis and opinion mining purposes. We perform linguistic analysis of the collected corpus and explain discovered phenomena.[10] We present a classifier to predict contextual polarity of subjective phrases in a sentence. [11] In this paper a novel approach for automatically classifying the sentiment of Twitter messages. These messages are classified as either positive or negative with respect to a query term. We present the results of machine learning algorithms for classifying the sentiment of Twitter messages using distant supervision. Our training data consists of Twitter messages with emoticons, which are used as noisy labels. We show that machine learning algorithms (Naive Bayes, Maximum Entropy, and SVM) have accuracy above 80% when trained with emoticon data.[12] By recovering the n-best parses using coarse to fine parsing.

[13] Traditional machine learning methods such naïve bayes, logistic regression and support vector machine(SVM) are widely used for large scale because they scale well.[14] Identifying sentiments (the affective parts of opinions) is a challenging problem. We present a system that, given a topic,

automatically finds the people who hold opinions about that topic and the sentiment of each opinion. The system contains a module for determining word sentiment and another for combining sentiments within a sentence.[15] The standard CKY(cocke-younger-kasami)algorithm is used for parsing PCFG .It is bottom up and make use of dynamic programming.

2.2 Conclusions on survey

From[1] we can conclude that by doing the analysis on those messages we can be able to detect which area might be more unsafe for women. From[2]we can conclude that the data which we collected undergoes various steps of pre-processing which makes it more machine sensible than its previous form. and by making use of NLTK .From[3] we can conclude that a probabilistic model which is based on Bayes' theorem is used for spelling correction, which is overlooked in other research studies. From[4] we can conclude that from python library NLTK (natural language toolkit)and other NLP library we use various machine learning techniques .From[5]we can conclude that we classify the sentiments of the tweets as positive ,negative or neutral According to the experimental results, our approach greatly improves the performance of target-dependent sentiment classification .From[6] we can conclude that by taking a supervised approach to the problem we evaluate the usefulness of existing lexical resources and leverage existing hashtags in the twitter data for building training data .From[7] we can conclude that the approach which is done automatically detects sentiments on tweets and we explore characteristics of how tweets are written and information is composed. From[8] we can conclude that we perform a number of observations for pertaining to the challenge of supervised learning for sentiment analysis . From[9] we can conclude that through Using the corpus, we build a sentiment classifier, that is able to determine positive, negative and neutral sentiments for a document.

From [10] we can conclude that our approach features lexical scoring derived from the Dictionary of Affect in Language(DAL) and extended through WordNet. From [11] we can conclude that this paper also describes the pre-processing steps needed in order to achieve high accuracy. The main contribution of this paper is the idea of using tweets with emoticons for distant supervised learning . From[12] we can conclude that by recovering the n-best parses using coarse to fine parsing . From[13] we can conclude that logistic regression and support vector machine(SVM) these are widely used for large scale because they scale well . From[14] we can conclude that the experiment done with various models of classifying and combining sentiment at word and sentence levels, with promising results . From[15] we can conclude that Markovization-by examining the vertical and horizontal ancestors of the current node ,tag splitting-internal and external annotation.

3. SOFTWARE AND HARDWARE REQUIREMENTS

3.1 Software Requirements:

- Python
- HTML/CSS
- Anaconda/ Colab
- MySQL

Operating Systems:

- Windows 10

3.2 Hardware Requirements:

- Processor: core i3
- RAM: 4GB

4. SOFTWARE DEVELOPMENT ANALYSIS

4.1 Overview of the problem

Staring at women and passing comments can be certain types of violence and harassments and these practices, which are unacceptable, are usually normal especially on the part of urban life. Many researches that have been conducted in India shows that women have reported sexual harassment and other practices as stated above. Such studies have also shown that in popular metropolitan cities like Delhi, Pune, Chennai and Mumbai, most women feel they are unsafe when surrounded by unknown people. On social media, people can freely express what they feel about the Indian politics, society and many other thoughts. Similarly, women can also share their experiences if they have faced any violence or harassment and this brings innocent people together in order to stand up against such incidents.

The crime rate against women is rapidly increasing in recent times. Many procedures have been introduced for women's safety for example 'SHE' and many other online applications which can be used by women for her safety. But these are not reliable and not much effective as they involve human errors and delays.

4.2 Define the problem

Twitter in this modern era has emerged as a ultimate microblogging social network consisting over hundred million users and generate over five hundred million messages known as 'Tweets' every day. Twitter with such a massive audience has magnetized users to emit their perspective and judgemental about every existing issue and topic of internet, therefore twitter is an informative source for all the zones like institutions, companies and organizations.

On the twitter, users will share their opinions and perspective in the tweets section. Many people express their opinions by using polysemy and sarcasm also. Hence twitter language can be termed as the unstructured. From the tweet, the sentiment behind the message is extracted. This extraction is done by using the sentimental analysis procedure. Results of the sentimental analysis can be used in many areas like sentiments regarding a particular brand or release of a product, analyzing public opinions on the government policies, people thoughts on women, etc. In order to perform classification of tweets and analyze the outcome, a lot of study has been done on the data obtained by the twitter. We also review some studies on machine learning in this paper and research on how to perform sentimental analysis using that domain on twitter data. The paper scope is restricted to machine learning algorithm and models.

From the analysis of tweets text collection obtained by the twitter, it includes names of people who has harassed the women and also names of women or innocent people who have stood against such violent acts or unethical behaviour of men and thus making them uncomfortable to walk freely in public.

The data set of the tweet will be used to process the machine learning algorithms and models. This algorithm will perform smoothening of the tweet data by eliminating zero values. Using Laplace and porter's theory, a method is developed in order to analyse the tweet data and remove redundant information from the data set. Huge numbers of people have been attracted to social media platform such as Twitter, Facebook, Instagram. People express their sentiments about society, politics, women, etc. via the text messages, emoticons and hash-tags through such platforms. There are some methods of sentiment that can be classified like machine leaning based and lexicon based learning.

4.3 Modules Overview

The aim of the project entitled as Analysis of Women Safety in Indian Cities Using Machine Learning on Tweets is to develop a platform where Users can post tweets, images or both and an Admin who manages the platform.

User is the person who can register and login and post tweets and images and also checks the progress of their tweets.

Admin is the person who manages the platform and can view all the users details and feedbacks given by them.

4.4 Defining the Modules

The two main modules of our project are

1.User

2.Admin

1.User

A new User can register in the platform and the existing user can directly login. A new user will have to give all the details asked like Name, Date of Birth, email, Phone Number, Address and password in order to be registered in the platform. Already existing user can directly login by using his email as login ID and password. In order to login the password must be authenticated.A user can post tweets and images and also check the progress of her/his tweet, update her/his personal information, give feedback. The tweets posted are cleansed using Sentimental Analysis.

2.Admin

Admin holds the accountability for managing and platform. After logging in, Admin can perform actions like View User Details, Today's Trending, Analysis Graph, View Feedback and then Logout. All the details of registered users are visible in View User Details. The metropolitan cities with more number of tweets are visible in Today's Trending page. Analysis Graph gives an overview of the negative and positive tweets after analysing the tweets using Sentimental Analysis.

4.5 Module Functionality

People communicate and share their opinion actively on social medias including Facebook and Twitter. Social network can be considered as a perfect platform to learn about people's opinion and sentiments regarding different events and issues. There exists several opinion-oriented information gathering and analytics systems that aim to extract people's opinion regarding different topics. For cleaning the dataset i.e., tweets, we use Sentimental Analysis. Sentiment Analysis is a 5 step procedure.

Implementation of sentiment Analysis

1. Data extraction: First step involved in analysis of sentiment is the collection of information from the social network website like twitter. This helps in extracting the tweet message but this message also includes extra data like tweets likes, dislikes and comments.

2. Text Cleaning: Once the data is extracted from the twitter source as the datasets, this information has to be passed to the classifier. The classifier cleans the dataset by removing redundant data like stop words, emoticons in order to make sure that non textual content is identified and removed before the analysis.

3. Sentiment Analysis: After the classifier cleans the dataset, the data is ready for the sentimental analysis process. Machine learning and Lexicon based learning and Hybrid learning are some of the approaches of sentimental analysis. There are also some other approaches such as Nero Linguistic Programming and Natural Language Processing. Training the dataset and then testing that trained dataset involves in machine learning approach. Training data and Testing data are useful for the classifier to perform the algorithm. Maximum Entropy, Naives Bayes classification, Bayesian Networks and Network Support Vector Machine are some of the algorithm which can be used to train the classifier. Testing data is used to identify the efficiency of the sentiment classifier. In case of Lexicon based leaning, training dataset is not used. This approach uses a built-in dictionary in which words associated with sentiments of human are present. The third approach, which is the Hybrid learning, combines both machine leaning approach and lexicon learning approach in order to improve the performance of classifier.

4. Sentiment Classification: At this step, the dataset is ready for the classification. Each and every sentence of the tweet will be examined and opinion will be formed accordingly for subjectivity.

Subjective expression sentences are retained and those of objective expression sentences are rejected. Techniques like Unigrams, Negation, and Lemmas and so on are used at different levels of sentimental analysis. Sentiments can be distinguished broadly into two groups – Positive and Negative. At this point of sentimental analysis, each of the subjective sentences which will be retained, are classified into good, bad or like, dislike or positive and negative.

5. Output Presentation: To generate useful and meaningful information out of the raw data, sentimental analysis plays vital role. Once the algorithm is completed, the outcome of the analysis can be visualized by creating different types of graphs. Bar graphs, Time series and Pie charts are some of the examples which can be used to display the output. To measure the sentiment of the tweets in terms of Positive and Negative, Bar graphs can be used. Similarly, to measure in terms of likes, dislikes, average length of tweet for a certain period, Time series can be used. To obtain the initial source of the tweet, pie charts can be used.

5. PROJECT SYSTEM DESIGN

5.1 Data Flow Diagrams

a. User

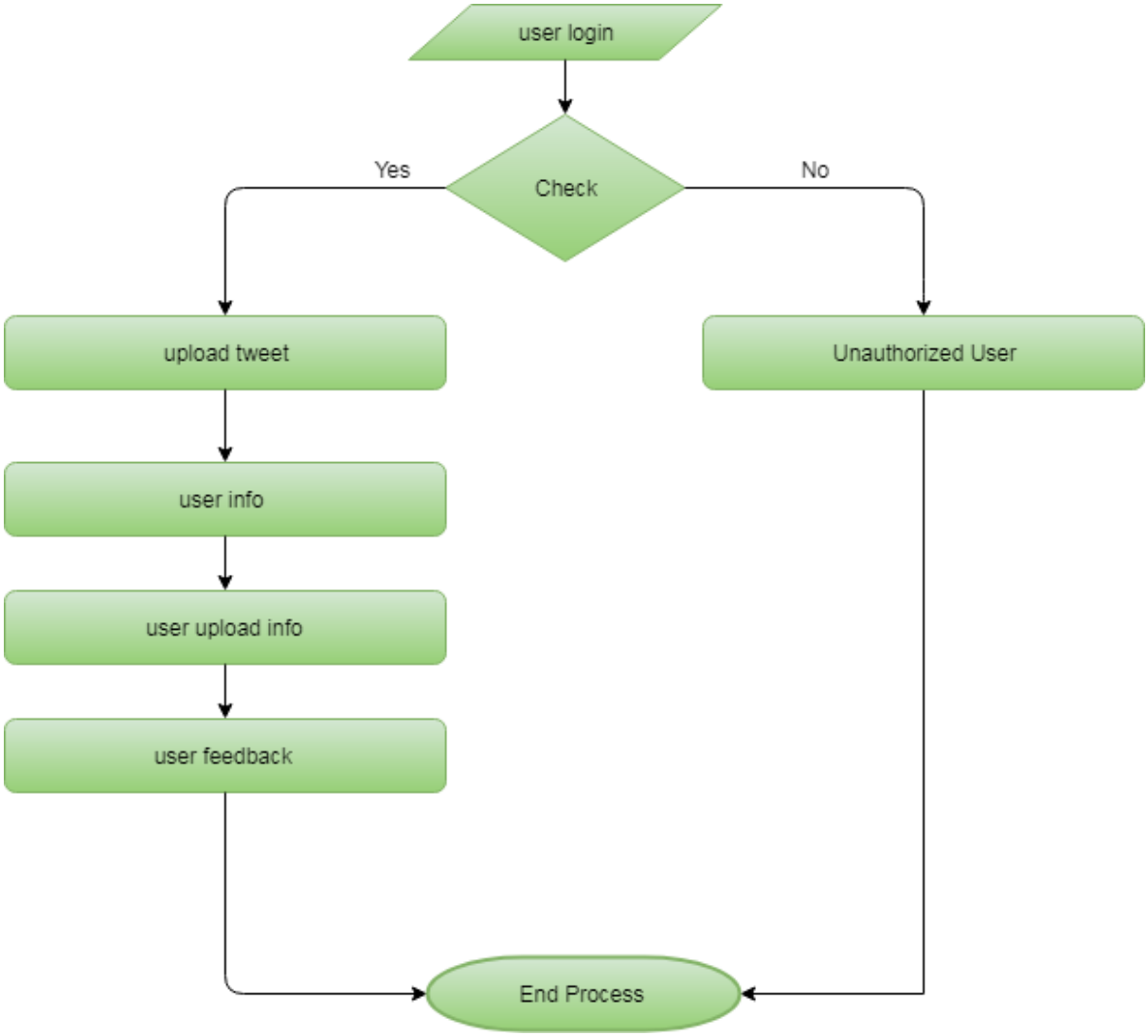


Figure 5.1: Data Flow Diagram- User

b. Admin

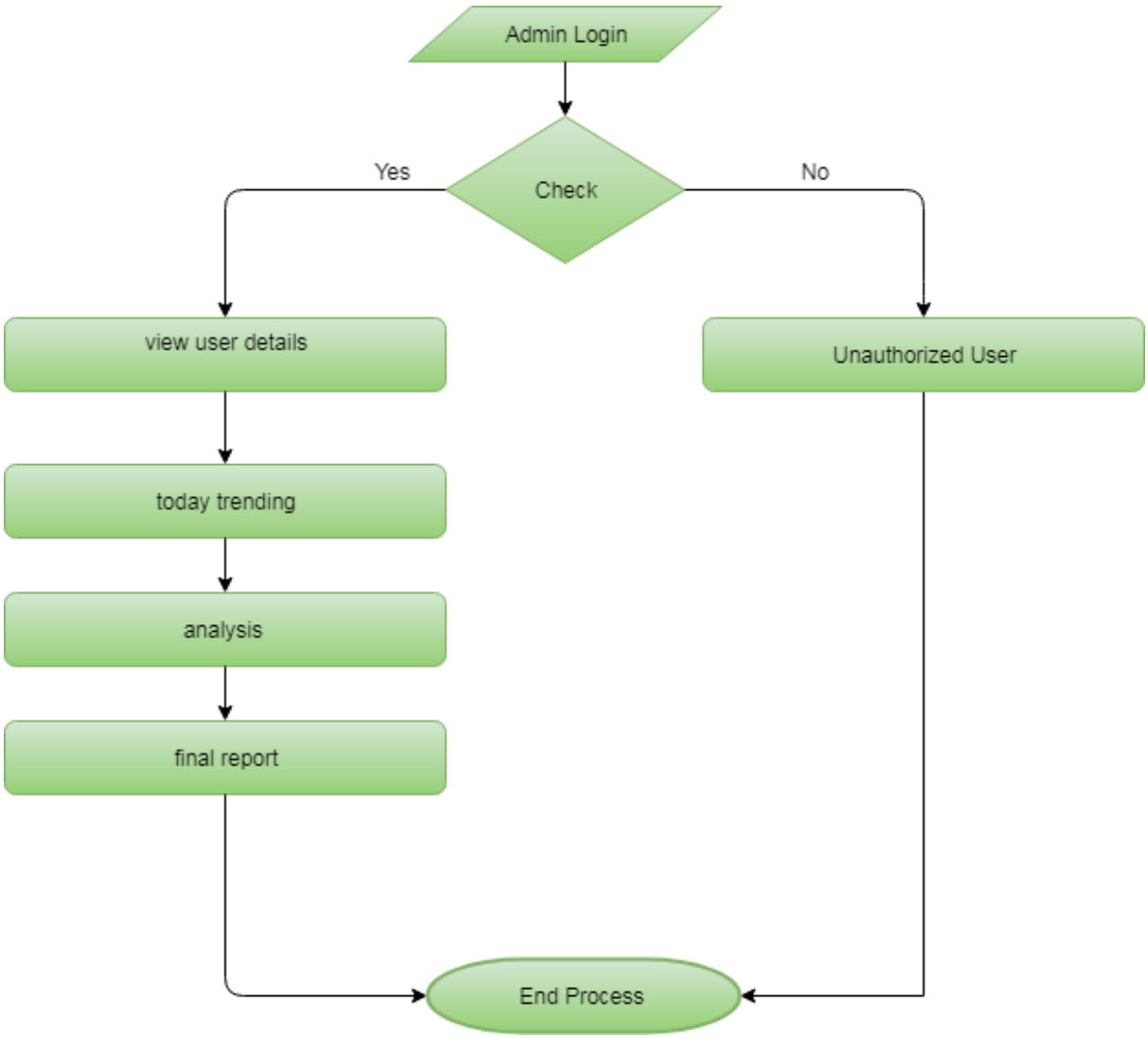


Figure 5.2: Data Flow Diagram-Admin

5.2 E-R Diagrams

a. User

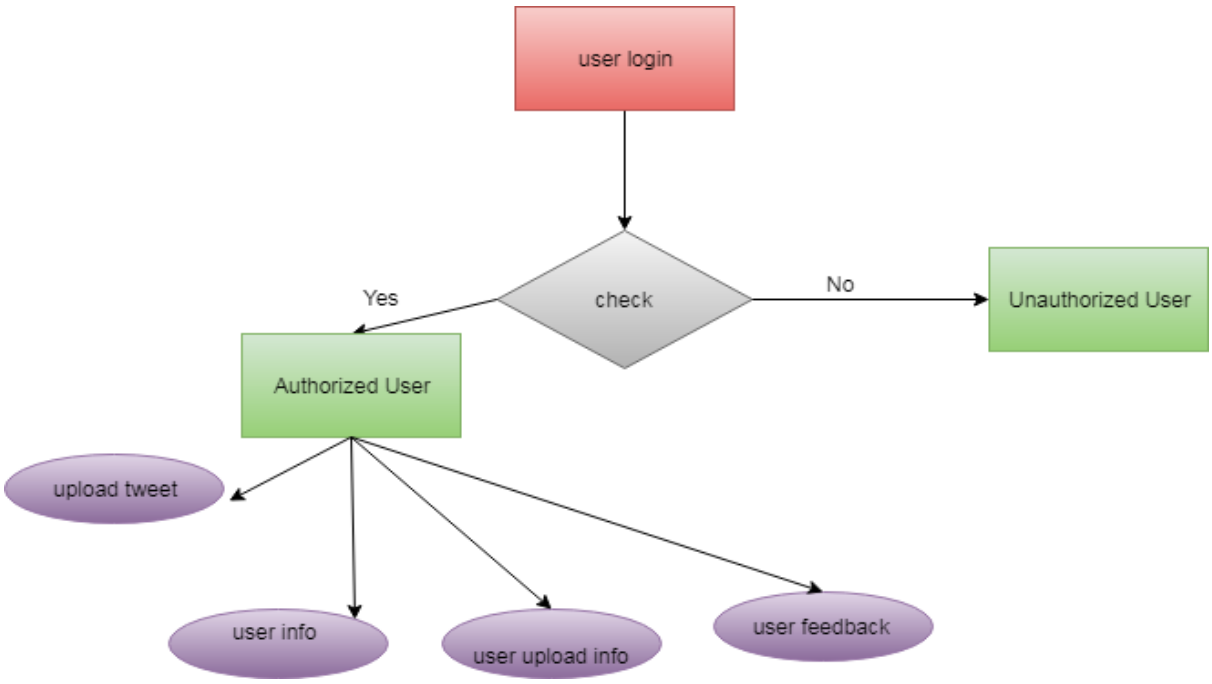


Figure 5.3: E-R Diagram-User

b. Admin

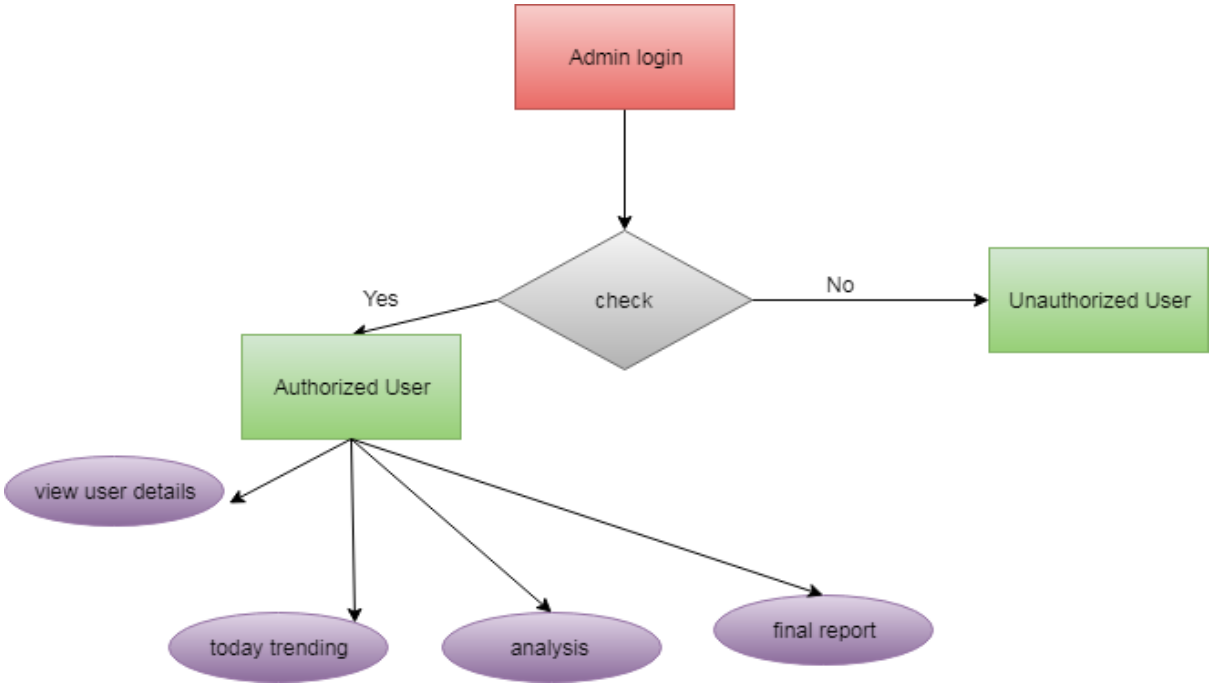


Figure 5.4: E-R Diagram-Admin

5.3 UML Diagrams

5.3.1 Class Diagram

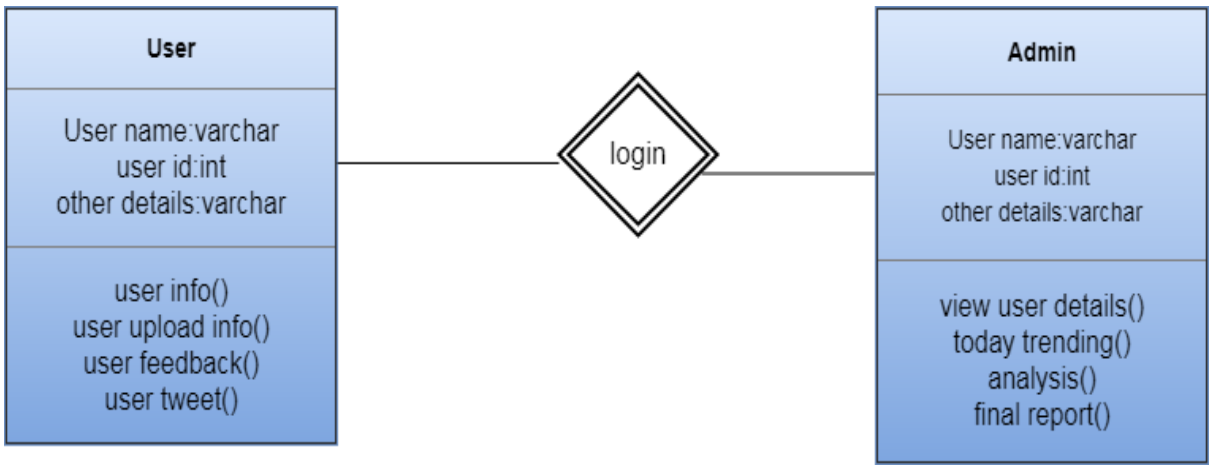


Figure 5.5: Class Diagram

5.3.2 Component Diagram

a. User

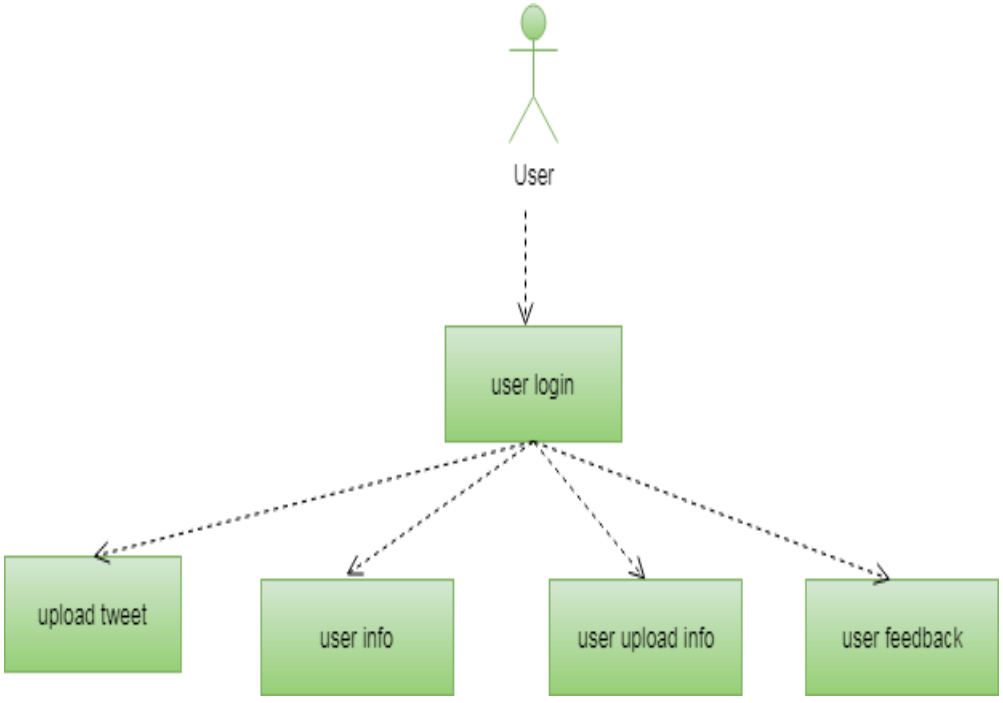


Figure 5.6: Component Diagram-User

b. Admin

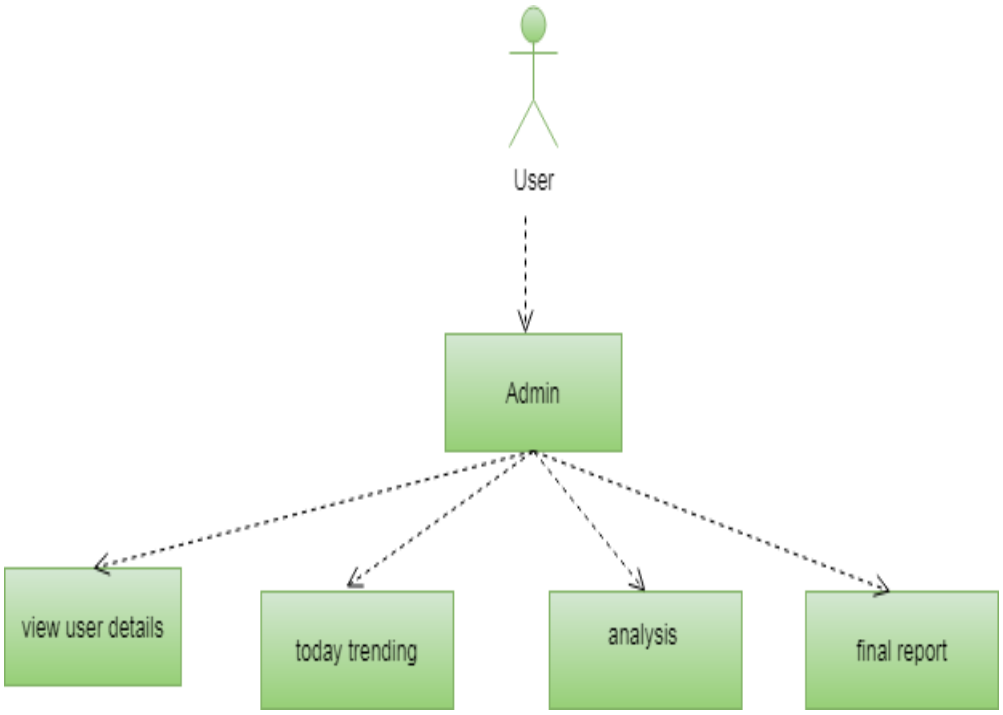


Figure 5.7: Component Diagram-Admin

5.3.3 Deployment Diagram

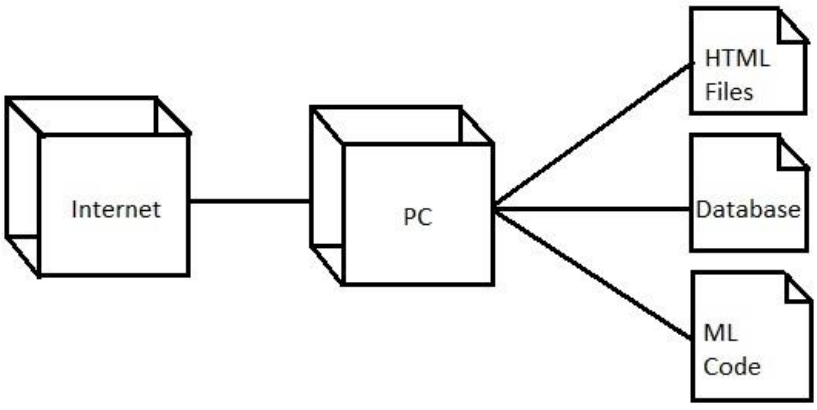


Figure 5.8: Deployment Diagram

5.3.4 Object Diagram

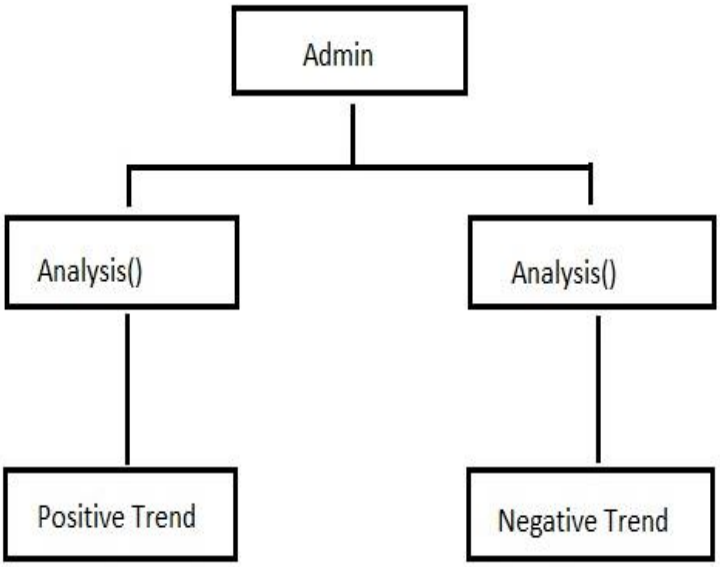


Figure 5.9: Object Diagram

5.3.4 Activity Diagram

a. User

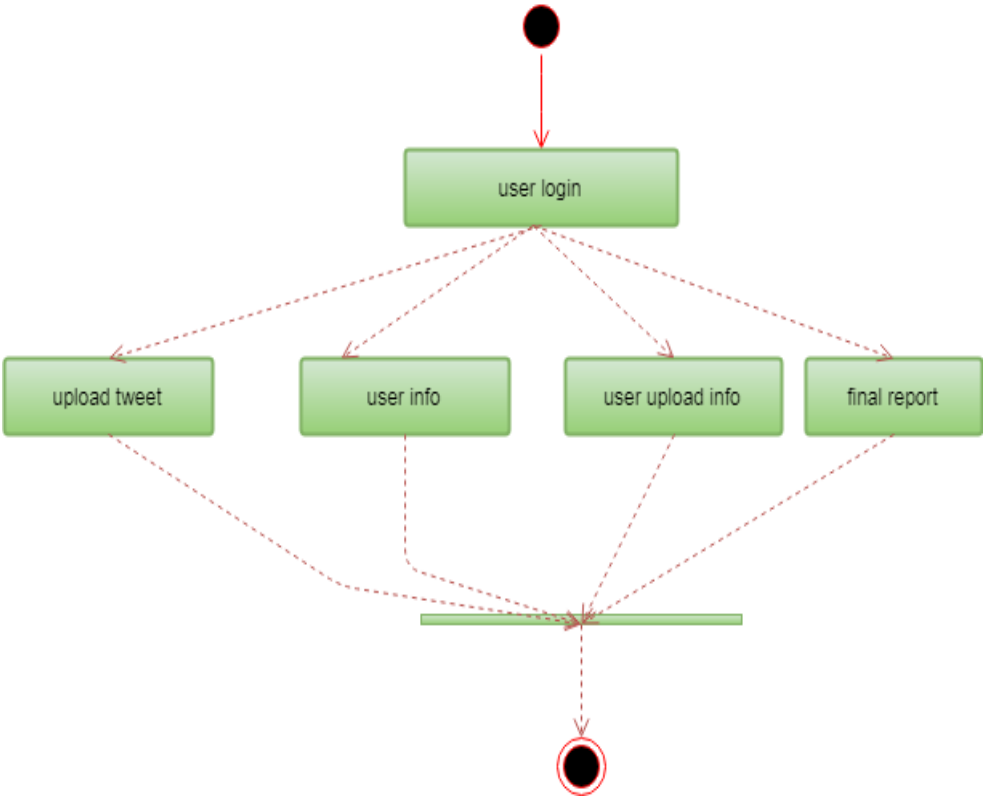


Figure 5.10: Activity Diagram

b. Admin

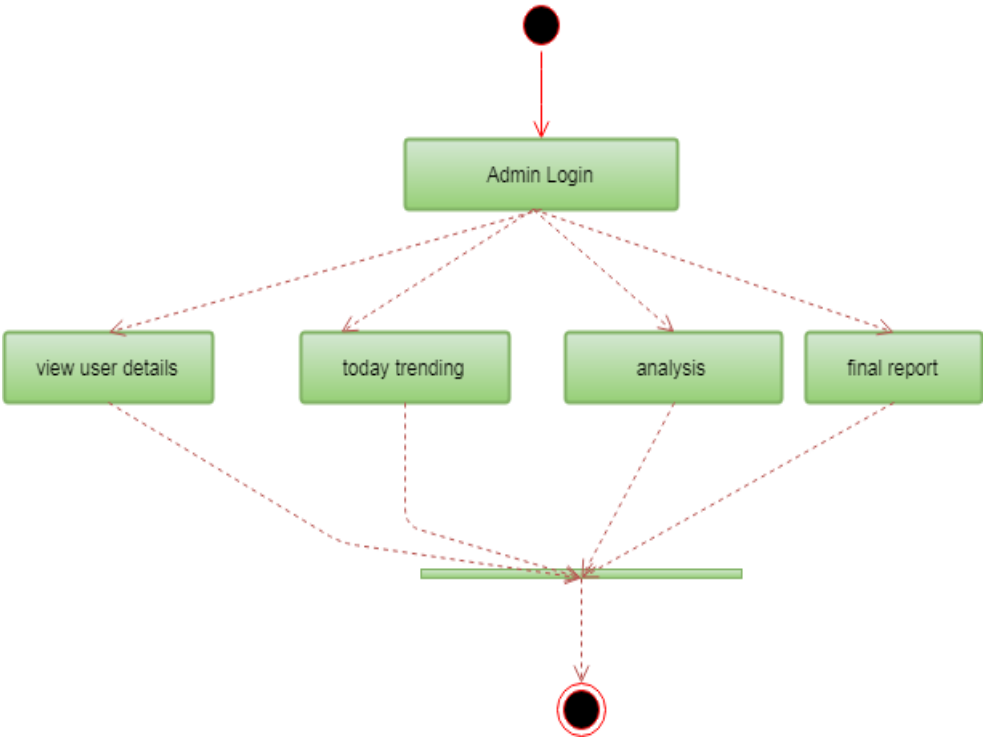


Figure 5.11: Activity Diagram-Admin

5.3.5 Sequence Diagram

a. User

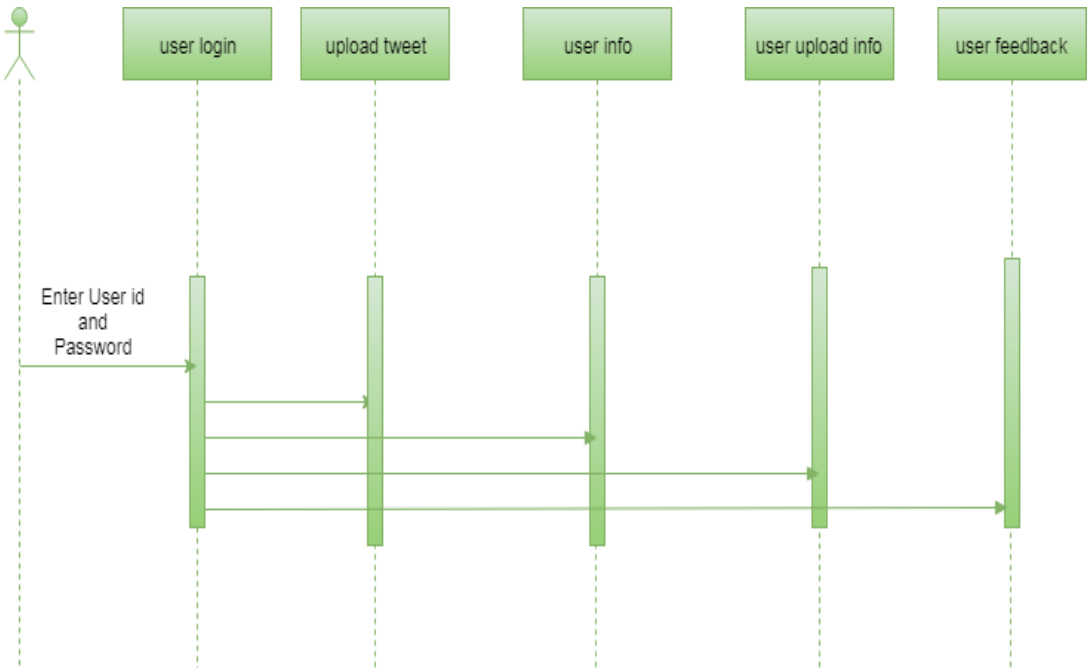


Figure 5.12: Sequence Diagram-User

b. Admin

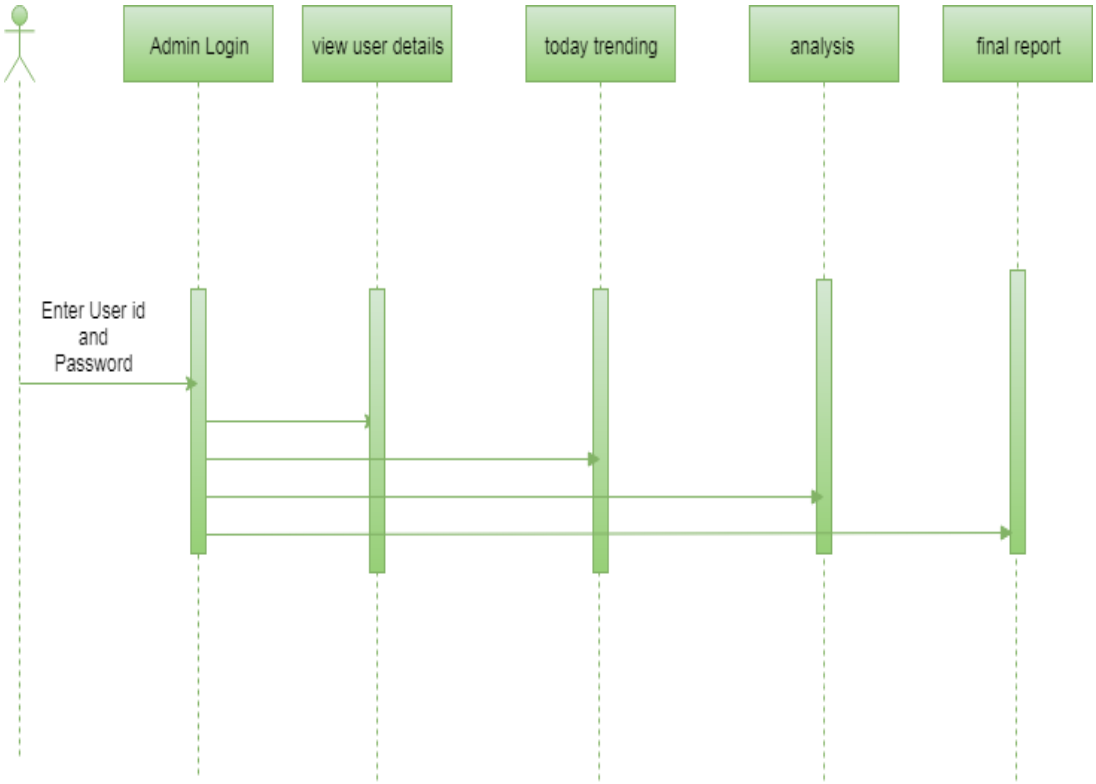


Figure 5.13: Sequence Diagram-Admin

5.3.6 Collaboration Diagram

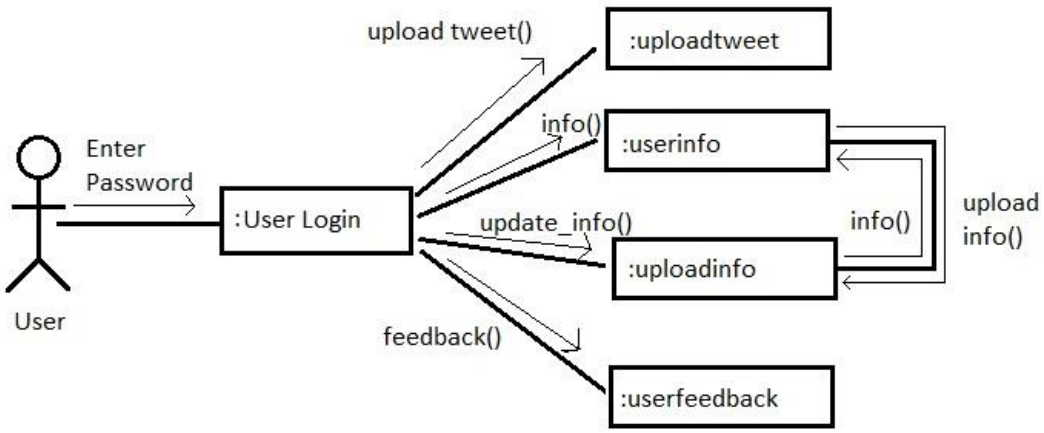


Figure5.14: Collaboration Diagram

5.3.7 State Chart Diagram

a. Admin

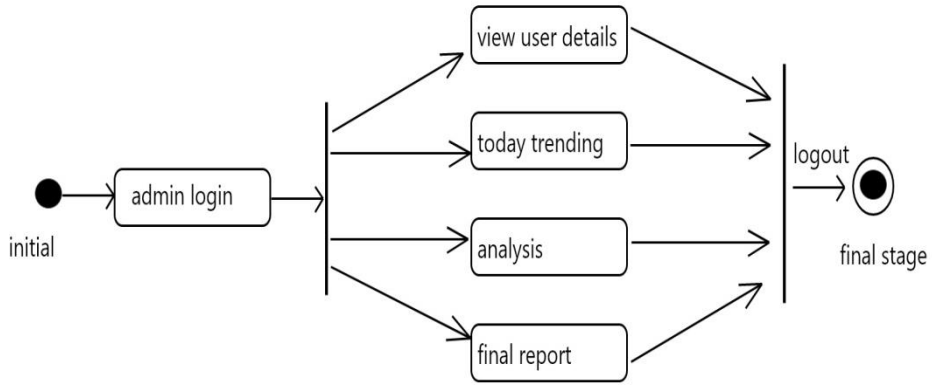


Figure 5.15: State Chart Diagram-Admin

b. User

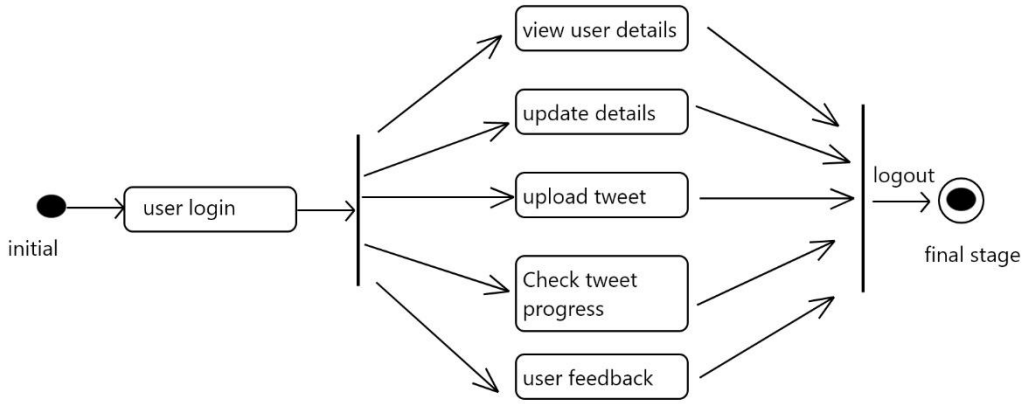


Figure 5.16: State Chart Diagram-User

6. PROJECT CODE

6.1 CODE TEMPLATES

Basic template for user background

```
<!DOCTYPE html>
{% load staticfiles %}
<html lang="en">
<head>
    <meta charset="UTF-8">
    <title>Title</title>
      <link rel="stylesheet"
href="https://stackpath.bootstrapcdn.com/bootstrap/4.3.1/css/bootstrap.min.css" integrity="sha384-
ggOyR0iXCbMQv3Xipma34MD+dH/1fQ784/j6cY/iJTQUOhcWr7x9JvoRxT2MZw1T"
crossorigin="anonymous">

    <link href="https://fonts.googleapis.com/css?family=Dancing+Script&display=swap"
rel="stylesheet">
    <link href="https://fonts.googleapis.com/css?family=Aldrich&display=swap" rel="stylesheet">
    <style>
        body{
        background: url("{% static 'bg.jpg' %}");
        background-size: cover;
        }
        .menu table{
        width:100%;
        text-align:center;
        }
```

```
.menu table td:hover{
background:rgb(0,0,0);
}
.menu table td{
background: #584b4f;
}
.menu table,.menu table th,.menu table td {
border: ;
border-collapse: collapse;
}
.menu table th,.menu table  td {
padding: 15px;
}
.topic h1{
 color:white;
 padding:10px;
 text-align:center;
 border-style:none;
 height:100px;
 width:1330px;
 float:left;
 font-family: 'Dancing Script', cursive;
 }
.giver {
 color: #ffd152;
```

```
    font-family: 'Aldrich', sans-serif;
     padding:;
     }
    .giver a{
        color: steelblue;
        font-family: cooper;
        cursor: pointer;
        text-decoration: none;
        background-color: whitesmoke;
        border-radius: 5px;
    }
    h1{
    font-family: 'Dancing Script', cursive;
    font-size:50px;
    font-color:;
    text-align:center;
       color: tomato;
background-color: snow;
    }
    h2{
    font-family: 'Aldrich', sans-serif;
    text-align:30px;
    text-align:center;
    margin-top:px;
    }
```

```
        .menu {
         border-radius:10px;
        }


</style>
</head>
<body>
{% csrf_token %}
<h1>Analysis_of_Women_Saftey</h1>
<h2></h2>
<div class ="menu">
   <table class="giver">
       <tr>
           <td><a href="{% url 'user_mydetails' %}">~MY DETAILS~</a></td>
           <td><a href="{% url 'user_updatedetails' %}">~UPDATE DETAILS~</a></td>
           <td><a href="{% url 'tweet' %}">~Upload Tweet~</a></td>
           <td><a href="{% url 'user_login'  %}">~LOGOUT~</a></td>
       </tr>
   </table>
</div>
<div class="mainholder">
{% block userblock %}
{% endblock %}
</div>
</body>
```

```
</html>
```

Basic template for admin background

```
<!DOCTYPE html>
{% load staticfiles %}
<html lang="en">
<head>
    <meta charset="UTF-8">
    <title>Title</title>
      <link rel="stylesheet"
href="https://stackpath.bootstrapcdn.com/bootstrap/4.3.1/css/bootstrap.min.css" integrity="sha384-
ggOyR0iXCbMQv3Xipma34MD+dH/1fQ784/j6cY/iJTQUOhcWr7x9JvoRxT2MZw1T"
crossorigin="anonymous">

    <link href="https://fonts.googleapis.com/css?family=Dancing+Script&display=swap"
rel="stylesheet">
    <link href="https://fonts.googleapis.com/css?family=Aldrich&display=swap" rel="stylesheet">
    <style>
        body{
        background: url("{% static 'bg.jpg' %}");
        background-size: cover;
        }
        .menu table{
        width:100%;
        text-align:center;
```

```
}
.menu table td:hover{
background:rgb(0,0,0);
}
.menu table td{
background: #584b4f;
}
.menu table,.menu table th,.menu table td {
border: ;
border-collapse: collapse;
}
.menu table th,.menu table  td {
padding: 15px;
}
.topic h1{
 color:white;
 padding:10px;
 text-align:center;
 border-style:none;
 height:100px;
 width:1330px;
 float:left;
 font-family: 'Dancing Script', cursive;
 }
.giver {
```

```
 color: #ffd152;
font-family: 'Aldrich', sans-serif;
 padding:;
 }
.giver a{
    color: steelblue;
    font-family: cooper;

    cursor: pointer;
    text-decoration: none;
    background-color: whitesmoke;
    border-radius: 5px;
}
h1{
font-family: 'Dancing Script', cursive;
font-size:50px;
font-color:;
text-align:center;
   color: tomato;
background-color: snow;
}
h2{
font-family: 'Aldrich', sans-serif;
text-align:30px;
text-align:center;
```

```
        margin-top:px;
        }
        .menu {
         border-radius:10px;
        }
</style>
</head>
<body>
{% csrf_token %}
<h1>Analysis_of_Women_Saftey - ADMIN</h1>
<h2></h2>
<div class ="menu">
   <table class="giver">
       <tr>
           <td><a href="{% url 'admin_viewpage' %}">~VIEW USER DETAILS~</a></td>
           <td><a href="{% url 'admin_viewtrending' %}">~TODAY TRENDING~</a></td>
           <td><a href="{% url 'viewtreandingtopics' 'column' %}">~ANALYSIS GRAPH~ </a></td>
           <td><a href="{% url 'admin_viewfeedback' %}">~VIEW FEEDBACK~</a></td>
           <td><a href="{% url 'user_login'  %}">~LOGOUT~</a></td>
       </tr>
   </table>

</div>
<div class="mainholder">
{% block adminblock %}
```

```
{% endblock %}

</div>

</body>

</html>
```

User Login

```
<!DOCTYPE html>

{% load staticfiles %}

<html lang="en">

<head>

    <meta charset="UTF-8">

    <title>Title</title>

    <link rel="stylesheet" href="https://stackpath.bootstrapcdn.com/bootstrap/4.3.1/css/bootstrap.min.css"
integrity="sha384-
ggOyR0iXCbMQv3Xipma34MD+dH/1fQ784/j6cY/iJTQUOhcWr7x9JvoRxT2MZw1T"
crossorigin="anonymous">


    <link href="https://fonts.googleapis.com/css?family=Dancing+Script&display=swap"
rel="stylesheet">

    <link href="https://fonts.googleapis.com/css?family=Aldrich&display=swap" rel="stylesheet">

    <style>

        body, html {

        height: 100%;

        margin: 0;

        }

        body{

        background: url("{% static 'twitter.jpg' %}");
```

```
    height: %;
    background-position: center;
    background-repeat: no-repeat;
    background-size: 100% 100%;
        border: 12px solid #dbe2e8;
    }
    input {
     padding: 7px -4px;
 border-radius: 5px;
 margin: 10px;
    }
    .table {
display: table;
 border-collapse: separate;
 border-spacing: 2px;
 border-color: grey;
 margin-top: 103px;
 margin-left: 943px;
 font-family: 'Aldrich', sans-serif;
 font-size: unset;
 color: black;
 border: 12px solid #dbe2e8;
 border-radius: 10px;
 background-color: white;
 width: 100px;
```

```
}
.table td, .table th {
        padding: 0px;
    vertical-align: unset;
    border-top: 1px solid #dee2e6;
}
    }
        .register{
        position:absolute;
        top:50%;
        left:73%;
        transform: translate(-50%, -50%);
        background-color:teal;
        font-family: 'Aldrich', sans-serif;

        color:white;
        font-size:16px;
        padding: 12px 24px;
        cursor: pointer;
        border-radius: 5px;
        text-align: center;
        margin-top: 191px;
        margin-left: -409px;
        }
```

```
    .register:hover{
    background-color: black;
    background: url("{% static 'hotel.jpg' %}");

    }

    h1{
    font-family: 'Dancing Script', cursive;

      text-shadow: 0 0 31px, 0 0 51px white;
text-align: center;
color: black;
font-size: 50px;
background-color: whitesmoke;
    }
    .log{
    background: url("{% static 'login.png' %} ");
    background-size:cover;
    width:100px;
    height:50px;
    }
    tbody {
display: table-row-group;
vertical-align: middle;
border-color: inherit;
```

```
}


   </style>
</head>
<body>
<h1>Analysis_of_Women_Saftey</h1>
<form method="POST">
   {% csrf_token %}
   <table class="table">
      <tr>
         <td>USERNAME:</td>
         <td><input type="text" name="name" value="" class=""></td>
      </tr>
      <tr>
         <td>PASSWORD:</td>
         <td><input type="password" name="password" class="" ></td>
      </tr>
      <tr>
         <td><input type="submit" class="log"   name="login" value="" >
      </tr>
      <tr>
         <td><p>Don't have an account?</p></td>
         <td><a href="user_register"><img src="static/signup.png" width="100" hight="70"></a></td>
      </tr>
      <tr>
```

```
        <td>ADMIN BLOCK</td>
        <td><a href="admin_login"><img src="static/admin.png" width="100" hight="70"></a></td>
      </tr>
   </table>
    <center class="message">
   {% if messages %}
      {% for message in messages %}
         {{ message }}
      {% endfor %}
   {% endif %}
   </center>
</form>
</body>
</html>
```

6.2 OUTLINE FOR VARIOUS FILES

Negative polarity values graph

```
{% extends 'research/base1.html' %}
{% block adminblock %}
{% load staticfiles %}
<style>
   .values{
   height:10px;
   width:10px;
   overflow:scroll;
```

```
    }
.any{
 border-style: double;
    height: 100px;
    width: 153px;
    background: whitesmoke;
    font-family: Aldrich;
    color: black;
    border-radius: 7px;
    text-align: center;
}
</style>
    <script>
window.onload = function() {
var chart= new CanvasJS.Chart("chartContainer", {
            animationEnabled: true,
            title: {
                        text: "ANALYSIS"
            },
            data: [{
                        type: "bar",
                        startAngle: 240,
                        dataPoints: [
                        {% for k,v in dd.items %}
```

```
        {y: {{v.1}},label:
"{{k.0}}{{k.1}}{{k.2}}{{k.3}}{{k.4}}{{k.5}}{{k.6}}{{k.7}}{{k.8}}{{k.9}}"},
    {% endfor %}
                ]
        }]
});
chart.render();
}
</script>
</head>
<div class="values">
<table>
  <tr>
    <th>Topics</th>
    <th>Poistive Comments</th>
    <th>Negative Comments</th>
    <th>Neutral Comments</th>
  </tr>
  {% for key,values in dd.items %}
  <tr>
    <td>{{key}}</td>
      <td>{{values.0}}</td>
      <td>{{values.1}}</td>
      <td>{{values.2}}</td>
  {% endfor %}
```

```
    </tr>
    </table>
    </div>
<div class="any">
    <a href="{% url 'viewtreandingtopics' 'bar' %}"><h4 style="color:green">Poistive</h4></a>
    <a href="{% url 'negativefeedbacktivechart' 'bar' %}"><h4 style="text-
align:center;color:red">Negative</h4></a>
</div>
<div id="chartContainer" class="graph"></div>
<script src="https://canvasjs.com/assets/script/canvasjs.min.js"></script>
{% endblock %}
```

This file gives us the overview of how much negativity against women is present in the respective cities. It also shows the extent to which each city contributes towards negativity or abuse against women.

Positive polarity values chart

```
{% extends 'research/base1.html' %}
{% block adminblock %}
{% load staticfiles %}
<style>
    .values{
    height:10px;
    width:10px;
    overflow:scroll;
    }
.any{
    border-style: double;
```

```
    height: 100px;
    width: 153px;
    background: whitesmoke;
    font-family: Aldrich;
    color: black;
    border-radius: 7px;
    text-align: center;
}
</style>
    <script>
window.onload = function() {
var chart= new CanvasJS.Chart("chartContainer", {
        animationEnabled: true,
        title: {
                text: "ANALYSIS"
        },
        data: [{
                type: "splineArea",
                startAngle: 240,
                dataPoints: [
                {% for k,v in dd.items %}
        {y: {{v.0}},label:
"{{k.0}}{{k.1}}{{k.2}}{{k.3}}{{k.4}}{{k.5}}{{k.6}}{{k.7}}{{k.8}}{{k.9}}"},
      {% endfor %}
                ]
```

```
        }]
});
chart.render();
}
</script>
<div class="values">
<table>
   <tr>
      <th>Topics</th>
      <th>Poistive Comments</th>
      <th>Negative Comments</th>
      <th>Neutral Comments</th>
   </tr>
   {% for key,values in dd.items %}
   <tr>
      <td>{{key}}</td>
         <td>{{values.0}}</td>
         <td>{{values.1}}</td>
         <td>{{values.2}}</td>
   {% endfor %}
   </tr>
   </table>
   </div>
<div class="any">
   <a href="{% url 'viewtreandingtopics' 'bar' %}"><h4 style="color:green">Poistive</h4></a>
```

```
    <a href="{% url 'negativefeedbacktivechart' 'bar' %}"><h4 style="color:red" >Negative</h4></a>
</div>
<div id="chartContainer" class="graph"></div>
<script src="https://canvasjs.com/assets/script/canvasjs.min.js"></script>
{% endblock %}
```

This file gives the overview about women safety in various cities showing the extent to which they feel safe in a particular city and all this is depicted in a graph.

View Trending

```
{% extends 'research/base1.html' %}
{% block adminblock %}
{% load staticfiles %}
<style>
   .viewtopic{
    position: absolute;
    top:132px;
    left:200px;
    padding:10px;
    width:500px;
    height:400px;
    overflow:scroll;
    float:left;
}
.viewtopic table{
```

```
    width:30em;
    text-align:center;
    border-collapse:collapse;
    border-spacing:1px;
    background:;
    float:left;
}
.viewtopic table tr th{
    color:;
}
.viewtopic table tr th{
background:
padding:10px;
}
.viewtopic table tr td{
background:rgb(0,0,0);
padding:10px;
}
.viewtopic table tr:hover td{
background:rgba(204, 0, 255);
}
.topicimage{
border-style:solid;
 border-width:1px;
 height:280px;
```

```
 width:380px;
 margin-top:70px;
 margin-left:840px;

 background: url("{% static 'bg1.gif' %}");
 background-size: 100%100%;
 float:left;

}
</style>
<div class="viewtopic">
<table>
   <tr>
   </tr>
   <tr>
      {% for object in objects %}
      <td style="color:yellow">{{object.topics}}</td>
   </tr>
   {% endfor %}
</table>
</div>
<div class="topicimage"></div>
{% endblock %}
```

This file shows the top 5 trending cities in India with highest number of women feeling unsafe on that particular day.

6.3 CLASS WITH FUNCTIONALITY

Settings

```
import os


# Build paths inside the project like this: os.path.join(BASE_DIR, ...)
BASE_DIR = os.path.dirname(os.path.dirname(os.path.abspath(__file__)))


# Quick-start development settings - unsuitable for production
# See https://docs.djangoproject.com/en/2.0/howto/deployment/checklist/
# SECURITY WARNING: keep the secret key used in production secret!
SECRET_KEY = 'tq^ad=hbc&40h6^3czmwvbczx_*o@u*pw7so#5x_&(cg08o-sb'
# SECURITY WARNING: don't run with debug turned on in production!
DEBUG = True


ALLOWED_HOSTS = []
# Application definition
INSTALLED_APPS = [
    'django.contrib.admin',
    'django.contrib.auth',
    'django.contrib.contenttypes',
    'django.contrib.sessions',
    'django.contrib.messages',
    'django.contrib.staticfiles',
    'Client',
    'Research',
```

```
]

MIDDLEWARE = [
    'django.middleware.security.SecurityMiddleware',
    'django.contrib.sessions.middleware.SessionMiddleware',
    'django.middleware.common.CommonMiddleware',
    'django.middleware.csrf.CsrfViewMiddleware',
    'django.contrib.auth.middleware.AuthenticationMiddleware',
    'django.contrib.messages.middleware.MessageMiddleware',
    'django.middleware.clickjacking.XFrameOptionsMiddleware',
]
ROOT_URLCONF = 'Analysis_of_Women_Safety.urls'
TEMPLATES = [
    {
        'BACKEND': 'django.template.backends.django.DjangoTemplates',
        'DIRS': [(os.path.join(BASE_DIR,'design/htmlfiles'))],
        'APP_DIRS': True,
        'OPTIONS': {
            'context_processors': [
                'django.template.context_processors.debug',
                'django.template.context_processors.request',
                'django.contrib.auth.context_processors.auth',
                'django.contrib.messages.context_processors.messages',
            ],
        },
```

```
    },
]
WSGI_APPLICATION = 'Analysis_of_Women_Safety.wsgi.application'
# Database
# https://docs.djangoproject.com/en/2.0/ref/settings/#databases
DATABASES = {
    'default': {
        'ENGINE': 'django.db.backends.mysql',
        'NAME': 'women',
        'USER': 'root',
        'PASSWORD': '',
        'HOST': '127.0.0.1',
        'PORT': '3306',
    }
}
# Password validation
# https://docs.djangoproject.com/en/2.0/ref/settings/#auth-password-validators
AUTH_PASSWORD_VALIDATORS = [
    {
        'NAME': 'django.contrib.auth.password_validation.UserAttributeSimilarityValidator',
    },
    {
        'NAME': 'django.contrib.auth.password_validation.MinimumLengthValidator',
    },
    {
```

```
        'NAME': 'django.contrib.auth.password_validation.CommonPasswordValidator',
    },
    {
        'NAME': 'django.contrib.auth.password_validation.NumericPasswordValidator',
    },
]
# Internationalization
# https://docs.djangoproject.com/en/2.0/topics/i18n/
LANGUAGE_CODE = 'en-us'
TIME_ZONE = 'UTC'
USE_I18N = True
USE_L10N = True
USE_TZ = True
# Static files (CSS, JavaScript, Images)
# https://docs.djangoproject.com/en/2.0/howto/static-files/

STATIC_URL = '/static/'
STATICFILES_DIRS = [os.path.join(BASE_DIR,'design/images')]
MEDIA_URL = '/media/'
MEDIA_ROOT = os.path.join(BASE_DIR, 'design/media')
```

Connecting URLS

```
from django.conf.urls import url
from django.contrib import admin
```

```
from django.urls import path
from django.conf.urls.static import static
from Analysis_of_Women_Safety import settings
from Client import views as user_view
from Research import views as admin_view

urlpatterns = [
    url('admin/', admin.site.urls),
    url('^$', user_view.user_login, name="user_login"),

    url(r'^user_register/$',user_view.user_register, name="user_register"),
    url(r'^user_mydetails/$',user_view.user_mydetails, name="user_mydetails"),
    url(r'^user_updatedetails/$',user_view.user_updatedetails, name="user_updatedetails"),
    url(r'^tweet/$',user_view.tweet, name="tweet"),
    url(r'^tweetview/$',user_view.tweetview, name="tweetview"),
    url(r'^feedback/$',user_view.feedback, name="feedback"),

    url(r'^admin_login/$', admin_view.admin_login, name="admin_login"),
    url(r'admin_viewpage/$',admin_view.admin_viewpage,name="admin_viewpage"),
    url(r'admin_viewfeedback/$',admin_view.admin_viewfeedback,name="admin_viewfeedback"),
    url(r'admin_viewtrending/$',admin_view.admin_viewtrending,name="admin_viewtrending"),
    url(r'^viewtreandingtopics/(?P<chart_type>\w+)/$',
admin_view.viewtreandingtopics,name="viewtreandingtopics"),
    url(r'^negativefeedbacktivechart/(?P<chart_type>\w+)/$',
admin_view.negativefeedbacktivechart,name="negativefeedbacktivechart"),
]+ static(settings.MEDIA_URL,document_root=settings.MEDIA_ROOT)
```

WSGI

"""

WSGI config for Analysis_of_Women_Safety project.

It exposes the WSGI callable as a module-level variable named ``application``.

"""

import os

from django.core.wsgi import get_wsgi_application

os.environ.setdefault("DJANGO_SETTINGS_MODULE", "Analysis_of_Women_Safety.settings")

application = get_wsgi_application()

6.4 MEHODS INPUT AND OUTPUT PARAMETERS

INPUT:

The dataset containing the details about the user and their tweets are taken as input for this program, which is then used to analyze the sentiment behind their tweets and classify them as safe or unsafe. Whenever a new user registers with the application, their details are saved into the database so that if they tweet about anything it can be recorded under their name and the city which later helps in the analysis procedure. The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the following things:

- What data should be given as input?
- How the data should be arranged or coded?
- The dialog to guide the operating personnel in providing input.
- Methods for preparing input validations and steps to follow when error occur.

OBJECTIVES

1.Input Design is the process of converting a user-oriented description of the input into a computer-based system. This design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerized system.

2. It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The goal of designing input is to make data entry easier and to be free from errors. The data entry screen is designed in such a way that all the data manipulates can be performed. It also provides record viewing facilities.

3.When the data is entered it will check for its validity. Data can be entered with the help of screens. Appropriate messages are provided as when needed so that the user will not be in maize of instant. Thus the objective of input design is to create an input layout that is easy to follow

OUTPUT:

The output for this application is given after the sentiment analysis and classification done by the algorithm in the form of two graphs. The first graph is a bar chart which depicts the lower safety levels of women based on each city or it can also be said that this graph gives information regarding the negative polarity rates of the sentiment which are considered as unsafe. The second graph depicts the positive polarity of sentiments which means that the level of safety assured for women by each city. A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making.

1. Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is designed so that people will find the system can use easily and effectively. When analysis design computer output, they should Identify the specific output that is needed to meet the requirements.

2.Select methods for presenting information.

3.Create document, report, or other formats that contain information produced by the system.

The output form of an information system should accomplish one or more of the following objectives.

- Convey information about past activities, current status or projections of the

- Future.
- Signal important events, opportunities, problems, or warnings.
- Trigger an action.
- Confirm an action.

7. PROJECT TESTING

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

7.1 VARIOUS TEST CASES

Test Id	Test Name	Input	Output	Expected Result	Status
1	User login	Uid and password	Success	Success	PASS
	User login	Uid and password	Failure	Failure	FAIL
2	Add post	Title description user privacy photo	Posted successfully if the post to is the user or request sent to other user if you are posting the image on their name	Success	PASS
	Add post	Title description user privacy photo	request sent to other user if you are posting the image on their name	If the post to someone	FAIL
3	Search post based on keyword	Keyword	Key word matching with post title will be retrived	Display post	PASS
	Search post based on keyword	Keyword	If no Key word matching with post title	Nothing to display	FAIL
4	Search post based on content	Content	Content matching with	Display post	PASS

			post description will be retrived		
	Search post based on content	Content	If no content matching with post description	Nothing to display	FAIL
5	Search friend	Friend name	Retrieved if available	Show friend details send request	PASS
	Search friend	Friend name	Doesn't Retrieve	No friend with this name	FAIL
6	Search history	Click on search HISTORY	Retrieve search history	Show search history	PASS
	Search history	Click on search HISTORY	No search history	Nothing to display	FAIL
7	View all requests	Accepted	Accept post request	Accepted	PASS
	View all requests	Rejected	Don't accept request	Not Accepted	FAIL

7.2 BLACK BOX TESTING

Black Box Testing is a software testing method in which the functionalities of software applications are tested without having knowledge of internal code structure, implementation details and internal paths. Black Box Testing mainly focuses on input and output of software applications and it is entirely based on software requirements and specifications. It is also known as Behavioural Testing.

The above Black-Box can be any software system you want to test. Under Black Box Testing, we can test these applications by just focusing on the inputs and outputs without knowing their internal code implementation

How to do BlackBox Testing

Here are the generic steps followed to carry out any type of Black Box Testing.

- Initially, the requirements and specifications of the system are examined.
- Tester chooses valid inputs (positive test scenario) to check whether SUT processes them correctly. Also, some invalid inputs (negative test scenario) are chosen to verify that the SUT is able to detect them.
- Tester determines expected outputs for all those inputs.
- Software tester constructs test cases with the selected inputs.
- The test cases are executed.
- Software tester compares the actual outputs with the expected outputs.
- Defects if any are fixed and re-tested.

Types of Black Box Testing

There are many types of Black Box Testing but the following are the prominent ones -

- **Functional testing** –

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals. Functional testing is centered on the following items:

Valid Input : identified classes of valid input must be accepted.

Invalid Input : identified classes of invalid input must be rejected.

Functions : identified functions must be exercised.

Output : identified classes of application outputs must be exercised.

Systems/Procedures : interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

Non-functional testing - This type of black box testing is not related to testing of specific functionality, but non-functional requirements such as performance, scalability, usability.

7.3 WHITE BOX TESTING

White Box Testing is software testing technique in which internal structure, design and coding of software are tested to verify flow of input-output and to improve design, usability and security. In white box testing, code is visible to testers so it is also called Clear box testing, Open box testing, Transparent box testing, Code-based testing and Glass box testing.

It is one of two parts of the Box Testing approach to software testing. Its counterpart, Blackbox testing, involves testing from an external or end-user type perspective. On the other hand, White box testing in software engineering is based on the inner workings of an application and revolves around internal testing.

The term "WhiteBox" was used because of the see-through box concept. The clear box or WhiteBox name symbolizes the ability to see through the software's outer shell (or "box") into its inner workings. Likewise, the "black box" in "Black Box Testing" symbolizes not being able to see the inner workings of the software so that only the end-user experience can be tested.

Types of White Box Testing

White box testing encompasses several testing types used to evaluate the usability of an application, block of code or specific software package. There are listed below --

- Unit Testing:

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

Testing for Memory Leaks:

Memory leaks are leading causes of slower running applications. A QA specialist who is experienced at detecting memory leaks is essential in cases where you have a slow running software application.

8. OUTPUT SCREENS

8.1 USER INTERFACES

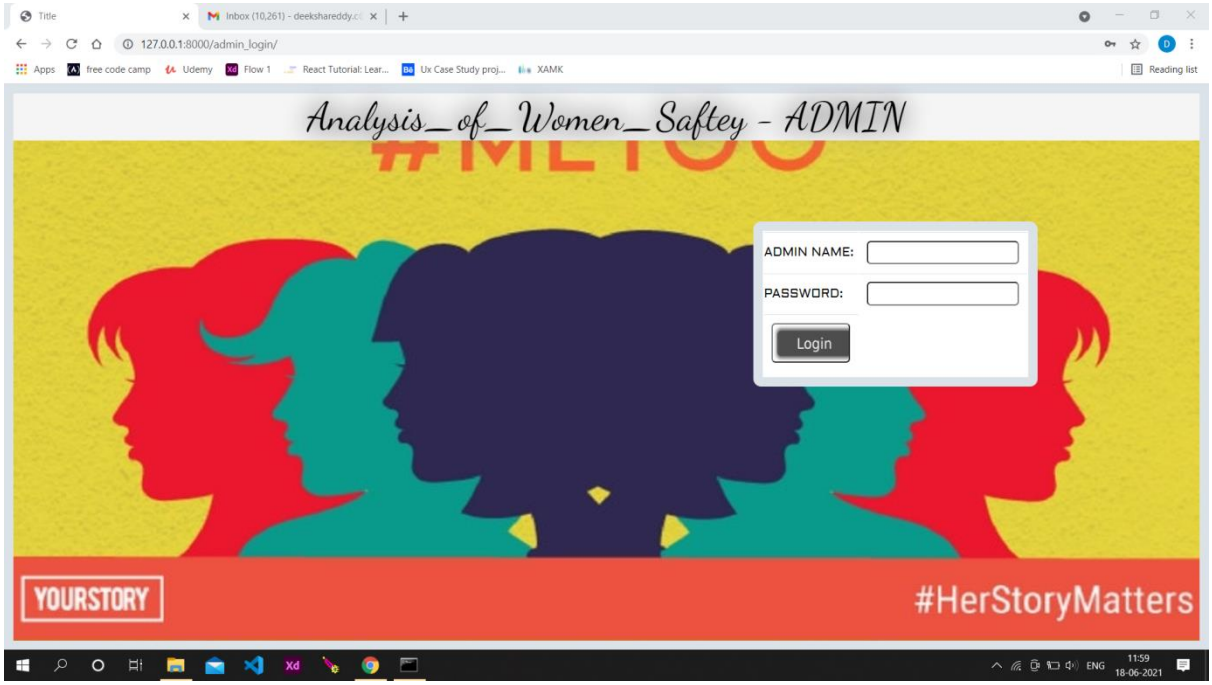


Figure 8.1: Admin Login

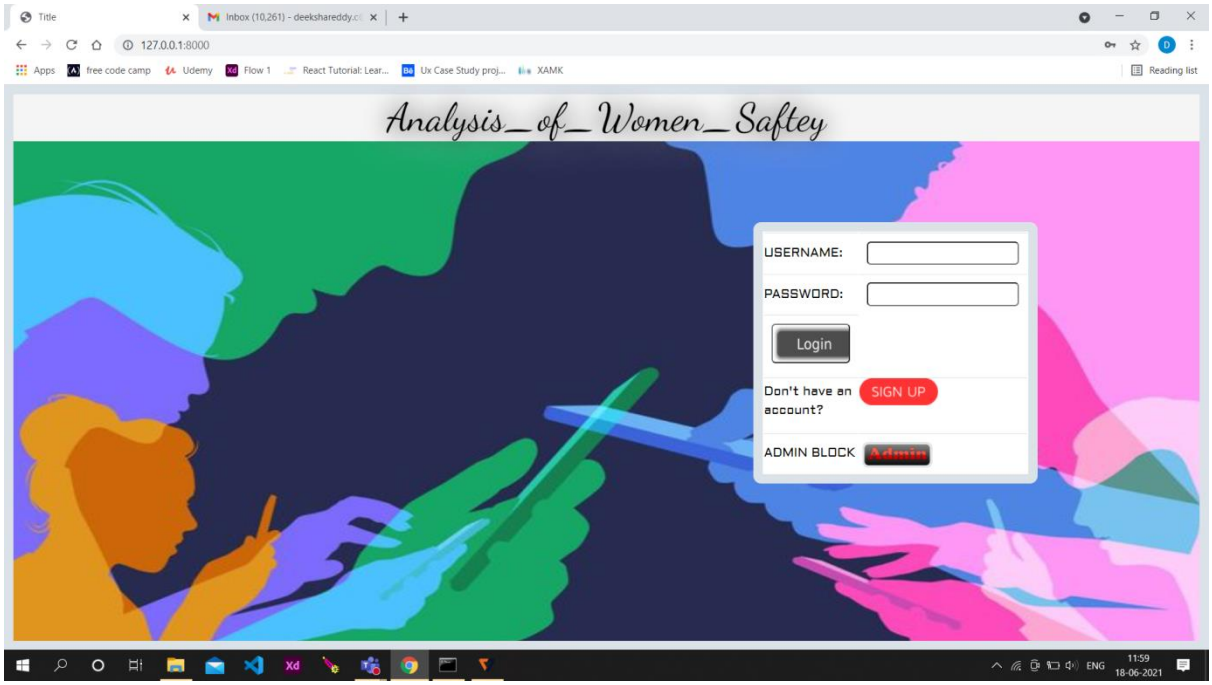


Figure 8.2 : User login

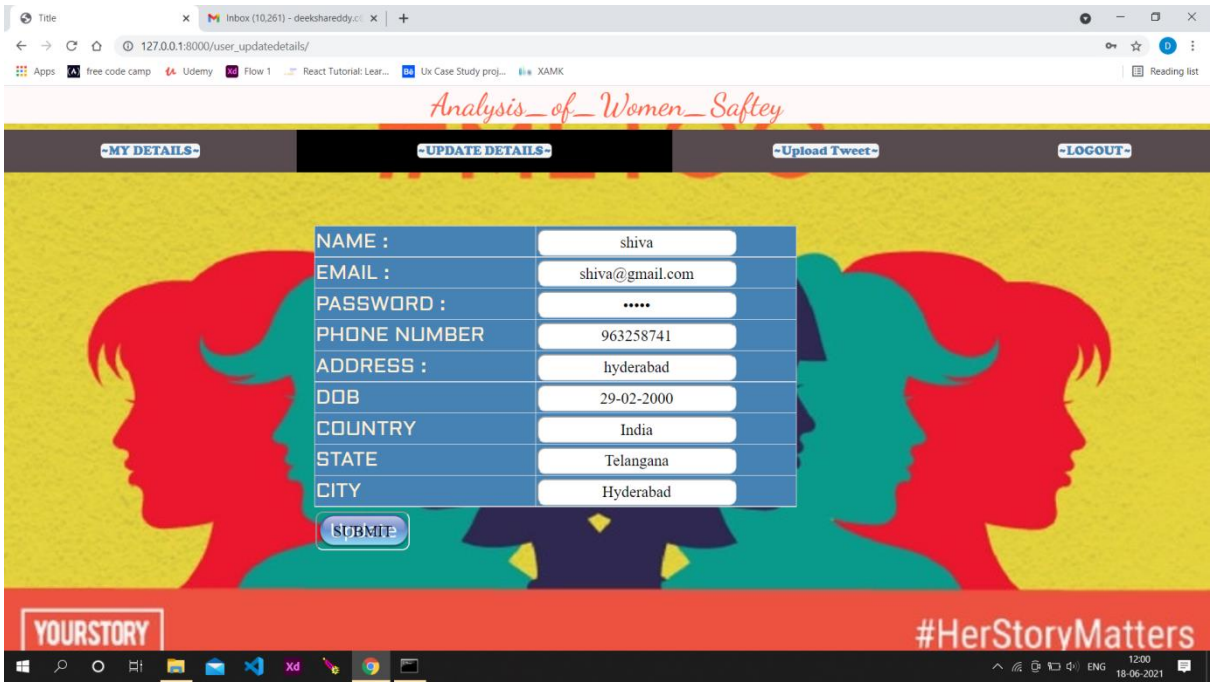


Figure 8.3: Updating user details

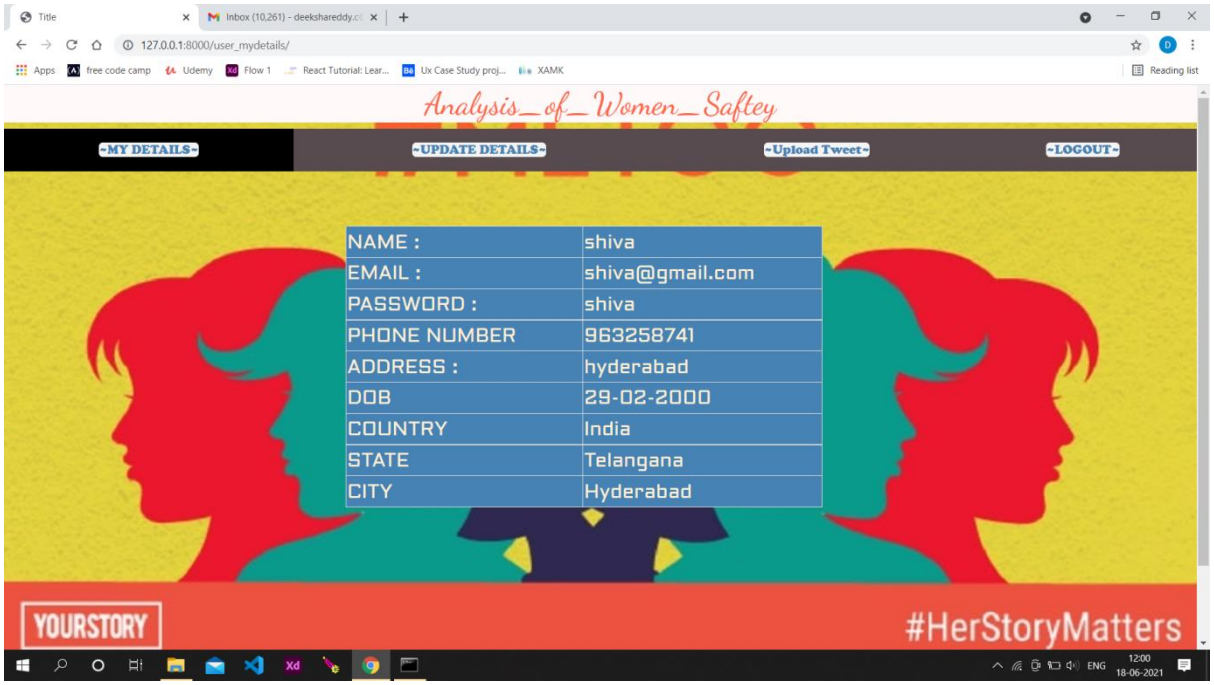


Figure 8.4: User details display

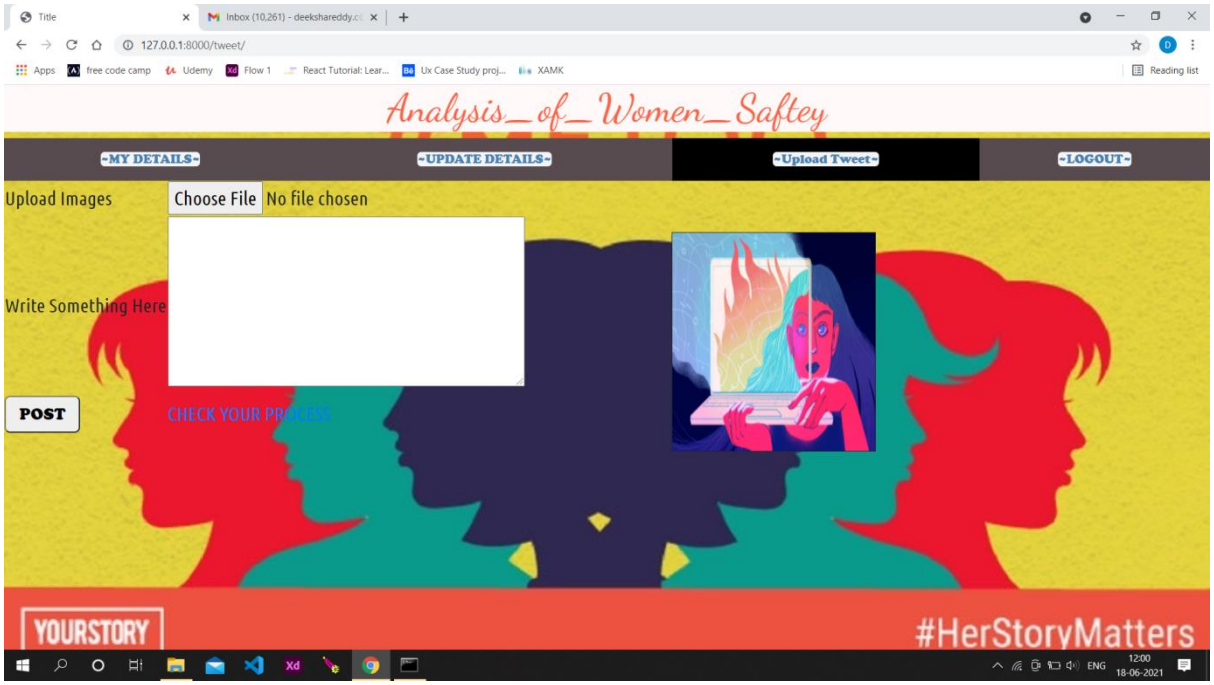


Figure 8.5: Uploading tweets

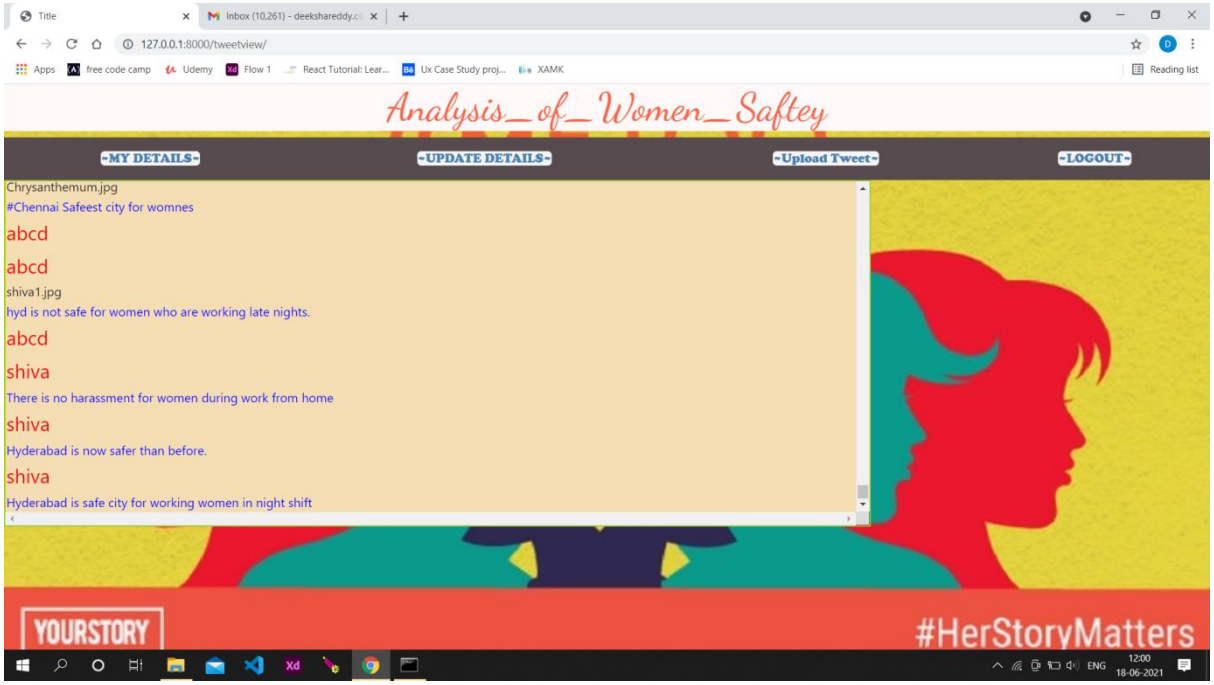


Figure 8.6: Displaying all the tweets

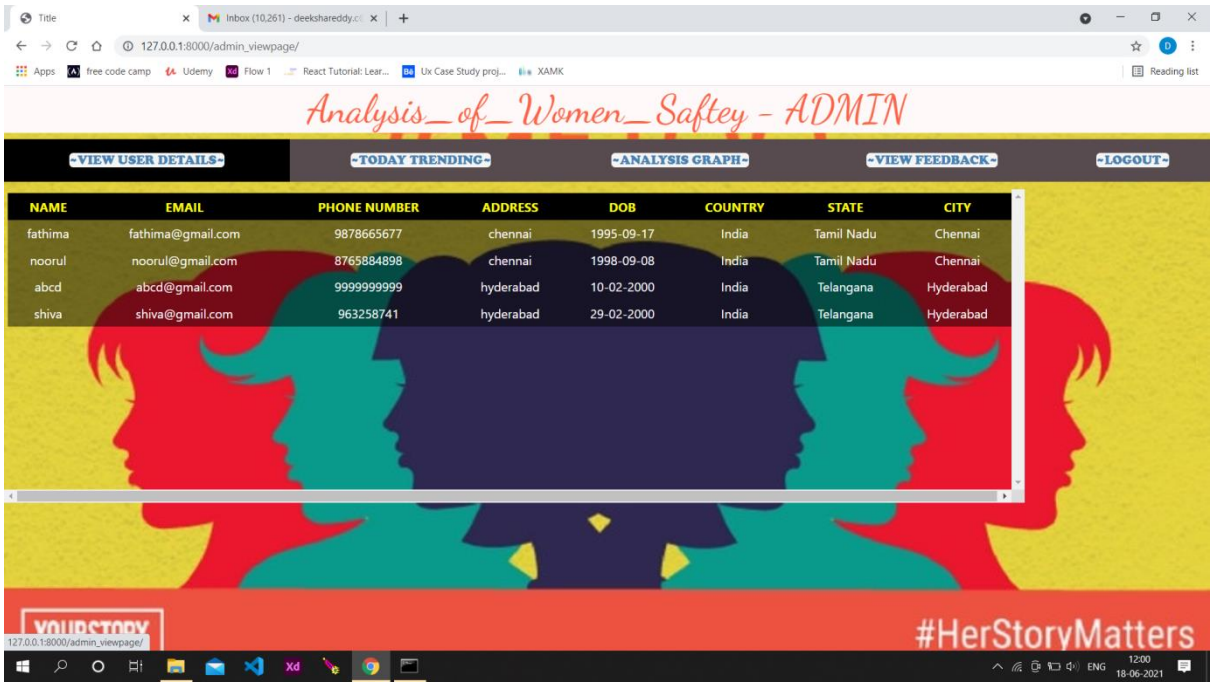


Figure 8.7: Viewing all the user's details

8.2 OUTPUT SCREENS

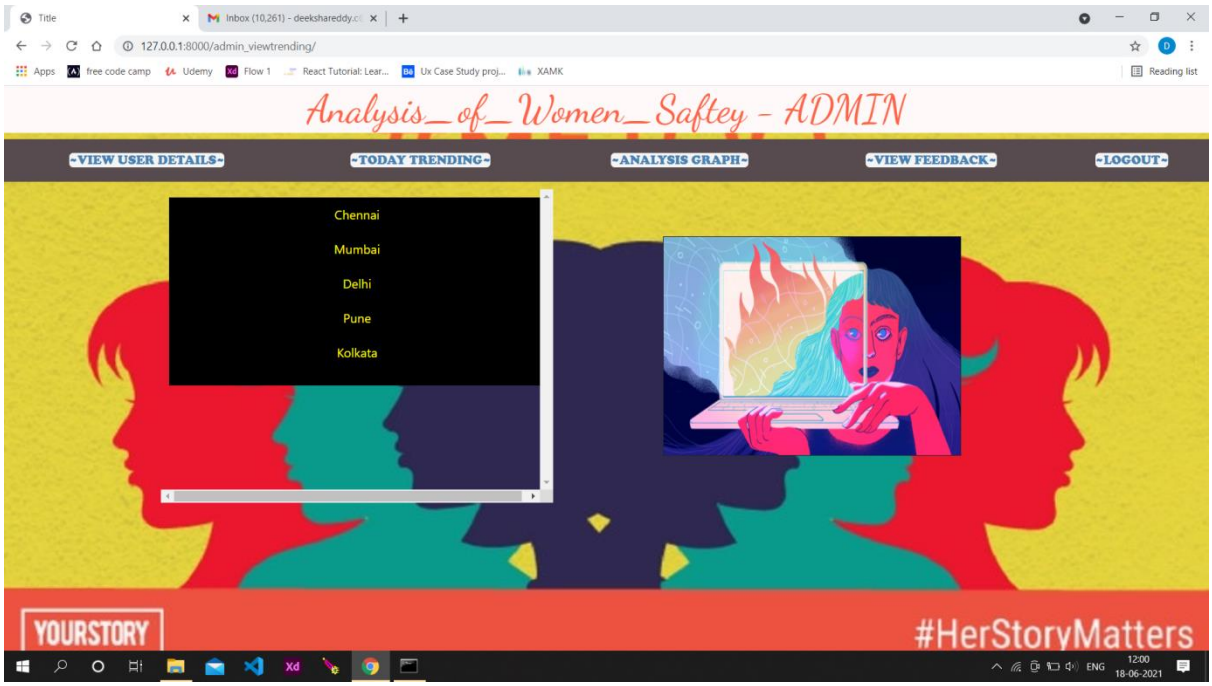


Figure 8.8: Showing today's trending cities

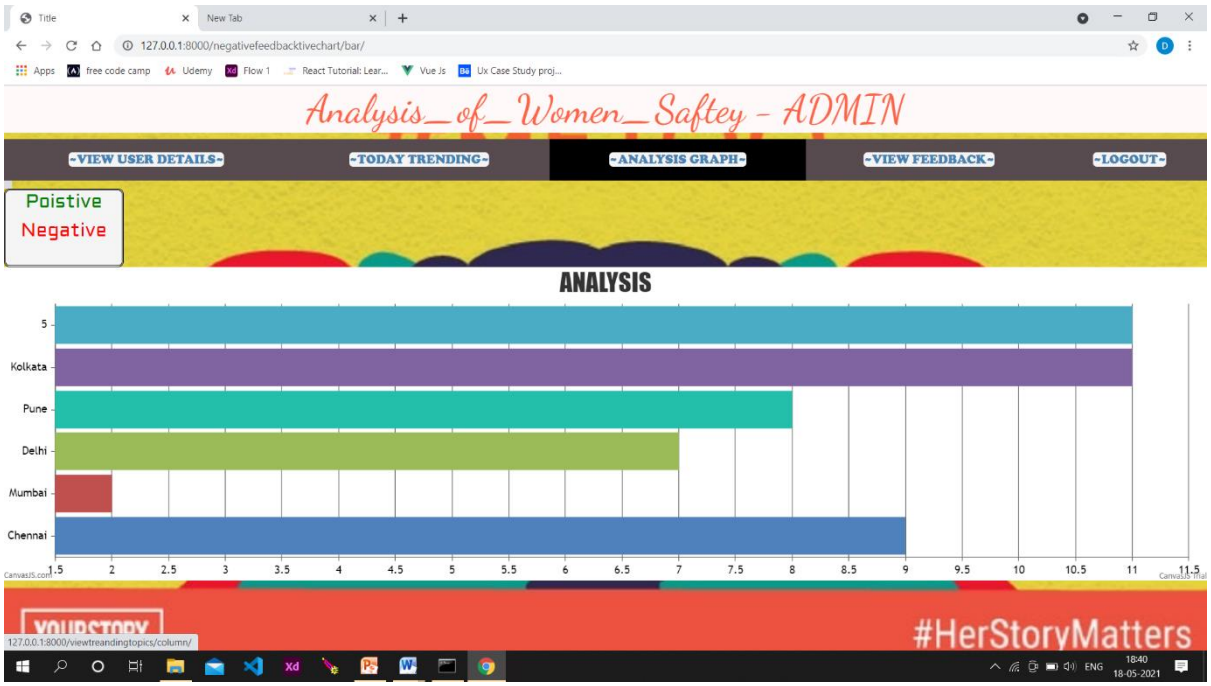


Figure 8.9 : Bar graph representing negative polarity rate

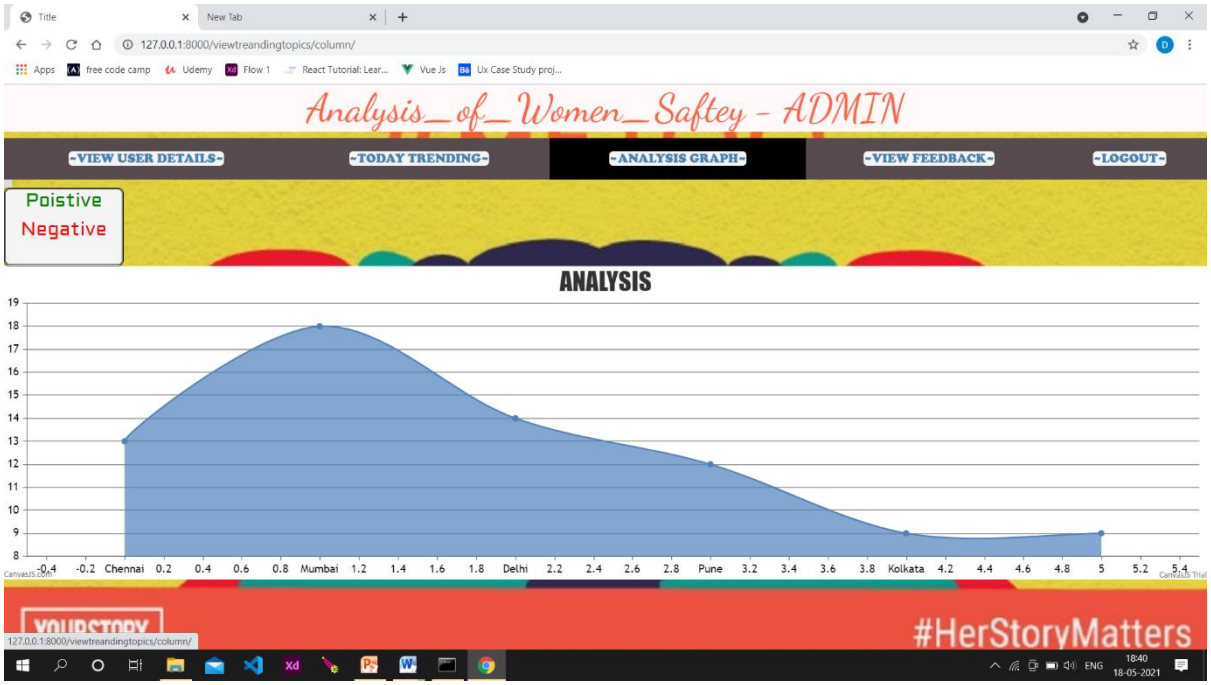


Figure 8.10: Graph representing positive polarity rate

9. EXPERIMENTAL RESULT

Analysing Experimental Data

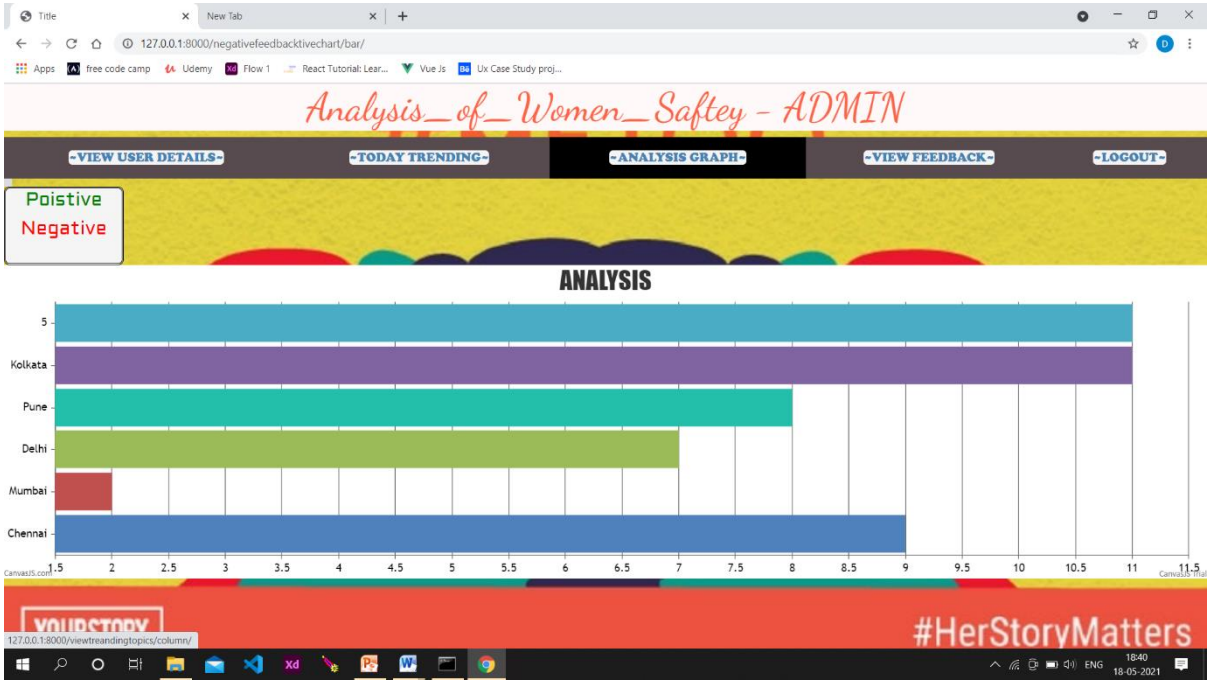


Figure 9.1 : Bar graph representing negative polarity rate

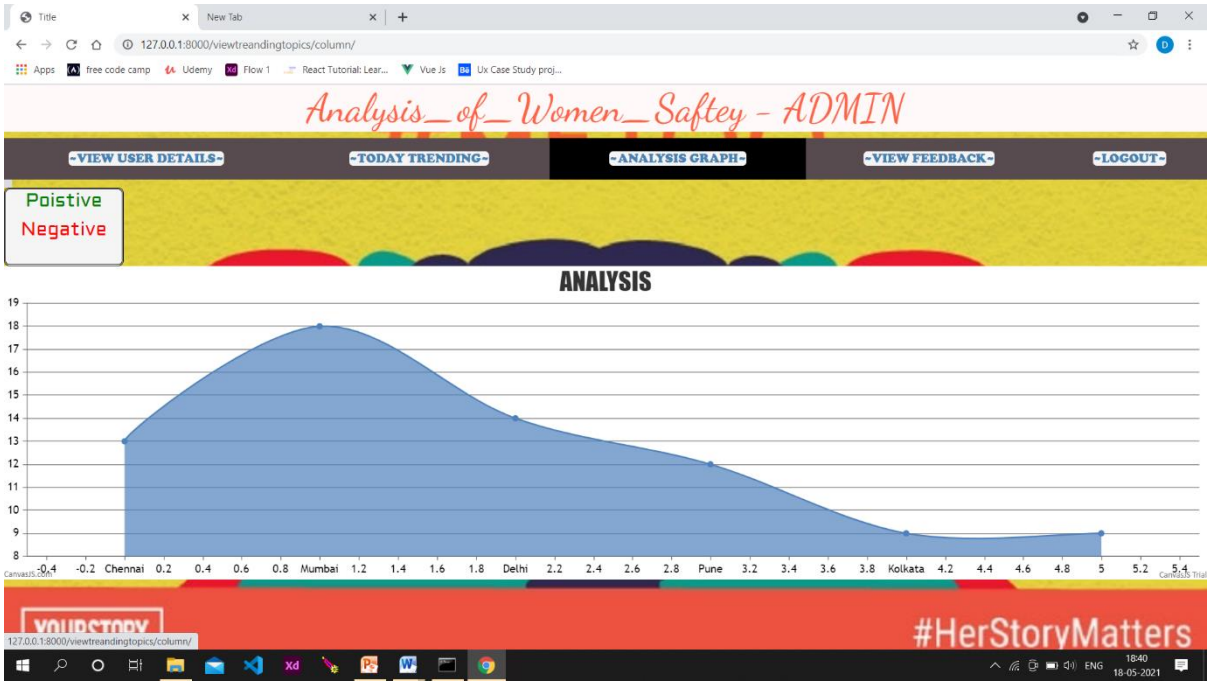


Figure 9.2 : Graph representing positive polarity rate

INTERPRETING EXPERIMENTAL RESULTS

Our model helps to analyze the sentiments behind the tweets and try to provide safety to women based on their condition. We use Support Vector Classifier algorithm to classify the whole data as either negative or positive and accordingly give the desired results .Based on the MeTooHate dataset, we performed the sentimental analysis first and then based on the polarity we classify it using the SVC algorithm. We got an accuracy of 88.3%. SVC is an appropriate algorithm for the dataset because of the linearity of the data.

10. CONCLUSION AND FUTURE ENHANCEMENT

Since the data is large, we used support vector algorithm and TEXTBLOB to achieve sentimental analysis based on which we categorized the places which are safe and unsafe for women. Due to the usage of the algorithm we could analysis the sentiment behind the tweets of women accurately up to 88.3% based on the MeToo datasetSince the accuracy is only 88.3% , there is a scope to increase the accuracy by using better models.Machine learning algorithm has been discussed throughout the project. For the twitter data that includes millions of tweet and messages every day, machine learning algorithm helps to organize and perform analysis. SPC algorithm, linear algebraic are some of the algorithms which are effective in analyzing the large data that provide categorization and convert into meaningful datasets. Hence we can perform machine learning algorithms to achieve sentimental analysis and bring more safety to women by spreading the awareness. For the future enhancement, we can extend to apply these machine learning algorithms on different social media platforms like facebook and instagram also since in our project only twitter is considered. Present ideology which is proposed can be integrated with the twitter application interface to reach larger extent and apply sentimental analysis on millions of tweet to provide more safety.

11. REFERENCES

[1] VIKRAM CHANDRA & RAMPUR SRINATH (2020). “Analysis of Women Safety using Machine Learning on Tweets”. International Research Journal of Engineering and Technology (IRJET) p-ISSN: 2395-0072

[2] Gupta B, Negi M, Vishwakarma K, Rawat G & Badhani P (2017). “Study of Twitter sentiment analysis using machine learning algorithms on Python.” International Journal of Computer Applications, 165(9) 0975-8887.

[3] Mamgain N, Mehta E, Mittal A & Bhatt G (2016, March). “Sentiment analysis of top colleges in India using Twitter data.” In Computational Techniques, in Information and Communication Technologies (ICCTICT), 2016 International Conference on (pp. 525-530). IEEE

[4] Sahayak V, Shete V & Pathan A (2015). “Sentiment analysis on twitter data.” International Journal of Innovative Research in Advanced Engineering (IJIRAE), 2(1), 178-183.

[5] Jiang, L., Yu, M., Zhou, M., Liu, X., & Zhao, T. (2011, June). Target-dependent twitter sentiment classification. In Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies (pp. 151-160).

[6] Kouloumpis, E., Wilson, T., & Moore, J. (2011, July). Twitter sentiment analysis: The good the bad and the omg!. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 5, No. 1).

[7] Barbosa, Luciano, and Junlan Feng. "Robust sentiment detection on twitter from biased and noisy data." Proceedings of the 23rd international conference on computational linguistics: posters. Association for Computational Linguistics, 2010.

[8] Bermingham, Adam, and Alan F. Smeaton. "Classifying sentiment in microblogs :is brevity an advantage?." Proceedings of the 19th ACM international conference on Information and knowledge management. ACM, 2010.

[9] Pak, A., & Paroubek, P. (2010, May). Twitter as a corpus for sentiment analysis and opinion mining. In *LREc* (Vol. 10, No. 2010, pp. 1320-1326).

[10] Agarwal, Apoorv, Fadi Biadsy, and Kathleen R. Mckeown. "Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams." Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2009.

[11] Go A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, *1*(12), 2009.

[12] Eugene Charniak and Mark Johnson."Coarse-to-fine n best parsing and MaxEnt discriminative reranking." Proceedings of the 43rd annual meeting on association for computational linguistics. Association for Computational Linguistics, 2005.

[13] Michael Gamon. "Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis." Proceedings of the 20th international conference on Computational Linguistics. Association for Computational Linguistics, 2004.

[14] Soo-Min Kim and Eduard Hovy. "Determining the sentiment of opinions." Proceedings of the 20th international conference on Computational Linguistics. Association for Computational Linguistics, 2004.

[15] Dan Klein and Christopher D. Manning. "Accurate unlexicalized parsing." Proceedings of the 41st Annual Meeting on Association for Computational LinguisticsVolume 1. Association for Computational Linguistics, 2003.

12. PUBLICATIONS

International Conference on “Innovations in Computers Networks, Computational Intelligence and IOT” (ICICCI-21)

Paper ID: ICICCI-21-0074

C. Deeksha Reddy is currently pursuing her Bachelor of Technology in the stream of Computer Science Engineering at St.Martin's Engineering College. She completed her intermediate from Sri Chaitanya Junior Kalasala and 10th class from Sri Chaitanya Techno School. Her technical skills include Python, C, HTML and CSS. She also has a basic understanding of Java. She took part in Employment Skill Development Program conducted by Zensar. She is also a student of Smart Interviews. Apart from programming, she is also interested in User Experience Designing. She participated in a state wide Design Hackathon conducted by HYSEA and JNTUH in which she stood as a Winner. She did a one moth internship i.e., from June 2019 to July 2019, in National Small Industries Corporation (NSIC), ECIL, where she was trained in Python programming language. Her participations include: Women online workshop on "Women in Cyber Security and Privacy 2020" which was conducted from 6th to 10th July, Online Two Day National Level Seminar on "Recent Trends in Cloud Computing Fog and Edge Computing" from 18th to 19th June, 2021. She completed few certification courses from online platforms like Udemy, Uxcel, Coursera, SoloLearn.

Chinthapula Sai Prasanna is currently pursuing her Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. She completed her intermediate from Sri Chaitanya Junior College and 10th class from Buds and Flowers High School. She is one of the volunteer in street cause hyderabad. Her technical skills include C, Python ,HTML. She also has a basic understanding of C++. Her participations include HTML & CSS Workshop of TAM event held from 5th January 2018 to 3rd february 2018 conducted in our college. Women online workshop on "Women in Cyber Security and Privacy 2020"which was conducted from 6th to 10th July, Online Two Day National Level Seminar on "Recent Trends in Cloud Computing Fog and Edge Computing" from 18th to 19th June 2021 . Her areas of interest are Python, Artificial Intelligence, Machine Learning. She completed few certification courses from online platforms like Coursera, CursaApp and SoloLearn.

T SHREYA is currently pursuing her Bachelor of Technology in the stream of Computer Science & Engineering at St. Martin's Engineering College. She completed her intermediate from Narayna Junior College and schooling from St. Ann's High School. She was a volunteer in the student run NGO, Street Cause, during the year 2017-2018. Her technical skills include C, C++, Python and Java. She took part in Employability Skill Development Program conducted by Zensar.She did a three months internship i.e. from May 2020 to August 2020,in Manac Infotech(P) Limited where she was trained in Web Application Development. Her participations include:, Workshop on "Arduino/Robotics" which was conducted in the college on 12th February 2019 and 13th February 2019, Workshop on "Ethical Hacking" which was conducted in the college on 31st January 2020 and 1st February 2020, National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020, Women online workshop on "Women in Cyber Security and Privacy in 2020" which was conducted from 6th to 10th July 2020.She was also a student organizing member during two days "National Level Hackathon-2020" held on 7th and 8th February 2020 at the college. She spends her free time taking online certification courses related to her field of study as well as personal interests from platform such as Coursera, Cursa and EdX. Her areas of interest are Python, Artificial Intelligence, Machine Learning and Deep Learning.

V. Meghana Reddy is currently pursuing her Bachelor of Technology with specialization in Computer Science Engineering at St. Martin's Engineering College. She completed her 12th class and 10th class from Bhavan's Sri Ramakrishna Vidhyalaya . Her technical skills include Python, MySQL, Machine Learning and Deep Learning. Also has an intermediary knowledge about C, C++ and Java. She worked with DATAI Analytics India Pvt Ltd. as a Data Science and Machine Learning Intern for 6 months from Nov 1, 2020 to April 30, 2021 and also participated in the online training provided by IIT Khanpur. She worked as a volunteer during 2017-18 in Street Cause SMEC division. Participated in Machine Learning workshop conducted on 8th and 9th May 2019 by TAM and also in National Level Project Expo and Competition "Technovation-2018" organized by Mechanical and Computer Science department of SMEC on 28th March 2018. Her participations include: Online workshop for Women on " Women in Cyber Security and privacy 2020" which was conducted from 6th to 10th July,2020 , Online Two Day National Level Seminar on "Recent Trends in Cloud Computing Fog and Edge Computing" from 18th to 19th June,2021. She completed few certification courses from online platforms like Udemy, Cousera, CursaApp.

14.APPENDICES

Feedback page code

```
{% extends 'client/base.html' %}

{% block userblock %}

{% load staticfiles %}


<style>

  .feedback{

   position: absolute;

   top:70px;

   left:200px;

   padding:10px;

   width:500px;

}


.feedback table{

   width:30em;

   text-align:center;

   //border-collapse:collapse;

   border-spacing:1px;

   background:;

}


.feedback table tr th{
```

```
    color:;
}
.feedback table tr th{
background:
padding:10px;
}
.feedback table tr td{
background:rgb(0,0,0);
padding:10px;
}
.feedback table tr:hover td{
background:rgba(204, 0, 255);
}
 .feedbackimage{
border-style:solid;
 border-width:1px;
 height:280px;
 width:260px;
 margin-top:-240px;
 margin-left:640px;

 background: url("{% static 'bg1.gif' %}");
 background-size: 100%100%;
}
</style>
```

```
<div class="feedback">
   <form method="post" autocomplete="off">
   {% csrf_token %}
   <table>
      <tr>
         <td style="color:yellow">Name </td>
         <td><input type="text" name="name"> </td>
      </tr>
      <tr>
         <td style="color:yellow">Mobilenumber</td>
         <td ><input type="text" name="mobilenumber"> </td>
      </tr>
      <tr>
         <td style="color:yellow">Feedback</td>
         <td><textarea name="feedback" cols="40" rows="4"></textarea></td>
      </tr>
      <tr>
      <td colspan="3" style="text-align:center"><input type="submit" name="submit" value="submit"
style="background-color:black;color:yellow;" ><br></td>
   </tr>
   </table>
</form>
   <div class="feedbackimage"></div>
</div>
```

{% endblock %}

User details updating page code

```
{% extends 'client/base.html' %}

{% block userblock %}

{% load staticfiles %}

<link rel="stylesheet" href="https://stackpath.bootstrapcdn.com/bootstrap/4.3.1/css/bootstrap.min.css" integrity="sha384-ggOyR0iXCbMQv3Xipma34MD+dH/1fQ784/j6cY/iJTQUOhcWr7x9JvoRxT2MZw1T" crossorigin="anonymous">


    <link href="https://fonts.googleapis.com/css?family=Dancing+Script&display=swap" rel="stylesheet">
    <link href="https://fonts.googleapis.com/css?family=Aldrich&display=swap" rel="stylesheet">
<style>
  #customers{
      font-family: 'Aldrich', sans-serif;
      border-collapse:collapse;
      width:40%;
      font-size:25px;
      color:papayawhip;
      background-color:steelblue;
            margin-left: 395px;
       margin-top: 68px;
   position: absolute;
```

```
        }
        #customers td, #customers th{
        border:1px solid #ddd;
        }
        #customers th {
    padding-top: 12px;
    padding-bottom: 12px;
    text-align: center;
    background-color: black  ;
    color: white;
}
input {
    -webkit-appearance: textfield;
    background-color: white;
    -webkit-rtl-ordering: logical;
    cursor: text;
    padding: 0px;
    border-width: 2px;
    border-style: inset;
    border-color: initial;
    border-image: initial;
    border-radius:10px;
    font-family:Abel;
    font-size:20px;
        text-align: center;
```

```
}
.imga{
background: url("{% static 'admin.jpeg' %}");
background-size:100% 100%;
height:-webkit-fill-available;
width:100%;
overflow: auto;

}
.imgs{
background: url("{% static 'update.png' %}");
width:100px:
height:200px;
}
input[type="submit" i],
{
}
.log{
        background: url("{% static 'update.png' %} ");
        background-size:cover;
        width:120px;
        height:50px;
            margin-left: 397px;
    margin-top: 433px;
```

```
    position: absolute;
        }


</style>
<body>
<form method="POST">
    {% csrf_token %}
    <div class="space">
         <div class="imga">
    <table id="customers">
        <tr>
            <td>NAME :</td>
            <td><input type="text" name="name" value="{{ form.name }}"></td>
        </tr>
        <tr>
            <td>EMAIL :</td>
            <td><input type="text" name="email" value="{{ form.email }}"></td>
        </tr>
        <tr>
            <td>PASSWORD :</td>
            <td><input type="password" name="password" value="{{ form.password }}"></td>
        </tr>
        <tr>
            <td>PHONE NUMBER</td>
            <td><input type="text" name="phoneno" value="{{ form.phoneno }}"></td>
```

```
    </tr>
    <tr>
      <td>ADDRESS :</td>
      <td><input type="text" name="address" value="{{ form.address}}"></td>
    </tr>
    <tr>
      <td>DOB</td>
      <td><input type="text" name="dob" value="{{ form.dob }}"></td>
    </tr>
     <tr>
      <td>COUNTRY</td>
      <td><input type="text" name="country" value="{{ form.country }}"></td>
    </tr>
     <tr>
      <td>STATE</td>
      <td><input type="text" name="state" value="{{ form.state }}"></td>
    </tr>
     <tr>
      <td>CITY</td>
      <td><input type="text" name="city" value="{{ form.city }}"></td>
    </tr>
      <input type="submit" class="log" value="SUBMIT" >
</table>
    </div>
</div>
```

```
</form>
</body>
{% endblock %}
```

Posting tweets page code

```
{% extends 'client/base.html' %}
{% block userblock %}
{% load staticfiles %}
<style>
   .message{

   font-family:'Ubuntu Condensed', sans-serif;
      padding: 63px;
   font-size: 25px;
   }
   textarea{
   font-family:'Ubuntu Condensed', sans-serif;
   font-size:18px;
   }
   .submit{
      padding: 5px;
   border-radius: 7px;
   width: 96px;
   font-family: cooper;
```

```
    height: 47px;
    font-size: 21px;
    }
    .feedbackimage{
    border-style:solid;
     border-width:1px;
     height:280px;
     width:260px;
     margin-top: -257px;
     margin-left: 850px;
     background: url("{% static 'bg1.gif' %}");
     background-size: 100%100%;

}
</style>
<link href="https://fonts.googleapis.com/css?family=Ubuntu+Condensed" rel="stylesheet">

<body>
<form method="POST">
    {% csrf_token %}

    <table class="message">
        <tr>
            <td>Upload Images</td>
            <td><input type="file" name="images"></td>
```

```
        </tr>
        <tr>
            <td>Write Something Here</td>
            <td> <textarea name="tweet" cols="60" rows="8" ></textarea> </td>
        </tr>
        <tr>
            <td><input type="submit" class="submit" value="POST"></td>
            <td><a href="{% url 'tweetview' %}">CHECK YOUR PROCESS</a> </td>
        </tr>
    </table>
</form>
<div class="feedbackimage"></div>
</div>
</body>
{% endblock %}
```

Viewing tweets page code

```
{% extends 'client/base.html' %}
{% block userblock %}
{% load staticfiles %}
<link rel="icon" href="images/icon.png" type="image/x-icon" />

        <link href="https://fonts.googleapis.com/css?family=Lobster" rel="stylesheet">
                <link href="https://fonts.googleapis.com/css?family=Righteous" rel="stylesheet">
```

```
<link href="https://fonts.googleapis.com/css?family=Fredoka+One" rel="stylesheet">
                    <style>
                            body {background-color:#eee;}
                            .container-fluid {padding:50px;}
                            .container{background-color:white;padding:50px;   }
                            #title{font-family: 'Fredoka One', cursive;
}
                            .text-uppercase{
                            font-family: 'Righteous', cursive;
                            }
                            .tweettext{

   border: 2px solid yellowgreen;
   width: 1104px;
   height: 442px;
   overflow: scroll;
  background-color: wheat;
}
                    </style>
          <body>
                      <div class="tweettext">
                         <table>
                            <tr>
                            {% for object in list_objects %}
                            <tr>
```

```
                    <td style="color:red; font-size:25px;">{{object.userId.name}}</td></tr>
                <tr>
                    <td >{{object.images}}</td>
                </tr>
                <tr>
                    <td style="color:blue">{{object.tweet}}</td></tr>
                    </tr>
                    {% endfor %}
                </table>
            </div>
                        <div class="col-md-2">
                                <!-------null------>
                        </div>
{% endblock %}
```

Database

Table structure for table `auth_group`

--

```
CREATE TABLE IF NOT EXISTS `auth_group` (
  `id` int(11) NOT NULL AUTO_INCREMENT,
  `name` varchar(80) NOT NULL,
  PRIMARY KEY (`id`),
  UNIQUE KEY `name` (`name`)
) ENGINE=InnoDB DEFAULT CHARSET=latin1 AUTO_INCREMENT=1 ;
```

```
-- --------------------------------------------------------

--
-- Table structure for table `auth_group_permissions`
--

CREATE TABLE IF NOT EXISTS `auth_group_permissions` (
  `id` int(11) NOT NULL AUTO_INCREMENT,
  `group_id` int(11) NOT NULL,
  `permission_id` int(11) NOT NULL,
  PRIMARY KEY (`id`),
  UNIQUE KEY `auth_group_permissions_group_id_permission_id_0cd325b0_uniq`
(`group_id`,`permission_id`),
  KEY `auth_group_permissio_permission_id_84c5c92e_fk_auth_perm` (`permission_id`)
) ENGINE=InnoDB DEFAULT CHARSET=latin1 AUTO_INCREMENT=1 ;

-- --------------------------------------------------------

--
-- Table structure for table `auth_permission`
--

CREATE TABLE IF NOT EXISTS `auth_permission` (
  `id` int(11) NOT NULL AUTO_INCREMENT,
```

```
  `name` varchar(255) NOT NULL,
  `content_type_id` int(11) NOT NULL,
  `codename` varchar(100) NOT NULL,
  PRIMARY KEY (`id`),
  UNIQUE KEY `auth_permission_content_type_id_codename_01ab375a_uniq`
(`content_type_id`,`codename`)
) ENGINE=InnoDB  DEFAULT CHARSET=latin1 AUTO_INCREMENT=28 ;
-- --------------------------------------------------------

--
-- Table structure for table `auth_user`
--

CREATE TABLE IF NOT EXISTS `auth_user` (
  `id` int(11) NOT NULL AUTO_INCREMENT,
  `password` varchar(128) NOT NULL,
  `last_login` datetime(6) DEFAULT NULL,
  `is_superuser` tinyint(1) NOT NULL,
  `username` varchar(150) NOT NULL,
  `first_name` varchar(30) NOT NULL,
  `last_name` varchar(150) NOT NULL,
  `email` varchar(254) NOT NULL,
  `is_staff` tinyint(1) NOT NULL,
  `is_active` tinyint(1) NOT NULL,
  `date_joined` datetime(6) NOT NULL,
```

```
  PRIMARY KEY (`id`),
  UNIQUE KEY `username` (`username`)
) ENGINE=InnoDB DEFAULT CHARSET=latin1 AUTO_INCREMENT=1 ;

-- --------------------------------------------------------

--
-- Table structure for table `auth_user_groups`
--

CREATE TABLE IF NOT EXISTS `auth_user_groups` (
  `id` int(11) NOT NULL AUTO_INCREMENT,
  `user_id` int(11) NOT NULL,
  `group_id` int(11) NOT NULL,
  PRIMARY KEY (`id`),
  UNIQUE KEY `auth_user_groups_user_id_group_id_94350c0c_uniq` (`user_id`,`group_id`),
  KEY `auth_user_groups_group_id_97559544_fk_auth_group_id` (`group_id`)
) ENGINE=InnoDB DEFAULT CHARSET=latin1 AUTO_INCREMENT=1 ;

-- --------------------------------------------------------

--
-- Table structure for table `auth_user_user_permissions`
--
```

```
CREATE TABLE IF NOT EXISTS `auth_user_user_permissions` (
  `id` int(11) NOT NULL AUTO_INCREMENT,
  `user_id` int(11) NOT NULL,
  `permission_id` int(11) NOT NULL,
  PRIMARY KEY (`id`),
  UNIQUE KEY `auth_user_user_permissions_user_id_permission_id_14a6b632_uniq`
(`user_id`,`permission_id`),
  KEY `auth_user_user_permi_permission_id_1fbb5f2c_fk_auth_perm` (`permission_id`)
) ENGINE=InnoDB DEFAULT CHARSET=latin1 AUTO_INCREMENT=1 ;

-- --------------------------------------------------------

--
-- Table structure for table `client_feedback_model`
--

CREATE TABLE IF NOT EXISTS `client_feedback_model` (
  `id` int(11) NOT NULL AUTO_INCREMENT,
  `name` varchar(100) NOT NULL,
  `mobilenumber` varchar(100) NOT NULL,
  `feedback` varchar(300) NOT NULL,
  PRIMARY KEY (`id`)
) ENGINE=InnoDB DEFAULT CHARSET=latin1 AUTO_INCREMENT=1 ;

-- --------------------------------------------------------
```

```
--
-- Table structure for table `client_tweetmodel`
--

CREATE TABLE IF NOT EXISTS `client_tweetmodel` (
  `id` int(11) NOT NULL AUTO_INCREMENT,
  `tweet` varchar(500) NOT NULL,
  `topics` varchar(300) NOT NULL,
  `sentiment` varchar(300) NOT NULL,
  `images` varchar(100) NOT NULL,
  `userId_id` int(11) NOT NULL,
  PRIMARY KEY (`id`),
  KEY `Client_tweetmodel_userId_id_cee682c4_fk_Client_us` (`userId_id`)
) ENGINE=InnoDB  DEFAULT CHARSET=latin1 AUTO_INCREMENT=206 ;
--
-- Table structure for table `client_userregister_model`
--

CREATE TABLE IF NOT EXISTS `client_userregister_model` (
  `id` int(11) NOT NULL AUTO_INCREMENT,
  `name` varchar(50) NOT NULL,
  `email` varchar(30) NOT NULL,
  `password` varchar(10) NOT NULL,
  `phoneno` varchar(15) NOT NULL,
```

```
  `address` varchar(500) NOT NULL,
  `dob` varchar(20) NOT NULL,
  `country` varchar(30) NOT NULL,
  `state` varchar(30) NOT NULL,
  `city` varchar(30) NOT NULL,
  PRIMARY KEY (`id`)
) ENGINE=InnoDB  DEFAULT CHARSET=latin1 AUTO_INCREMENT=3 ;
```

A

PROJECT REPORT

On

CHARACTERIZING AND PREDICTING EARLY REVIEWERS FOR EFFECTIVE PRODUCT MARKETING ON E-COMMERCE WEBSITES

Submitted by

1)Ms.V.Srija(17K81A0557) 2)Ms. S.Sowmya(17K81A0543)

3)Mr. N. Sandeep(18K85A0501) 4)Mr. M. Rahul(16K81A0536)

in partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Under the Guidance of

Mrs.P.Sabitha

Assistant Professor(M.Tech)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ST.MARTIN'S ENGINEERING COLLEGE
An Autonomous Institute

Dhulapally, Secunderabad – 500 100

JUNE 2021

BONAFIDE CERTIFICATE

This is to certify that the project entitled **CHARACTERIZING AND PREDICTING EARLY REVIEWERS FOR EFFECTIVE PRODUCT MARKETING ON E-COMMERCE WEBSITES,** is being submitted by **1.Ms.V.Srija 17K81A0557, 2.Ms.S.Sowmya 17K81A0543, 3.Mr.N.Sandeep 18K85A0501, 4.Mr.M.Rahul 16K81A0536** in partial fulfillment of the requirement for the award of the degreeof **BACHELOROFTECHNOLOGYINDEPARTMENTOFCOMPUTERSCIENCE AND ENGINEERIN**G is recorded of bonafide work carried out by them. The result embodied in this report have been verified and found satisfactory.

P. Sabitha

Assistant Professor
(M.TECH)

Department of CSE

Head of the Department

Dr.M.NARAYANAN

Department of CSE

InternalExaminer

ExternalExaminer

Place:

Date:

DECLARATION

We, the student of **Bachelor of Technology** in Department of Computer Science and Engineering', session: <2017 – 2021>, St. Martin's Engineering College, Dhulapally, Kompally, Secunderabad, hereby declare that work presented in this Project Work entitled ‒CHARACTERIZINGANDPREDICTINGEARLYREVIEWERSFOREFFECTIVE PRODUT MARKETING ON E-COMMERCE WEBSITES‖ is the outcome of our own bonafide work and is correct to the best of our knowledge and this work has been undertaken taking care of Engineering Ethics. This result embodied in this project report has not been submitted in any university for award of any degree.

Ms.V.Srija 17K81A0557

Ms.S.Sowmya 17K81A0543

Mr.N.Sandeep 18K85A0501

Mr.M.Rahul 16K81A0536

ABSTRACT

Online reviews have become an important source of instruction for users before manufacture an informed procure decision. Early reviews of a product tend to have a high effect on the ensuing product sales. In this paper, we take the initiative to study the behaviour characteristics of early reviewers through their posted reviews on two real-world large e-commerce platforms, i.e., Amazon and Yelp. In specific, we divide product lifetime into three uninterrupted phase and quantitatively characterize early reviewers based on their rating behaviours, the helpfulness scores received from others and the correlation of their reviews with product popularity. By viewing review posting process as a multiplayer competition game, we present a novel margin-based embedding model for early reviewer divination. Extensive experiments on two different e-commerce datasets have shown that our proposed approach outperforms a number of aggressivebaselines.

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible and whose encouragements and guidance have crowded effects with success.

We extended our deep sense of gratitude to Principal, **Dr. P. SANTOSH KUMARPATRA**, St. Martin's Engineering College, Dhulapally, for permitting us to undertake this project.

We are also thankful to **Dr. M.NARAYANAN**, Head of the Department, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout ourproject.

We would like to thank **Dr. T. POONGOTHAI,** Department of Computer Science and Engineering (AI & ML) for her encouragement and insightful comments and as well as our project coordinators **Dr. B.RAJALINGAM**, Associate Professor and **Mr. J.SUDHAKAR**, Assistant Professor, in Department of Computer Science and Engineering for their valuablesupport.

We would like to express our sincere gratitude and indebtedness to our project supervisor P.Sabitha, Assistant Professor (M.Tech), Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

Finally, we express thanks to all those who have helped us successfully to completing this project. Furthermore, we would like to thank our family and friends for their moral support and encouragement.

We express thanks to all those who have helped us in successfully completing the project.

Ms.V.Srija 17K81A0557

Ms.S.Sowmya 17K81A0543

Mr.N.Sandeep 18K85A0501

Mr.M.Rahul 16K81A0536

TABLE OF CONTENTS

LIST OF OUTPUTSCREENS

LIST OF FIGURES

LIST OF ACRONYMS

MERM	Margin Based Embedding Ranking Model
SVM	Supervised Vector Machine
SGD	Stochastic Gradient Descent
NLP	Natural Language Processing
TS	Time Based Spamming
RT	Review Text Spamming

1.INTRODUCTION

1 . INTRODUCTION

1.1 PROJECTOVERVIEW

The emergence of e-commerce websites has enabled users to publish or share purchase experiences by posting product reviews, which usually contain useful opinions, comments and feedback towards a product. As such, a majority of customers will read online reviews before making an informed purchase decision . It has been reported about 71% of global online shoppers read online reviews before purchasing a product . Product reviews, especially the early reviews (i.e., the reviews posted in the early stage of a product), have a high impact on subsequent product sales . We call the users who posted the early reviews early reviewers. Although early reviewers contribute only a small proportion of reviews, their opinions can determine the success or failure of new products and services . It is important for companies to identify early reviewers since their feedbacks can help companies to adjust marketing strategies and improve product designs, which can eventually lead to the success of their new products. For this reason, early reviewers become the emphasis to monitor and attract at the early promotion stage of a company. The pivotal role of early reviews has attracted extensive attention from marketing practitioners to induce consumer purchase intentions . For example, Amazon, one of the largest e-commerce company in the world, has advocated the Early Reviewer Program1, which helps to acquire early reviews on products that have few or no reviews. With this program, Amazon shoppers can learn more about products and make smarter buying decisions. As another related program, Amazon Vine2 invites the most trusted reviewers on Amazon to post opinions about new and prerelease items to help their fellow customers make informed purchasedecisions.

1.2 PROJECTOBJECTIVES

Data Cleaning:

Our data cleaning contains two main steps as follows.

1.Preprocessing 2.Review Spammer Detection and Removal **Preprocessing:**

We first remove reviews from anonymous users, since we would like to associate each review with a unique user. We then remove duplicate reviews often caused by multiple versions of the same product. We also remove inactive users and unpopular products: we

only keep the users who have posted at least ten and five reviews, and products which have received at least ten and five reviews in Amazon and Yelp datasets respectively. For review text, we remove stopwords and very infrequentwords.

Review Spammer Detection and Removal:

Our focus is to study the early adoption behaviors of genuine Amazon and Yelp users. However, The number of spam reviews has increasingly grown on ecommerce websites, and it was found that about 10% to 15% of reviews echoed earlier reviews and might be posted by review spammers. It is possible that spam reviews are posted to give biased or false opinions on some products so as to influence the consumers' perception of the products by directly or indirectly inflating or damaging the product's reputation. The existence of spam reviews could lead to erroneous conclusions in our study. Therefore, we need to remove review spammers as part of our data cleaning process.

ORGANIZATION OF CHAPTERS

1 . Introduction

The emergence of e-commerce websites has enabled users to publish or share purchase experiences by posting product reviews, which usually contain useful opinions, comments and feedback towards a product. As such, a majority of customers will read online reviews before making an informed purchase decision . It has been reported about 71% of global online shoppers read online reviews before purchasing a product . Product reviews, especially the early reviews (i.e., the reviews posted in the early stage of a product), have a high impact on subsequent product sales . We call the users who posted the early reviews early reviewers. Although early reviewers contribute only a small proportion of reviews, their opinions can determine the success or failure of new products andservices

2. Literaturesurvey

In this section,we predicted that early reviews are responsible for increasing product sales. Some additional investigations additionally reveal that product evaluations from previous adopters, like star ratings and sales volume, influence customers' on-line product decisions.

3 .Software and Hardware Requirements

In this chapter, we specified the Software and Hardware components required to develop our project. The Software and Hardware requirements specify the intended purpose, requirements, and nature of software/application/project to be developed. By selecting the dataset that most resembles the usage requirements in our environment, we can use the recommended topology and associated hardware requirements for our topology as a starting point when we plan for hardware of our project.. Requirements may varybased on utilization and observing performance of pilot projects is recommended prior toscale out.

4. Software DevelopmentAnalysis

In this project, we discussed about development and implementation of the project indetail. we developed in our roles as front-end, back-end and database administrator bycollecting relevant data and testing it in required cases.

5. project systemdesign

Thischapterreportsontheanalysisanddesignofourproposedapplication.Thischapterdescribes the system design architecture and database design and is organized in a sequence included with data gathering and system design. Stakeholders will discuss factors such as risk levels, team composition, applicable technologies, time, budget, project limitations, method and architectural design.

Chapter 6: Project Coding

Thischapterisasystemimplementationoftheproject.Wewilldiscussbrieflytheimplementation of our project. This section describes some of the coding templates, outline of various files, class with functionalities, the various methods of input and outputparameters.

Chapter 7: Project Testing

In this chapter, we will discuss briefly the testing of each functionality of our proposed

application in the project. We performed various tasting's like whitebox, blackbox, unit testing, integration testing and many more to check the accuracy and performance of our output. They notify developers of defects in the code. If developers confirm the flaws are valid, they improve the program, and the testers repeat the process until the software is free of bugs and behaves according to requirements.

Chapter 8: Output screens

In this chapter, we captured the screenshots of our project output. We considered few sample inputs and obtained desired outputs for our data with related database.

Chapter 9: Experimental Results

In this chapter, we conclude the performance analysis of our proposed project by comparing it withtheexistingproject.Inthischapter,wediscussbrieflytheconclusionofeachchapterwiththe progressofourproposedsystem.

2.**LITERATURE SURVEY**

2. LITERATURESURVEY

2.1 SURVEY ONBACKGROUND

Our current study is especially associated with the subsequent 3 lines of analysis. Early parent Detection The term of early parent originates from the classic theory for Diffusion of Innovations. AN early parent may seek advice from a trendsetter, e.g., AN early client of a given company, product and technology. The importance of early adopters has been wide studied in social science and political economy. it's been shown that early adopters square measure vital in trend prediction, infective agent selling, product promotion, and so on. Moreover, the influence of early adopters is closely associated with the studies of herd behaviour that describes that people square measure powerfully influenced by the choices of others, like available market bubbles, decision-making, social selling and products success. As for product selling, customers oftentimes choose in style brands as a result of they believe that quality indicates higher quality. As an example, in digital auctions, patrons tend to bid for listings that others already bid for, whereas ignoring similar or additional enticing un-bid-for listings. Similarly, AN experimental study shows that the social influence of early adopters' decisions of songs ends up in each difference and unpredictability of the songs in terms of transfer counts. Some additional investigations additionally reveal that product evaluations from previous adopters, like star ratings and sales volume, influence customers' on-line product decisions. The analysis and detection of early adopters within the diffusion of innovations have attracted a lot of attention from the analysis community. Typically speaking, 3 parts of a diffusion method are studied: attributes of AN innovation, communication channels, and social network structures. Early studies square measure in the main theoretical analysis at the macro level. With the rising of on-line social platforms and also the accessibility of a high volume of social networking information, studies of the diffusion of innovations are mostly conducted on social networks, together with resource-constrained networks, following or re tweet networks, user-click graphs and text-based innovation networks. Modelling Comparison-based preference Comparison-based preference has been studied for many decades and a survey of the classic approaches and strategies was given. By modelling comparison based mostly preference, we will basically perform any ranking task. As an example, in data retrieval (IR), learning to rank aims to be told the ranking for a listing of candidate things with manually designated options. 3 classes of wide used learning to rank approaches embrace purpose wise, pair wise and list wise strategies. Excluding IR, thecompetition-based

ranking strategies have additionally been wide studied in games and matches, wherever the aim is to gauge the talent level of every concerned player. These studies usually solely use a scalar price because the live of the talent rating of a private player. as an example, supported the two-player model,

2.2 CONCLUSIONS ON SURVEY

To predict early reviewers, we propose a novel approach by viewing review posting process as a multiplayer competition game. Only the most competitive users can become the early reviewer's w.r.t. to a product. The competition process can be further decomposed into multiple pairwise comparisons between two players. In a two-player competition, the winner will beat the loser with an earlier timestamp. Inspired by the recent progress in distributed representation learning, we propose to use a margin-based embedding model by first mapping both users and products into the same embedding space, and then determining the order of a pair of users given a product based on their respective distance to the productrepresentation

3.SOFTWARE AND HARDWARE REQUIREMENTS

3. SOFTWARE AND HARDWARE REQUIREMENTS

2.1 SOFTWARE REQUIREMENTS

Operating system : Windows 7 Ultimate

CodingLanguage :Python 3.9

Front-End : Python 3.9

Designing :Html,css,javascript.

Data Base :MySQL

2.2 HARDWAREREQUIREMENTS

System : Pentium IV 2.4GHz.

HardDisk : 40 GB.

FloppyDrive : 1.44 Mb.

Monitor : 14’ ColourMonitor.

Mouse : OpticalMouse.

Ram : 512 Mb.

4. SOFTWARE DEVELOPMENT

ANALYSIS

4. SOFTWARE DEVELOPMENT ANALYSIS

4.1 OVERVIEW OFPROBLEM

Previous studies have highly emphasized the phenomenon that individuals are strongly influenced by the decisions of others, which can be explained by herd behaviour. The influence of early reviews on subsequent purchase can be understood as a special case of herding effect. Early reviews contain important product evaluations from previous adopters, which are valuable reference resources for subsequent purchase decisions. As shown in, when consumers use the product evaluations of others to estimate product quality on the Internet, herd behaviour occurs in the online shopping process. Different from existing studies on herd behaviour, we focus on quantitatively analyzing the overall characteristics of early reviewers using large-scale real-world datasets. In addition, we formalize the early reviewer prediction task as a competition problem and propose a novel embedding based ranking approach to this task. To our knowledge, the task of early reviewer prediction itself has received very little attention in the literature. Our contributions are summarized asfollows:

4.2 DEFINE THEPROBLEM

We present a first study to characterize early reviewers on an e-commerce website using two real-world large datasets. We quantitatively analyze the characteristics of early reviewers and their impact on product popularity. Our empirical analysis provides support to a series of theoretical conclusions from the sociology and economics. We v develop an embedding-based ranking model for the prediction of early reviewers. Our model can deal with the cold-start problem by incorporating side information of products. Extensive experiments on two real-world large datasets, i.e., Amazon and Yelp have demonstrated the effectiveness of our approach for the prediction of earlyreviewers.

4.3 MODULESOVERVIEW

The aim of the project entitled as characterizing and predicting early reviews for effective product marketing on e-commerce websites is used to characterize and Predict the early reviews using two large ecommerce datasets called amazon and yulp.we used k means

clustering algorithm to cluster the datasets into one group based on similarity.

4.4 DEFINE THEMODULES

There are three modules can be divided here for this project they are listed as below

- Uploadproducts
- Product Review BasedOrder
- Rating andReviews
- Data Analysis

From the above three modules, project is implemented

4.5 MODULEFUNCTIONALITY

1. UPLOAD PRODUCTS

Uploading the products is done by admin. Authorized person is uploading the new arrivals to system that are listed to users. Product can be uploaded with its attributes such as brand, color, and all other details of warranty. The uploaded products are able to block or unblock byusers.

2. PRODUCT REVIEW BASEDORDER

The suggestion to user's view of products is listed based on the review by user and rating to particular item. Naïve bayes algorithm is used in this project to develop the whether the sentiment of given review is positive or negative. Based on the output of algorithm suggestion to users is given. The algorithm is applied and lists the products in user side based on the positive andnegative.

3.RATINGS ANDREVIEWS

Ratings and reviews are main concept of the project in order to find effective product marketing. The main aim of the project is to get the user reviews based on how they purchased or whether they purchased or not. The major find out of the project is when they give the ratings and

how effective it is. And this will helpful for the users who are willing to buy the same kind of product.

4. DATA ANALYSIS

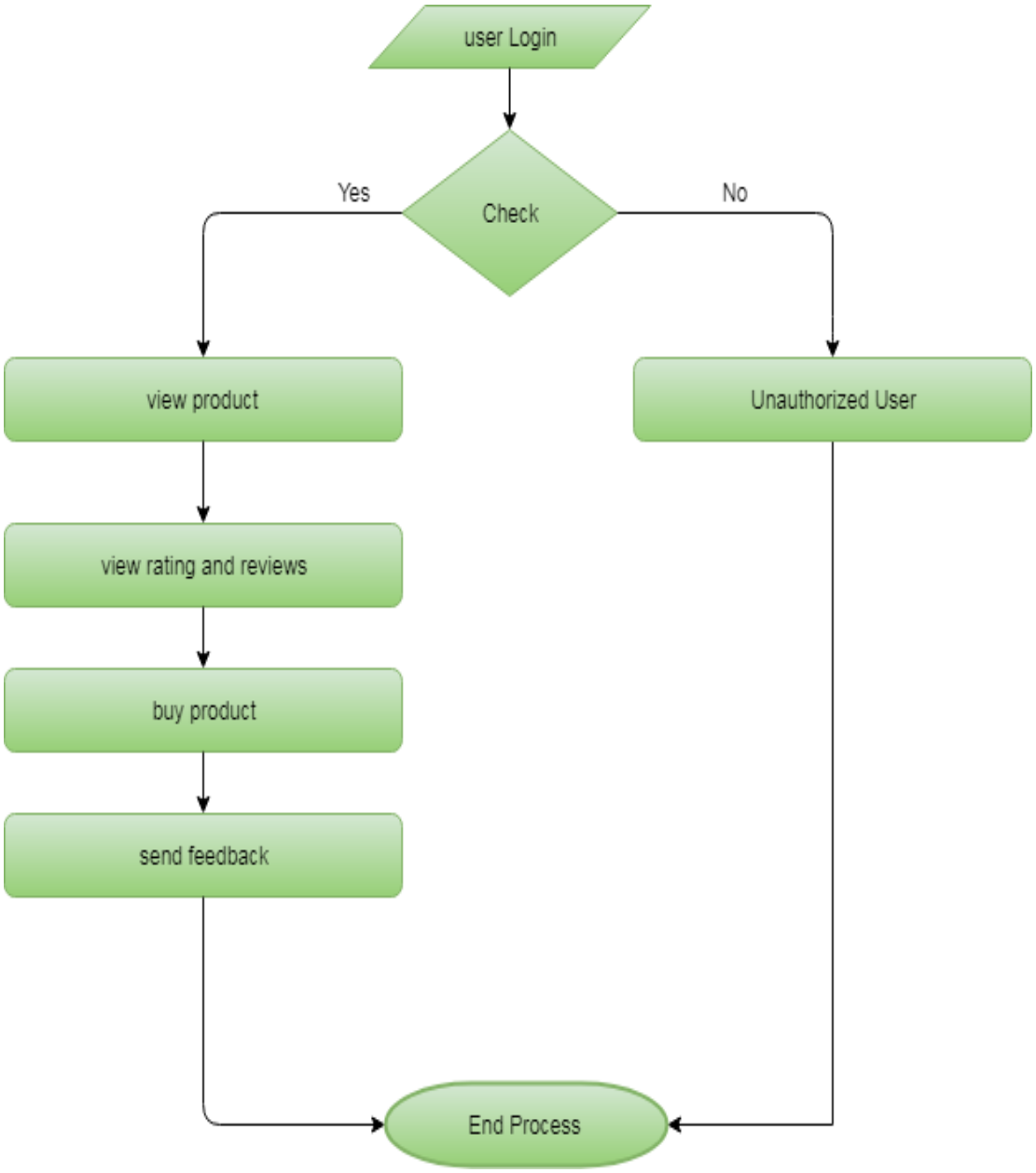
The main part of the project is to analysis the ratings and reviews that are given by the user. The products can be analysis based on the numbers which are given by user. The user data analysis of the data can be done by charts format. The graphs may vary like pie chart, bar chart or some othercharts.

5.PROJECT SYSTEM DESIGN

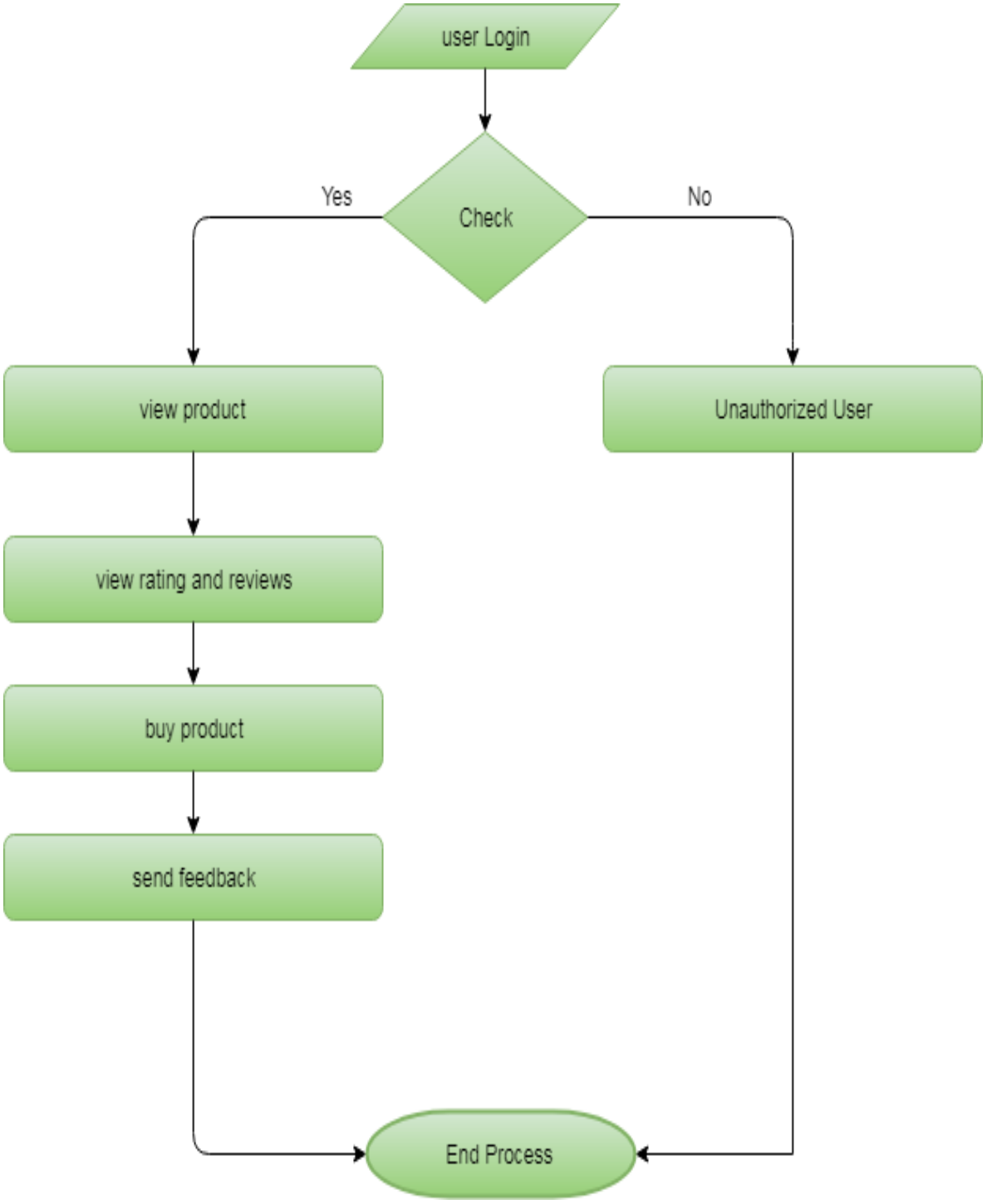
5. PROJECT SYSTEMDESIGN

5.1 DFDS IN CASE OF DATABASEPROJECTS

a) user

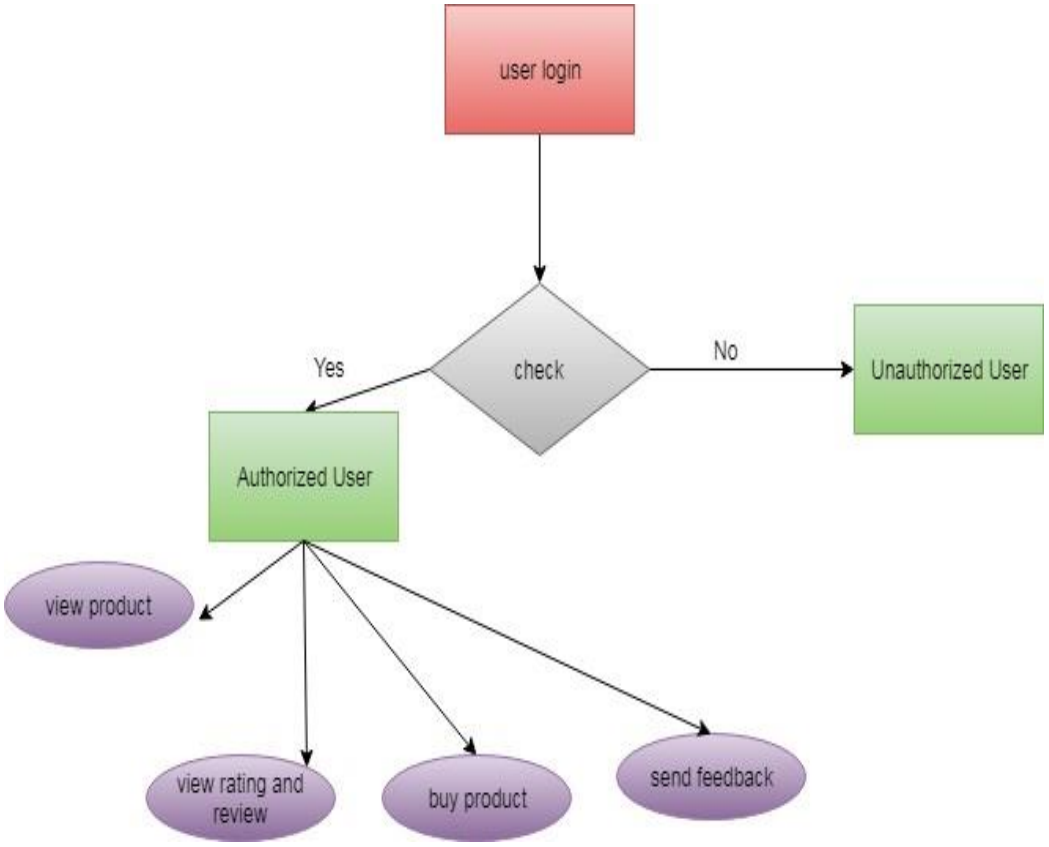


b)admin

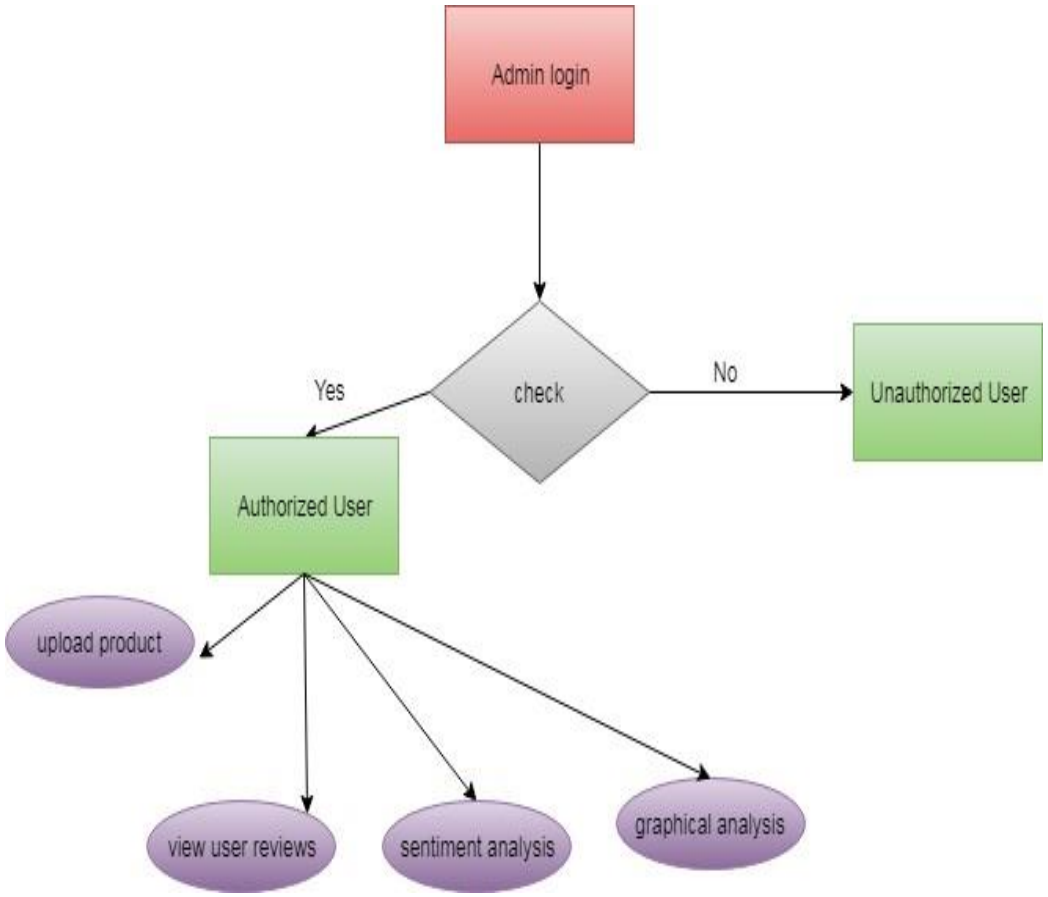

user Login
Check
Yes
No
view product
Unauthorized User
view rating and reviews
buy product
send feedback
End Process

5.2E-RDIAGRAMS

a)user

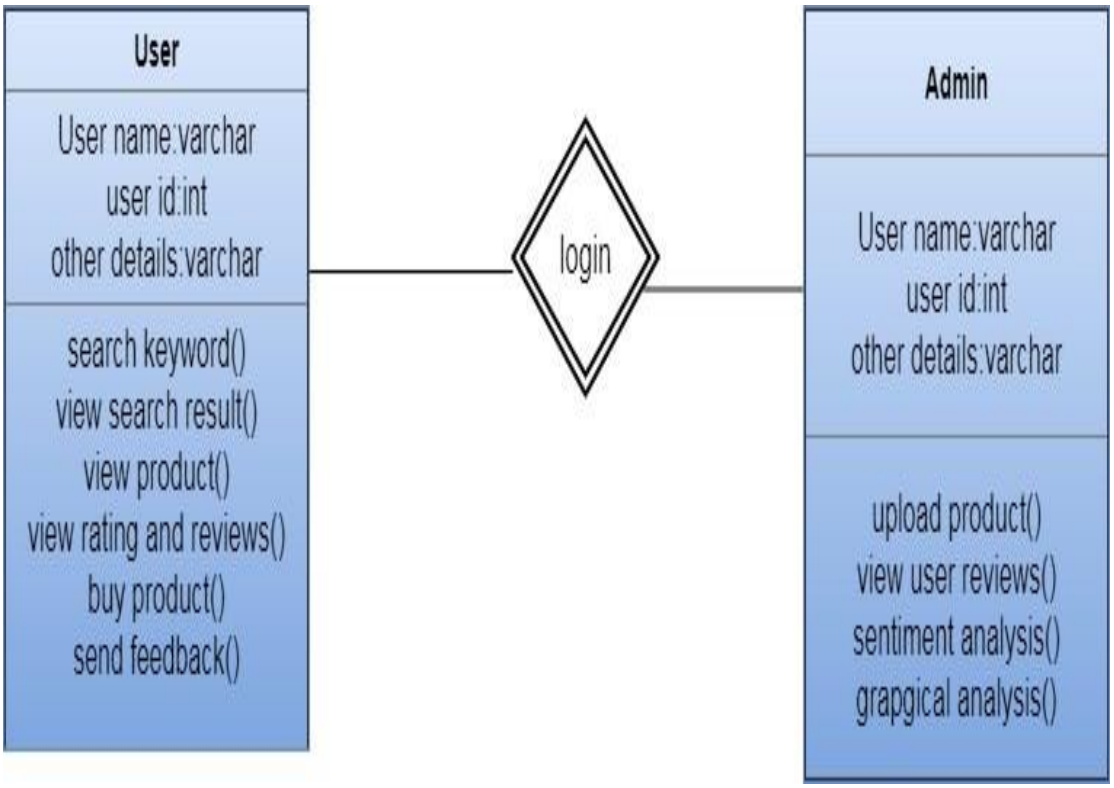


b)admin


Admin login
check
Yes
No
Unauthorized User
Authorized User
upload product
view user reviews
sentiment analysis
graphical analysis

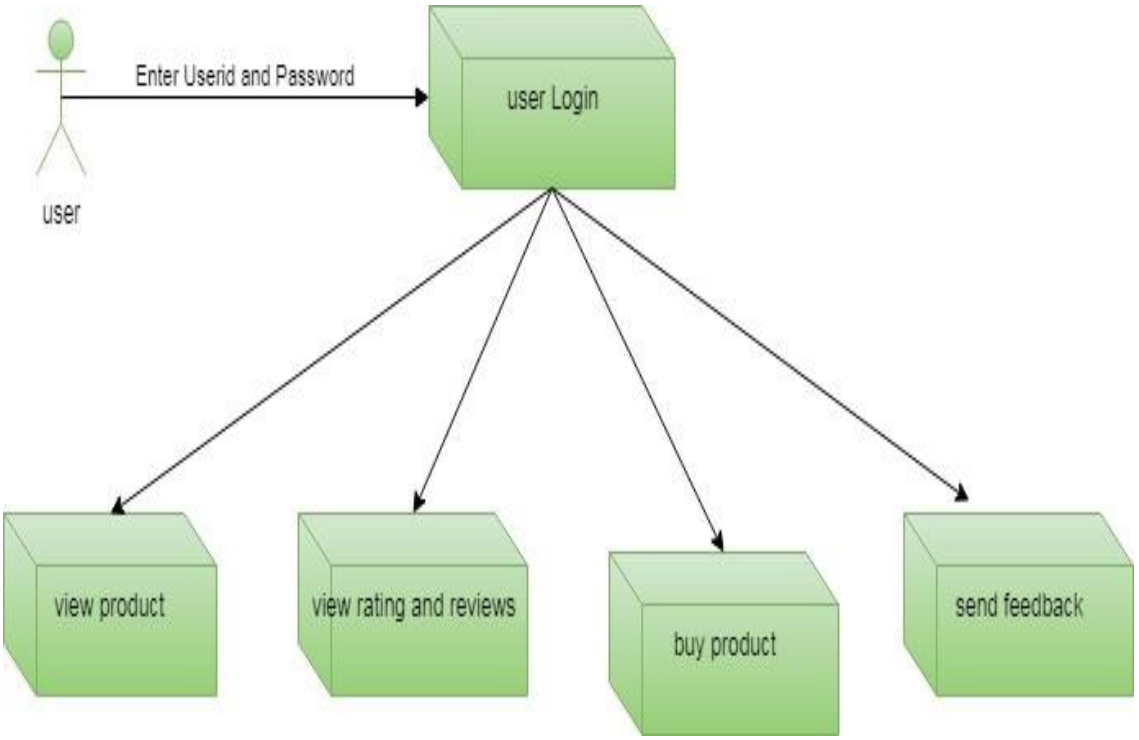
5.3 UMLDIAGRAMS

1. CLASSDIAGRAM

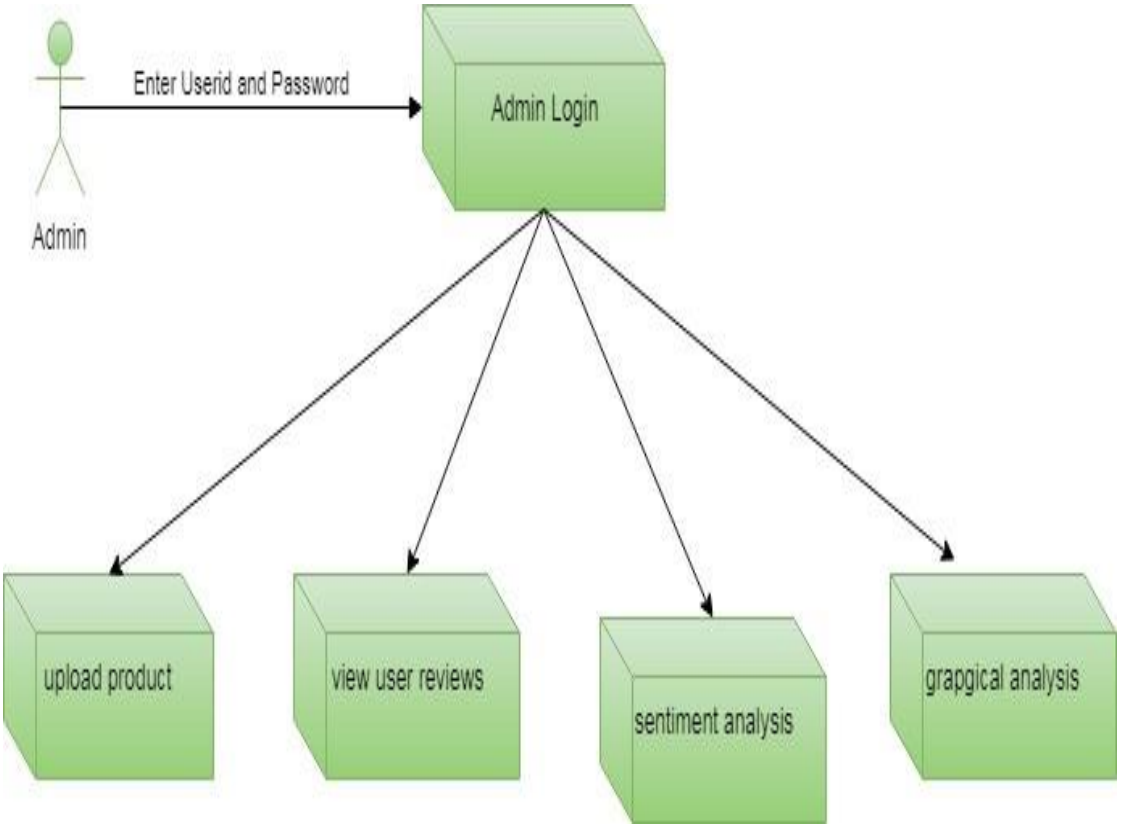


2. COMPONENTDIAGRAM

a) User

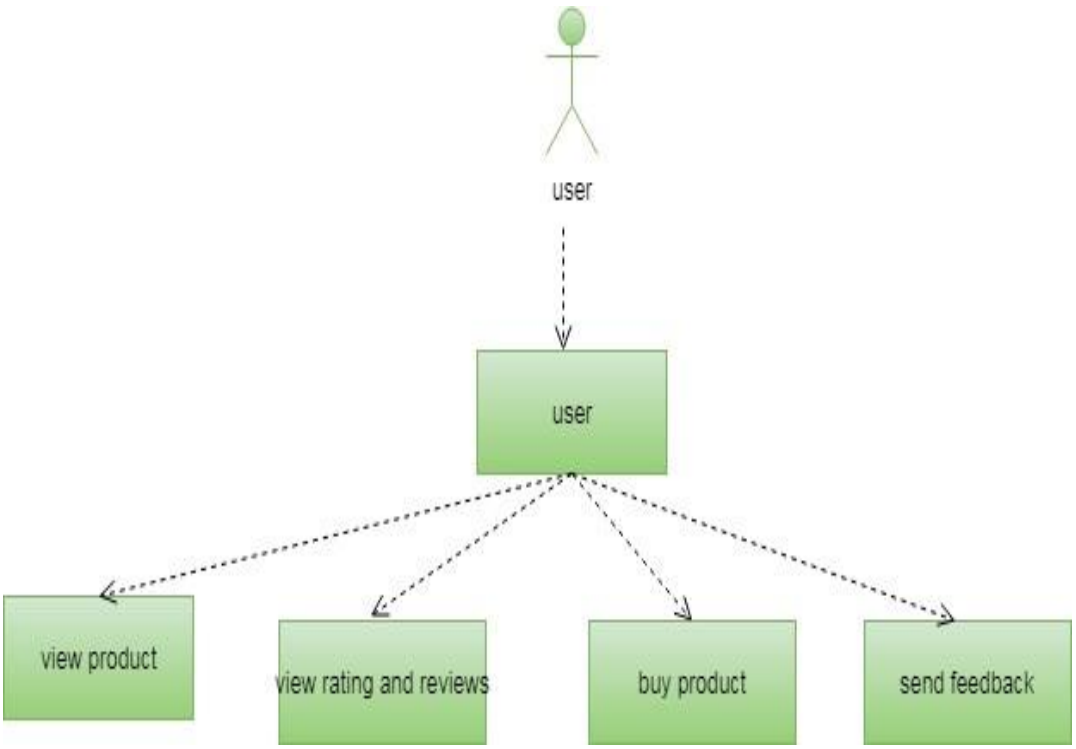


b) Admin

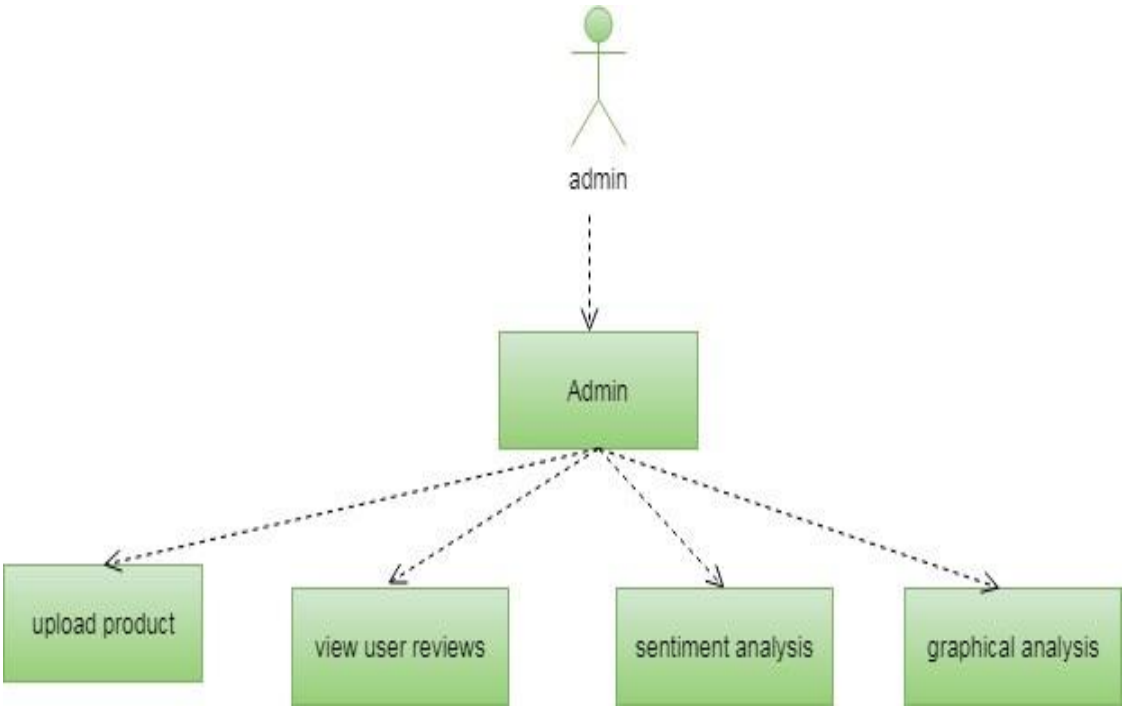

Enter Userid and Password
Admin Login
Admin
upload product
view user reviews
sentiment analysis
grapgical analysis

3. USECASEDIAGRAM

a) User

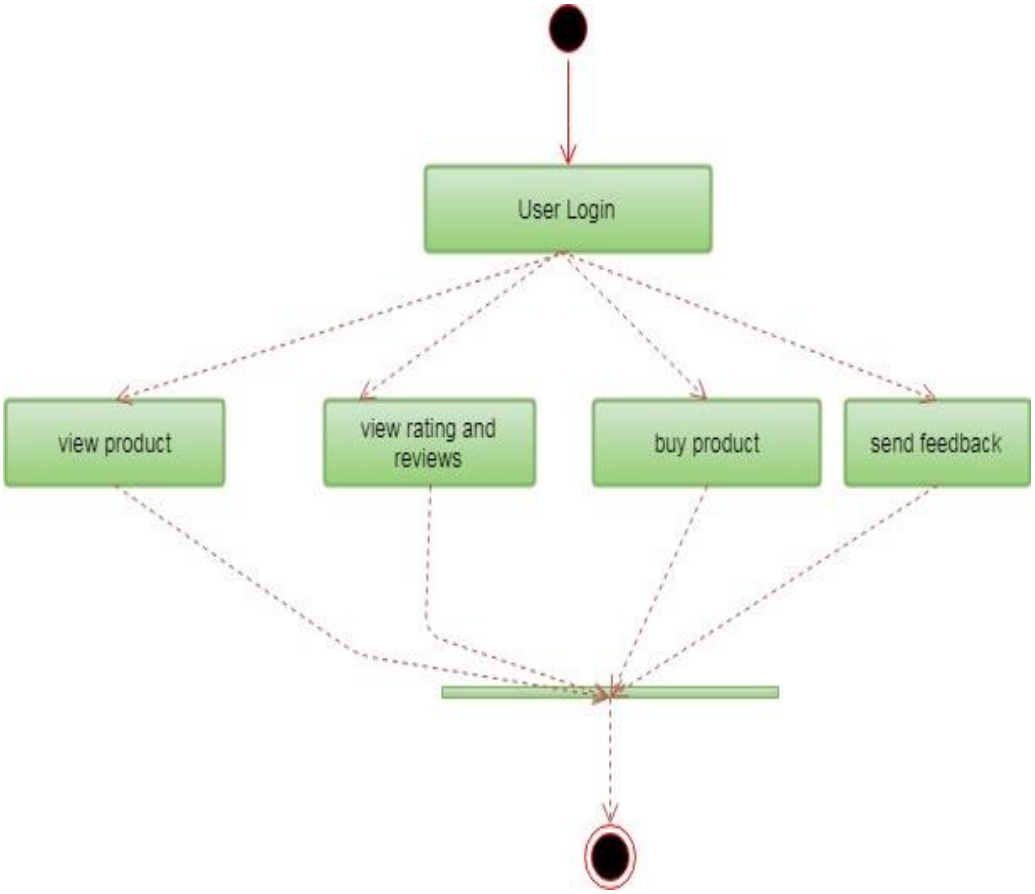


b) admin

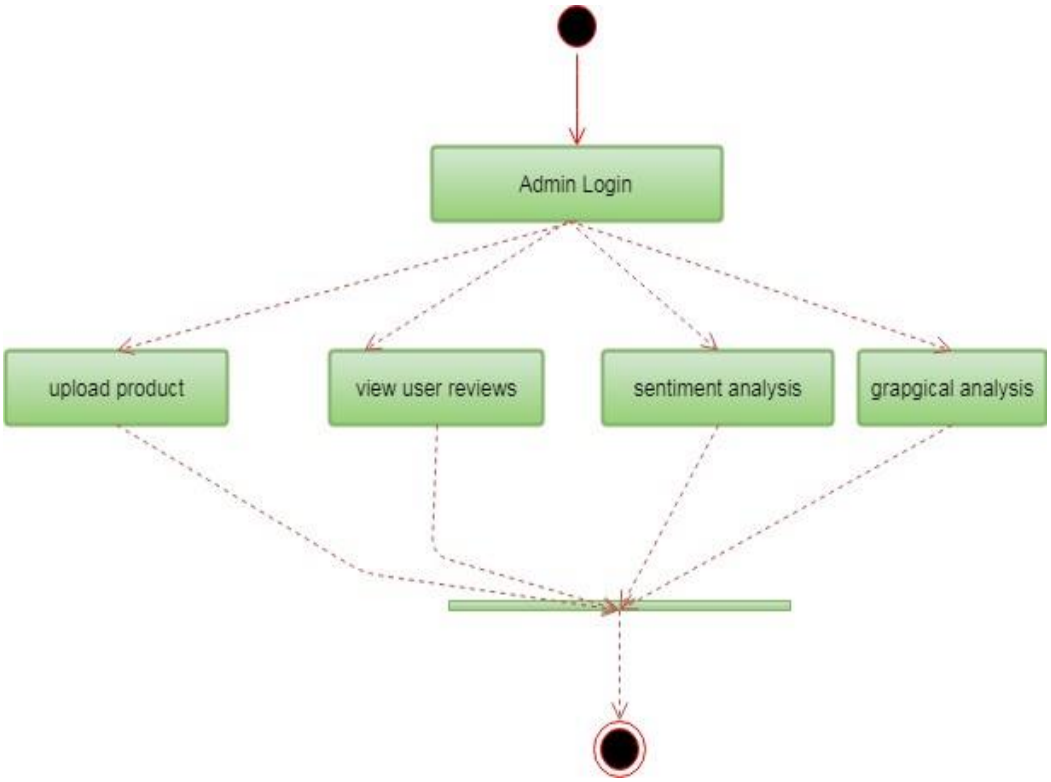

admin
Admin
upload product
view user reviews
sentiment analysis
graphical analysis

4. ACTIVITY DIAGRAM

a)User

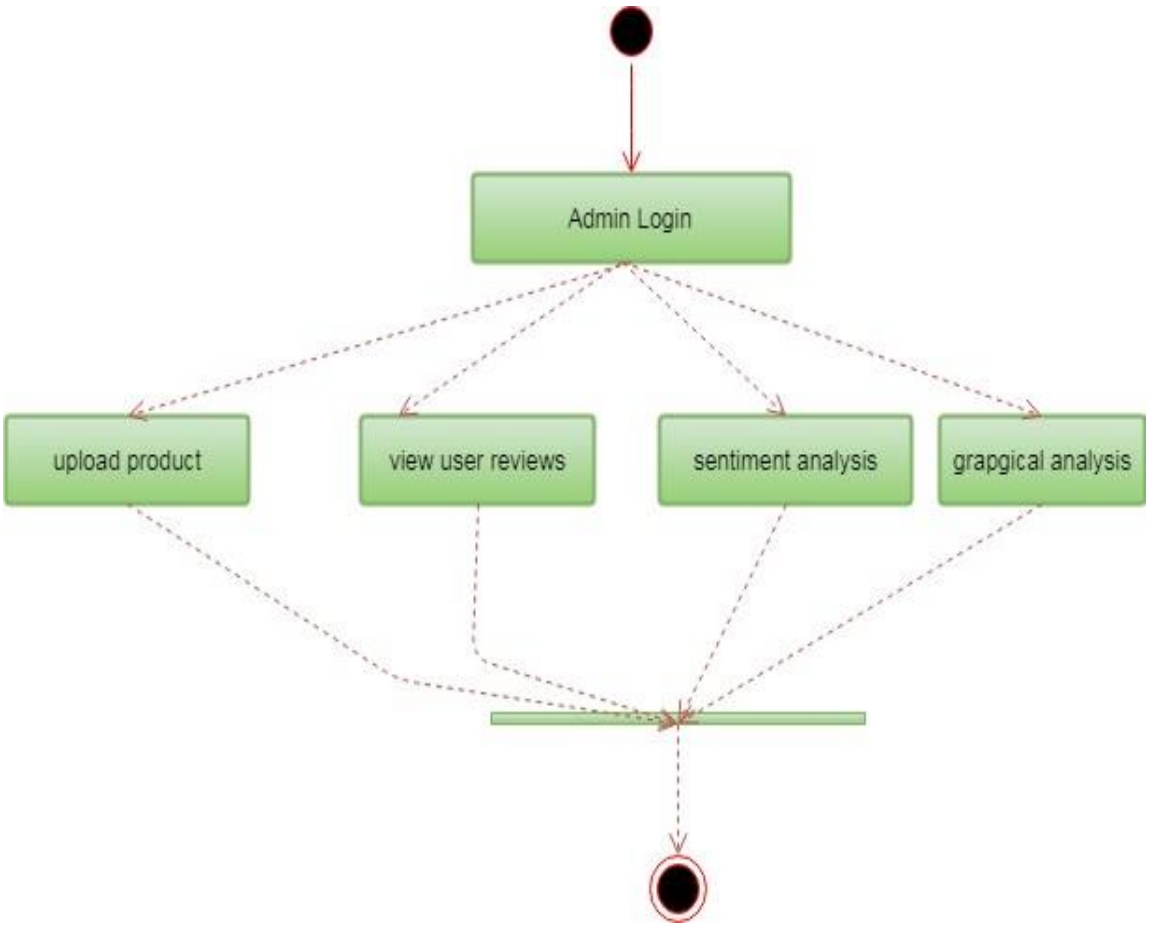


b)Admin

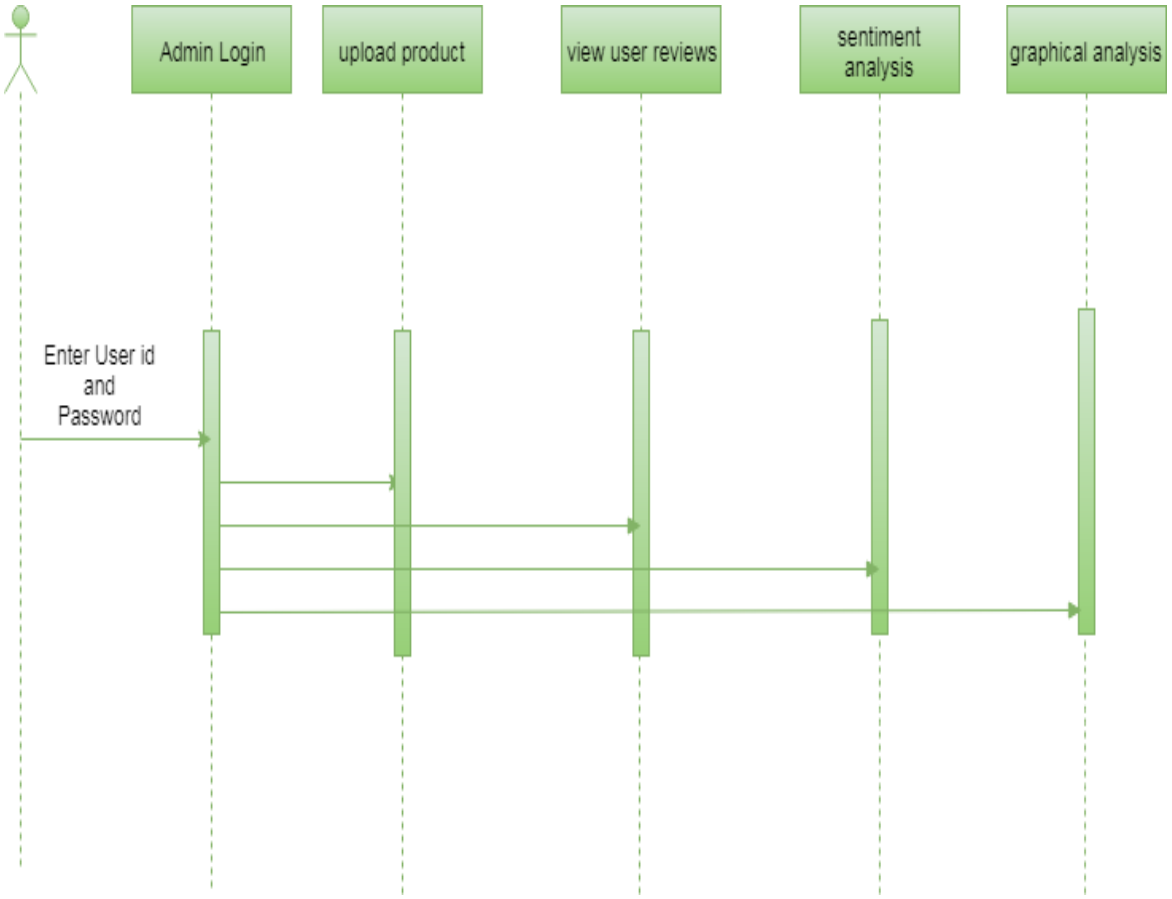

Admin Login
upload product
view user reviews
sentiment analysis
grapgical analysis

5. SEQUENCE DIAGRAM

a)User



a) Admin


Admin Login
upload product
view user reviews
sentiment analysis
graphical analysis
Enter User id and Password

6.PROJECT CODING

6. PROJECTCODING

6.1 CODETEMPLATES

creating a function to load the text and labels from train and test set

```
def get_labels_and_texts(file):
labels=[]
texts=[]
for line in bz2.bz2F(file(file)):
x=line.decode("utf-8")
labels.append(intx[9]-1)
texts.append(x[10:].strip())
return np.array(labels),texts
train_labels,train_texts=get_labels_and_texts("train.ft.txt.bz2")
train_labels,train_texts=get_labels_and_texts("test`.ft.txt.bz2")
```

#Convert from raw binary strings to strings that can be parsed

```
train_file_lines = [x.decode('utf-8') for x in train_file_lines]
test_file_lines = [x.decode('utf-8') for x in test_file_lines]
print(type(train_file_lines), type(test_file_lines), "\n")

print("Train Data Volume:", len(train_file_lines), "\n")
print("Test Data Volume:", len(test_file_lines), "\n\n")

print("Demo: ", "\n")
for x in train_file_lines[:2]:
    print(x, "\n")
print(type(train_file_lines), type(test_file_lines), "\n")
```

#Extracting Labels from the data

```
train_labels = [0 if x.split(' ')[0] == 'label1' else 1 for x in train_file_lines]
test_labels = [0 if x.split(' ')[0] == 'label1' else 1 for x in test_file_lines]
sns.countplot(train_labels)
plt.title('Train Labels distribution')
sns.countplot(test_labels)
plt.title('Test Labels distribution')
```

#Extracting Reviews from the data

```
train_sentences = [x.split(' ', 1)[1][:-1] for x in train_file_lines]
test_sentences = [x.split(' ', 1)[1][:-1] for x in test_file_lines]
```

#Let’s count number of words in reviews and see it distribution

```
train_sentences_size = list(map(lambda x: len(x.split()), train_sentences))

sns.distplot(train_sentences_size)
plt.xlabel("#words in reviews")
plt.ylabel("Frequency")
plt.title("Word Frequency Distribution in Reviews")
train_label_len = pd.DataFrame({"labels": train_labels, "len": train_sentences_size})
train_label_len.head()
```

 # Now we’ll divide it by sentiment and calculate average values

```
def sentiment(n):
    if n>=4:
      return 1
      else 0
products['sentiment']=products['rating'].apply(sentiment)
neg_mean_len = train_label_len.groupby('labels')['len'].mean().values[0]
pos_mean_len = train_label_len.groupby('labels')['len'].mean().values[1]

print(f"Negative mean length: {neg_mean_len:.2f}")
print(f"Positive mean length: {pos_mean_len:.2f}")
print(f"Mean Difference: {neg_mean_len-pos_mean_len:.2f}")
sns.catplot(x='labels', y='len', data=train_label_len, kind='box')
plt.xlabel("labels (0->negative, 1->positive)")
plt.ylabel("#words in reviews")
plt.title("Review Size Categorization")
```

 # Clean URLs

```
for i in range(len(train_sentences)):
   if 'www.' in train_sentences[i] or 'http:' in train_sentences[i] or 'https:' in
train_sentences[i] or '.com' in train_sentences[i]:
       train_sentences[i] = re.sub(r"([^ ]+(?<=\.[a-z]{3}))", "<url>",
train_sentences[i])

for i in range(len(test_sentences)):
    if 'www.' in test_sentences[i] or 'http:' in test_sentences[i] or 'https:' in
test_sentences[i] or '.com' in test_sentences[i]:
       test_sentences[i] = re.sub(r"([^ ]+(?<=\.[a-z]{3}))", "<url>",
test_sentences[i])
```

6.2 OUTLINE FOR VARIOUSFILES

_label<X>label<Y> ...<Text>

where X and Y are the class names.

In this case, the classes arelabel1 andlabel2, and there is only one class per row.

label1 corresponds to 1- and 2-star reviews, and
label2 corresponds to 4- and 5-star reviews.

train.ft.txt.bz2= it is a file which contains all labels and textdatain raw format

test.ft.txt.bz2 = it is a file which stores the labels of all reviews present in dataset

Dataset (folder which contains reviews for training andtesting)

6.3 CLASS WITHFUNCTIONALITY

1.UPLOAD PRODUCTS

Uploading the products is done by admin. Authorized person is uploading the new arrivals to system that are listed to users. Product can be uploaded with its attributes such as brand, color, and all other details of warranty. The uploaded products are able to block or unblock byusers.

2.PRODUCT REVIEW BASEDORDER

The suggestion to user's view of products is listed based on the review by user and rating to particular item. K MEANS CLUSTERING algorithm is used in this project to develop the whether the sentiment of given review is positive or negative. Based on the output of algorithm suggestion to users is given.The
algorithm is applied and lists the products in user side based on the positive and negative.

3.RATINGS ANDREVIEWS

Ratings and reviews are main concept of the project in order to find effective product marketing. The main aim of the project is to get the user reviews based on how they purchased or whether they purchased or not. The major find out of the project is when they give the ratings and how effective it is.

label1 corresponds to 1- and 2-star reviews, and
label2 corresponds to 4- and 5-star reviews.

train.ft.txt.bz2= it is a file which contains all labels and textdatain raw format

test.ft.txt.bz2 = it is a file which stores the labels of all reviews present in dataset

Dataset (folder which contains reviews for training andtesting)

6.3 CLASS WITHFUNCTIONALITY

1.UPLOAD PRODUCTS

Uploading the products is done by admin. Authorized person is uploading the new arrivals to system that are listed to users. Product can be uploaded with its attributes such as brand, color, and all other details of warranty. The uploaded products are able to block or unblock byusers.

2.PRODUCT REVIEW BASEDORDER

The suggestion to user's view of products is listed based on the review by user and rating to particular item. K MEANS CLUSTERING algorithm is used in this project to develop the whether the sentiment of given review is positive or negative. Based on the output of algorithm suggestion to users is given.The
algorithm is applied and lists the products in user side based on the positive and negative.

3.**RATINGS ANDREVIEWS**

Ratings and reviews are main concept of the project in order to find effective product marketing. The main aim of the project is to get the user reviews based on how they purchased or whether they purchased or not. The major find out of the project is when they give the ratings and how effective it is. And this will helpful for the users who are willing to buy the same kind ofproduct.

4.DATA ANALYSIS

The main part of the project is to analysis the ratings and reviews that are given by the user. The products can be analysis based on the numbers which are given by user. The user data analysis of the data can be done by charts format. The graphs may vary like pie chart, bar chart or some othercharts.

6.4 METHODS INPUT AND OUTPUTPARAMETERS

A) REQUEST.METHOD: This method plays a major role in getting user id and password we can observe the request method in codebelow

```
#Code
from django.db.models import Count, Avg
from django.shortcuts import render, redirect, get_object

from admins.forms import UploadForm
from admins.models import Prodcuts
from user.models import Purchase, Feedback

#Code
def index(request):
    if request.method=="POST":
        username=request.POST.get('username','')
        password=request.POST.get('password','')
        if username=='admin' and password=='admin':
            request.session['userid']=1
            request.session['username']='admin'
            return redirect('admins:home')
    return render(request,'admins/index.html',)
#code

def home(request):
    products=Prodcuts.objects.all()
    return render(request,'admins/home.html',{'products':products})
```

#Code

```
def uploadproducts(request):
    if request.method=="POST":
        forms=UploadForm(request.POST, request.FILES)
        if forms.is_valid():
            forms.save()
            return redirect('admins:home')
    else:
        forms = UploadForm()
    return render(request,'admins/uploadproducts.html',{'form':forms})
```

#Code

```
def charts(request,chart_type):
    d=None
    if chart_type=='all':
        d=Feedback.objects.values('product').annotate(dcount=Count('rating'))
    elif chart_type=='mobile':

d=Feedback.objects.filter(productproduct_name='mobile').values('productvendor_na
me').annotate(dcount=Count('rating'))
    elif chart_type=='laptop':

d=Feedback.objects.filter(productproduct_name='laptop').values('productvendor_nam
e').annotate(
            dcount=Count('rating'))
    elif chart_type=='mobileaccessories':
        d=Feedback.objects.filter(productproduct_name='mobile
accessories').values('productvendor_name').annotate(
            dcount=Count('rating'))
    elif chart_type=='watches':

d=Feedback.objects.filter(productproduct_name='watches').values('productvendor_na
me').annotate(
            dcount=Count('rating'))
    elif chart_type=='shoes':
```

```
d=Feedback.objects.filter(productproduct_name='shoes').values('productvendor_name ').annotate(
dcount=Count('rating'))
    return render(request,'admins/charts.html',{'chart_type':chart_type,'d':d})
```

#Code

```
def charts1(request,chart_type):
    d=Feedback.objects.values('userprofession').annotate(dcount=Count('rating'))
    return render(request,'admins/charts1.html',{'chart_type':chart_type,'d':d})

def charts2(request,chart_type):
    d=Feedback.objects.values('userlocation').annotate(dcount=Count('rating'))
    return render(request,'admins/charts2.html',{'chart_type':chart_type,'d':d})

def charts3(request,chart_type):

d1=Feedback.objects.filter(sentiment='positive').values('productproduct_name').annotate
(dcount=Count('sentiment'))

d2=Feedback.objects.filter(sentiment='negative').values('productproduct_name').annotat
e(dcount=Count('sentiment'))
    return render(request,'admins/charts3.html',{'chart_type':chart_type,'d1':d1,'d2':d2})

def logout(request):

return redirect('admins:index')
```

#code

```
from sklearn.datasets import load_amazondataset

from sklearn.cluster import kmeans

amazon=load_amazon()

kmeans=kmeans(n_clusters=3)

KMmodel=kmeans.fit(amazon.data)

kMmodels.labels_
```

#Code

```
from sklearn.metrics import accuracy_score


# To predict our tags (i.e. whether requesters get their pizza),
# we feed the vectorized `test_set` to .predict()
predictions_valid = clf.predict(test_set)


print('Amazon Sentiment Analysis Accuracy = {}'.format(
        accuracy_score(predictions_valid, test_labels[:1000]) * 100)
    )
```

7.PROJECTTESTING

7. PROJECT TESTING

7.1 VARIOUS TESTCASES

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

TYPES OF TESTS

Unit testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expectedresults.

Integration testing

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

Functional test

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items:

Valid Input : identified classes of valid input must be accepted.

Invalid Inpu : identified classes of invalid input must be rejected.

Functions : identified functions must be exercised.

Output: identified classes of application outputs must be exercised

System: interfacing systems or procedures must beinvoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests isdetermined.

System Test

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

7.2 BLACK BOX

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box .you cannot –see into it. The test provides inputs and responds to outputs without considering how the software works.

7.3 WHITE BOXTESTING

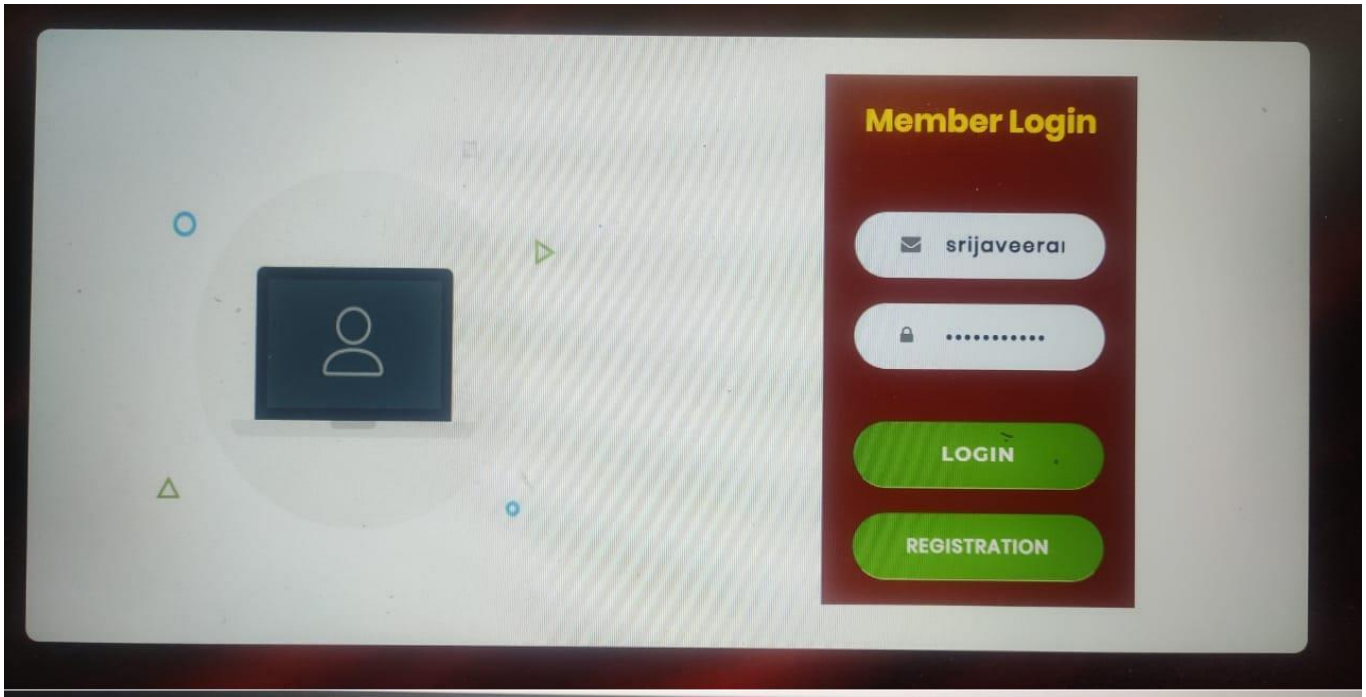
White Box Testing is a testing in which in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is purpose. It is used to test areas that cannot be reached from a black box level.

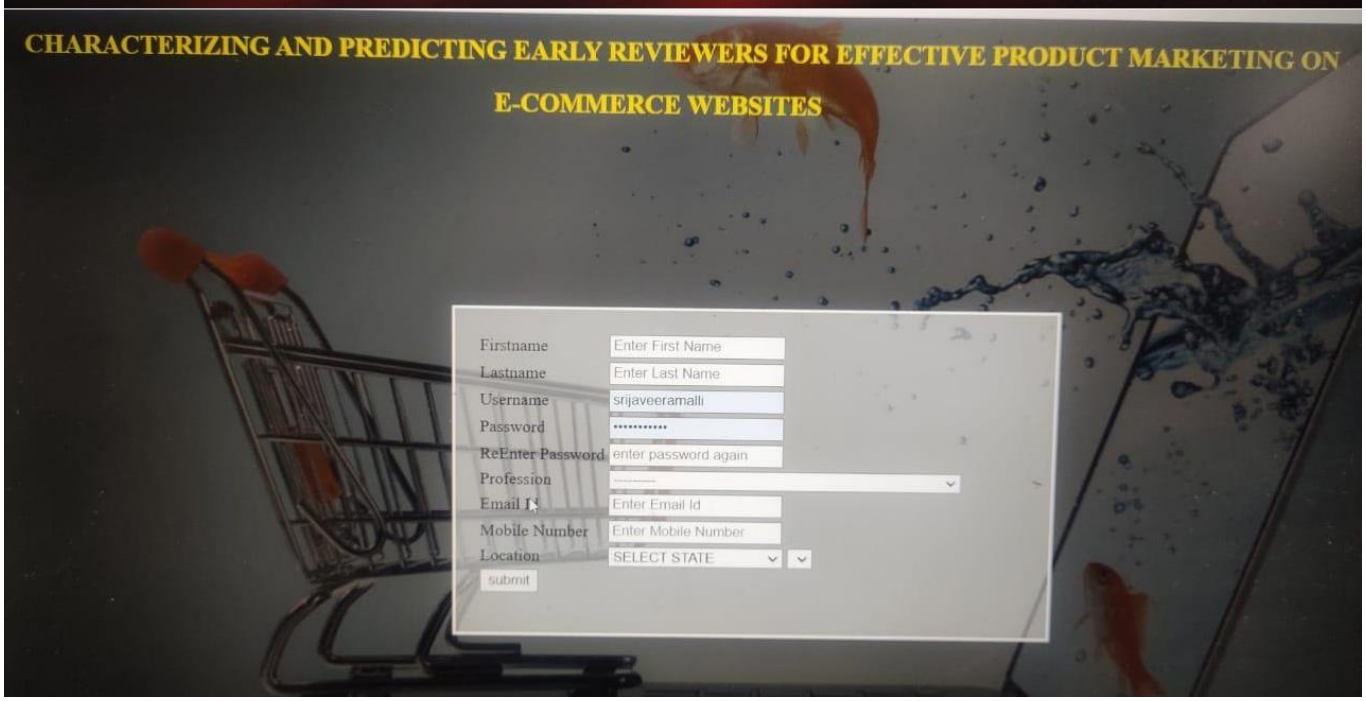
Test Results: All the test cases mentioned above passed successfully. No defects encountered.

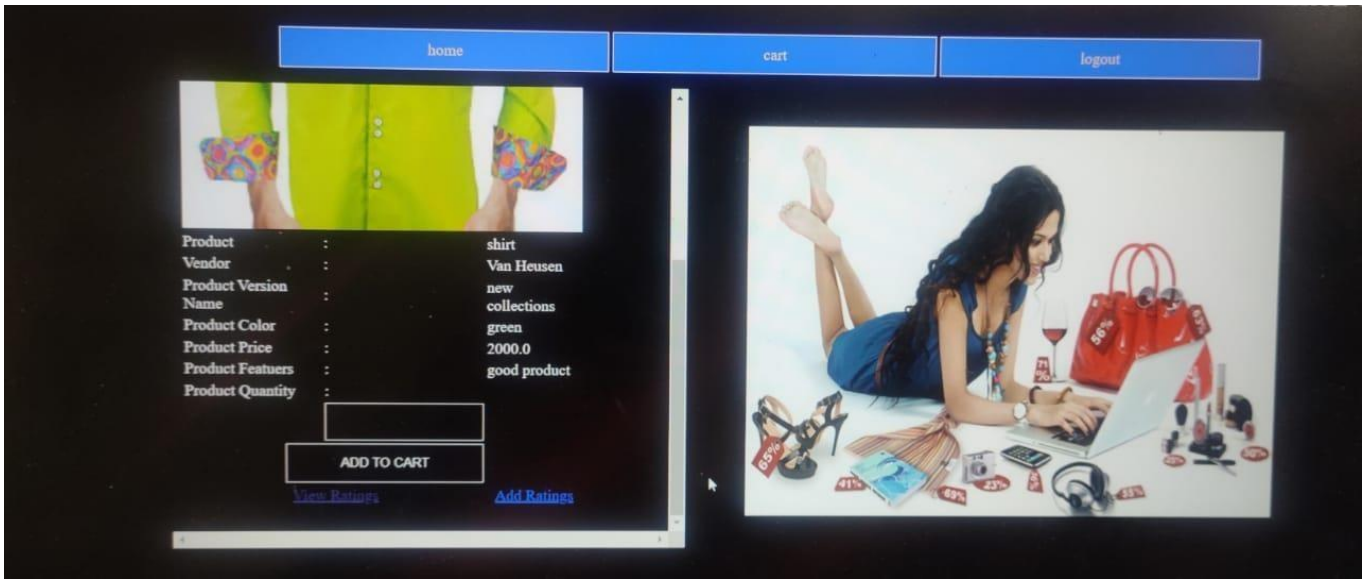
8.OUTPUTSCREENS

8. OUTPUTSCREENS

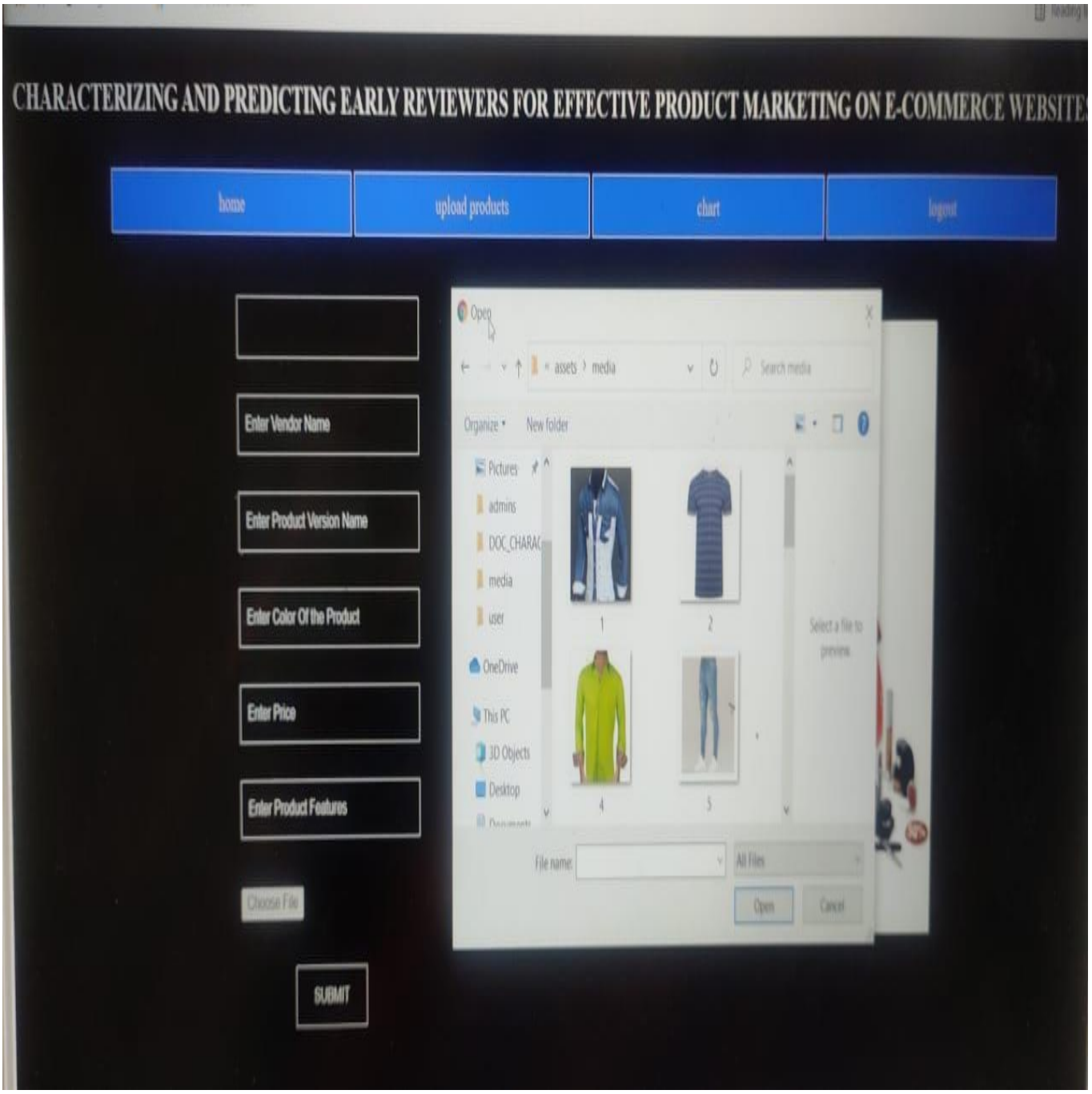
8.1 USER INTERFACES



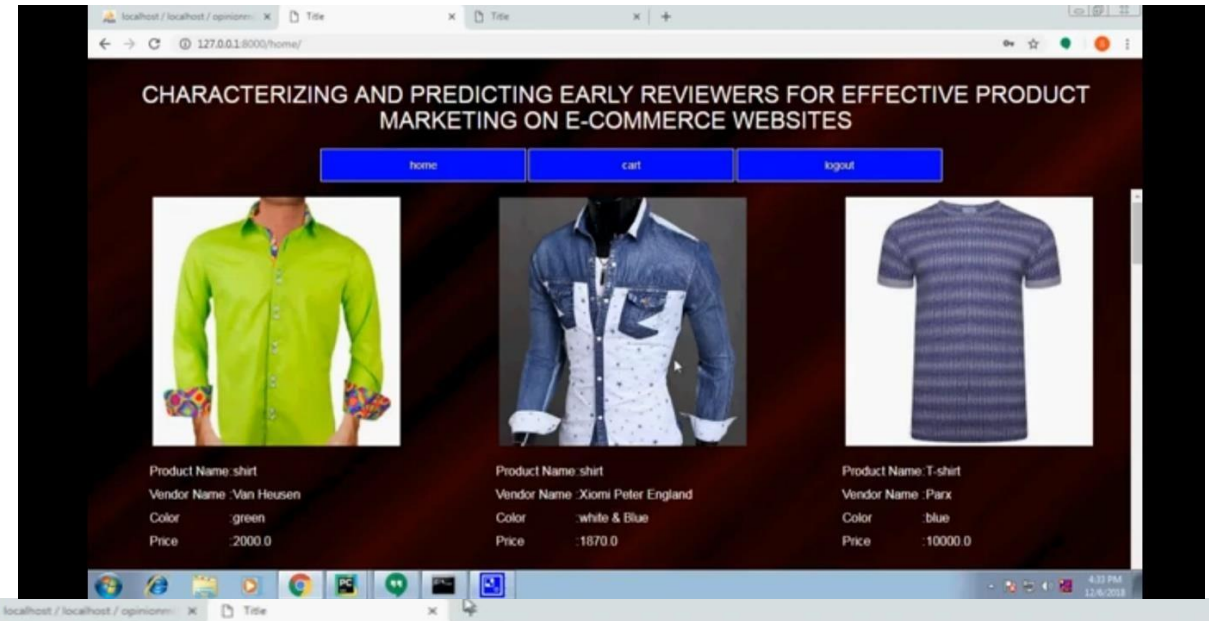


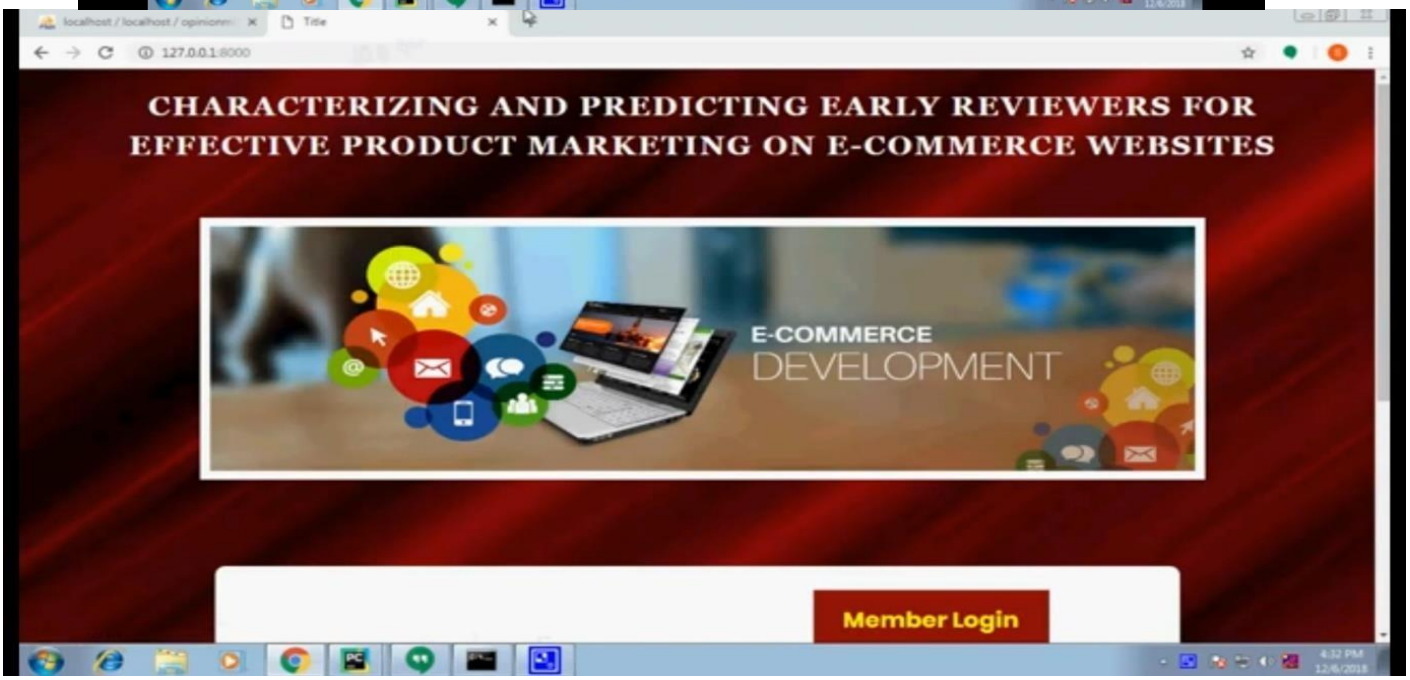


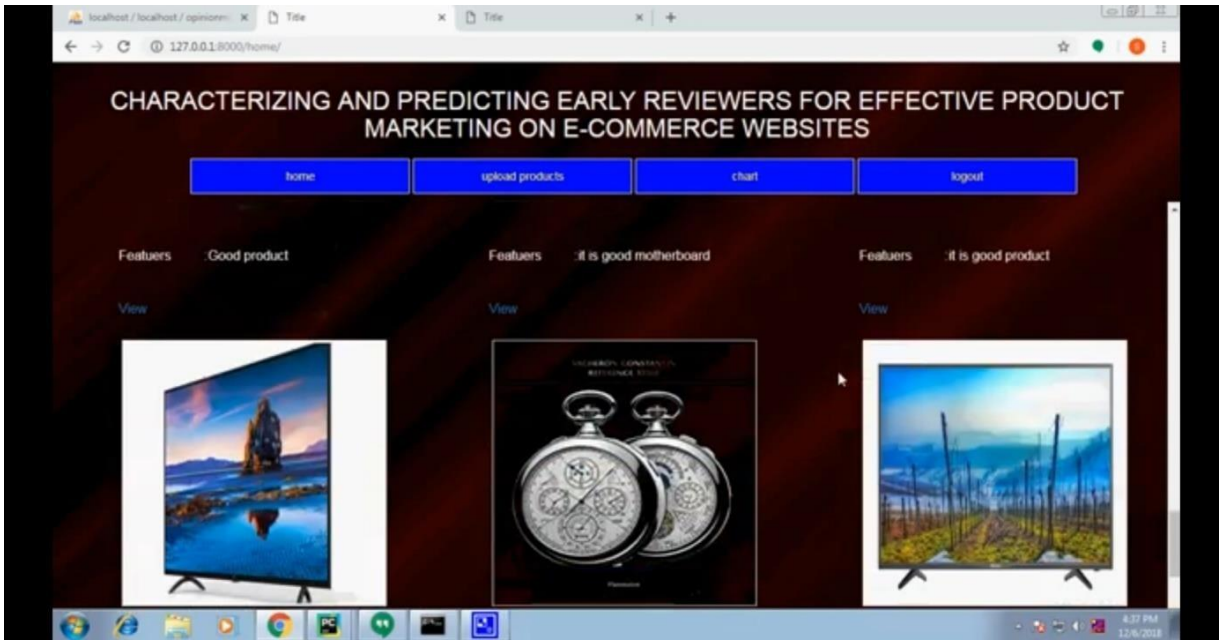
Customer Name	Product	Purchasing Status	Rating	Review
siva	shirt-new collections- Van Heusen	purchased	2	good
vinay	shirt-new collections- Van Heusen	purchased	5	worst product I have ever purchased
vinay	shirt-new collections- Van Heusen	purchased	3	not good
vinay	shirt-new collections- Van Heusen	not purchased	1	it is good
vinay	shirt-new collections- Van Heusen	not purchased	2	worst product I have ever purchased
vinay	shirt-new collections- Van Heusen	purchased	3	good
vinay	shirt-new collections- Van Heusen	purchased	5	poor product i have ever seen

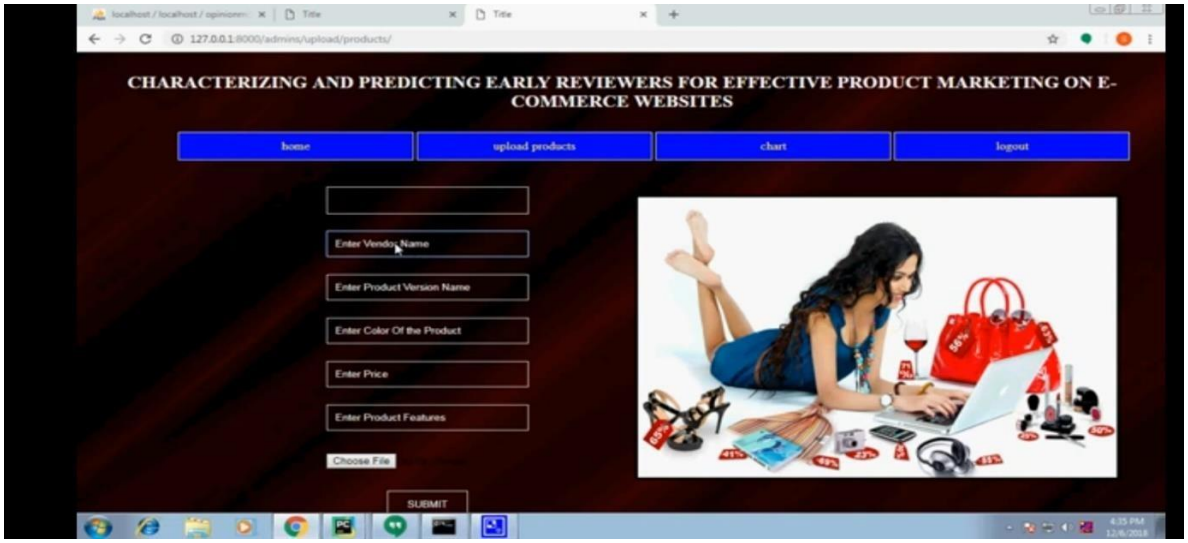

CHARACTERIZING AND PREDICTING EARLY REVIEWERS FOR EFFECTIVE PRODUCT MARKETING ON E-COMMERCE WEBSITE
home
upload products
chart
logout
Enter Vendor Name
Enter Product Version Name
Enter Color Of the Product
Enter Price
Enter Product Features
Choose File
SUBMIT
Open
assets › media
Search media
Organize
New folder
Pictures
admins
media
user
OneDrive
This PC
3D Objects
Desktop
Select a file to preview
File name:
All Files
Open
Cancel

8.2 OUTPUT SCREENS



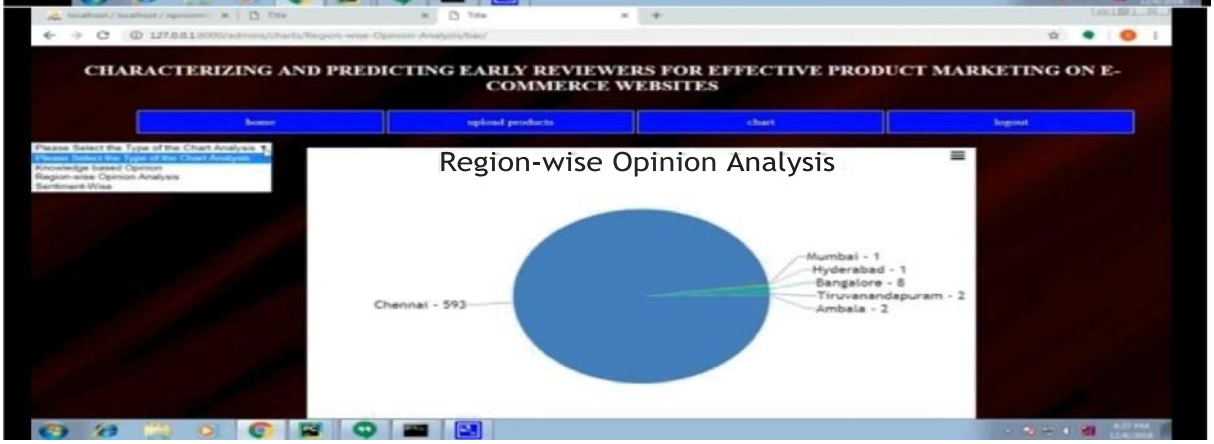



CHARACTERIZING AND PREDICTING EARLY REVIEWERS FOR EFFECTIVE PRODUCT MARKETING ON E-COMMERCE WEBSITES
home
upload products
chart
logout
Featuers :Good product
View
Featuers :it is good motherboard
View
Featuers :it is good product
View


CHARACTERIZING AND PREDICTING EARLY REVIEWERS FOR EFFECTIVE PRODUCT MARKETING ON E-COMMERCE WEBSITES
home
upload products
chart
logout
Enter Vendor Name
Enter Product Version Name
Enter Color Of the Product
Enter Price
Enter Product Features
Choose File
SUBMIT


CHARACTERIZING AND PREDICTING EARLY REVIEWERS FOR EFFECTIVE PRODUCT MARKETING ON E-COMMERCE WEBSITES
home
upload products
chart
logout
Please Select the Type of the Chart Analysis
Comparision of Various Vendors For Product : all
Ratings
Vendor-wise Gross Ratings for Product

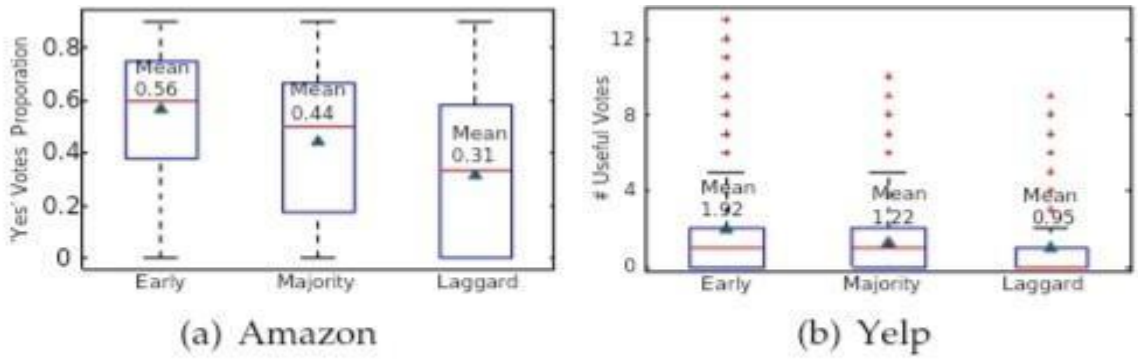

CHARACTERIZING AND PREDICTING EARLY REVIEWERS FOR EFFECTIVE PRODUCT MARKETING ON E-COMMERCE WEBSITES
home
upload products
chart
logout
Please Select the Type of the Chart Analysis
Com par is ion of Products Based on Sentiments
Ratings
shirt
T-shirt
jeans pant
cotten pant
sudithar
mobile accesories
mobile


CHARACTERIZING AND PREDICTING EARLY REVIEWERS FOR EFFECTIVE PRODUCT MARKETING ON E-COMMERCE WEBSITES
Region-wise Opinion Analysis
Chennai - 593
Mumbai - 1
Hyderabad - 1
Bangalore - 8
Tiruvanandapuram - 2
Ambala - 2

9.EXPERIMENTAL RESULTS

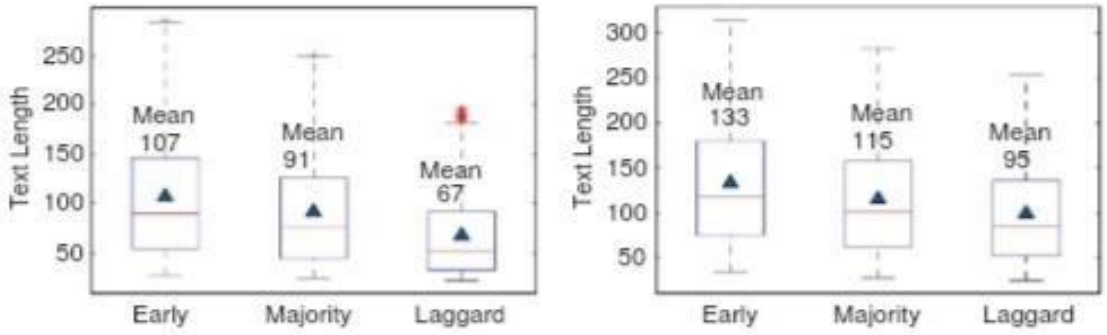
9.EXPERIMENTALRESULTS

9.1 ANALYASING EXPERIMENTALDATA



(a) Amazon (b) Yelp

Fig. 7. Comparisons of the helpfulness scores by the three categories of reviews.



(a) Amazon (b) Yelp

Fig. 8. Comparisons of review text length by the three categories of reviews.

INTERPRETING EXPERIMENTALRESULT

Category	No.of yes	No.of no	Normalized yes	Normalized no
Early	15.5	3.28	0.72	0.28
Majority	4.98	2.28	0.68	0.32
Laggards	2.23	1.44	0.67	0.33

Amazon and Yelp datasets indicate that early reviews are more helpful with higher helpfulness scores than those from the other two categories. This might be caused bytheaccumulation time of review data: early reviews themselves tend to receive more attention. To reduce the effect of time span, in Amazon dataset, we report both the count of Yesand No votes and the normalized Yes and No votes (i.e., the proportion of Yes and No votes) by the three categories in Table 2. It can be observed that the counts of both Yes and No votes for early reviews are significantly higher than those of the other two categories, especially the count of Yes votes. The higher normalized Yes votes of earlyreviews indicates that early reviewers tend to post more helpful reviews.

10.CONCLUSION AND FUTURE ENHANCEMENT

10.CONCLUSION AND FUTUREENHANCEMENT

we have studied the novel task of early reviewer characterization and prediction on two real-world online review datasets. Our empirical analysis strengthens a series of theoretical conclusions from sociology and economics. We found that (1) an early reviewer tends to assign a higher average rating score; and (2) an early reviewer tends to post more helpful reviews. Our experiments also indicate that early reviewers' ratings and their received helpfulness scores are likely to influence product popularity at a later stage. We have adopted a competition-based viewpoint to model the review posting process, and developed a margin based embedding ranking model (MERM) for predicting early reviewers in a cold-start setting.

In our current work, the review content is not considered. In the future, we will explore effective ways in incorporating review content into our early reviewer prediction model. Also, we have not studied the communication channel and social network structure in diffusion of innovations partly due to the difficulty in obtaining the relevant information from our review data. We will look into other sources of data such as Flixster in which social networks can be extracted and carry out more insightful analysis. Currently, we focus on the analysis and prediction of early reviewers, while there remains an important issue to address, i.e., how to improve product marketing with the identified early reviewers. We will investigate this task with real e-commerce cases in collaboration with e-commerce companies in thefuture.

11.REFERENCES

11.REFERENCES

[1] N. Aaraj, S. Ravi, S. Raghunathan, and N. K. Jha, ―Architectures for efficient face authentication in embedded systems,‖ inProc. Design, Autom. Test Eur., Mar.2006, vol. 2, pp.1–6.

[2] M.D.Marsico,M.Nappi,andD.Riccio,―FARO:Facerecognitionagainstocclusions and expression variations,‖ IEEE Trans. Syst., Man, Cybern. A, Syst.,Humans, vol. 40,no. 1, pp.121–132, Jan.2010.

[3] F.Abate,M.Nappi,D.Riccio,andG.Tortora,―RBS:Arobust bimodalsystemfor face recognition,‖Int. J. Softw. Eng. Knowl. Eng., vol. 17, no. 4, pp.497–514,2007.

[4] N.J. Belkin, P. B. Kantor, E. A. Fox, and J. A. Shaw, ―Combining evidence of multiple query representation for information retrieval,‖ Inf. Process. Manag., vol.3, no. 31, pp. 431–448,1995

[5] R.M.Bolle,J.H.Connell,S.Pananti,N.K.Ratha,andA.W.Senior,―Therelation between the ROC curve and the CMC,‖ inProc. 4th IEEE Work. Automat.Identification Adv. Technol., 2005, pp.15–20.

[6] D. Delgado-Gomez, F. Sukno, D. Aguado, C. Santacruz, and A.ArtesRodriguez, ―Individual identification using personality traits,‖J. Netw. Comput. Appl., vol.33, no. 3, pp. 293–299, May 2010.

[7] M. D. Marsico, M. Nappi, and D. Riccio, ―HERO: Human ear recognition against occlusions,‖ inProc. IEEE Comput. Soc. Workshop Biometrics—In Assoc. IEEE Conf. Comput. Vis. Pattern Recognit.— CVPR, San Francisco, CA, 18 Jun. 2010, pp.320–325.

[8] R. Distasi, M. Nappi, and D. Riccio, ―A range/domain approximation error based approach for fractal image compression,‖ IEEE Trans. Image Process., vol. 15,no. 1,pp. 89–97, Jan.2006.

[9] K.SarkarandH.Sundaram,―Howdowe findearlyadopters whowillguidea resource constrained network towards a desired distribution of behaviors?‖ in CoRR, 2013, p. 1303.

[10] D. Imamori and K. Tajima, ―Predicting popularityoftwitter accountsthroughthe discovery of link-propagating early adopters,‖ in CoRR, 2015, p.1512.

[11] X. Rong and Q. Mei, ―Diffusion of innovations revisited: from social network to innovation network,‖ in CIKM, 2013, pp. 499–508.

[12] Mele,F.Bonchi,andA.Gionis,–Theearly-adoptergraphanditsapplicationtoweb- page recommendation,‖in CIKM, 2012, pp. 1682–1686. [13] Y.-F. Chen, –Herd behavior in purchase books online,‖ Computers in Human Behavior, vol. 24(5), pp. 1977– 1992,2008.

[14] Banerjee,–Asimplemodelofherdbehaviour,‖QuarterlyJournalofEconomics,vol. 107, pp. 797– 817,1992.

[15] A. S.E, –Studiesofindependence and conformity: I. a minorityofone against a unanimousmajority,

Psychological monographs:General and applied, vol. 70(9), p. 1, 1956.

[16] T. Mikolov, K. Chen, G. S. Corrado, and J. Dean, –Efficient estimation of word representations in vector space,‖ in ICLR,2013.

[17] A. Bordes, N. Usunier, A. Garc´ıa-Dur´an, J. Weston, and O.Yakhnenko, –Translating embeddings for modeling multirelational data,‖ in NIPS, 2013, pp. 2787–2795.

[18] R.M.Bolle,J.H.Connell,S.Pananti,N.K.Ratha,andA.W.Senior,–Therelation between the ROC curve and the CMC,‖ inProc. 4th IEEE Work. Automat.Identification Adv. Technol., 2005, pp.15–20.

[19] D. Delgado-Gomez, F. Sukno, D. Aguado, C. Santacruz, and A.ArtesRodriguez, –Individual identification using personality traits,‖J. Netw. Comput. Appl., vol.33, no. 3, pp. 293–299, May 2010.

[20] M. D. Marsico, M. Nappi, and D. Riccio, –HERO: Human ear recognition against occlusions,‖ inProc. IEEE Comput. Soc. Workshop Biometrics—In Assoc. IEEE Conf. Comput. Vis. Pattern Recognit.— CVPR, San Francisco, CA, 18 Jun. 2010, pp.320–325.

[21] R. Distasi, M. Nappi, and D. Riccio, –A range/domain approximation error based approach for fractal image compression,‖ IEEE Trans. Image Process., vol. 15,no. 1,pp. 89–97, Jan.2006.

[22] K.SarkarandH.Sundaram,–Howdowe findearlyadopters whowillguidea resource constrained network towards a desired distribution of behaviors?‖ in CoRR, 2013, p. 1303.

[23] D. Imamori and K. Tajima, –Predicting popularityoftwitter accountsthroughthe discovery of link-propagating early adopters,‖ in CoRR, 2015, p.1512.

[24] X. Rong and Q. Mei, –Diffusion of innovations revisited: from social network to innovation network,‖ in CIKM, 2013, pp. 499–508.

[12] Mele,F.Bonchi,andA.Gionis,–Theearly-adoptergraphanditsapplicationtoweb- page recommendation,‖in CIKM, 2012, pp. 1682–1686. [13] Y.-F. Chen, –Herd behavior in purchase books online,‖ Computers in Human Behavior, vol. 24(5), pp. 1977– 1992,2008.

PUBLICATIONS

CONFERENCE

- InternationalConferenceon–Characterizingand PredictingEarlyReviewersfor Effective Product Marketing on E-CommerceWebsites‖(ICICCI-21-0144)
- Paper ID: ICICCI-21-0144

ALL FOUR STUDENTS ONE PAGE PROFILE

1.VEERAMALLISRIJA(17K81A0557)

VEERAMALLI SRIJA is currently pursuing her graduation from St. Martin's Engineering College in thestreamof Computer Science. She completed her intermediate from Sri Chaitanya Junior College and 10th class from Krishnaveni Talent School. She participated in various events, seminars and workshops during her graduation, some of themare:

S.NO	EVENTS/COURSES/SEMINARS
1	Participated in National Level Three Day Online Workshop on "AI & ML in Speech and AudioProcessing".
2	Certification in JavaScript in SoloLearn
3	Participated in Women online workshop on "Women in Cyber Security and Privacy in 2020
4	Certification in MySQL database by Thenewboston in Cursa
5	Certification in Python Core in SoloLearn
6	Certification in basics of AWS Fundamentals Going Cloud-Native in Coursera
7	Certification in Managing Project Risks And Changes in Coursera
8	Participated in National Level Seminar on Recent Trends in Cloud Computing,Fog and Edge Computing
9	Certification in AI For Everyone in Coursera
10	Participated in Anti-Drug Campaign conducted by Lush life Bistro

2.SABA SOWMYA(17K81A0543)

SABA SOWMYA is currently pursuing her graduation from St. Martin's Engineering College in the stream of Computer Science. She completed her intermediate from Sri Chaitanya Junior College and 10thclassfrom Sanghamitra High School. She participated in various events, seminars and workshops during her graduation, some of themare:

S.NO	EVENTS/COURSES/SEMINARS
1	Participated in Employability Skill development Program conducted by Zensar.
2	Participated in National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing".
3	Certification in JavaScript in SoloLearn
4	Certification in Python for Beginners in SoloLearn
5	Participated in Women online workshop on "Women in Cyber Security and Privacy in 2020"
6	Certification in MySQL database by Thenewboston in Cursa
7	Certification in Cyber Security by Packethacks
8	Certification in Managing Project Risks And Changes in Coursera
9	Participated in National Level Seminar on Recent Trends in Cloud Computing,Fog and EdgeComputing
10	Certification in AI For Everyone in Coursera

3. N.SANDEEP(18K85A0501)

N.SANDEEP is currently pursuing his graduation from St.Martin's Engineering College in the stream of Computer Science. Hecompletedhis intermediate from VNR Vignana jyothi institute of technology and 10th class from R S K High School. He participated In various events, seminars and workshops during her graduation, some of themare:

S.NO	EVENTS/SEMINARS/WORKSHOPS
1	Certification in Angular 2 for Beginners by The new boston
2	Certification in Basic of AWS Concepts by ExamPro
3	Certification in Wordpress for beginners by Design Tuts
4	Certification in React JS for Beginners by The new boston
5	Certification in HTML by EJ Media
6	Participated in National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing".
7	Participated in MHRD INNOVATION CELL
8	Certification in Data Science Math Skills

4. M.RAHUL(16K81A0536)

M.RAHUL is currently pursuing his graduation from St.Martin's Engineering College in the stream of Computer Science. He completed his intermediatefrom Urbane Junior College and 10th class from Nrayana High School. He participated In various events, seminars and workshops during her graduation, some of themare:

SNO	EVENTS/SEMINARS/WORKSHOPS
1	Certification in Leadership and Emotional Intelligence
2	Certification in AWS Fundamentals:Going Cloud-Native
3	Certification in Managing Project Risks and Changes
4	Certification in Matrix Algebra for Engineers
5	Certification in AI For Everyone

A

PROJECT REPORT

On

DETECTING AT-RISK STUDENTS WITH EARLY INTERVENTIONS USING MACHINE LEARNING TECHNIQUES

Submitted by

Mr. K. Harshit (17K81A0531)
Ms. S. Sushma (17K81A0551)
Mr. P. Keerthan Srichakra (17K81A0540)
Mr. CM. Vardhan (17K81A0556)

In partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Under the Guidance of

Dr. G. JawaherlalNehru M.E., Ph.D.,

Assistant Professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ST.MARTIN'S ENGINEERING COLLEGE

An Autonomous Institute

Dhulapally, Secunderabad – 500 100

JUNE 2021

BONAFIDE CERTIFICATE

This is to certify that the project entitled "DETECTING AT-RISK STUDENTS WITH EARLY INTERVENTIONS USING MACHINE LEARNING TECHNIQUES" is being submitted by **1. Mr. K. Harshit 17K81A0531, 2. Mr. P. Keerthan SriChakra 17K81A0540, 3. Mr. CM. Vardhan 17K81A0556, 4. Ms. S. Sushma 17K81A0551** in partial fulfillment of the requirement for the award of the degree of **BACHELOR OFTECHNOLOGY INCOMPUTER SCIENCE AND ENGINEERING** is recorded of bonafide work carried out by them. The result embodied in this report have been verified and found satisfactory.

Dr. G. JawaherlalNehru

Department of CSE

Head of the Department

Dr.M.NARAYANAN

Department of CSE

Internal Examiner

External Examiner

Place:

Date:

DECLARATION

We, the student of **Bachelor of Technology** in Department of Computer Science and Engineering', session: 2017 – 2021, St. Martin's Engineering College, Dhulapally, Kompally, Secunderabad, hereby declare that work presented in this Project Work entitled "Detecting At-risk Students with early interventions using Machine Learning Techniques" is the outcome of our own bonafide work and is correct to the best of our knowledge and this work has been undertaken taking care of Engineering Ethics. This result embodied in this project report has not been submitted in any university for award of any degree.

K. Harshit	17K81A0531
P. Keerthan Srichakra	17K81A0540
S. Sushma	17K81A0551
CM. Vardhan	17K81A0556

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible and whose encouragements and guidance have crowded effects with success.

We extended our deep sense of gratitude to Principal**, Dr. P. SANTOSH KUMARPATRA**, St. Martin's Engineering College, Dhulapally, for permitting us to undertake this project.

We are also thankful to **Dr. M.NARAYANAN**, Head of the Department, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project. 3

We would like to thank **Dr. T. POONGOTHAI,** Department of Computer Science and Engineering (AI & ML) for her encouragement and insightful comments and as well as our project coordinators **Dr. B. RAJALINGAM**, Associate Professor and **Mr. J. SUDHAKAR**, Assistant Professor, in Department of Computer Science and Engineering for their valuable support.

We would like to express our sincere gratitude and indebtedness to our project supervisor **Dr. G. JawaherlalNehru**, Assistant Professor, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

Finally, we express thanks to all those who have helped us successfully to completing this project. Furthermore, we would like to thank our family and friends for their moral support and encouragement.

We express thanks to all those who have helped us in successfully completing the project.

K. Harshit 17K81A0531
P. Keerthan 17K81A0540
S. Sushma 17K81A0551
CM. Vardhan 17K81A0556

TABLEOFCONTENTS

ABSTRACT

Massive Open Online Courses (MOOCs) have shown rapid development in recent years, allowing learners to access high-quality digital material. Because of facilitated learning and the flexibility of the teaching environment, the number of participants is rapidly growing. However, extensive research reports that the high attrition rate and low completion rate are major concerns. In this, the early identification of students who are at risk of withdrew and failure is provided. Therefore, we have used 3 models with better accuracy. They are Random forest, Naïve Bayes, Support Vector Machine out of which SVM gave us the precise accuracy.

LISTOFTABLES

LIST OF FIGURES

LIST OF ACRONYMS

SVM	Support Vector Machine
MOOC	Massive open online courses
ICT	Information Communication Technology
LMS	Learning management system

CHAPTER 1: INTRODUCTION

1. INTRODUCTION

1.1 Project Overview:

The use of Information Communication Technology (ICT) has become widespread and plays a vital role in education. ICT has contributed to the support of the academic curriculum and allows for the creation of a virtual classroom. ICT could improve student outcomes and enables instructors to aid students in solving exercises. Therefore, high-quality teaching could be delivered through virtual learning. There are many studies in e-learning field that investigated the ways of applying machine learning techniques for various educational purposes. One of the focuses of these studies is the predicting dropout rates or at-risk students in distance courses by majorly examining log data obtained from learning management systems (LMSs).

The large number of students participating in MOOCs provides the opportunity to perform rich analysis of large-scale online interaction and behavioural data. This analysis can help improve student engagement in MOOCs by identifying patterns, suggesting new feedback mechanisms, and guiding instructor interventions.

Additionally, insights gained by analysing online student engagement can also help validate and refine our understanding of engagement in traditional classrooms. Educational evaluation and measurements have been increasingly enhanced recently. The evaluation of student performance has become a necessary and significant criterion in higher education assessment. Higher education committees consider the quality of higher education from the perspectives of student performance improvement, and these committees give considerable attention to student learning outcomes based on evaluation dimensions.

Massive Open Online Courses (MOOCs) have become an alternative educational platform that allows learners from dispersed geographic locations access the same quality of learning through the web. Coursera, HarvardX, and Khan Academy are some examples of MOOCs. Since 2012, MOOCs modalities have received widespread usage by top Universities. Investigations undertaken by such institutions indicated that the use of MOOCs have attracted many participants towards engagement in the space of courses offered, due to the removal of financial, geographical, and educational barriers. A large volume of data can be collected and captured from MOOCs platforms during Student interaction with learning activities, such as viewing of video lectures, undertaking of quizzes, posting in discussion forums, and interacting with the courseware.

To build an accurate at-risk student prediction model, researchers investigated the reasons behind course withdrawal. This has been attributed to a number of factors. The main reason for students dropping out of online courses is the lack of motivation. Researchers suggested that students' motivational levels in online courses either decrease or increase according to social, cognitive and environmental factors. The motivational trajectory is an important indicator of student dropout. Motivational trajectories can be measured by exploring changes in learner behaviour across courses.

The study conducted by Kotsiantis et al [5] is one of the initial studies which investigated application of machine learning techniques in distance learning for dropout prediction. In this study, time-invariant and time-varying data were included and totally six machine learning techniques was employed, which are Decision Trees, Neural Networks, Naïve Bayes algorithm, Instance-Based Learning Algorithms, Logistic Regression and Support Vector Machines. This study was composed of two experimental stages, training and testing. During these stages, number of attributes was increased step-by-step. For example, while only demographic data was included in the first step, data from the first face-to-face meeting was added in the next step. Six algorithms were tested for each these subsequent steps and then they were compared.

The important conclusion of this study is that Naïve Bayes algorithm is very successful in the prediction of dropouts; it predicts with 83% accuracy.

1.2 Project Objective:

Detecting At-Risk Students with Early Interventions Using Machine Learning Techniques. Detecting at-risk students in a timely manner could help educators deliver instructional interventions and improve the structure of courses.

1.3 Organization of Chapters:

The thesis is organized in the following chapters:

Chapter 1: Introduction
One of the focuses of these studies is the predicting dropout rates or at-risk students in distance courses by majorly examining log data obtained from learning management systems (LMSs). The large number of students participating in MOOCs provides the opportunity to perform rich analysis of large-scale online interaction and behavioural data. This analysis can help improve student engagement in MOOCs by identifying patterns, suggesting new feedback mechanisms, and guiding instructor interventions.

Chapter 2: Literature Survey
Student withdrawal and learning achievements are a major concern in MOOCs. In this section, we provide a review of the state-of-the-art researches in the detection of at-risk students with respect to dropout and failure. The model can also learn specific transitions between quiz assessment date and submission date. The research concluded that high performing students have fewer latent behavioural states since they have sufficient knowledge, and thus, they do not need further support.

Chapter 3: Software and Hardware Requirements
We used Microsoft Windows (also referred to as Windows or Win) which is a graphical operating system developed and published by Microsoft. It provides a way to store files, run software, and connect to the Internet. It is widely available and economical. It helped us for enhancing the working of our project. We used Python programming language as it emphasises code readability and is user friendly, as such it can be used for serving machine learning applications. To write the code we used Jupyter as it is a project and community whose goal is to "develop open-source software, open-standards, and services for interactive computing across dozens of programming languages”.

Chapter 4: Software Development Analysis
The development and implementation of the design parameters. Developer’s code based on the product specifications and requirements agreed upon in the previous stages. Following company procedures and guidelines, front-end developers build interfaces and back-ends while database administrators create relevant data in the database. The programmers also test and review each other's code.

Chapter 5: Project System Design

The System Design is a required document for every project. It should include a high level description of why the System Design Document has been created, provide what the new system is intended for or is intended to replace and contain detailed descriptions of the architecture and system components.

Chapter 6: Project Coding

A programming project produces a well-designed executing system that solves a specified distributed programming problem. A project code is used to represent a one-time, or intermittent departmental event or activity. Any person can use a project code on a transaction, regardless of the project manager or home organization. This section describes some of the coding templates, outline of various files, class with functionalities, the various methods of input and output parameters.

Chapter 7: Project Testing

The purpose of the testing phase is to evaluate and test declared requirements, features, and expectations regarding the project prior to its delivery in order to ensure the project matches initial requirements stated in specification documents.

Chapter 8: Output Screens

The output of the programmed project is being displayed in the form of screenshots. The data from Excel file has been taken and necessary operations were performed to get the final input. The results have been captured and projected.

Chapter 9: Experimental Results

The results obtained helps us to compare which algorithm works better with good accuracy so as to overcome the hurdles faced in existing systems. In the end, Support Vector Machine gave us the good results and accuracy.

CHAPTER 2:
LITERATURE SURVEY

2. LITERATURE SURVEY

2.1 Survey on background:

Student withdrawal and learning achievements are a major concern in MOOCs. In this section, we provide a review of the state-of-the-art researches in the detection of at-risk students with respect to dropout and failure. Feedforward neural networks were implemented in to detect at-risk students in MOOCs, using student sentiments and clickstream as baseline features. The data was collected from 3 million student click logs in addition to 5,000 forum posts via the Coursera platform in 2014.

Dealing with an imbalanced dataset was one of the main concerns in this study. This was overcome by employing Cohen's Kappa criteria instead of accuracy. The results demonstrated an accuracy of 74%, when both sets of features were employed. This reduced to 70%, when sentiment features were excluded.

In at-risk students were identified by applying various machine learning algorithms, including regularized logistic regression, support vector machines, random forest, decision tree and Naïve Bayes. A set of features were captured from behavioural log data, including the number of times students visited the home page and the length of the session. The results illustrated that regularized logistic regression models achieved the highest AUC.

The ConRec Network model, a type of deep neural network, was proposed in. In this work, Convolutional Neural Networks (CNN) were combined with Recurrent Neural Networks (RNN) to predict whether students are at risk of withdrawal from the online course ``XuetangX" in the next ten days. Student records were structured according to a sequence of time-stamps and contained various attributes such as event time, event type and student enrolment date. The hybrid neural network model consists of two parts, namely, the lower and upper parts. In the lower part, the hidden layer of CNN was utilized to extract features automatically. In the upper part, RNN was used to make a prediction by aggregating and combining the extracted features at each time.

The model was compared with various baseline methods. The results indicated similar performance across all models. The F1-score results were reported in the range of 90.74-92.48. Although there was similarity in performance, the authors argued that the ConRec Network model is more efficient than baseline methods, as it has the ability to extract the features automatically from student records without the need of feature engineering.

2.2 Conclusion on Survey

A number of features have been considered by researchers to identify the level of student learner achievement in the online setting, such as how long students interact with digital resources when students submitted assessments and the total number of attempts undertaken, educational level, geographical location and gender. In, Genetic Algorithms (GA) were used to optimize the feature set. The findings indicated that high ranked features are related to behavioural attributes instead of demographic features. Four classifiers were considered to predict student performance, namely decision tree, neural network, Naïve Bayes and k-nearest neighbour. Simulation results indicated that accuracy was improved by 12% when using the GA-optimized feature set. Using the decision tree with the complete feature set led to an accuracy of 83.87%, while when the GA-optimized feature set was used, accuracy jumped to 94.09%. Hidden Markov models were used to measure how latent variables in conjunction with observed variables could impact student performance in virtual learning environments. A two-layer hidden Markov model (TL-HMM) was proposed in to infer latent student behavioural patterns. TL-HMM differs from conventional HMM in its capacity to discover the micro-behavioural patterns of students in more detail and detect transition between latent states. For instance, when students undertake quizzes, they would tend to participate in forum discussions. The model can also learn specific transitions between quiz assessment date and submission date. The research concluded that high performing students have fewer latent behavioural states since they have sufficient knowledge, and thus, they do not need further support.

CHAPTER 3:

SOFTWARE AND HARDWARE REQUIREMENTS

3. SOFTWARE AND HARDWARE REQUIREMENTS

To be used efficiently, all computer software needs certain hardware components or other software resources to be present on a computer. These prerequisites are known as (computer) system requirements and are often used as a guideline as opposed to an absolute rule. Most software defines two sets of system requirements: minimum and recommended. With increasing demand for higher processing power and resources in newer versions of software, system requirements tend to increase over time.

Industry analysts suggest that this trend plays a bigger part in driving upgrades to existing computer systems than technological advancements. A second meaning of the term of system requirements, is a generalisation of this first definition, giving the requirements to be met in the design of a system or sub-system.

3.1 Software Requirements:

Operating System: Windows

Programming Language: Python

IDE: Jupyter Notebook

3.2 Hardware Requirements:

Processor: i3 or Above

RAM: 2GB

Hard Disk: 10GB

CHAPTER 4:

SOFTWARE DEVELOPMENT ANALYSIS

4. SOFTWARE DEVELOPMENT ANALYSIS

The software development process involves the creation and maintenance of applications, frameworks and other components for software design, design, programming, documentation, testing and problem remediation. The development of software is a process of creating and keeping source code, but it encompasses everything from the idea of the intended software to the last manifestation of the programme, often in a planned and organised process in a larger context. Software development may therefore encompass research, creation of new software products, prototype, modification, reuse, reengineering, maintenance, or any other software-production activity.

A life-cycle "model" is sometimes considered a more general term for a category of methodologies and a software development "process" a more specific term to refer to a specific process chosen by a specific organization.[citation needed] For example, there are many specific software development processes that fit the spiral lifecycle model. For example, there are many specific software development processes that fit the spiral life-cycle model. The field is often considered a subset of the systems development life cycle.

4.1 Overview of the Problem:

Massive Open Online Courses (MOOCs) have shown rapid development in recent years, allowing learners to access high-quality digital material. Because of facilitated learning and the flexibility of the teaching environment, the number of participants is rapidly growing. However, extensive research reports that the high attrition rate and low completion rate are major concerns.

Predicting student retention in MOOCs can provide valuable information to help educators to early recognise at-risk students. Although a number of works were reported in the literature proposing robust learning frameworks for online courses, it is still challenging to achieve high prediction accuracy of student performance in the long term over multiple datasets.

4.2 Define the Problem:

To build an accurate at-risk student prediction model, researchers investigated the reasons behind course withdrawal. This has been attributed to a number of factors. The main reason for students dropping out of online courses is the lack of motivation. Researchers suggested that students' motivational levels in online courses either decrease or increase according to social, cognitive and environmental factors. The motivational trajectory is an important indicator of student dropout. Motivational trajectories can be measured by exploring changes in learner behaviour across courses. Until now, most researchers did not pay attention in examining the association between motivational trajectories, student learning achievement and at-risk students in the online setting.

4.3 Module Overview:

The aim of the project entitled as Detecting at-risk students with early interventions using Machine learning techniques is to develop a procedure where we can predict the risk of students so as to apply appropriate interventions. Doing which the instructor can guide the students.

4.4 Defining the Modules:

There are two modules in our project:

A. User:

In this the user obtains the dataset from a particular university. This dataset includes all the features which directly effects the motivational/sentimental status of the students. Using this it allows the instructors to have idea about how effective the courses are.

B. Application:

In this the dataset is read as an input and all the pre-processing and PCA are performed so as to tune the dataset(explore the important features in the dataset). Along with it the selected algorithms are applied so as to detect the accuracy in order to obtain the best results.

4.5 Module functionality:

A. Load the Dataset:

Using this module we will load dataset from internet.

Two datasets are utilised in our experiments. The rst set is obtained from Harvard University and Massachusetts Institute of Technology online courses, while the second set is related to Open University online courses. Harvard University collaborated with Massachusetts Institute of Technology (MIT) in developing online courses. The primary attribute of the Harvard dataset is the clickstream, which represents the number of events that correspond to user interaction with courseware. Qualifying events include clicking on a chapter or on forum posts and accessing the home page of videos. The user must register on each course before the actual enrolment date. To complete the registration process, the user must click on five web pages. The ``Nchapters" feature is the number of chapters that learners are required to read. ``Nplay_video" represents the number of events during which the learner viewed a particular video. The ``Explored" feature is a binary discretisation of exploratory learners. To be classied as an explorer, a learner must have accessed more than half of the course contents. The ``Viewed" feature is also a binary feature, which is set to 1 when a student accessed the home page of assignments an related videos. The date of learner registration for a specific course is recorded in the dataset in addition to the date of the learners' last interaction with the courseware. The ``LoE_DI" feature is a demographic feature, which represents the learners educational level. ``age" and ``gender" are other types of demographic features, which are also recorded. The assignment grade is an indicator attribute that represents the failure/success rate of participants.

B. Data Pre-processing:

Data Pre-Processing is applied to the extracted behavioural features and demographic variables of the dataset, with the aim to achieve the best performance. The first step in pre-processing the data is to investigate highly correlated variables. We set a correlation cut off value of 0.8, i.e., if the correlation between two features is greater than 0.8, then these features are considered highly correlated. Highly correlated features are removed from the model, given that the problem of feature redundancy could be solved. Moreover, the occurrence of overfitting may also be reduced. The zero and near-zero variance predictors are also investigated in this database; the features with the same values that appear frequently become zero variance predictors when the data is split into training and test. These features, which have a ``near-zero-variance'' are diagnosed and eliminated during the pre-processing procedure.

C. Exploratory Data Analysis (EDA):

Exploratory Data Analysis (EDA) is implemented in this study in order to gain insight into the learners' motivational trajectories in conjunction with their dropout rate. EDA is an important step in machine learning, providing intuition about the structure and relationships within the dataset. With regards to the first case study, the objective of data visualization is to provide information and understanding of the type of motivational status at the first-time interval, which is more relevant to at-risk students.

D. Support Vector Machine:

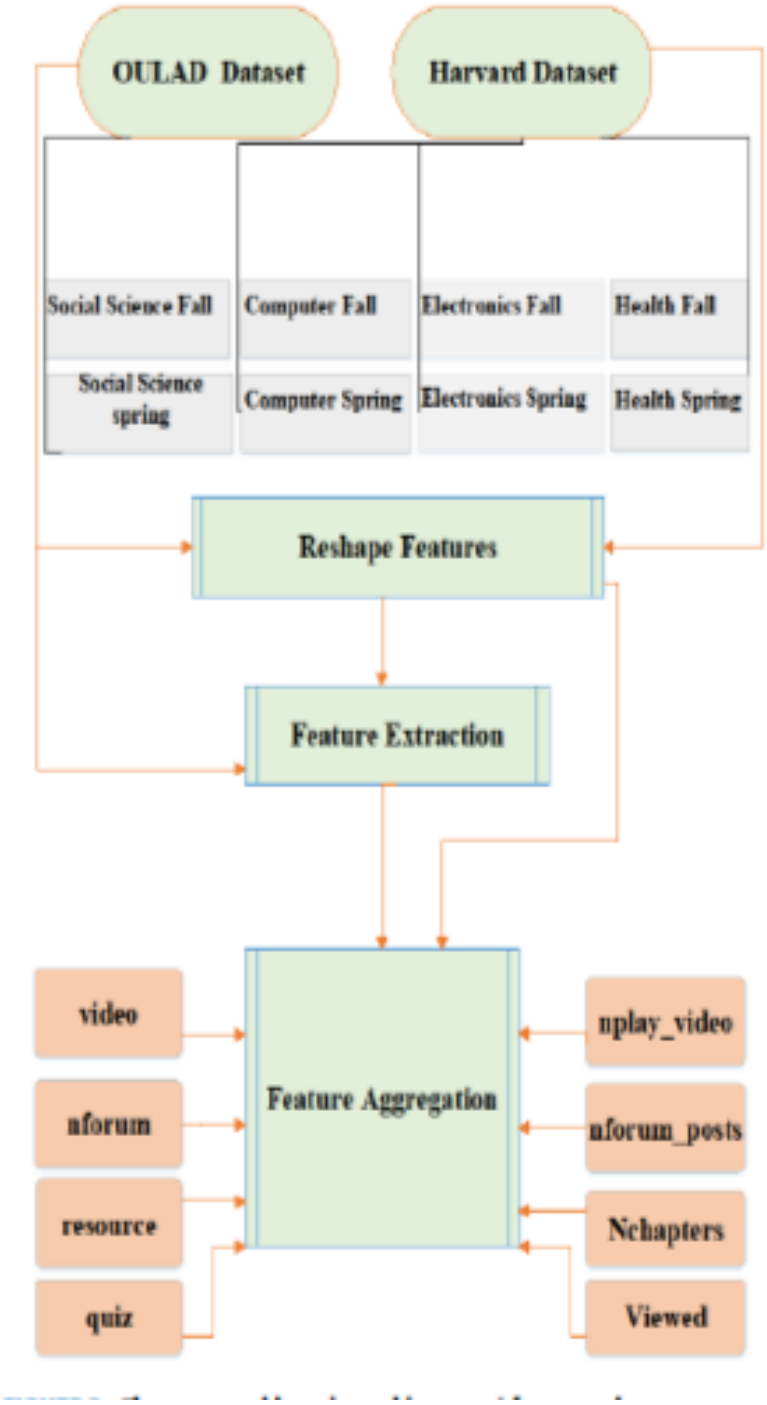
In machine learning, Support Vector Machines (SVM) also called as support vector networks are supervised learning models with associated learning algorithms that analyse data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible.

CHAPTER 5: PROJECT SYSTEM DESIGN

5. PROJECT SYSTEM DESIGN

5.1 Data-Flow diagram:

1. The proposed learning achievement framework.



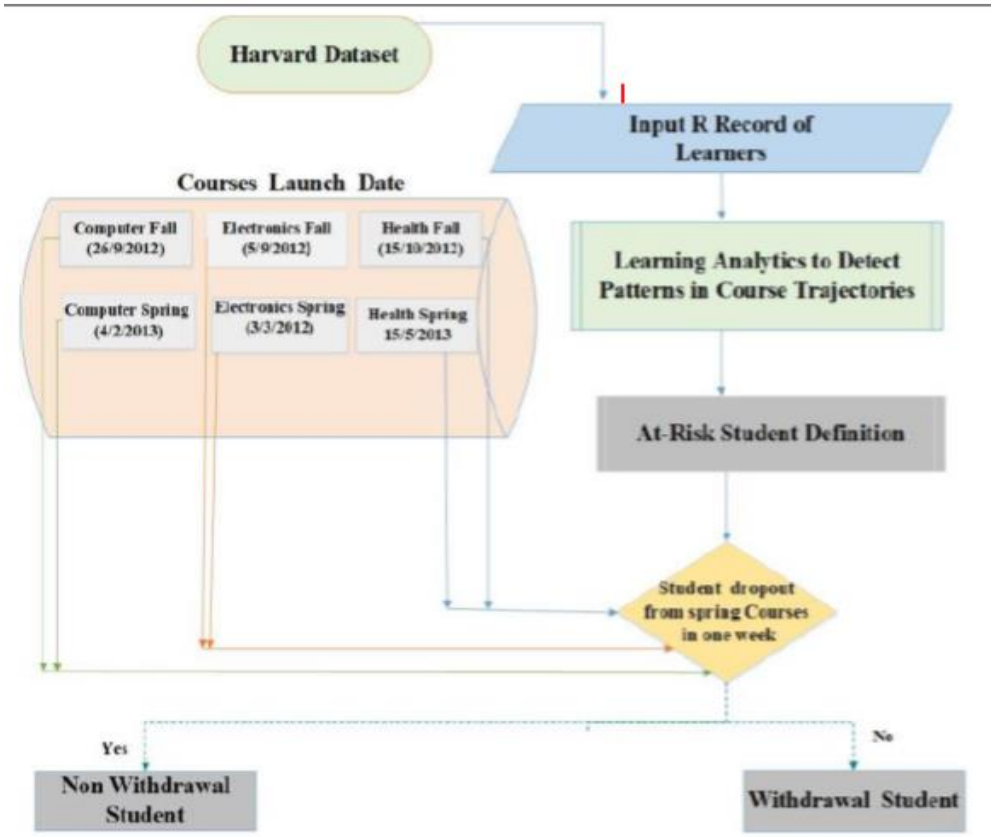
Learning achievement is considered a vital indicator of the effectiveness of the MOOCs platform [22]. A student performance predictive model is proposed to predict whether students will pass or fail in online courses. The framework aims to measure poor student performance and investigate the impact of learning activities that influence student decisions to complete a future course. This will assist instructors in drawing inferences about student performance and will offer deeper insights into the learning process. Additionally, it could support instructors in racking student progress for each tier of learning. Hence, effective teaching can be delivered.

The key challenge in building a The Harvard dataset does not provide a granular record structure for student activity over time. Instead, summary values are provided, which incorporate totals, with the intermediate structure discarded. On the other hand, daily learning activities are collected in the OULAD dataset.

Clickstreams information is employed to acquire a common set of attributes across the two datasets. Specifically, the daily VLE activities are used to construct summative behavioural features across the OULAD dataset. Only four activities are considered, i.e., “nforum”, “resource”, “quiz” and “videos”. Next, the extracted features are aggregated with OULAD behavioural features these are “nfroum_posts”, “Nchapters”, “Viewed” and “nplay_vedio” Thus, similar behavioural attributes can be extracted from the two datasets.

With regards to temporal features, the number of days that learners interact with the OULAD online courses is extracted by computing the difference between the dates of student registration and deregistration from MOOCs. The same feature extraction process is performed in the Harvard dataset. Due to the weak association between learning outcomes and demographic features, demographic characteristics are excluded in this analysis**.**

2. At-risk student framework



Learning Analytics (LA) tools were utilized.

Since students in the OULAD courses are required to participate in assessments, intrinsically motivated and amotivatied students cannot be evaluated for this dataset [21].Therefore, the at-risk student detection framework is only considered with the Harvard dataset, as the aim is to assess how motivation trajectories could impact at-risk students.

Learning trajectories can facilitate online course analysis by tracing student activities over time. In this study, LA is utilized in the tracking of learning trajectories across multiple courses. Figure 1 illustrates the at-risk student framework.

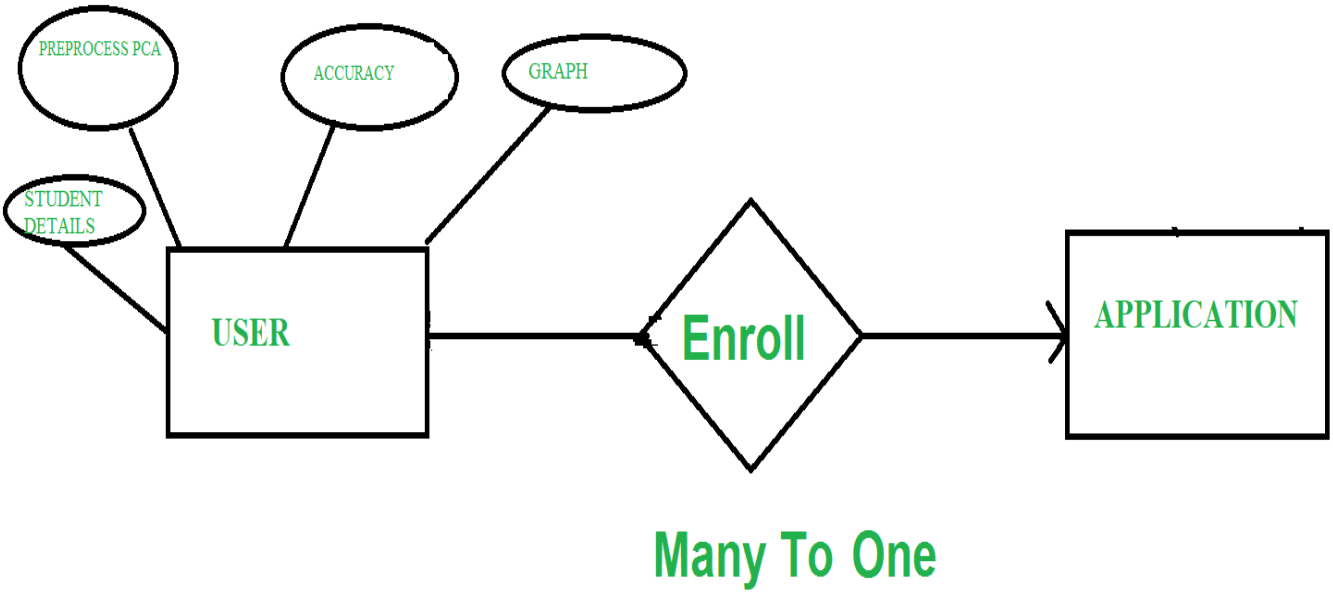
It has been suggested that low student performance and learning achievement outcomes are important factors for students withdrawal from online courses. However, in the current case study, students are defined as at risk if they withdraw from spring courses within the period of one week. This is because it is not possible to perform a reliable evaluation of student learning in such a short period.

Although intrinsically motivated students can attain learning outcomes within one week, in the Harvard dataset, it is not possible to measure student performance for such students,

since relevant information, e.g., student feedback is not captured [23]. A data-driven approach should be considered when investigating the most critical factors which impact on student learning outcomes. To examine how such factors influence students who are at risk of failure, a student learning achievement model is proposed.

5.2 ER diagram:

ER Diagram stands for Entity Relationship Diagram, also known as ERD is a diagram that displays the relationship of entity sets stored in a database. In other words, ER diagrams help to explain the logical structure of databases. ER diagrams are created based on three basic concepts: entities, attributes and relationships.ER Diagrams contain different symbols that use rectangles to represent entities, ovals to define attributes and diamond shapes to represent relationships.



\

5.3 UML Diagrams:

Unified Modelling Language (UML) is a general purpose modelling language. The main aim of UML is to define a standard way to visualize the way a system has been designed. It is quite similar to blueprints used in other fields of engineering. UML is not a programming language, it is rather a visual language.

We use UML diagrams to portray the behaviour and structure of a system. UML helps software engineers, businessmen and system architects with modelling, design and analysis. The Object Management Group (OMG) adopted Unified Modelling Language as a standard in 1997. Its been managed by OMG ever since. International Organization for Standardization (ISO) published UML as an approved standard in 2005. UML has been revised over the years and is reviewed periodically.

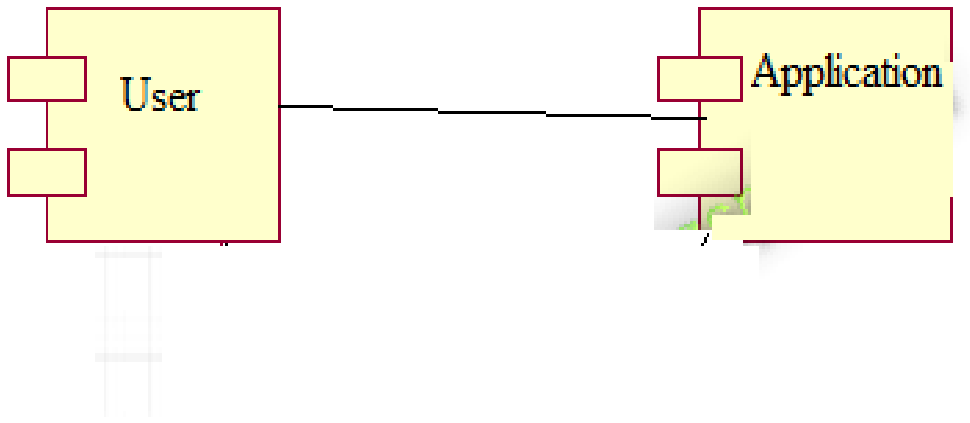
UML is linked with object oriented design and analysis. UML makes the use of elements and forms associations between them to form diagrams. Diagrams in UML can be broadly classified as:

Structural Diagrams – Capture static aspects or structure of a system. Structural Diagrams include: Component Diagrams, Object Diagrams, Class Diagrams and Deployment Diagrams.

Behaviour Diagrams – Capture dynamic aspects or behaviour of the system. Behaviour diagrams include: Use Case Diagrams, State Diagrams, Activity Diagrams and Interaction Diagrams.

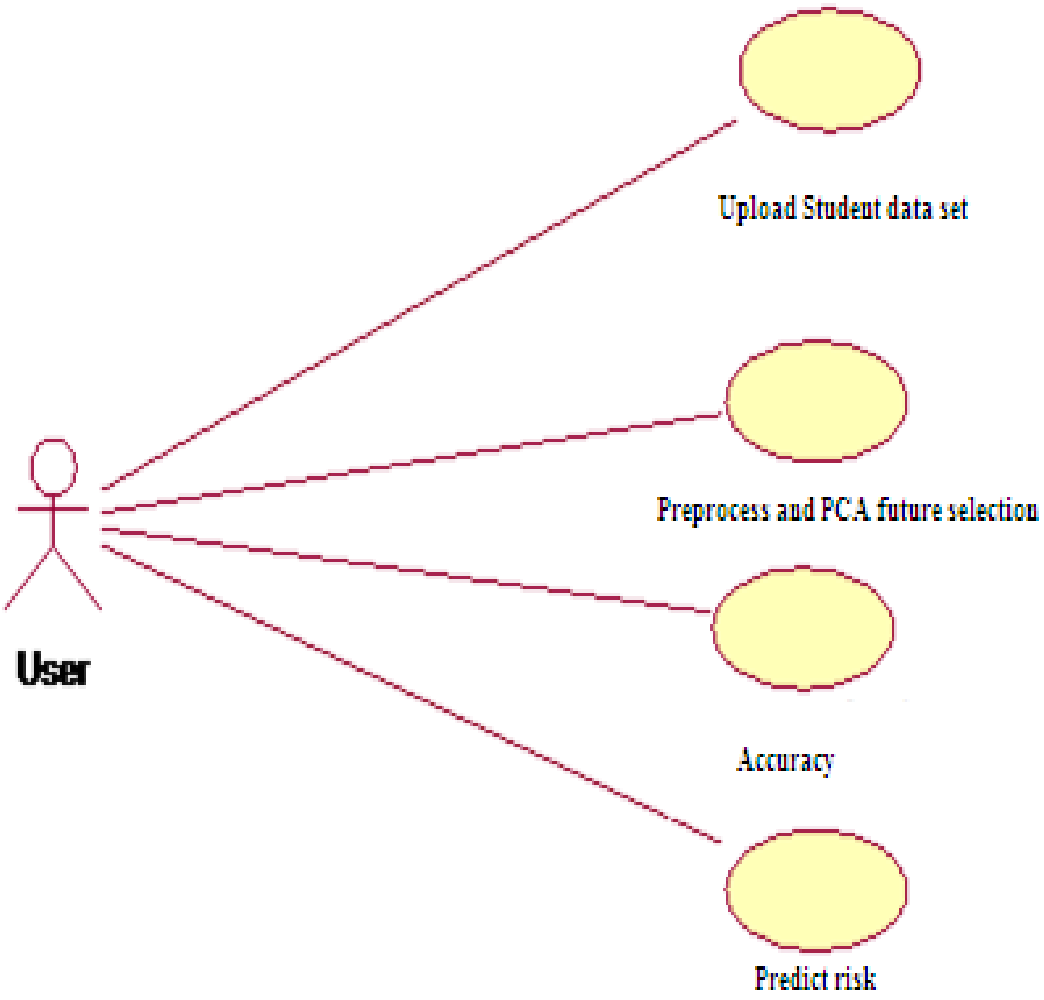
A) Component:

Component diagrams are used to represent the how the physical components in a system have been organized. We use them for modelling implementation details. Component Diagrams depict the structural relationship between software system elements and help us in understanding if functional requirements have been covered by planned development. Component Diagrams become essential to use when we design and build complex systems. Interfaces are used by components of the system to communicate with each other.



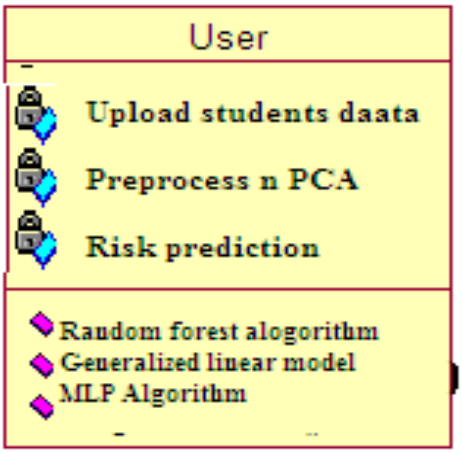
B) Use-case:

Use Case Diagrams are used to depict the functionality of a system or a part of a system. They are widely used to illustrate the functional requirements of the system and its interaction with external agents(actors). A use case is basically a diagram representing different scenarios where the system can be used. A use case diagram gives us a high level view of what the system or a part of the system does without going into implementation details.



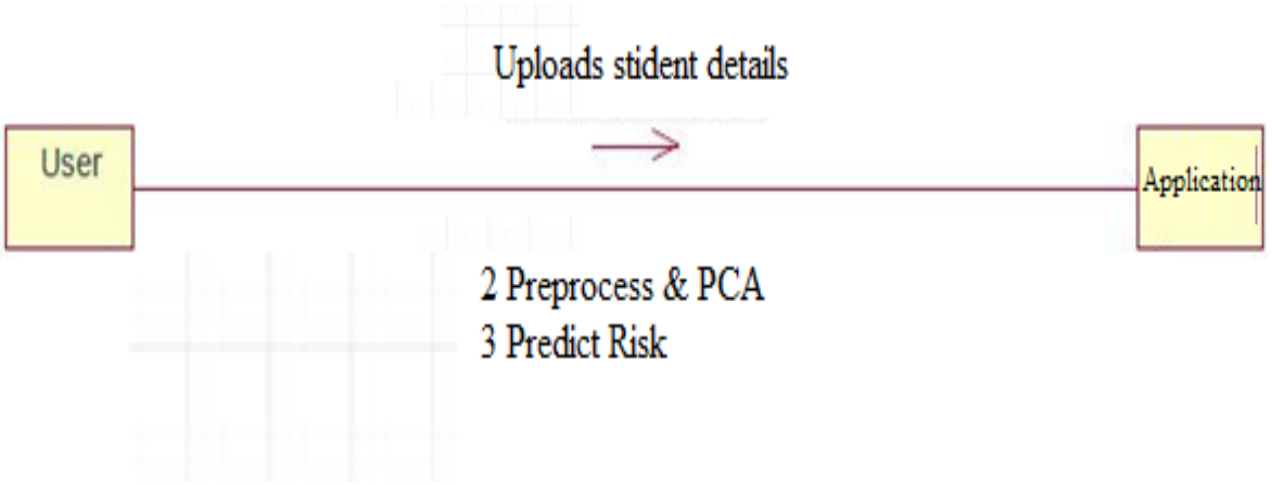
C) Class:

The most widely use UML diagram is the class diagram. It is the building block of all object oriented software systems. We use class diagrams to depict the static structure of a system by showing system's classes, their methods and attributes. Class diagrams also help us identify relationship between different classes or objects.



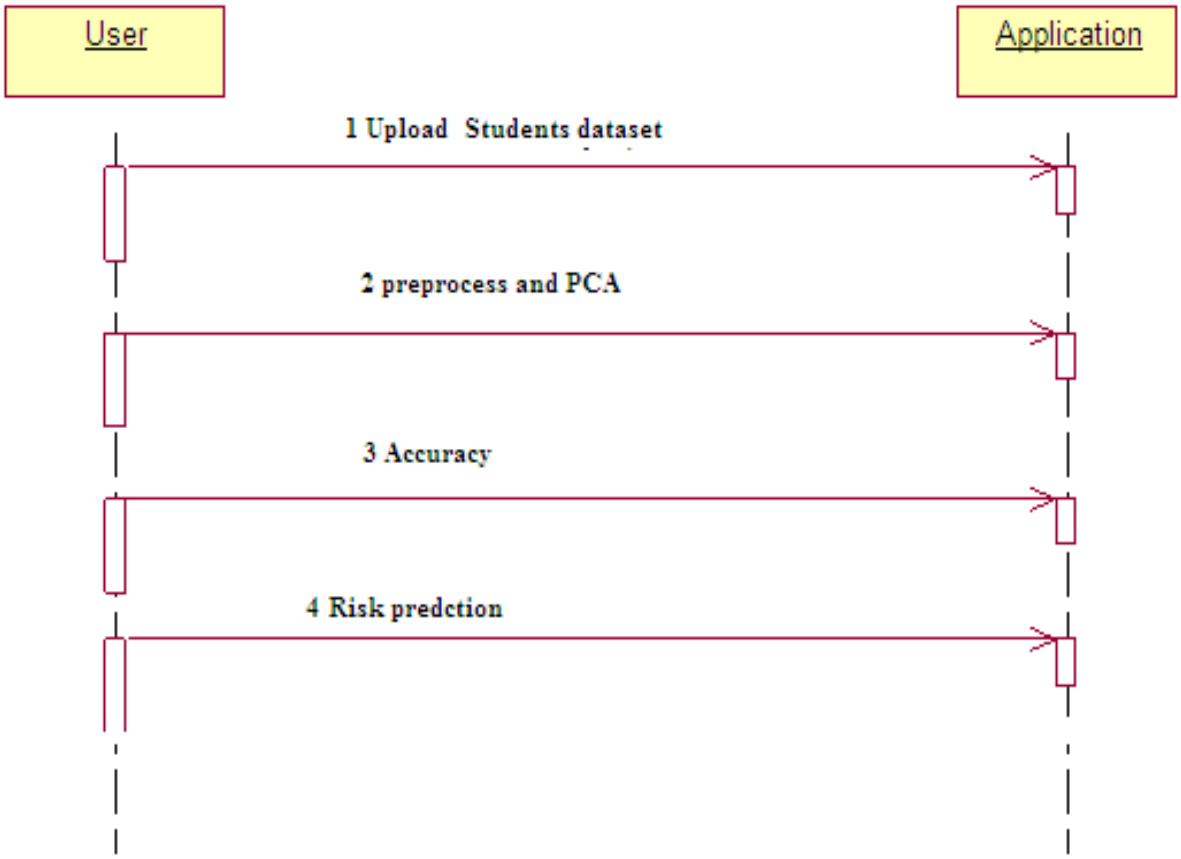
D) Collaboration:

A collaboration diagram, also known as a communication diagram, is an illustration of the relationships and interactions among software objects in the Unified Modelling Language (UML). These diagrams can be used to portray the dynamic behaviour of a particular use case and define the role of each object.



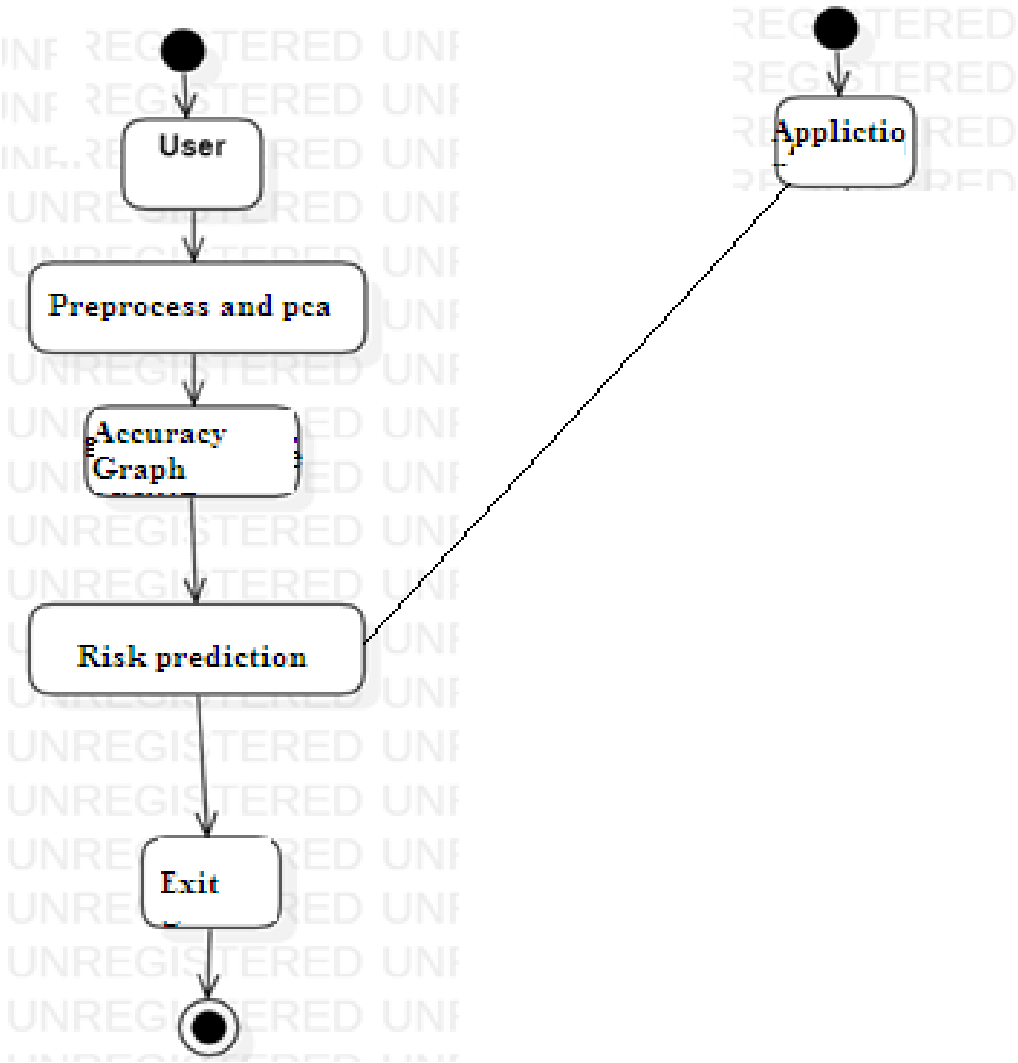
E) Sequence:

Sequence diagram simply depicts interaction between objects in a sequential order i.e. the order in which these interactions take place. We can also use the terms event diagrams or event scenarios to refer to a sequence diagram. Sequence diagrams describe how and in what order the objects in a system function. These diagrams are widely used by businessmen and software developers to document and understand requirements for new and existing systems.



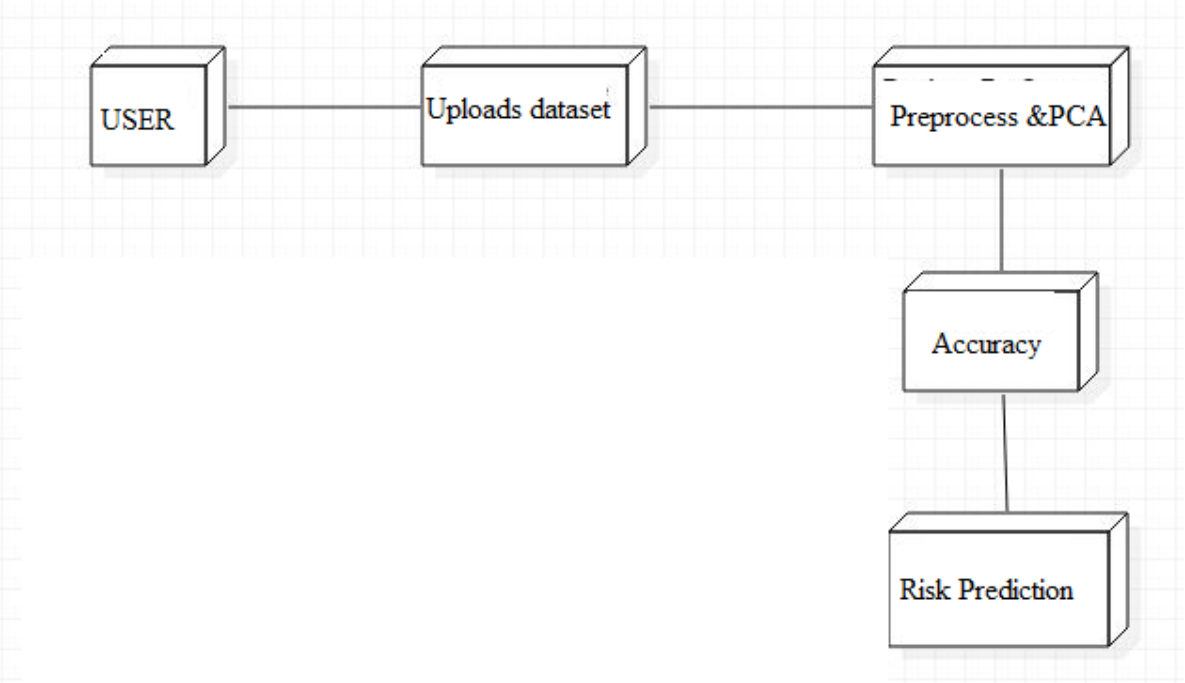
F) Activity:

We use Activity Diagrams to illustrate the flow of control in a system. We can also use an activity diagram to refer to the steps involved in the execution of a use case. We model sequential and concurrent activities using activity diagrams. So, we basically depict workflows visually using an activity diagram. An activity diagram focuses on condition of flow and the sequence in which it happens. We describe or depict what causes a particular event using an activity diagram.



G) Deployment:

Deployment Diagrams are used to represent system hardware and its software. It tells us what hardware components exist and what software components run on them. We illustrate system architecture as distribution of software artifacts over distributed targets. An artifact is the information that is generated by system software. They are primarily used when a software is being used, distributed or deployed over multiple machines with different configurations.



CHAPTER 6: PROJECT CODING

6. PROJECT CODING

6.1 Coding Template:

A) Setting the libraries

```
import numpy as np
import pandas as pd
import csv
import time from sklearn
import cross_validation from sklearn.cross_validation
 import cross_val_score from sklearn.cross_validation
import train_test_split, KFold from sklearn.cross_validation
import StratifiedShuffleSplit from sklearn.naive_bayes
 import GaussianNB from sklearn.ensemble
import RandomForestClassifier from sklearn.svm
import SVC from sklearn.grid_search
import GridSearchCV from sklearn.metrics
import f1_score
# First, decide how many training vs test samples you want
from sklearn.cross_validation import train_test_split
import pylab as pl
import matplotlib.pyplot as pl
 from sklearn.preprocessing import scale
```

B) Explore the data

```
student_data = pd.read_csv("student-data.csv")
n_students = student_data.shape[0]
n_features = student_data.shape[1]
n_passed = student_data[student_data['passed'] =='yes'].shape[0]
 n_failed = student_data[student_data['passed'] == 'no'].shape[0]
grad_rate = (n_passed*1.0) / (n_students*1.0) * 100
 print "Total number of students: {}".format(n_students)
```

```
print "Number of students who passed: {}".format(n_passed)
print "Number of students who failed: {}".format(n_failed)
 print "Number of features: {}".format(n_features)
print "Graduation rate of the class: {:.2f}%".format(float(grad_rate))
```

C) Preparing the data

```
# Extract feature (X) and target (y)
columns feature_cols = list(student_data.columns[:-1])
# all columns but last are features
target_col = student_data.columns[-1]
# last column is the target/label
print "Feature column(s):\n{}".format(feature_cols)
print "Target column: {}".format(target_col)
X_all = student_data[feature_cols]
 # feature values for all students
y_all = student_data[target_col]
 # corresponding targets/labels
 print "\nFeature values:"
print X_all.head() # print the first 5 rows
```

D) Preprocess feature column

```
def preprocess_features(X):
outX = pd.DataFrame(index=X.index)
# output dataframe, initially empty
# Check each column
for col, col_data in X.iteritems():
# If data type is non-numeric, try to replace all yes/no values with 1/0
if col_data.dtype == object:
      col_data = col_data.replace(['yes', 'no'], [1, 0])
 # Note: This should change the data type for yes/no columns to int
```

```
# If still non-numeric, convert to one or more dummy variables
 if col_data.dtype == object:
# Splits Columns Up if non-numeric into 1 or 0;
 col_data = pd.get_dummies(col_data, prefix=col)
outX = outX.join(col_data)
 # collect column(s) in output dataframe
return outX
 X_all = preprocess_features(X_all)
 print "Processed feature columns ({}):-\n{}".format(len(X_all.columns), list(X_all.columns))
 #Format Target yes/no values with 1/0
 y = pd.DataFrame(y_all)
 y = y.replace(['yes', 'no'], [1, 0])
#in the form (X, 1), but the method expects a 1d array and has to be in the form (X, )
 y_all = np.ravel(y)
#join dataset student_data = pd.concat([X_all, y], axis = 1)
print(student_data)
```

E) Splitting data into testing and training set

```
# First, decide how many training vs test samples you want
num_all = student_data.shape[0]
# same as len(student_data)
num_train = 300
 # about 75% of the data
num_test = num_all - num_train y = student_data['passed']
#print y
def Stratified_Shuffle_Split(X,y,num_train):
 sss = StratifiedShuffleSplit(y, 3, train_size=num_train, random_state = 0)
 for train_index, test_index in sss:
 X_train, X_test = X.iloc[train_index], X.iloc[test_index]
y_train, y_test = y.iloc[train_index], y.iloc[test_index]
```

return X_train, X_test, y_train, y_test X_train, X_test, y_train, y_test = Stratified_Shuffle_Split(X_all, y, num_train)

print "Training Set: {0:.2f} Samples".format(X_train.shape[0])

print "Testing Set: {0:.2f} Samples".format(X_test.shape[0])

F) Training and evaluating models

Predict on Training Set and Compute F1 Score

#Train Model

def train_classifier(clf, X_train, y_train):

print "Training {}:".format(clf.__class__.__name__)

start = time.time() clf.fit(X_train, y_train)

end = time.time()

train_clf_time = end - start

print "Training Time (secs): {:.3f}".format(train_clf_time)

return train_clf_time

Predict on Training Set and Compute F1 Score

def predict_labels(clf, features, target):

print "Predicting labels using {}:".format(clf.__class__.__name__)

start = time.time()

y_pred = clf.predict(features)

end = time.time()

prediction_time = end - start

print "Prediction Time (secs): {:.3f}".format(prediction_time)

return (f1_score(target.values, y_pred, pos_label='yes'), prediction_time)

G) Choosing the best model

def reformat(col_data):

return col_data.replace(['yes', 'no'], [1, 0])

def iterate_fit_predict(number_runs):

f1_scores = []

```
gamma = []

C = []

 # Get the features and labels from the Boston housing data

y = reformat(student_data['passed'])

for num in range(0,number_runs):

X_train, X_test, y_train, y_test = Stratified_Shuffle_Split(X_all, y, 300)

 clf_SVC = SVC()

parameters = [{'C':[1,10,50,100,200,300,400,500,1000,], 'gamma':[1e-4, 1e-3, 1e-2, 1e-1, 1e0, 1e1],
'kernel': ['rbf']}]

 clf = GridSearchCV(clf_SVC, parameters, scoring = 'f1')

 # Fit the learner to the training data to obtain the best parameter set

clf.fit(X_train, y_train)

f1_scores.append(clf.score(X_test, y_test))

gamma.append(clf.best_params_['gamma'])

C.append(clf.best_params_['C'])

clf = clf.best_estimator_

#print clf

df_f1 = pd.Series(f1_scores)

 df_gamma = pd.Series(gamma)

df_C = pd.Series(C)

 print clf print "\nF1 Scores:"

print df_f1

 print "\nAverage F1 Test Scores:"

 print df_f1.mean()

print "\nC:" print df_C

print "\nGamma:"

 print df_gamma
```

H) Models final F1 score

SVC(C=200, cache_size=200, class_weight=None, coef0=0.0, decision_function_shape=None, degree=3, gamma=0.0001, kernel='rbf', max_iter=-1, probability=False, random_state=None, shrinking=True, tol=0.001, verbose=False)

6.2 Outline for various files:

This project includes two files:

A) **Code file (student_intervention-v2):** In this it contains the code
B) **Dataset file (Harvard.excel):** It includes the students dataset that we are taking as the input

6.3 Class with functionalities:

1. **Reading the dataset:** In this method we read the dataset and extract it into the code successfully.

2. **Exploring the dataset:** In this method we explore all the features that are included which directly effects the sentimental status of the students.

3. **Pre-processing the data:** In this method we tune the dataset by eliminating redundancy so as to have purity in the dataset. This thereby taken as input in an application. The rst step in pre-processing is cleaning the data by detecting the occurrence of missing values. Several variables in the Harvard dataset have null values; examples of these include ``Nevent", ``nplay_video", ``Nchapters", ``nforum_post", ``YOB", ``Gender" and ``LoE_DI" attributes. The data is cleaned by removing missing values and others. In addition, student records with duplicated rows are also removed. The Harvard dataset is non-normally distributed. In order to address this problem, transformation methods were applied. The BOX_COST transformation [25] was used to transform the data distribution into normal. As seen in Table 5, the Box Cox method transformed ten features with skewed distributions. The scaling and centring transforms were also applied , and results show that all features are centred to a mean value of 0 and scaled to a standard deviation of 1. Data Pre-Processing is applied to the extracted behavioural features and demographic variables of the OULAD dataset, with the aim to achieve the best performance. The rst step in pre-processing the data is to investigate highly correlated variables. We set a correlation cut off value of 0.8, i.e., if the correlation between two features is greater than 0.8, then these features are considered highly correlated. Highly correlated features are removed from the model, given that the problem of feature redundancy could be solved. Moreover, the occurrence of over-tting may also be reduced. The zero and near-zero variance predictors are also investigated in this database; the features with the same values that appear frequently become zero variance predictors when the data is split into training and test. These features, which have a ``near-zero-variance" are diagnosed and eliminated during the pre-processing procedure. The Open University dataset is non-normally distributed; in order to address this problem, transformation methods are applied. Yeo-Johnson [26] is one of the data transformations methods and performs a similar function to the Box-Cox transformation, in which a continuous variable that has a raw value equal to zero is applied [26]. In our case, when a student did not participate in a particular activity, the value of the extracted features become zero. To this end, Yeo-Johnson is more useful than Box-Cox.

4. **Training and Testing:** In this we take the data samples so as to test the algorithms accuracy. This is done so as to come up with the better algorithm to give accurate precision.

6.4 Methods INPUT and OUTPUT parameters:

INPUT:

The dataset containing the details of students whose risk level is to be determined is taken as input. And from this dataset we explore the features that directly effects the students motivational/sentimental status. Later we pre-process the dataset so as to reduce the size by eliminating unnecessary records. Since the dataset is ready to test or train the model, we finally give this dataset to the application that we proposed.

OUTPUT:

After reading the input we finally run it through the algorithms so as to measure the scores in terms of accuracy. Thereby allowing us to chose which algorithm is best fit for this model. By selecting it makes the instructors to determine if their students require any interventions or not.

CHAPTER 7:
PROJECT TESTING

7. PROJECT TESTING

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

7.1 Various Test Cases:

A) Unit Testing

The Unit testing is the testing at the code level and helps eliminate issues at an early stage, mainly the developer is responsible to perform the unit test for his code while not all the defects cannot be discovered at the unit testing.

B) Functional Testing

Functional testing is associated with the low-level design phase which ensures that collections of codes and units are working together probably to execute new function or service.

C) Integration Testing

Integration testing is associated with the high-level design phase. Integration testing ensures the integration between all system modules after adding any new functions or updates.

D) System Testing

System testing is associated with the system requirements and design phase. It combines the software, hardware, and the integration of this system with the other external systems.

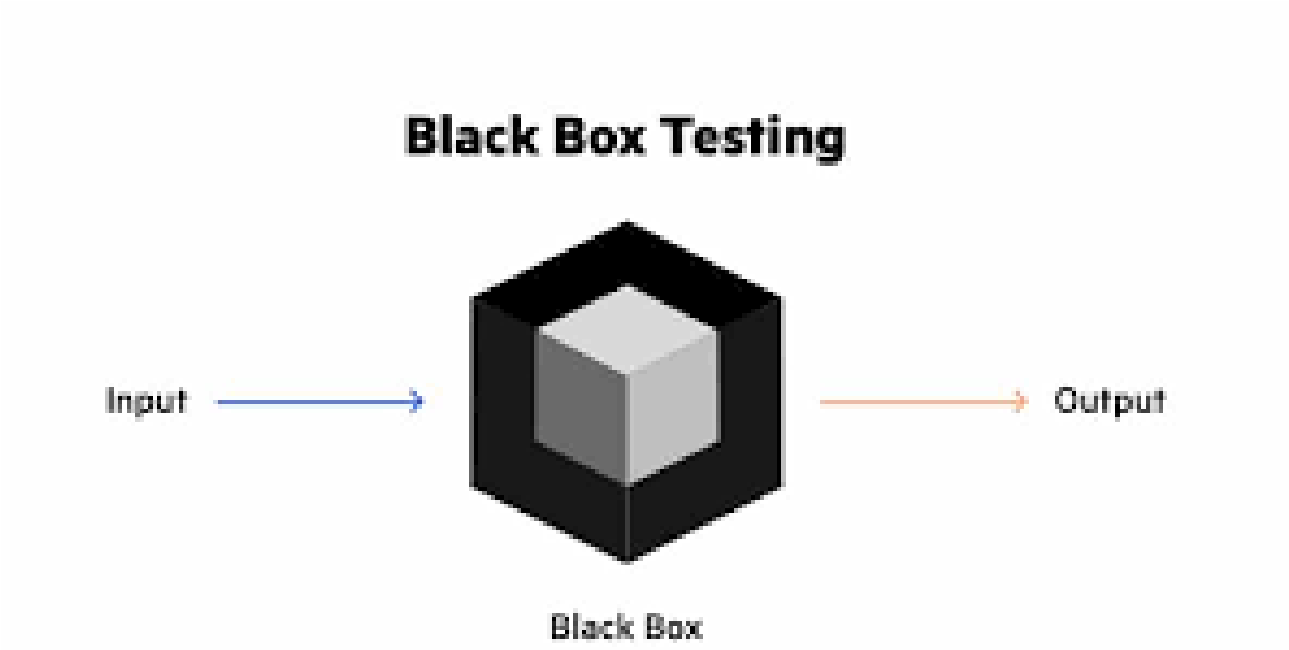
E) User Acceptance Testing

User Acceptance testing is associated with the business and operations analysis phase. The customer users are the main performers of this testing based on test cases and scenarios that cover the business requirements to ensure that they have delivered the right software as per the specifications.

7.2 Blackbox Testing:

Black Box Testing is a software testing method in which the functionalities of software applications are tested without having knowledge of internal code structure, implementation details and internal paths. Black Box Testing mainly focuses on input and output of software applications and it is entirely based on software requirements and specifications. It is also known as Behavioural Testing.

The above Black-Box can be any software system you want to test. Under Black Box Testing, we can test these applications by just focusing on the inputs and outputs without knowing their internal code implementation.



How to do BlackBox Testing?
Here are the generic steps followed to carry out any type of Black Box Testing.

- Initially, the requirements and specifications of the system are examined.
- Tester chooses valid inputs (positive test scenario) to check whether SUT processes them correctly. Also, some invalid inputs (negative test scenario) are chosen to verify that the SUT is able to detect them.
- Tester determines expected outputs for all those inputs.
- Software tester constructs test cases with the selected inputs.
- The test cases are executed.
- Software tester compares the actual outputs with the expected outputs.
- Defects if any are fixed and re-tested.

Types of Black Box Testing
There are many types of Black Box Testing but the following are the prominent ones -

Functional testing – Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals. Functional testing is centered on the following items:

Valid Input	: identified classes of valid input must be accepted.
Invalid Input	: identified classes of invalid input must be rejected.
Functions	: identified functions must be exercised.
Output	: identified classes of application outputs must be exercised.
Systems/Procedures	: interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

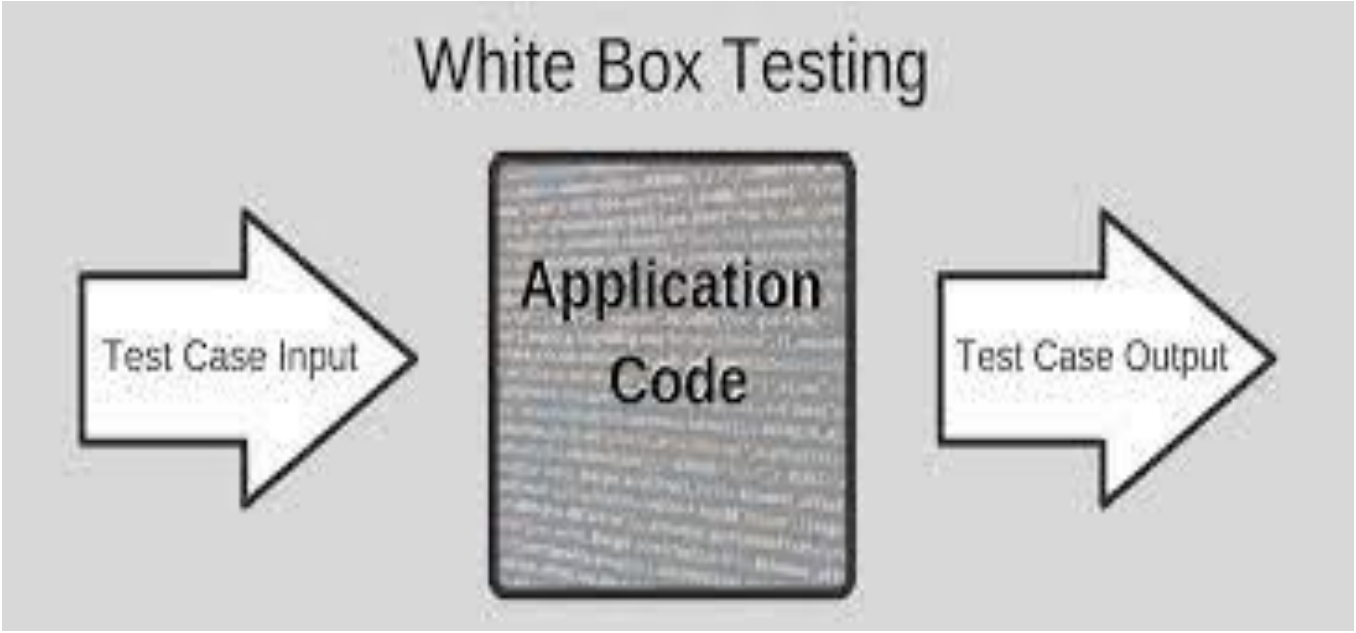
Non-functional testing - This type of black box testing is not related to testing of specific functionality, but non-functional requirements such as performance, scalability, usability. Non-functional testing is the testing of a software application or system for its non-functional requirements: the way a system operates, rather than specific behaviours of that system.
This is in contrast to functional testing, which tests against functional requirements that describe the functions of a system and its components. The names of many non-functional tests are often used interchangeably because of the overlap in scope between various non-functional requirements. For example, software performance is a broad term that includes many specific requirements like reliability and scalability.
Characteristics of Non-functional testing:
Non-functional testing should be measurable, so there is no place for subjective characterization like good, better, best, etc.
Exact numbers are unlikely to be known at the start of the requirement process
Important to prioritize the requirements
Ensure that quality attributes are identified correctly in Software Engineering.

7.3 Whitebox Testing:

White Box Testing is software testing technique in which internal structure, design and coding of software are tested to verify flow of input-output and to improve design, usability and security. In white box testing, code is visible to testers, so it is also called Clear box testing, Open box testing, Transparent box testing, Code-based testing and Glass box testing.

It is one of two parts of the Box Testing approach to software testing. Its counterpart, Blackbox testing, involves testing from an external or end-user type perspective. On the other hand, White box testing in software engineering is based on the inner workings of an application and revolves around internal testing.

The term "White Box" was used because of the see-through box concept. The clear box or White Box name symbolizes the ability to see through the software's outer shell (or "box") into its inner workings. Likewise, the "black box" in "Black Box Testing" symbolizes not being able to see the inner workings of the software so that only the individual experience can be tested.



Testing for Memory Leaks:

Memory leaks are leading causes of slower running applications. A QA specialist who is experienced at detecting memory leaks is essential in cases where you have a slow running software application.

Memory leaks can be catastrophic to an application, resulting in hangs, buffering, or crashes. In the worst-case scenario, they are reported by a customer.

While developers may routinely run their own memory tests on the code they are actively developing, memory testing later in the development cycle is often reduced, or overlooked entirely in the rush to deliver a new release.

Advantages:

Continuous Integration systems enable constant feedback of the impact of changes on a codebase. Using the automated capabilities of a CI system to conduct routine memory-use analysis delivers the same advantages as it does for regular functional testing:

- New code changes are immediately tested, making them easier to fix.
- Any new illegal uses of memory and memory leaks are detected and reported through the CI system.
- The CI system’s test reports provide transparency to a problem that is often undetected until it is too late.
- Generation of memory result files allows for follow-on detailed analysis of the application's use of memory.

CHAPTER 8:
OUTPUT SCREENS

8. OUTPUT SCREEN

1. Reading data:

To load necessary Python libraries and load the student data.

```
# Import Libraries
from __future__ import division
import numpy as np
import pandas as pd
from time import time
from sklearn.metrics import f1_score


# Read student data
student_data = pd.read_csv("Harvard.csv")
print("Student data read successfully!")
```

```
Student data read successfully!
```

2. Exploring the data:

To determine how many students we have information on, and learn about the graduation rate among these students.

```
Total number of students: 395
Number of features: 30
Number of students who passed: 265
Number of students who failed: 130
Graduation rate of the class: 67.09%
```

3. Preparing the data:

In this section, we will prepare the data for modelling, training and testing.

- Identify feature and target columns:

It is often the case that the data you obtain contains non-numeric features.
This can be a problem, as most machine learning algorithms expect numeric data to perform computations with.
Run we separate the student data into feature and target columns to see if any features are non-numeric.

```
Feature columns:
['school', 'sex', 'age', 'address', 'famsize', 'Pstatus', 'Medu', 'Fedu', 'Mjob', 'Fjob', 'reason', 'guardian', 'traveltime',
'studytime', 'failures', 'schoolsup', 'famsup', 'paid', 'activities', 'nursery', 'higher', 'internet', 'romantic', 'famrel', 'f
reetime', 'goout', 'Dalc', 'Walc', 'health', 'absences']

Target column: passed

Feature values:
  school sex  age address famsize Pstatus  Medu  Fedu     Mjob      Fjob  ... \
0     GP   F   18       U     GT3       A     4     4  at_home   teacher  ...
1     GP   F   17       U     GT3       T     1     1  at_home     other  ...
2     GP   F   15       U     LE3       T     1     1  at_home     other  ...
3     GP   F   15       U     GT3       T     4     2   health  services  ...
4     GP   F   16       U     GT3       T     3     3    other     other  ...

  higher internet  romantic  famrel  freetime goout Dalc Walc health absences
0    yes       no        no       4         3     4    1    1      3        6
1    yes      yes        no       5         3     3    1    1      3        4
2    yes      yes        no       4         3     2    2    3      3       10
3    yes      yes       yes       3         2     2    1    1      5        2
4    yes       no        no       4         3     2    1    2      5        4

[5 rows x 30 columns]
```

4. Pre-process feature columns:

As you can see, there are several non-numeric columns that need to be converted! Many of them are simply yes/no, e.g. internet. These can be reasonably converted into 1/0 (binary) values.

Other columns, like Mjob and Fjob, have more than two values, and are known as *categorical variables*. The recommended way to handle such a column is to create as many columns as possible values (e.g. Fjob_teacher, Fjob_other, Fjob_services, etc.), and assign a 1 to one of them and 0 to all others.

These generated columns are sometimes called *dummy variables*, and we will use the pandas.get_dummies() function to perform this transformation.

```
Processed feature columns (48 total features):
['school_GP', 'school_MS', 'sex_F', 'sex_M', 'age', 'address_R', 'address_U', 'famsize_GT3', 'famsize_LE3', 'Pstatus_A', 'Pstat
us_T', 'Medu', 'Fedu', 'Mjob_at_home', 'Mjob_health', 'Mjob_other', 'Mjob_services', 'Mjob_teacher', 'Fjob_at_home', 'Fjob_heal
th', 'Fjob_other', 'Fjob_services', 'Fjob_teacher', 'reason_course', 'reason_home', 'reason_other', 'reason_reputation', 'guard
ian_father', 'guardian_mother', 'guardian_other', 'traveltime', 'studytime', 'failures', 'schoolsup', 'famsup', 'paid', 'activi
ties', 'nursery', 'higher', 'internet', 'romantic', 'famrel', 'freetime', 'goout', 'Dalc', 'Walc', 'health', 'absences']
```

5. Training and Test split:

So far, we have converted all *categorical* features into numeric values. For the next step, we split the data (both features and corresponding labels) into training and test sets. In the following code cell below, you will need to implement the following:

- Randomly shuffle and split the data (X_all, y_all) into training and testing subsets.
  - Use 300 training points (approximately 75%) and 95 testing points (approximately 25%).
  - Set a random_state for the function(s) you use, if provided.
  - Store the results in X_train, X_test, y_train, and y_test.

In [16]:
```
# TODO: Import any additional functionality you may need here
from sklearn.model_selection import train_test_split
# TODO: Set the number of training points
num_train = 300

# Set the number of testing points
num_test = X_all.shape[0] - num_train

# TODO: Shuffle and split the dataset into the number of training and testing points above
X_train, X_test, y_train, y_test = train_test_split(X_all,y_all,train_size=num_train,random_state=53453)
#random_state=53453
# Show the results of the split
print ("Training set has {} samples.".format(X_train.shape[0]))
print ("Testing set has {} samples.".format(X_test.shape[0]))
```

```
Training set has 300 samples.
Testing set has 95 samples.
```

6. Model performance metrics:

To initialize three helper functions which you can use for training and testing the three supervised learning models you've chosen above. The functions are as follows:

- train_classifier - takes as input a classifier and training data and fits the classifier to the data.
- predict_labels - takes as input a fit classifier, features, and a target labeling and makes predictions using the F_1 score.
- train_predict - takes as input a classifier, and the training and testing data, and performs train_clasifier and predict_labels.
    - This function will report the F_1 score for both the training and testing data separately.

With the predefined functions above, you will now import the three supervised learning models of your choice and run the train_predict function for each one. Remember that you will need to train and predict on each classifier for three different training set sizes: 100, 200, and 300. Hence, you should expect to have 9 different outputs below — 3 for each model using the varying training set sizes. In the following code cell, you will need to implement the following:

- Import the three supervised learning models you've discussed in the previous section.
- Initialize the three models and store them in clf_A, clf_B, and clf_C.
    - Use a random_state for each model you use, if provided.
    - **Note:** Use the default settings for each model — you will tune one specific model in a later section.
- Create the different training set sizes to be used to train each model.
    - *Do not reshuffle and resplit the data! The new training points should be drawn from X_train and y_train.*
- Fit each model with each training set size and make predictions on the test set (9 in total). **Note:** Three tables are provided after the following code cell which can be used to store your results.

```
RandomForestClassifier:

Training a RandomForestClassifier using a training set size of 100. . .
Trained model in 0.0269 seconds
Made predictions in 0.0050 seconds.
F1 score for training set: 1.0000.
Made predictions in 0.0040 seconds.
F1 score for test set: 0.6822.
Training a RandomForestClassifier using a training set size of 200. . .
Trained model in 0.0130 seconds
Made predictions in 0.0020 seconds.
F1 score for training set: 0.9895.
Made predictions in 0.0020 seconds.
F1 score for test set: 0.7353.
Training a RandomForestClassifier using a training set size of 300. . .
Trained model in 0.0160 seconds
Made predictions in 0.0030 seconds.
F1 score for training set: 0.9951.
Made predictions in 0.0020 seconds.
F1 score for test set: 0.7188.

SVC:

Training a SVC using a training set size of 100. . .
Trained model in 0.0030 seconds
Made predictions in 0.0010 seconds.
F1 score for training set: 0.9324.
Made predictions in 0.0020 seconds.
F1 score for test set: 0.7432.
Training a SVC using a training set size of 200. . .
Trained model in 0.0070 seconds
Made predictions in 0.0070 seconds.
F1 score for training set: 0.8875.
Made predictions in 0.0020 seconds.
F1 score for test set: 0.7785.
Training a SVC using a training set size of 300. . .
Trained model in 0.0150 seconds
Made predictions in 0.0060 seconds.
F1 score for training set: 0.8779.
Made predictions in 0.0030 seconds.
F1 score for test set: 0.7671.
```

```
GaussianNB:

Training a GaussianNB using a training set size of 100. . .
Trained model in 0.0020 seconds
Made predictions in 0.0020 seconds.
F1 score for training set: 0.8382.
Made predictions in 0.0010 seconds.
F1 score for test set: 0.6140.
Training a GaussianNB using a training set size of 200. . .
Trained model in 0.0020 seconds
C:\Users\Harshith\Anaconda3\lib\site-packages\sklearn\ensemble\forest.py:246: FutureWarning: The default value of n_estimators
will change from 10 in version 0.20 to 100 in 0.22.
  "10 in version 0.20 to 100 in 0.22.", FutureWarning)
C:\Users\Harshith\Anaconda3\lib\site-packages\sklearn\svm\base.py:196: FutureWarning: The default value of gamma will change fr
om 'auto' to 'scale' in version 0.22 to account better for unscaled features. Set gamma explicitly to 'auto' or 'scale' to avoi
d this warning.
  "avoid this warning.", FutureWarning)
C:\Users\Harshith\Anaconda3\lib\site-packages\sklearn\svm\base.py:196: FutureWarning: The default value of gamma will change fr
om 'auto' to 'scale' in version 0.22 to account better for unscaled features. Set gamma explicitly to 'auto' or 'scale' to avoi
d this warning.
  "avoid this warning.", FutureWarning)
C:\Users\Harshith\Anaconda3\lib\site-packages\sklearn\svm\base.py:196: FutureWarning: The default value of gamma will change fr
om 'auto' to 'scale' in version 0.22 to account better for unscaled features. Set gamma explicitly to 'auto' or 'scale' to avoi
d this warning.
  "avoid this warning.", FutureWarning)

Made predictions in 0.0030 seconds.
F1 score for training set: 0.8175.
Made predictions in 0.0030 seconds.
F1 score for test set: 0.6720.
Training a GaussianNB using a training set size of 300. . .
Trained model in 0.0030 seconds
Made predictions in 0.0040 seconds.
F1 score for training set: 0.8116.
Made predictions in 0.0020 seconds.
F1 score for test set: 0.7481.
```

7. Model tuning:

Fine tune the chosen model. Use grid search (GridSearchCV) with at least one important parameter tuned with at least 3 different values. You will need to use the entire training set for this. In the code cell below, you will need to implement the following:

- Import sklearn.grid_search.gridSearchCV and sklearn.metrics.make_scorer.
- Create a dictionary of parameters you wish to tune for the chosen model.
  - Example: parameters = {'parameter' : [list of values]}.
- Initialize the classifier you've chosen and store it in clf.
- Create the F_1 scoring function using make_scorer and store it in f1_scorer.
  - Set the pos_label parameter to the correct value!
- Perform grid search on the classifier clf using f1_scorer as the scoring method, and store it in grid_obj.
- Fit the grid search object to the training data (X_train, y_train), and store it in grid_obj.

```
Made predictions in 0.0050 seconds.
Tuned model has a training F1 score of 0.8344.
Made predictions in 0.0030 seconds.
Tuned model has a testing F1 score of 0.7919.
```

8. Tabular results:

Choosing the best model by comparing the F1 scores of the algorithms run on different sizes of training sets.

** Classifer 1 - Random Forest Classifier**

Training Set Size	Training Time	Prediction Time (test)	F1 Score (train)	F1 Score (test)
100	0.0908	0.0013	1.0000	0.6822
200	0.0340	0.0012	0.9895	0.7353
300	0.0280	0.0011	0.9951	0.7188

** Classifer 2 - Support Vector Machine**

Training Set Size	Training Time	Prediction Time (test)	F1 Score (train)	F1 Score (test)
100	0.0018	0.0012	0.9324	0.7432
200	0.0032	0.0012	0.8875	0.7785
300	0.0062	0.0016	0.8779	0.7671

** Classifer 3 - Naive Bayes**

Training Set Size	Training Time	Prediction Time (test)	F1 Score (train)	F1 Score (test)
100	0.0033	0.0004	0.8382	0.6140
200	0.0023	0.0004	0.8175	0.6720
300	0.0009	0.0004	0.8116	0.7481

9. Model final F1 score:

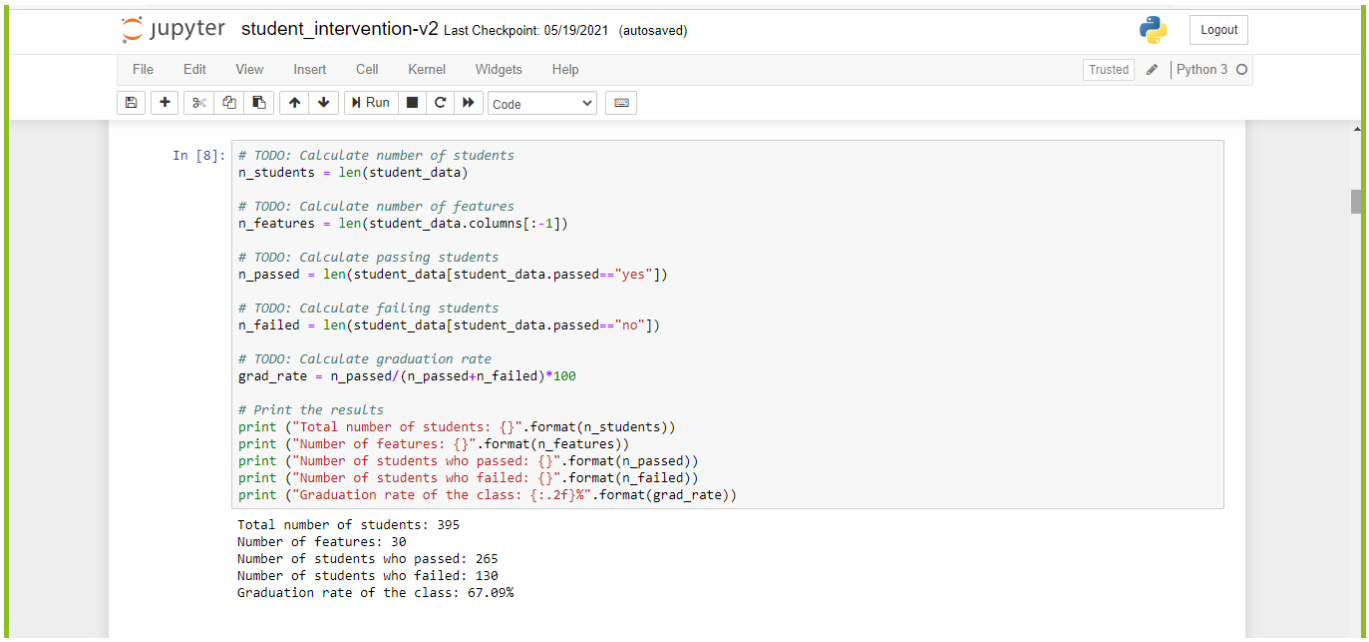
Multiple iterations are desired to be performed in order to provide a conclusive F1 score (performance of model). After running 10 iterations, the average F1 score of the tuned SVC model is: 0.813.

```
SVC(C=0.027825594022071126, cache_size=200, class_weight=None, coef0=0.0,
  decision_function_shape='ovr', degree=3, gamma=1, kernel='linear',
  max_iter=-1, probability=False, random_state=765, shrinking=True,
  tol=0.001, verbose=False)
```

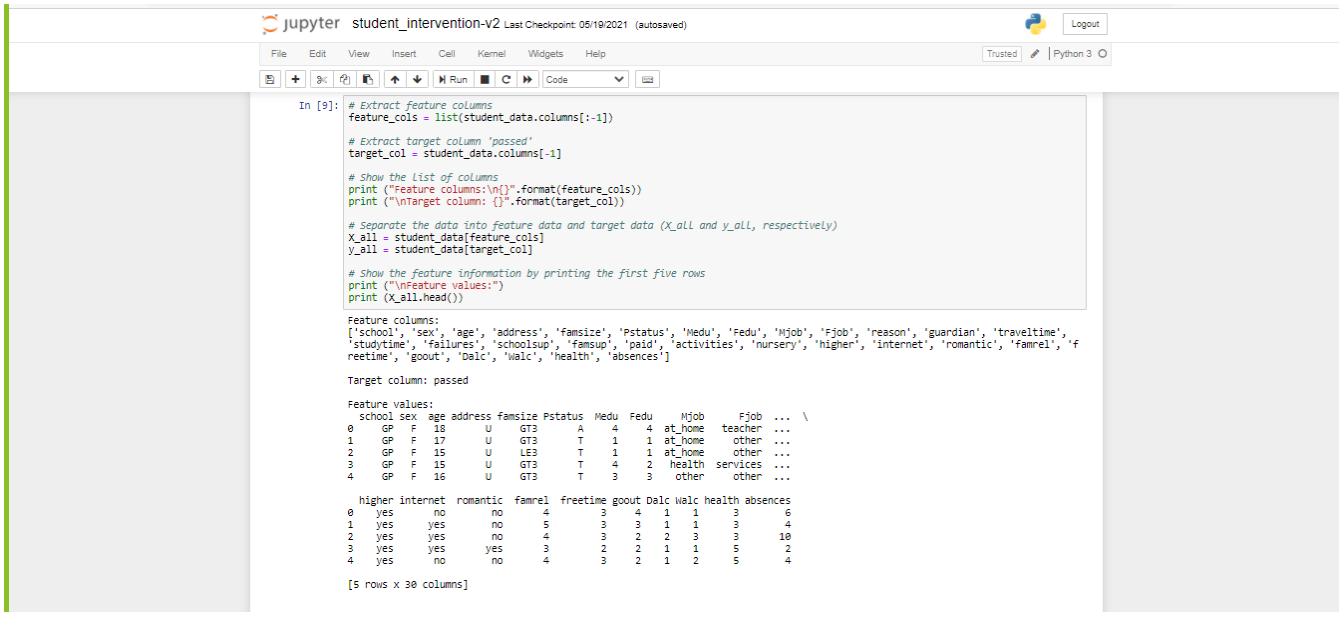
CHAPTER 9: EXPERIMENTAL RESULTS

9. EXPERIMENTAL RESULTS

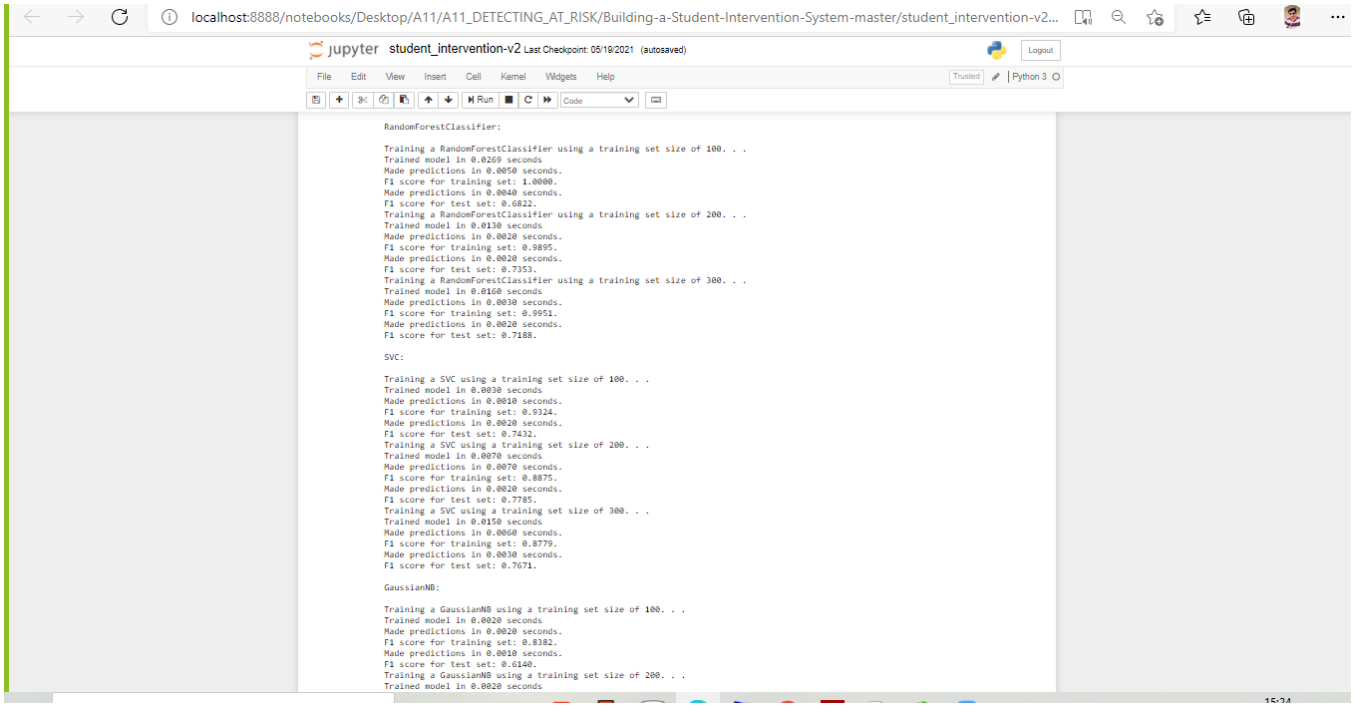
Exploring features:



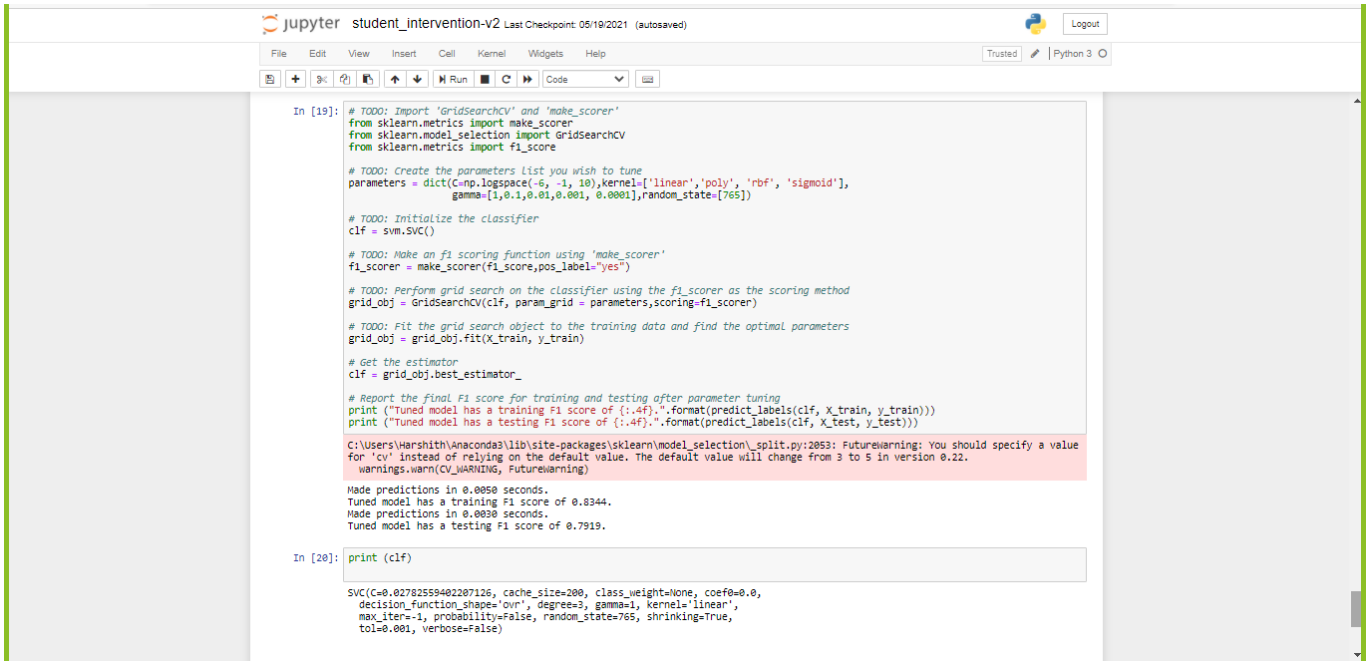
Feature Extraction:



Testing the algorithms and selecting the best:



Final:



CHAPTER 10: CONCLUSIONS AND FUTURE ENHANCEMENTS

10. CONCLUSIONS AND FUTURE ENHANCEMENT

Two case studies were conducted in this work, with the aim of offering decision-makers the opportunity for early intervention and provision of timely assistance to students who are at risk of withdrawal and failure. In the first case study, the relationship between engagement level and motivational status with withdrawal rates was examined. In the second case study, a learning achievement model was proposed to identify at-risk students and analyse the factors affecting student failure.

The dropout prediction model can facilitate educators in delivering early intervention support for at-risk students. The findings show that student motivation trajectories are the main reason for student withdrawal in online courses. Feature selection enhances the predictive capacity of machine learning models while reducing the associated computational costs. Furthermore, the filter method for feature selection is a promising solution for tackling the overfitting problem. The results of this study could assist educators in monitoring changes in student motivational status, thus enabling them to identify those students who require additional support.

In regards to future research, we intend to consider the validation of the proposed framework with additional datasets. It will be interesting to capture online datasets from different providers, delivering courses on the same topics, to evaluate subject trends. Machine learning can also be used to automatically predict students who are in danger of dropout from courses. Machine learning can extract features from student records by inferring the sequences of temporal events across various MOOCs datasets. As such, SVM can be used to track student behaviour and motivational status and discover the impact of these characteristics on at-risk students.

CHAPTER 11: REFERENCES

11. REFERENCES

1. S. Nawar and A. M. Mouazen, ``Comparison between random forests, artificial neural networks and gradient boosted machines methods of online vis-NIR spectroscopy measurements of soil total nitrogen and total carbon," Sensors, vol. 17, no. 10, p. 2428, 2017.
2. J. Sinclair and S. Kalvala, ``Student engagement in massive open online courses," Int. J. Learn. Technol., vol. 11, no. 3, pp. 218237,2016.
3. H. B. Shapiro, C. H. Lee, N. E.W. Roth, K. Li, M. Çetinkaya-Rundel, and D. A. Canelas, ``Understanding the massive open online course (MOOC) student experience: An examination of attitudes, motivations, and barriers," Comput. Educ., vol. 110, pp. 3550, Jul. 2017.
4. J.-L. Hung, M. C. Wang, S. Wang, M. Abdelrasoul, Y. Li, and W. He, ``Identifying at-risk students for early interventions A time-series clustering approach," IEEE Trans. Emerg. Topics Comput., vol. 5, no. 1, pp. 4555, Jan./Mar. 2017.
5. S. Kotsiantis, C. Pierrakeas, and P. Pintelas, "Preventing student dropout in distance learning systems using machine learning techniques," AI Techniques in Web-Based Educational Systems at Seventh International Conference on Knowledge-Based Intelligent Information & Engineering Systems, pp. 3-5, September 2003.
6. A. Acharya and D. Sinha, "Early prediction of students' performance using machine learning techniques," *International Journal of Computer Applications*, vol. 107, no. 1, pp. 37–43, 2014.
7. M. Koutina and K. L. Kermanidis, "Predicting postgraduate students' performance using machine learning techniques," in *Proceedings of the International Conference on Artificial Intelligence Applications and Innovations*, pp. 159–168, Springer, Corfu, Greece, September 2011.
8. E. B. Belachew and F. A. Gobena, "Student performance prediction model using machine learning approach: the case of Wolkite university," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 7, no. 2, pp. 46–50, 2017.
9. S. Kotsiantis, C. Pierrakeas, and P. E. Pintelas, "Predicting students' performance in distance learning using machine learning techniques," *Applied Artificial Intelligence*, vol. 18, no. 5, pp. 411–426, 2004.
10. O. El Aissaoui, Y. E. M. El Alami, L. Oughdir, and Y. El Allioui, "A hybrid machine learning approach to predict learning styles in adaptive E-learning system," in *Proceedings of the International Conference on Advanced Intelligent Systems for Sustainable Development*, vol. 5, pp. 772–786, Springer, Tetouan, Morocco, January 2019.
11. J. Xu, K. H. Moon, and M. Van Der Schaar, "A machine learning approach for tracking and predicting student performance in degree programs," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 5, pp. 742–753, 2017.
12. G. De'ath, ``Boosted trees for ecological modeling and prediction," Ecol- ogy, vol. 88, no. 1, pp. 243251, 2007.
13. S. Nawar and A. M. Mouazen, ``Comparison between random forests, artificial neural networks and gradient boosted machines methods of online vis-NIR spectroscopy measurements of soil total nitrogen and total carbon," Sensors, vol. 17, no. 10, p. 2428, 2017.
14. R. L. T. Hahnloser, ``On the piecewise analysis of networks of linear threshold neurons," Neural Netw., vol. 11, no. 4, pp. 691697, 1998.
15. G. L. Marcialis and F. Roli, ``Fusion of multiple fingerprint matchers by single-layer perceptron with class-separation loss function," Pattern Recognit. Lett., vol. 26, no. 12, pp. 18301839, 2005.

16. H. G. Hosseini, D. Luo, and K. J. Reynolds, ``The comparison of different feed forward neural network architectures for ECG signal diagnosis," Med. Eng. Phys., vol. 28, no. 4, pp. 372378, 2006.
17. J. A. Bullinaria, ``Learning in multi-layer perceptrons-back-propagation," Neural Comput., Lect., vol. 7, no. 8, pp. 116, 2015.
18. I. Guyon and A. Elisseeff, ``An introduction to variable and feature selection," J. Mach. Learn. Res., vol. 3, pp. 11571182, Jan. 2003.
19. G. Chandrashekar and F. Sahin, ``A survey on feature selection methods," Comput. Elect. Eng., vol. 40, no. 1, pp. 1628, Jan. 2014.149478
20. D. R. Tobergte, S. Curtis, B. Lantz, D. R. Tobergte, S. Curtis, and B. Lantz,
Machine Learning with R Cookbook, vol. 53, no. 9. 2013.

12. PUBLICATIONS

International Conference on “Innovations in Computers Networks, Computational Intelligence and IOT” (ICICCI-21)

Paper ID: ICICCI-21-0119

13. PROFILE OF STUDENTS:

S.Sushma 17K81A0551:

S. Sushma is currently pursuing her Bachelor of Technology in the stream of Computer Science Engineering at St. Martin's Engineering College. She completed her intermediate from Sri Chaitanya Junior College and 10th class from Hindu Public School. Her technical skills include Python, C, HTML and CSS. She also has a basic understanding of Java. She took part in Employment Skill Development Program conducted by Zensar. She is also a student of Smart Interviews. She did a one moth internship i.e., from June 2019 to July 2019, in National Small Industries Corporation (NSIC), ECIL, where she was trained in Python programming language. Her participations include: Women online workshop on "Women in Cyber Security and Privacy 2020" which was conducted from 6th to 10th July, Online Two Day National Level Seminar on "Recent Trends in Cloud Computing Fog and Edge Computing" from 18th to 19th June, 2021. She completed few certification courses from online platforms like Udemy, Uxcel, Coursera, SoloLearn.

P.Keerthan Srichakra 17K81A0540:

P. Keerthan Sri Chakra is currently pursuing her Bachelor of Technology in the stream of Computer Science Engineering at St. Martin's Engineering College. He completed his intermediate from New Vision Junior College and 10th class from Vani Vidhyalayam School. His technical skills include Python, C, HTML and CSS. He also has a basic understanding of Java and DBMS. He took part in Employment Skill Development Program conducted by Zensar. He is also a student of Smart Interviews. Apart from programming, He is also interested in User Experience Designing. He participated in a state wide Design Hackathon conducted by HYSEA and JNTUH in which he stood as a Winner. He did a one moth internship i.e., from June 2019 to July 2019, with Verzeo where he was trained in basics of Machine Learning. His participations include: Online International Conference on "Innovations in Computer Networks, Computational Intelligence and loT" [ICICCI-21] On 25th June, 2021 and Online Two Day National Level Seminar on "Recent Trends in Cloud Computing Fog and Edge Computing" from 18th to 19th June, 2021. He completed few certification courses from online platforms like Udemy, Uxcel, Coursera, SoloLearn.

Vardhan.CM 17K81A0556:

C.M Vardhan is currently pursuing her Bachelor of Technology in the stream of Computer Science Engineering at St. Martin's Engineering College. He completed his intermediate from Sri Chaitanya Junior Kalasala and 10th class from Pallavi Model School. His technical skills include Python, C, HTML and CSS. He also has a basic understanding of Java and DBMS. He took part in Employment Skill Development Program conducted by Zensar. He is also a student of Smart Interviews. Apart from programming, He is also interested in User Experience Designing. He participated in a state wide Design Hackathon conducted by HYSEA and JNTUH in which he stood as a Winner. He did a one moth internship i.e., from June 2019 to July 2019, with Verzeo where he was trained in basics of Machine Learning. His participations include: Online International Conference on "Innovations in Computer Networks, Computational Intelligence and loT" [ICICCI-21] On 25th June, 2021 and Online Two Day National Level Seminar on "Recent Trends in Cloud Computing Fog and Edge Computing" from 18th to 19th June, 2021. He completed few certification courses from online platforms like Udemy, Uxcel, Coursera, SoloLearn.

K.Harshit 17K81A0531:

K. Harshit is currently pursuing her Bachelor of Technology in the stream of Computer Science Engineering at St. Martin's Engineering College. He completed his intermediate from Sri Chaitanya Junior Kalasala and 10th class from Vignan Vidyalaya School. His technical skills include Python, C, HTML. He also has a basic understanding of Java and DBMS. He took part in Employment Skill Development Program conducted by Zensar. He is also a student of Smart Interviews. Apart from programming, He is also interested in sports like Volleyball, Soccer and has won various certificates. He did a one moth internship i.e., from June 2019 to July 2019, with Verzeo where he was trained in basics of Machine Learning. His participations include: Online International Conference on "Innovations in Computer Networks, Computational Intelligence and loT" [ICICCI-21] On 25th June, 2021 and Online Two Day National Level Seminar on "Recent Trends in Cloud Computing Fog and Edge Computing" from 18th to 19th June, 2021. He completed few certification courses from online platforms like Udemy, Uxcel, Coursera, SoloLearn.

A

PROJECT REPORT

On

SIGN LANGUAGE RECOGNITION USING CONVOLUTIONAL NEURAL NETWORK AND COMPUTER VISION

Submitted by

1)Ms.D.Preethi(17K81A0512)

2)Ms.Fouzia begum (17K81A0513)

3)Ms.G.Chetna Varma (17K81A0521)

4)Ms.J.Nishitha(17K81A0524)

in partial fulfillment for the award of the

degreeof

BACHELOR OF TECHNOLOGY

IN

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Under the Guidance of

Dr. B. RAJALINGAM

Associate professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ST.MARTIN'S ENGINEERING COLLEGE
An Autonomous Institute

Dhulapally, Secunderabad – 500 100

JUNE 2021

BONAFIDE CERTIFICATE

This is to certify that the project entitled Sign language recognition using convolutional neural network and computer vision , is being submitted by 1.**Ms. D.Preethi 17K81A0512, 2. Fouzia begum 17K81A0513, 3. G.Chetna varma 17K81A0521 4. J.Nishitha 17K81A0524** in partial fulfillment of the requirement for the award of the degree of **BACHELOR OF TECHNOLOGY IN COMPUTER SCIENCE AND ENGINEERING** is recorded of bonafide work carried out by them. The result embodied in this report have been verified and found satisfactory.

Dr. B.RAJALINGAM
Department of CSE

Head of the Department
Dr. M.NARAYANAN
Department of CSE

Internal Examiner **External Examiner**

Place:

Date:

DECLARATION

We, the student of **Bachelor of Technology** in Department of Computer Science and Engineering', session: 2017 – 2021, St. Martin's Engineering College, Dhulapally, Kompally, Secunderabad, hereby declare that work presented in this Project Work entitled "**Sign language recognition using convolutional neural network and computer vision"** is the outcome of our own bonafide work and is correct to the best of our knowledge and this work has been undertaken taking care of Engineering Ethics. This result embodied in this project report has not been submitted in any university for award of any degree.

D.Preethi 17K81A0512

Fouzia Begum 17K81A0513

G.Chetna varma 17K81A0521

J.Nishitha 17K81A0524

ABSTRACT

Sign Language Recognition (SLR) targets on interpreting the sign language into text or speech, so as to facilitate the communication between deaf-mute people and ordinary people. This task has broad social impact but is still very challenging due to the complexity and large variations in hand actions. Existing methods for SLR use hand-crafted features to describe sign language motion and build classification models based on those features. However, it is difficult to design reliable features to adapt to the large variations of hand gestures. To approach this problem, we propose a novel convolutional neural network (CNN) which extracts discriminative spatial-temporal features from raw video stream automatically without any prior knowledge, avoiding designing features. To boost the performance, multi-channels of video streams, including colour information, depth clue, and body joint positions, are used as input to the CNN in order to integrate colour, depth and trajectory information. We validate the proposed model on a real dataset collected with Microsoft Kinect and demonstrate its effectiveness over the traditional approaches based on hand-crafted features.

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible and whose encouragements and guidance have crowded effects with success.

We extended our deep sense of gratitude to Principal**,** **Dr. P. SANTOSH KUMAR PATRA**, St. Martin's Engineering College, Dhulapally, for permitting us to undertake this project.

We are also thankful to **Dr. M.NARAYANAN**, Head of the Department, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

We would like to thank **Dr. T. POONGOTHAI,** Department of Computer Science and Engineering (AI & ML) for her encouragement and insightful comments and as well as our project coordinators **Dr. B.RAJALINGAM**, Associate Professor and **Mr. J.SUDHAKAR**, Assistant Professor, in Department of Computer Science and Engineering for their valuable support.

We would like to express our sincere gratitude and indebtedness to our project supervisor **Dr. B.RAJALINGAM**, Associate Professor, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

Finally, we express thanks to all those who have helped us successfully to completing this project. Furthermore, we would like to thank our family and friends for their moral support and encouragement.

We express thanks to all those who have helped us in successfully completing the project.

D.Preethi	17K81A0512
Fouzia Begum	17K81A0513
G.Chetna varma	17K81A0521
J.Nishitha	17K81A0524

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

1. INTRODUCTION

1.1 PROJECT OVERVIEW

“Talk to a man in a language he knows, that goes to his head,” as Nelson Mandela[1] put it. Talk to him in his own language; it will reach his heart.” Language is undeniably important in human interaction and has existed since the dawn of civilization. It is a medium through which individuals communicate in order to express themselves and comprehend real-world concepts. No books, no cell phones, and certainly no word I'm writing would be meaningful without it. It is so deeply ingrained in our daily lives that we frequently take it for granted and overlook its significance. Unfortunately, in today's fast-paced society, those with hearing impairments are frequently neglected and excluded. They must strive to communicate themselves to others who are different from them, to bring up their thoughts, to speak their opinions, and to express themselves. Even though sign language is a means of communication for the deaf, it has no significance when communicated to a non-sign language user.

As a result, the communication gap is being bridged. We're putting in place a sign language recognition system to prevent this from happening. It will be a fantastic tool for persons with hearing impairments to convey their thoughts, as well as a great way for non-sign language users to grasp what the latter is saying. Many countries have their own set of sign motions and interpretations. An alphabet in Korean sign language, for example, will not be the same as an alphabet in Indian sign language. While this emphasises the diversity of sign languages, it also emphasises their complexity. Deep learning must be well-versed in gestures in order to achieve a reasonable level of accuracy. The datasets in our proposed system are created using American Sign Language. The alphabets of American Sign Language (ASL) are shown in Figure 1.

Either of the two ways can be used to identify a sign gesture. The first is a glove-based approach, in which the signer wears a pair of data gloves while the hand movements are captured. The second way is based on vision, which is further divided into static and dynamic recognition [2]. Static is concerned with the portrayal of motions in two dimensions, whereas dynamic is concerned with the recording of motions in real time. Gloves, though having an accuracy of over 90% [3] are uncomfortable to wear and cannot be used in rainy conditions. They are cumbersome to transport because their use necessitates the use of a computer.

We chose static recognition of hand movements in this example because it improves accuracy when compared to dynamic hand movements, such as those for the letters J and Z. We propose this study in order to improve accuracy by utilising Convolution Neural Networks (CNN).

Fig. 1.1.American Sign Language Alphabets

Sign language, as one of the most widely used communication means for hearing-impaired people, is expressed by variations of handshapes, body movement, and even facial expression. Since it is difficult to collaboratively exploit the information from handshapes and body movement trajectory, sign language recognition is still a very challenging task. This paper proposes an effective recognition model to translate sign language into text or speech in order to help the hearing impaired communicate with normal people through sign language.

Technically speaking, the main challenge of sign language recognition lies in developing descriptors to express handshapes and motion trajectory. In particular, hand-shape description involves tracking hand regions in video stream, segmenting hand-shape images from complex background in each frame and gestures recognition problems. Motion trajectory is also related to tracking of the key points and curve matching. Although lots of research works have been conducted on these two issues for now, it is still hard to obtain satisfying result for SLR due to the variation and occlusion of hands and body joints. Besides, it is a nontrivial issue to integrate the hand-shape features and trajectory features

together. To address these difficulties, we develop a CNNs to naturally integrate hand-shapes, trajectory of action and facial expression. Instead of using commonly used color images as input to networks like [1, 2], we take color images, depth images and body skeleton images simultaneously as input which are all provided by Microsoft Kinect.

Kinect is a motion sensor which can provide color stream and depth stream. With the public Windows SDK, the body joint locations can be obtained in real-time as shown in Fig.1. Therefore, we choose Kinect as capture device to record sign words dataset. The change of color and depth in pixel level are useful information to discriminate different sign actions. And the variation of body joints in time dimension can depict the trajectory of sign actions. Using multiple types of visual sources as input leads CNNs paying attention to the change not only in color, but also in depth and trajectory. It is worth mentioning that we can avoid the difficulty of tracking hands, segmenting hands from background and designing descriptors for hands because CNNs have the capability to learn features automatically from raw data without any prior knowledge [3].

CNNs have been applied in video stream classification recently years. A potential concern of CNNs is time consuming. It costs several weeks or months to train a CNNs with million-scale in million videos. Fortunately, it is still possible to achieve real-time efficiency, with the help of CUDA for parallel processing. We propose to apply CNNs to extract spatial and temporal features from video stream for Sign Language Recognition (SLR). Existing methods for SLR use hand-crafted features to describe sign language motion and build classification model based on these features. In contrast, CNNs can capture motion information from raw video data automatically, avoiding designing features. We develop a CNNs taking multiple types of data as input. This architecture integrates color, depth and trajectory information by performing convolution and subsampling on adjacent video frames. Experimental results demonstrate that 3D CNNs can significantly outperform Gaussian mixture model with Hidden Markov model (GMM-HMM) baselines on some sign words recorded by ourselves.

1.2 PROJECT OBJECTIVES

Conversing to a person with hearing disability is always a major challenge.

Sign language has indelibly become the ultimate panacea and is a very powerful tool for individuals with hearing and speech disability to communicate their feelings and opinions to the world.

Main objective of the project is to broadening the communication gap between normal people and people with hearing disability Sign Language helps people with hearing disability to communicate their thoughts as well as a very good interpretation for non sign language user to understand what the latter is saying.

1.3 ORGANIZATION OF CHAPTERS

Besides the introduction, the thesis is organized in other six chapters as follows:

Chapter 1: Introduction

Conversing to a person with hearing disability is always a major challenge. Sign language has indelibly become the ultimate panacea and is a very powerful tool for individuals with hearing and speech disability to communicate their feelings and opinions to the world.Main objective of the project is to broadening the communication gap between normal people and people with hearing disability

Chapter 2: Literature Survey

A survey of the literature for our proposed system reveals that many attempts have been made to solve sign identification in videos and photos using various methodologies and algorithms.

Chapter 3: Software and Hardware Requirements

This chapter discuss about the software and hardware required for the execution of the project.

Chapter 4: Software Development Analysis

This chapter explains the assumptions and technical specifications of the project.

Chapter 5: Project System Design

This chapter explains all the software development process with dfd, E-R diagrams, UML diagrams clearly.

Chapter 6: Project Coding

A programming project produces a well-designed executing system that solves a specified distributed programming problem. A project code is used to represent a one-time, or intermittent departmental event or activity. Any person can use a project code on a transaction, regardless of the project manager or home organization. This section describes some of the coding templates, outline of various files, class with functionalities, the various methods of input and output parameters.

Chapter 7: Project Testing

The testing phase checks the software for bugs and verifies its performance before delivery to users. In this stage, expert testers verify the product's functions to make sure it performs according to the requirements analysis document.Testers use exploratory testing if they have experience with that software or a test script to validate the performance of individual components of the software. They notify developers of defects in the code. If developers

confirm the flaws are valid, they improve the program, and the testers repeat the process until the software is free of bugs and behaves according to requirements.

Chapter 8: Output screens

The output of the programmed project is being screened with the screenshots. Front end development is done which is connected with the back-end servers database and the operations are done with the final input. The various test case results are captured and projected some sample outputs.

Chapter 9: Experimental Results

Tests and results are shown and explained in this chapter. The results are analyzed in the context of the thesis project and followed by discussion on systems throughput and resiliency, as well as the approaches to testing and analysis

Chapter 10: CONCLUSION AND FUTURE ENHANCEMENT

The chapter ends the project with a short summary of the main concepts mentioned in the thesis as well as the relevant results.

CHAPTER 2

LITERATURE SURVEY

2 LITERATURE SURVEY

A systematic and thorough search of all types of published literature as well as other sources including dissertation, theses in order to identify as many items as possible that are relevant to a particular topic.

2.1 SURVEY ON BACKGROUND

A survey of the literature for our proposed system reveals that many attempts have been made to solve sign identification in videos and photos using various methodologies and algorithms. Siming He[4] suggested a system using a 40-word dataset and 10,000 sign language graphics. Faster R-CNN with an incorporated RPN module is utilised to locate the hand regions in the video frame. In terms of accuracy, it enhances performance. When compared to single stage target detection algorithms like YOLO, detection and template classification can be done at a faster rate When compared to Fast-RCNN, the detection accuracy of Faster R-CNN improves from 89.0 percent to 91.7 percent in the paper. For the language image sequences, a 3D CNN is employed for feature extraction, and a sign-language recognition framework comprising of long and short time memory (LSTM) coding and decoding networks is created. The paper combines the hand locating network, 3D CNN feature extraction network, and LSTM encoding and decoding to develop an extraction technique for the problem of RGB sign language picture or video recognition in practical scenarios. In the common vocabulary dataset, this paper attained a recognition rate of 98%.

Let's have a look at Rekha's research, J[5]. The skin region of the hand motions was detected and fragmented using the YCbCr skin model. The visual characteristics are retrieved and categorised with Multiclass SVM, DTW, and non-linearKNN using the Principal Curvature based Region Detector. For training, a dataset of 23 Indian Sign Language static alphabet signs was employed, and 25 videos were used for testing The static result was 94.4 percent, while the dynamic result was 86.4 percent. For image processing, a low-cost technique was adopted in [6]. The photographs were taken with a green background so that the green colour could be readily eliminated from the RGB colorspace during processing and the image could be transformed to black and white. The sign gestures were made in Sinhala. The strategy presented in the paper is to use the centroid approach to map the signs. It can map the input gesture to a database regardless of the size or position of the hands. 92 percent of the sign gestures were accurately recognized by the prototype.

M. Geetha and U. C. Manjusha used 50 examples of each letter and digit in a vision-based recognition of Indian Sign Language characters and numerals using B-Spine approximations in their paper[7]. The sign gesture's region of interest is analyzed, and the boundary is removed. The acquired boundary is then converted to a B-spline curve utilizing the Maximum Curvature Points (MCPs) as Control points. A number

of smoothing processes are applied to the B-spline curve in order to extract features. The photos are classified using a support vector machine, which has a 90.00 percent accuracy. Pigou utilized CLAP14 as his dataset [9] in [8]. It is made up of 20 Italian hand gestures. He utilized a Convolutional Neural Network model with six layers for training after preprocessing the photos. It's worth noting that his model isn't a three-dimensional CNN, and all of the kernels are two-dimensional. Rectified linear Units (ReLU) were employed as activation functions. Feature extraction is performed by the CNN while classification uses ANN or fully connected layer. His work has achieved an accuracy of 91.70% with an error rate of 8.30% .

J Huang [10] did something similar. Using Kinect, he developed his own dataset, resulting in a total of 25 vocabularies that are utilised in everyday life. After that, he used a 3D CNN with all kernels in 3D. His model's input included five crucial channels: color-r, color-b, color-g, depth, and body skeleton. He was 94.2 percent accurate on average. Another study work on the issue of action recognition by J.Carriera [11] has some parallels with sign gesture recognition. For his research, he applied a transfer learning strategy. He used ImageNet[12] and KineticDataset [9] as his pre-trained datasets. He then mixed the RGB datasets after training the pertained models with additional two datasets, UCF-101 [13] and HMDB-51 [14] , flow model, pre-trained Kinetic, and pre-trained ImageNet are all examples of pre-trained models

2.2 CONCLUSIONS ON SURVEY

Many breakthroughs have been made in the field of artificial intelligence, machine learning and computer vision.They have immensely contributed in how we perceive things around us and improve the way in which we apply their techniques in our everyday lives.Many researches have been conducted on sign gesture recognition using different techniques like ANN, LSTM and 3D CNN. However, most of them require extra computing power . On the other hand, our project requires low computing power and gives a remarkable accuracy of above 90%.We proposed to normalize and rescale our images to 64 pixels in order to extract features (binary pixels) and make the system more robust. We use CNN to classify the 10 alphabetical American sign gestures and successfully achieve an accuracy of 98%.

CHAPTER 3

SOFTWARE AND HARDWARE REQUIREMENTS

3 SOFTWARE AND HARDWARE REQUIREMENTS

To be used efficiently, all computer software needs certain hardware components or other software resources to be present on a computer. These prerequisites are known as (computer) system requirements and are often used as a guideline as opposed to an absolute rule. Most software defines two sets of system requirements: minimum and recommended. With increasing demand for higher processing power and resources in newer versions of software, system requirements tend to increase over time. Industry analysts suggest that this trend plays a bigger part in driving upgrades to existing computer systems than technological advancements. A second meaning of the term of system requirements, is a generalisation of this first definition, giving the requirements to be met in the design of a system or sub-system.

3.1 SOFTWARE REQUIREMENTS

For developing the application the following are the software requirements:

- Python
- Operating Systems supported windows
- Technologies and Languages used to Develop python

3.2 HARDWARE REQUIREMENTS

- **Processor -** Pentium–III
- **Speed** – 2.4GHz
- **RAM -** 512 MB(min)
- **Key Board -** Standard Keyboard
- **Monitor** – 15 VGAColour

CHAPTER 4

SOFTWARE DEVELOPMENT

ANALYASIS

4 SOFTWARE DEVELOPMENT ANALYSIS

The software development process involves the creation and maintenance of applications, frameworks and other components for software design, design, programming, documentation, testing and problem remediation. The development of software is a process of creating and keeping source code, but it encompasses everything from the idea of the intended software to the last manifestation of the programme, often in a planned and organised process in a larger context. Software development may therefore encompass research, creation of new software products, prototype, modification, reuse, reengineering, maintenance, or any other software-production activity.

4.1 OVERVIEW OF PROBLEM

- Conversing to a person with hearing disability is always a major challenge.
- Sign language has indelibly become the ultimate panacea and is a very powerful tool for individuals with hearing and speech disability to communicate their feelings and opinions to the world.
- In this the user must be able to capture images of the hand gesture using web camera and the system shall predict and display the name of the captured image.
- Sign Language Recognition (SLR) targets on interpreting the sign language into text or speech, so as to facilitate the communication between deaf-mute people and ordinary people.

4.2 DEFINE THE PROBLEM

We intend to create an sign language recognition system to facilitate the communication between deaf-mute people and ordinary people. As a result, the communication gap is being bridged. We're putting in place a sign language recognition system to prevent this from happening. It will be a fantastic tool for persons with hearing impairments to convey their thoughts, as well as a great way for non-sign language users to grasp what the latter is saying. Many countries have their own set of sign motions and interpretations. An alphabet in Korean sign language, for example, will not be the same as an alphabet in Indian sign language. While this emphasises the diversity of sign languages, it also emphasises their complexity. . Deep learning must be well-versed in gestures in order to achieve a reasonable level of accuracy. The datasets in our proposed system are created using American Sign Language.

4.3 MODULES OVERVIEW

A module allows you to logically organize your Python code. Grouping related code into a module makes the code easier to understand and use. A module is a Python object with arbitrarily named attributes that you can bind and reference.

Simply, a module is a file consisting of Python code. A module can define functions, classes and variables. A module can also include runnable code. The modules used in this project are as mentioned below.

1. numpy

2. pandas

3. matplotlib

4. keras

5.tensorflow

6.opencv-python

7. scikit-image

4.4 DEFINE THE MODULES

Tensorflow

TensorFlow is a free and open-source software library for dataflow and differentiable programming across a range of tasks. It is a symbolic math library, and is also used for machine learning applications

Numpy

Numpy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools

Pandas

Pandas is an open-source Python Library providing high-performance data manipulation and analysis tool using its powerful data structures.

Matplotlib

Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms.

Scikit – learn

Scikit-learn provides a range of supervised and unsupervised learning algorithms via a consistent interface in Python.

4.5 MODULE FUNCTIONALITY

Tensorflow

TensorFlow is a free and open-source software library for dataflow and differentiable programming across a range of tasks. It is a symbolic math library, and is also used for machine learning applications such as neural networks. It is used for both research and production at Google.

TensorFlow was developed by the Google Brain team for internal Google use. It was released under the Apache 2.0 open-source license on November 9, 2015.

Numpy

Numpy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays.

It is the fundamental package for scientific computing with Python. It contains various features including these important ones:

- A powerful N-dimensional array object
- Sophisticated (broadcasting) functions
- Tools for integrating C/C++ and Fortran code
- Useful linear algebra, Fourier transform, and random number capabilities

Besides its obvious scientific uses, Numpy can also be used as an efficient multi-dimensional container of generic data. Arbitrary data-types can be defined using Numpy which allows Numpy to seamlessly and speedily integrate with a wide variety of databases.

Pandas

Pandas is an open-source Python Library providing high-performance data manipulation and analysis tool using its powerful data structures. Python was majorly used for data munging and preparation. It had very little contribution towards data analysis. Pandas solved this problem. Using Pandas, we can accomplish five typical steps in the processing and analysis of data, regardless of the origin of data load, prepare, manipulate, model, and analyze. Python with Pandas is used in a wide range of fields including academic and commercial domains including finance, economics, Statistics, analytics, etc.

Matplotlib

Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms. Matplotlib can be used in Python scripts, the Python and IPython shells, the Jupyter Notebook, web application servers, and four graphical user interface toolkits. Matplotlib tries to make easy things easy and hard things possible. You can generate plots, histograms, power spectra, bar charts, error charts, scatter plots, etc., with just a few lines of code. For examples, see the sample plots and thumbnail gallery.

For simple plotting the pyplot module provides a MATLAB-like interface, particularly when combined with IPython. For the power user, you have full control of line styles, font properties, axes properties, etc, via an object oriented interface or via a set of functions familiar to MATLAB users.

Scikit – learn

Scikit-learn provides a range of supervised and unsupervised learning algorithms via a consistent interface in Python. It is licensed under a permissive simplified BSD license and is distributed under many Linux distributions, encouraging academic and commercial use. **Python**

Python is an interpreted high-level programming language for general-purpose programming. Created by Guido van Rossum and first released in 1991, Python has a design philosophy that emphasizes code readability, notably using significant whitespace.

Python features a dynamic type system and automatic memory management. It supports multiple programming paradigms, including object-oriented, imperative, functional and procedural, and has a large and comprehensive standard library.

- Python is Interpreted − Python is processed at runtime by the interpreter. You do not need to compile your program before executing it. This is similar to PERL and PHP.
- Python is Interactive − you can actually sit at a Python prompt and interact with the interpreter directly to write your programs.

Python also acknowledges that speed of development is important. Readable and terse code is part of this, and so is access to powerful constructs that avoid tedious repetition of code. Maintainability also ties into this may be an all but useless metric, but it does say something about how much code you have to scan, read and/or understand to troubleshoot problems or tweak behaviors. This speed of development, the ease with which a programmer of other languages can pick up basic Python skills and the huge standard library is key to another area where Python excels. All its tools have been quick to implement, saved a lot of time, and several of them have later been patched and updated by people with no Python background - without breaking.

CHAPTER 5

PROJECT SYSTEM DESIGN

5 PROJECT SYSTEM DESIGN

The first phase is to collect data. To capture hand movements, many researchers have employed sensors or cameras. The hand motions are captured using the web camera in our system. The photographs go through a series of steps in which the backgrounds are recognized and removed using the HSV colour extraction technique (Hue, Saturation, Value). Following that, segmentation is used to identify the skin tone zone. A mask is applied to the images using morphological processes, and a series of dilation and erosion using an elliptical kernel is performed. The photographs obtained with open CV are all cropped to the same size, therefore there is no difference between the photos of different gestures . Our dataset contains 2000 photos of American sign gestures, of which 1600 are for training and 400 are for testing. It has an 80:20 ratio. Each frame's binary pixels are retrieved, and a Convolutional Neural Network is used to train and classify them. After that, the model is evaluated, and the system is able to predict the alphabets.

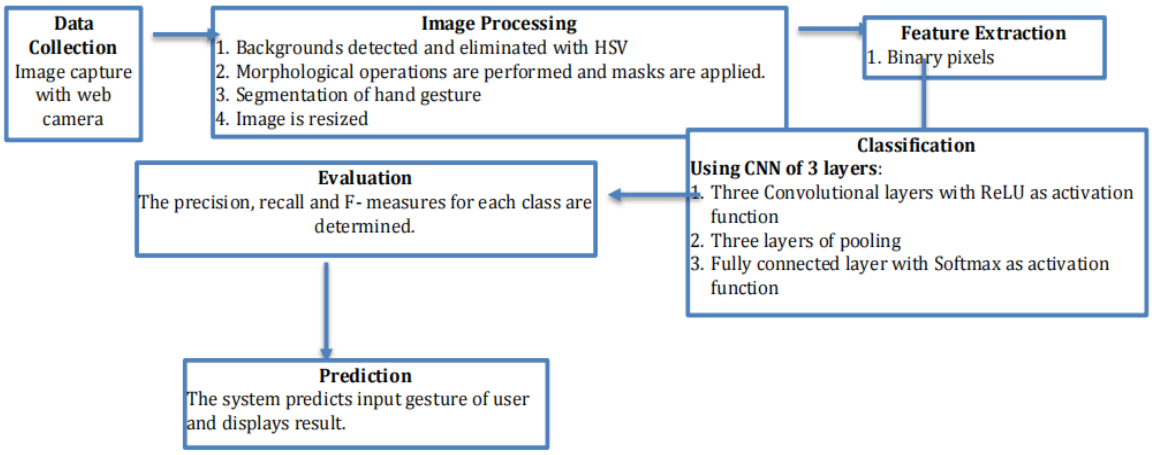


Fig. 5. System Architecture

(i) DATA COLLECTION

Data collecting is an indisputable component of this study, as our outcome is greatly dependent on it. As a result, we developed our own ASL dataset with 2000 photos of 10 static alphabet signs. A,B,C,D,K,N,O,T, and Y are the ten classes of static alphabets. Two distinct signers created two different datasets. In different lighting situations, each of them has made one alphabetical motion 200 times. The alphabetic sign motions dataset folder is further divided into two folders, one for training and the other for testing. 1600 photographs are utilised for training and the remaining photos are used for testing. To ensure consistency, we took images with a webcam in the same background each time a command was issued. The resulting images are saved in the png format. It should be noted that when a png image is opened, closed, and saved again, there is no loss of quality. PNG is also capable of handling images with high contrast and detail.

(ii) DATA PROCESSING

Because the photos are in RGB colour spaces, segmenting the hand gesture only on the basis of skin colour becomes more challenging. As a result, we convert the photos to HSV colour space. It's a model that divides an image's colour into three pieces, each of which has its own colour.: Hue, saturation, and value are all important factors to consider. HSV is a useful tool for improving image stability by separating

brightness from chromaticity [15]. Because the Hue element is unaffected by any form of light, shadows, or shadings[16], it can be used to remove the background.

To detect the hand gesture and set the background to black, a track-bar with H values ranging from 0 to 179, S values ranging from 0-255, and V values ranging from 0 to 255 is utilised. Dilation and erosion operations using elliptical kernels are performed on the hand gesture regionThe first image is obtained when the two masks are applied, as illustrated in fig 5.1. (b)

Fig. 5.1 (a) Image captured from web-camera (b) Image after background is set to black

A. SEGMENTATION

After that, the first image is converted to grayscale. While this technique may result in a loss of colour in the skin gesture region, it will also improve our system's resiliency to changes in lighting or illumination. The non-black pixels in the modified image are binarized, while the others stay black. The hand gesture is split in two ways: first, by removing all of the image's attached components, and then, by allowing only the part that is really related, in this case, the hand gesture. The frame has been shrunk to a 64 by 64 pixel size. After the segmentation process, binary pictures of 64 by 64 pixels are created, with the white area representing the hand gesture and the black coloured area representing the rest.

Fig. 5.2 (a) Image after binarise. (b) Image after segmentation and resizing.

B. FEATURE EXTRACTION

The ability to identify and extract relevant elements from an image is one of the most significant aspects of image processing. When images are recorded and saved as a dataset, they typically take up a lot of space since they contain a lot of data. Feature extraction assists us in solving this challenge by automatically decreasing the data after the key features have been extracted. It also helps to preserve the classifier's accuracy while simultaneously reducing its complexity. The binary pixels of the photographs were

determined to be critical in our case. We were able to gather enough characteristics by scaling the photos to 64 pixels to effectively classify the American Sign Language gestures. We have 4096 features in total, which we obtained by multiplying 64 by 64 pixels.

5.1. E-R DIAGRAMS

An E-R model is usually the result of systematic analysis to define and describe what is important to process in an area of a business. It does not define the business processes; it only presents a business data schema in graphical form. It is usually drawn in a graphical form as boxes (entities) that are connected by lines (relationships) which express the associations and dependencies between entities. An ER model can also be expressed in a verbal form, for example: one building may be divided into zero or more apartments, but one apartment can only be located in one building. Entities may be characterized not only by relationships, but also by additional properties (attributes), which include identifiers called "primary keys". Diagrams created to represent attributes as well as entities and relationships may be called entity-attribute-relationship diagrams, rather than entity–relationship models.

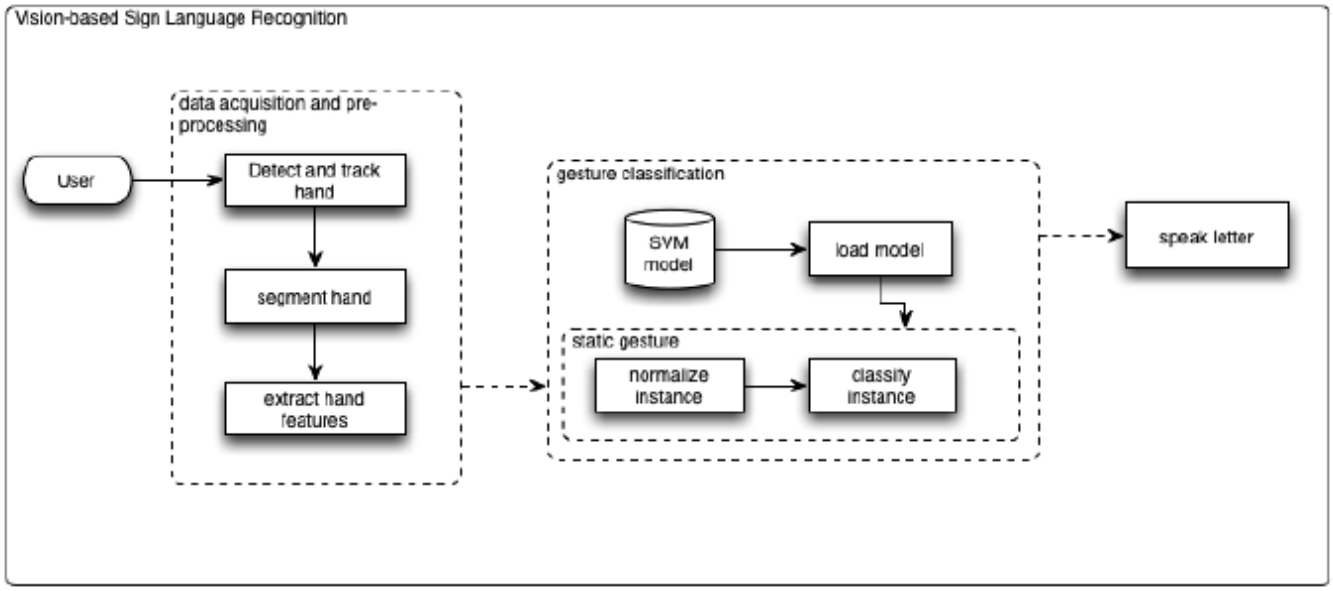


Fig.5.3. E-R Diagrams

An ER model is typically implemented as a database. In a simple relational database implementation, each row of a table represents one instance of an entity type, and each field in a table represents an attribute type. In a relational database a relationship between entities is implemented by storing the primary key of one entity as a pointer or "foreign key" in the table of another entity. There is a tradition for ER/data models to be built at two or three levels of abstraction. Note that the conceptual-logical-physical hierarchy below is used in other kinds of specification, and is different from the three schema approach to software engineering.

5.2.UML DIAGRAMS

USE CASEDIAGRAM:

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.

UML is a modern approach to modelling and documenting software. In fact, it’s one of the most popular business process modelling techniques. It is based on diagrammatic representations of software components. As the old proverb says: “a picture is worth a thousand words”. By using visual representations, we are able to better understand possible flaws or errors in software or business processes. Mainly, UML has been used as a general-purpose modelling language in the field of software engineering. However, it has now found its way into the documentation of several business processes or workflows. For example, activity diagrams, a type of UML diagram, can be used as a replacement for flowcharts. They provide both a more standardized way of modelling workflows as well as a wider range of features to improve readability and efficacy.

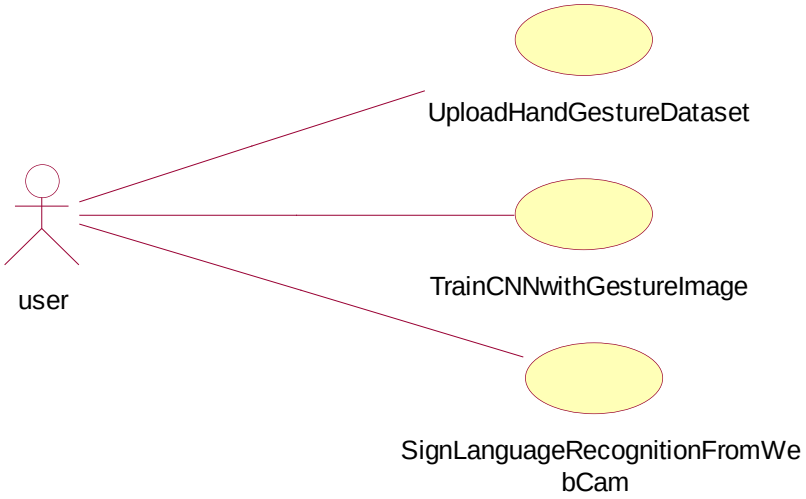


Fig.5.4 User use case diagram

While a use case itself might drill into a lot of detail about every possibility, a use-case diagram can help provide a higher-level view of the system. It has been said before that "Use case diagrams are the blueprints for your system".

Due to their simplistic nature, use case diagrams can be a good communication tool for stakeholders. The drawings attempt to mimic the real world and provide a view for the stakeholder to understand how the system is going to be designed. Siau and Lee conducted research to determine if there was a valid situation for use case diagrams at all or if they were unnecessary. What was found was that the use case diagrams conveyed the intent of the system in a more simplified manner to stakeholders and that they were "interpreted more completely than class diagrams". The purpose of a use case diagram is to capture the dynamic aspect of a system. They provide a simplified graphical representation of what the system should do in a use case. Further diagrams and documentation are needed for a complete functional and technical outlook on the system.

CLASS DIAGRAM :

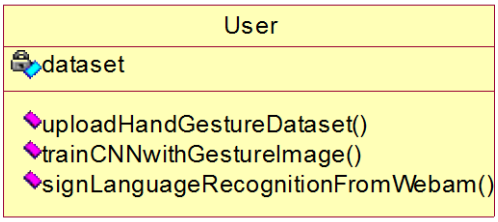


Fig.5.5 Class diagram

In software engineering, a class diagram in the Unified Modelling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among objects. The class diagram is the main building block of object-oriented modelling. It is used for general conceptual modelling of the structure of the application, and for detailed modelling, translating the models into programming code. Class diagrams can also be used for data modelling. The classes in a class diagram represent both the main elements, interactions in the application, and the classes to be programmed. In the diagram, classes are represented with boxes that contain three compartments: 28 • The top compartment contains the name of the class. It is printed in bold and cantered, and the first letter is capitalized. • The middle compartment contains the attributes of the class. They are left-aligned, and the first letter is lowercase. • The bottom compartment contains the operations the class can execute. They are also leftaligned, and the first letter is lowercase. A class with three compartments. In the design of a system, a number of classes are identified and grouped together in a class diagram that helps to determine the static relations between them. In detailed modelling, the classes of the conceptual design are often split into subclasses. In order to further describe the behaviour of systems, these class diagrams can be complemented by a state diagram or UML state machine.

SEQUENCE DIAGRAM :

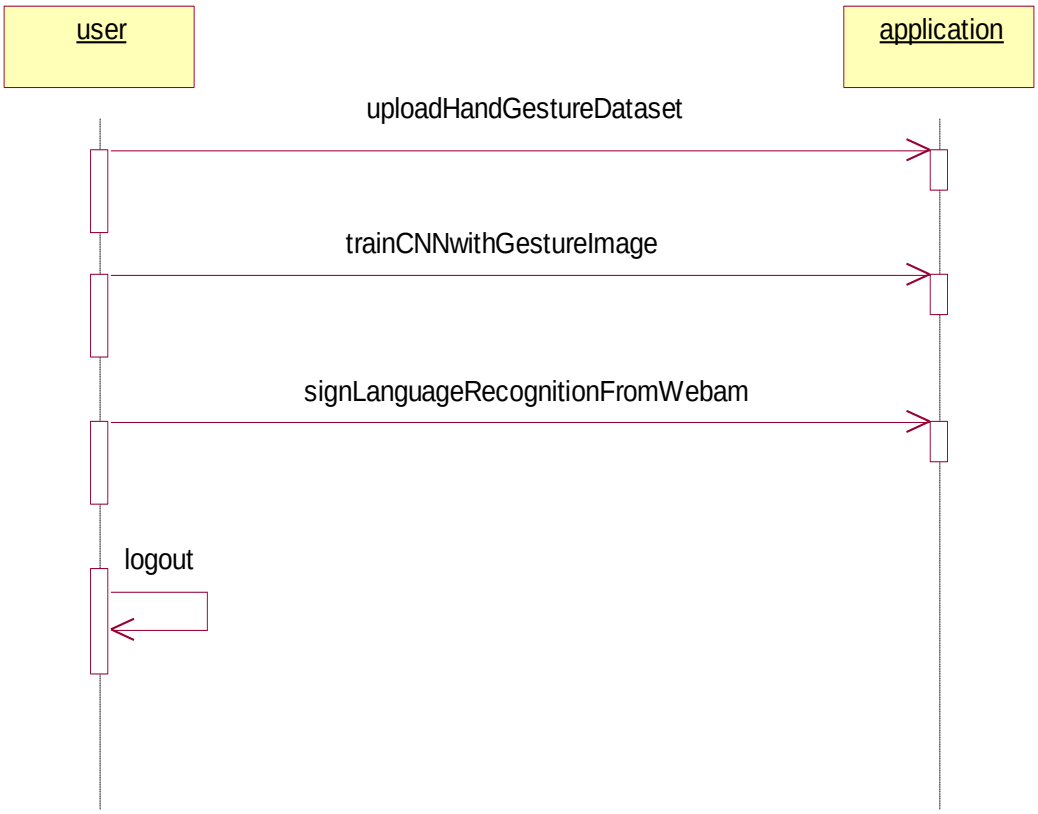


Fig.5.6 Sequence diagram

A sequence diagram is a type of interaction diagram because it describes how—and in what order—a group of objects works together. These diagrams are used by software developers and business professionals to understand requirements for a new system or to document an existing process. Sequence diagrams are sometimes known as event diagrams or event scenarios.

COLLABRATION DIAGRAM :

The collaboration diagram is used to show the relationship between the objects in a system. Both the sequence and the collaboration diagrams represent the same information but differently. Instead of showing the flow of messages, it depicts the architecture of the object residing in the system as it is based on object-oriented programming. An object consists of several features. Multiple objects present in the system are connected to each other. The collaboration diagram, which is also known as a communication diagram, is used to portray the object's architecture in the system.

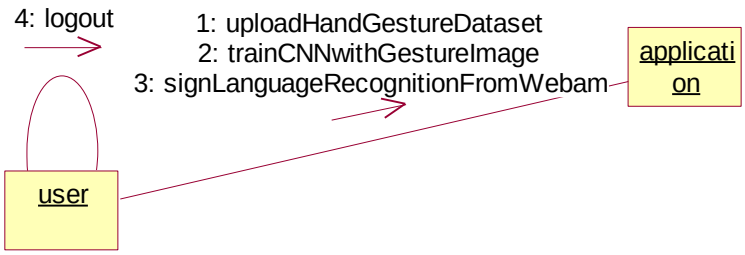


Fig.5.7 Collabration diagram

Unlike a sequence diagram, a collaboration diagram shows the relationships among the objects. Sequence diagrams and collaboration diagrams express similar information but show it in different ways. Because of the format of the collaboration diagram, they tend to better suit for analysis activities (see Activity: Use-Case Analysis). Specifically, they tend to be better suited to depicting simpler interactions of smaller numbers of objects. However, if the number of objects and messages grows, the diagram becomes increasingly hard to read. In addition, it is difficult to show additional descriptive information such as timing, decision points, or other unstructured information that can be easily added to the notes in a sequence diagram.

ACTIVITY DIAGRAM :

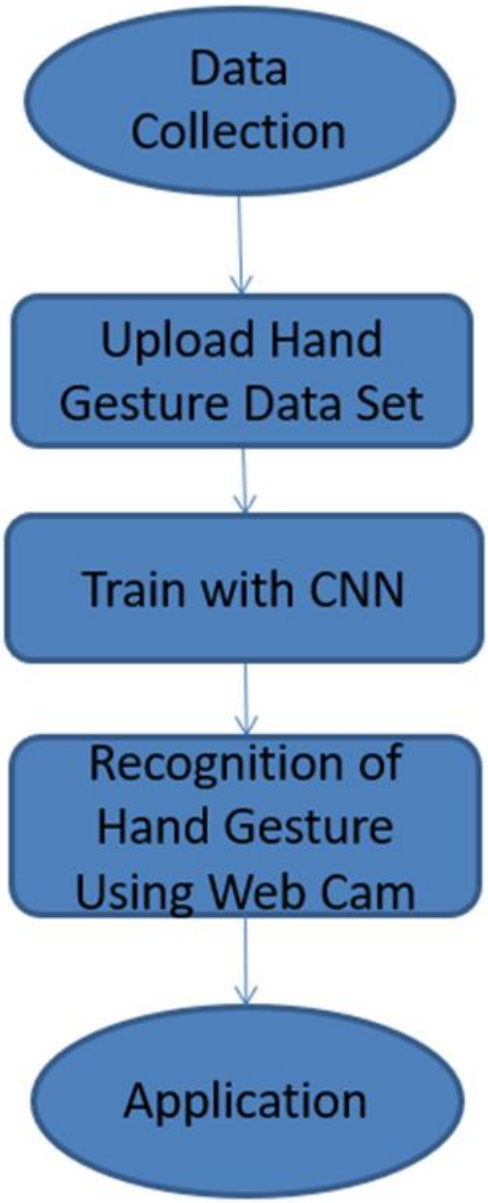


Fig.5.8 Activity Diagram

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modelling Language, activity diagrams are intended to model both computational and organizational processes (i.e., workflows), as well as the data flows intersecting with the related activities. Although activity diagrams primarily show the overall flow of control, they can also include elements showing the flow of data between activities through one or more data stores

Activity Diagrams describe how activities are coordinated to provide a service which can be at different levels of abstraction. Typically, an event needs to be achieved by some operations, particularly where the

operation is intended to achieve a number of different things that require coordination, or how the events in a single use case relate to one another, in particular, use cases where activities may overlap and require coordination. It is also suitable for modelling how a collection of use cases coordinates to represent business workflows. 1. Identify candidate use cases, through the examination of business workflows. 2. Identify pre- and post-conditions (the context) for use cases. 3. Model workflows between/within use cases. 25 4. Model complex workflows in operations on objects. 5. Model in detail complex activities in a high-level activity Diagram

DEPLOYMENT DIAGRAM :

A deployment diagram in the Unified Modelling Language models the physical deployment of artifacts on nodes. [1] To describe a web site, for example, a deployment diagram would show what hardware components ("nodes") exist (e.g., a web server, an application server, and a database server), what software components ("artifacts") run on each node (e.g., web application, database), and how the different pieces are connected (e.g. JDBC, REST, RMI). The nodes appear as boxes, and the artifacts allocated to each node appear as rectangles within the boxes.

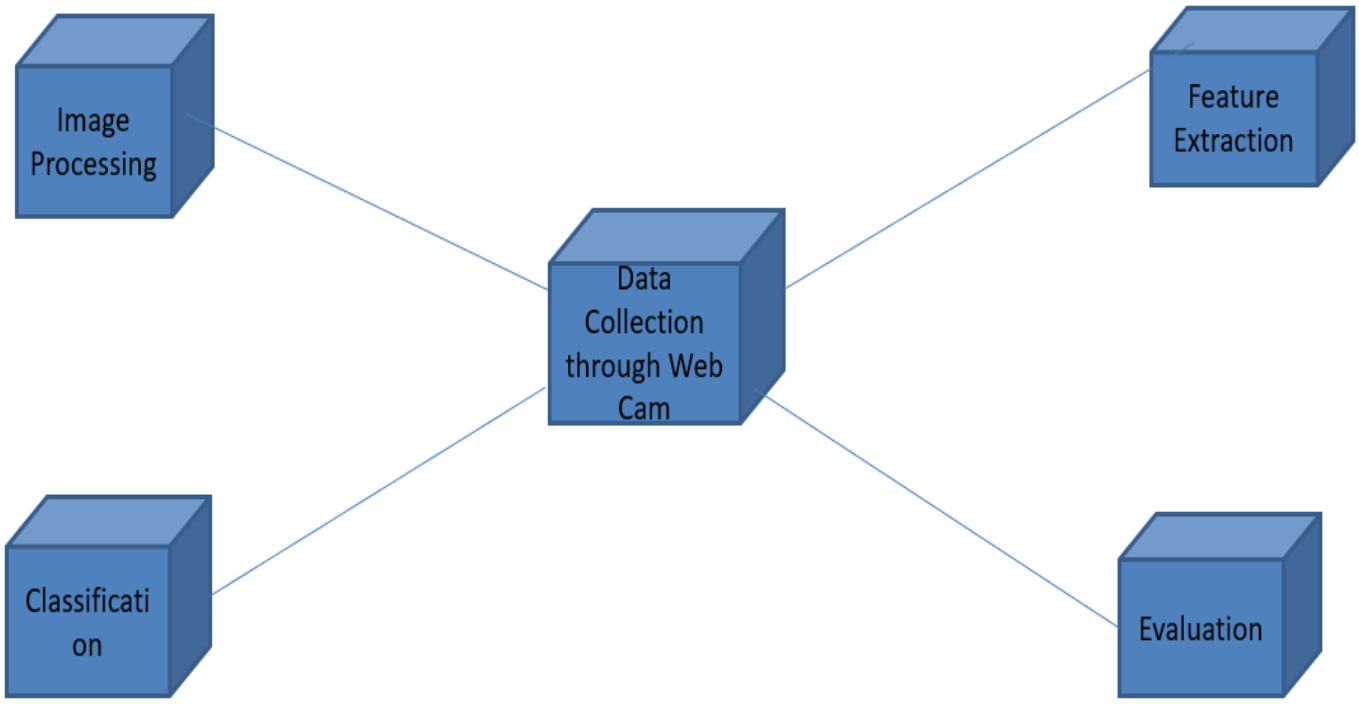


Fig.5.9 Deployment Diagram

Nodes may have sub nodes, which appear as nested boxes. A single node in a deployment diagram may conceptually represent multiple physical nodes, such as a cluster of database servers. There are two types of Nodes: 1. Device Node 2. Execution Environment Node Device nodes are physical computing

resources with processing memory and services to execute software, such as typical computers or mobile phones. An execution environment node (EEN) is a software computing resource that runs within an outer node and which itself provides a service to host and execute other executable software elements

CHAPTER 6
PROJECT CODING

6.1 CODE TEMPLATES

```
def trainCNN():
    global classifier
    text.delete('1.0', END)
    X_train = np.load('model1/X.txt.npy')
    Y_train = np.load('model1/Y.txt.npy')
    text.insert(END,"CNN is training on total images : "+str(len(X_train))+"\n")


    if os.path.exists('model1/model.json'):
        with open('model1/model.json', "r") as json_file:
            loaded_model_json = json_file.read()
            classifier = model_from_json(loaded_model_json)
        classifier.load_weights("model1/model_weights.h5")
        classifier._make_predict_function()
        print(classifier.summary())
        f = open('model1/history.pckl', 'rb')
        data = pickle.load(f)
        f.close()
        acc = data['accuracy']
        accuracy = acc[9] * 100
        text.insert(END,"CNN Hand Gesture Training Model Prediction Accuracy = "+str(accuracy))
    else:
        classifier = Sequential()
        classifier.add(Convolution2D(32, 3, 3, input_shape = (64, 64, 3), activation = 'relu'))
        classifier.add(MaxPooling2D(pool_size = (2, 2)))
        classifier.add(Convolution2D(32, 3, 3, activation = 'relu'))
```

```
classifier.add(MaxPooling2D(pool_size = (2, 2)))

classifier.add(Flatten())

classifier.add(Dense(output_dim = 256, activation = 'relu'))

classifier.add(Dense(output_dim = 5, activation = 'softmax'))

print(classifier.summary())

classifier.compile(optimizer = 'adam', loss = 'categorical_crossentropy', metrics = ['accuracy'])

hist = classifier.fit(X_train, Y_train, batch_size=16, epochs=10, shuffle=True, verbose=2)

classifier.save_weights('model1/model_weights.h5')

model_json = classifier.to_json()

with open("model1/model.json", "w") as json_file:

    json_file.write(model_json)

f = open('model1/history.pckl', 'wb')

pickle.dump(hist.history, f)

f.close()

f = open('model1/history.pckl', 'rb')

data = pickle.load(f)

f.close()

acc = data['accuracy']

accuracy = acc[9] * 100

text.insert(END,"CNN Hand Gesture Training Model Prediction Accuracy = "+str(accuracy))
```

6.2 OUTLINE FOR VARIOUS FILES

Model1 file :

1. history.pckl
2. model
3. model_weight.h5
4. X.txt.npy
5. Y.txt.npy

6.3 CLASS WITH FUNCTIONALITY

LABLES

font = ('times', 16, 'bold')

title = Label(main, text='Sign Language Recognition to Text & Voice using CNN Advance',anchor=W, justify=CENTER)

title.config(bg='yellow4', fg='white')

title.config(font=font)

title.config(height=3, width=120)

title.place(x=0,y=5)

UPLOAD HAND GESTURE DATASET BUTTON :

font1 = ('times', 13, 'bold')

upload = Button(main, text="Upload Hand Gesture Dataset", command=uploadDataset)

upload.place(x=50,y=100)

upload.config(font=font1)

TRAIN CNN WITH GESTURE IMAGES BUTTON :

markovButton = Button(main, text="Train CNN with Gesture Images", command=trainCNN)

```
markovButton.place(x=50,y=200)

markovButton.config(font=font1)
```

SIGN LANGUAGE RECOGNITION FROM WEBCAM BUTTON :

```
predictButton = Button(main, text="Sign Language Recognition from Webcam",
command=webcamPredict)

predictButton.place(x=50,y=250)

predictButton.config(font=font1)
```

6.4 METHODS INPUT AND OUTPUT PARAMETERS.

```
def webcamPredict():
    global playcount
    oldresult = 'none'
    count = 0
    fgbg2 = cv2.createBackgroundSubtractorKNN();
    aWeight = 0.5
    camera = cv2.VideoCapture(0)
    top, right, bottom, left = 10, 350, 325, 690
    num_frames = 0
    while(True):
        (grabbed, frame) = camera.read()
        frame = imutils.resize(frame, width=700)
        frame = cv2.flip(frame, 1)
        clone = frame.copy()
        (height, width) = frame.shape[:2]
        roi = frame[top:bottom, right:left]
        gray = cv2.cvtColor(roi, cv2.COLOR_BGR2GRAY)
```

```
gray = cv2.GaussianBlur(gray, (41, 41), 0)
if num_frames < 30:
    run_avg(gray, aWeight)
else:
    temp = gray
    hand = segment(gray)
    if hand is not None:
        (thresholded, segmented) = hand
        cv2.drawContours(clone, [segmented + (right, top)], -1, (0, 0, 255))
        #cv2.imwrite("test.jpg",temp)
        #cv2.imshow("Thesholded", temp)
        #ret, thresh = cv2.threshold(temp, 150, 255, cv2.THRESH_BINARY + cv2.THRESH_OTSU)
        #thresh = cv2.resize(thresh, (64, 64))
        #thresh = np.array(thresh)
        #img = np.stack((thresh,)*3, axis=-1)
        roi = frame[top:bottom, right:left]
        roi = fgbg2.apply(roi);
        cv2.imwrite("test.jpg",roi)
        #cv2.imwrite("newDataset/Fist/"+str(count)+".png",roi)
        #count = count + 1
        #print(count)
        img = cv2.imread("test.jpg")
        img = cv2.resize(img, (64, 64))
        img = img.reshape(1, 64, 64, 3)
        img = np.array(img, dtype='float32')
        img /= 255
```

```
            predict = classifier.predict(img)
            value = np.amax(predict)
            cl = np.argmax(predict)
            result = names[np.argmax(predict)]
            if value >= 0.99:
                print(str(value)+" "+str(result))
                cv2.putText(clone, 'Gesture Recognize as : '+str(result), (10, 25),
cv2.FONT_HERSHEY_SIMPLEX,0.5, (0, 255, 255), 2)
                if oldresult != result:
                    play(playcount,result)
                    oldresult = result
                    playcount = playcount + 1
            else:
                cv2.putText(clone, '', (10, 25),  cv2.FONT_HERSHEY_SIMPLEX,0.5, (0, 255, 255), 2)
            cv2.imshow("video frame", roi)
        cv2.rectangle(clone, (left, top), (right, bottom), (0,255,0), 2)
        num_frames += 1
        cv2.imshow("Video Feed", clone)
        keypress = cv2.waitKey(1) & 0xFF
        if keypress == ord("q"):
            break
    camera.release()
    cv2.destroyAllWindows()
```

CHAPTER 7
PROJECT TESTING

7.PROJECT TESTING

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub-assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of tests. Each test type addresses a specific testing requirement.

Software Testing is a method to check whether the actual software product matches expected requirements and to ensure that software product is Defect free. It involves execution of software/system components using manual or automated tools to evaluate one or more properties of interest. The purpose of software testing is to identify errors, gaps or missing requirements in contrast to actual requirements.

Some prefer saying Software testing definition as a White Box and Black Box Testing. In simple terms, Software Testing means the Verification of Application Under Test (AUT). This Software Testing course introduces testing software to the audience and justifies the importance of software testing.

It depends on the process and the associated stakeholders of the project(s). In the IT industry, large companies have a team with responsibilities to evaluate the developed software in context of the given requirements. Moreover, developers also conduct testing which is called Unit Testing. In most cases, the following professionals are involved in testing a system within their respective capacities:

- Software Tester
- Software Developer
- Project Lead/Manager
- End User

Different companies have different designations for people who test the software on the basis of their experience and knowledge such as Software Tester, Software Quality Assurance Engineer, QA Analyst, etc. It is not possible to test the software at any time during its cycle. The next two sections state when testing should be started and when to end it during the SDLC.

An early start to testing reduces the cost and time to rework and produce error-free software that is delivered to the client. However, in Software Development Life Cycle (SDLC), testing can be started from the Requirements Gathering phase and continued till the deployment of the software. It also depends on the development model that is being used. For example, in the Waterfall model, formal testing is conducted in the testing phase; but in the incremental model, testing is performed at the end of every increment/iteration and the whole application is tested at the end.

7.1 VARIOUS TEST CASES

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

TYPES OF TESTS

Unit testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

Integration testing

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

Functional test

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items:

Valid Input : identified classes of valid input must be accepted.

Invalid Input : identified classes of invalid input must be rejected.

Functions : identified functions must be exercised.

Output : identified classes of application outputs must be exercised.

Systems/Procedures : interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

System Test

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

Unit Testing

Unit testing is usually conducted as part of a combined code and unit test phase of the software lifecycle, although it is not uncommon for coding and unit testing to be conducted as two distinct phases.

Test strategy and approach

Field testing will be performed manually and functional tests will be written in detail.

Test objectives

- All field entries must work properly.
- Pages must be activated from the identified link.
- The entry screen, messages and responses must not be delayed.

Features to be tested

- Verify that the entries are of the correct format
- No duplicate entries should be allowed
- All links should take the user to the correct page.

Integration Testing

Software integration testing is the incremental integration testing of two or more integrated software components on a single platform to produce failures caused by interface defects.

The task of the integration test is to check that components or software applications, e.g. components in a software system or – one step up – software applications at the company level – interact without error.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

Acceptance Testing

User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements.

7.2 BLACK BOX

Black box testing is a technique of software testing which examines the functionality of software without peering into its internal structure or coding. The primary source of black box testing is a specification of requirements that is stated by the customer.

In this method, tester selects a function and gives input value to examine its functionality, and checks whether the function is giving expected output or not. If the function produces correct output, then it is passed in testing, otherwise failed. The test team reports the result to the development team and then tests the next function. After completing testing of all functions if there are severe problems, then it is given back to the development team for correction.

Generic steps of black box testing

- The black box test is based on the specification of requirements, so it is examined in the beginning.
- In the second step, the tester creates a positive test scenario and an adverse test scenario by selecting valid and invalid input values to check that the software is processing them correctly or incorrectly.
- In the third step, the tester develops various test cases such as decision table, all pairs test, equivalent division, error estimation, cause-effect graph, etc.
- The fourth phase includes the execution of all test cases.
- In the fifth step, the tester compares the expected output against the actual output.
- In the sixth and final step, if there is any flaw in the software, then it is cured and tested again

7.3 WHITE BOX TESTING

The box testing approach of software testing consists of black box testing and white box testing. We are discussing here white box testing which also known as glass box is testing, structural testing, clear box testing, open box testing and transparent box testing. It tests internal coding and infrastructure of a software

focus on checking of predefined inputs against expected and desired outputs. It is based on inner workings of an application and revolves around internal structure testing. In this type of testing programming skills are required to design test cases. The primary goal of white box testing is to focus on the flow of inputs and outputs through the software and strengthening the security of the software.

The term 'white box' is used because of the internal perspective of the system. The clear box or white box or transparent box name denote the ability to see through the software's outer shell into its inner workings.

Developers do white box testing. In this, the developer will test every line of the code of the program. The developers perform the White-box testing and then send the application or the software to the testing team, where they will perform the black box testing and verify the application along with the requirements and identify the bugs and sends it to the developer. The developer fixes the bugs and does one round of white box testing and sends it to the testing team. Here, fixing the bugs implies that the bug is deleted, and the particular feature is working fine on the application.

CHAPTER 8

OUTPUT SCREENS

8.1 USER INTERFACES

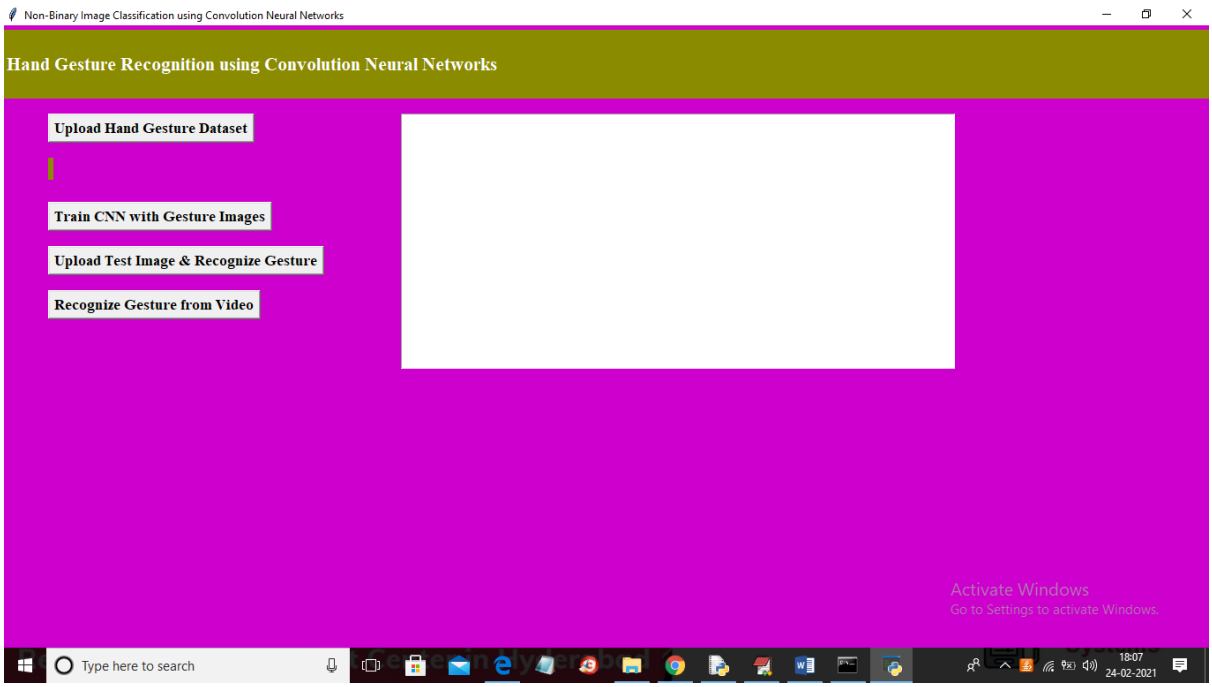


Fig. 8.1: Uploading Hand Gesture Dataset

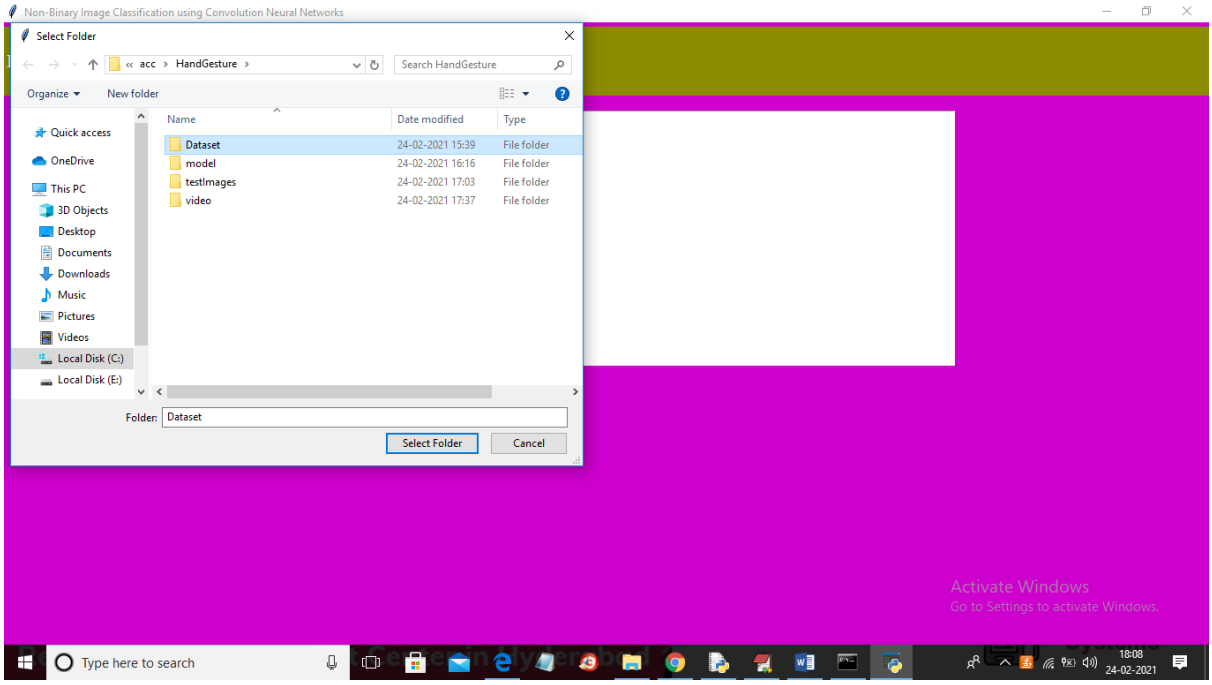


Fig. 8.2: Selecting and uploading Dataset folder.

8.2 OUTPUT SCREENS

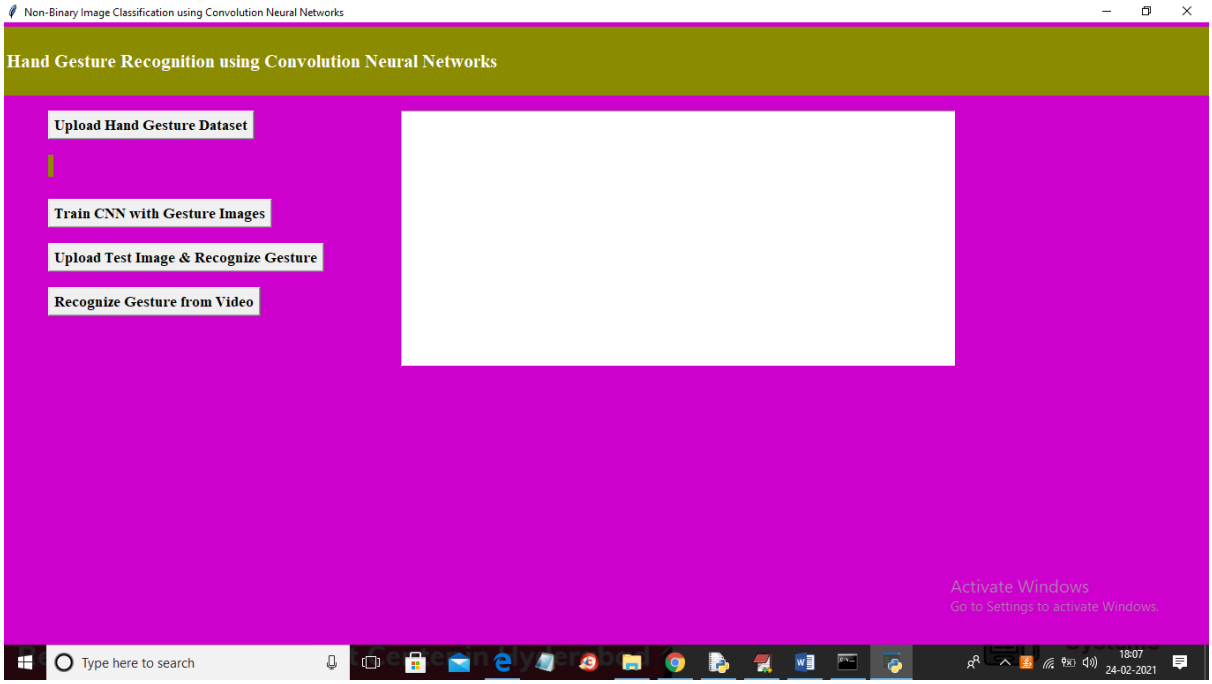


Fig. 8.3: Training CNN with Gesture Images

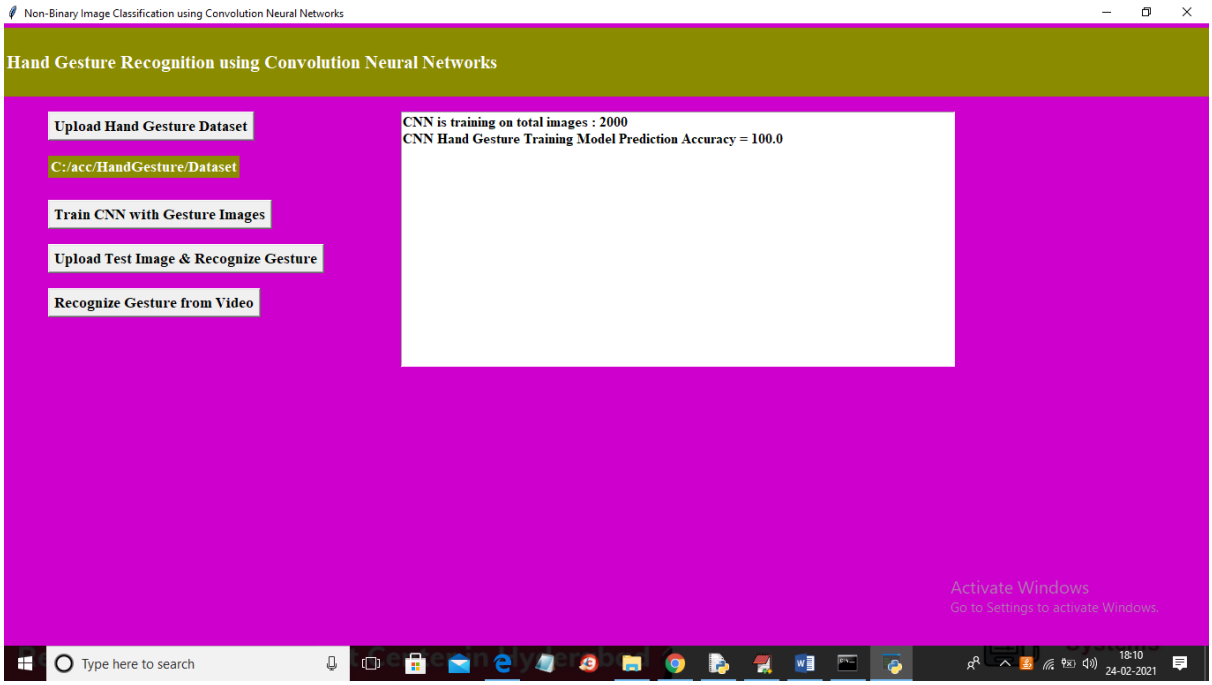


Fig. 8.4: Prediction accuracy

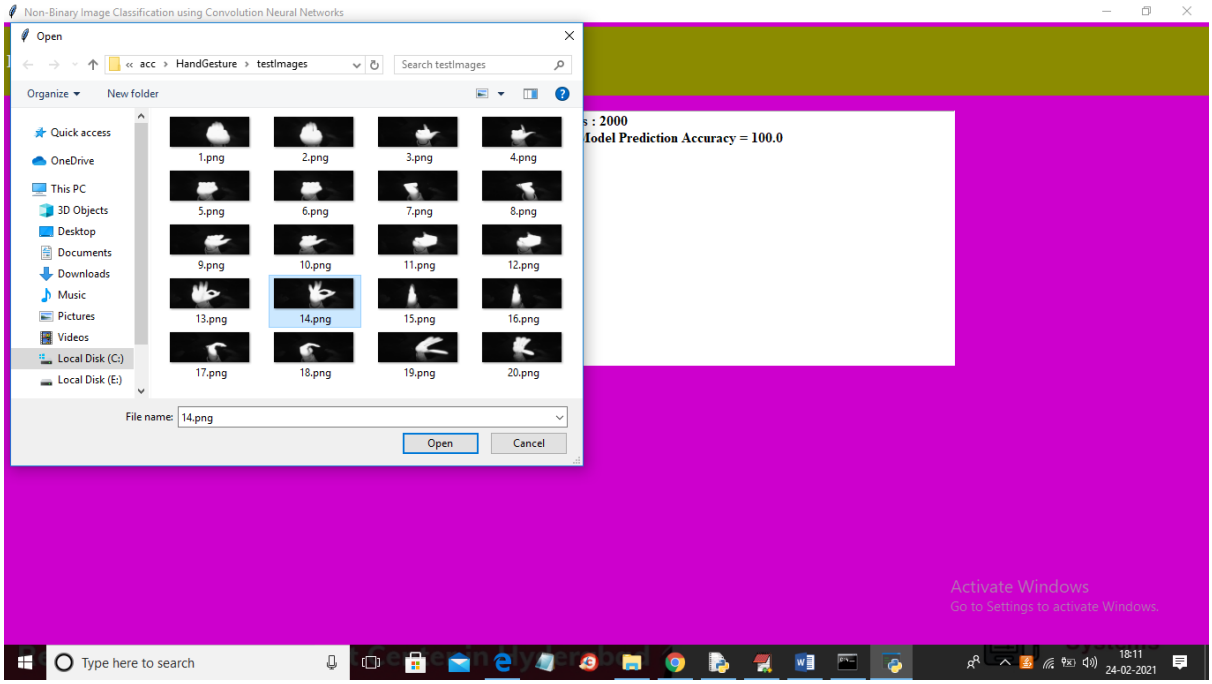


Fig. 8.5: Selecting and uploading

CHAPTER 9

EXPERIMENTAL RESULTS

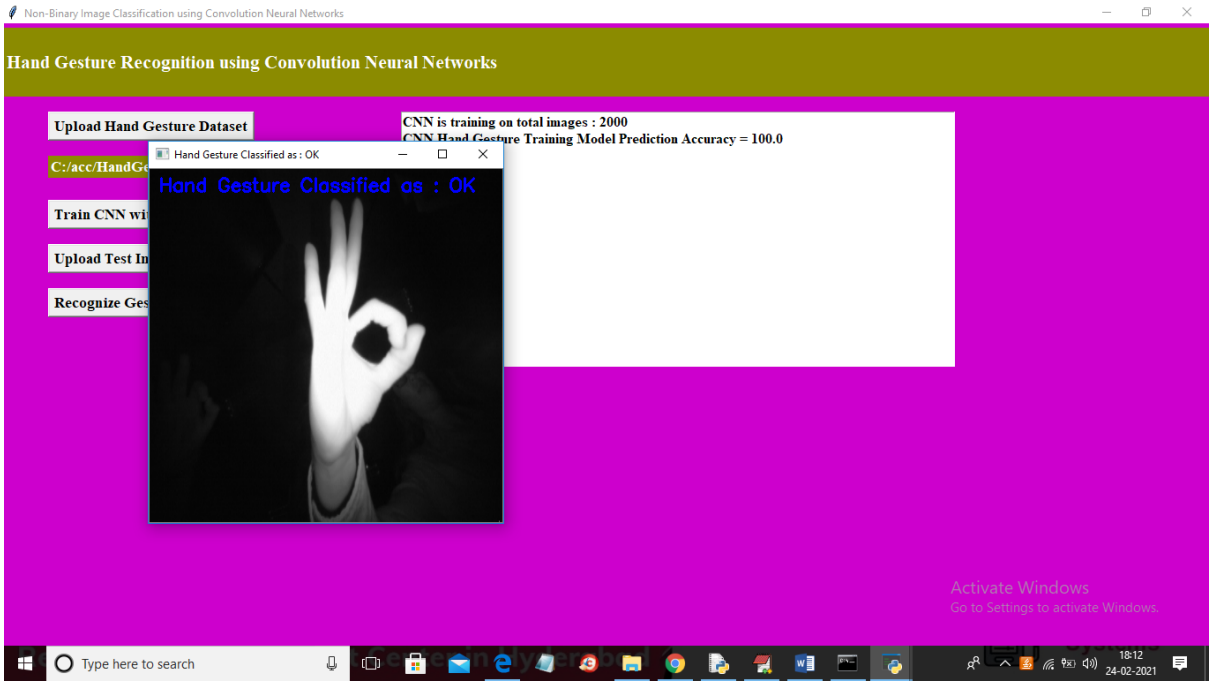


Fig. 9.1: Recognize Gesture from Video.

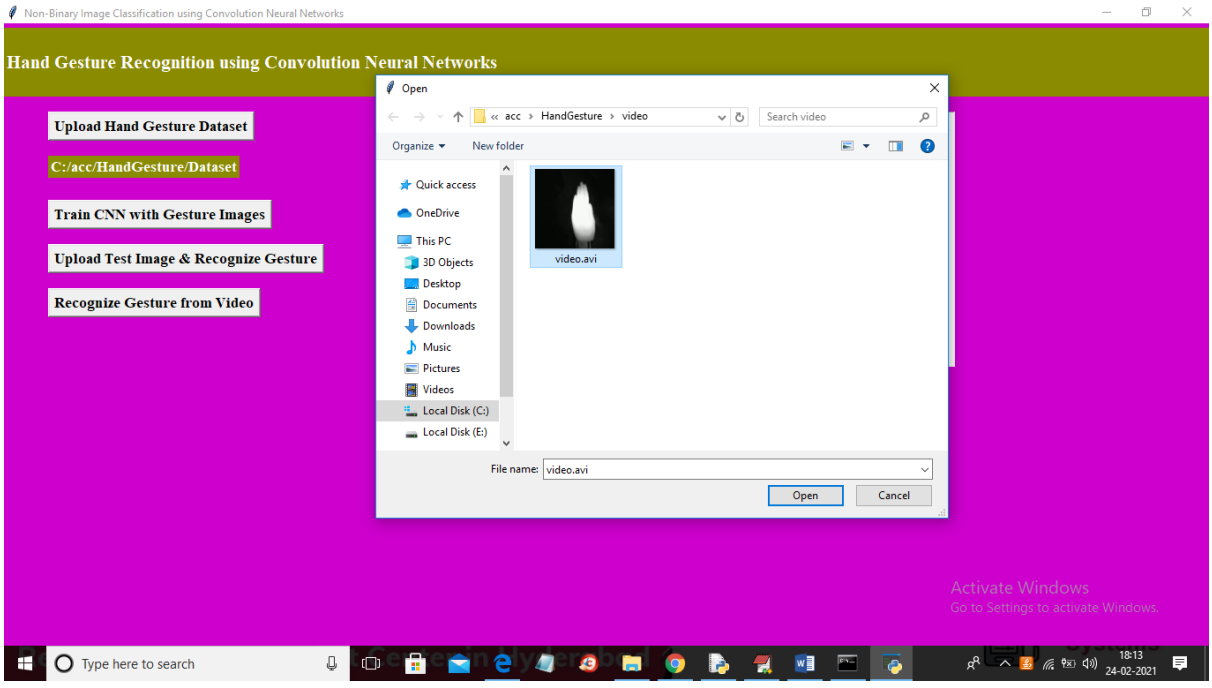


Fig. 9.2 : Selecting and uploading video.avi

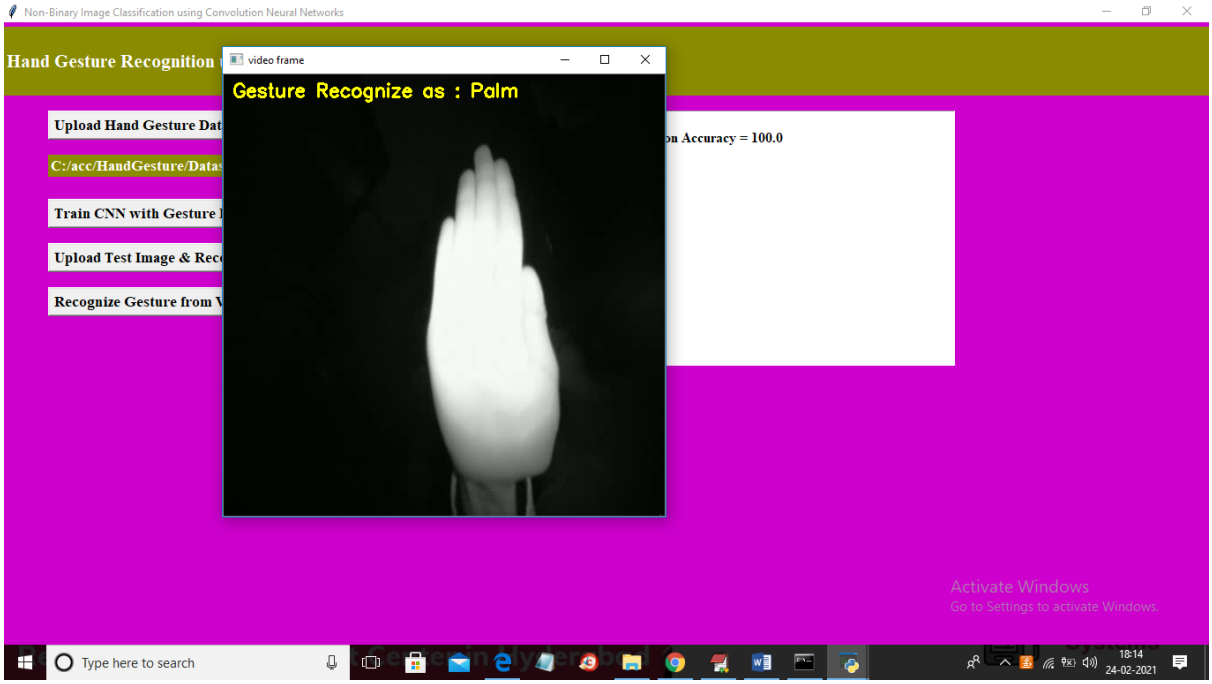


Fig. 9.3 : Gesture Recognize as Palm.

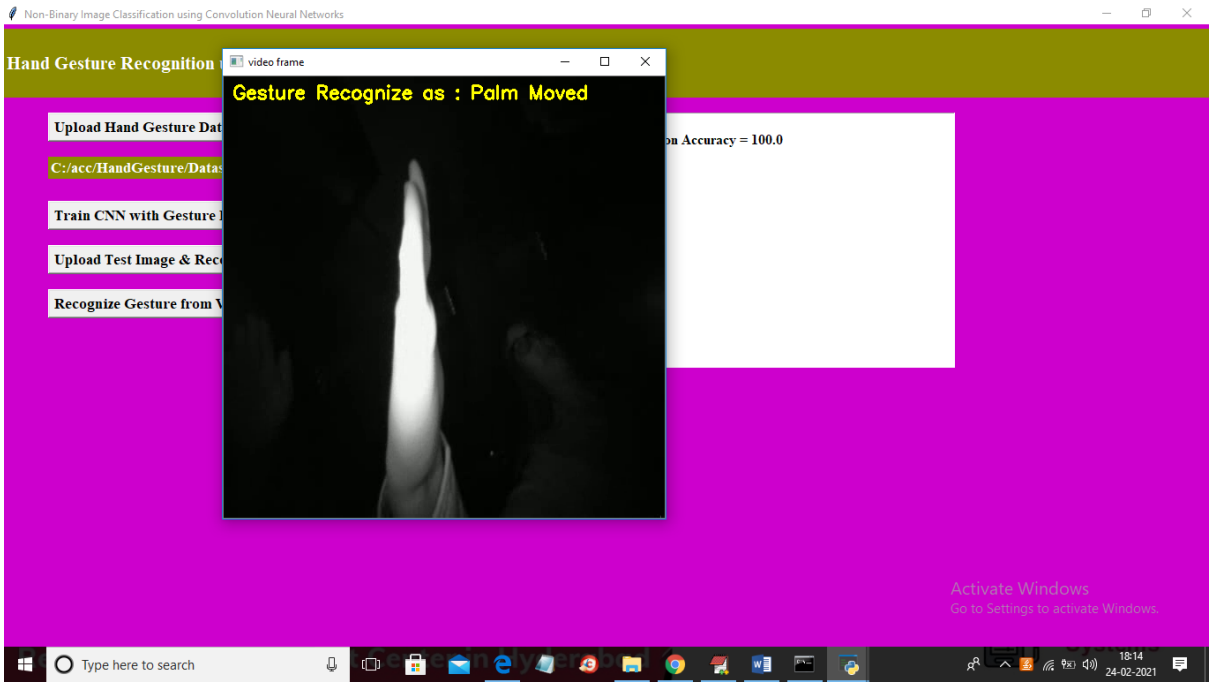


Fig. 9.4 : Gesture Recognize as Palm moved

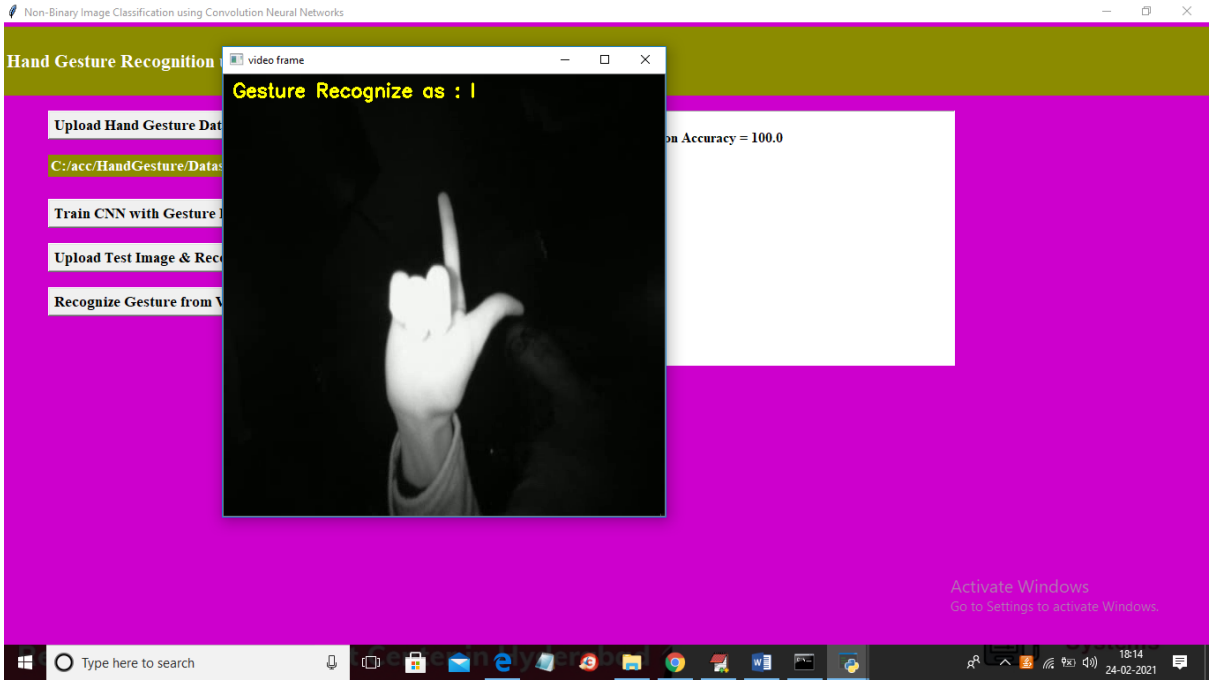


Fig. 9.5 : Gesture Recognize as I

CONCLUSION

Many breakthroughs have been made in the field of artificial intelligence, machine learning and computer vision.They have immensely contributed in how we perceive things around us and improve the way in which we apply their techniques in our everyday lives. Many researches have been conducted on sign gesture recognition using different techniques like ANN, LSTM and 3D CNN. However, most of them require extra computing power . On the other hand, our project requires low computing power and gives a remarkable accuracy of above 90%.We proposed to normalize and rescale our images to 64 pixels in order to extract features (binary pixels) and make the system more robust. We use CNN to classify the 10 alphabetical American sign gestures and successfully achieve an accuracy of 98%.

FUTURE ENHANCEMENT

We look forward to use more alphabets in our datasets and improve the model so that it recognizes more alphabetical features. while at the same time get a high accuracy. We would also like to enhance the system by adding speech recognition so that blind people can benefit as well.

REFERENCES

1. Rastgoo, R., Kiani, K., & Escalera, S. (2020). Hand sign language recognition using multi-view hand skeleton. *Expert Systems with Applications*, *150*, 113336.
2. Bird, J. J., Ekárt, A., & Faria, D. R. (2020). British sign language recognition via late fusion of computer vision and leap motion with transfer learning to american sign language. Sensors, 20(18), 5151.
3. Wadhawan, A., & Kumar, P. (2020). Deep learning-based sign language recognition system for static signs. Neural Computing and Applications, 32(12), 7957-7968.
4. Deshpande, A. M., & Kalbhor, S. R. (2020). Video-based marathi sign language recognition and text conversion using convolutional neural network. In Emerging Trends in Electrical, Communications, and Information Technologies (pp. 761-773). Springer, Singapore.
5. Rastgoo, R., Kiani, K., & Escalera, S. (2020). Video-based isolated hand sign language recognition using a deep cascaded model. *Multimedia Tools and Applications*, *79*, 22965-22987.
6. Albanie, S., Varol, G., Momeni, L., Afouras, T., Chung, J. S., Fox, N., & Zisserman, A. (2020, August). BSL-1K: Scaling up co-articulated sign language recognition using mouthing cues. In *European Conference on Computer Vision* (pp. 35-53). Springer, Cham.
7. Parelli, M., Papadimitriou, K., Potamianos, G., Pavlakos, G., & Maragos, P. (2020, August). Exploiting 3D hand pose estimation in deep learning-based sign language recognition from RGB videos. In *European Conference on Computer Vision* (pp. 249-263). Springer, Cham.
8. Saggio, G., Cavallo, P., Ricci, M., Errico, V., Zea, J., & Benalcázar, M. E. (2020). Sign language recognition using wearable electronics: implementing k-nearest neighbors with dynamic time warping and convolutional neural network algorithms. *Sensors*, *20*(14), 3879.
9. De Coster, M., Van Herreweghe, M., & Dambre, J. (2020). Sign language recognition with transformer networks. In *12th International Conference on Language Resources and Evaluation*.
10. Wangchuk, K., Riyamongkol, P., & Waranusast, R. (2020, October). Bhutanese Sign Language Alphabets Recognition Using Convolutional Neural Network. In *2020-5th International Conference on Information Technology (InCIT)* (pp. 44-49). IEEE.
11. Rathi, P., Kuwar Gupta, R., Agarwal, S., & Shukla, A. (2020). Sign Language Recognition Using ResNet50 Deep Neural Network Architecture. *Available at SSRN 3545064*.
12. Lim, Z. J. (2020). *Computer Vision Based Sign Language Recognition System* (Doctoral dissertation, Tunku Abdul Rahman University College).
13. Jiang, X. (2020). Isolated Chinese sign language recognition using gray-level Co-occurrence Matrix and parameter-optimized Medium Gaussian support vector machine. In *Frontiers in*

Intelligent Computing: Theory and Applications (pp. 182-193). Springer, Singapore.

14. Camgoz, N. C., Koller, O., Hadfield, S., & Bowden, R. (2020). Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10023-10033).
15. Rastgoo, R., Kiani, K., & Escalera, S. (2020). Sign language recognition: A deep survey. *Expert Systems with Applications*, 113794.

PUBLICATIONS

1. Presented paper in Online Mega International Conference “Innovations In Computers Networks, Computational Intelligence and IOT” (ICICCI-21) titled “Sign Language Recognition System Using Neural Network and Computer Vision",

2. Paper Accepted in UGC Approved Journal

I am Fouzia Begum, pursuing my Bachelor of Technology in the stream of Computer Science and Engineering from St. Martin's Engineering College. I have done my Board of Intermediate from Sri chaitanya junior college and SSC from Triveni talent school. I do have many Leadership qualities with good communication skills. My technical skills include C, C++, Java, Python, and basic understanding in Web Development.My participation in technical workshop include five days exclusively for women online workshop on "**Women in Cyber Security and Privacy in 2020**" from 6th to 10th july . I have also completed various certification courses like "Managing project risks and changes", "AWS fundamentals", **"**AI for everyone", "Leadership and emotional intellegence", and many more from professional Platforms like Coursera . My areas of interests include Cybersecurity, Networking and Webdevelopment Technologies etc. I have also got offers from "**Altruista health inc"**

My name is Devanaka Preethi, currently I am pursuing my Bachelor of Technology in the stream of Computer Science and Engineering from St. Martin's Engineering College. I have done my Board of Intermediate from Narayana Junior College and SSC from Sri Nayaki Model High School. My technical skills include C++, Python, and basic understanding in Java. I am one of the members who got shortlisted and trained in the Employment Skill Development Program provided by Zensar Technologies. My participation in technical workshops include five days exclusively for women online workshop on "**Women in Cyber Security and Privacy in 2020**" from 6th to 10th july 2020 . I have Successfully completed one month internship on "**Machine Learning**" by codemania . I have also completed various certification courses like "AI by crashcourse", "MYSQL databases by the newboston", "Ethical hacking from scratch", "Javascript by net ninja", and many more from professional Platforms like Coursera and Udemy. My areas of interests include Data science, Machine Learning, Artificial Intelligence and IOT Technology. I have also got offer from "**Altruisa health Inc**".

My name **is G.Chetna Varma.**I am currently pursuing Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. I completed intermediate from Nano Junior College and 10th class from Sri Chaitanya High School. My technical skills include C, Python,C++ and Java. .My participations are: five days exclusively for women online workshop on "**Women in Cyber Security and Privacy in 2020**" from 6th to 10th july 2020 and National Level Two Day seminar on "Recent Trends in Cloud Computing, Fog and Edge Computing" which was conducted from 18th to 19th June 2021. I have Successfully completed Two months Internship Program organized by Computer Society of India & Global Cyber Security Forum and received course completion certificate on "Python". I have also completed various certification courses like "AWS", "MYSQL database by Thenewboston", "Java Script by net ninja" "3D modeling with blender by blender sensei" and many more from professional Platforms like, Cursa, Coursera and Udemy. My areas of interest are Python, Artificial Intelligence, Machine Learning and Deep Learning, web development. I have also got offer from "**NNIIT**".

My name is Jukanti Nishitha, currently I am pursuing my Bachelor of Technology in the stream of Computer Science and Engineering from St. Martin's Engineering College. I have done my Board of Intermediate from Narayana Junior College and SSC from Dilsukhnagar Public School. My technical skills include C, C++, Python, and basic understanding in Java. I am one of the member in Smart Interviews. My participation in technical workshops include five days exclusively for women online workshop on "**Women in Cyber Security and Privacy in 2020**" from 6th to 10th july 2020 . I have Successfully completed one month internship on "**Machine Learning**" by codemania . I have also completed various certification courses like "AI by crashcourse", "MYSQL databases by the newboston", "Ethical hacking from scratch", "Javascript by net ninja", and many more from professional Platforms like Coursera and Udemy. My areas of interests include Web Development, Machine Learning, Artificial Intelligence and IOT Technology. I have also got offer from "**TCS**" , "**ACCENTURE**" and "**WIPRO**".

A

PROJECT REPORT

On

CREDIT CARD FRAUD DETECTION USING PREDICTIVE MODELLING

Submitted by

1)D.Grishma(17K81A0511) **2)K.Swathi (17K81A0532)**

3)J.Supriya (17K81A0525) **4)T.Satwika(17K81A0554)**

in partial fulfillment for the award of the

degree of

BACHELOR OF TECHNOLOGY

IN

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Under the Guidance of

Dr.T.Poongothai

(B.E.,M.E.,Ph.D)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ST.MARTIN'S ENGINEERING COLLEGE
An Autonomous Institute

Dhulapally, Secunderabad – 500 100

JUNE 2021

BONAFIDE CERTIFICATE

This is to certify that the project entitled **CREDIT CARD FRAUD DETECTION USING PREDICTIVE MODELLING** submitted by **Ms. D. Grishma (17K81A0511), Ms. K. Swathi (17K81A0532), Ms. J. Supriya (17K81A0525), Ms. T. Satwika (17K81A0554)** in partial fulfillment of the requirement for the award of the degree of **BACHELOR OF TECHNOLOGY IN Computer Science And Engineering** is recorded of bonafide work carried out by them. The result embodied in this report have been verified and found satisfactory.

Dr.T.Poongothai
(B.E.,M.E.,Ph.D)
Department of CSE

Head of the Department
Dr.M.NARAYANAN
Department of CSE

Internal Examiner

External Examiner

Place:

Date:

DECLARATION

We, the student of **Bachelor of Technology** in Department of Computer Science and Engineering', session: 2017 – 2021, St. Martin's Engineering College, Dhulapally, Kompally, Secunderabad, hereby declare that work presented in this Project Work entitled **Credit Card Fraud Detection Using Predective Modelling** is the outcome of our own bonafide work and is correct to the best of our knowledge and this work has been undertaken taking care of Engineering Ethics. This result embodied in this project report has not been submitted in any university for award of any degree.

D.Grishma(17K81A0511)
K.Swathi(17K81A0532)
J.Supriya(17K81A0525)
T.R.Satwika(17K81A0554)

ABSTRACT

Fraud is a set of illegal activities that are used to take money or property using false pretenses. Transaction fraud using credit card is one of the growing issue in the world of finance. A huge financial loss has significantly affected individuals using credit cards and furthermore vendors and banks. One of the most successful techniques to identify such fraud is Machine learning. This project proposes a fraud detection algorithm using Logistic Regression which can help in solving this real world problem. The accuracy of detecting fraud in credit card transaction is increased using this proposed system. To evaluate the model efficiency, a publicly available credit card dataset is used. Then, a real world from a financial institution is analyzed. In addition, we used Random Forest and Decision tree for the collection and representation of data. The credit card fraud detection features uses user behavior and location scanning to check for unusual patterns. These patterns include user characteristics such as user spending patterns as well as usual user geographic locations to verify his identity. The Credit Card Fraud Detection Problem includes modelling past credit card transactions with the data of the ones that turned out to be fraud. This model is then used to recognize whether a new transaction is fraudulent or not. Our target here is to detect 100% of the fraudulent transactions while minimizing the incorrect fraud classifications.

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible and whose encouragements and guidance have crowded effects with success.

We extended our deep sense of gratitude to Principal**, Dr. P. SANTOSH KUMAR PATRA**, St. Martin's Engineering College, Dhullapally, for permitting us to undertake this project.

We are also thankful to **Dr. M.NARAYANAN**, Head of the Department, Computer Science and Engineering, St. Martin's Engineering College, Dhullapally, for his support and guidance throughout our project.

We would like to thank **Dr. T. POONGOTHAI,** Department of Computer Science and Engineering (AI & ML) for her encouragement and insightful comments and as well as our project coordinators **Dr. B.RAJALINGAM**, Associate Professor and **Mr. J.SUDHAKAR**, Assistant Professor, in Department of Computer Science and Engineering for their valuable support.

We would like to express our sincere gratitude and indebtedness to our project supervisor Dr.T.Poongothai(B.E.,M.E.,Ph.D), Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

Finally, we express thanks to all those who have helped us successfully to completing this project. Furthermore, we would like to thank our family and friends for their moral support and encouragement.

We express thanks to all those who have helped us in successfully completing the project.

D.Grishma(17K81A0511)

K.Swathi(17K81A0532)

J.Supriya(17K81A0525)

T.R.satwika(17K81A0554)

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

<AVI>	Audio Video Interlace
<BMP>	Bitmap
<CPU>	Central Processing Unit
<GB>	Giga Bytes
<GUI>	Graphical User Interface

CHAPTER 1

1. INTRODUCTION

Fraud is a wrongful or criminal deception aimed to bring financial or personal gain . In avoiding loss from fraud, two mechanisms can be used: fraud prevention and fraud detection. Fraud prevention is a proactive method, where it stops fraud from happening in the first place. On the other hand, fraud detection is needed when a fraudulent transaction is attempted by a fraudster. Credit card fraud is concerned with the illegal use of credit card information for purchases. Credit card transactions can be accomplished either physically or digitally . In physical transactions, the credit card is involved during the transactions. In digital transactions, this can happen over the telephone or the internet. Cardholders typically provide the card number, expiry date, and card verification number through telephone or website. With the rise of e-commerce in the past decade, the use of credit cards has increased dramatically . The number of credit card transactions in 2011 in Malaysia were at about 320 million, and increased in 2015 to about 360 million. Along with the rise of credit card usage, the number of fraud cases have been constantly increased. While numerous authorization techniques have been in place, credit card fraud cases have not hindered effectively. Fraudsters favour the internet as their identity and location are hidden.

The rise in credit card fraud has a big impact on the financial industry. The global credit card fraud in 2015 reached to a staggering USD $21.84 billion . Loss from credit card fraud affects the merchants, where they bear all costs, including card issuer fees, charges, and administrative charges. Since the merchants need to bear the loss, some goods are priced higher, or discounts and incentives are reduced. Therefore, it is imperative to reduce the loss, and an effective fraud detection system to reduce or eliminate fraud cases is important. There have been various studies on credit card fraud detection. Machine learning and related methods are most commonly used, which include artificial neural networks, rule-induction techniques, decision trees, logistic regression, and support vector machines . These methods are used either standalone or by combining several methods together to form hybrid models. In this paper, a total of twelve machine learning algorithms are used for detecting credit card fraud. The algorithms range from standard neural networks to deep learning models. They are evaluated using both benchmark and realworld credit card data sets. In addition, the Adaboost and majority voting methods are applied for forming hybrid models. To further evaluate the robustness and reliability of the models, noise is added to the real-world data set.

The key contribution of this paper is the evaluation of a variety of machine learning models with a real-world credit card data set for fraud detection.

1.1 PROJECT OVERVIEW

Fraud detection is a critical problem affecting large financial companies that have increased due to the growth in credit card transactions. This paper presents detection of frauds in credit card transactions, using data mining techniques of Predictive modeling, logistic Regression, and Decision Tree. In this paper, we propose to detect credit card transaction using available data set and data mining techniques of predictive modeling, Decision tree, and Logistic Regression. Predictive modeling splits the data into two partitions 70% of testing and 30%of training check output class distribution to predict the outcome. The decision tree to get the result as a tree with root node describes the best predictor in the data, the combination of two or more branches is denoted by decision node (non leaf nodes) and each branch represents a value for the attribute which is tested. The leaf node may be 1 in the case of fraud and 0 otherwise. Logistic regression or logistic model is a regression model, where the dependent variable is categorical of a linear generalized model.The rest of the paper is organized as explained. Section II describes fraud detection methods. Section III explains Dataset description for credit card transaction. Section IV consists of experimental results of fraud detection methods, and finally, the conclusion of this work .

1.2 PROJECT OBJECTIVES

The design of efficient fraud detection algorithms is key for reducing these losses, and more and more algorithms rely on advanced machine learning techniques to assist fraud investigators. The design of fraud detection algorithms is however particularly challenging due to the non-stationary distribution of the data, the highly unbalanced classes distributions and the availability of few transactions labeled by fraud investigators. At the same time public data are scarcely available for confidentiality issues, leaving unanswered many questions about what is the best strategy. In this thesis we aim to provide some answers by focusing on crucial issues such as: i) why and how under sampling is useful in the presence of class imbalance (i.e. frauds are a small percentage of the transactions), ii) how to deal with unbalanced and evolving data streams (non-stationarity due to fraud evolution and change of spending behavior), iii) how to assess performances in a way which is relevant for detection and iv) how to use feedbacks provided by investigators on the fraud alerts generated. Finally, we design and assess a prototype of a Fraud Detection System able to meet real-world working conditions and that is able to integrate investigators’ feedback to generate accurate alerts.

1.3 CHAPTERS OVERVIEW

Chapter 1 : INTRODUCTION

Banks collect a lot of historical records corresponding to millions of customer's transactions. They are credit card and debit card operations, but unfortunately, only a small portion, if any, is open access. Fraud detection is a critical problem affecting large financial companies that have increased due to the growth in credit card transactions [1]. The proposed method consists of the Predictive modeling and Logistic Regression.

Chapter 2 : LITERATURE SURVEY

For credit card fraud detection, Random Forest (RF) and Logistic Regression (LOR) were examined . The data set consisted of one-year transactions. Data under-sampling was used to examine the algorithm performances, with RF demonstrating a better performance as compared with LOR. An Artificial Immune Recognition System (AIRS) for credit card fraud detection was proposed in. AIRS is an improvement over the standard AIS model, where negative selection was used to achieve higher precision. This resulted in an increase of accuracy by25% and reduced system response time by 40% .

Chapter 3 : SOFTWARE AND HARD WARE REQUIREMENTS

The project involved analyzing the design of few applications so as to make the application more users friendly. To do so, it was really important to keep the navigations from one screen to the other well ordered and at the same time reducing the amount of typing the user needs to do. In order to make the application more accessible, the browser version had to be chosen so that it is compatible with most of the Browsers.

Chapter 4 : SOFTWARE DEVELPOMENT ANALYSIS

The development and implementation of the design parameters. Developer's code based on the product specifications and requirements agreed upon in the previous stages. Following company procedures and guidelines, front-end developers build interfaces and back-ends while database administrators create relevant data in the database. The programmers also test and review each other's code.

Chapter 5 : PROJECT SYSTEM DESIGN

Design is the stage of the software development process. Here, architects and developers draw up advanced technical specifications they need to create the software to requirements. Stakeholders will discuss factors such as risk levels, team team composition, applicable technologies, time, budget, project limitations, method and architectural design.

Chapter 6 : PROJECT CODING

A programming project produces a well-designed executing system that solves a specified distributed programming problem. A project code is used to represent a one-time, or intermittent departmental event or activity. Any person can use a project code on a transaction, regardless of the project manager or home organization. This section describes some of the coding templates, outline of various files, class with functionalities, the various methods of input and output parameters.

Chapter 7 : PROJECT TESTING

The testing phase checks the software for bugs and verifies its performance before delivery to users. In this stage, expert testers verify the product's functions to make sure it performs according to the requirements analysis document.Testers use exploratory testing if they have experience with that software or a test script to validate the performance of individual components of the software. They notify developers of defects in the code. If developers confirm the flaws are valid, they improve the program, and the testers repeat the process until the software is free of bugs and behaves according to requirements.

Chapter 8 : OUTPUT SCREENS

The output of the programmed project is being screened with the screenshots. Operations are done with the final input. The various test case results are captured and projected some sample output.

Chapter 9 : EXPERIMENTAL RESULTS

The code prints out the number of false positives it detected and compares it with the actual values. This is used to calculate the accuracy score and precision of the algorithms. The fraction of data we used for faster testing is 10% of the entire dataset. The complete dataset is also used at the end and both the results are printed. These results along with the classification report for each algorithm is given in the output as follows, where class 0 means the transaction was determined to be valid and 1 means it was determined as a fraud transaction. This result matched against the class values to check for false positives.

Chapter 10 : CONCLUSION

This process is used to detect the credit card transaction, which are fraudulent or genuine. Data mining techniques of Predictive modeling, Decision trees and Logistic Regression are used to predict the fraudulent or genuine credit card transaction. In predictive modeling to detect and check output class distribution.

CHAPTER 2

2. LITERATURE SURVEY

2.1 SURVEY ON BACKGROUND

The fraud detection is a complex task and there is no system that correctly predicts any transaction as fraudulent. The properties for a good fraud detection system are:

1. Should identify the frauds accurately.

2. Should detect the frauds quickly.

3. Should not classify a genuine transaction as fraud.

Outlier detection is a critical task as outliers indicate abnormal running conditions from which significant performance degradation may happen. Techniques used in fraud detection can be divided into two:

1) Supervised techniques where past known legitimate/fraud cases are used to build a model which will produce a suspicion score for the new transactions .

2) Unsupervised are those where there are no prior sets in which the state of the transactions are known to be fraud or legitimate .

Fraud is a wrongful or criminal deception aimed to bring financial or personal gain . In avoiding loss from fraud, two mechanisms can be used: fraud prevention and fraud detection. Fraud prevention is a proactive method, where it stops fraud from happening in the first place. On the other hand, fraud detection is needed when a fraudulent transaction is attempted by a fraudster. Credit card fraud is concerned with the illegal use of credit card information for purchases. Credit card transactions can be accomplished either physically or digitally . In physical transactions, the credit card is involved during the transactions. In digital transactions, this can happen over the telephone or the internet. Cardholders typically provide the card number, expiry date, and card verification number through telephone or website. With the rise of e-commerce in the past decade, the use of credit cards has increased dramatically . The number of credit card transactions in 2011 in Malaysia were at about 320 million, and increased in 2015 to about 360 million. Along with the rise of credit card usage, the number of fraud cases have been constantly increased. While numerous authorization techniques have been in place, credit card fraud cases have not hindered effectively. Fraudsters favour the internet as their identity and location are hidden.

The rise in credit card fraud has a big impact on the financial industry. The global credit card fraud in 2015 reached to a staggering USD $21.84 billion . Loss from credit card fraud affects the merchants, where they bear all costs, including card issuer fees, charges, and administrative charges. Since the merchants need to bear the loss, some goods are priced higher, or discounts and incentives are reduced. Therefore, it is imperative to reduce the loss, and an effective fraud detection system to reduce or eliminate fraud cases is important. There have been various studies on credit card fraud detection. Machine learning and related methods are most commonly used, which include artificial neural networks, rule-induction techniques, decision trees, logistic regression, and support vector machines. These methods are used either standalone or by combining several methods together to form hybrid models. In this paper, a total of twelve machine learning algorithms are used for detecting credit card fraud. The algorithms range from standard neural networks to deep learning models. They are evaluated using both benchmark and realworld credit card data sets. In addition, the Adaboost and majority voting methods are applied for forming hybrid models. To further evaluate the robustness and reliability of the models, noise is added to the real-world data set.

The key contribution of this paper is the evaluation of a variety of machine learning models with a real-world credit card data set for fraud detection. While other researchers have used various methods on publicly available data sets, the data set used in this paper are extracted from actual credit card transaction information over three months.For transaction monitoring by bank employees the clustering model was developed. This model allows provision of fast analysis of transactions by attributes.

2.2 CONCLUSION ON SURVEY

This process is used to detect the credit card transaction, which are fraudulent or genuine. Data mining techniques of Predictive modeling, Decision trees and Logistic Regression are used to predict the fraudulent or genuine credit card transaction. In predictive modeling to detect and check output class distribution. The prediction model predicts continuous valued functions. We have to detect 148 may be fraud and other are genuine. In decision tree generate a tree with root node, decision node and leaf nodes. The leaf node may be 1 becomes fraud and 0 otherwise. Logistic Regression is same as linear regression but interpret curve is different. To generalize the linear regression model, when dependent variable is categorical and analyzes relationship between multiple independent variables. Fraud is a wrongful or criminal deception aimed to bring financial or personal gain . In avoiding loss from fraud, two mechanisms can be used: fraud prevention and fraud detection. Fraud prevention is a proactive method, where it stops fraud from happening in

the first place. On the other hand, fraud detection is needed when a fraudulent transaction is attempted by a fraudster. Credit card fraud is concerned with the illegal use of credit card information for purchases. Credit card transactions can be accomplished either physically or digitally. In physical transactions, the credit card is involved during the transactions. In digital transactions, this can happen over the telephone or the internet. Cardholders typically provide the card number, expiry date, and card verification number through telephone or website. With the rise of e-commerce in the past decade, the use of credit cards has increased dramatically. The number of credit card transactions in 2011 in Malaysia were at about 320 million, and increased in 2015 to about 360 million. Along with the rise of credit card usage, the number of fraud cases have been constantly increased. While numerous authorization techniques have been in place, credit card fraud cases have not hindered effectively. Fraudsters favour the internet as their identity and location are hidden. The rise in credit card fraud has a big impact on the financial industry. The global credit card fraud in 2015 reached to a staggering USD $21.84 billion. Loss from credit card fraud affects the merchants, where they bear all costs, including card issuer fees, charges, and administrative charges. Since the merchants need to bear the loss, some goods are priced higher, or discounts and incentives are reduced. Therefore, it is imperative to reduce the loss, and an effective fraud detection system to reduce or eliminate fraud cases is important. There have been various studies on credit card fraud detection. Machine learning and related methods are most commonly used, which include artificial neural networks, rule-induction techniques, decision trees, logistic regression, and support vector machines. These methods are used either standalone or by combining several methods together to form hybrid models. In this paper, a total of twelve machine learning algorithms are used for detecting credit card fraud. The algorithms range from standard neural networks to deep learning models. They are evaluated using both benchmark and realworld credit card data sets. In addition, the Adaboost and majority voting methods are applied for forming hybrid models. To further evaluate the robustness and reliability of the models, noise is added to the real-world data set.The key contribution of this paper is the evaluation of a variety of machine learning models with a real-world credit card data set for fraud detection. While other researchers have used various methods on publicly available data sets, the data set used in this paper are extracted from actual credit card transaction information over three months.For transaction monitoring by bank employees the clustering model was developed. This model allows provision of fast analysis of transactions by attributes.

CHAPTER 3

3. SOFTWARE AND HARD WARE REQUIREMENTS

The project involved analyzing the design of few applications so as to make the application more users friendly. To do so, it was really important to keep the navigations from one screen to the other well ordered and at the same time reducing the amount of typing the user needs to do. In order to make the application more accessible, the browser version had to be chosen so that it is compatible with most of the Browsers.

REQUIREMENT SPECIFICATION

Functional Requirements:

Graphical User interface with the User.

3.1 SOFTWARE REQUIREMENTS:

For developing the application the following are the Software Requirements:

1. Python

2. Pycharm

3. Jupyter Notebook(Annaconda 3)

Operating Systems supported:

1. Windows 10

2. Debugger and Emulator:

3. Any Browser (Particularly Chrome)

3.2 HARDWARE REQUIREMENTS

For developing the application the following are the Hardware Requirements:

1. RAM: 256 MB

2. Space on Hard Disk: minimum 512MB

CHAPTER 4

4. SOFTWARE DEVELOPMENT ANALYSIS

The software development process involves the creation and maintenance of applications,frameworks and other components for software design, design, programming, documentation, testing and problem remediation. The development of software is a process of creating and keeping source code, but it encompasses everything from the idea of the intended software to the last manifestation of the programme, often in a planned and organised process in a larger context. Software development may therefore encompass research, creation of new software products, prototype, modification, reuse, reengineering, maintenance, or any other software-production activity.

4.1 OVER VIEW OF PROBLEM

Fraud detection is a critical problem affecting large financial companies that have increased due to the growth in credit card transactions. To detect credit card fraud, data mining techniques- Predictive modeling and Logistic Regression are used. In prediction model to predict the continuous valued functions. The fraud transaction detection is the major issue of prediction due to a frequent and large number of transactions. The fraud transaction prediction has the two phases which are feature extraction and classification. In the first phase, the feature extraction technique is applied and in the second phase, classification is applied for the fraud transaction detection. In this review paper various techniques of credit card fraud detection are reviewed. In future hybrid approach will be designed for the credit card fraud detection.

4.2 DEFINE THE PROBLEM

Credit card fraud is any dishonest act and behaviour to obtain information without the proper authorization from the account holder for financial gain. Among different ways of frauds, Skimming is the most common one, which is the way of duplicating of information located on the magnetic strip of the card. In this process, If we swipe our credit card in the terminal then the terminal will check whether the pin is correct or not, or sufficient balance is available or not. If all the information is correct then it will accept or it will reject. If the pin number is correct, it is not meant that the transaction is valid. Because the other person also use the known pin number or other details. So we have a Predictive model which collects the fraud score and also using investigator it can investigate and check fraud detection history using different methods and give feedback to the Predictivemodel.

4.3 MODULES OVERVIEW

A python module can be defined as a python program file which contains a python code including python functions, class, or variables. In other words, we can say that our python code file saved with the extension is treated as the module. We may have a runnable code inside the python module.

Modules in Python provides us the flexibility to organize the code in a logical way. To use the functionality of one module into another, we must have to import the specific module. The module was created to enable cross-platform copy-pasting in Python which was earlier absent. The pyperclip module has copy() and paste() functions that can send text to and receive text from your computer's clipboard. Sending the output of your program to the clipboard will make it easy to paste it on an email, word processor, or some other software. NumPy is an open source library available in Python, which helps in mathematical, scientific, engineering, and data science programming. It is a very useful library to perform mathematical and statistical operations in Python. It works perfectly for multi-dimensional arrays and matrix multiplication. It is easy to integrate with C/C++ and Fortran. For any scientific project, NumPy is the tool to know.

It has been built to work with the N-dimensional array, linear algebra, random number, Fourier transform, etc. NumPy is a programming language that deals with multi-dimensional arrays and matrices. On top of the arrays and matrices, NumPy supports a large number of mathematical operations. Python provides us the flexibility to import some module with a specific name so that we can use this name to use that module in our python source file.

4.4 DEFINE THE MODULES

Credit Card Fraud:

Credit Card Fraud can be authorized, where the genuine customer themselves processes a payment to another account which is controlled by a criminal , or unauthorized. The account holder doesnot provide authorization for the payment to proceed and the transaction is carried out by a third party. The challenge is to recognize fraudulent credit card transactions so that the customers of credit card companies are not charged for items that they did not purchase.

Predictive Model:

Predictive modeling, also called predictive analytics, is a mathematical process that seeks to predict future events or outcomes by analyzing patterns that are likely to forecast future results. This module will conclude by considering more complicated models for Logistic Regression and data observed over time.

Preparing the Data:

This will help in combining the data to prepare it for analysis. Data Preparation is pre-processing step in which data from one or more sources is cleaned and transformed to improve its quality prior to its use.

Pandas:

Pandas is an open-source, Python library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language. Python with Pandas is used in a wide range of fields including academic and commercial domains including finance, economics, statistics, analytics etc..

Numpy:

NumPy, which stands for Numerical Python, is a library consisting of multidimensional array objects and a collection of routines for processing those arrays. Using NumPy, mathematical and logical operations on arrays can be performed. This tutorial explains the basics of NumPy such as its architecture and environment. It also discusses the various array functions, types of indexing, etc. An introduction to Matplotlib is also provided. All this is explained with the help of examples for better understanding.

CHAPTER 5

5. PROJECT SYSTEM DESIGN

Credit card fraud is any dishonest act and behaviour to obtain information without the proper authorization from the account holder for financial gain. Among different ways of frauds, Skimming is the most common one, which is the way of duplicating of information located on the magnetic strip of the card.

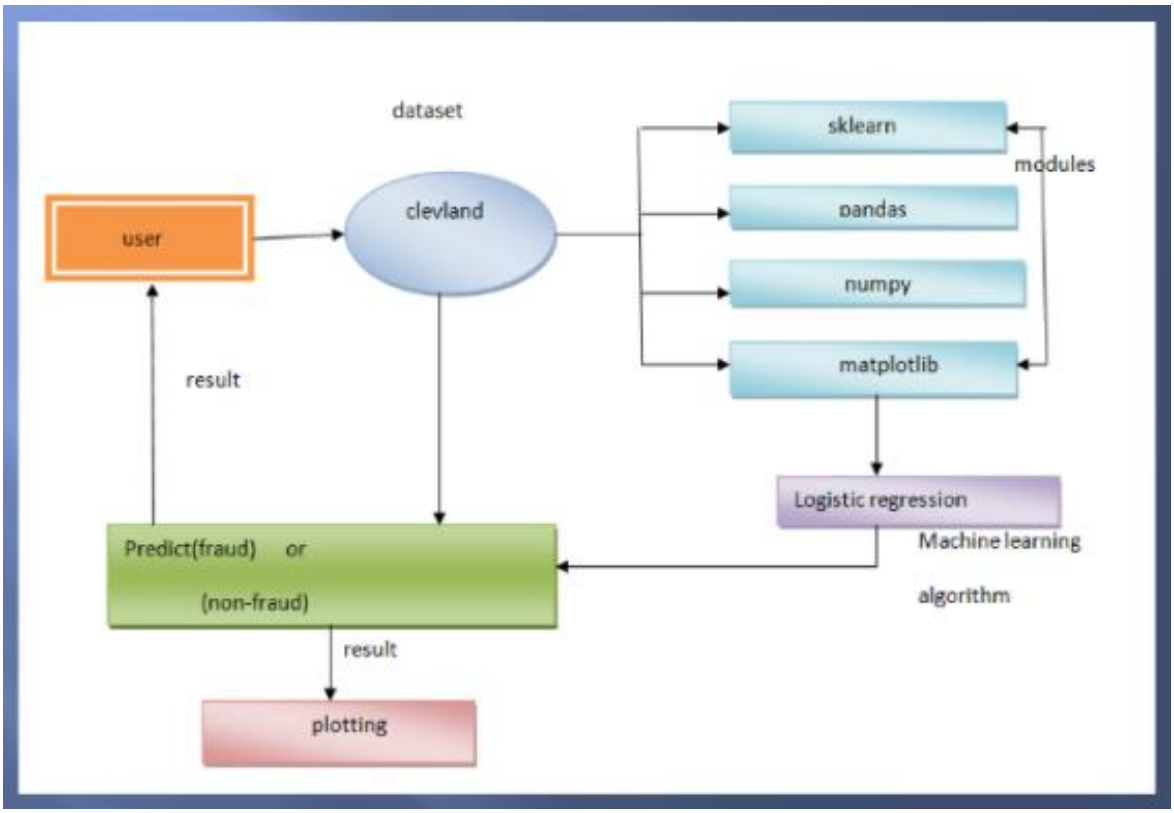


Fig 5 System Design

5.1 DATA FLOW DIAGRAM

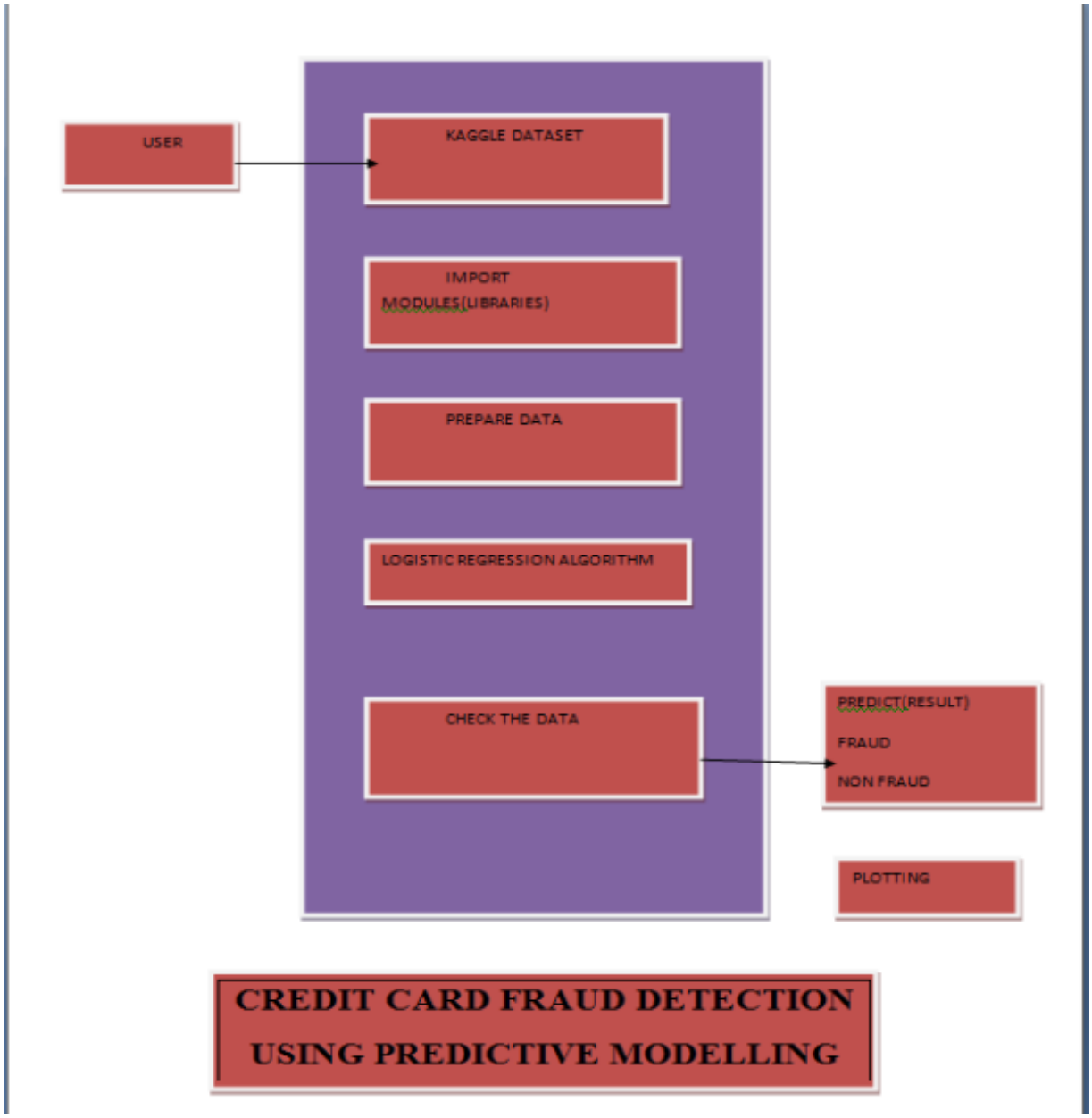


Fig 5.1 Data Flow Diagram

5.2 E-R DIAGRAMS

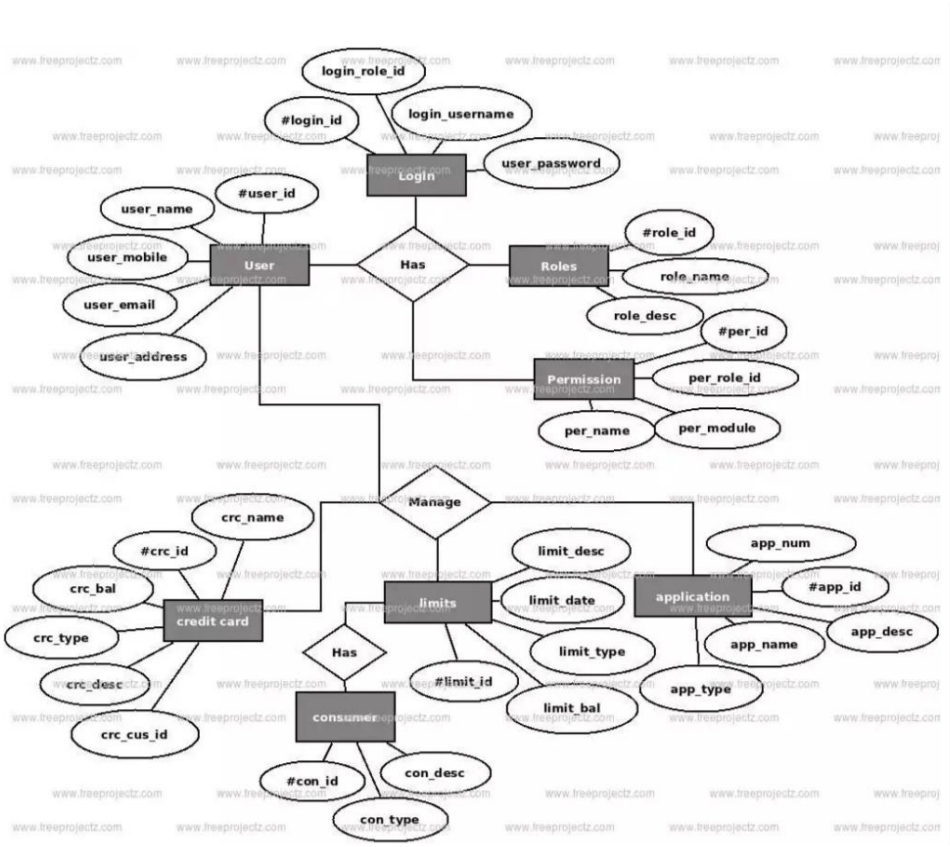


Fig 5.2 ER Diagram

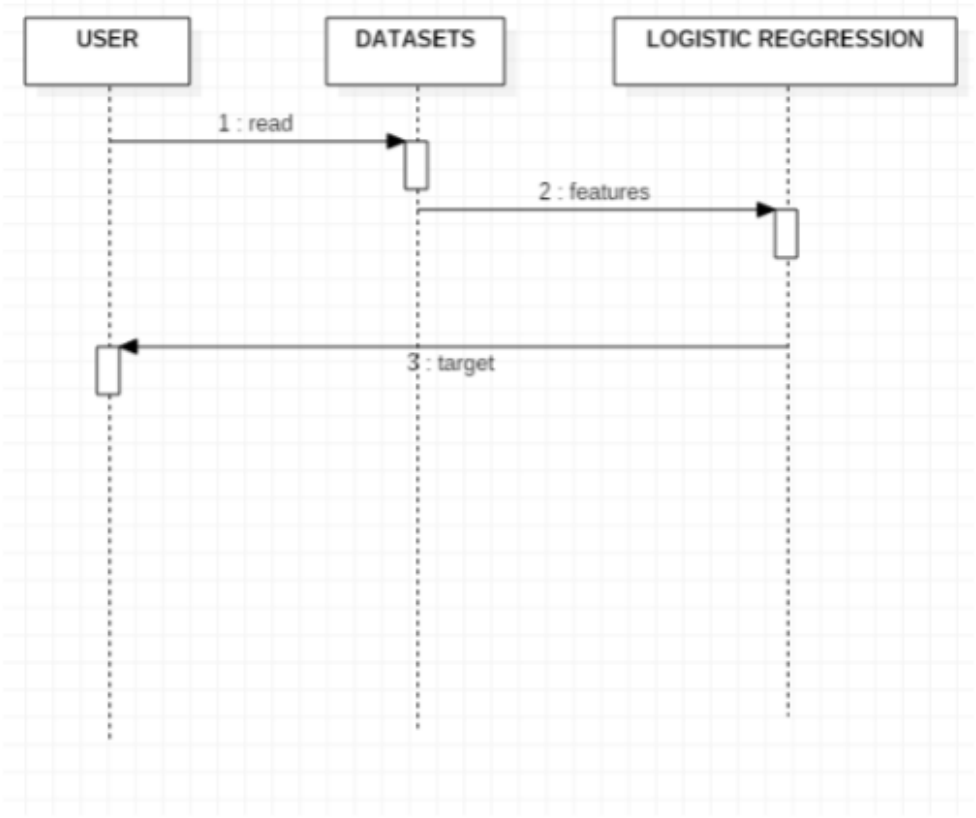
5.3 UML CASE DIAGRAM

Use case Diagram:

Fig 5.3.1 Use Case Diagram

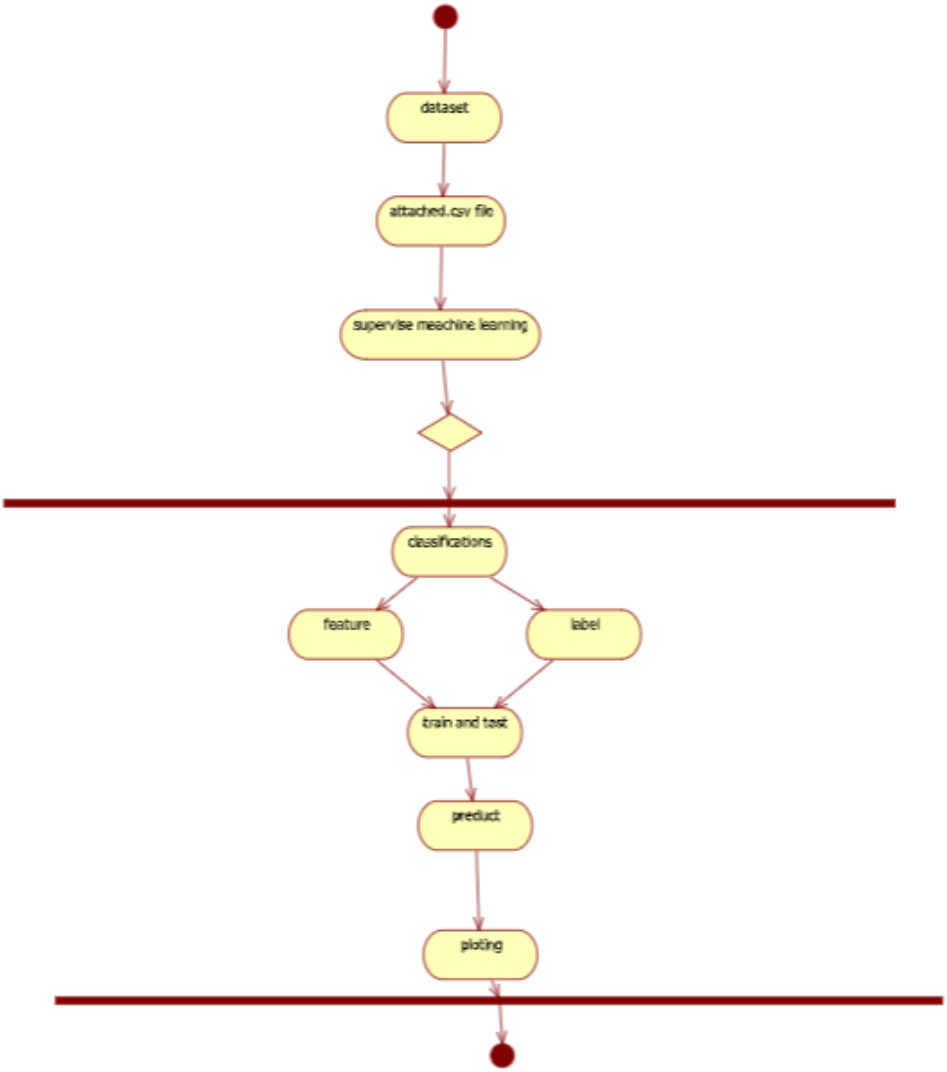
Sequence Diagram:

Fig 5.3.2 Sequence Diagram



Activity Diagram:

Fig 5.3.3Activity Diagram



Component Diagram:

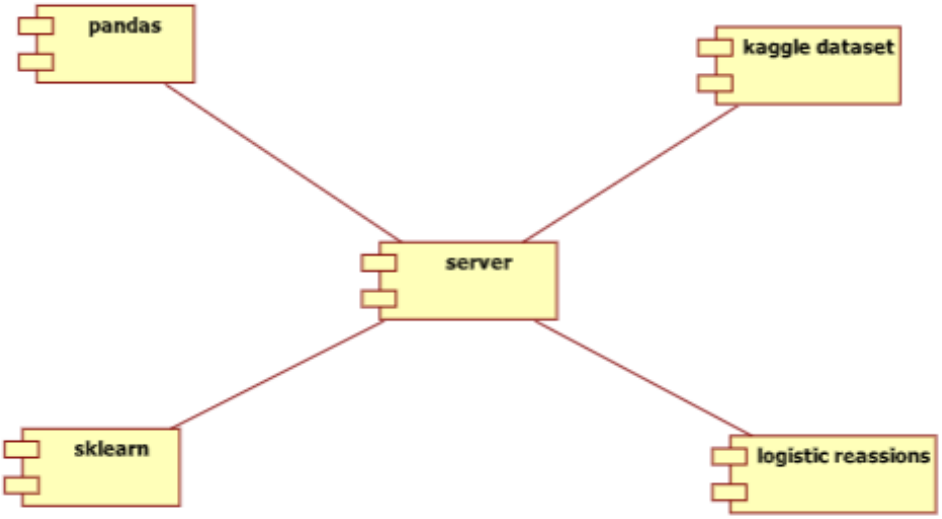


Fig5.3.4 Component Diagram

Object Diagrams:

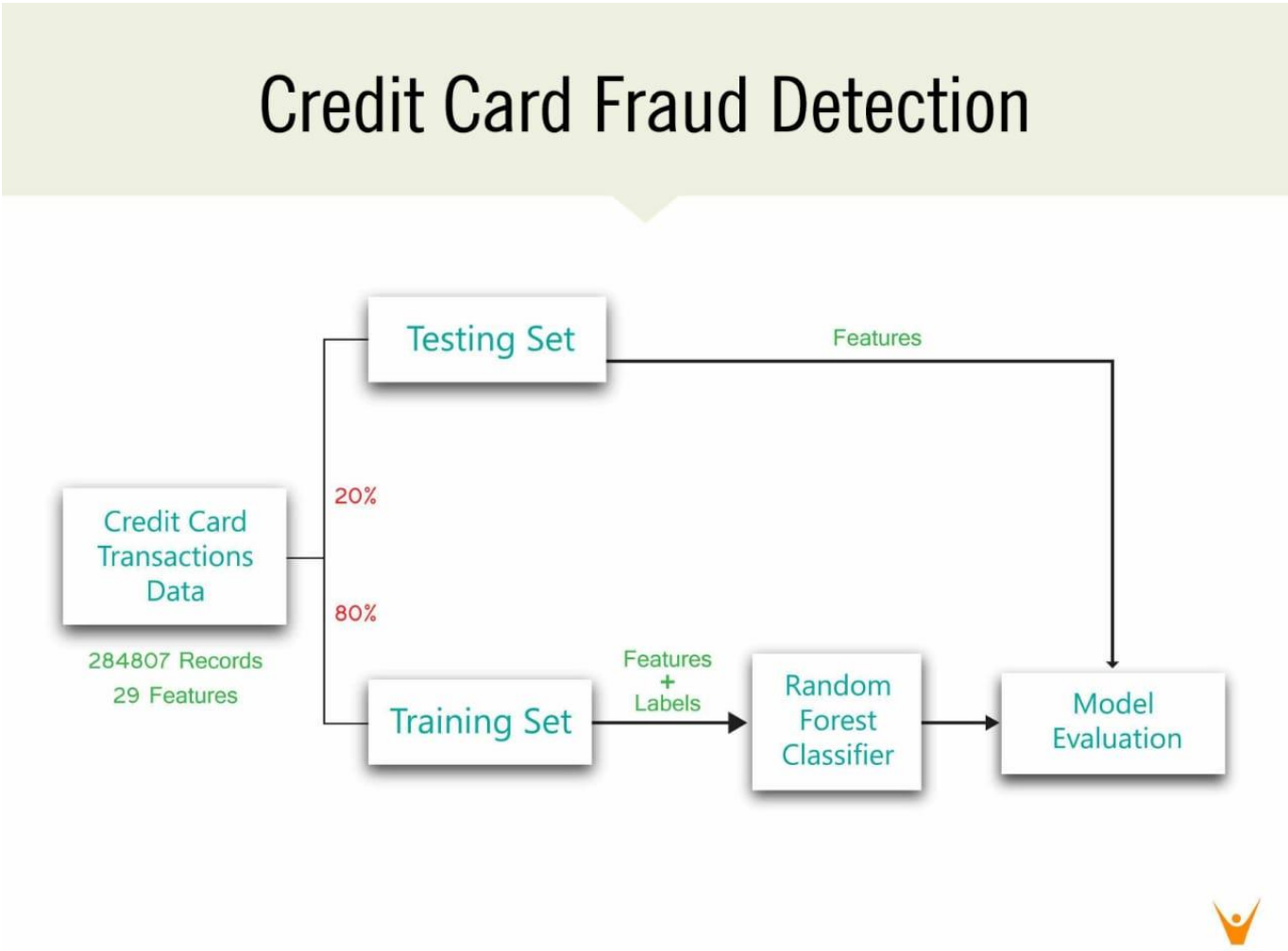
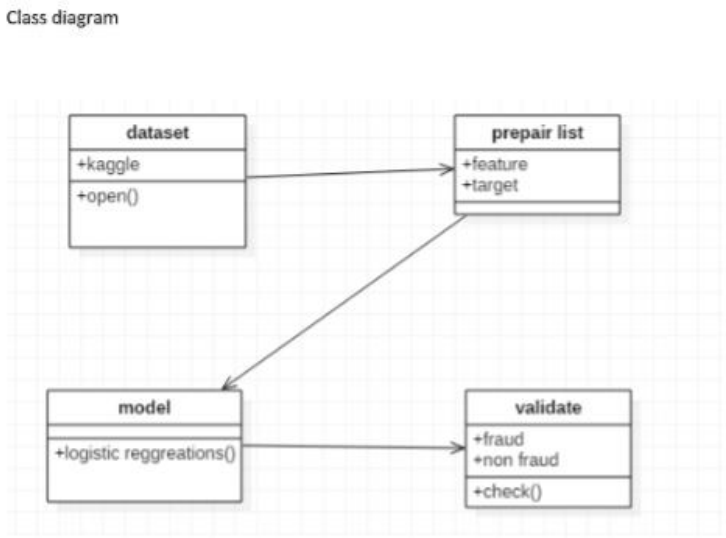


Fig5.3.5 Object Diagram

Class Diagram:

Fig5.3.6 Class Diagram



Deployment Diagram:

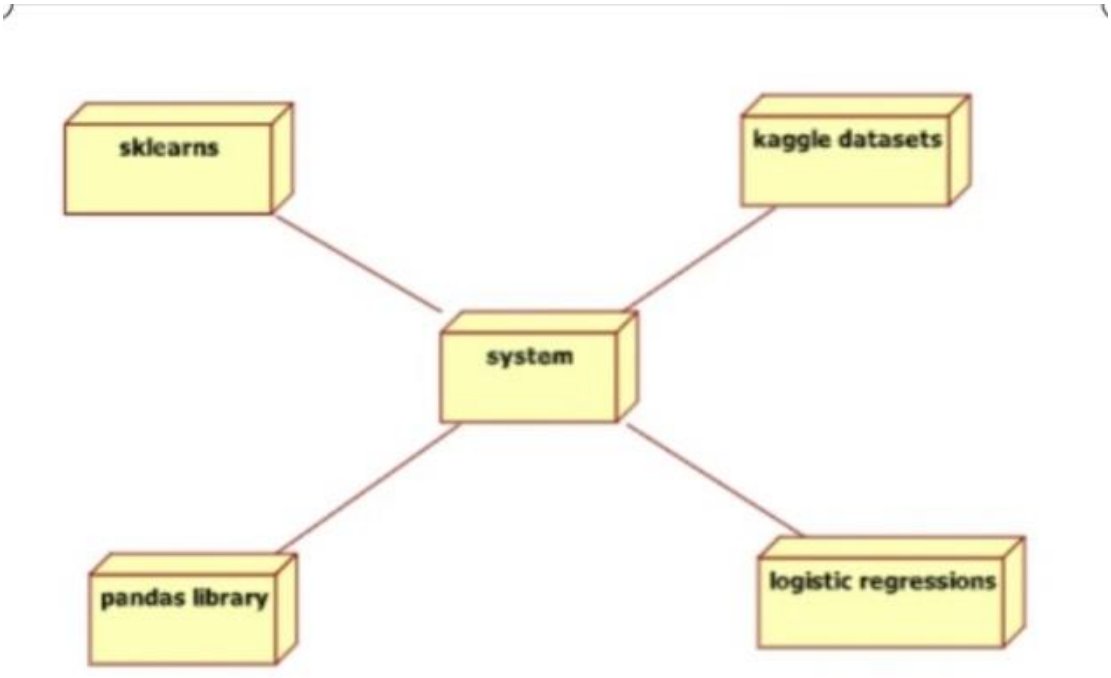


Fig5.3.7Deployment Diagram

Collaboration Diagram:

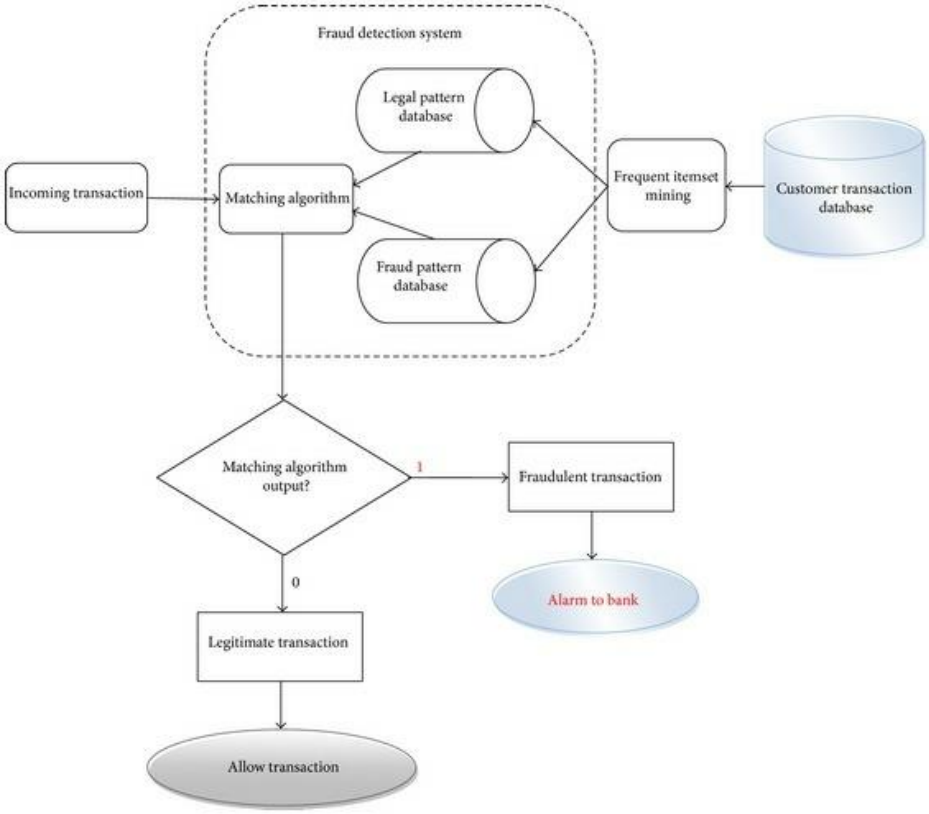


Fig5.3.8 Collaboration Diagram

CHAPTER 6

6. PROJECT CODING

6.1 CODE TEMPLATES:

```
#importing the modules

import numpy as np

#import sklearn python machine learning modules

import sklearn as sk

#import pandas dataframes

import pandas as pd

#import matplotlib for plotting

import matplotlib.pyplot as plt

#import datasets and linear_model from sklearn module

from sklearn import datasets, linear_model

#import Polynomial features from sklearn module

from sklearn.preprocessing import PolynomialFeatures

#import train_test_split data classification

from sklearn.model_selection import train_test_split

#import ConfusionMatrix from pandas_ml

from pandas_ml import ConfusionMatrix

#reading the csv file from C:/Python27

dataframe = pd.read_csv(&#39;C:/Python27/creditcard.csv&#39;, low_memory=False)

#dataframe.sample Returns a random sample of items from an axis of object.

#The frac keyword argument specifies the fraction of rows to return in the random sample, so frac=1

means return all rows (in random order).
```

```
# If you wish to shuffle your dataframe in-place and reset the index

dataframe = dataframe.sample(frac=1).reset_index(drop=True)

#dataframe.head(n) returns a DataFrame holding the first n rows of dataframe.

dataframe.head()

print dataframe
```

6.2 OUTLINE FOR VARIOUS FILES:

The data set that is used here came from superdatascience.com. Huge shout out to them for providing amazing courses and content on their website which motivates people like me to pursue a career in Data Science. Preprocessing is a crucial part to be done at the very beginning of any data science project (unless someone has already done that for you). It includes dealing with NULL values, detecting outliers, removing irrelevant columns through analysis, and cleaning the data in general. With the enhancements to Scenario Outlines, it is now possible to use multiple sources for test parameter data. Previously, parameters were set in the Feature in an Examples section specified and maintained by the user. The addition of Example Row, CSV, SQL, MOCA, and Datastore sources provide more dynamic permutations and less maintenance.

CSV is an commonly utilized file format. CSV files for test parameter data can be used to iterate through many rows of static data. This file can be created manually or exported from another system to swiftly create many test permutations. CSV files are easily edited and maintained, small in size and usually do not require special connections or credentials to use. In short, log files allow you to anticipate upcoming issues before they actually occur. Monitoring and analyzing all of them can be a challenging task. The sheer volume of logs can sometimes make it frustrating just to drill down and find the right file that contains the desired information.

An IPYNB file is a notebook document used by Jupyter Notebook, an interactive computational environment designed to help scientists work with the **Python** language and their data. It contains all the content from the Jupyter Notebook web application session, which includes the inputs and outputs of computations, mathematics, images, and explanatory text. IPYNB files can be exported to **.HTML**, **.PDF**, reStructuredText, and LaTeX formats. More Information IPYNB notebook documents are stored in the **JSON** plain text format, which makes it easier for them to be shared with colleagues and controlling versions. Also, IPYNB notebook documents available from a publicly accessible URL can be shared using the Jupyter Notebook Viewer with other colleagues without requiring them to install Jupyter Notebook on their machines. Jupyter notebooks were formerly known as IPython notebooks, which is where the "ipynb" extension got its name. The file was created by IPython but is now used by the Jupyter Notebook app.

6.3 CLASS WITH FUNCTIONALITIES

```
import pandas as pd
import matplotlib.pyplot as plt
from matplotlib.patches import Rectangle
import numpy as np
from pprint import pprint as pp
import csv
from pathlib import Path
import seaborn as sns
from itertools import product
import string

import nltk
from nltk.corpus import stopwords
from nltk.stem.wordnet import WordNetLemmatizer

from imblearn.over_sampling import SMOTE
from imblearn.over_sampling import BorderlineSMOTE
from imblearn.pipeline import Pipeline

from sklearn.linear_model import LinearRegression, LogisticRegression
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import r2_score, classification_report, confusion_matrix, accuracy_score, roc_auc_s
core, roc_curve, precision_recall_curve, average_precision_score
from sklearn.metrics import homogeneity_score, silhouette_score
from sklearn.ensemble import RandomForestClassifier, VotingClassifier
from sklearn.preprocessing import MinMaxScaler
from sklearn.cluster import MiniBatchKMeans, DBSCAN

import gensim
from gensim import corpora
```

A typical organization loses an estimated 5% of its yearly revenue to fraud. In this course, learn to fight fraud by using data. Apply supervised learning algorithms to detect fraudulent behavior based upon past fraud, and use unsupervised learning methods to discover new types of fraud activities. Fraudulent transactions are rare compared to the norm. As such, learn to properly classify imbalanced datasets. The course provides technical and theoretical insights and demonstrates how to implement fraud detection models. Finally, get tips and advice from real-life experience to help prevent common mistakes in fraud analytics.

6.4 METHODS INPUT AND OUTPUT PARAMETERS:

Machine learning involves predicting and classifying data and to do so, you employ various machine learning models according to the dataset. Machine learning models are parameterized so that their behavior can be tuned for a given problem. These models can have many parameters and finding the best combination of parameters can be treated as a search problem. But this very term called parameter may appear unfamiliar to you if you are new to applied machine learning. But don't worry! You will get to know about it in the very first place of this blog, and you will also discover what the difference between a parameter and a hyperparameter of a machine learning model is. This blog consists of following sections: What are a parameter and a hyperparameter in a machine learning model? Why hyperparameter optimization/tuning is vital in order to enhance your model's performance?Two simple strategies to optimize/tune the hyperparameters A simple case study in Python with the two strategies

A model parameter is a configuration variable that is internal to the model and whose value can be estimated from the given data.

They are required by the model when making predictions.

- Their values define the skill of the model on your problem.
- They are estimated or learned from data.
- They are often not set manually by the practitioner.
- They are often saved as part of the learned model.

So your main take away from the above points should be parameters are crucial to machine learning algorithms. Also, they are the part of the model that is learned from historical training data. Let's dig it a bit deeper. Think of the function parameters that you use while programming in general. You may pass a parameter to a function. In this case, a parameter is a function argument that could have one of a range of

values. In machine learning, the specific model you are using is the function and requires parameters in order to make a prediction on new data. Whether a model has a fixed or variable number of parameters determines whether it may be referred to as *"parametric"* or *"nonparametric"*.

Some examples of model parameters include:

- The weights in an artificial neural network.
- The support vectors in a support vector machine.
- The coefficients in a linear regression or logistic regression.

A model hyperparameter is a configuration that is external to the model and whose value cannot be estimated from data.

- They are often used in processes to help estimate model parameters.
- They are often specified by the practitioner.
- They can often be set using heuristics.
- They are often tuned for a given predictive modeling problem.

You cannot know the best value for a model hyperparameter on a given problem. You may use rules of thumb, copy values used on other issues, or search for the best value by trial and error. When a machine learning algorithm is tuned for a specific problem then essentially you are tuning the hyperparameters of the model to discover the parameters of the model that result in the most skillful predictions.

According to a very popular book called "Applied Predictive Modelling" - Model hyperparameters are often referred to as model parameters which can make things confusing. A good rule of thumb to overcome this confusion is as follows: Some examples of model hyperparameters include:

- The learning rate for training a neural network.
- The C and sigma hyperparameters for support vector machines.
- The k in k-nearest neighbors.

In the next section, you will discover the importance of the right set of hyperparameter values in a machine learning model.

The best way to think about hyperparameters is like the settings of an algorithm that can be adjusted to optimize performance, just as you might turn the knobs of an AM radio to get a clear signal. When creating a machine learning model, you'll be presented with design choices as to how to define your model architecture. Often, you don't immediately know what the optimal model architecture should be for a given model, and thus you'd like to be able to explore a range of possibilities. In a true machine learning fashion, you'll ideally ask the machine to perform this exploration and select the optimal model architecture automatically.

You will see in the case study section on how the right choice of hyperparameter values affect the performance of a machine learning model.

Models can have many hyperparameters and finding the best combination of parameters can be treated as a search problem.

Although there are many hyperparameter optimization/tuning algorithms now, this post discusses two simple strategies: 1. grid search and 2. Random Search.Note that, the array of values of that you are defining for the hyperparameters has to be legitimate in a sense that you cannot supply floating type values to the array if the hyperparameter only takes Integer values.

Random search differs from a grid search. In that you longer provide a discrete set of values to explore for each hyperparameter; rather, you provide a statistical distribution for each hyperparameter from which values may be randomly sampled. In Statistics, by distribution, it is essentially meant an arrangement of values of a variable showing their observed or theoretical frequency of occurrence.On the other hand, Sampling is a term used in statistics. It is the process of choosing a representative sample from a target population and collecting data from that sample in order to understand something about the population as a whole.

Machine learning involves predicting and classifying data and to do so, you employ various machine learning models according to the dataset. Machine learning models are parameterized so that their behavior can be tuned for a given problem. These models can have many parameters and finding the best combination of parameters can be treated as a search problem. But this very term called parameter may appear unfamiliar to you if you are new to applied machine learning. But don't worry! You will get to know about it in the very first place of this blog, and you will also discover what the difference between a parameter and a hyperparameter of a machine learning model is. A model parameter is a configuration variable that is internal to the model and whose value can be estimated from the given data.

- They are required by the model when making predictions.
- Their values define the skill of the model on your problem.
- They are estimated or learned from data.
- They are often not set manually by the practitioner.
- They are often saved as part of the learned model.

So your main take away from the above points should be parameters are crucial to machine learning algorithms. Also, they are the part of the model that is learned from historical training data. Let's dig it a bit deeper. Think of the function parameters that you use while programming in general. You may pass a parameter to a function. In this case, a parameter is a function argument that could have one of a range of values. In machine learning, the specific model you are using is the function and requires parameters in order to make a prediction on new data. Whether a model has a fixed or variable number of parameters determines whether it may be referred to as *"parametric"* or *"nonparametric"*.

Some examples of model parameters include:

- The weights in an artificial neural network.
- The support vectors in a support vector machine.
- The coefficients in a linear regression or logistic regression.

A model hyperparameter is a configuration that is external to the model and whose value cannot be estimated from data.

- They are often used in processes to help estimate model parameters.
- They are often specified by the practitioner.
- They can often be set using heuristics.
- They are often tuned for a given predictive modeling problem.

You cannot know the best value for a model hyperparameter on a given problem. You may use rules of thumb, copy values used on other issues, or search for the best value by trial and error. When a machine learning algorithm is tuned for a specific problem then essentially you are tuning the hyperparameters of

the model to discover the parameters of the model that result in the most skillful predictions. Model hyperparameters are often referred to as model parameters which can make things confusing. A good rule of thumb to overcome this confusion is as follows: “If you have to specify a model parameter manually, then it is probably a model hyperparameter. ” Some examples of model hyperparameters include:

- The learning rate for training a neural network.
- The C and sigma hyperparameters for support vector machines.
- The k in k-nearest neighbors.

The best way to think about hyperparameters is like the settings of an algorithm that can be adjusted to optimize performance, just as you might turn the knobs of an AM radio to get a clear signal. When creating a machine learning model, you'll be presented with design choices as to how to define your model architecture. Often, you don't immediately know what the optimal model architecture should be for a given model, and thus you'd like to be able to explore a range of possibilities. In a true machine learning fashion, you’ll ideally ask the machine to perform this exploration and select the optimal model architecture automatically. In this context, choosing the right set of values is typically known as Models can have many hyperparameters and finding the best combination of parameters can be treated as a search problem. Although there are many hyperparameter optimization/tuning algorithms now, this post discusses two simple strategies: 1. grid search and 2. Random Search.nRandom searching of hyperparameters: The idea of random searching of hyperparameters was proposed by James Bergstra & Yoshua Bengio. Random search differs from a grid search. In that you longer provide a discrete set of values to explore for each hyperparameter; rather, you provide a statistical distribution for each hyperparameter from which values may be randomly sampled.Before going any further, let’s understand what distribution and sampling mean: In Statistics, by distribution, it is essentially meant an arrangement of values of a variable showing their observed or theoretical frequency of occurrence.

On the other hand, Sampling is a term used in statistics. It is the process of choosing a representative sample from a target population and collecting data from that sample in order to understand something about the population as a whole.

CHAPTER 7

7. PROJECT TESTING

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub-assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of tests. Each test type addresses a specific testing requirement.

7.1 VARIOUS TEST CASES:

In the simplest form, a test case is a set of conditions or variables under which a tester determines whether the software satisfies requirements and functions properly. A test case is a single executable test which a tester carries out. It guides them through the steps of the test.

7.1.1 UNIT TESTING

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

Unit testing is usually conducted as part of a combined code and unit test phase of the software lifecycle, although it is not uncommon for coding and unit testing to be conducted as two distinct phases.

Test strategy and approach

Field testing will be performed manually, and functional tests will be written in detail.

Test objectives

- All field entries must work properly.
- Pages must be activated from the identified link.
- The entry screen, messages and responses must not be delayed.

Features to be tested:

- Verify that the entries are of the correct format.
- No duplicate entries should be allowed.

7.1.2 INTEGRATION TESTING

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

Software integration testing is the incremental integration testing of two or more integrated software components on a single platform to produce failures caused by interface defects.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

7.1.3 FUNCTIONAL TESTING

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals. Functional testing is centered on the following items:

- Valid Input : identified classes of valid input must be accepted.
- Invalid Input : identified classes of invalid input must be rejected.
- Functions : identified functions must be exercised.
- Output : identified classes of application outputs must be exercised.

Systems/Procedures: Interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

7.1.4 SYSTEM TESTING

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration-oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

Software once validated must be combined with other system elements (e.g., Hardware, people, database). System testing verifies that all the elements are proper, and that overall system function performance is achieved. It also tests to find discrepancies between the system and its original objective, current specifications and system documentation.

7.1.5 ACCEPTANCE TESTING

User Acceptance of a system is the key factor for the success of any system. The system under consideration is tested for user acceptance by constantly keeping in touch with the prospective system users at the time of developing and making changes wherever required. The system developed provides a friendly user interface that can easily be understood even by a person who is new to the system. User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

7.1.6 OUTPUT TESTING

After performing the validation testing, the next step is output testing of the proposed system, since no system could be useful if it does not produce the required output in the specified format. Asking the users about the format required by them tests the outputs generated or displayed by the system under consideration. Hence the output format is considered in 2 ways – one is on screen and another in printed format.

7.1.7 VALIDATION CHECKING

Validation checks are performed on the following fields.

• **Text Field:**

The text field can contain only the number of characters lesser than or equal to its size. The text fields are alphanumeric in some tables and alphabetic in other tables. Incorrect entry always flashes and error message.

• **Numeric Field:**

The numeric field can contain only numbers from 0 to 9. An entry of any character flashes an error message. The individual modules are checked for accuracy and what it has to perform. Each module is subjected to test run along with sample data. The individually tested modules are integrated into a single system. Testing involves executing the real data information is used in the program the existence of any program defect is inferred from the output. The testing should be planned so that all the requirements are individually tested.

7.2 BLACK BOX TESTING

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box. you cannot "see" into it. The test provides inputs and responds to outputs without considering how the software works.

7.3 WHITE BOX TESTING

White Box Testing is a testing in which in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is purpose. It is used to test areas that cannot be reached from a black box level.

White-box testing is a method of testing the application at the level of the source code. These test cases are derived through the use of the design techniques mentioned above: control flow testing, data flow testing, branch testing, path testing, statement coverage and decision coverage as well as modified condition/decision coverage. White-box testing is the use of these techniques as guidelines to create an error-free environment by examining all code. These white-box testing techniques are the building blocks of white-box testing, whose essence is the careful testing of the application at the source code level to reduce hidden errors later on. These different techniques exercise every visible path of the source code to minimize errors and create an error-free environment. The whole point of white-

box testing is the ability to know which line of the code is being executed and being able to identify what the correct output should be.

Working process of white box testing:

• **Input:** Requirements, Functional specifications, design documents, source code.

• **Processing:** Performing risk analysis for guiding through the entire process.

• **Proper test planning:** Designing test cases so as to cover entire code. Execute rinse-repeat until error-free software is reached. Also, the results are communicated.

• **Output:** Preparing final report of the entire testing process.

CHAPTER 8

8. OUTPUT SCREENS

Fig 8.1 transaction class distribution

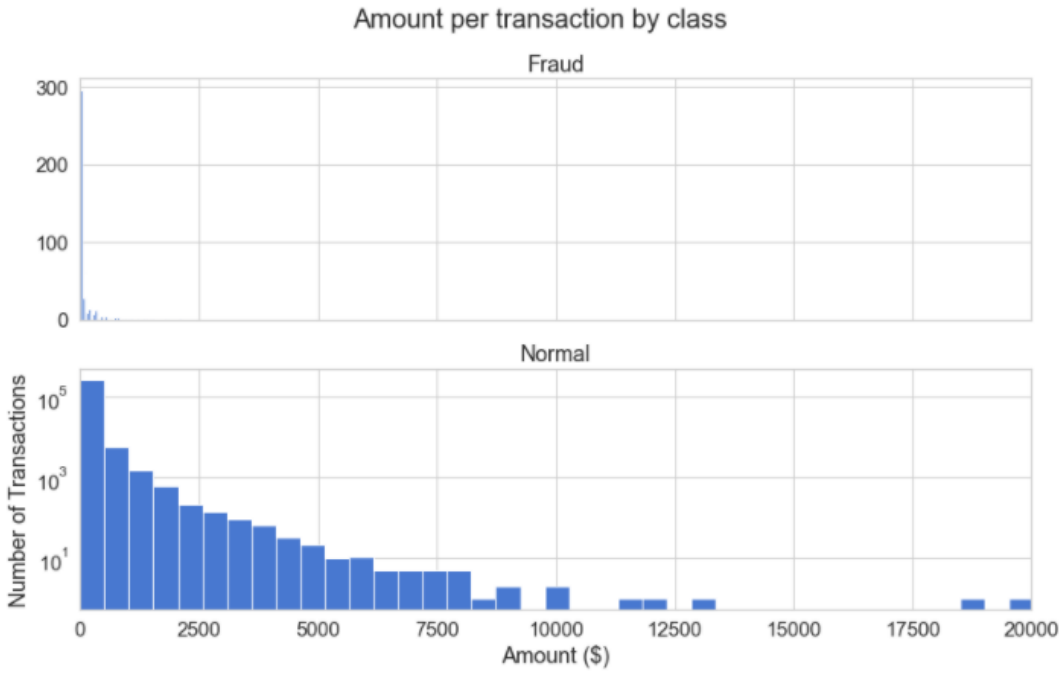


Fig 8.2 Amount per Transaction class

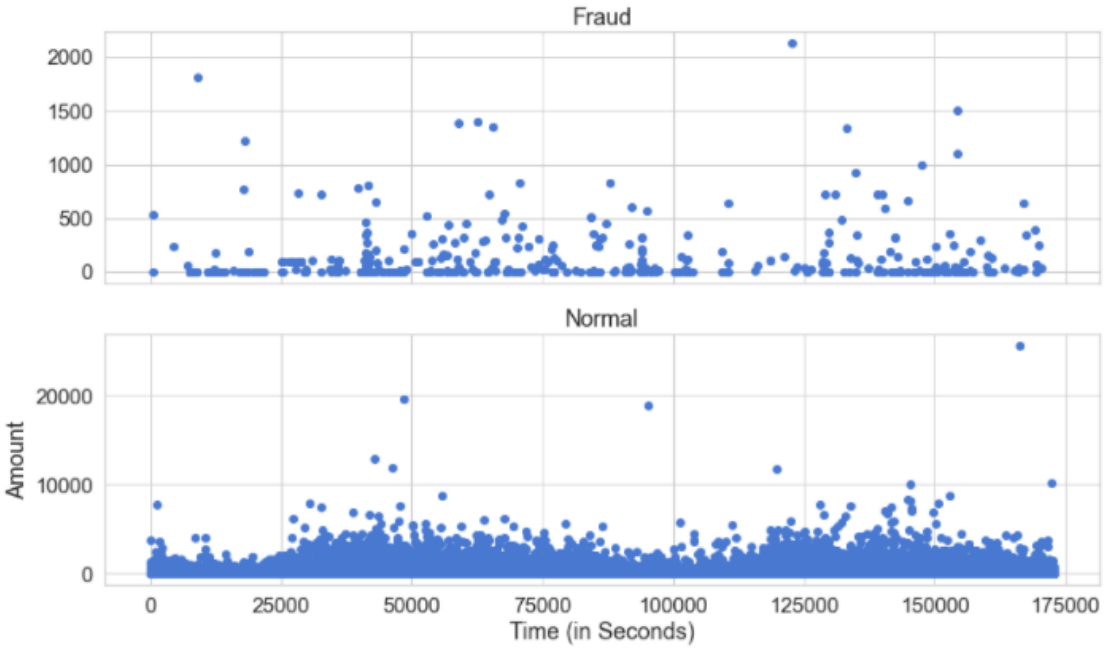


Fig 8.3 Time of Transaction Vs Amount of Time

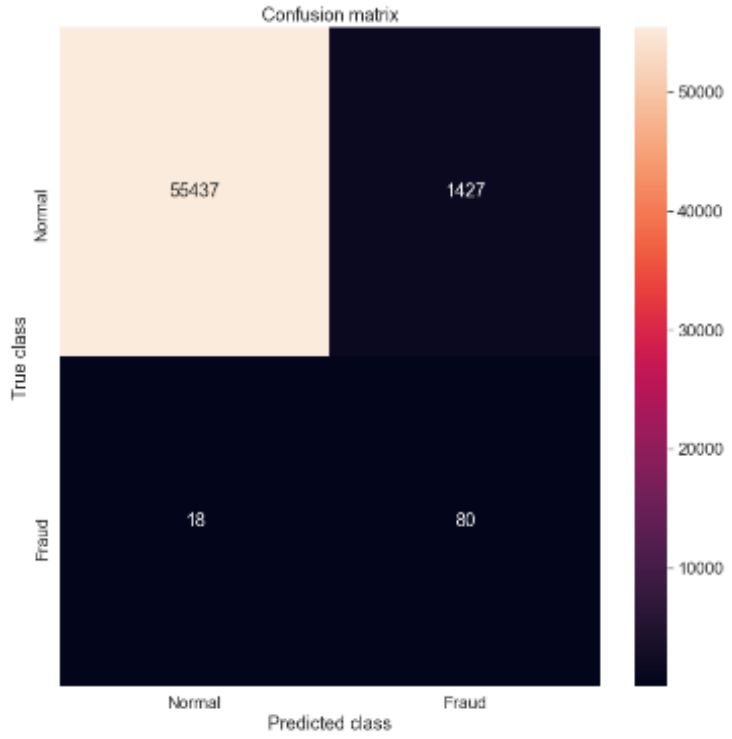


Fig 8.4 Confusion Matrix

CHAPTER 9

9. EXPERIMENTAL RESULTS

The code prints out the number of false positives it detected and compares it with the actual values. This is used to calculate the accuracy score and precision of the algorithms. The fraction of data we used for faster testing is 10% of the entire dataset. The complete dataset is also used at the end and both the results are printed. These results along with the classification report for each algorithm is given in the output as follows, where class 0 means the transaction was determined to be valid and 1 means it was determined as a fraud transaction. This result matched against the class values to check for false positives.

Predictive modeling of output Class distribution In Predictive modeling, to split the data set into 70% of testing and 30% of training then check output class distribution as shown above table 1, 148 frauds out of 284,807 transactions. To predict an outcome of frauds in credit card transaction that occurred in two days. We have to analyzed 85295 not frauds in credit card transaction.

0	1
85295	148

Table9.1 Predictive modeling of output Class distribution

Logistic Regression for credit card fraud detection Logistic Regression or logit model is a regression model, where the dependent variable is categorical. In case of logistic regression as shown above table 2, 79 frauds out of 85279 transactions. The categorical variable is class variable may be 0 or 1, to generalize a model with 1 that is fraud, otherwise 0. Decision tree model Decision tree uses tree structure to build regression model, the final result is a tree as shown Fig .1.with top most node is root node. Root node describes the best predictor in the data, and decision node is a combination of two or more branches, each branch represents a value for the attribute which is tested, leaf node holds class label. The leaf node may be 1 means fraud and 0 otherwise.

	FALSE	TRUE
0	85279	16
1	69	79

Table9.2 Logistic Regression for credit card fraud detection

CONCLUSIONS

- This project thus proposed a system to classify alerts in fraud detection system using supervised learning technique to classify alert as fraudulent or non fraudulent.
- Further we have also used learning to rank approach to rank the fraudulent alert generated by classifier based on priority. The performances of all this techniques are examined based on precision, F1 score, recall and accuracy.
- A comparative study is also done where proposed system is compared with Decision Tree and Naïve Bayes Technique.
- It showed that the proposed system better accuracy for a large dataset. Future work concerns the classification of alerts by applying semi-supervised learning methods in FDS(functional Desing System).

FUTURE ENHANCEMENTS

This process is used to detect the credit card transaction, which are fraudulent or genuine. Data mining techniques of Predictive modeling, Decision trees and Logistic Regression are used to predict the fraudulent or genuine credit card transaction. In predictive modeling to detect and check output class distribution. The prediction model predicts continuous valued functions. We have to detect 148 may be fraud and other are genuine. In decision tree generate a tree with root node, decision node and leaf nodes. The leaf node may be 1 becomes fraud and 0 otherwise. Logistic Regression is same as linear regression but interpret curve is different. To generalize the linear regression model, when dependant variable is categorical and analyzes relationship between multiple independent variables.

REFERENCES

[1] Y. Sahin, S. Bulkan, and E. Duman, “A cost-sensitive decision tree approach for fraud detection,” Expert Systems with Applications, vol. 40, no. 15, pp. 5916–5923, 2013.

[2] A. O. Adewumi and A. A. Akinyelu, “A survey of machine-learning and nature-inspired based credit card fraud detection techniques,” International Journal of System Assurance Engineering and Management, vol. 8, pp. 937–953, 2017.

[3] A. Srivastava, A. Kundu, S. Sural, A. Majumdar, “Credit card fraud detection using hidden Markov model,” IEEE Transactions on Dependable and Secure Computing, vol. 5, no. 1, pp. 37–48, 2008.

[4] J. T. Quah, and M. Sriganesh, “Real-time credit card fraud detection using computational intelligence,” Expert Systems with Applications, vol. 35, no. 4, pp. 1721–1732, 2008.

[5] S. Bhattacharyya, S. Jha, K. Tharakunnel, and J. C., “Data mining for credit card fraud: A comparative study,” Decision Support Systems, vol. 50, no. 3, pp. 602–613, 2011.

[6] N. S. Halvaiee and M. K. Akbari, “A novel model for credit card fraud detection using Artificial Immune Systems,” Applied Soft Computing, vol. 24, pp. 40–49, 2014.

[7] S. Panigrahi, A. Kundu, S. Sural, and A. K. Majumdar, “Credit card fraud detection: A fusion approach using Dempster–Shafer theory and Bayesian learning,” Information Fusion, vol. 10, no. 4, pp. 354–363, 2009.

[8] N. Mahmoudi and E. Duman, “Detecting credit card fraud by modified Fisher discriminant analysis,” Expert Systems with Applications, vol. 42, no. 5, pp. 2510–2516, 2015.

[9] D. Sánchez, M. A. Vila, L. Cerda, and J. M. Serrano, “Association rules applied to credit card fraud detection,” Expert Systems with Applications, vol. 36, no. 2, pp. 3630–3640, 2009.

[10] E. Duman and M. H. Ozcelik, “Detecting credit card fraud by genetic algorithm and scatter search,” Expert Systems with Applications, vol. 38, no. 10, pp. 13057–13063, 2011.

[11] P. Ravisankar, V. Ravi, G. R. Rao, a[19] Rhodes, J., 2020. Bulding your First Data Science project n Microsoft Azure.

[12] Srinivasulu, R., 2019. Preparing Data For Feature Engineering And Machine Learning in Microsoft Azure

[13] P. Ravisankar, V. Ravi, G. R. Rao, and I. Bose, “Detection of financial statement fraud and feature selection using data mining.

[14] N.Mahmoudi and E.Duman,”Detecting credit card fraud by modified fisher discriminant analysis,”Expert System.Appl,vol.42,no 5pp.2510-2516,2015

[15] ”Detecting Credit card Fraud using periodic features,”in proc.14th Int.conf.Mach.Learn Appl.Dec.2015,pp.208-213.

[16] A.C.BahnsenD.Aouada,and B.Ottersten,”Example-dependent cost-sensitive decision trees,”Expert Syst.Appl,vol.42,no.19,pp.6609-6619,2015

PUBLICATIONS

- JOURNAL : UGC Journal.
- CONFERENCE : *Online Mega International Conference " Innovations in Computer Networks, Computational Intelligence and IOT" (ICICCI-21)" On 25th & 26th June, 2021.*

Dasari.Grishma is currently pursuing her Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin’s Engineering College. She completed her intermediatefrom Narayana Junior College and 10th class from The Indo English High School. Her areas of interest are HTML,CSS,JavaScript and its frameworks(React JS ,Angular JS) . She completed a couple of certification courses from online platforms like Coursera, CursaApp and SoloLearn and Hacker Rank. Also have an Internship Certificate on completing Internship as a UI Developer in Syncor Solutions Private Limited.

K.Swathi is currently pursuing her Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin’s Engineering College. She completed her intermediate from Sri Chaitanya Junior College and 10th class from The Dhatrak Model School. Her areas of interest are HTML, CSS, JavaScript, Python. she also have a Java language certificate from Lasya Infotech. She completed a couple of certification courses from online platforms like Coursera, CursaApp and Hacker Rank.

Supriya jangam is currently parsuing bachelor's of technology (B Tech) in the stream of computer science and engineering at St.Martin's engineering college .And completed her board of intermediate education at Sri Chaitanya junior college and Secondary school certificate (SSC) at Vijay high school . And also she participated in certified for a course called "A1 for everyone" in Coursera, participated in certified for a course called "Matrix algebra of engineers", participated in certified for a course called " introduction and programming with iot boards", participated in certified for a course called " Introduction to cybersecurity tools and cyber attacks".The area of interest are python and HTML. And also completed few certificate courses from online platform like Coursera, Solo learn and Cursaapp.

Sathwika Tallarekula is Currently parsuring bachelor's of technology (B.tech) in the Stream of Computer Science and Engineering at St. Martin's Engineering college And I complete bored of Intermediate education at Sri Chaitanya Junior College And Central Board of Secondary (CBSE) At Jawahar Navodaya Vidyalaya Nizamsagar. And also She participate in certificate for a Course called "A1 for Everyone" in Coursera, participated in Certificate for a course Called " Leadership and Emotional intelligence in Coursera, participate in Certificate for a Cursa called "MYSQL database by Thenewbaston in Cursa, participate in certificate for a course called"Basic of AWS Concept by Exampro", The area of interest in c, c++, and HTML. And also complete certificate courses from online platform like Cursa, courses.

A

PROJECT REPORT

On

STOCK MARKET TREND USING K NEAREST NEIGHBOR (KNN) ALGORITHM

Submitted by

N.Santosh Goud (17K81A0538) M.Sai Kumar(17K81A0537)

M.Sai Chand (16K81A0542) K.Sujith (16K81A0531)

In partial fulfillment for

the award of the degree

of

BACHELOR OF TECHNOLOGY

IN

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Under the Guidance of

Mr. P.R.K AYYAPPA

Assistant Professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ST.MARTIN'S ENGINEERING COLLEGE
An Autonomous Institute

Dhulapally, Secunderabad – 500 100

JUNE 2021

BONAFIDE CERTIFICATE

This is to certify that the project entitled **STOCK MARKET TREND USING**

K NEAREST NEIGHBOR(KNN) ALGORITHM is being submitted by **1. Mr.Santosh Goud (17K81A0538) , 2. Mr. Sai Kumar (17K81A0537), 3. Mr.Sai Chand (16K81A0542) 4. Mr. Sujith (16K81A0531)** in partial fulfillment of the requirement for the award of the degree of **BACHELOR OF TECHNOLOGY IN COMPUTER SCIENCE AND ENGINEERING** is recorded of bonafide work carried out by them. The result embodied in this report have been verified and found satisfactory.

Mr.P.R.K.AYYAPPA	**Head of the Department**
Assistant Professor	**Dr.M.NARAYANAN**
Department of CSE	**Department of CSE**

Internal Examiner External Examiner

Place:

Date:

DECLARATION

We, the student of **Bachelor of Technology** in Department of Computer Science and Engineering', session: 2017 – 2021, St. Martin's Engineering College, Dhulapally, Kompally, Secunderabad, hereby declare that work presented in this Project Work entitled STOCK MARKET TREND USING K NEAREST NEIGHBOR(KNN) ALGORITHM is the outcome of our own bonafide work and is correct to the best of our knowledge and this work has been undertaken taking care of Engineering Ethics. This result embodied in this project report has not been submitted in any university for award of any degree.

N.Santosh Goud (17K81A0538)
M.Sai Kumar (17K81A0537)
M.Sai Chand (16K81A0542)
K.Sujith (16K81A0531)

ABSTRACT

This project examines a hybrid model which combines a K-Nearest Neighbors (KNN) approach with a probabilistic method for the prediction of stock price trends. One of the main problems of KNN classification is the assumptions implied by distance functions. The assumptions focus on the nearest neighbors which are at the centroid of data points for test instances. This approach excludes the non-centric data points which can be statistically significant in the problem of predicting the stock price trends. For this it is necessary to construct an enhanced model that integrates KNN with a probabilistic method which utilizes both centric and non-centric data points in the computations of probabilities for the target instances. The embedded probabilistic method is derived from Bayes' theorem. The prediction outcome is based on a joint probability where the likelihood of the event of the nearest neighbors and the event of prior probability occurring together and at the same point in time where they are calculated. The proposed hybrid KNN Probabilistic model was compared with the standard classifiers that include KNN, Naive Bayes, One Rule (OneR) and Zero Rule (ZeroR). The test results showed that the proposed model outperformed the standard classifiers which were used for the comparisons.

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible and whose encouragements and guidance have crowded effects with success.

We extended our deep sense of gratitude to Principal**, Dr. P. SANTOSH KUMARPATRA**, St. Martin's Engineering College, Dhulapally, for permitting us to undertake this project.

We are also thankful to **Dr. M.NARAYANAN**, Head of the Department, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

We would like to thank **Dr. T. POONGOTHAI,** Department of Computer Science and Engineering (AI & ML) for her encouragement and insightful comments and as well as our project coordinators **Dr. B.RAJALINGAM**, Associate Professor and **Mr. J.SUDHAKAR**, Assistant Professor, in Department of Computer Science and Engineering for their valuable support.

We would like to express our sincere gratitude and indebtedness to our project supervisor Mr. P.R.Kayyapa, Assistant professor, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

Finally, we express thanks to all those who have helped us successfully to completing this project. Furthermore, we would like to thank our family and friends for their moral support and encouragement.

We express thanks to all those who have helped us in successfully completing the project.

N.Santosh Goud (17K81A0538)

M.Sai Kumar (17K81A0537)

M.Sai Chand (16K81A0542)

K.Sujith (16K81A0531)

TABLEOFCONTENTS

LIST OF FIGURES

LIST OF OUTPUT SCREENS

LIST OF ACRONYMS

KNN	K-NEAREST NEIGHBORS
NSS	NEAREST NEIGHBOR SEARCH
OHLC	OPEN-HIGH-LOW-CLOSE
SES	SINGLE EXPONENTIAL SMOOTHING
WNN	WEIGHTLESS NEURAL NETWORK
NLP	NATURAL LANGUAGE PROCESSING

CHAPTER 1
INTRODUCTION

1. INTRODUCTION

Analyzing financial data in securities has been an important and challenging issue in the investment community. Stock price efficiency for public listed firms is difficult to achieve due to the opposing effects of information competition among major investors and the adverse selection costs imposed by their information advantage. There are two main schools of thought in analyzing the financial markets. The first approach is known as fundamental analysis.

The methodology used in fundamental analysis evaluates a stock by measuring its intrinsic value through qualitative and quantitative analysis. This approach examines a company's financial reports, management, industry, micro and macro-economic factors. The second approach is known as technical analysis. The methodology used in technical analysis for forecasting the direction of prices is through the study of historical market data. Technical analysis uses a variety of charts to anticipate what are likely to happen. The stock charts include candlestick charts, line charts, bar charts, point and figure charts, OHLC (open-high-low-close) charts and mountain charts. The charts are viewable in different time frames with price and volume. There are many types of indicators used in the charts, including resistance, support, breakout, trending and momentum. Several alternatives to approach this type of problem have been proposed, which range from traditional statistical modelling to methods based on computational intelligence and machine learning.

Vanstone and Tan surveyed the works in the domain of applying soft computing to financial trading and investment. They categorized the papers reviewed in the following areas: time series, optimization, hybrid methods, pattern recognition and classification. Within the context of financial trading discipline, the survey showed that most of the research was being conducted in the field of technical analysis. An integrated fundamental and technical analysis model was examined to evaluate the stock price trends by focusing on macro-economic analysis. It also analyzed the company behaviour and the associated industry in relation to the economy which in turn provide more information for investors in their investment decisions. A nearest neighbor search (NNS) method produced an intended result by the use of KNN technique with technical analysis.

1.1 PROJECT OVERVIEW

Describe this project or product and its intended audience, or provide a link or reference to the project charter.

Purpose and Scope of this Specification

Describe the purpose of this specification and its intended audience. Include a description of what is within the scope what is outside of the scope of these specifications. For example:

In scope

This document addresses requirements related to phase 2 of Project A:

modification of Classification Processing to meet legislative mandate ABC.

modification of Labor Relations Processing to meet legislative mandate ABC.

Out of Scope

The following items in phase 3 of Project A are out of scope:

modification of Classification Processing to meet legislative mandate XYZ.

modification of Labor Relations Processing to meet legislative mandate XYZ.

(Phase 3 will be considered in the development of the requirements for Phase 2, but the Phase 3 requirements will be documented separately.)

Product/Service Description

In this section, describe the general factors that affect the product and its requirements. This section should contain background information, not state specific requirements (provide the reasons why certain specific requirements are later specified).

1.2 PROJECT OBJECTIVES

1. Input Design is the process of converting a user-oriented description of the input into a computer-based system. This design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerized system.

2. It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The goal of designing input is to make data entry easier and to be free from errors. The data entry screen is designed in such a way that all the data manipulates can be performed. It also provides record viewing facilities.

3. When the data is entered it will check for its validity. Data can be entered with the help of screens. Appropriate messages are provided as when needed so that the user will not be in maize of instant. Thus the objective of input design is to create an input layout that is easy to follow

1.3 ORGANIZATION OF CHAPTERS

The thesis is organized in the following chapters:

Chapter 1: Introduction

This section discusses about Stock price efficiency for public listed firms is difficult to achieve due to the opposing effects of information competition among major investors and the adverse selection costs imposed by their information advantage. There are two main schools of thought in analyzing the financial markets. The first approach is known as fundamental analysis. The second approach is known as technical analysis. The methodology used in technical analysis for forecasting the direction of prices is through the study of historical market data.

Chapter 2: Literature Survey

This section discusses about there are a large amount of financial information sources in the world that can be valuable research areas, one of these areas is stock prediction and also called stock market mining. Stock prediction becomes increasingly important especially if number of rules could be created to help making better investment decisions in different stock markets.

Chapter 3: Software and Hardware Requirements

To be used efficiently , all computer software needs certain hardware components are other software resources to be present on a computer.These prerequisites are known as system requirements and are often used as guidline as opposed to an absolute rule. Most Software defines two sets of system requirements ; minimum and recomemded .This section outlines minimum software and hardware requirements for deploying the project. Requirements may vary based on utilization and observing performance of pilot projects is recommended prior to scale out.

Chapter 4: Software Development Analysis

The development and implementation of the design parameters. Developer’s code based on the product specifications and requirements agreed upon in the previous stages. Following company procedures and guidelines, front-end developers build interfaces and back-ends while database administrators create relevant data in the database. The programmers also test and review each other's code.

Chapter 5: Project System Design

Design is the stage of the software development process. Here, architects and developers draw up advanced technical specifications they need to create the software to requirements. Stakeholders will discuss factors such as risk levels, team composition, applicable technologies, time, budget, project limitations, method and architectural design.

Chapter 6: Project Coding

A programming project produces a well-designed executing system that solves a specified distributed programming problem. A project code is used to represent a one-time, or intermittent departmental event or activity. Any person can use a project code on a transaction, regardless of the project manager or home organization. This section describes some of the coding templates, outline of various files, class with functionalities, the various methods of input and output parameters.

Chapter 7: Project Testing

The testing phase checks the software for bugs and verifies its performance before delivery to users. In this stage, expert testers verify the product's functions to make sure it performs according to the requirements analysis document. Testers use exploratory testing if they have experience with that software or a test script to validate the performance of individual components of the software. They notify developers of defects in the code. If developers confirm the flaws are valid, they improve the program, and the testers repeat the process until the software is free of bugs and behaves according to requirements.

Chapter 8: Output screens

The output of the programmed project is being screened with the screenshots. This section will contain screenshots of the execution at the intermediate stages of the execution. It will contain all interfaces and final output screens of the project.

Chapter 9: Experimental Results

This section will contain about the experimental results of our project.

CHAPTER 2

LITERATURE SURVEY

2. **LITERATURE SURVEY**

2.1 SURVEY ON BACKGROUND

Financial services companies are developing their products to serve future prediction. There are a large amount of financial information sources in the world that can be valuable research areas, one of these areas is stock prediction and also called stock market mining. Stock prediction becomes increasingly important especially if number of rules could be created to help making better investment decisions in different stock markets.

The genetic algorithm had been adopted by Shin et al. (2005); the number of trading rules was generated for Korea Stock Price Index 200 (KOSPI 200), in Sweden Hellestrom and Homlstrom (1998) used a statistical analysis based on a modified kNN to determine where correlated areas fall in the input space to improve the performance of prediction for the period 1987-1996. Both models mentioned were provided in the Zimbabwe stock exchange to predict the stock prices which included Weightless Neural Network (WNN) model and single exponential smoothing (SES) model Mpofu (2004). Clustering stocks approach was provided by Gavrilov et al. (2004) to group 500 stocks from the Standard & Poor. The data represented a series of 252 numbers including the opening stock price. A fuzzy genetic algorithm was presented by Cao (1977) to discover pair relationship in stock data based on user preferences. The study developed potential guidelines to mine pairs of stocks, stock-trading rules, and markets; it also showed that such approach is useful for real trading. Moreover, other studies adopted kNN as prediction techniques such as (Subha et al., 2012; Liao et al. 2010; Tsai and Hsiao 2010; Qian and Rasheed, 2007)

2.2 CONCLUSIONS ON SURVEY:

The objective of this project is to construct a model to predict stock value movement using the opinion mining and clustering method to predict National Stock Exchange (NSE). It used domain specific approach to predict the stocks from each domain and taken some stock with maximum capitalization. Topics and related opinion of share holders are automatically extracted from the writings in a message board by utilizing

our proposed strategy along side isolating clusters of comparable sort of stocks from others using clustering algorithms. Proposed methodology will give two output set i.e. one from sentiment analysis and another from clustering based prediction with respect to some specialized parameters of stock exchange. By examining both the results an efficient prediction is produced.

CHAPTER 3
SOFTWARE AND HARDWARE REQUIREMENTS

3.SOFTWARE AND HARDWARE REQUIREMENTS:

3.1 SOFTWARE REQUIREMENTS:

Operating System	:	Windows Family or higher version
Techniques	:	JDK 1.7
Data Bases	:	Mysql
Server	:	Apache Tomcat

3.2 HARDWARE REQUIREMENTS:

Processor	:	Pentium-III (or) Higher
Ram	:	64MB (or) Higher
Cache	:	512MB
Hard disk	:	10GB

CHAPTER 3
SOFTWARE DEVELOPMENT ANALYASIS

4. SOFTWARE DEVELOPMENT ANALYASIS:

4.1 OVERVIEW OF PROBLEM:

The rapid progress in digital data acquisition has led to the fast-growing amount of data stored in databases, data warehouses, or other kinds of data repositories. Although valuable information may be hiding behind the data, the overwhelming data volume makes it difficult for human beings to extract them without powerful tools. Easy and quick availability to news information was not possible until the beginning of the last decade. In this age of information, news is now easily accessible, as content providers and content locators such as online news services have sprouted on the World Wide Web. Continuous availability of more news articles in digital form, the latest developments in Natural Language Processing (NLP) and the availability of faster computers lead to the question how to extract more information out of news articles. Financial analysts who invest in stock markets usually are not aware of the stock market behavior. They are facing the problem of stock trading as they do not know which stocks to buy and which to sell in order to gain more profits. All these users know that the progress of the stock market depends a lot on relevant news and they have to deal daily with vast amount of information. They have to analyze all the news that appears on newspapers, magazines and other textual resources. But analysis of such amount of financial news and articles in order to extract useful knowledge exceeds human capabilities. Text mining techniques can help them automatically extracting the useful knowledge out of textual resources.

4.2 DEFINE THE PROBLEM

Methodology for NLP module To exactly predict the stock price is very complex task till the date. Here we are proposing to make a prediction based on news articles using one of the Text Mining concepts like sentiment analysis. We would like to make the prediction system for Indian Stock market. Implementation steps to be followed to make a prediction system are:

1. Gathering of news articles.
2. Perform sentiment analysis on news articles
3. Get Polarity of the text

4. Make a prediction based on current stock price and calculated polarity of the text.

4.3 MODULES OVERVIEW

The sample data contains the apple_yahoo dataset and competitor dataset . The study sample included stock data of five randomly selected companies as a sample training dataset from the period December 31, 2009 to January 4, 2010. Each of these companies has attributes including closing price, low price, and high price , volume, open etc.A brief data analysis is presented with the fundamental concepts of data attributes. The attributes for each company are included in the data analysis. Closing price is the main factor that affects the prediction process for a specific stock based on kNN algorithm.

4.4 DEFINE THE MODULES

In the proposed system the user will be able to perform the following operations.

A.DOWNLOAD DATASET:

B. CORRELATION FOR DATA:

C.DATA PREPROCESSING:

D.KNN WITH UNIFORM WEIGHTS:

E.KNN WITH DISTANCE WEIGHTS:

F.KNN ACCURACY:

4.5 MODULE FUNCTIONALITY

A.DOWNLOAD DATASET:

This is the first step involved in the process. In this we upload the apple_yahoo dataset and competitor dataset from yahoo finance dataset. Hence the datasets are downloaded.

B. CORRELATION FOR DATA:

Once the datasets are uploaded to the system it will correlate the data and find the correlation between Apple and competitor Stock market Dataset.

C.DATA PREPROCESSING:

Once the datasets are correlated it will preprocess the data. The actions such as drop missing values, split labels split train and test are performed by preprocessing. we can also see that the dataset contains total 1752 records and 1226 used for training and 526 used for testing.

D.KNN WITH UNIFORM WEIGHTS:

Run KNN with Uniform Weights' to generate KNN model with uniform weights and calculate its model accuracy

E.KNN WITH DISTANCE WEIGHTS:

Run KNN with distance weights' to calculate accuracy.

F.KNN ACCURACY:

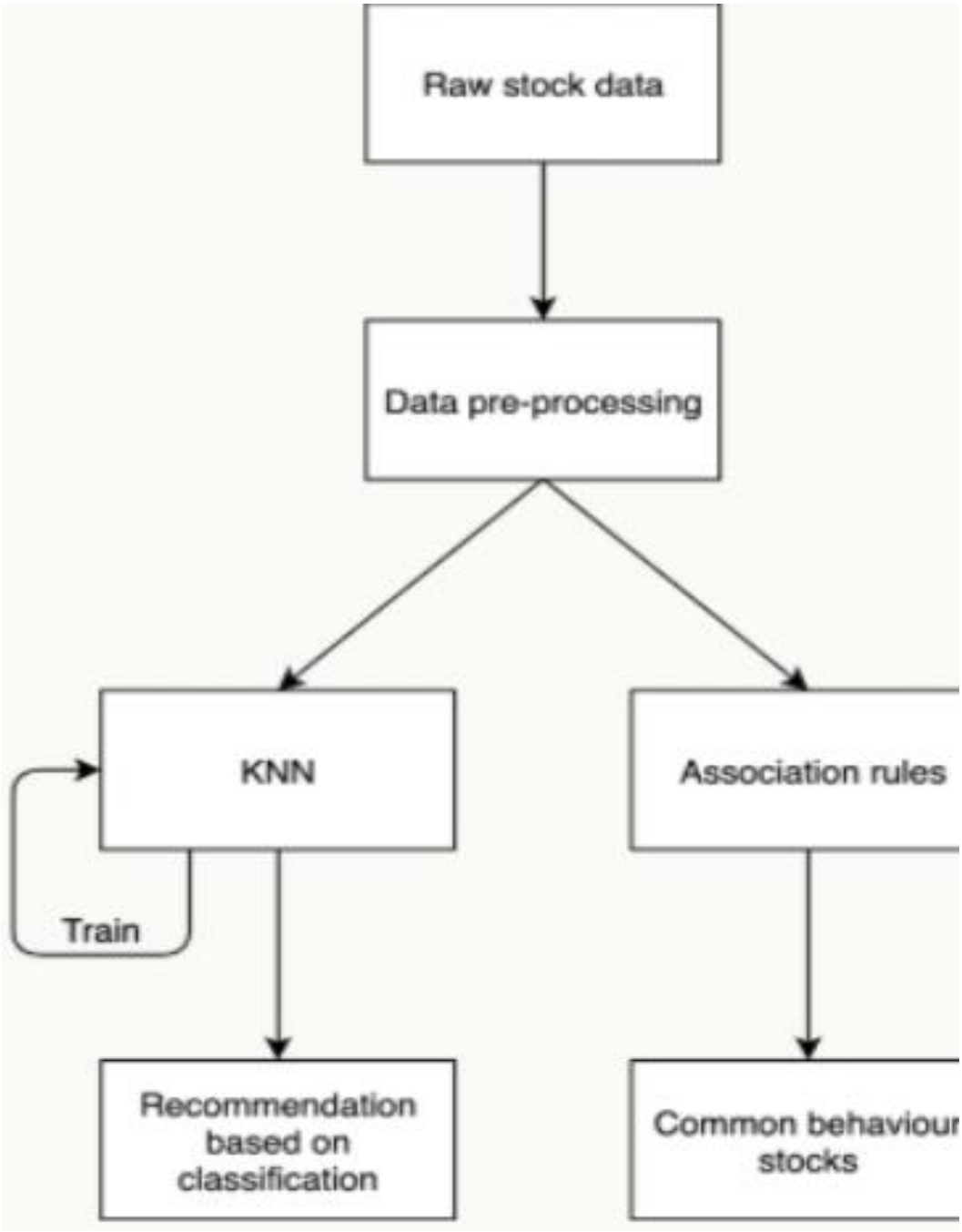
Finally we will get KNN accuracy.

CHAPTER 5
PROJECT SYSTEM DESIGN

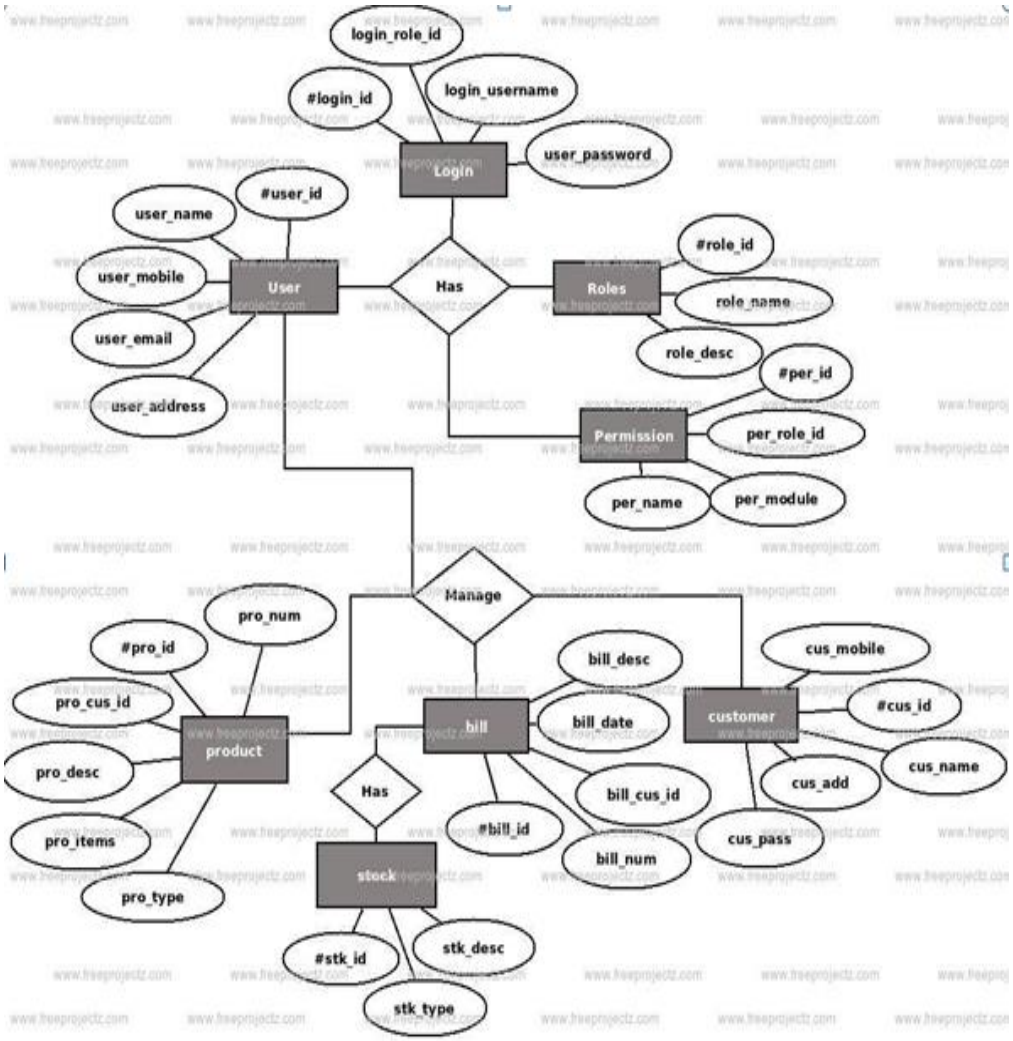
5. PROJECT SYSTEM DESIGN:

5. DFDS IN CASE OF DATABASE PROJECTS

1

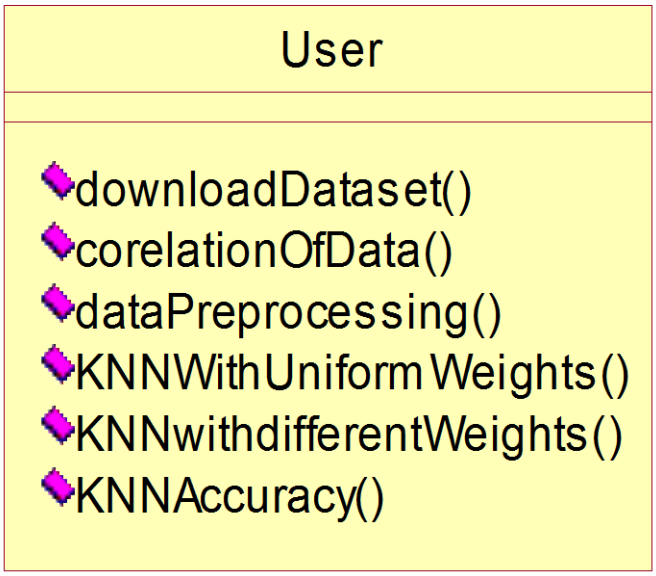


5.2 E-R DIAGRAM

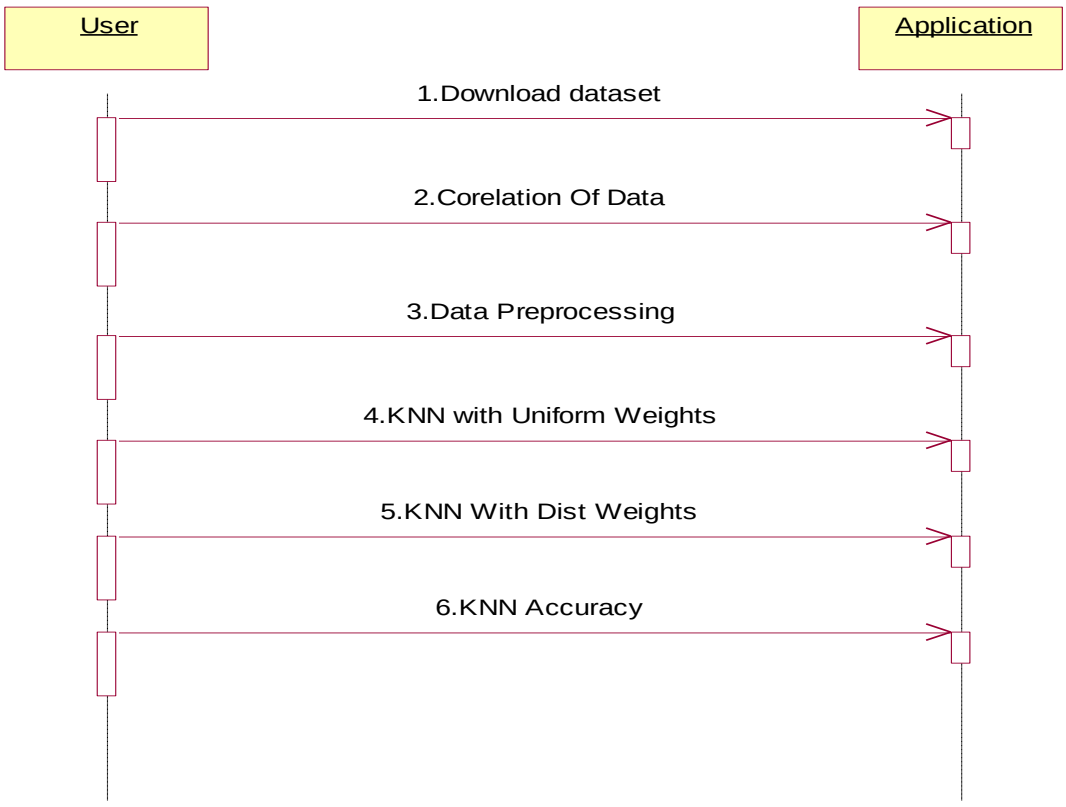


5.3UML DIAGRAMS

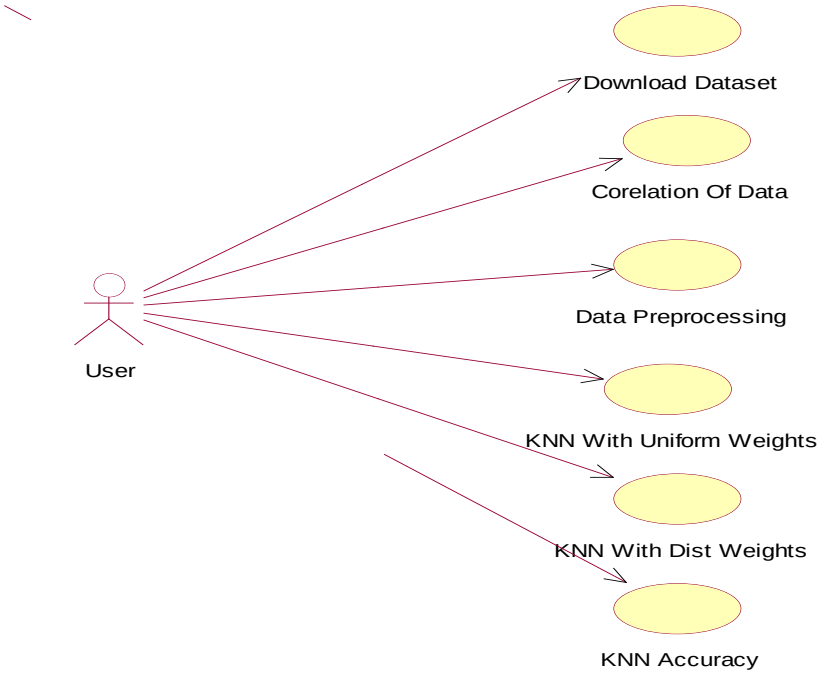
5.3.1 Class Diagram :



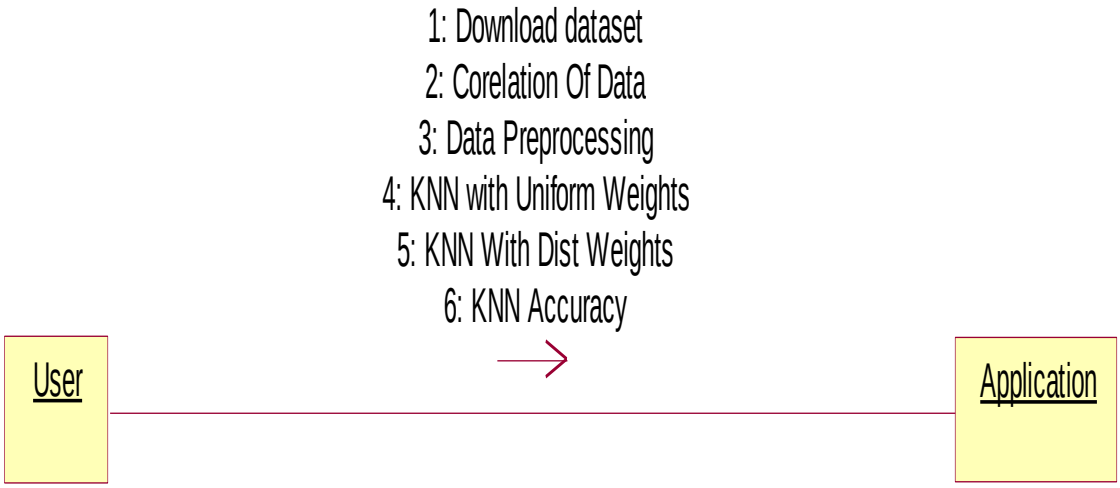
5.3.2Sequence Diagram:



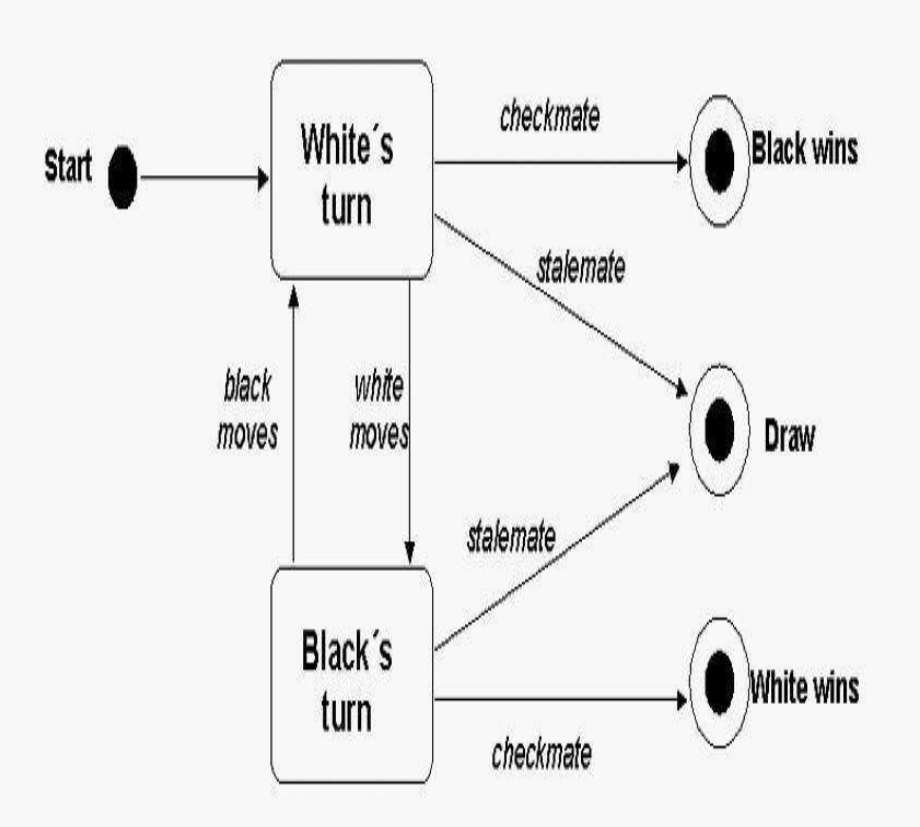
5.3.3Use Case Diagram :



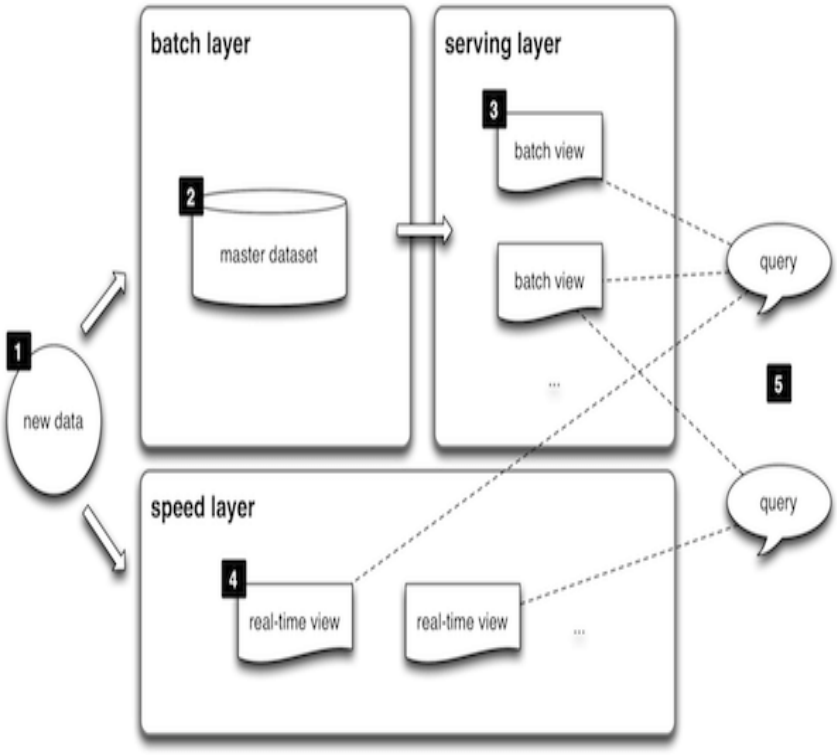
5.3.4Collaboration Diagram :



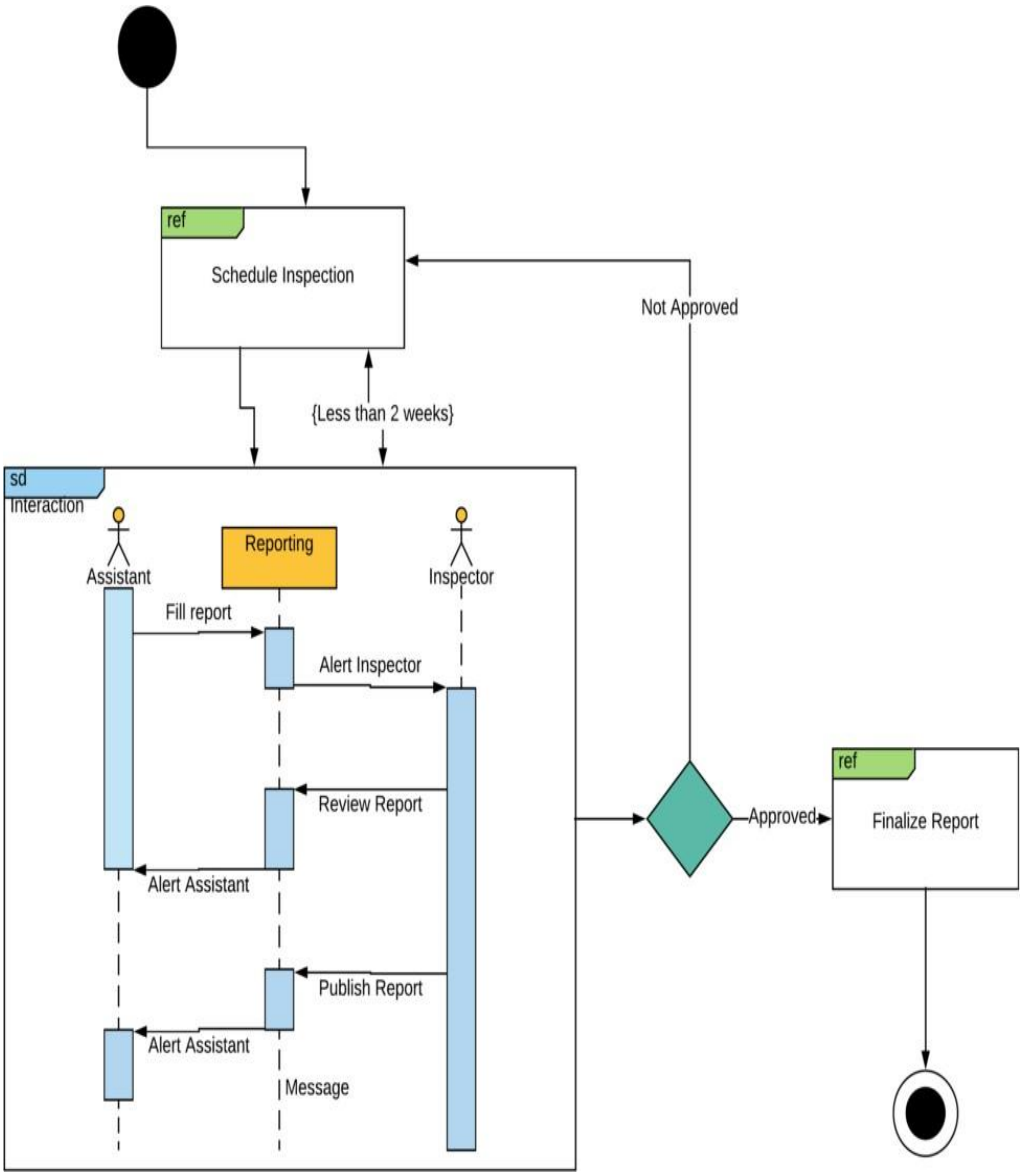
5.3.5State Machine diagram:



5.3.6 Deployement Diagram:



5.3.7 Interaction Overview Diagram:



CHAPTER 6
PROJECT CODING:

6. PROJECT CODING:

6.1 CODE TEMPLATES

```
def loadDataset():
    text.delete('1.0', END)
    global dataFrame
    global dfcomp
    start = datetime.datetime(2010, 1, 1)
    end = datetime.datetime(2017, 1, 11)

    dataFrame = web.DataReader("AAPL", 'yahoo', start, end)

    text.insert(END, "Shape of Apple Stock Dataset: "+str(dataFrame.shape)+"\n\n")
    text.insert(END, "Sample of Apple Stock Data: \n"+str(dataFrame.head(2))+"\n\n")

    dfcomp = web.DataReader(['AAPL', 'GE', 'GOOG', 'IBM', 'MSFT'], 'yahoo', start=start, end=end)['Adj Close']
    text.insert(END, "Shape of Apple Competitor Stock Dataset: " + str(dfcomp.shape) + "\n\n")
    text.insert(END, "Sample of Apple Competitor Stock Data: \n" + str(dfcomp.head(2)) + "\n\n")

    text.insert(END, "Dataset Downloaded from Yahoo Finance Dataset\n\n")

def dfcorr():
    text.delete('1.0', END)
    global dfcomp
```

```
    text.insert(END, "Correlation form Apple Competitor Stock\n\n")
    retscomp = dfcomp.pct_change()
    corr = retscomp.corr()
    text.insert(END, "correlation: \n"+str(corr)+"\n\n")


def dataPreProcess():
    text.delete('1.0', END)
    global dataFrame,dfreg
    global X, y, X_train, X_test, y_train, y_test,X_pred

    text.insert(END,"Data PreProcessing for Apple Stock Dataset\n\n")
    dfreg = dataFrame.loc[:,["Adj Close","Volume"]]
    dfreg["HL_PCT"] = (dataFrame["High"] - dataFrame["Low"]) / dataFrame["Close"] *
100.0
    dfreg["PCT_change"] = (dataFrame["Close"] - dataFrame["Open"]) / dataFrame["Open"] *
100.0

    # Drop missing value
    dfreg.fillna(value=-99999, inplace=True)
    # We want to separate 1 percent of the data to forecast
    forecast_out = int(math.ceil(0.01 * len(dfreg)))

    # Separating the label here, we want to predict the AdjClose
    forecast_col = 'Adj Close'
    dfreg['label'] = dfreg[forecast_col].shift(-forecast_out)
```

```
    X = np.array(dfreg.drop(['label'], 1))


    # Scale the X so that everyone can have the same distribution for linear regression
    X = preprocessing.scale(X)
    # Finally We want to find Data Series of late X and early X (train) for model generation and evaluation
    X_pred = X[-forecast_out:]
    X = X[:-forecast_out]
    # Separate label and identify it as y
    y = np.array(dfreg['label'])
    y = y[:-forecast_out]


    text.insert(END, "X lablels : \n"+str(X)+"\n\n")
    text.insert(END, "Y lablels : \n"+str(y)+"\n\n")
    text.insert(END, "Data spliting into Train and Test")
    X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.3)


    text.insert(END, "number of Train Samples : " + str(len(X_train)) + "\n")
    text.insert(END, "number of Test Sample: " + str(len(X_test)) + "\n")


    text.insert(END, "Data Preprocessing Completed\n\n")


def uniformKNN():
    text.delete('1.0',END)
    global clfknn
    global uniknn
```

```
    # KNN Regression
    clfknn = KNeighborsRegressor(n_neighbors=5)
    clfknn.fit(X_train, y_train)

    uniknn = clfknn.score(X_train, y_train)

    text.insert(END, "Accuracy of KNN with Uniform weights : "+str(uniknn*100)+"\n\n")

def distKNN():
    text.delete('1.0', END)
    global clfknndist,knndistpred
    global distknn,knnunipred
    # KNN Regression
    clfknndist = KNeighborsRegressor(n_neighbors=5,weights='distance')
    clfknndist.fit(X_train, y_train)
    distknn = clfknndist.score(X_train, y_train)

    text.insert(END, "Accuracy of KNN with Uniform weights : "+str(distknn*100)+"\n\n")
def predModel():
    text.delete('1.0', END)
    global clfknndist,clfknn,knnunipred,knndistpred
    global X, y, X_train, X_test, y_train, y_test

    filename = filedialog.askopenfilename(initialdir="Yahoo-Finance-Dataset")
```

```
    test = pd.read_csv(filename)
    text.insert(END, filename + " test file loaded\n"+str(test.columns)+"\n");
    x_pred = np.array(test.drop(['Unnamed: 0'],1))

    text.insert(END, "test Dataset: \n"+str(x_pred)+"\n\n");

    knndistpred = clfknndist.predict(x_pred)

    text.insert(END, "Predict values for KNN with Dist weights: \n" + str(knndistpred) +
"\n\n");

    knnunipred = clfknn.predict(x_pred)

    text.insert(END, "Predict values for KNN with Uni Wights: \n" + str(knnunipred) + "\n\n");

def graph():
    text.delete('1.0', END)

    global uniknn,distknn
    global knnunipred,knndistpred
    global dfreg

    dfreg['Forecast'] = np.nan
    last_date = dfreg.iloc[-1].name
    last_unix = last_date
    next_unix = last_unix + datetime.timedelta(days=1)
```

```
for i in knnunipred:
    next_date = next_unix
    next_unix += datetime.timedelta(days=1)
    dfreg.loc[next_date] = [np.nan for _ in range(len(dfreg.columns) - 1)] + [i]
dfreg['Adj Close'].tail(500).plot()
dfreg['Forecast'].tail(500).plot()
plt.legend(loc=4)
plt.xlabel('Date')
plt.ylabel('Price')
plt.savefig('knnUniformPredGraph.png')
plt.close()

for i in knndistpred:
    next_date = next_unix
    next_unix += datetime.timedelta(days=1)
    dfreg.loc[next_date] = [np.nan for _ in range(len(dfreg.columns) - 1)] + [i]
dfreg['Adj Close'].tail(500).plot()
dfreg['Forecast'].tail(500).plot()
plt.legend(loc=4)
plt.xlabel('Date')
plt.ylabel('Price')
plt.savefig('knnDistPredGraph.png')
plt.close()

height = [uniknn,distknn]
```

```
    bars = ('KNN with uniform weights Accuracy', 'KNN with distance weights Accuracy')
    y_pos = np.arange(len(bars))
    plt.bar(y_pos, height)
    plt.xticks(y_pos, bars)
    plt.show()


font = ('times', 16, 'bold')
title = Label(main, text='Stock Trend Prediction Using KNN')
title.config(bg='PaleGreen2', fg='Khaki4')
title.config(font=font)
title.config(height=3, width=120)
title.place(x=0,y=5)

font1 = ('times', 14, 'bold')
uploadButton = Button(main, text="Download Dataset", command=loadDataset)
uploadButton.place(x=700,y=100)
uploadButton.config(font=font1)

corrButton = Button(main, text="Correlation for Data", command=dfcorr)
corrButton.place(x=700,y=150)
corrButton.config(font=font1)

ppButton = Button(main, text="Data Preprocessing", command=dataPreProcess)
ppButton.place(x=700,y=200)
ppButton.config(font=font1)
```

```
uniformButton = Button(main, text="Run KNN with Uniform Weights",
command=uniformKNN)

uniformButton.place(x=700,y=250)

uniformButton.config(font=font1)


distButton = Button(main, text="Run KNN with Dist Weights", command=distKNN)

distButton.place(x=700,y=300)

distButton.config(font=font1)


predButton = Button(main, text="Predict the Test Data ", command=predModel)

predButton.place(x=700,y=350)

predButton.config(font=font1)


graphButton = Button(main, text="KNN Accuracy", command=graph)

graphButton.place(x=700,y=400)

graphButton.config(font=font1)


font1 = ('times', 12, 'bold')

text=Text(main,height=30,width=80)

scroll=Scrollbar(text)

text.configure(yscrollcommand=scroll.set)

text.place(x=10,y=100)

text.config(font=font1)
```

```
main.config(bg='PeachPuff2')

main.mainloop()
```

6.2 OUTLINE FOR VARIOUS FILES

```
from tkinter import messagebox

from tkinter import *

from tkinter import simpledialog

import tkinter

from tkinter import filedialog

from imutils import paths

from tkinter.filedialog import askopenfilename


import pandas

import datetime

import pandas_datareader.data

from pandas import Series, DataFrame

import matplotlib.pyplot

from matplotlib import style

import matplotlib

from matplotlib import cm

import math

import numpy

from sklearn import preprocessing

from sklearn.model_selection import train_test_split

from sklearn.neighbors import KNeighborsRegressor
```

```
import seaborn

main = tkinter.Tk()

main.title

main.geometry
```

6.3 CLASS WITH FUNCTIONALITY

Let us look into the brief overview of each module

6.3.1.DOWNLOAD DATASET:

This is the first step involved in the process. In this we upload the apple_yahoo dataset and competitor dataset from yahoo finance dataset. Hence the datasets are downloaded.

6.3.2.CORRELATION FOR DATA:

Once the datasets are uploaded to the system it will correlate the data and find the correlation between Apple and competitor Stock market Dataset.

6.3.3.DATA PREPROCESSING:

Once the datasets are correlated it will preprocess the data. The actions such as drop missing values, split labels split train and test are performed by preprocessing. we can also see that the dataset contains total 1752 records and 1226 used for training and 526 used for testing.

6.3.4.KNN WITH UNIFORM WEIGHTS:

Run KNN with Uniform Weights' to generate KNN model with uniform weights and calculate its model accuracy

6.3.5.KNN WITH DISTANCE WEIGHTS:

Run KNN with distance weights' to calculate accuracy.

6.3.6.KNN ACCURACY:

Finally we will get KNN accuracy.

6.4 METHODS INPUT AND OUTPUT PARAMETERS.

We implemented various methods, which include :

loadDataset()

dfcorr()

dataPreProcess

uniformKNN()

distKNN()

preModel()

grapgh()

localDataset method is used to take stock dataset as input, dfcorr method is used to find correlation between various stocks, dataPreProcess takes the downloaded dataset and input process it and split data into training and training heads, unifromKNN using training and testing data generate a KNN Model and calculates accuracy, distKNN using training and testing data generate a KNN Model and calculates accuracy, preModel predicts future values and grapgh takes the predicted values as input and plot graph comparing accuracy of different models.

CHAPTER 7
PROJECT TESTING

7.PROJECT TESTING:

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub-assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

7.1 VARIOUS TEST CASES

Unit testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a 48 business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

Integration testing

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

Functional test

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals. Functional testing is centred on the following items:

• Valid Input : identified classes of valid input must be accepted.

• Invalid Input : identified classes of invalid input must be rejected.

• Functions : identified functions must be exercised.

• Output : identified classes of application outputs must be exercised.

• Systems/Procedures : interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

System Test

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of 49 system testing is the configuration oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

7.2 BLACK BOX TESTING

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box .you cannot “see” into it. The test provides inputs and responds to outputs without considering how the software works.

7.3 WHITE BOX TESTING

Testing White Box Testing is a testing in which in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is purpose. It is used to test areas that cannot be reached from a black box level.

CHAPTER 8
OUTPUT SCREENS

8. OUTPUT SCREENS

8.1 USER INTERFACES

Fig. 8.1.1.Main interface

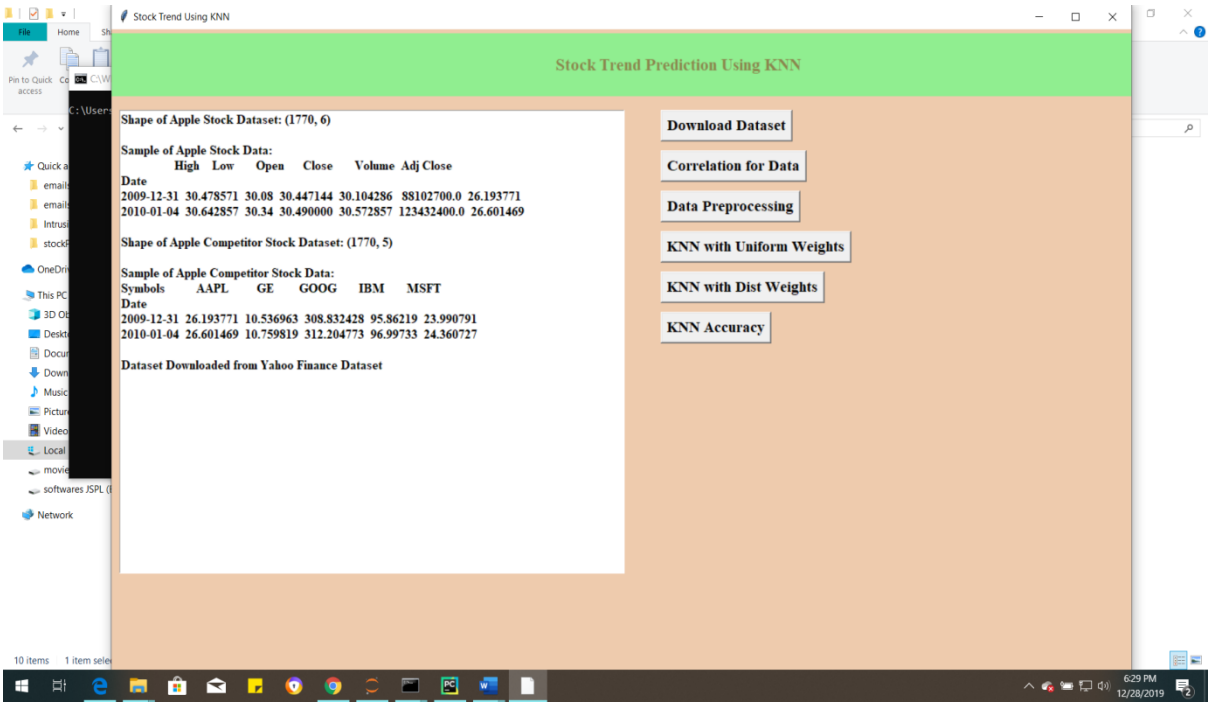


Fig. 8.1.2.Download dataset

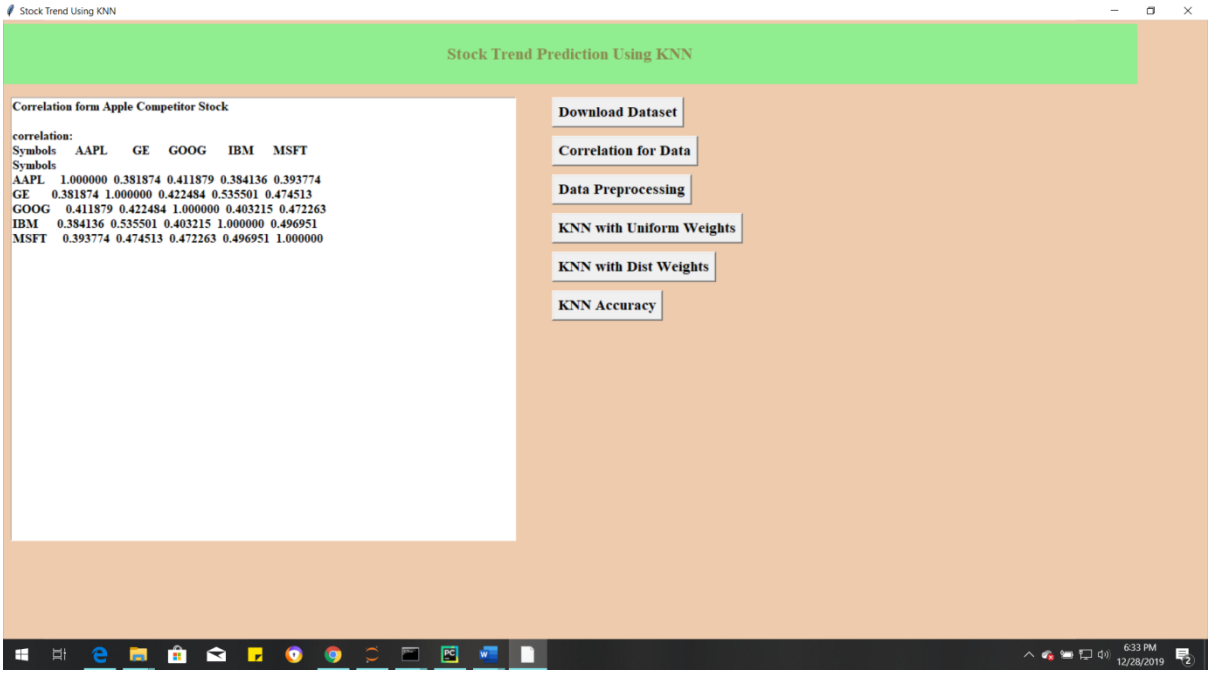


Fig. 8.1.3.Correlation for data

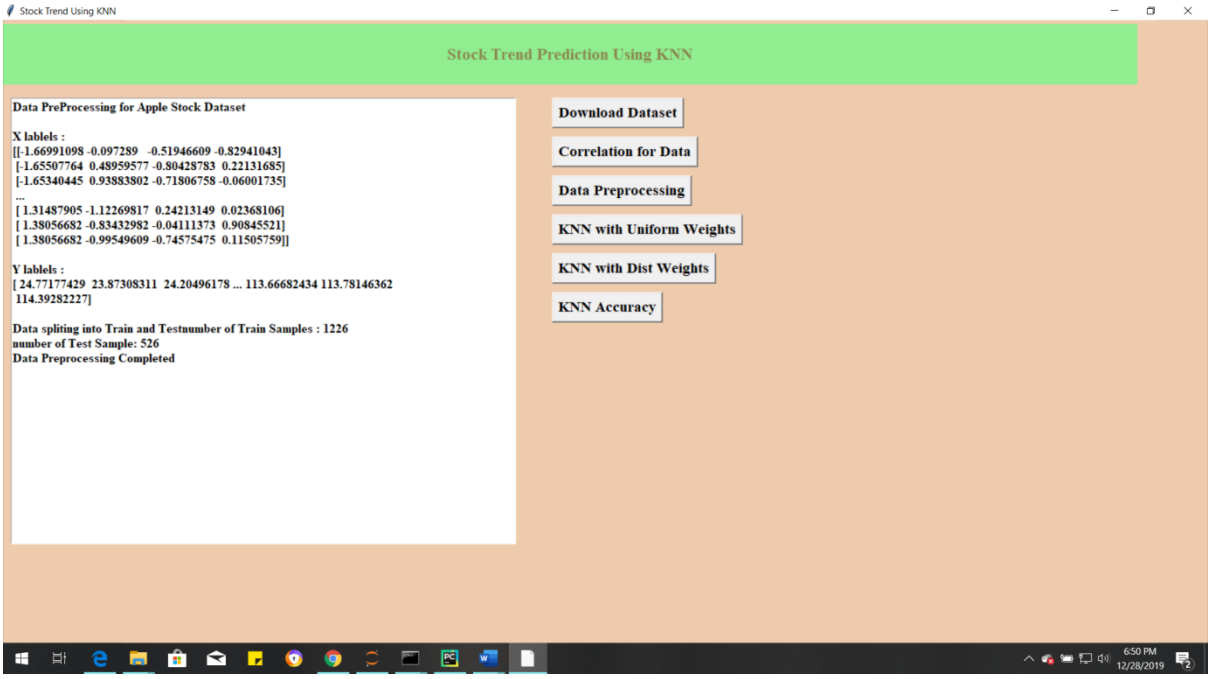


Fig. 8.1.4.Data preprocessing

8.2 OUTPUT SCREENS:

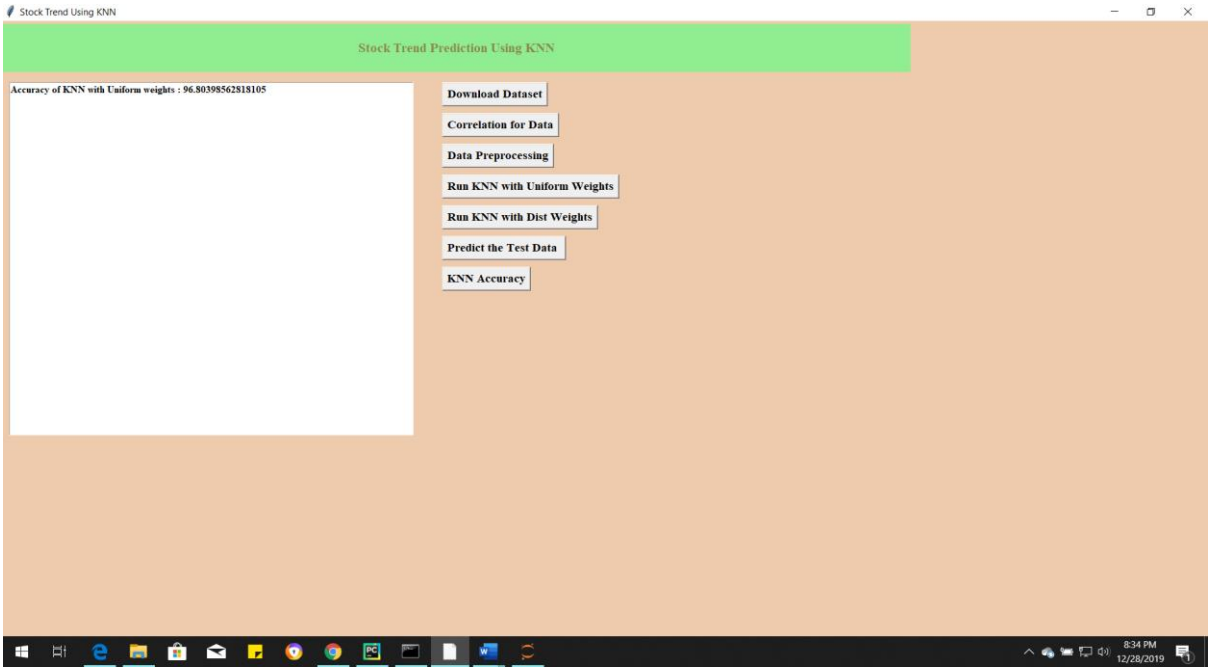


Fig. 8.2.1.KNN with uniform weights

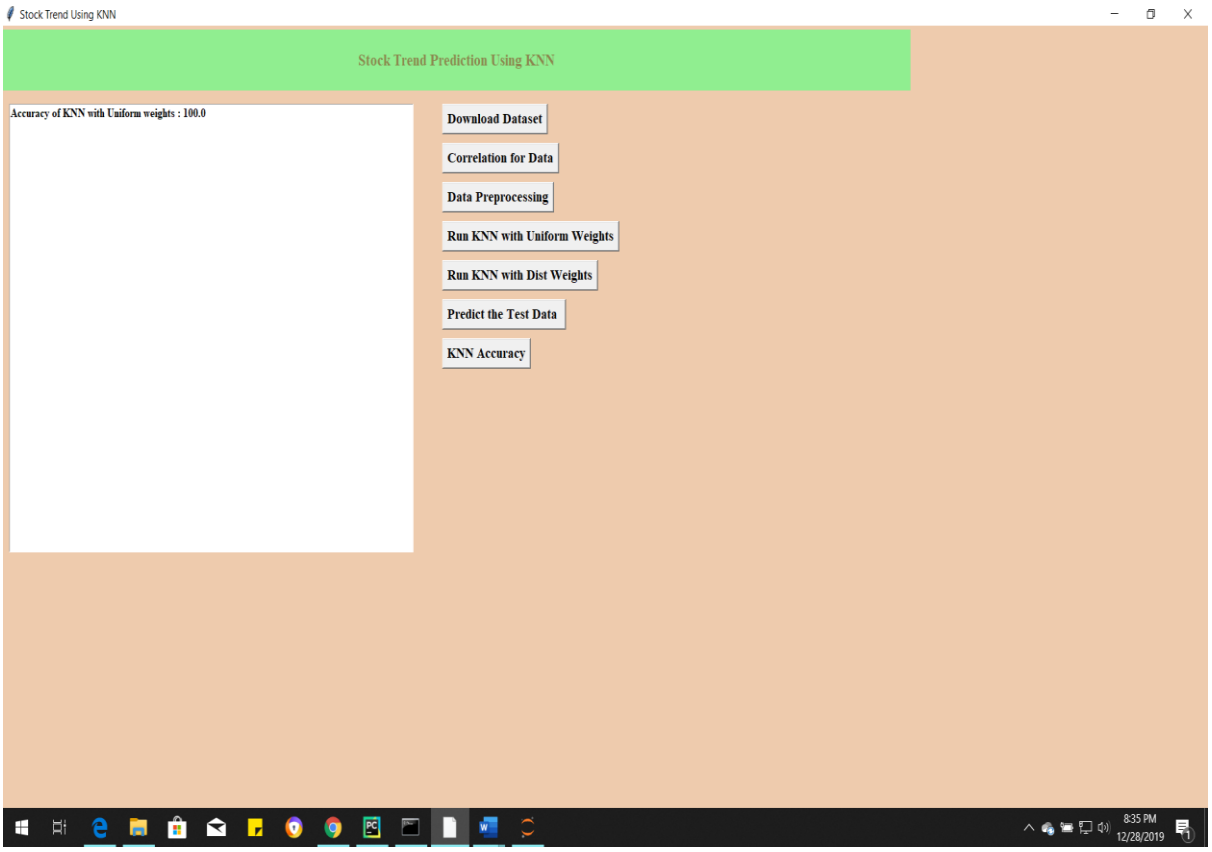


Fig. 8.2.1.KNN with distance weights

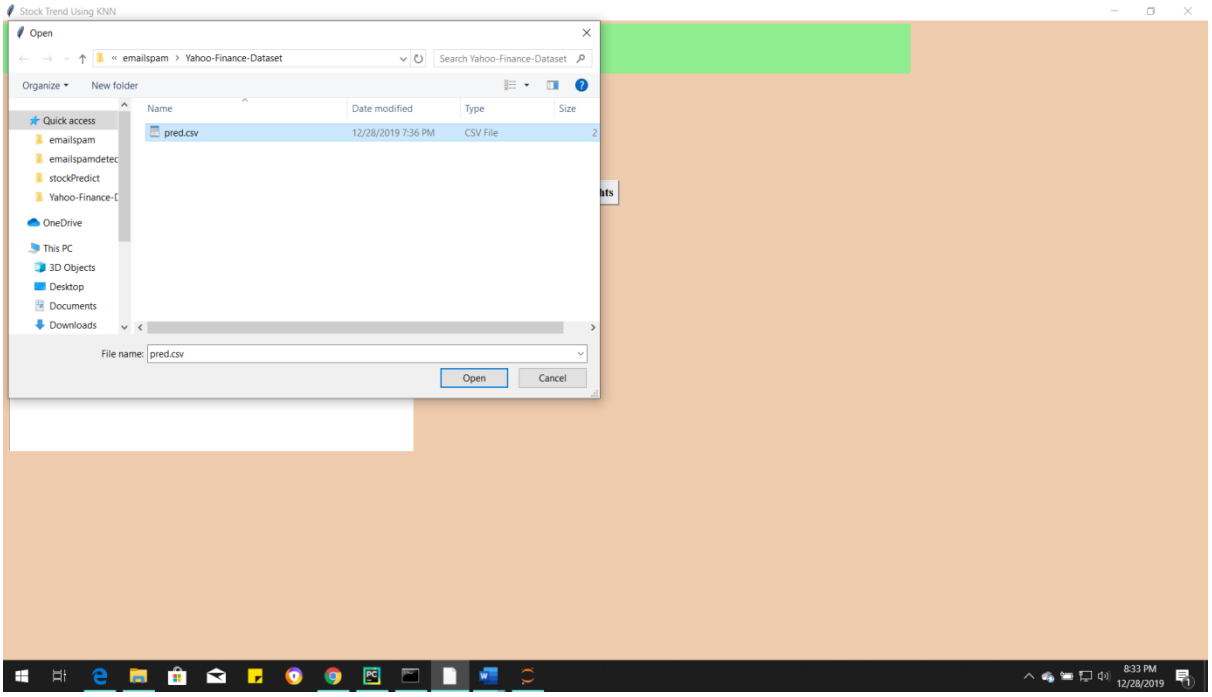


Fig. 8.2.3.Uploading pred.csv file

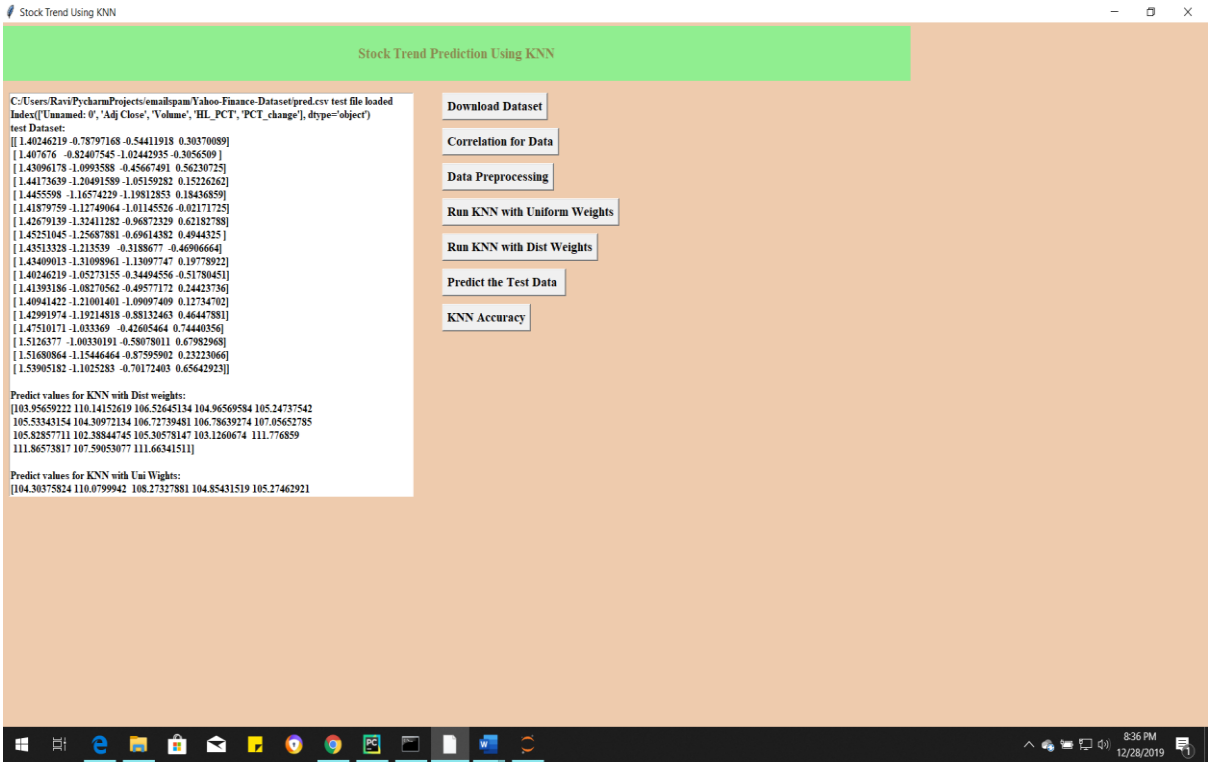


Fig. 8.2.4.Prediction values

CHAPTER 9
EXPERIMENTAL RESULTS

9 EXPERIMENTAL RESULTS:

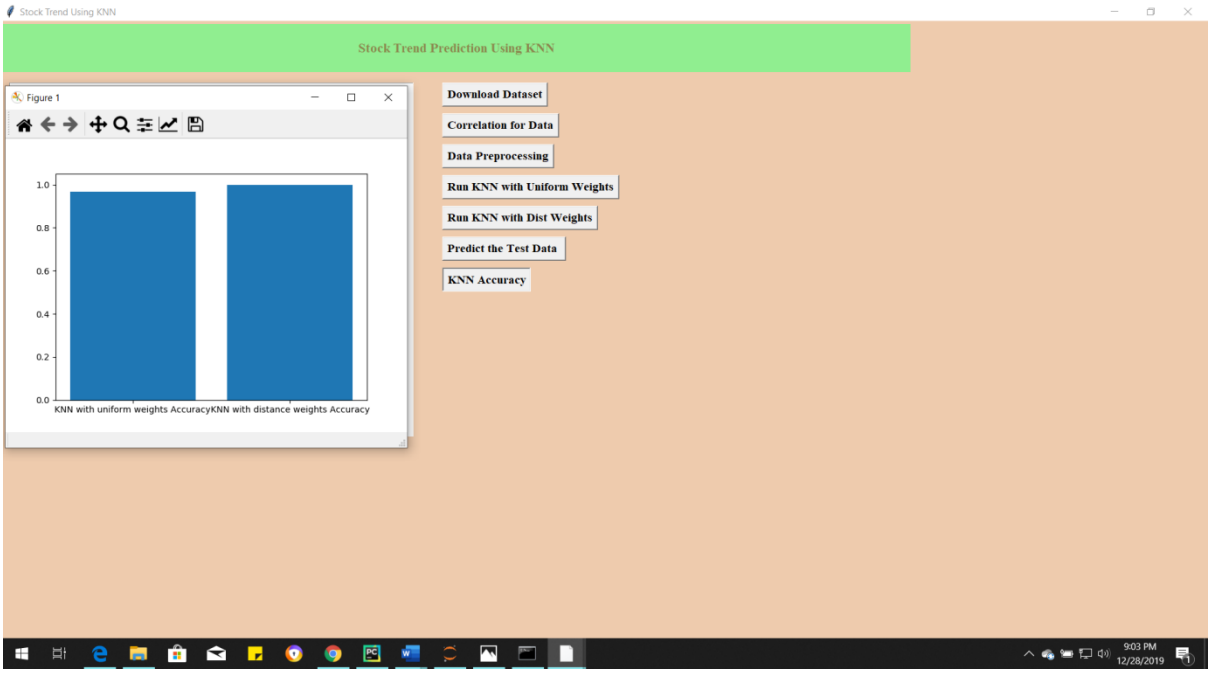


Fig. 9.1.Result

In the above graph we can see that distance weights has little bit better accuracy compare to Uniform weights, in above graph x-axis contains algorithm name and y-axis represents accuracy of that algorithms.

CHAPTER 10
CONCLUSION AND FUTURE ENHANCEMENT

10. CONCLUSION AND FUTURE ENHANCEMENT

The aim of this research is to improve the statistical fitness of the proposed model to overcome a KNN problem due to its computation approach. The KNN classifier can compute the empirical distribution over the Profit and Loss class values in the k number of nearest neighbors. However, the outcome is less than adequate due to sparse data. The KNN classifier has under fitting issue as it does not cater to generalization of sparse data outside the range of nearest neighborhood. We have compared a hybrid KNN-Probabilistic model with four standard algorithms on the problem of predicting the stock price trends. Our results showed that the proposed KNN-Probabilistic model leads to significantly better results compared to the standard KNN algorithm and the other classification algorithms. The limitation of the proposed model is that it applies a binary classification technique. The actual output of this binary classification model is a prediction score in twoclass. The score indicates the model's certainty that the given observation belongs to either the Profit class or Loss class.

For future work, the knowledge component is to transform the binary classification into multiclass classification. The multiclass classification involves observation and analysis of more than the existing two statistical class values. Additional research will include the application of the probabilistic model to multiclass data in order to provide more specific information of each class value. The newly formed multiclass classification will contain five class labels named “Sell”, “Underperform”, “Hold”, “Outperform”, and “Buy”. In numerical values for mapping purpose, we will convert “Sell” to -2 which implies strongly unfavorable; “Underperform” to -1 which implies moderately unfavorable; “Hold” to 0 which implies neutral; “Outperform” to 1 which implies moderately favorable; and “Buy” to 2 which implies strongly favourable

CHAPTER 11
REFERENCES

REFERENCES

1.Choudhry, Rohit, and Kumkum Garg. "A hybrid machine learning system for stock market forecasting." World Academy of Science, Engineering and Technology Volume 39 Issue. 3 pp: 315-31, 2008.

2. Mittermayer, Marc-André. "Forecasting intraday stock price trends with text mining techniques." System Sciences, 2004. Proceedings of the 37th Annual Hawaii International Conference on. IEEE, 2004.

3. Schumaker, Robert P., H. Chen. "Textual analysis of stock market prediction using breaking financial news: The AZFin text system." ACM Transactions on Information Systems (TOIS) Volume 27, Issue 2, pp: 12, 2009.

4. Kim, Kyoung-jae. "Financial time series forecasting using support vector machines." Neurocomputing Volume 55, Issue 1, pp:307-319, 2003.

5. Huang, Wei, Y. Nakamori, S. Wang. "Forecasting stock market movement direction with support vector machine." Computers & Operations Research, Volume 32, Issue 10, pp: 2513-2522, 2005.

6. D. E. Rapach, M. E. Wohar. "In-sample vs. out-ofsample tests of stock return predictability in the context of data mining." Journal of Empirical Finance, Volume 13, Issue 2, pp: 231-247, 2006.

7. B. Wuthrich,V. Cho, S, Leung, J. Zhang, W. Lam. "Daily stock market forecast from textual web data." Systems, Man, and Cybernetics. 1998 IEEE International Conference. Volume 3, 1998.

8. Bao, Yu-Kun, Liu, Guo, Wang "Forecasting stock composite index by fuzzy support vector machines regression." Machine Learning and Cybernetics. Proceedings of 2005 International Conference on. Vol. 6. IEEE, 2005.

9. Fenghua, W. E. N., Jihong, X.I.A.O, Zhifang. "Stock Price Prediction based on SSA and SVM." Procedia Computer Science 31 (2014): 625-631.

10. Kara, Yakup, M. A. Boyacioglu, Ö. K. Baykan. "Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange." Expert systems with Applications, Volume 38, Issue 5,pp: 5311-5319,2011.

11. Cao, Lijuan, and Francis EH Tay. "Financial forecasting using support

vector machines." Neural Computing & Applications, Volume 10, Issue , pp: 184- 192,2001

12. Kim, Kyoung-jae, and I. Han. "Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index." Expert systems with Applications, Volume 19, Issue 2,pp: 125-132,2000.

13. Huang, Cheng-Lung, C. Tsai. "A hybrid SOFM-SVR with a filter-based feature selection for stock market forecasting." Expert Systems with Applications, Volume 36, Issue 2, pp: 1529-1539,2009.

14. Bollen, Johan, H. Mao, X. Zeng. "Twitter mood predicts the stock market." Journal of Computational Science, Volume 2, Issue 1 , pp: 1-8,2011.

15. Dase R. K., Pawar D. D. "Application of Artificial Neural Network for stock market predictions: A review of literature." International Journal of Machine Intelligence, Volume 2, Issue 2, pp: 14-17, 2010.

16. L.J. Cao, F. E.H. Tay. "Support vector machine with adaptive parameters in financial time series forecasting." Neural Networks, IEEE Transactions on 14.6 pp: 1506- 1518,2003.

17. Shen, Shunrong, H. Jiang, T. Zhang. "Stock market forecasting using machine learning algorithms." , 2012.

18. Hegazy, Osman, O. S. Soliman, M. A. Salam. "A Machine Learning Model for Stock Market Prediction." arXiv preprint arXiv, pp:1402.7351, 2014.

19. Olaniyi, S. A. Sulaiman, K. S. Adewole, R. G. Jimoh. "Stock trend prediction using regression analysis–a data mining approach." ARPN Journal of Systems and Software, Volume 1, Issue 4 ,pp: 154-157, 2011.

20. Boser, B., Guyon, I., Vapnik, V. “A training algorithm for optimal Margin classifiers.” Fifth Annual Workshop on Computational Learning Theory, New York: ACM Press 1992.

21. Srivastava, Durgesh K., Lekha Bhambhu. "Data classification using support vector machine." ,2010

PUBLICATIONS:

"Innovations in Computer Networks ,Computer Intelligence and IOT"[: ICICCI-21]

Paper ID: ICICCI-21- 0045

A

PROJECT REPORT

On

MODELLING AND PREDICTING CYBER HACKING BREACHES

Submitted by

1)Mr. D. Abhishek Reddy(17K81A0575) **2) Ms. S. Advaitha Reddy(17K81A05A9)**

3)Mr. K. Pradhyun Reddy(17K81A0588) **4) Ms. N. Rakshitha (17K81A0599)**

in partial fulfillment for the award of the degreeof

BACHELOR OF TECHNOLOGY

IN

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Under the Guidance of

Dr. M. Narayanan

Professor & HOD(CSE)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ST.MARTIN'S ENGINEERING COLLEGE
An Autonomous Institute

Dhulapally, Secunderabad – 500 100

JUNE 2021

BONAFIDE CERTIFICATE

This is to certify that the project entitled Modelling and Predicting Cyber Hacking Breaches, is being submitted by **1**. **D. Abhishek Reddy (17K81A0575), 2. S. Advaitha Reddy (17K81A05A9), 3. K. Pradhyun Reddy (17K81A0588) 4 .N. Rakshitha (17K81A0599)** in partial fulfillment of the requirement for the award of the degree of **BACHELOR OF TECHNOLOGY IN COMPUTER SCIENCE AND ENGINEERING** is recorded of bonafide work carried out by them. The result embodied in this report have been verified and found satisfactory.

<Signature>

Dr. M. NARAYANAN

Department Of CSE

Head of the Department

Dr. M. NARAYANAN

Department Of CSE

Internal Examiner

External Examiner

Place:
Date:

DECLARATION

We, the student of **Bachelor of Technology** in Department of Computer Science and Engineering', session: <2017 – 2021>, St. Martin's Engineering College, Dhulapally, Kompally, Secunderabad, hereby declare that work presented in this Project Work entitled Modelling and Predicting Cyber Hacking Breaches is the outcome of our own bonafide work and is correct to the best of our knowledge and this work has been undertaken taking care of Engineering Ethics. This result embodied in this project report has not been submitted in any university for award of any degree.

D. Abhishek Reddy	17K81A0575
S. Advaitha Reddy	17K81A05A9
K. Pradhyun Reddy	17K81A0588
N. Rakshitha	17K81A0599

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible and whose encouragements and guidance have crowded effects with success.

We extended our deep sense of gratitude to Principal**, Dr. P. SANTOSH KUMAR PATRA**, St. Martin's Engineering College, Dhulapally, for permitting us to undertake this project.

We are also thankful to **Dr. M.NARAYANAN**, Head of the Department, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

We would like to thank **Dr. T. POONGOTHAI,** Department of Computer Science and Engineering (AI & ML) for her encouragement and insightful comments and as well as our project coordinators **Dr. B.RAJALINGAM**, Associate Professor and **Dr. G. GOVINDARAJULU,** Associate Professor, in Department of Computer Science and Engineering for their valuable support.

We would like to express our sincere gratitude and indebtedness to our project supervisor Dr.M.Narayanan, Professor & HOD(CSE), Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

Finally, we express thanks to all those who have helped us successfully to completing this project. Furthermore, we would like to thank our family and friends for their moral support and encouragement.

We express thanks to all those who have helped us in successfully completing the project.

D. Abhishek Reddy 17K81A0575
S. Advaitha Reddy 17K81A05A9
K. Pradhyun Reddy 17K81A0588
N. Rakshitha 17K81A0599

ABSTRACT

Analysing cyber incident data sets is an important method for deepening our understanding of the evolution of the threat situation. This is a relatively new research topic, and many studies remain to be done. In this paper, we report a statistical analysis of a breach incident data set corresponding to 12 years (2005–2017) of cyber hacking activities that include malware attacks. We show that, in contrast to the findings reported in the literature, both hacking breach incident inter-arrival times and breach sizes should be modelled by stochastic processes, rather Than by distributions because they exhibit autocorrelations. Then, we propose particular stochastic process models to, respectively, fit the inter-arrival times and the breach sizes. We also show that these models can predict the inter-arrival times and the breach sizes. In order to get deeper insights into the evolution of hacking breach incidents, we conduct both qualitative and quantitative trend analyses on the data set. We draw a set of cyber security insights, including that the threat of cyber hacks is indeed getting worse in terms of their frequency, but not in terms of the magnitude of their damage.

t

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF OUTPUT SCREENS

LIST OF ACRONYMS

<UML>	Unified Modeling Language
<AI>	Artificial Intelligence
<GB>	Giga Bytes
<RAM>	Random Access Memory
<SSD>	Solid State Drive
<ANN>	Artificial Neural Network
<SVM>	Support Vector Machine

1. INTRODUCTION

1.1 PROJECT OVERVIEW

Cyber hacking is an effort to take advantage of a computing system or a personal network inside a computer.it is the unauthorised access to regulate over network security system for a few illicit purposes. The data breaches are sensitive, confidential or otherwise protected data has been accessed in an unauthorized fashion. cyber-attack is an assault launched by cybercriminals using one or multiple computers or networks. A data breach is a confirmed incident in which sensitive, confidential protected data has been accessed or disclosed in an unauthorized fashion. Data breaches may involve personal health information, trade secrets. Breach of privacy laws can expose individuals to risks such as embarrassment, loss of employment opportunity, loss of business opportunity, physical risks to safety and identity theft. Data breaches are becoming more and more common and some of the most recent data breaches have been the largest on record to date. Data breaches are one of the most devastating cyber incidents

1.2 PROJECT OBJECTIVES

Modelling and predicting cyber hacking breaches is an important, yet challenging, problems. In this paper, we initiate the study of modelling and predicting cyber hacking breaches.in the present study we proposed a stochastic process model to predict the both hacking breach incident inter arrival times and breach sizes. We are using stochastic process model because it does not exhibit auto corelations. Here we will use both qualitative and quantitative trend analysis on the data set.

1.3 ORGANIZATION OF CHAPTERS

In introduction we include this chapter covers the overview of our project and its objectives. Literature Survey – This includes the details of our survey. Software and Hardware Requirements – We specify our software and hardware requirements here. Software Development Analysis – This section includes the problem definition and details of the modules we used in our project. Project System Design – This chapter includes the design part of our project which includes UML diagrams. Project Coding – This section contains the details of our project code. Project Testing – The details of test cases and testing are included in this chapter. Output Screens – This contains the screenshots of how our project looks like when executed. In Experimental Results – This chapter contains the screenshots of our results. Conclusion and Future Enhancements – This covers the conclusion of our project and the possible future developments.

2. LITERATURE SURVEY

2.1 SURVEY ON BACKGROUND

1. What do we know about cyber risk and cyber risk insurance

AUTHORS: Martin Eling, Werner Schnell

Purpose This paper aims to provide an overview of the main research topics in the emerging fields of cyber risk and cyber risk insurance. The paper also illustrates future research directions, from both academic and practical points of view. Design/methodology/approach the authors conduct a literature review on cyber risk and cyber risk insurance using a standardized search and identification process that has been used in various academic articles. Based upon this selection process, a database of 209 papers is created. The main research results findings are extracted and organized in seven clusters. Findings The results illustrate the immense difficulties to insure cyber risk, especially due to a lack of data and modelling approaches, the risk of change and incalculable accumulation risks. The authors discuss various ways to overcome these insurability limitations, such as mandatory reporting requirements, pooling of data or public–private partnerships in which the government covers parts of the risk. Originality/value Despite its increasing relevance for businesses at present, research on cyber risk is limited. Many papers can be found in the IT domain, but relatively little research has been done in the business and economics literature. The authors illustrate where research stands currently and outline directions for future research

2. Heavy-tailed distribution of cyber-risks

AUTHORS: Thomas Maillart, Didier Sornette

With the development of the Internet, new kinds of massive epidemics, distributed attacks, virtual conflicts and criminality have emerged. We present a study of some striking statistical properties of cyber-risks that quantify the distribution and time evolution of information risks on the Internet, to understand their mechanisms, and create opportunities to mitigate, control, predict and insure them at a global scale. First, we report an exceptional stable power-law tail distribution of personal identity losses per event, $\Pr(\text{ID loss} \geq V) \sim 1/V^b$, with $b = 0.7 \pm 0.1$. This result is robust against a surprising strong non-stationary growth of ID losses culminating in July 2006 followed by a more stationary phase. Moreover, this distribution is identical for different types and sizes of targeted organizations. Since $b < 1$, the cumulative number of all losses over all events up to time t increases faster-than-linear with time according to $\simeq t^{1/b}$, suggesting that privacy, characterized by personal identities, is necessarily becoming more and more insecure. We also show the existence of a size effect, such that

the largest possible ID losses per event grow faster-than-linearly as ~S1.3 with the organization size S. The small value b ≃ 0.7 of the power law distribution of ID losses is explained by the interplay between Zipf's law and the size effect. We also infer that compromised entities exhibit basically the same probability to incur a small or large loss.

3. Hype and heavy tails: A closer look at data breaches

AUTHORS: Thomas Maillart , Didier Sornette

With the development of the Internet, new kinds of massive epidemics, distributed attacks, virtual conflicts and criminality have emerged. We present a study of some striking statistical properties of cyber-risks that quantify the distribution and time evolution of information risks on the Internet, to understand their mechanisms, and create opportunities to mitigate, control, predict and insure them at a global scale. First, we report an exceptional stable power-law tail distribution of personal identity losses per event, Pr(ID loss ≥ V) ~ 1/Vb, with b = 0.7 ± 0.1. This result is robust against a surprising strong non-stationary growth of ID losses culminating in July 2006 followed by a more stationary phase. Moreover, this distribution is identical for different types and sizes of targeted organizations. Since b < 1, the cumulative number of all losses over all events up to time t increases faster-than-linear with time according to ≃ t1/b, suggesting that privacy, characterized by personal identities, is necessarily becoming more and more insecure. We also show the existence of a size effect, such that the largest possible ID losses per event grow faster-than-linearly as ~S1.3 with the organization size S. The small value b ≃ 0.7 of the power law distribution of ID losses is explained by the interplay between Zipf's law and the size effect. We also infer that compromised entities exhibit basically the same probability to incur a small or large loss.

4. The Extreme Risk of Personal Data Breaches & The Erosion of Privacy

AUTHORS: Spencer Wheatley , Thomas Maillart , Didier Sornette

Personal data breaches from organizations, enabling mass identity fraud, constitute an extreme risk. This risk worsens daily as an ever-growing amount of personal data are stored by organizations and on-line, and the attack surface surrounding this data becomes larger and harder to secure. Further, breached information is distributed and accumulates in the hands of cyber criminals, thus driving a cumulative erosion of privacy. Statistical modelling of breach data from 2000 through 2015 provides insights into this risk: A current maximum breach size of about 200 million is detected, and is expected to grow by fifty percent over the next five years. The breach sizes are found to be well modelled by an extremely heavy tailed truncated Pareto distribution, with tail exponent parameter decreasing linearly

from 0.57 in 2007 to 0.37 in 2015. With this current model, given a breach contains above fifty thousand items, there is a ten percent probability of exceeding ten million. Projections indicate that the total amount of breached information is expected to double from two to four billion items within the next five years, eclipsing the population of users of the Internet. This massive and uncontrolled dissemination of personal identities raises fundamental concerns about privacy.

5. Modelling Extremal Events: For Insurance and Finance

AUTHORS: Embrechts, Paul, Klüppelberg, Claudia, Mikosch, Thomas

Both in insurance and in finance applications, questions involving extremal events (such as large insurance claims, large fluctuations, in financial data, stock-market shocks, risk management, ...) play an increasingly important role. This much awaited book presents a comprehensive development of extreme value methodology for random walk models, time series, certain types of continuous-time stochastic processes and compound Poisson processes, all models which standardly occur in applications in insurance mathematics and mathematical finance. Both probabilistic and statistical methods are discussed in detail, with such topics as ruin theory for large claim models, fluctuation theory of sums and extremes of aid sequences, extremes in time series models, point process methods, statistical estimation of tail probabilities. Besides summarising and bringing together known results, the book also features topics that appear for the first time in textbook form, including the theory of sub exponential distributions and the spectral theory of heavy-tailed time series. A typical chapter will introduce the new methodology in a rather intuitive (tough always mathematically correct) way, stressing the understanding of new techniques rather than following the usual "theorem-proof" format. Many examples, mainly from applications in insurance and finance, help to convey the usefulness of the new material. A final chapter on more extensive applications and/or related fields broadens the scope further. The book can serve either as a text for a graduate course on stochastics, insurance or mathematical finance, or as a basic reference source. Its reference quality is enhanced by a very extensive bibliography, annotated by various comments sections making the book broadly and easily accessible.

6. Models and measures for correlation in cyber-insurance

AUTHORS: Rainer B¨ohme, Gaurav Kataria

High correlation in failure of information systems due to worms and viruses has been cited as major impediment to cyber-insurance. However, of the many cyber-risk classes that influence failure of information systems, not all exhibit similar correlation properties. In this paper, we introduce a new

classification of correlation properties of cyber-risks based on a twin-tier approach. At the first tier, is the correlation of cyber-risks within a firm i.e., correlated failure of multiple systems on its internal network. At second tier, is the correlation in risk at a global level i.e., correlation across independent firms in an insurer's portfolio. Various classes of cyber-risks exhibit different level of correlation at two tiers, for instance, insider attacks exhibit high internal but low global correlation. While internal risk correlation within a firm influences its decision to seek insurance, the global correlation influences insurers' decision in setting the premium. Citing real data, we study the combined dynamics of the two-step risk arrival process to determine conditions conducive to the existence of cyber-insurance market. We address technical, managerial and policy choices influencing the correlation at both steps and the business implications thereof.

7. Copula-based actuarial model for pricing cyber-insurance policies

AUTHORS: Hemantha S.B. Herath (Canada), Tejaswini C. Herath (Canada)

Cyber-insurance is often suggested as a tool to manage IT security residual risks but the accuracy of premiums is still an open question. Thus, practitioners and academics have argued for more robust and innovative cyber-insurance pricing models. The paper fills this important gap in the literature by developing a cyber-insurance model using the emerging copula methodology. The premiums for first party losses due to virus intrusions are estimated using three types of insurance policy models. Our approach is the first in the information security literature to integrate standard elements of insurance risk with the robust copula methodology to determine cyber insurance premiums.

8. Cyber-risk decision models: To insure it or not?"

AUTHORS: Arunabha Mukhopadhyay , Samir Chatterjee

Security breaches adversely impact profit margins, market capitalization and brand image of an organization. Global organizations resort to the use of technological devices to reduce the frequency of a security breach. To minimize the impact of financial losses from security breaches, we advocate the use of cyber-insurance products. This paper proposes models to help firms decide on the utility of cyber-insurance products and to what extent they can use them. In this paper, we propose a Copula-aided Bayesian Belief Network (CBBN) for cyber-vulnerability assessment (C-VA), and expected loss computation. Taking these as an input and using the concepts of collective risk modelling theory, we also compute the premium that a cyber risk insurer can charge to indemnify cyber losses. Further, to assist cyber risk insurers and to effectively design products, we propose a utility based preferential pricing (UBPP) model. UBPP takes into account risk profiles and wealth of the prospective insured

firm before proposing the premium. Display Omitted Proposed Cyber risk insurance products to minimize the impact of financial loss of security breach. Cyber risk insurance products complement security technology. Our proposed Copula aided Bayesian Belief networks model helps to asses cyber risk. Collective risk & Utility Theory used to compute premium for Cyber risk insurance products. Cyber risks mode for to decide to opt for cyber insurance or not for organizations

2.2 CONCLUSION OF SURVEY

The Privacy Rights Clearinghouse reports 7,730 data breaches between 2005 and 2017, accounting for 9,919,228,821 breached records. The Identity Theft Resource Centre and Cyber Scout reports 1,093 data breach incidents in 2016, which is 40% higher than the 780 data breach incidents in 2015.Data breaches expose 4.1 billion records in first six month of 2019.the first six month of 2019 have seen more than 3800 publicly disclosed breaches exposing an incredible 4.1 billion compromised records. In 2019, the number of data breaches in the United States amounted to 1,473 with over 164.68 million sensitive records exposed. Data breaches have gained attention with the increasing use of digital files and companies and users' large reliance on digital data. State of breach January 2020:at least 7.9 billion records, including credit card numbers, home addresses, phone numbers and other highly sensitive information, have been exposed through data breaches since 2019.

3. SOFTWARE AND HARDWARE REQUIREMENTS

3.1 SOFTWARE REQUIREMENTS

- **Operating system** **:** Windows 7 Ultimate.
- **Coding Language** **:** Python.
- **Front-End** **:** Python.
- **Designing** **:** Html, CSS, JavaScript.
- **Data Base** **:** MySQL.

3.2 HARDWARE REQUIREMENTS

- **System** **:** Intel(R) Core (TM) i5-7300HQ CPU @ 2.50GHz
- **Hard Disk** **:** 1TB
- **Mouse** **:** Optical Mouse.
- **Ram** **:** 4GB.

4. SOFTWARE DEVELOPMENT ANALYASIS

4.1 OVERVIEW OF PROBLEM

Modelling and predicting cyber hacking breaches is an important, yet challenging, problem. In this project, we initiate the study of modeling and predicting cyber hacking breaches. Here we are using a ARMAGARCH process for the evaluation of breaching incidents. We used Support Vector Machine Algorithm for classification purpose. In the present project we proposed a stochastic process model to predict the both hacking breach incident inter arrival times and breach sizes. We show that stochastic processes should be used, instead of distributions because they exhibit auto-correlation. We conducted both qualitative and quantitative analysis to draw further insights. We draw a set of cybersecurity insights, including that the threat of cyber hacks is indeed getting worse in terms of their frequency, but not in terms of the magnitude of their damage.

4.2 DEFINE THE PROBLEM

Cyber hacking is an effort to take advantage of a computing system or a personal network inside a computer. It is the unauthorized access to regulate over network security system for a few illicit purposes. Data breaches have gained attention with the increasing use of digital files and companies and users large reliance in digital data. A data breach occurs when a cybercriminal successfully infiltrates a data source and extracts sensitive information. This can be done physically by accessing a computer or network to steal local files or by bypassing network security remotely. In this project we attempt to model and predict these breaches so that we can prevent any damage that can occur to the data. However, there are many open problems that are left for future research. For example, it is challenging to investigate how to predict the extremely large values and how to deal with missing data. It is also worthwhile to estimate the exact occurring times of breach incidents. Finally, more research needs to be conducted towards understanding the predictability of breach incidents. The contemporary attacks are done so as to increase raised or higher access benefits. Through the cotemporary assaults, the aggressor can increase managerial benefits of the framework enduring an onslaught.

4.3 MODULES OVERVIEW

This project develops a framework which provides protection to the data of the user. There are four modules present in this project. Using these the authorized user and admin can upload the data to the database. The access of data from the database can be given by administrators. Uploaded data are managed by admin and admin is the only person to provide the rights to process

the accessing details. If user is access the data with wrong attempts then, users are blocked accordingly It also maintains the secrecy of the data using a key. Finally, data analysis is done with the help of a graph. Data is applied to the graph in order to get the best analysis and prediction of dataset.

4.4 DEFINE THE MODULES

This project mainly consists of 4 modules. They are: -

- Upload data
- Access details
- User permissions
- Data analysis

4.5 MODULE FUNCTIONALITY

1. Upload Data: -

The data resource to database can be uploaded by both administrator and authorized user. The data can be uploaded with key in order to maintain the security of the data that is not released without knowledge of user. The users are authorized based on their details that are shared to admin and admin can authorize each user. Only Authorized users are allowed to access the system and upload or request for files. Before uploading the data the user has to create a username and password in order to login. After logging in the user can upload the data regarding the entitites and organization.

2. Access Details: -

The access of data from the database can be given by administrators. Uploaded data are managed by admin and admin is the only person to provide the rights to process the accessing details and approve or unapproved users based on their details. The user can send a request to the admin to regarding access to data. If the admin approves the request, then the user can view their details or make any changes to their details if necessary. The user can upload or download data only by the approval of the admin.

3. User Permission: -

The data from any resources are allowed to access the data with only permission from administrator. Prior to access data, users are allowed by admin to share their data and verify the

details which are provided by user. If user is accessing the data with wrong attempts, then, users are blocked accordingly. If user is requested to unblock them, based on the requests and previous activities admin is unblock users. Only admin can approve users so that the secrecy of the data is maintained and unauthorized users do not get access to any sensitive data.

4. Data Analysis: -

Data analyses are done with the help of graph. The collected data are applied to graph in order to get the best analysis and prediction of dataset and given data policies. In the graph, the number of breaches are mapped on the x-axis and the year of occurance is mapped on the y-axis. The dataset can be analyzed through this pictorial representation in order to better understand of the data details.

5. PROJECT SYSTEM DESIGN

5.1 ARCHITECTURE DIAGRAM

An architectural diagram is a diagram of a system that is used to abstract the overall outline of the software system and the relationships, constraints, and boundaries between components. It is an important tool as it provides an overall view of the physical deployment of the software system and its evolution roadmap. Below architecture diagram is a three tier architecture. The data resource to database can be uploaded by both administrator and authorized user. The users are authorized based on their details that are shared to admin and admin can authorize each user. Only authorized users are allowed to access the system and upload or request for files. Prior to access data, users are allowed by admin to share their data and verify the details which are provided by user. If user is access the data with wrong attempts then, users are blocked accordingly. If user is requested to unblock them, based on the requests and previous activities admin is unblock users.

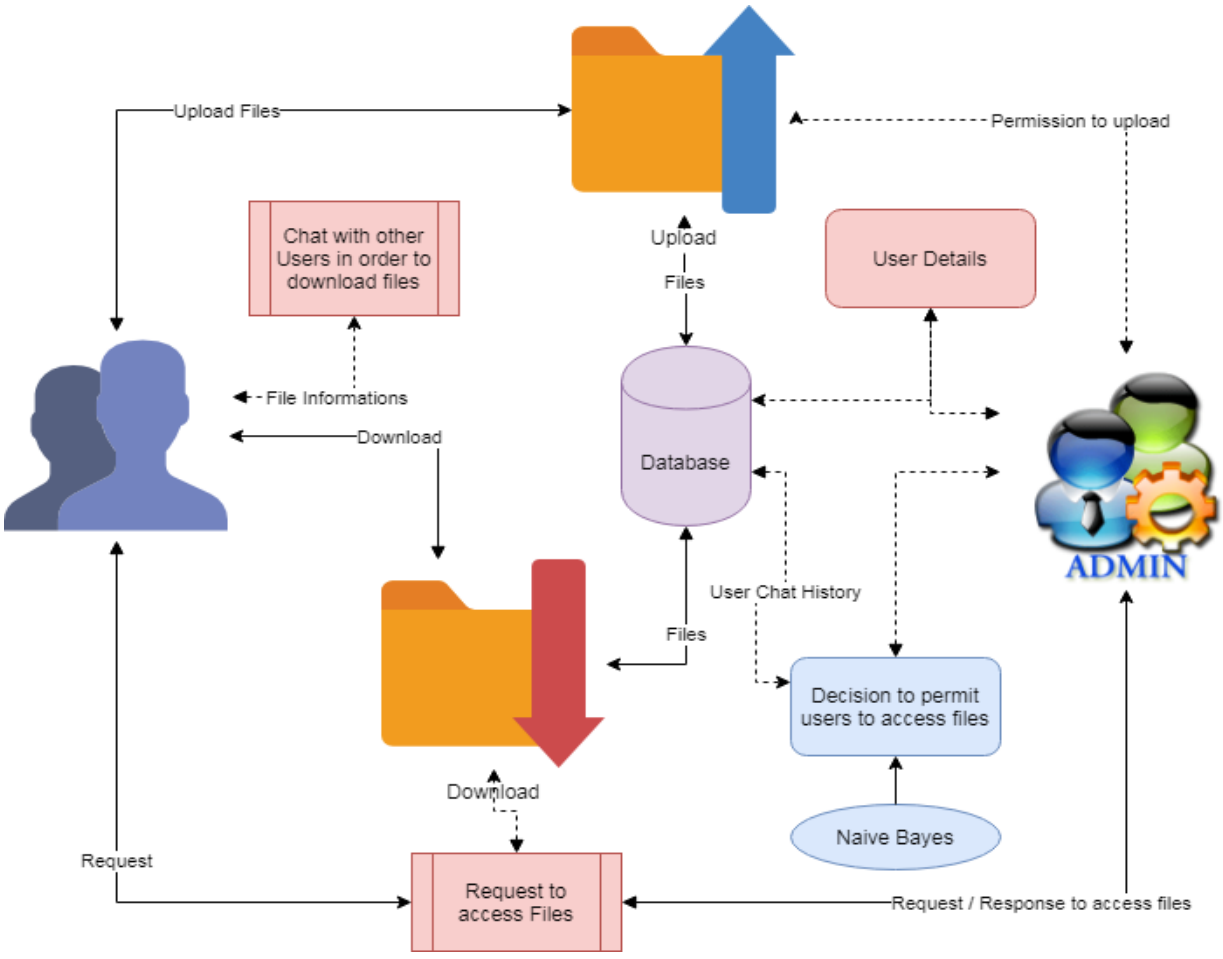


Figure 5.1- Architecture Diagram

5.2 DATA FLOW DIAGRAMS

A data flow diagram (DFD) maps out the flow of information for any process or system. It uses defined symbols like rectangles, circles and arrows, plus short text labels, to show data inputs, outputs, storage points and the routes between each destination.

User:

Figure 5.2 shows the data flow diagram for the user. First the user will login and is verified by the username and password entered. If wrong username and password are entered, then the user is termed as unauthorized. If the right username and password are entered then the user can upload documents and view their details. User can also send a download request to the admin for downloading their data. Finally, the user can also send feedback to the admin.

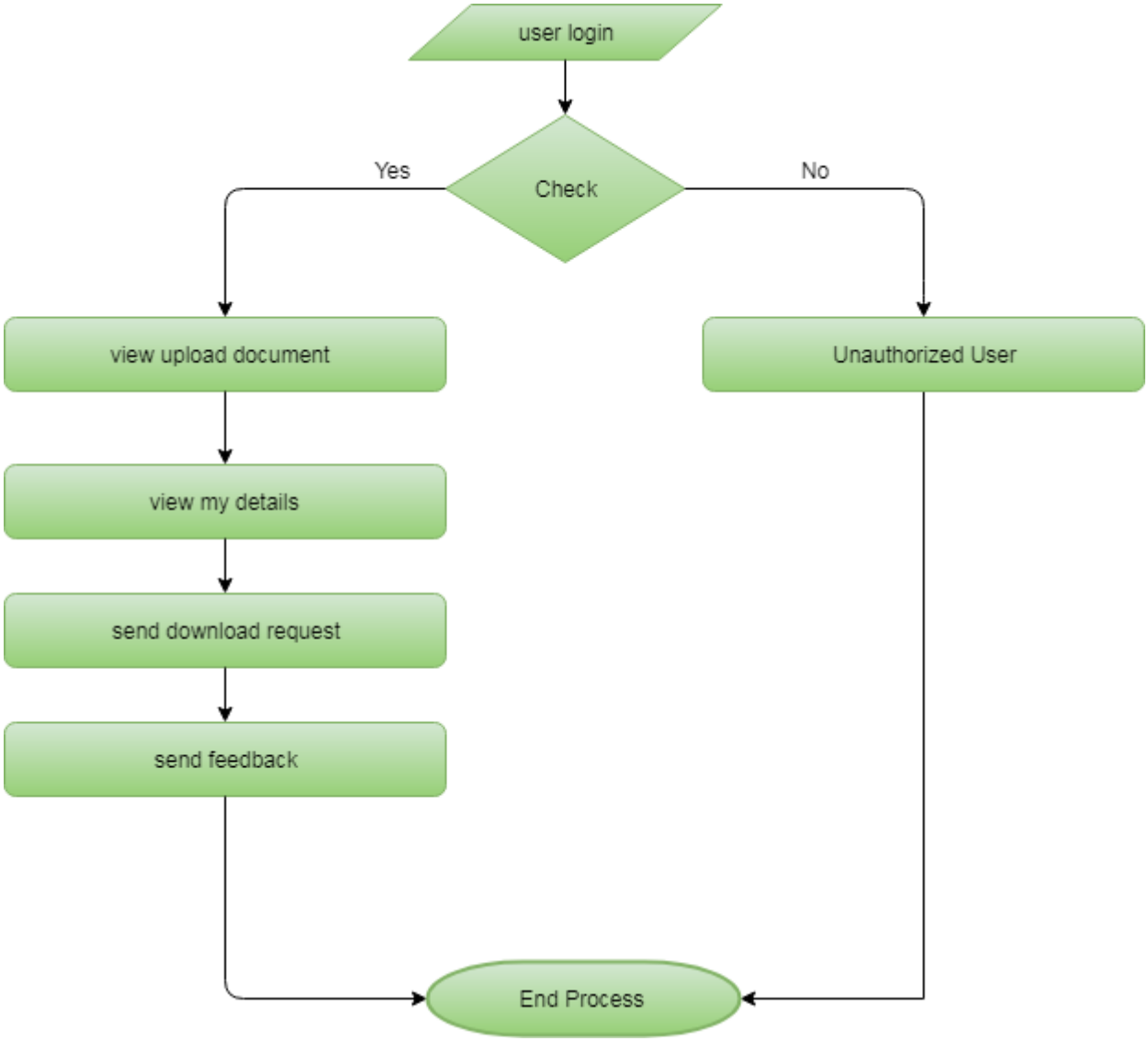


Figure 5.2 Data Flow Diagram (User)

Admin:

Figure 5.3 shows the data flow diagram for the admin. The admin will login and is verified by the username and password entered. If wrong username and password are entered, then the admin is termed as unauthorized. If the right username and password are entered then the admin can view the uploaded documents and view the user details. Admin can also view the download request sent by the user for downloading their data. Finally, the admin can also view feedback sent by the user.

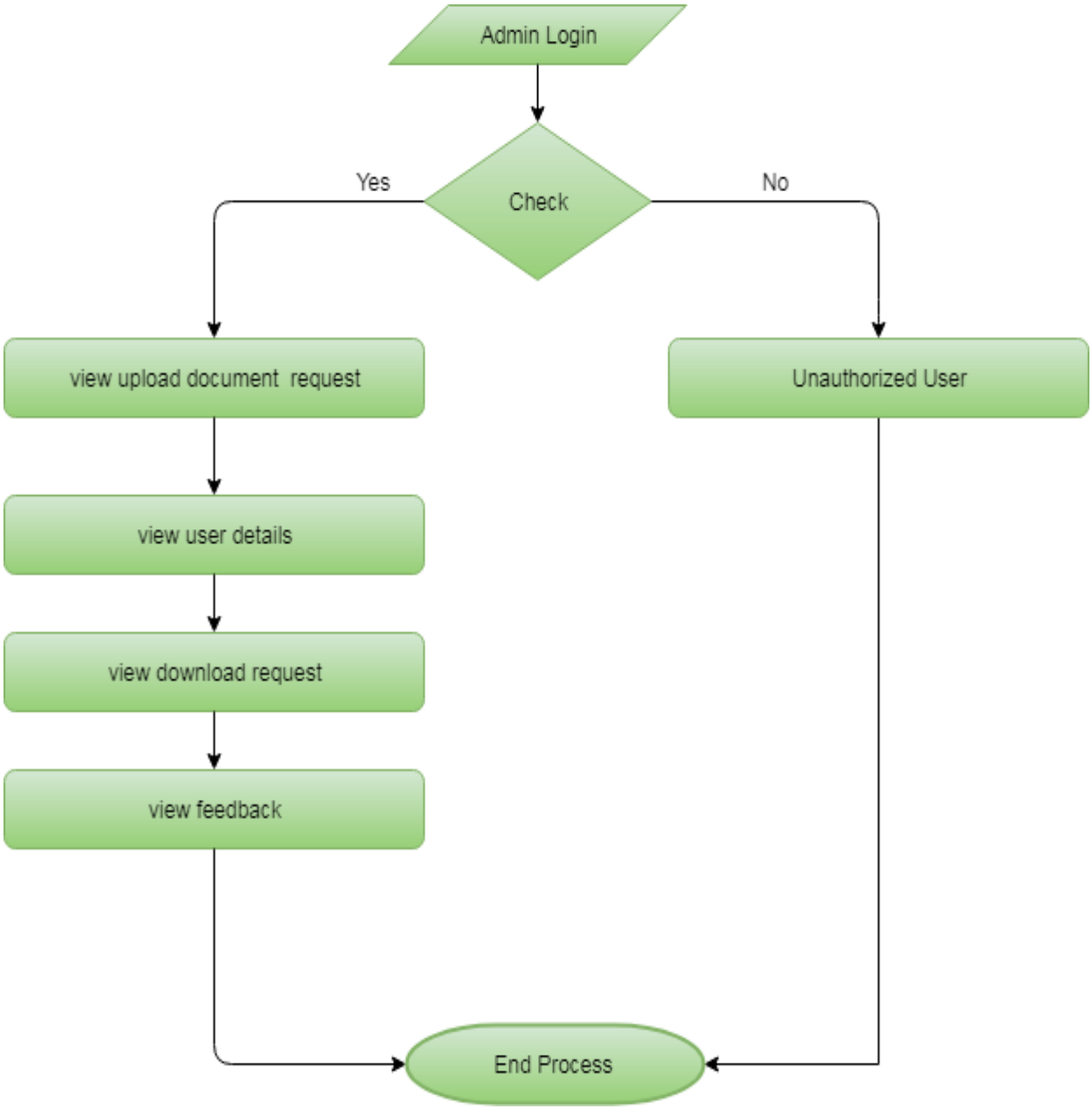


Figure 5.3- Data Flow Diagram (Admin)

5.3 E-R DIAGRAMS

An Entity Relationship (ER) Diagram is a type of flowchart that illustrates how “entities” such as people, objects or concepts relate to each other within a system. ER Diagrams are most often used to design or debug relational databases in the fields of software engineering, business information systems, education and research.

User:

Figure 5.4 shows the E-R diagram for the user. First the user will login and is verified by the username and password entered. If wrong username and password are entered, then the user is termed as unauthorized. If the right username and password are entered then the user can upload documents and view their details. User can also send a download request to the admin for downloading their data. Finally, the user can also send feedback to the admin.

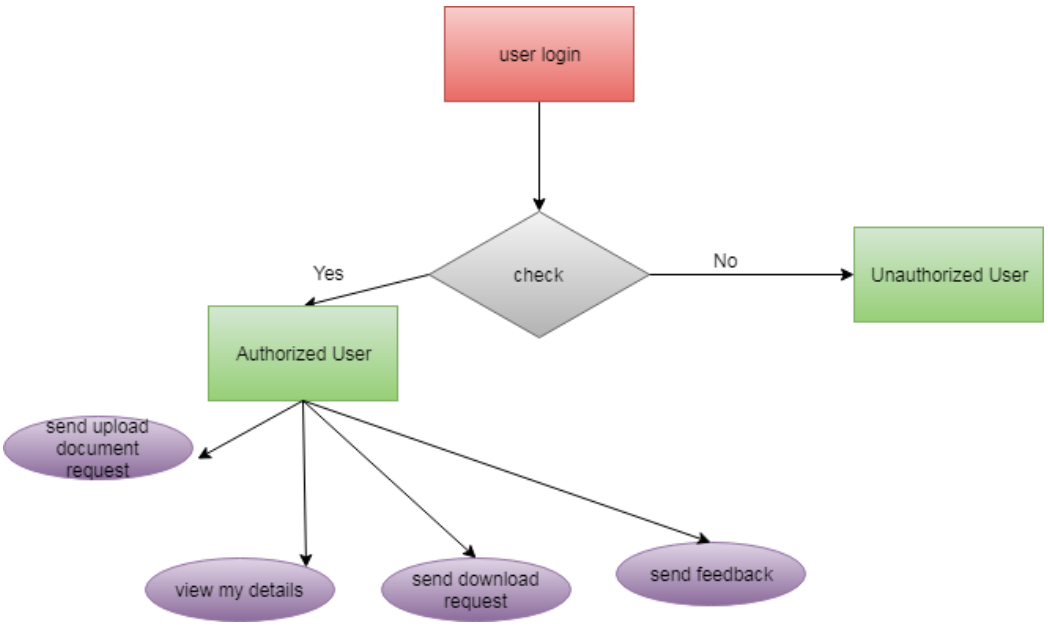


Figure 5.4- ER Diagram (User)

Admin:

Figure 5.5 shows the E-R diagram for the admin. The admin will login and is verified by the username and password entered. If wrong username and password are entered, then the admin is termed as unauthorized. If the right username and password are entered then the admin can view the uploaded documents and view the user details. Admin can also view the download request sent by the user for downloading their data. Finally, the admin can also view feedback sent by the user.

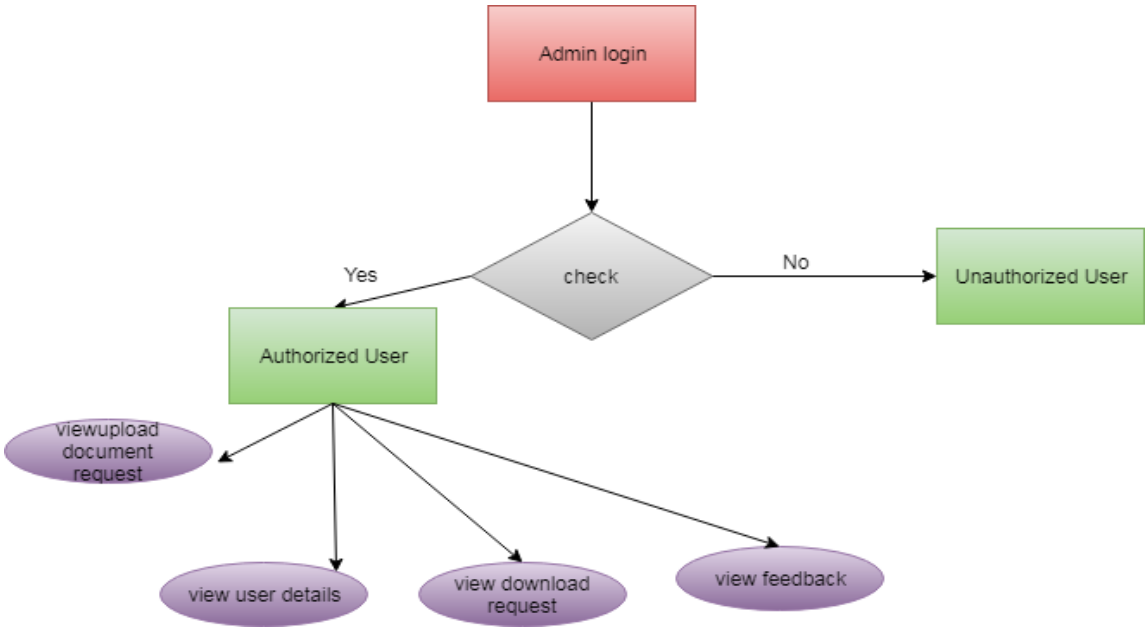


Figure 5.5- ER Diagram (Admin)

5.4 UML DIAGRAMS

UML stands for Unified Modeling Language. UML is a standardized general-purpose modeling language in the field of object-oriented software engineering. The standard is managed, and was created by, the Object Management Group.

The goal is for UML to become a common language for creating models of object-oriented computer software. In its current form UML is comprised of two major components: a Meta-model and a notation. In the future, some form of method or process may also be added to; or associated with, UML.

The Unified Modeling Language is a standard language for specifying, Visualization, Constructing and documenting the artifacts of software system, as well as for business modeling and other non-software systems.

The UML represents a collection of best engineering practices that have proven successful in the modeling of large and complex systems.

The UML is a very important part of developing objects-oriented software and the software development process. The UML uses mostly graphical notations to express the design of software projects.

GOALS:

The Primary goals in the design of the UML are as follows:

1. Provide users a ready-to-use, expressive visual modeling Language so that they can develop and exchange meaningful models.
2. Provide extendibility and specialization mechanisms to extend the core concepts.
3. Be independent of particular programming languages and development process.
4. Provide a formal basis for understanding the modeling language.
5. Encourage the growth of OO tools market.
6. Support higher level development concepts such as collaborations, frameworks, patterns and components.
7. Integrate best practices.

USE CASE DIAGRAM

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.

User:

Figure 5.6 shows the use case diagram for the user. First the user will login and is verified by the username and password entered. If wrong username and password are entered, then the user is termed as unauthorized. If the right username and password are entered then the user can upload documents and view their details. User can also send a download request to the admin for downloading their data. Finally, the user can also send feedback to the admin.

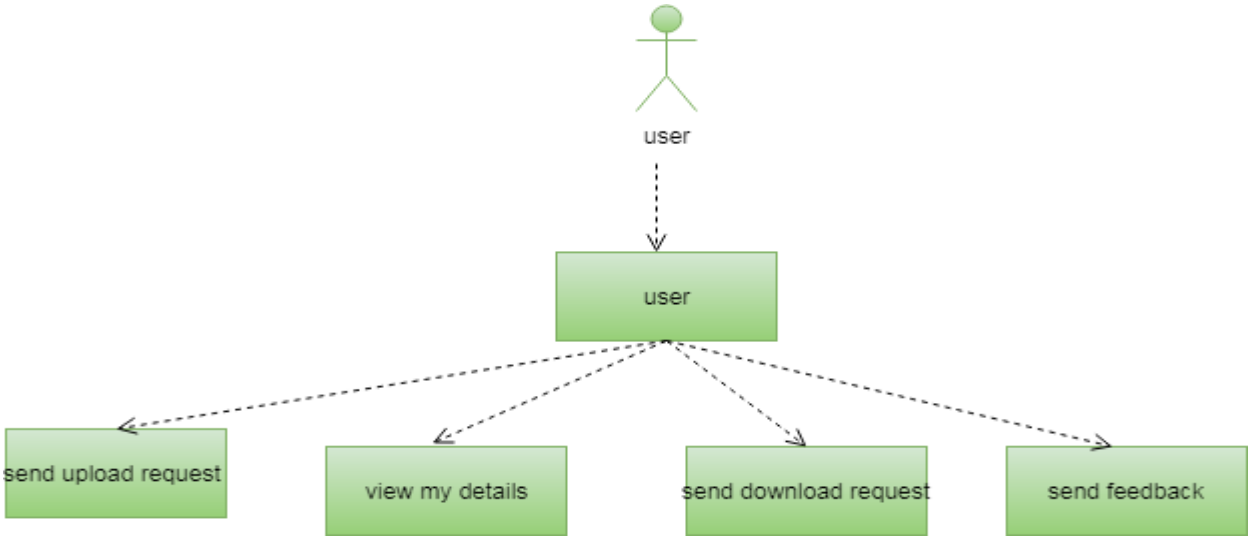


Figure 5.6- Use Case Diagram (User)

Admin:

Figure 5.7 shows the use case diagram for the admin. The admin will login and is verified by the username and password entered. If wrong username and password are entered, then the admin is termed as unauthorized. If the right username and password are entered then the admin can view the uploaded documents and view the user details. Admin can also view the download request sent by the user for downloading their data. Finally, the admin can also view feedback sent by the user.

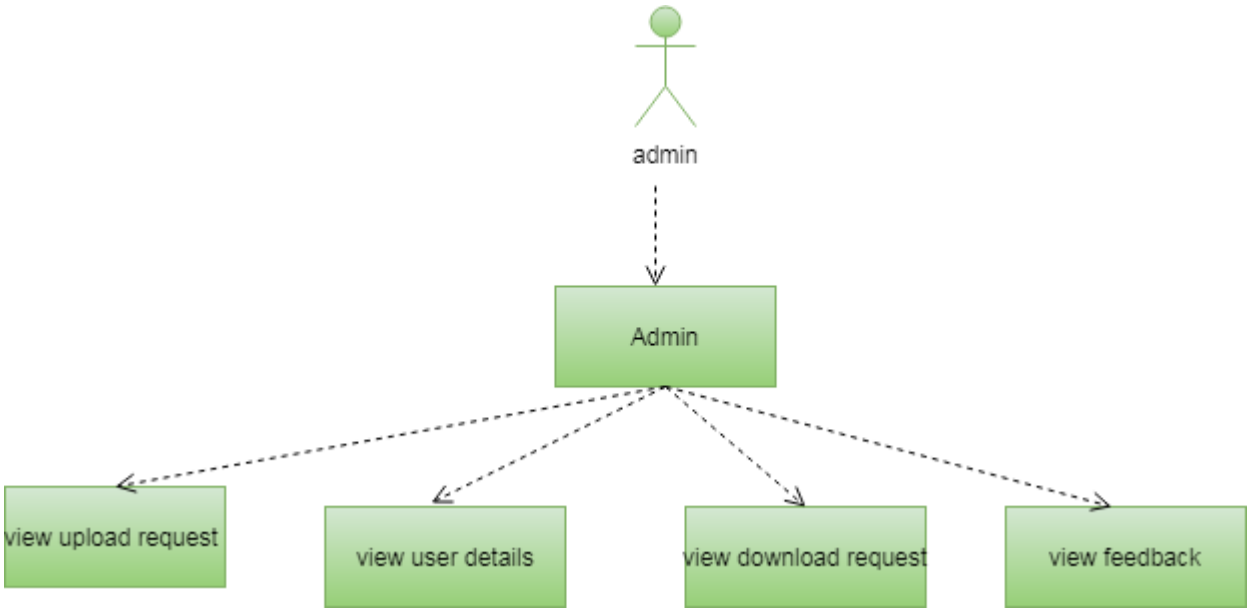


Figure 5.7- Use Case Diagram (Admin)

CLASS DIAGRAM

In software engineering, a class diagram in the Unified Modelling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.

Figure 5.8 shows the class diagram. It consists of two classes; user and admin. The attributes of user are username and user id. The operations of user are upload document (), send feedback (), view my details (), send download request () etc. The attributes of admin are user name and user id. The operations of admin are view upload document file (), view feedback (), view user details (), view user request ().

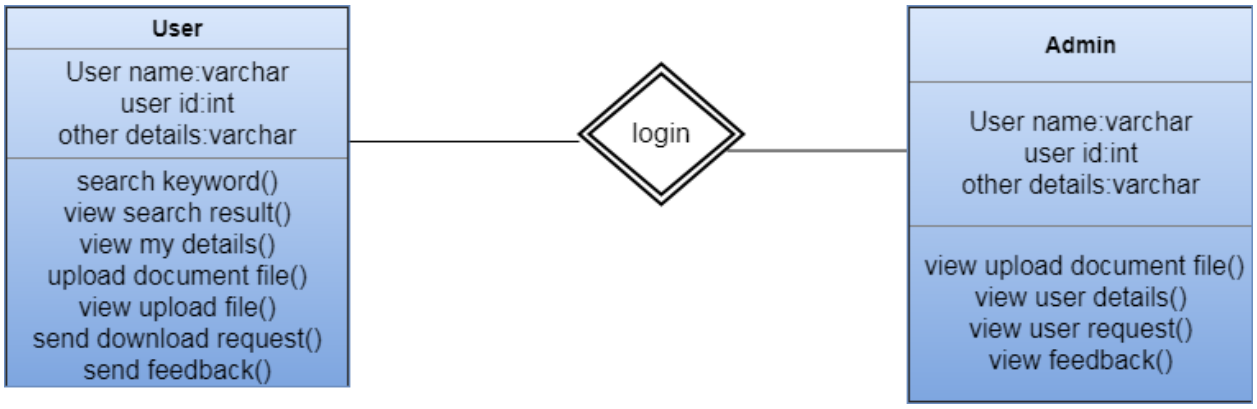


Figure 5.8- Class Diagram

SEQUENCE DIAGRAM

A sequence diagram in Unified Modelling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.

User:

Figure 5.9 shows the sequence diagram for the user. First the user will login and is verified by the username and password entered. If wrong username and password are entered, then the user is termed as unauthorized. If the right username and password are entered then the user can upload documents and view their details. User can also send a download request to the admin for downloading their data. Finally, the user can also send feedback to the admin.

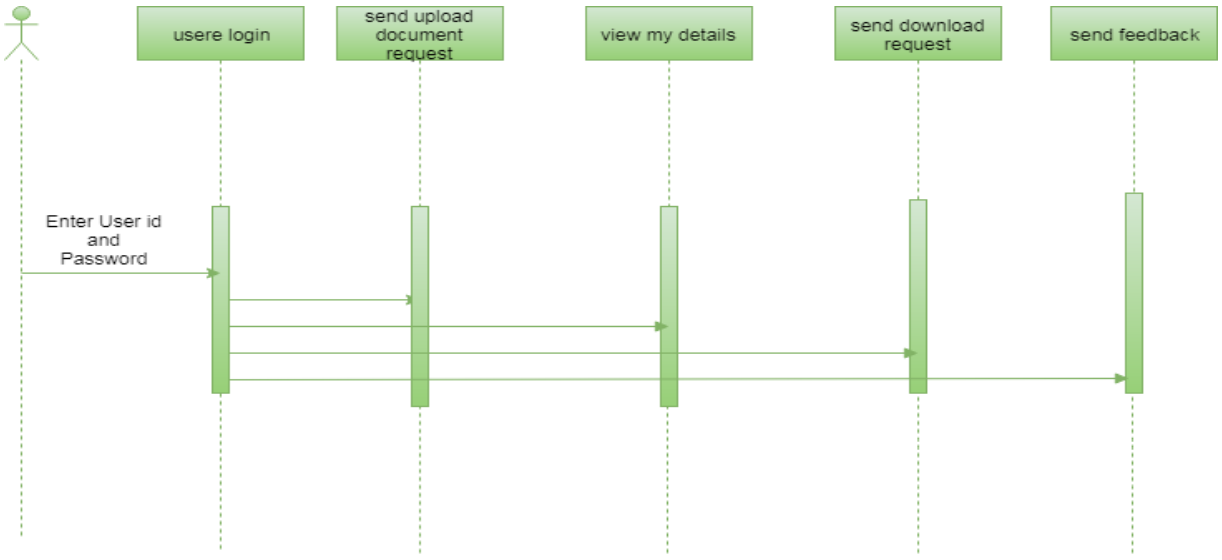


Figure 5.9- Sequence Diagram (User)

Admin:

Figure 5.10 shows the sequence diagram for the admin. The admin will login and is verified by the username and password entered. If wrong username and password are entered, then the admin is termed as unauthorized. If the right username and password are entered then the admin can view the uploaded documents and view the user details. Admin can also view the download request sent by the user for downloading their data. Finally, the admin can also view feedback sent by the user.

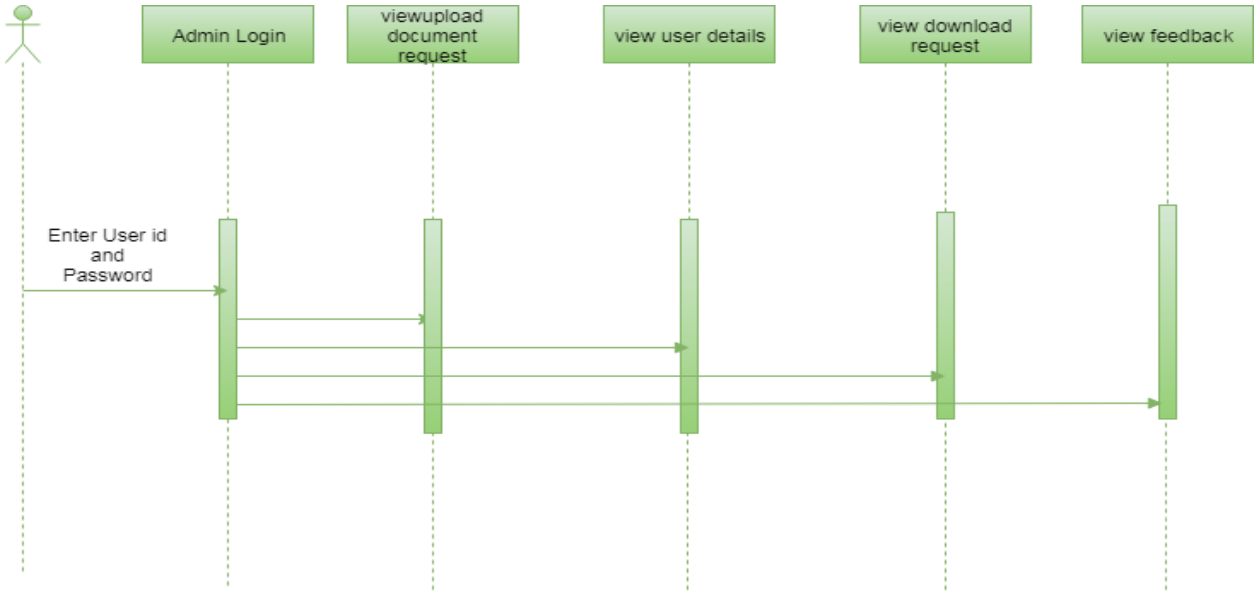


Figure 5.10- Sequence Diagram (Admin)

ACTIVITY DIAGRAM

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modelling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.

User:

Figure 5.11 shows the activity diagram for the user. First the user will login and is verified by the username and password entered. If wrong username and password are entered, then the user is termed as unauthorized. If the right username and password are entered then the user can upload documents and view their details. User can also send a download request to the admin for downloading their data. Finally, the user can also send feedback to the admin.

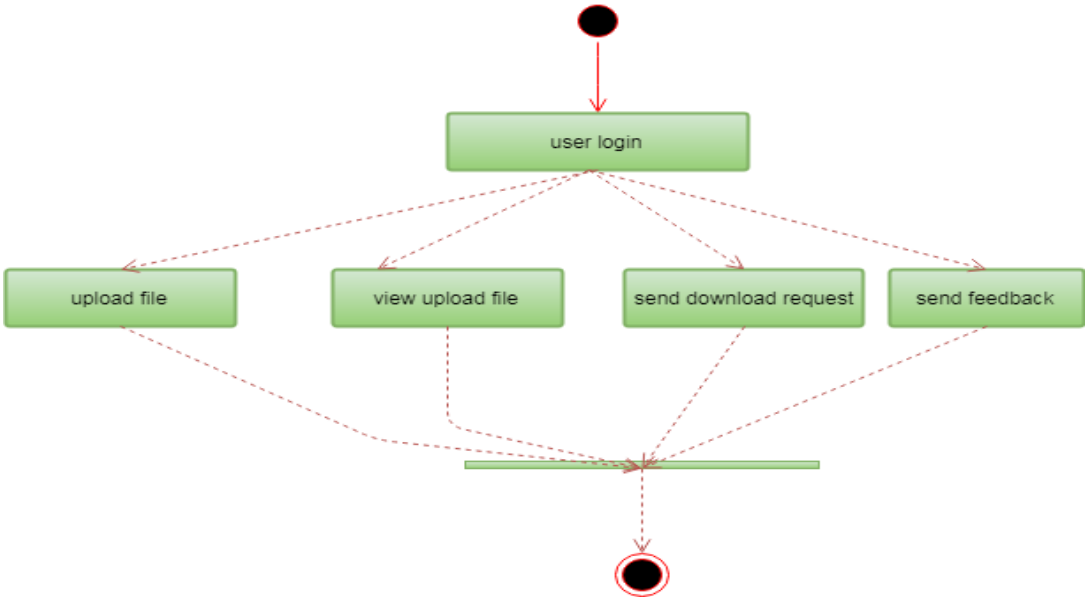


Figure 5.11- Activity Diagram (User)

Admin:

Figure 5.12 shows the activity diagram for the admin. The admin will login and is verified by the username and password entered. If wrong username and password are entered, then the admin is termed as unauthorized. If the right username and password are entered then the admin can view the uploaded documents and view the user details. Admin can also view the download request sent by the user for downloading their data. Finally, the admin can also view feedback sent by the user.

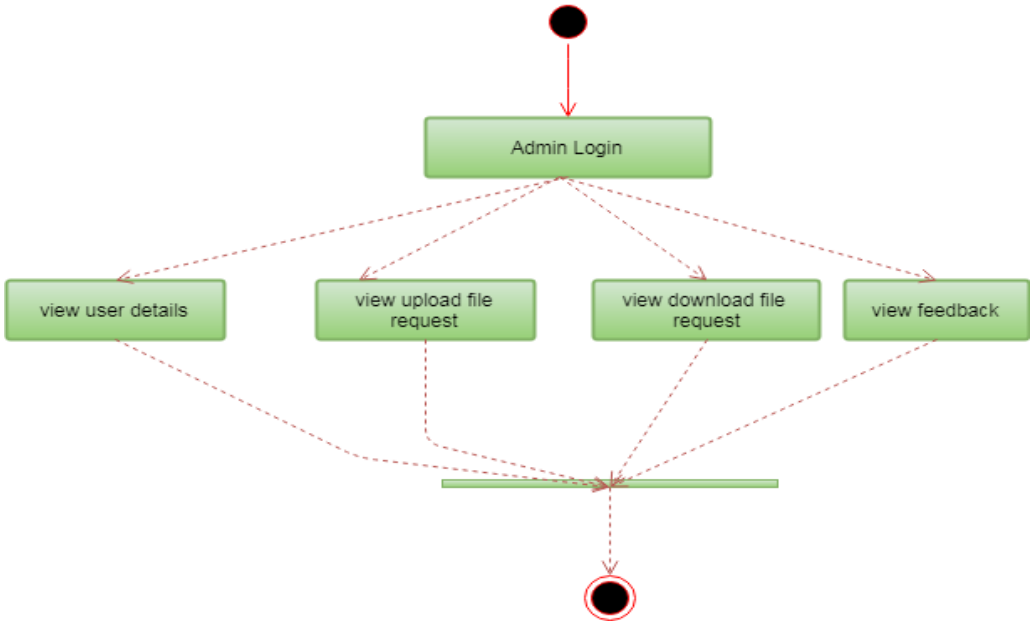


Figure 5.12- Activity Diagram (Admin)

COMPONENT DIAGRAM

In Unified Modelling Language (UML), a component diagram depicts how components are wired together to form larger components or software systems. They are used to illustrate the structure of arbitrarily complex systems.

User:

Figure 5.13 shows the data flow diagram for the user. First the user will login and is verified by the username and password entered. If wrong username and password are entered, then the user is termed as unauthorized. If the right username and password are entered then the user can upload documents and view their details. User can also send a download request to the admin for downloading their data. Finally, the user can also send feedback to the admin.

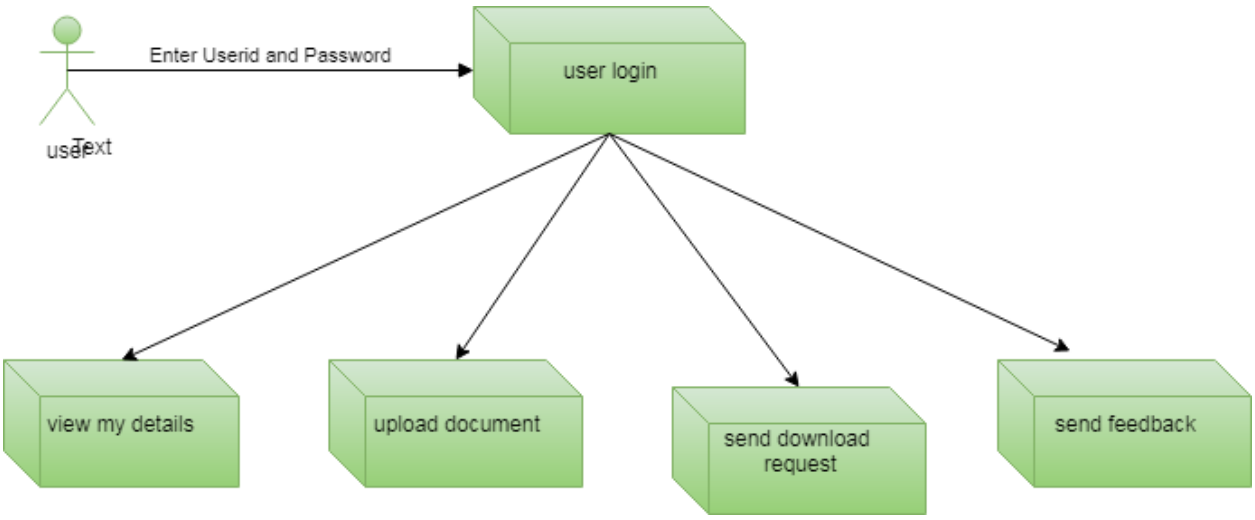


Figure 5.13- Component Diagram (User)

Admin:

Figure 5.14 shows the component diagram for the admin. The admin will login and is verified by the username and password entered. If wrong username and password are entered, then the admin is termed as unauthorized. If the right username and password are entered then the admin can view the uploaded documents and view the user details. Admin can also view the download request sent by the user for downloading their data. Finally, the admin can also view feedback sent by the user.

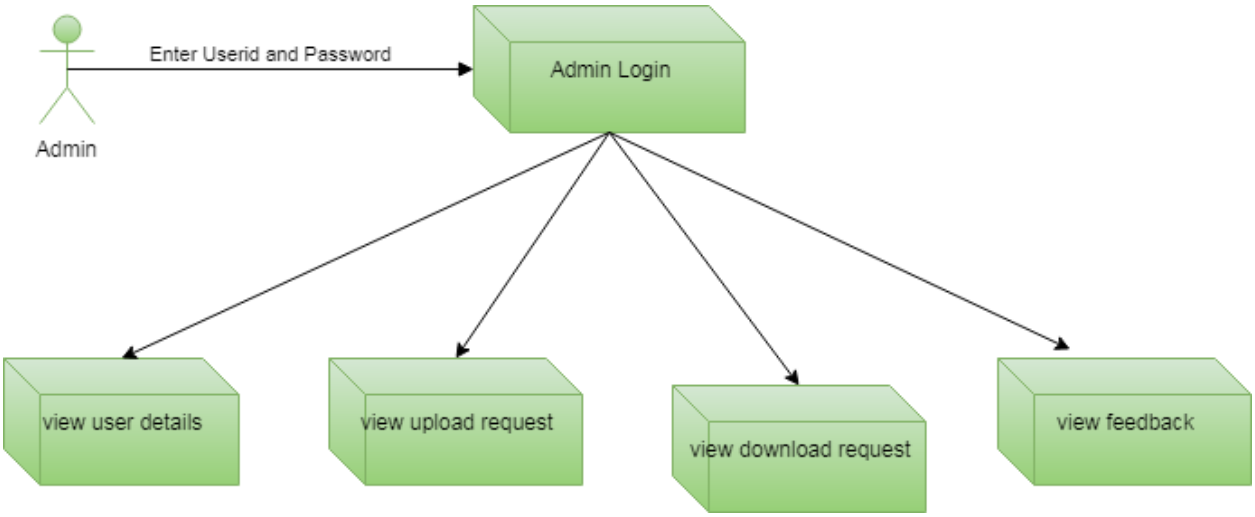


Figure 5.14- Component Diagram (Admin)

6. PROJECT CODING

6.1 SUPPORT VECTOR MACHINE

"Support Vector Machine" (SVM) is a supervised machine learning algorithm which can be used for both classification and regression challenges. However, it is mostly used in classification problems. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiate the two classes very well (look at the below snapshot). Support Vectors are simply the co-ordinates of individual observation. Support Vector Machine is a frontier which best segregates the two classes (hyper-plane/ line). More formally, a support vector machine constructs a hyper plane or set of hyper planes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks like outlier's detection. Intuitively, a good separation is achieved by the hyper plane that has the largest distance to the nearest training-data point of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier. Whereas the original problem may be stated in a finite dimensional space, it often happens that the sets to discriminate are not linearly separable in that space. For this reason, it was proposed that the original finite-dimensional space be mapped into a much higher-dimensional space, presumably making the separation easier in that space

6.2 OUTLINE FOR VARIOUS FILES

What is Python?

Below are some facts about Python.

Python is currently the most widely used multi-purpose, high-level programming language.

Python allows programming in Object-Oriented and Procedural paradigms. Python programs generally are smaller than other programming languages like Java.

Programmers have to type relatively less and indentation requirement of the language, makes them readable all the time.

Python language is being used by almost all tech-giant companies like – Google, Amazon, Facebook, Instagram, Dropbox, Uber… etc.

The biggest strength of Python is huge collection of standard libraries which can be used for the following –

Machine Learning

GUI Applications (like Kivy, Tkinter, PyQt etc.)

Web frameworks like Django (used by YouTube, Instagram, Dropbox)

Image processing (like Opencv, Pillow)

Web scraping (like Scrapy, BeautifulSoup, Selenium)

Test frameworks

Multimedia

Advantages of Python:

Let's see how Python dominates over other languages.

1. Extensive Libraries

Python downloads with an extensive library and it *contain code for various purposes like regular expressions, documentation-generation, unit-testing, web browsers, threading, databases, CGI, email, image manipulation, and more.* So, we don't have to write the complete code for that manually.

2. Extensible

As we have seen earlier, Python can be **extended to other languages**. You can write some of your code in languages like C++ or C. This comes in handy, especially in projects.

3. Embeddable

Complimentary to extensibility, Python is embeddable as well. You can put your Python code in your source code of a different language, like C++. This lets us add **scripting capabilities** to our code in the other language.

4. Improved Productivity

The language's simplicity and extensive libraries render programmers **more productive** than languages like Java and C++ do. Also, the fact that you need to write less and get more things done.

5. IOT Opportunities

Since Python forms the basis of new platforms like Raspberry Pi, it finds the future bright for the Internet of Things. This is a way to connect the language with the real world.

6. Simple and Easy

When working with Java, you may have to create a class to print **'Hello World'**. But in Python, just a print statement will do. It is also quite **easy to learn, understand,** and **code.** This is why when people pick up Python, they have a hard time adjusting to other more verbose languages like Java.

7. Readable

Because it is not such a verbose language, reading Python is much like reading English. This is the reason why it is so easy to learn, understand, and code. It also does not need curly braces to define blocks, and **indentation is mandatory.** This further aids the readability of the code.

8. Object-Oriented

This language supports both the **procedural and object-oriented** programming paradigms. While functions help us with code reusability, classes and objects let us model the real world. A class allows the **encapsulation of data** and functions into one.

9. Free and Open-Source

Like we said earlier, Python is **freely available.** But not only can you **download Python** for free, but you can also download its source code, make changes to it, and even distribute it. It downloads with an extensive collection of libraries to help you with your tasks.

10. Portable

When you code your project in a language like C++, you may need to make some changes to it if you want to run it on another platform. But it isn't the same with Python. Here, you need to **code only once**, and you can run it anywhere. This is called **Write Once Run Anywhere (WORA)**. However, you need to be careful enough not to include any system-dependent features.

11. Interpreted

Lastly, we will say that it is an interpreted language. Since statements are executed one by one, **debugging is easier** than in compiled languages.

Advantages of Python over Other Languages

1. Less Coding

Almost all of the tasks done in Python require less coding when the same task is done in other languages. Python also has an awesome standard library support, so you don't have to search for any third-party libraries to get your job done. This is the reason that many people suggest learning Python to beginners.

2. Affordable

Python is free therefore individuals, small companies or big organizations can leverage the free available resources to build applications. Python is popular and widely used so it gives you better community support.

3. Python is for Everyone

Python code can run on any machine whether it is Linux, Mac or Windows. Programmers need to learn different languages for different jobs but with Python, you can professionally build web apps, perform data analysis and **machine learning**, automate things, do web scraping and also build games and powerful visualizations. It is an all-rounder programming language.

Disadvantages of Python

So far, we've seen why Python is a great choice for your project. But if you choose it, you should be aware of its consequences as well. Let's now see the downsides of choosing Python over another language.

1. Speed Limitations

We have seen that Python code is executed line by line. But since Python is interpreted, it often results in **slow execution**. This, however, isn't a problem unless speed is a focal point for the project. In other words, unless high speed is a requirement, the benefits offered by Python are enough to distract us from its speed limitations.

2. Weak in Mobile Computing and Browsers

While it serves as an excellent server-side language, Python is much rarely seen on the **client-side**. Besides that, it is rarely ever used to implement smartphone-based applications. One such application is called **Carbonnelle**.

The reason it is not so famous despite the existence of Brython is that it isn't that secure.

3. Design Restrictions

As you know, Python **is** dynamically-typed. This means that you don't need to declare the type of variable while writing the code. It uses duck-typing. But wait, what's that? Well, it just means that if it looks like a duck, it must be a duck. While this is easy on the programmers during coding, it can **raise** run-time errors**.**

4. Underdeveloped Database Access Layers

Compared to more widely used technologies like JDBC (Java Database Connectivity) and ODBC (Open Database Connectivity), Python's database access layers are a bit underdeveloped. Consequently, it is less often applied in huge enterprises.

5. Simple

No, we're not kidding. Python's simplicity can indeed be a problem. Take my example. I don't do Java, I'm more of a Python person. To me, its syntax is so simple that the verbosity of Java code seems unnecessary.

This was all about the Advantages and Disadvantages of Python Programming Language.

History of Python:

What do the alphabet and the programming language Python have in common? Right, both start with ABC. If we are talking about ABC in the Python context, it's clear that the programming language ABC is meant. ABC is a general-purpose programming language and programming environment, which had been developed in the Netherlands, Amsterdam, at the CWI (Centrum Wiskunde &Informatica). The greatest achievement of ABC was to influence the design of Python.Python was conceptualized in the late 1980s. Guido van Rossum worked that time in a project at the CWI, called Amoeba, a distributed operating system. In an interview with Bill Venners[1], Guido van Rossum said: "In the early 1980s, I worked as an implementer on a team building a language called ABC at Centrum voor Wiskunde en Informatica (CWI). I don't know how well people know ABC's influence on Python. I try to mention ABC's influence because I'm indebted to everything I learned during that project and to the people who worked on it."Later on in the same Interview, Guido van Rossum continued: "I remembered all my experience and some of my frustration with ABC. I decided to try to design a simple scripting language that possessed some of ABC's better properties, but without its problems. So I started typing. I created a simple virtual machine, a simple parser, and a simple runtime. I made my own version of the various ABC parts that I liked. I created a basic syntax, used indentation for statement grouping instead of curly braces or begin-end blocks, and developed a small number of powerful data types: a hash table (or dictionary, as we call it), a list, strings, and numbers."

What is Machine Learning:

Before we take a look at the details of various machine learning methods, let's start by looking at what machine learning is, and what it isn't. Machine learning is often categorized as a subfield of artificial

intelligence, but I find that categorization can often be misleading at first brush. The study of machine learning certainly arose from research in this context, but in the data science application of machine learning methods, it's more helpful to think of machine learning as a means of *building models of data*.

Fundamentally, machine learning involves building mathematical models to help understand data. "Learning" enters the fray when we give these models *tunable parameters* that can be adapted to observed data; in this way the program can be considered to be "learning" from the data. Once these models have been fit to previously seen data, they can be used to predict and understand aspects of newly observed data. I'll leave to the reader the more philosophical digression regarding the extent to which this type of mathematical, model-based "learning" is similar to the "learning" exhibited by the human brain. Understanding the problem setting in machine learning is essential to using these tools effectively, and so we will start with some broad categorizations of the types of approaches we'll discuss here.

Categories of Machine Learning:

At the most fundamental level, machine learning can be categorized into two main types: supervised learning and unsupervised learning.

Supervised learning involves somehow modeling the relationship between measured features of data and some label associated with the data; once this model is determined, it can be used to apply labels to new, unknown data. This is further subdivided into *classification* tasks and *regression* tasks: in classification, the labels are discrete categories, while in regression, the labels are continuous quantities. We will see examples of both types of supervised learning in the following section.

Unsupervised learning involves modeling the features of a dataset without reference to any label, and is often described as "letting the dataset speak for itself." These models include tasks such as *clustering* and *dimensionality reduction.* Clustering algorithms identify distinct groups of data, while dimensionality reduction algorithms search for more succinct representations of the data. We will see examples of both types of unsupervised learning in the following section.

Need for Machine Learning

Human beings, at this moment, are the most intelligent and advanced species on earth because they can think, evaluate and solve complex problems. On the other side, AI is still in its initial stage and hasn’t surpassed human intelligence in many aspects. Then the question is that what is the need to make machine learn? The most suitable reason for doing this is, “to make decisions, based on data, with efficiency and scale”.

Lately, organizations are investing heavily in newer technologies like Artificial Intelligence, Machine Learning and Deep Learning to get the key information from data to perform several real-world tasks and

solve problems. We can call it data-driven decisions taken by machines, particularly to automate the process. These data-driven decisions can be used, instead of using programing logic, in the problems that cannot be programmed inherently. The fact is that we can't do without human intelligence, but other aspect is that we all need to solve real-world problems with efficiency at a huge scale. That is why the need for machine learning arises.

Challenges in Machines Learning:

While Machine Learning is rapidly evolving, making significant strides with cybersecurity and autonomous cars, this segment of AI as whole still has a long way to go. The reason behind is that ML has not been able to overcome number of challenges. The challenges that ML is facing currently are −

Quality of data − Having good-quality data for ML algorithms is one of the biggest challenges. Use of low-quality data leads to the problems related to data preprocessing and feature extraction.

Time-Consuming task − Another challenge faced by ML models is the consumption of time especially for data acquisition, feature extraction and retrieval.

Lack of specialist persons − As ML technology is still in its infancy stage, availability of expert resources is a tough job.

No clear objective for formulating business problems − Having no clear objective and well-defined goal for business problems is another key challenge for ML because this technology is not that mature yet.

Issue of overfitting & underfitting − If the model is overfitting or underfitting, it cannot be represented well for the problem.

Curse of dimensionality − Another challenge ML model faces is too many features of data points. This can be a real hindrance.

Difficulty in deployment − Complexity of the ML model makes it quite difficult to be deployed in real life.

Applications of Machines Learning:

Machine Learning is the most rapidly growing technology and according to researchers we are in the golden year of AI and ML. It is used to solve many real-world complex problems which cannot be solved with traditional approach. Following are some real-world applications of ML −

Emotion analysis

Sentiment analysis

Error detection and prevention

Weather forecasting and prediction

Stock market analysis and forecasting

Speech synthesis

Speech recognition

Customer segmentation

Object recognition

Fraud detection

Fraud prevention

Recommendation of products to customer in online shopping

How to Start Learning Machine Learning?

Arthur Samuel coined the term “Machine Learning” in 1959 and defined it as a “Field of study that gives computers the capability to learn without being explicitly programmed”.

And that was the beginning of Machine Learning! In modern times, Machine Learning is one of the most popular (if not the most!) career choices. According to Indeed, Machine Learning Engineer Is the Best Job of 2019 with a *344%* growth and an average base salary of $146,085 per year.

In case you are a genius, you could start ML directly but normally, there are some prerequisites that you need to know which include Linear Algebra, Multivariate Calculus, Statistics, and Python. And if you don’t know these, never fear! You don’t need a Ph.D. degree in these topics to get started but you do need a basic understanding.

(a) Learn Linear Algebra and Multivariate Calculus

Both Linear Algebra and Multivariate Calculus are important in Machine Learning. However, the extent to which you need them depends on your role as a data scientist. If you are more focused on application heavy machine learning, then you will not be that heavily focused on maths as there are many common libraries available. But if you want to focus on R&D in Machine Learning, then mastery of Linear Algebra and Multivariate Calculus is very important as you will have to implement many ML algorithms from scratch.

(b) Learn Statistics

Data plays a huge role in Machine Learning. In fact, around 80% of your time as an ML expert will be spent collecting and cleaning data. And statistics is a field that handles the collection, analysis, and presentation of data. So it is no surprise that you need to learn it!!!
Some of the key concepts in statistics that are important are Statistical Significance, Probability Distributions, Hypothesis Testing, Regression, etc. Also, Bayesian Thinking is also a very important part of ML which deals with various concepts like Conditional Probability, Priors, and Posteriors, Maximum Likelihood, etc.

(c) Learn Python

Some people prefer to skip Linear Algebra, Multivariate Calculus and Statistics and learn them as they go along with trial and error. But the one thing that you absolutely cannot skip is Python! While there are other languages you can use for Machine Learning like R, Scala, etc. Python is currently the most popular language for ML. In fact, there are many Python libraries that are specifically useful for Artificial Intelligence and Machine Learning such as Keras, TensorFlow, Scikit-learn, etc.

So, if you want to learn ML, it's best if you learn Python! You can do that using various online resources and courses such as Fork Python available Free on GeeksforGeeks.

Step 2 – Learn Various ML Concepts

Now that you are done with the prerequisites, you can move on to actually learning ML (Which is the fun part!!!) It's best to start with the basics and then move on to the more complicated stuff. Some of the basic concepts in ML are:

(a) Terminologies of Machine Learning

Model – A model is a specific representation learned from data by applying some machine learning algorithm. A model is also called a hypothesis.

Feature – A feature is an individual measurable property of the data. A set of numeric features can be conveniently described by a feature vector. Feature vectors are fed as input to the model. For example, in order to predict a fruit, there may be features like colour, smell, taste, etc.

Target (Label) – A target variable or label is the value to be predicted by our model. For the fruit example discussed in the feature section, the label with each set of input would be the name of the fruit like apple, orange, banana, etc.

Training – The idea is to give a set of inputs(features) and it's expected outputs(labels), so after training, we will have a model (hypothesis) that will then map new data to one of the categories trained on.

Prediction – Once our model is ready, it can be fed a set of inputs to which it will provide a predicted output (label).

(b) Types of Machine Learning

Supervised Learning – This involves learning from a training dataset with labelled data using classification and regression models. This learning process continues until the required level of performance is achieved.

Unsupervised Learning – This involves using unlabelled data and then finding the underlying structure in the data in order to learn more and more about the data itself using factor and cluster analysis models.

Semi-supervised Learning – This involves using unlabelled data like Unsupervised Learning with a small amount of labelled data. Using labelled data vastly increases the learning accuracy and is also more cost-effective than Supervised Learning.

Reinforcement Learning – This involves learning optimal actions through trial and error. So the next action is decided by learning behaviours that are based on the current state and that will maximize the reward in the future.

Advantages of Machine learning:

1. Easily identifies trends and patterns -

Machine Learning can review large volumes of data and discover specific trends and patterns that would not be apparent to humans. For instance, for an e-commerce website like Amazon, it serves to understand the browsing behaviors and purchase histories of its users to help cater to the right products, deals, and reminders relevant to them. It uses the results to reveal relevant advertisements to them.

2. No human intervention needed (automation)

With ML, you don't need to babysit your project every step of the way. Since it means giving machines the ability to learn, it lets them make predictions and also improve the algorithms on their own. A common example of this is anti-virus software's; they learn to filter new threats as they are recognized. ML is also good at recognizing spam.

3. Continuous Improvement

As **ML algorithms** gain experience, they keep improving in accuracy and efficiency. This lets them make better decisions. Say you need to make a weather forecast model. As the amount of data you have keeps growing, your algorithms learn to make more accurate predictions faster.

4. Handling multi-dimensional and multi-variety data

Machine Learning algorithms are good at handling data that are multi-dimensional and multi-variety, and they can do this in dynamic or uncertain environments.

5. Wide Applications

You could be an e-tailer or a healthcare provider and make ML work for you. Where it does apply, it holds the capability to help deliver a much more personal experience to customers while also targeting the right customers.

Disadvantages of Machine Learning:

1. Data Acquisition

Machine Learning requires massive data sets to train on, and these should be inclusive/unbiased, and of good quality. There can also be times where they must wait for new data to be generated.

2. Time and Resources

ML needs enough time to let the algorithms learn and develop enough to fulfill their purpose with a considerable amount of accuracy and relevancy. It also needs massive resources to function. This can mean additional requirements of computer power for you.

3. Interpretation of Results

Another major challenge is the ability to accurately interpret results generated by the algorithms. You must also carefully choose the algorithms for your purpose.

4. High error-susceptibility

Machine Learning is autonomous but highly susceptible to errors. Suppose you train an algorithm with data sets small enough to not be inclusive. You end up with biased predictions coming from a biased training set. This leads to irrelevant advertisements being displayed to customers. In the case of ML, such blunders can set off a chain of errors that can go undetected for long periods of time. And when they do get noticed, it takes quite some time to recognize the source of the issue, and even longer to correct it.

Python Development Steps :

Guido Van Rossum published the first version of Python code (version 0.9.0) at alt.sources in February 1991. This release included already exception handling, functions, and the core data types of list, dict, str and others. It was also object oriented and had a module system.
Python version 1.0 was released in January 1994. The major new features included in this release were the functional programming tools lambda, map, filter and reduce, which Guido Van Rossum never liked. Six and a half years later in October 2000, Python 2.0 was introduced. This release included list comprehensions, a full garbage collector and it was supporting unicode. Python flourished for another 8 years in the versions 2.x before the next major release as Python 3.0 (also known as "Python 3000" and "Py3K") was released. Python 3 is not backwards compatible with Python 2.x. The emphasis in Python 3 had been on the removal of duplicate programming constructs and modules, thus fulfilling or coming close to fulfilling the 13th law of the Zen of Python: "There should be one -- and preferably only one -- obvious way to do it."Some changes in Python 7.3:

Print is now a function

Views and iterators instead of lists

The rules for ordering comparisons have been simplified. E.g. a heterogeneous list cannot be sorted, because all the elements of a list must be comparable to each other.

There is only one integer type left, i.e. int. long is int as well.

The division of two integers returns a float instead of an integer. "//" can be used to have the "old" behaviour.

Text Vs. Data Instead of Unicode Vs. 8-bit

Purpose:

We demonstrated that our approach enables successful segmentation of intra-retinal layers—even with low-quality images containing speckle noise, low contrast, and different intensity ranges throughout—with the assistance of the ANIS feature.

Python

Python is an interpreted high-level programming language for general-purpose programming. Created by Guido van Rossum and first released in 1991, Python has a design philosophy that emphasizes code readability, notably using significant whitespace.

Python features a dynamic type system and automatic memory management. It supports multiple programming paradigms, including object-oriented, imperative, functional and procedural, and has a large and comprehensive standard library.

Python is Interpreted − Python is processed at runtime by the interpreter. You do not need to compile your program before executing it. This is similar to PERL and PHP.

Python is Interactive − you can actually sit at a Python prompt and interact with the interpreter directly to write your programs.

Python also acknowledges that speed of development is important. Readable and terse code is part of this, and so is access to powerful constructs that avoid tedious repetition of code. Maintainability also ties into this may be an all but useless metric, but it does say something about how much code you have to scan, read and/or understand to troubleshoot problems or tweak behaviors. This speed of development, the ease with which a programmer of other languages can pick up basic Python skills and the huge standard library is key to another area where Python excels. All its tools have been quick to implement, saved a lot of time, and several of them have later been patched and updated by people with no Python background - without breaking.

Python is Interpreted − Python is processed at runtime by the interpreter. You do not need to compile your program before executing it. This is similar to PERL and PHP.

Python is Interactive − you can actually sit at a Python prompt and interact with the interpreter directly to write your programs.

Python also acknowledges that speed of development is important. Readable and terse code is part of this, and so is access to powerful constructs that avoid tedious repetition of code. Maintainability also ties into this may be an all but useless metric, but it does say something about how much code you have to scan, read and/or understand to troubleshoot problems or tweak behaviors. This speed of development, the ease with which a programmer of other languages can pick up basic Python skills and the huge standard library is key to another area where Python excels. All its tools have been quick to implement, saved a lot of time, and several of them have later been patched and updated by people with no Python background - without breaking.

Install Python Step-by-Step in Windows and Mac:

Python a versatile programming language doesn't come pre-installed on your computer devices. Python was first released in the year 1991 and until today it is a very popular high-level programming language. Its style philosophy emphasizes code readability with its notable use of great whitespace.

The object-oriented approach and language construct provided by Python enables programmers to write both clear and logical code for projects. This software does not come pre-packaged with Windows.

How to Install Python on Windows and Mac:

There have been several updates in the Python version over the years. The question is how to install Python? It might be confusing for the beginner who is willing to start learning Python but this tutorial will solve your query. The latest or the newest version of Python is version 3.7.4 or in other words, it is Python 3.

Note: The python version 3.7.4 cannot be used on Windows XP or earlier devices.

Before you start with the installation process of Python. First, you need to know about your System Requirements. Based on your system type i.e., operating system and based processor, you must download the python version. My system type is a Windows 64-bit operating system. So, the steps below are to install python version 3.7.4 on Windows 7 device or to install Python 3. Download the Python Cheatsheet here. The steps on how to install Python on Windows 10, 8 and 7 are divided into 4 parts to help understand better.

Download the Correct version into the system

Step 1: Go to the official site to download and install python using Google Chrome or any other web browser. OR Click on the following link: https://www.python.org

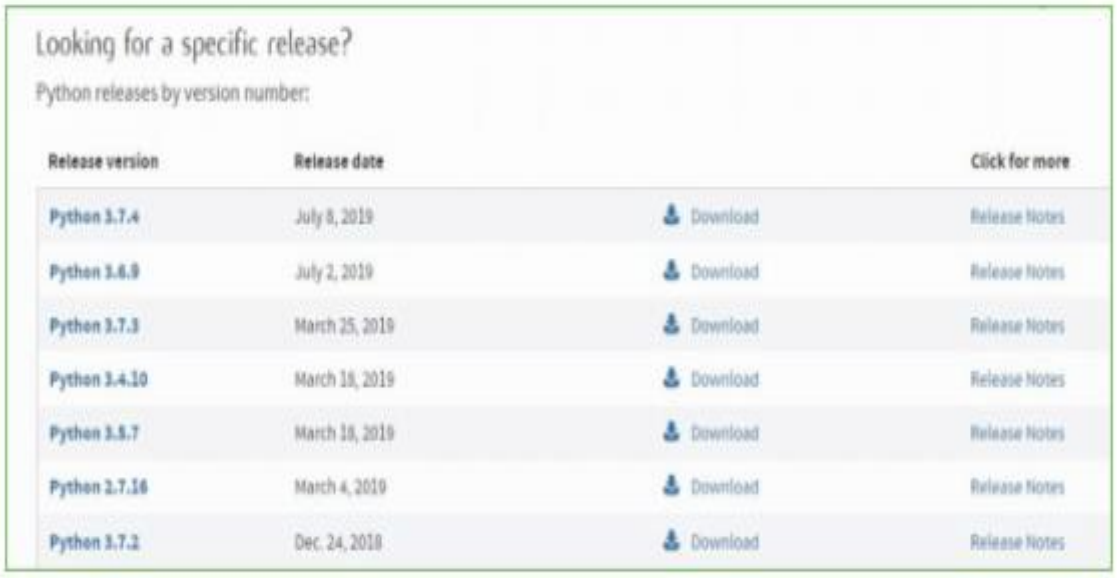
Fig 5.7 : Google Chrome or any other web browser.

Now, check for the latest and the correct version for your operating system.

Step 2: Click on the Download Tab.

Fig 5.8: Python original page

Step 3: You can either select the Download Python for windows 3.7.4 button in Yellow Color or you can scroll further down and click on download with respective to their version. Here, we are downloading the most recent python version for windows 3.7.4



Looking for a specific release?

Python releases by version number:

Release version	Release date		Click for more
Python 3.7.4	July 8, 2019	Download	Release Notes
Python 3.6.9	July 2, 2019	Download	Release Notes
Python 3.7.3	March 25, 2019	Download	Release Notes
Python 3.4.10	March 18, 2019	Download	Release Notes
Python 3.5.7	March 18, 2019	Download	Release Notes
Python 2.7.16	March 4, 2019	Download	Release Notes
Python 3.7.2	Dec. 24, 2018	Download	Release Notes

Fig 5.9: Selecting Python version

Step 4: Scroll down the page until you find the Files option.

Step 5: Here you see a different version of python along with the operating system.

Files

Version	Operating System	Description	MD5 Sum	File Size	GPG
Gzipped source tarball	Source release		68111671e5b2db4aef7b9ab01bf0f9be	23017663	SIG
XZ compressed source tarball	Source release		d33e4aae66097051c2eca45ee3604803	17131432	SIG
macOS 64-bit/32-bit installer	Mac OS X	for Mac OS X 10.6 and later	6428b4fa7583daff1a442cbalcee08e6	34898416	SIG
macOS 64-bit installer	Mac OS X	for OS X 10.9 and later	5dd605c38217a45773bf5e4a936b241f	28082845	SIG
Windows help file	Windows		d639995732a2c56b2ac56cade6b4f7cd2	8131761	SIG
Windows x86-64 embeddable zip file	Windows	for AMD64/EM64T/x64	9b00c8cf6d9ec0b9a8e83184a40729a2	7504391	SIG
Windows x86-64 executable installer	Windows	for AMD64/EM64T/x64	a702b4b0ad76deb0db3043a583e563400	26680368	SIG
Windows x86-64 web-based installer	Windows	for AMD64/EM64T/x64	28cb1c608b6d73ae8e53a3bd351b4bd2	1362904	SIG
Windows x86 embeddable zip file	Windows		9fab3b61f8841879fda941335741330d8	6741628	SIG
Windows x86 executable installer	Windows		33cc602942a54446a3d8451476394789	25663848	SIG
Windows x86 web-based installer	Windows		1b670cfa5d317df82c30983ea371d87c	1324608	SIG

Fig 5.10 Different types of python versions along with operating system

• To download Windows 32-bit python, you can select any one from the three options: Windows x86 embeddable zip file, Windows x86 executable installer or Windows x86 web-based installer.

•To download Windows 64-bit python, you can select any one from the three options: Windows x86-64 embeddable zip file, Windows x86-64 executable installer or Windows x86-64 web-based installer.

Here we will install Windows x86-64 web-based installer. Here your first part regarding which version of python is to be downloaded is completed. Now we move ahead with the second part in installing python i.e. Installation

Note: To know the changes or updates that are made in the version you can click on the Release Note Option.

Installation of Python

Step 1: Go to Download and Open the downloaded python version to carry out the installation process.

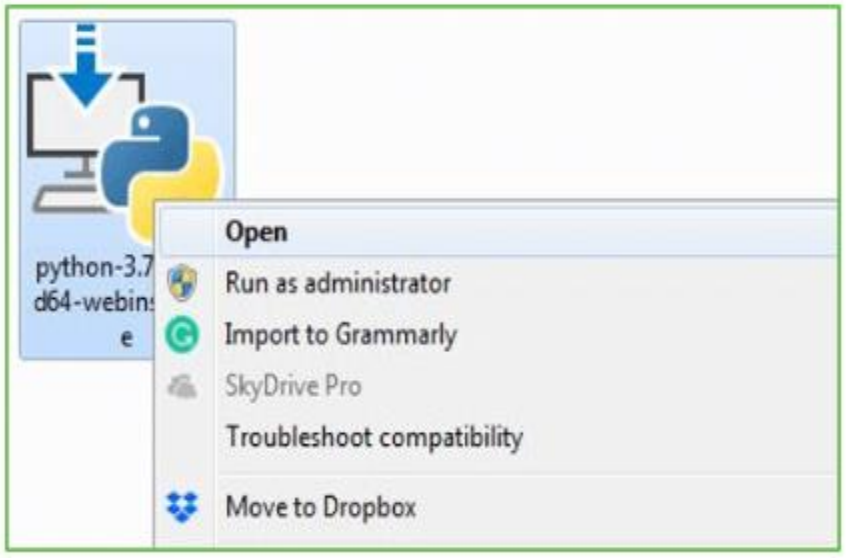


Fig 5.11: Python file

Step 2: Before you click on Install Now, make sure to put a tick on Add Python 3.7 to PATH.

Fig 5.12:Path setting

Step 3: Click on Install NOW After the installation is successful. Click on Close.

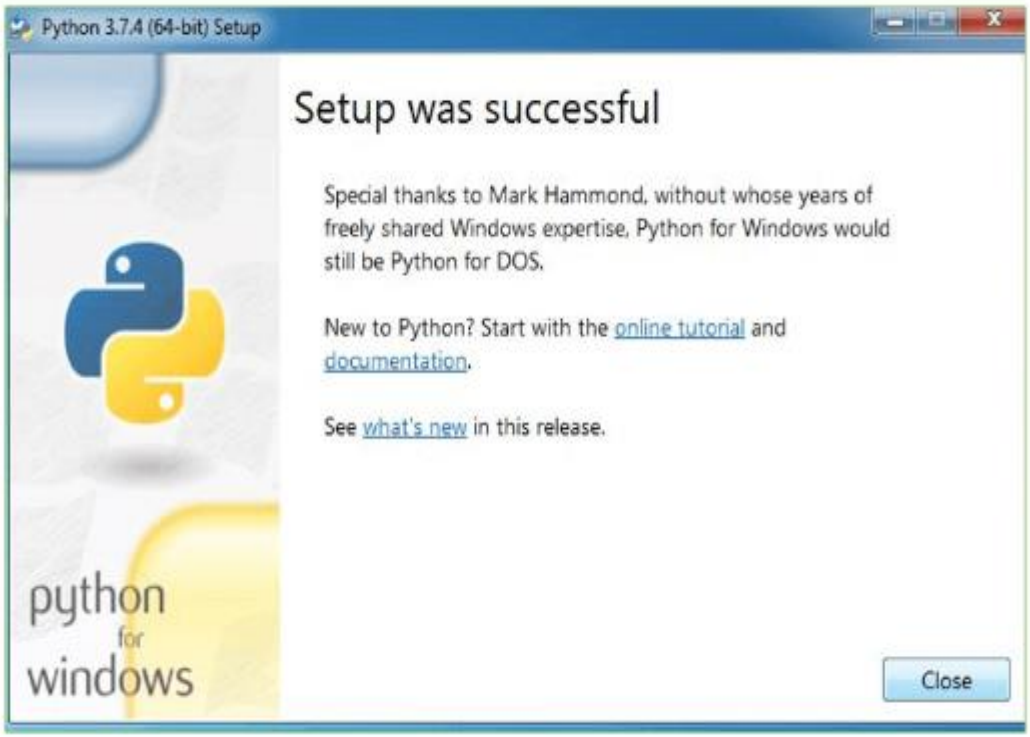


Fig 5.13:Setup completed

With these above three steps on python installation, you have successfully and correctly installed Python. Now is the time to verify the installation.

Note: The installation process might take a couple of minutes.

Verify the Python Installation

Step 1: Click on Start

Step 2: In the Windows Run Command, type “cmd”.

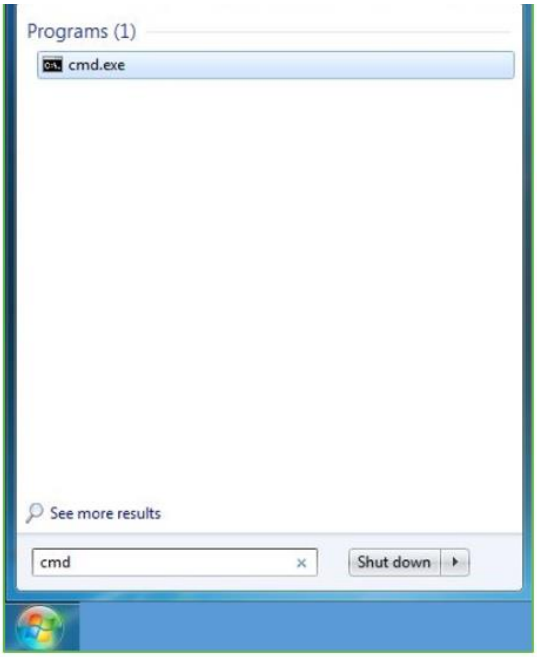


Fig 5.14 Opening command prompt to check python installation

Step 3: Open the Command prompt option.

Step 4: Let us test whether the python is correctly installed. Type **python –V** and press Enter.

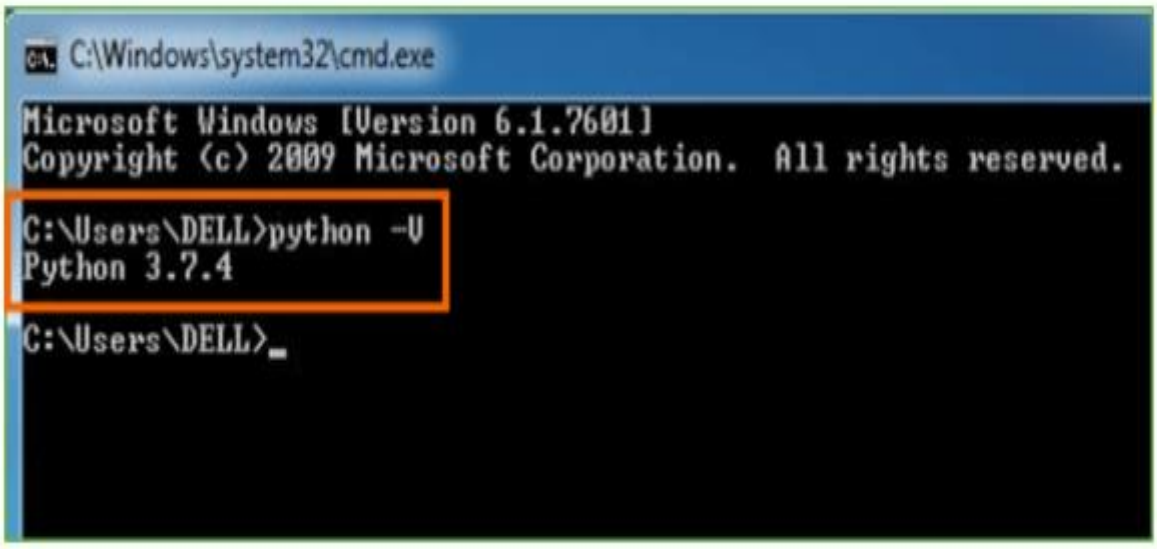


Fig 5.15:Checking python version in command prompt

Step 5: You will get the answer as 3.7.4

Note: If you have any of the earlier versions of Python already installed. You must first uninstall the earlier version and then install the new one.

Check how the Python IDLE works

Step 1: Click on Start

Step 2: In the Windows Run command, type "python idle".

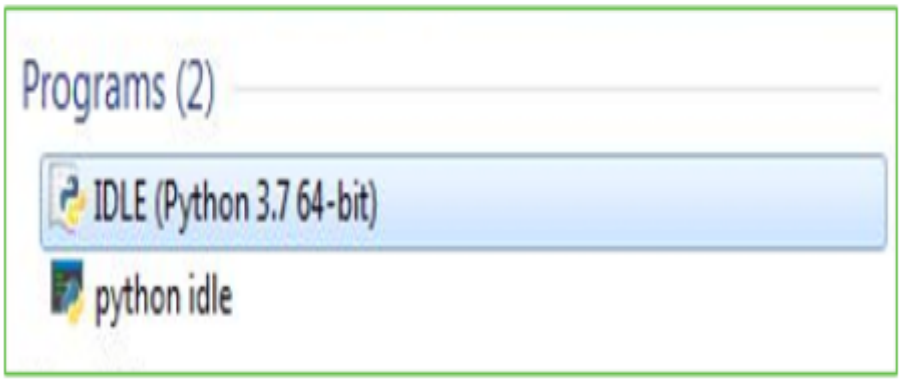


Fig 5.16:Opening python idle

Step 3: Click on IDLE (Python 3.7 64-bit) and launch the program

Step 4: To go ahead with working in IDLE you must first save the file. **Click on File > Click on Save**

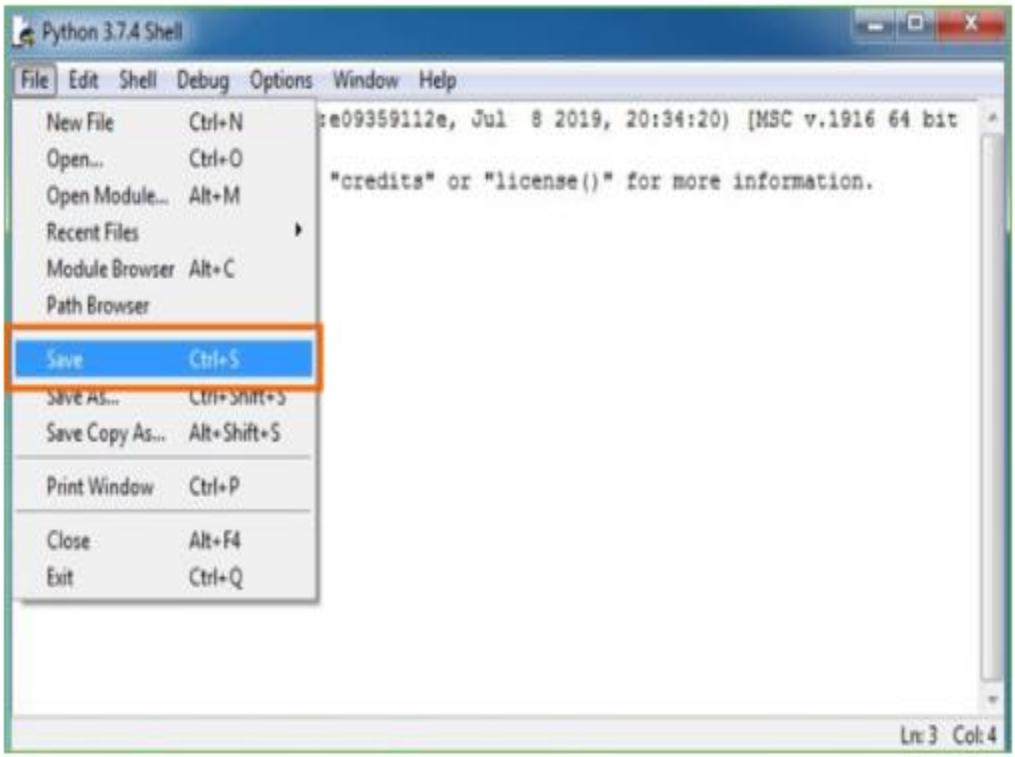


Fig 5.17:Saving the file

Step 5: Name the file and save as type should be Python files. Click on SAVE. Here I have named the files as Hey World.

Step 6: Now for e.g. **enter print**

6.3 METHODS OF INPUT AND OUTPUT PARAMETERS

Input Design

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the following things:

What data should be given as input?

How the data should be arranged or coded?

The dialog to guide the operating personnel in providing input.

Methods for preparing input validations and steps to follow when error occur.

Objectives

1.Input Design is the process of converting a user-oriented description of the input into a computer-based system. This design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerized system.

2. It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The goal of designing input is to make data entry easier and to be free from errors. The data entry screen is designed in such a way that all the data manipulates can be performed. It also provides record viewing facilities.

3.When the data is entered it will check for its validity. Data can be entered with the help of screens. Appropriate messages are provided as when needed so that the user will not be in maize of instant. Thus, the objective of input design is to create an input layout that is easy to follow

Output Design

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making.

Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is designed so that people will find the system can use easily and effectively. When analysis design computer output, they should Identify the specific output that is needed to meet the requirements. Select methods for presenting information. Create document, report, or other formats that contain information produced by the system.

The output form of an information system should accomplish one or more of the following objectives. Convey information about past activities, current status or projections of the future. Signal important events, opportunities, problems, or warnings. Trigger an action. Confirm an action.

7. PROJECT TESTING

7.1 VARIOUS TEST CASES

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub-assemblies, assemblies and/or a finished product. It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of tests. Each test type addresses a specific testing requirement.

TYPES OF TESTS

Unit testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

Integration testing

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

Functional test

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centred on the following items:

Valid Input	: identified classes of valid input must be accepted.
Invalid Input	: identified classes of invalid input must be rejected.
Functions	: identified functions must be exercised.
Output	: identified classes of application outputs must be exercised.
Systems/Procedures	: interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

System Test

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration-oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

Unit Testing

Unit testing is usually conducted as part of a combined code and unit test phase of the software lifecycle, although it is not uncommon for coding and unit testing to be conducted as two distinct phases.

Test strategy and approach

Field testing will be performed manually and functional tests will be written in detail.

Test objectives

- All field entries must work properly.
- Pages must be activated from the identified link.
- The entry screen, messages and responses must not be delayed.

Features to be tested

- Verify that the entries are of the correct format
- No duplicate entries should be allowed
- All links should take the user to the correct page.

Integration Testing

Software integration testing is the incremental integration testing of two or more integrated software components on a single platform to produce failures caused by interface defects.

The task of the integration test is to check that components or software applications, e.g. components in a software system or – one step up – software applications at the company level – interact without error.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

Acceptance Testing

User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

7.2 BLACK BOX TESTING

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box .you cannot “see” into it. The test provides inputs and responds to outputs without considering how the software works.

7.3 WHITE BOX TESTING

White Box Testing is a testing in which in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is purpose. It is used to test areas that cannot be reached from a black box level.

OUTPUT SCREENS

8.1 USER INTERFACE

Figure 8.1 shows the web page using which the user can login or create an account. The user can login using the username and password. If the user enters wrong username and password then the user is blocked after certain attempts. Only the admin has the right to unblock a user.

Figure 8.1 Login Page

8.2 OUTPUT SCREENS

Figure 8.2 shows the web page where the user can upload data. User can upload the data regarding the entities and organizational data. Uploaded data are managed by admin and admin is the only person to provide the accessing rights and approve or unapproved users based on their details.

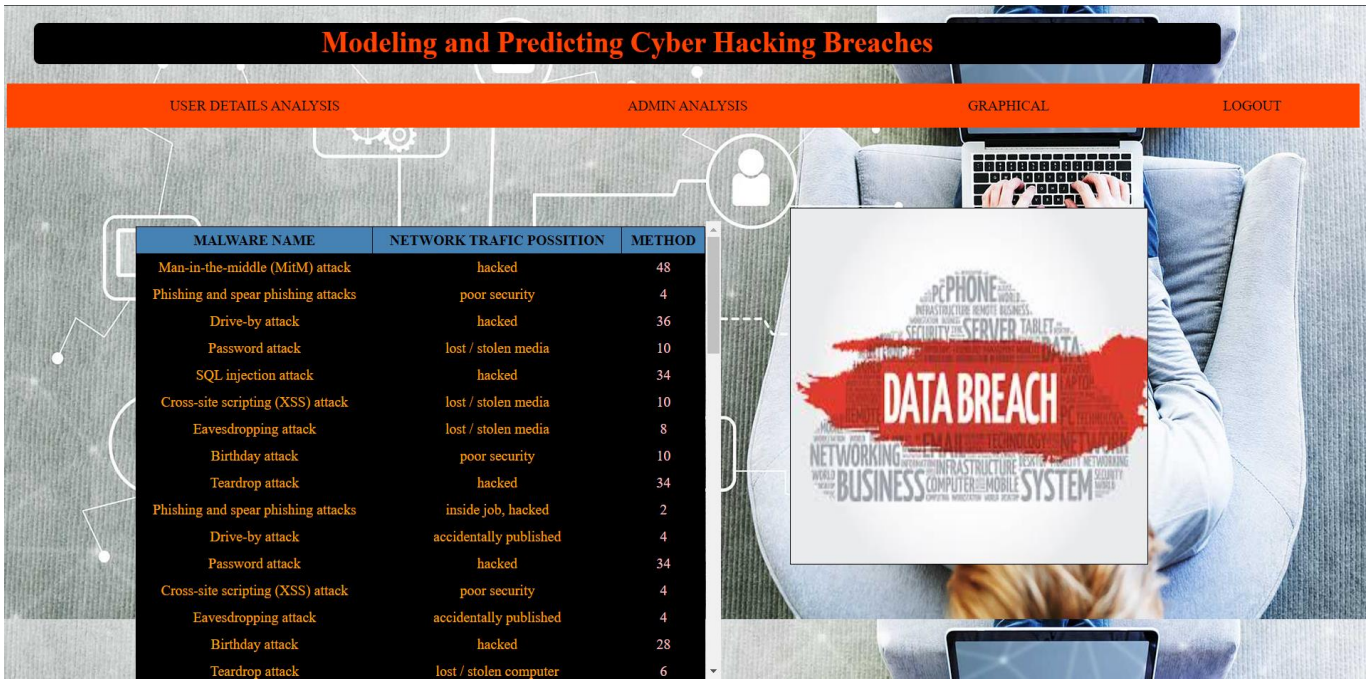
Figure 8.2 Upload Request Page

Figure 8.3 gives an analysis of the different organizations and their types and also mentions the number of records in each organization. It displays whether these organizations were hacked or not and also shows if they have poor security.

ENTITY	YEAR	RECORDS	ORGANIZATION TYPE	METHOD	DATA
21st Century Oncology	2016	2,200,000	healthcare	hacked	https://www.linkedin.com/jobs/view/930124877/?refId=d3493ec8-privateid.37f8-4218-92b2-4656b383a955&trk card&midToken=serverattack/AQHBnYxQHAJchw&trkEmail=eml-jobs_jymbii_digest-null-4-null-null-9xzoen jobs~view&lipi=urn%3Ali%3Apage%3Aemail_jobs_jymbii_digest%3BCxmcCwrxR62ABhqSrl2dYA%3D%3D
Accendo Insurance Co.	2011	175,350	healthcare	poor security	https://www.bayt.com/en/job-seekers/create-account/?url_id=1&utm_medium=associate&utm_source=walkinu/
Adobe Systems	2013	152,000,000	tech	hacked	https://www.google.co.in/search?ei=9pzSW4rJA8zWvgSzuYDwBg&q=brainmagic+infotech+NLOM]pvt+ltd+g ab.1.0.0i71k1l4.0.0.0.709767.0.0.0.0.0.0.0.0..0.0....0...1c..64.psy-ab..0.0.0....0.YV8QKntrcq4
Advocate Medical Group	2013	4,000,000	healthcare	lost / stolen media	https://stackoverflow.com/questions/43727583/expected-string-or-bytes-like-object/ECSID/getmonlist
AerServ (subsidiary of InMobi)	2018	75,000	advertising	hacked	https://www.google.co.in/search?q=python+free+online+course+certification&oq=p&aqs/2F4=chrome.0.69i59j6
Affinity Health Plan, Inc.	2009	344,579	healthcare	lost / stolen media	http://localhost/phpmyadmin/index.php?token=27a2ceb5cb82727c1f0b63f93f1d0c4f#PMAURL-/x25/NLOM;2: db=malware_detection&table=user_malware_recogition_model&server=1&target=&token=27a2ceb5cb82727c1
Ameritrade	2005	200,000	financial	lost / stolen media	https://realpython.com/pypi-publish-python-package/ECSID/ICMPID
Ancestry.com	2015	300,000	web	poor security	https://mail.google.com/mail/u/0/#inbox/ECSID/getmonlist

Figure 8.3 Unmalware Data

Figure 8.4 gives an analysis of the different kinds of attacks that have occurred and also the network traffic position. It shows whether the attack occurred accidentally or was intentional.



MALWARE NAME	NETWORK TRAFIC POSSITION	METHOD
Man-in-the-middle (MitM) attack	hacked	48
Phishing and spear phishing attacks	poor security	4
Drive-by attack	hacked	36
Password attack	lost / stolen media	10
SQL injection attack	hacked	34
Cross-site scripting (XSS) attack	lost / stolen media	10
Eavesdropping attack	lost / stolen media	8
Birthday attack	poor security	10
Teardrop attack	hacked	34
Phishing and spear phishing attacks	inside job, hacked	2
Drive-by attack	accidentally published	4
Password attack	hacked	34
Cross-site scripting (XSS) attack	poor security	4
Eavesdropping attack	accidentally published	4
Birthday attack	hacked	28
Teardrop attack	lost / stolen computer	6

Figure 8.4 Admin Report

Data analyses are done with the help of graph. Figure 8.5 gives the graphical analysis of all the breaches that have occurred with the number of breaches on x-axis and the year of occurrence on the y-axis. The collected data are applied to graph in order to get the best analysis and prediction of dataset and given data policies.

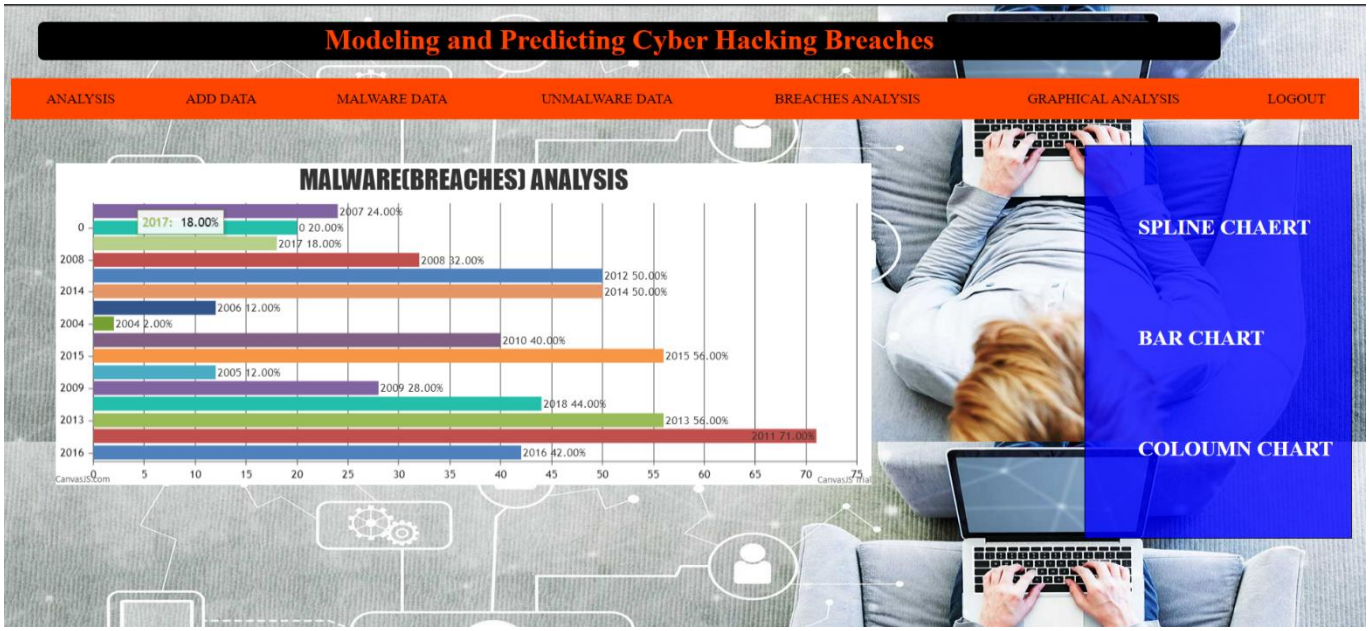


Figure 8.5

In figure 8.6, the year of occurance is mapped on the x-axis and the quantity of malware breaches that have occurred are mapped on the y-axis. The dataset can be analyzed through this pictorial representation in order to better understand of the data details.

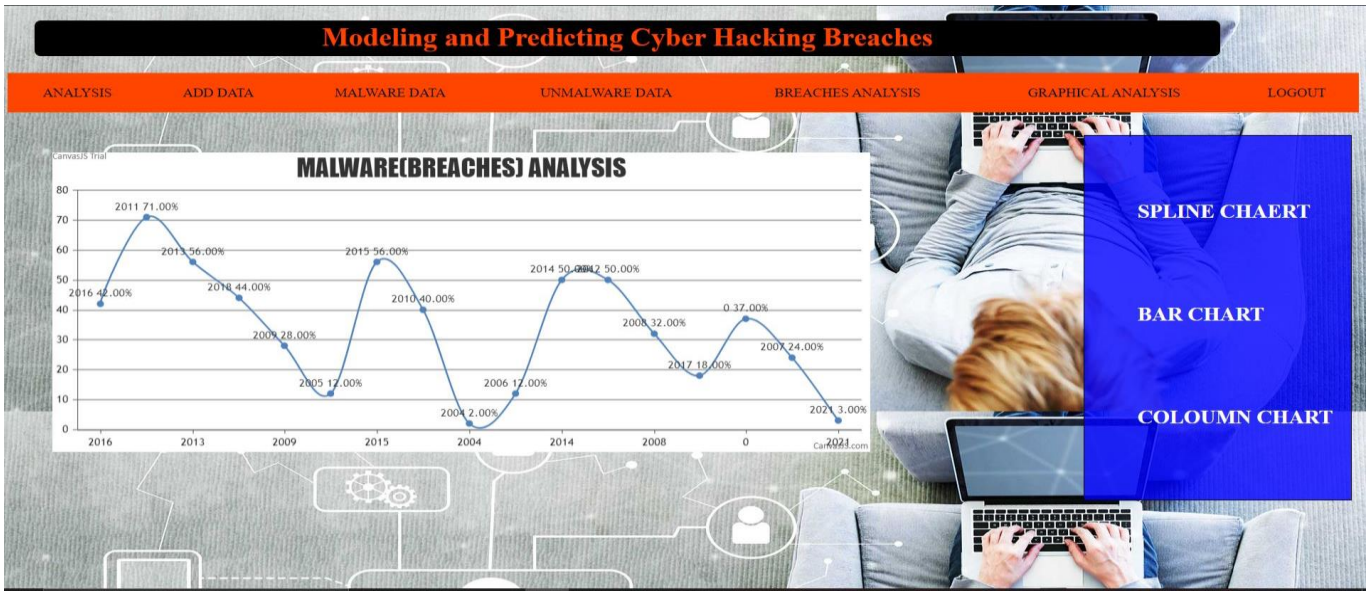


Figure 8.6

CONCLUSION AND FUTURE ENHANCEMENT

We analysed a hacking breach dataset from the points of view of the incidents inter-arrival time and the breach size, and showed that they both should be modelled by stochastic processes rather than distributions. The statistical models developed in this paper show satisfactory fitting and prediction accuracies. In particular, we propose using a copula-based approach to predict the joint probability that an incident with a certain magnitude of breach size will occur during a future period of time. Statistical tests show that the methodologies proposed in this paper are better than those which are presented in the literature, because the latter ignored both the temporal correlations and the dependence between the incidents inter-arrival times and the breach sizes. We conducted qualitative and quantitative analyses to draw further insights. We drew a set of cybersecurity insights, including that the threat of cyber hacking breach incidents is indeed getting worse in terms of their frequency, but not the magnitude of their damage. The methodology presented in this paper can be adopted or adapted to analyse datasets of a similar nature.

REFERENCES

1. R. R. Subramanian, R. Avula, P. S. Surya and B. Pranay, "Modeling and Predicting Cyber Hacking Breaches," 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), 2021, pp. 288-293, doi: 10.1109/ICICCS51141.2021.9432175.
2. T Manikandan, B Balamurugan, C Senthilkumar, RRA Harinarayan and RR Subramanian, "Cyber War is Coming" in Cyber Security in Parallel and Distributed Computing: Concepts Techniques Applications and Case Studies, John Wiley & Sons, Inc, pp. 79-89, Mar. 2019.
3. M. Xu, K. M. Schweitzer, R. M. Bateman and S. Xu, "Modeling and Predicting Cyber Hacking Breaches," in IEEE Transactions on Information Forensics and Security, vol. 13, no. 11, pp. 2856-2871, Nov. 2018, doi: 10.1109/TIFS.2018.2834227.
4. Z. Mohammed, aNITDA Raises Alarm over Potential Cyber Attacks to Banks. Govt Agencies, 2018.
5. C. R. Centre. Cybersecurity Incidents. Accessed: Nov. 2020. [Online]. Available: https://www.opm.gov/cybersecurity/cybersecurity-incidents.
6. IBM Security. Accessed: Nov. 2019. [Online]. Available: https://www.ibm.com/security/data-breach/index.html
7. P. R. Clearinghouse. Privacy Rights Clearinghouse’s Chronology of Data Breaches. Accessed: Nov. 2017. [Online]. Available: https://www.privacyrights.org/data-breaches
8. R. B. Security. Datalossdb. Accessed: Nov. 2018. [Online]. Available: https://blog.datalossdb.org.

PUBLICATIONS

Modeling and Predicting Cyber Hacking Breaches using Stochastic Process Models

[1]Saddi Advaitha Reddy, [2]Nalla Rakshitha, [3]Abhishek Reddy, [4]Katta Pradhyun Reddy,

[5]Dr.M. Narayanan, [6]Dr.P. Santosh Kumar Patra, [7]Dr. G. Jawaharlal Nehru

[1234]UG Scholar, [5]Professor, [6]Principal & Professor in CSE, [7]Assistant Professor

Department of Computer Science and Engineering

St. Martin's Engineering College, Secunderabad – 500 100, India

E-Mail: [1]advaithareddy2201@gmail.com, [2]rakshitha2349@gmail.com, [3]d.abhishekreddy11@gmail.com, [4]kattapradhyun@gmail.com

Abstract

The selection of parameters greatly affects the prediction accuracy of support vector machine. Analyzing cyber incident data sets is an important method for deepening our understanding of the evolution of the threat situation. This is a relatively new research topic, and many studies remain to be done. In this paper, we report a statistical analysis of a breach incident data set corresponding to 12 years (2005–2017) of cyber hacking activities that include malware attacks. We show that, in contrast to the findings reported in the literature, both hacking breach incident inter-arrival times and breach sizes should be modeled by stochastic processes, rather than by distributions because they exhibit autocorrelations. Then, we propose particular stochastic process models to, respectively, fit the inter-arrival times and the breach sizes. We also show that these models can predict the inter-arrival times and the breach sizes. In order to get deeper insights into the evolution of hacking breach incidents, we conduct both qualitative and quantitative trend analyses on the data set. We draw a set of cyber security insights, including that the threat of cyber hacks is indeed getting worse in terms of their frequency, but not in terms of the magnitude of their damage.

Keywords: Hacking breach, data breach, cyber threats, cyber risk analysis, breach prediction, SVM.

1. INTRODUCTION

Data breaching is an act of disseminating the highly sensitive data which are intended to be kept secret. Disclosing the data to the un secured domain either with a motive or unintentionally. It occurs when the third-party or an unauthorized individual tries to steal or access the data which may comprises of top secrets, company shares, transaction details or legal information. There are different types of data breaching which includes phishing, denial of service attack, malware and exhilaration. From time to time

we hear about several companies and industries announcing that their systems have been breached. This might happen by illegal action of the intruders [1]. Or also by an individual within the organization. They might even belong to an organized group of criminals whose main target is money. This is known as cyber attack and those who perform such illegitimate practices are known as cyber criminals. Overcoming this situation is not an easy task. But, several steps of prevention towards data security can be followed. Maintenance of data can be improved by adapting to new technologies [2]. There are several threats and consequences in data breaching. Cyber Threat Management can be taken into consideration. Figure 1 depicts a architecture for predicting cyber hacking

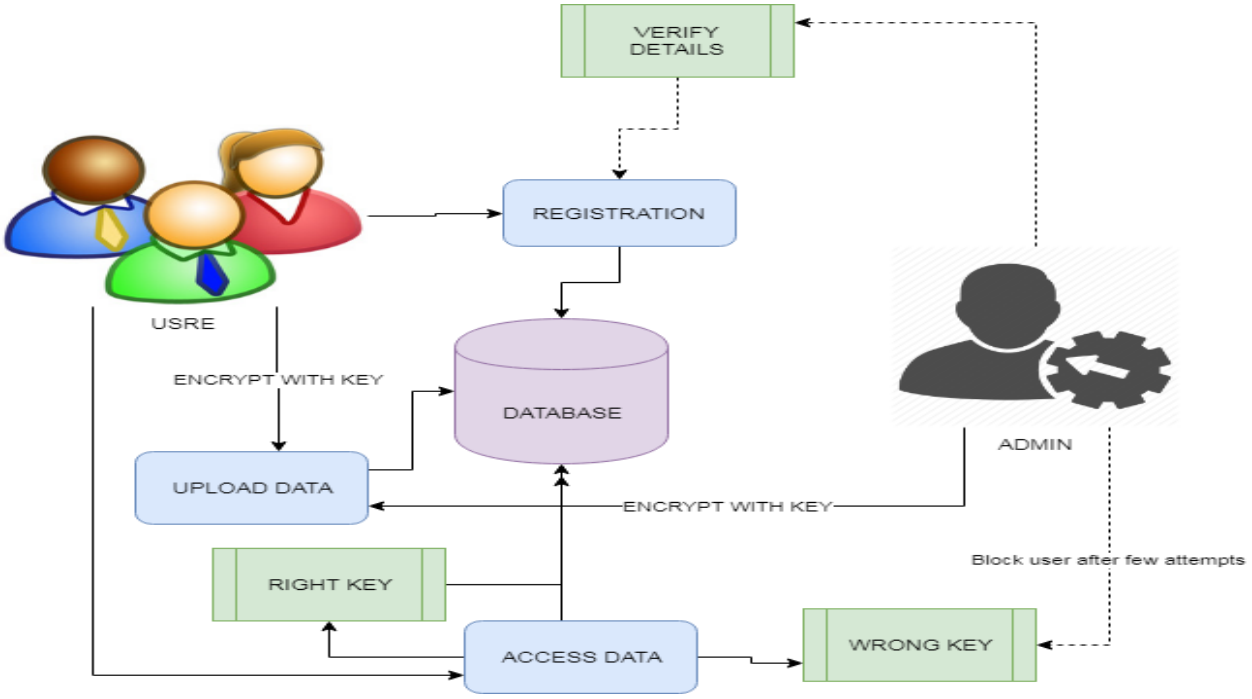


Fig.1: Block Diagram for Predicting Cyber Hacking

An information rupture is a security occurrence in which delicate, ensured or secret information is duplicated, transmitted, saw, stolen or utilized by an individual unapproved to do as such." An information break is the purposeful or accidental arrival of secure or private/classified data to an untrusted domain. Different expressions for this marvel incorporate inadvertent data divulgence, information spill and furthermore information spill [3]. This may incorporate occurrences, for example, robbery or loss of advanced media, for example, PC tapes, hard drives, or smart phones such media whereupon such data is put away decoded, posting such data on the internet or on a PC generally available from the Internet without legitimate data security safeguards, exchange of such data to a framework which isn't totally open yet isn't fittingly or formally authorize for security at the affirmed dimension, for example, decoded email - or exchange of such data to the data frameworks of a conceivably unfriendly office, for example, a

contending organization or a remote country, where it might be presented to increasingly serious unscrambling strategies [4].

While mechanical arrangements can solidify digital frameworks against assaults, information breaks keep on being a major issue. This propels us to describe the development of information rupture occurrences. This not exclusively will profound our comprehension of information breaks, yet in addition shed light on different methodologies for relieving the harm, for example, protection [5]. Many trust that protection will be valuable, however the advancement of accurate cyber hazard measurements to control the task of protection rates is past the compass of the present comprehension of information breaks In this paper, we make the accompanying commitments. We shoe that as opposed to by circulating the ruptures we should demonstrate by stochastic procedure both the hacking break occurrence entomb entry times and rupture sizes. We demonstrate that these stochastic procedure models can foresee the between landing times and the rupture sizes. To the best of our knowledge, this is the primary paper appearing stochastic procedures, instead of circulations, ought to be utilized to show these digital danger factors. We demonstrate that the reliance between the episode's entry time and the break sizes can be satisfactorily depicted by a specific copula [6].

This the primary works demonstrating the presence of this reliance and the results of disregarding it. We additionally demonstrate that it is important to consider the reliance while foreseeing entomb entry times and break sizes generally the outcomes are not accurate. We hope the present study will inspire more investigations, which can offer deep insights into alternate risk mitigation approaches. Such insights are useful to insurance companies, government agencies, and regulators because they need to deeply understand the nature of data breach risks. We hope the present study will inspire more investigations, which can offer deep insights into alternate risk mitigation approaches. Such insights are useful to insurance companies, government agencies, and regulators because they need to deeply understand the nature of data breach risks [7].

2. LITERATURE SURVEY

The nature of the system breaches and the attacks on the system affects the state of operation and working of the system. A system may incur active or passive attack which makes the whole system collapse. When a system is attacked, the data security is breached and all the information contained in the system are hacked or obtained by the hacker in the successful attack. When a system is under attack and if the access to the system is granted, all the potential information will be lost or damaged depending on the intention of the attacker [8]

System States & Cyber-attacks

In order to know the details of the current state of the system, the changes that are made by the cyber attacks must be analysed and the ways in which system has experienced the attack with respective to the changes to the operating system. The purpose and intention of the attacker is to intrude into the system and gain unauthorized access to the system or the information and the resources contained in the system under attack. A malicious code will be sent to the system without the knowledge of the system's owner which can be able to write or transmit the data from the system to the attacker's system through which he can exploit its resources

Contemporary Attacks

These types of attacks are carried out in order to gain elevated or higher access privileges. Through the cotemporary attacks, the attacker can gain administrative privileges of the system under attack. Any modification, changes that are intended by the attacker can be carried out at once he has access to the administrative privileges of the system. The third type of the cotemporary attack can make the system in operable and isolate the system by flooding the information and data contained in the system. This will make the system unresponsive the administrative privileges. The system will respond to the attacker rather than the owner of the system

Determining the breach probability

By comparing the statistics of the attacks in the past on the system and similar type of attacks across the world and the respective models are taken into account for determining the probability of the attacks across the system. Analysing the breach probability is an important objective for the system security and protection. It analyses the attacks that succeeded inspite of the different counter measures taken by the system administrator and it assess the risks and threats that are posed by the cyber-attacks. If the counter measures are involved during the cyber-attack, then the overall breach probability will be able to compute the breach probability.

Determining the Access Matrix

We can identify the nature of the access granted to the system to an attacker by listing the attack matrix and the access matrix is determined by coupling with the task of the attack matrix. The privileges that are granted to the attacker are enlisted in the form of matrix and the different types of attacks that are made to breach the security of the system and the combination of the modality is listed in the access matrix

Advanced Persistent Threat

An attack in a network in which a person extracts a network and access important and highly confidential information rather than doing any actual damage to the network or an organization.

3. SECURITY ISSUES

Due to these various breaches and cyber-attacks that take place in various systems this has led to a significant financial loss as these hackers stole account information and breach security to relocate money to their account. These threats can range from small losses to an entire information loss. These threats can affect at various levels also like some affect confidentiality of data and others affect the entire system [9]. Many people and organizations are struggling to understand what sought of breach or threat has occurred to their systems and how can they protect their information from such other attacks causing massive losses. There are various types of attackers that attack in different methods. Some such attackers are briefed below.

Bot-network Operators

Bot-network operators are hackers that penetrate into the networks. They do so to take over multiple systems. Like this the whole organization can be brought down and malicious attacks can be executed. These network services are made available to shady markets and hence can be misused.

Criminal Groups

These group of people or hackers attack the systems for getting financial profits. Different groups use various ways tod a malicious attack and acquire all the confidential information to commit identity theft and online fraud.

Hackers

These group of people breach into systems to challenge or for bragging rights. This requires a good skill or computer knowledge to breach into the systems or securities. They pose a high threat causing massive damage world-wide. Once they understand the algorithm to crack the security of any site then they can do anything they want to the system.

Insiders:

These are the people who are already working inside the organization. They have all the liberty to access to the system, hence they can easily understand the system and can use it for their own use. They can steal crucial information. The insider threat also includes outsourcing of data and inception of malware into systems.

Phishers

Phishers are groups that use the phishing scheme so that they can steal information for own financial profit. They may also introduce divers' ways as spam in pursuance of their objectives.

4. PROPOSED SYSTEM

Dataset

The hacking breach dataset we analyse in this paper was obtained from the Privacy Rights Clearinghouse (PRC) [1], which is the largest and most extensive dataset that is also publicly available. Since we focus on hacking breaches, we disregard the negligent breaches and the other sub-categories of malicious breaches (i.e., insider, payment card fraud, and unknown). From the remaining raw hacking breaches data, we further disregard the incomplete records with unknown/unreported/missing hacking breach sizes because breach size is one of the objects for our study. The resulting dataset contains 600 hacking breach incidents in the United States between January 1st, 2005 and April 7th, 2017. The hacking breach victims span over 7 industries: businesses-financial and insurance services (BSF); businesses retail/merchant including online retail (BSR); businesses-other (BSO); educational institutions (EDU); government and military (GOV); healthcare, medical providers and medical insurance services (MED); and non-profit organizations (NGO).

Here we make the following three contributions: -

First, we show that both the hacking breach incident inter arrival times (reflecting incident frequency) and breach sizes should be modelled by stochastic processes, instead of distributions. Because they exhibit autocorrelation. We can describe the evolution of the hacking breach incidents inter-arrival times and that a particular ARMA-GARCH model can adequately describe the evolution of the hacking breach sizes. Where ARMA is acronym for “Auto Regressive and Moving Average” and GARCH is acronym for “Generalized Auto Regressive Conditional Heteroskedasticity”. We show that these stochastic process models can predict the inter-arrival times and the breach sizes. Here we are using those stochastic processes, rather than distributions, should be used to model these cyber threat factors.

Second, we discover a positive dependence between the incidents inter-arrival times and therefore the breach sizes.

Third, we conduct both qualitative and quantitative trend analyses of the cyber hacking breach incidents. We find that the situation is indeed getting worse in terms of the incidents inter-arrival time because hacking breach incidents become more and more frequent, but the situation is stabilizing in terms of the incident breach size, indicating that the damage of individual hacking breach incidents will not get much worse.

This is the first paper showing that the stochastic process model rather than distribution. It will help for the reducing inter-arrival time and breach sizes. We also show that when predicting inter-arrival times

and breach sizes, it is necessary to consider the dependence; otherwise, the prediction are not accurate. The third we conduct both qualitative and quantitative breach analysis of cyber hacking breach incidents. Here we use a SUPPORT VECTOR MACHINE algorithm to solve the problems. “Support Vector Machine” (SVM) is a supervised machine learning algorithm which can be used for both classification and regression challenges. It is mostly used in classification.

SUPPORT VECTOR MACHINE

Support Vector Machines (SVM) [10] were first introduced in the mid of 1990s, and have since been established as one of standard tools for machine learning and data mining. SVM were originally designed for binary classification. However, cyber-attack detection is a problem of multi-class classification. How to effectively extend SVM for multi-class classification is still an ongoing research issue. Currently there are two types of approaches for multi-class SVM. One is by combining several binary classifiers while the other is by directly considering all training samples into one optimization formulation.

The SVM identifies the best separating hyper plane (the plan with maximum margins) between the two classes of the training samples within the feature space by focusing on training cases placed at the edge of the class descriptors. In this way, not only an optimal hyper plane is fitted, but also less training samples are effectively used; thus high classification accuracy is achieved with small training sets.

We trust the current examination will rouse more examinations, which can offer profound bits of knowledge into exchange chance alleviation draws near. Such experiences are helpful to insurance agencies, government offices, and controllers since they have to profoundly comprehend the idea of information penetrate dangers.

There are many open issues that are left for future exploration. For instance, it is both intriguing and testing to research how to anticipate the amazingly huge qualities and how to manage missing information (i.e., penetrate episodes that are not revealed). It is likewise advantageous to appraise the specific happening times of penetrate episodes. At last, more exploration should be led towards understanding the consistency of break occurrences (i.e., the upper bound of expectation exactness).

5. MODULES

UPLOAD DATA

The data resource to database can be uploaded by both administrator and authorized user. The data can be uploaded with key in order to maintain the secrecy of the data that is not released without knowledge of

user. The users are authorized based on their details that are shared to admin and admin can authorize each user. Only Authorized users are allowed to access the system and upload or request for files.

ACCESS DETAILS

The access of data from the database can be given by administrators. Uploaded data are managed by admin and admin is the only person to provide the rights to process the accessing details and approve or unapproved users based on their details.

USER PERMISSIONS

The data from any resources are allowed to access the data with only permission from administrator. Prior to access data, users are allowed by admin to share their data and verify the details which are provided by user. If user is accessing the data with wrong attempts, then users are blocked accordingly. If user is requested to unblock them, based on the requests and previous activities admin is unblock users.

DATA ANALYSIS

Data analyses are done with the help of graph. The collected data are applied to graph in order to get the best analysis and prediction of dataset and given data policies. The dataset can be analysed through this pictorial representation in order to better understand of the data details

6. Experimental Result

This experiment uses Windows as its operating system. Python is the programming language, and Django is the web framework. Below table shows the specific equipment arrangement.

Table 1: Equipment Arrangement

Developing Tool	PyCharm
Database	MySQL
CPU	Core i5
Operating System	Windows 10
Hard Disk	128GB SSD
RAM	8GB

Strategy-

We can describe the evolution of the hacking breach incidents inter-arrival times and that a particular ARMA-GARCH model can adequately describe the evolution of the hacking breach sizes. First, we show

that both the hacking breach incident inter arrival times and breach sizes should be modeled by stochastic processes, instead of distributions. Because they exhibit auto-correlation. The steps in the procedure are as follows.

1) Normalization of every dataset.

2) Convert that dataset into the testing and training.

3) Form Intrusion Detection System models with the help of SVM algorithms.

4) Evaluate every model's performance.

Implementation-

In this paper, we make the following three contributions. First, we show that both the hacking breach incident interarrival times (reflecting incident frequency) and breach sizes should be modeled by stochastic processes, rather than by distributions. We find that a particular point process can adequately describe the evolution of the hacking breach incidents inter-arrival times and that a particular ARMA-GARCH model can adequately describe the evolution of the hacking breach sizes, where ARMA is acronym for "Auto Regressive and Moving Average" and GARCH is acronym for "Generalized Auto Regressive Conditional Heteroskedasticity".

Fig.2 displays the web page using which data can be uploaded to the database. Uploaded data are managed by admin and admin is the only person to provide the accessing rights and approve or unapproved users based on their details.

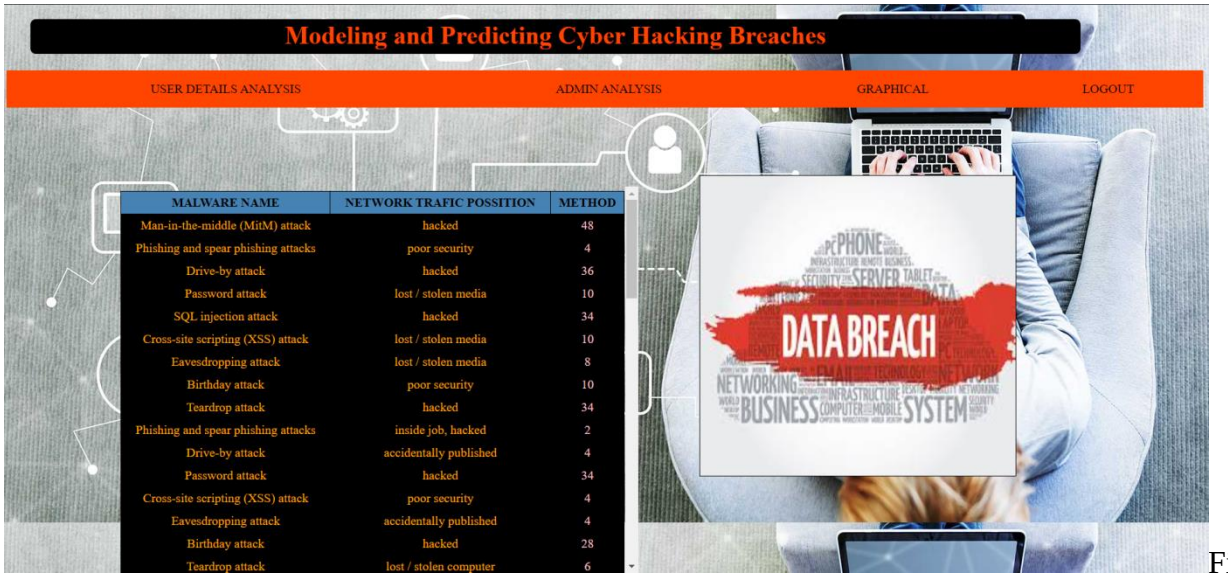
Fig.2: Analysis of the Different Organizations

Fig.3 gives an analysis of the different organizations and their types and also mentions the number of records in each organization. It displays whether these organizations were hacked or not and also shows if they have poor security.

ENTITY	YEAR	RECORDS	ORGANIZATION TYPE	METHOD	DATA
21st Century Oncology	2016	2,200,000	healthcare	hacked	https://www.linkedin.com/jobs/view/930124877/?refId=d3493ec8-privateid.37f8-4218-92b2-4656b383a955&trk card&midToken=serverattack/AQHBnYxQHAJchw&trkEmail=eml-jobs_jymbii_digest-null-4-null-null-9xzoen jobs~view&lipi=urn%3Ali%3Apage%3Aemail_jobs_jymbii_digest%3BCxmcCwrxR62ABhqSrl2dYA%3D%3D
Accendo Insurance Co.	2011	175,350	healthcare	poor security	https://www.bayt.com/en/job-seekers/create-account/?url_id=1&utm_medium=associate&utm_source=walkinu/
Adobe Systems	2013	152,000,000	tech	hacked	https://www.google.co.in/search?ei=9pzSW4rJA8zWvgSzuYDwBg&q=brainmagic+infotech+NLOM]pvt+ltd+g ab.1.0.0i71k1l4.0.0.0.709767.0.0.0.0.0.0.0.0..0.0....0...1c..64.psy-ab..0.0.0....0.YV8QKntrcq4
Advocate Medical Group	2013	4,000,000	healthcare	lost / stolen media	https://stackoverflow.com/questions/43727583/expected-string-or-bytes-like-object/ECSID/getmonlist
AerServ (subsidiary of InMobi)	2018	75,000	advertising	hacked	https://www.google.co.in/search?q=python+free+online+course+certification&oq=p&aqs/2F4=chrome.0.69i59j6
Affinity Health Plan, Inc.	2009	344,579	healthcare	lost / stolen media	http://localhost/phpmyadmin/index.php?token=27a2ceb5cb82727c1f0b63f93f1d0c4f#PMAURL-/x25/NLOM;2: db=malware_detection&table=user_malware_recogition_model&server=1&target=&token=27a2ceb5cb82727c1
Ameritrade	2005	200,000	financial	lost / stolen media	https://realpython.com/pypi-publish-python-package/ECSID/ICMPID
Ancestry.com	2015	300,000	web	poor security	https://mail.google.com/mail/u/0/#inbox/ECSID/getmonlist

Fig.3: Analysis of the Different Organizations

Fig.4 gives an analysis of the different kinds of attacks that have occurred and also the network traffic position. It shows whether the attack occurred accidentally or was intentional.



MALWARE NAME	NETWORK TRAFIC POSSITION	METHOD
Man-in-the-middle (MitM) attack	hacked	48
Phishing and spear phishing attacks	poor security	4
Drive-by attack	hacked	36
Password attack	lost / stolen media	10
SQL injection attack	hacked	34
Cross-site scripting (XSS) attack	lost / stolen media	10
Eavesdropping attack	lost / stolen media	8
Birthday attack	poor security	10
Teardrop attack	hacked	34
Phishing and spear phishing attacks	inside job, hacked	2
Drive-by attack	accidentally published	4
Password attack	hacked	34
Cross-site scripting (XSS) attack	poor security	4
Eavesdropping attack	accidentally published	4
Birthday attack	hacked	28
Teardrop attack	lost / stolen computer	6

Fig.4: Analysis of the Different kinds of Attacks

Fig.5 gives the graphical analysis of all the breaches that have occurred with the number of breaches on x-axis and the year of occurrence on the y-axis.

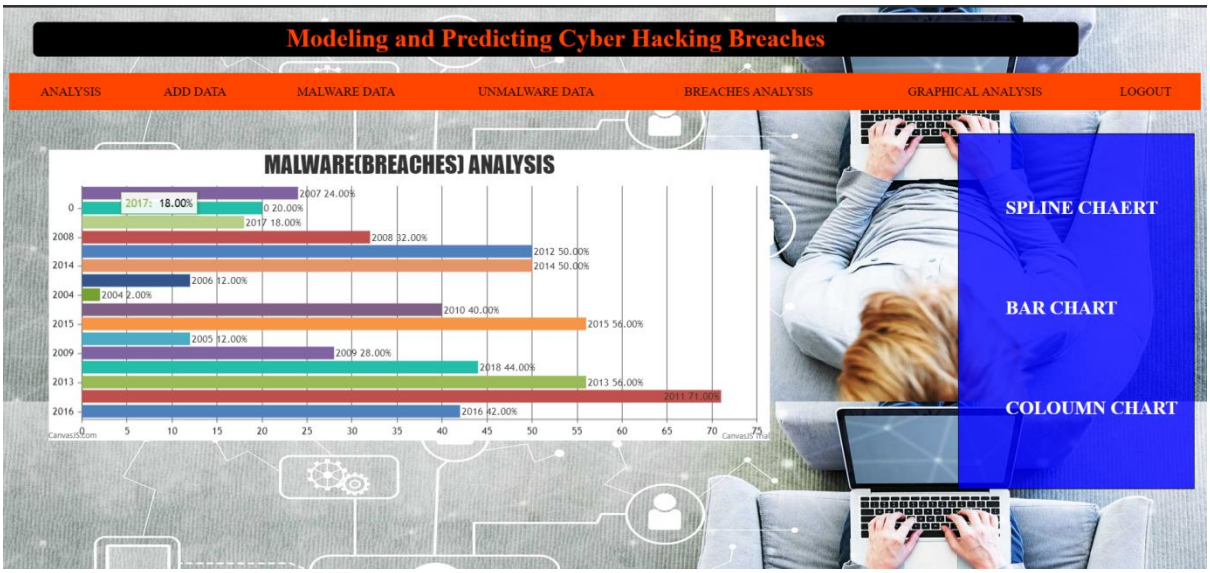


Fig.5: Graphical Analysis of all the Breaches

7. CONCLUSIONS

We analyzed a hacking breach dataset from the points of view of the incidents inter-arrival time and the breach size, and showed that they both should be modeled by stochastic processes rather than distributions. The statistical models developed in this paper show satisfactory fitting and prediction accuracies. In particular, we propose using a copula-based approach to predict the joint probability that an incident with a certain magnitude of breach size will occur during a future period of time. Statistical tests show that the methodologies proposed in this paper are better than those which are presented in the literature, because the latter ignored both the temporal correlations and the dependence between the incidents inter-arrival times and the breach sizes. We conducted qualitative and quantitative analyses to draw further insights. We drew a set of cybersecurity insights, including that the threat of cyber hacking breach incidents is indeed getting worse in terms of their frequency, but not the magnitude of their damage. The methodology presented in this paper can be adopted or adapted to analyze datasets of a similar nature.

REFERENCES

[1]. P. R. Clearinghouse. Privacy Rights Clearinghouse's Chronology of Data Breaches. Accessed: Nov. 2017.

[2] ITR Center. Data Breaches Increase 40 Percent in 2016, Finds New Report From Identity Theft Resource Center and CyberScout. Accessed: Nov. 2017.

[3] C. R. Center. Cybersecurity Incidents. Accessed: Nov. 2017.

[4] IBM Security. Accessed: Nov. 2017.

[5] NetDiligence. The 2016 Cyber Claims Study. Accessed: Nov. 2017.

[6] M. Eling and W. Schnell, “What do we know about cyber risk and cyber risk insurance?” J. Risk Finance, vol. 17, no. 5, pp. 474–491, 2016.

[7] T. Maillart and D. Sornette, “Heavy-tailed distribution of cyber-risks,” Eur. Phys. J. B, vol. 75, no. 3, pp. 357–364, 2010.

[8] R. B. Security. Datalossdb. Accessed: Nov. 2017.

[9] B. Edwards, S. Hofmeyr, and S. Forrest, “Hype and heavy tails: A closer look at data breaches,” J. Cybersecur., vol. 2, no. 1, pp. 3–14, 2016.

[10] S. Wheatley, T. Maillart, and D. Sornette, “The extreme risk of personal data breaches and the erosion of privacy,” Eur. Phys. J. B, vol. 89, no. 1, p. 7, 2016

STUDENT PROFILES

Donthi Reddy Abhishek Reddy is currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. He completed his intermediate from Sri Chaitanya Junior College and 10th class from Sri Chaitanya Techno School. His technical skills include Java, Python. He also has a basic understanding of C. He is one of the students of Smart Interviews and participated in few tests conducted by them. His participations include: National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020, and he has completed few courses on coursera in the year 2020 and have also participated in Two-Day National Level Seminar On "Recent Trends in Cloud Computing, Fog and Edge Computing" scheduled on 18th June to 19th June 2021 April to 22nd May 2020. His areas of interest are Python, Java, Cloud Computing, Cyber Security, Machine Learning. He completed few certification courses from online platforms like Coursera, Data Camp.

Saddi Advaitha Reddy is currently pursuing her Bachelor of Technology in the stream of Computer Science & Engineering at St. Martin's Engineering College. She completed her intermediate from Sri Chaitanya Junior College and schooling from St. Peter's High School. She was a member of Events Department for two consecutive years in Technology Awareness Month (TAM) in our college. Her responsibilities in that group included orchestrating the work of the crew and binding the members together to work as an operationally active team. Apart from this she was also a volunteer in the student run NGO, Street Cause, during the year 2017-2018. Her technical skills include C, C++, Python and Java. She took part in Employability Skill Development Program conducted by Zensar. Her participations include: Workshop on "HTML & CSS" which was conducted in the college on 29th January 2018 and 30th January 2018, Workshop on "Arduino/Robotics" which was conducted in the college on 12th February 2019 and 13th February 2019, Workshop on "Ethical Hacking" which was conducted in the college on 31st January 2020 and 1st February 2020, National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020, Women online workshop on "Women in Cyber Security and Privacy in 2020" which was conducted from 6th to 10th July 2020.She was also a student organizing member during two days "National Level Hackathon-2020" held on 7th and 8th February 2020 at the college. She spends her free time taking online certification courses related to her field of study as well as personal interests from platform such as Coursera, Cursa and EdX. Her areas of interest are Python, Artificial Intelligence, Machine Learning and Deep Learning.

Katta Pradhyun Reddy is pursuing his Bachelor of Technology in the stream of Computer Science & Engineering at St. Martin's Engineering College. He completed his intermediate from Narayana Junior College and schooling from RBVRR High School. His technical skills include C, C++, Java, HTML and Python. He has completed few certificate courses from online platforms like Coursera on Python Programming, AI for Everyone, HTML5, CSS, Data Analysis, Managing Project Risks and Changes. His participations include Workshop on "Ethical Hacking" which was conducted by college on 31st January to 1st February 2020, National Level Three Day Online Workshop on "AI & ML in speech and audio processing" which was conducted from 10th and 12th December, 2020, Leadership Talk with Mr. Mahesh Babu CEO Mahindra Electric Mobility Ltd. His areas of interest are Machine Learning and Deep Learning.

Nalla Rakshitha is currently pursuing her Bachelor of Technology in the stream of Computer Science & Engineering at St. Martin's Engineering College. She completed her intermediate from Narayana Junior College and schooling from Delhi Public School. She was an active volunteer in the student run NGO, Street Cause, during the year 2017-2018. Her technical skills include C, C++, Python and Java. She took part in Employability Skill Development Program conducted by Zensar. Her participations include: Workshop on "HTML & CSS" which was conducted in the college on 29th January 2018 and 30th January 2018, Workshop on "Arduino/Robotics" which was conducted in the college on 12th February 2019 and 13th February 2019, "Circuitronics Quiz" which was conducted by TAM from 21st January 2020 to 24th February 2020 and secured second position in the quiz , National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020, Women online workshop on "Women in Cyber Security and Privacy in 2020" which was conducted from 6th to 10th July 2020. She spends her free time taking online certification courses related to her field of study as well as personal interests from platform such as Coursera, Cursa and EdX. Her areas of interest are Python, Artificial Intelligence, Machine Learning and Deep Learning.

APPENDICES

```
from django.contrib import messages
from django.contrib.auth import authenticate
from django.db.models import Q, Count
from django.shortcuts import render, redirect

# Create your views here.
from Cyber_Users.forms import UserRegister_Form
from Cyber_Users.models import UserRegister_Model, UserAdd_Model



def user_login(request):
    if request.method == "POST":
        name = request.POST.get('name')
        password = request.POST.get('password')
        try:

            check = UserRegister_Model.objects.get(name=name,password=password)
            request.session['userid'] = check.id
            return redirect('user_adddata')
        except:
            pass
        user = authenticate(name=name,password=password)
        if user is not None:
            if user.is_active:



                return redirect('user_adddata')
        else:
            messages.error(request, 'username or password are not match')

            return redirect('user_login')

    return render(request, 'users/user_login.html')
```

```
def user_register(request):
    if request.method == "POST":
        forms = UserRegister_Form(request.POST)
        if forms.is_valid():
            forms.save()
            messages.success(request, 'You have been successfully registered')
            return redirect('user_login')
    else:
        forms = UserRegister_Form()

    return render(request,'users/user_register.html',{'form':forms})



def user_adddata(request):
    userid = request.session["userid"]
    obj = UserRegister_Model.objects.get(id=userid)
    attack1 = []
    attack2, attack3, attack4, attack5, attack6, attack7, attack8, attack9 = [], [], [], [], [], [], [], []

    splt = ''
    Entity = ''
    Year = 0
    Records = ''
    Organizationtype = ''
    Method = ''
    txt =''
    Adddata = ''
    ans = ''
    Time = ''
    if request.method == "POST":
        Entity = request.POST.get("entity")
        Year = request.POST.get("year")
        Records = request.POST.get("records")
```

```
    Organizationtype = request.POST.get("organizationtype")
    Method = request.POST.get("method")
    txt = request.POST.get("name")
    Time = request.POST.get("time")




    splt = (re.findall(r"[\w']+", str(txt)))

for f in splt:
    if f in ('IPid', 'FDDI', 'x25', 'rangingdistance'):
        attack1.append(f)
    elif f in ('tcpchecksum', 'mtcp', 'controlflags', 'tcpoffset', 'tcpport'):
        attack2.append(f)
    elif f in ('ICMPID', 'udptraffic', 'udpunicorn', 'datagramid', 'NTP', 'RIP', 'TFTP'):
        attack3.append(f)
    elif f in ('GETID', 'POSTID', 'openBSD', 'appid', 'sessionid', 'transid', 'physicalid'):
        attack4.append(f)
    elif f in ('SYN', 'ACK', 'synpacket', 'sycookies'):
        attack5.append(f)
    elif f in ('serverattack', 'serverid', 'blockbankwidth'):
        attack6.append(f)
    elif f in ('monlist', 'getmonlist', 'NTPserver'):
        attack7.append(f)
    elif f in ('portid', 'FTPID', 'tryion', 'fragflag'):
        attack8.append(f)
    elif f in ('malwareid', 'gethttpid', 'httpid'):
        attack9.append(f)

if len(attack1) > len(attack2) and len(attack1) > len(attack3) and len(attack1) > len(attack4) and len(
        attack1) > len(attack5) and len(attack1) > len(attack6) and len(attack1) > len(attack7) and len(
    attack1) > len(attack8) and len(attack1) > len(attack9):
    ans = "Man-in-the-middle Attack"
elif len(attack2) > len(attack1) and len(attack2) > len(attack3) and len(attack2) > len(attack4) and len(
```

```
        attack2) > len(attack5) and len(attack2) > len(attack6) and len(attack2) > len(attack7) and len(
    attack2) > len(attack8) and len(attack2) > len(attack9):
    ans = "Phishing and spear phishing attacks"
elif len(attack3) > len(attack2) and len(attack3) > len(attack1) and len(attack3) > len(attack4) and len(
        attack1) > len(attack5) and len(attack1) > len(attack6) and len(attack1) > len(attack7) and len(
    attack1) > len(attack8) and len(attack1) > len(attack9):
    ans = "Drive-by attack"
elif len(attack4) > len(attack2) and len(attack4) > len(attack3) and len(attack4) > len(attack1) and len(
        attack4) > len(attack5) and len(attack4) > len(attack6) and len(attack4) > len(attack7) and len(
    attack4) > len(attack8) and len(attack4) > len(attack9):
    ans = "Password attack"
elif len(attack5) > len(attack2) and len(attack5) > len(attack3) and len(attack5) > len(attack4) and len(
        attack5) > len(attack1) and len(attack5) > len(attack6) and len(attack5) > len(attack7) and len(
    attack5) > len(attack8) and len(attack5) > len(attack9):
    ans = "SQL injection attack"
elif len(attack6) > len(attack2) and len(attack6) > len(attack3) and len(attack6) > len(attack4) and len(
        attack6) > len(attack5) and len(attack6) > len(attack1) and len(attack6) > len(attack7) and len(
    attack6) > len(attack8) and len(attack6) > len(attack9):
    ans = "Cross-site scripting (XSS) attack"
elif len(attack7) > len(attack2) and len(attack7) > len(attack3) and len(attack7) > len(attack4) and len(
        attack7) > len(attack5) and len(attack7) > len(attack6) and len(attack7) > len(attack1) and len(
    attack7) > len(attack8) and len(attack7) > len(attack9):
    ans = "Eavesdropping attack"
elif len(attack8) > len(attack2) and len(attack8) > len(attack3) and len(attack8) > len(attack4) and len(
        attack8) > len(attack5) and len(attack8) > len(attack6) and len(attack8) > len(attack7) and len(
    attack8) > len(attack1) and len(attack8) > len(attack9):
    ans = "Birthday attack"
elif len(attack9) > len(attack2) and len(attack9) > len(attack3) and len(attack9) > len(attack4) and len(
        attack9) > len(attack5) and len(attack9) > len(attack6) and len(attack9) > len(attack7) and len(
    attack9) > len(attack8) and len(attack9) > len(attack1):
    ans = "Teardrop attack"


else:
    ans = "Unmalware"
```

```
UserAdd_Model.objects.create(uregid=obj,entity=Entity,year=Year,records=Records,organizationtype=O
rganizationtype,method=Method,adddata=txt,attackresult=ans,time=Time)


    return render(request,'users/user_adddata.html')

def user_page(request):
    obj = UserAdd_Model.objects.all()
    return render(request,'users/user_page.html',{'object':obj})


def malware(request):
    obj = UserAdd_Model.objects.filter(Q(attackresult='Man-in-the-middle (MitM) attack') |
Q(attackresult='Phishing and spear phishing attacks') | Q(
        attackresult='Drive-by attack') | Q(attackresult='Password attack') | Q(
        attackresult='SQL injection attack') | Q(attackresult='Cross-site scripting (XSS) attack') |
Q(attackresult='Eavesdropping attack') | Q(
        attackresult='Birthday attack') | Q(attackresult='Teardrop attack'))
    return render(request,'users/malware.html',{'object':obj})

def unmalware(request):
    obj = UserAdd_Model.objects.filter(attackresult='Unmalware')
    return render(request,'users/unmalware.html',{'object':obj})

def breaches_analysis(request):
    chart = UserAdd_Model.objects.values('attackresult','method').annotate(dcount=Count('attackresult'))
    return render(request,'users/breaches_analysis.html',{'objects':chart})

def chart_page(request,chart_type):
    chart = UserAdd_Model.objects.values('year').annotate(dcount=Count('organizationtype'))
    return render(request,'users/chart_page.html',{'chart_type':chart_type,'objects':chart})
```

A

PROJECT REPORT

On

MOVIE RECOMMENDATION SYSTEM USING SENTIMENT ANALYSIS FROM MICROBLOGGING DATA

Submitted by

Ms. D. Varshitha (17K81A0574) **Mr. G. Vinay Prasad (17K81A05B7)**

Mr. R. Sai Chaitanya (17K81A05B0) **Ms. D. Charmitha Rao (17K81A0573)**

in partial fulfillment for the award of the

degreeof

BACHELOR OF TECHNOLOGY

IN

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Under the Guidance of

Mr. V.L. Kartheek

Assistant Professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ST.MARTIN'S ENGINEERING COLLEGE
An Autonomous Institute

Dhulapally, Secunderabad – 500 100

JUNE 2021

BONAFIDE CERTIFICATE

This is to certify that the project entitled **Movie Recommendation Systen using Sentiment Analysis from Microblogging Data**, is being submitted by **D.Varshitha 17K81A0574, G. Vinay Prasad 17K81A05B7, R. Sai Chaitanya 17K81A05B0**, **D. Charmitha Rao 17K81A0573,** in partial fulfillment of the requirement for the award of the degree of **BACHELOR OF TECHNOLOGY** in **Computer Science Of Engineering** is recorded of bonafide work carried out by them. The result embodied in this report have been verified and found satisfactory.

.

Guide

Mr. V.L. Kartheek

Department of CSE

Head of the Department

Dr. M. NARAYANAN

Department of CSE

Internal Examiner

External Examiner

Place:

Date:

DECLARATION

We, the student of **Bachelor of Technology** in Department of Computer Science and Engineering', session: 2017 – 2021, St. Martin's Engineering College, Dhulapally, Kompally, Secunderabad, hereby declare that work presented in this Project Work entitled Movie Recommendation System using Sentiment Analysis from Microblogging Data is the outcome of our own bonafide work and is correct to the best of our knowledge and this work has been undertaken taking care of Engineering Ethics. This result embodied in this project report has not been submitted in any university for award of any degree.

D. Varshitha	17K81A0574
G. Vinay Prasad	17K81A05B7
R. Sai Chaitanya	17K81A05B0
D. Charmitha Rao	17K81A0573

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible and whose encouragements and guidance have crowded effects with success.

We extended our deep sense of gratitude to Principal**,** **Dr. P. SANTOSH KUMAR PATRA**, St. Martin's Engineering College, Dhulapally, for permitting us to undertake this project.

We are also thankful to **Dr. M.NARAYANAN**, Head of the Department, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

We would like to thank **Dr. T. POONGOTHAI,** Department of Computer Science and Engineering (AI & ML) for her encouragement and insightful comments and as well as our project coordinators **Dr. B.RAJALINGAM**, Associate Professor and **Dr. GOVINDA RAJULU. G** Associate Professor, in Department of Computer Science and Engineering for their valuable support.

We would like to express our sincere gratitude and indebtedness to our project supervisor **Mr. V.L. Kartheek ,** Assistant Professor, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

Finally, we express thanks to all those who have helped us successfully to completing this project. Furthermore, we would like to thank our family and friends for their moral support and encouragement.

We express thanks to all those who have helped us in successfully completing the project.

D. Varshitha 17K81A0574

G. Vinay Prasad 17K81A05B7

R. Sai Chaitanya 17K81A05B0

D. Charmitha Rao 17K81A0573

ABSTRACT

Recommendation systems (RSs) have garnered immense interest for applications in e-commerce and digital media. Traditional approaches in RSs include such as collaborative filtering (CF) and content-based filtering (CBF) through these approaches that have certain limitations, such as the necessity of prior user history and habits for performing the task of recommendation. To minimize the effect of such limitation, this article proposes a hybrid RS for the movies that leverage the best of concepts used from CF and CBF along with sentiment analysis of tweets from microblogging sites. The purpose to use movie tweets is to understand the current trends, public sentiment, and user response of the movie. Experiments conducted on the public database have yielded promising results.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF OUTPUT SCREENS

LIST OF ACRONYMS

<RS>	Recommendation System
<CF>	Collaborative Filtering
<CBF>	Content Based Filtering

1.INTRODUCTION

In today's world, internet has become an important part of the human life. Users often face the problem of excessive available information. Recommendation systems (RS) are deployed to help users cope with this information explosion. RS are mostly usedin e-commerce applications and knowledge management systems such as tourism, entertainment and online shopping portals. In this paper, we focus on RS for moviesare an important source of recreation and entertainment in our life.

Movie suggestions for users depend on web-based portals. Movies can be easily differentiated through their genres like comedy, thriller, animation, and action. Another possible way to categorize movies can be achieved on the basis of metadata such as year, language, director or by cast. Most online video-streaming services provide a number of similar movies to the user to utilising the user's previously viewed or rated history. Movie Recommendation Systems help us to search our preferred moviesand also reduce the trouble of spending a lot of time searching for favorable movies.

The primary requirement of a movie recommendation system is that, it should be very reliable and provide the user with the recommendation of movies which are similar to their preferences. In recent times, with exponential increase in amount of online- data, RS are very beneficial for taking decisions in different activities of day-to-day life. RS are broadly classified into two categories: Collaborative filtering (CF) and Content-based filtering (CBF).

The main contributions are as follows:

1. We propose a hybrid recommendation system by combining collaborative filtering and content-based filtering.
2. Sentiment analysis is used to boost up this recommendation system.
3. A detailed analysis of proposed recommendation system is presented through extensive experiment. Finally, a qualitative as well as quantitative comparison with other baselines models is also demonstrated.

1.1. PROJECT OVERVIEW

Users often face the problem of excessive available information. Recommendation systems (RSs) are deployed to help users cope up with the information explosion. RS is mostly used in digital entertainment, such as Netflix, Prime Video, and IMDB, and e-commerce portals such as Amazon, Flipkart, and eBay. In this article, we focus on RS for movies, which is an important source of recreation and entertainment in our life. Movie suggestions for users depend on Web-based portals. Movies can be easily differentiated through their genres, such as comedy, thriller, animation, and action. Another possible way to categorize the movies based on its metadata, such as release year, language, director, or cast. Most online video-streaming services , provide personalized user experience by utilizing the user's historical data, such as previously viewed or rated history.

1.2. PROJECT OBJECTIVE

The purpose to use movie tweets is to understand the current trends, public sentiment, and user response of the movie. Experiments conducted on the public database have yielded promising results.

1.3. ORGANIZATION OF CHAPTERS

This documentation consists of 10 different chapter and they are:

1. Introduction – This chapter covers the overview of our project and its objectives.

2. Literature Survey – This includes the details of our survey.

3. Software and Hardware Requirements – We specify our software and hardware requirements here.

4. Software Development Analysis – This section includes the problem definition and details of the modules we used in our project.

5. Project System Design – This chapter includes the design part of our project which includes uml diagrams.

6. Project Coding – This section contains the details of our project code.

7. Project Testing – The details of test cases and testing are included in this chapter.

8. Output Screens – This contains the screenshots of how our project looks like when executed.

9. Experimental Results – This chapter contains the screenshots of our results.

10. Conclusion and Future Enhancements – This covers the conclusion of our project and the possible future developments.

2.LITERATURE SURVEY

2.1 SURVEY ON BACKGROUND

1. Analyzing user modeling on Twitter for personalized news recommendations

AUTHORS: F. Abel, Q. Gao, G.-J. Houben, and K. Tao.

How can micro-blogging activities on Twitter be leveraged for user modeling and personalization? In this paper we investigate this question and introduce a framework for user modeling on Twitter which enriches the semantics of Twitter messages (tweets) and identifies topics and entities (e.g. persons, events, products) mentioned in tweets. We analyze how strategies for constructing hashtag-based, entity-based or topic-based user profiles benefit from semantic enrichment and explore the temporal dynamics of those profiles. We further measure and compare the performance of the user modeling strategies in context of a personalized news recommendation system. Our results reveal how semantic enrichment enhances the variety and quality of the generated user profiles. Further, we see how the different user modeling strategies impact personalization and discover that the consideration of temporal profile patterns can improve recommendation quality.

2. Twitter-based user modeling for news recommendations

AUTHORS: F. Abel, Q. Gao, G.-J. Houben, and K. Tao.

How can micro-blogging activities on Twitter be leveraged for user modeling and personalization? In this paper we investigate this question and introduce a framework for user modeling on Twitter which enriches the semantics of Twitter messages (tweets) and identifies topics and entities (e.g. persons, events, products) mentioned in tweets. We analyze how strategies for constructing hashtag-based, entity-based or topic-based user profiles benefit from semantic enrichment and explore the temporal dynamics of those profiles. We further measure and compare the performance of the user modeling strategies in context of a personalized news recommendation system. Our results reveal how semantic enrichment enhances the variety and

quality of the generated user profiles. Further, we see how the different user modeling strategies impact personalization and discover that the consideration of temporal profile patterns can improve recommendation quality.

3. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions.

AUTHORS: G. Adomavicius and A. Tuzhilin.

This paper presents an overview of the field of recommender systems and describes the current generation of recommendation methods that are usually classified into the following three main categories: content-based, collaborative, and hybrid recommendation approaches. This paper also describes various limitations of current recommendation methods and discusses possible extensions that can improve recommendation capabilities and make recommender systems applicable to an even broader range of applications. These extensions include, among others, an improvement of understanding of users and items, incorporation of the contextual information into the recommendation process, support for multicriteria ratings, and a provision of more flexible and less intrusive types of recommendations.

4. Enhancing deep learning sentiment analysis with ensemble techniques in social applications.

AUTHORS: O. Araque, I. Corcuera-Platas, J. F. Sánchez-Rada, and C. A. Iglesias.

Deep learning techniques for Sentiment Analysis have become very popular. They provide automatic feature extraction and both richer representation capabilities and better performance than traditional feature based techniques (i.e., surface methods). Traditional surface approaches are based on complex manually extracted features, and this extraction process is a fundamental question in feature driven methods. These long-established approaches can yield strong baselines, and their predictive capabilities can be used in conjunction with the arising deep learning methods. In this paper we seek to improve the performance of deep learning techniques integrating them with traditional surface approaches based on manually extracted

features. The contributions of this paper are sixfold. First, we develop a deep learning based sentiment classifier using a word embeddings model and a linear machine learning algorithm. This classifier serves as a baseline to compare to subsequent results. Second, we propose two ensemble techniques which aggregate our baseline classifier with other surface classifiers widely used in Sentiment Analysis. Third, we also propose two models for combining both surface and deep features to merge information from several sources. Fourth, we introduce a taxonomy for classifying the different models found in the literature, as well as the ones we propose. Fifth, we conduct several experiments to compare the performance of these models with the deep learning baseline. For this, we use seven public datasets that were extracted from the microblogging and movie reviews domain. Finally, as a result, a statistical study confirms that the performance of these proposed models surpasses that of our original baseline on F1-Score.

5. Hybrid recommender systems based on content feature relationship

AUTHORS: E. Aslanian, M. Radmanesh, and M. Jalili

Recommendation systems get ever-increasing importance due to their applications in both academia and industry. The most popular type of these systems, known as collaborative filtering algorithms, employ user-item interactions to perform the recommendation tasks. With growth of additional information sources other than the rating (or purchase) history of users on items, such as item descriptions and social media information, further extensions of these systems have been proposed, known as hybrid recommendation algorithms. Hybrid recommenders use both user-item interaction data and their contextual information. In this work, we propose new hybrid recommender algorithms by considering the relationship between content features. This relationship is embedded into the hybrid recommenders to improve their accuracy. We first introduce a novel method to extract the content feature relationship matrix, and then the collaborative filtering recommender is modified such that this relationship matrix can be effectively integrated within the algorithm. The proposed algorithm can better deal with the cold-start problem than the state-of-art algorithms. We also propose a novel content-based hybrid recommender system. Our experiments on a benchmark movie dataset show that the proposed approach significantly improves the accuracy of the system,

while resulting in satisfactory performance in terms of novelty and diversity of the recommendation lists.

2.2. CONCLUSION ON SURVEY

Recommendation Systems are the most effective knowledge management systems that help users to filter unusable data and deliver personalized ideas based on their past historical data and similar items which user are looking over the internet. Many Recommendation Systems have been developed over the past decades. These systems used different approaches like Collaborative Filtering, Content-Based Filtering , hybrid and sentiment analysis system to recommend items.

3.SOFTWARE AND HARDWARE REQUIREMENTS

The project involved analysing the design of few applications so as to make the application more users friendly. To do so, it was really important to keep the navigations from one screen to the other well ordered and at the same time reducing the amount of typing the user needs to do.

3.1. SOFTWARE REQUIREMENTS

For developing the application the following are the Software Requirements:

- **Technology** **:** Python
- **Frameworks** **:** Tkinter, Pandas
- **Editor** **:** Python IDE
- **Operating System** **:** Microsoft Windows, Linux or Mac any version

3.2. HARDWARE REQUIREMENTS

For developing the application the following are the Hardware Requirements:

- System : Pentium IV 2.4 GHz.
- Hard Disk : 40 GB.
- Floppy Drive : 1.44 Mb.

4.SOFTWARE DEVELOPMEN ANALYSIS

4.1. OVERVIEW OF PROBLEM

The project involved analysing the design of few applications so as to make the application more users friendly. To do so, it was really important to keep the navigations from one screen to the other well-ordered and at the same time reducing the amount of typing the user needs to do. In order to make the application more accessible, the browser version had to be chosen so that it is compatible with most of the Browsers.

4.2. DEFINE THE PROBLEM

Users often face the problem of excessive available information. Recommendation systems (RSs) are deployed to help users cope up with the information explosion. RS is mostly used in digital entertainment, such as Netflix, Prime Video, and IMDB, and e-commerce portals such as Amazon, Flipkart, and eBay. In this article, we focus on RS for movies, which is an important source of recreation and entertainment in our life. Movie suggestions for users depend on Web-based portals. Movies can be easily differentiated through their genres, such as comedy, thriller, animation, and action. Another possible way to categorize the movies based on its metadata, such as release year, language, director, or cast. Most online video-streaming services , provide personalized user experience by utilizing the user's historical data, such as previously viewed or rated history.

4.3. MODULES OVERVIEW

This project has two modules Admin and User and two modules are designed for the interactions between users and application. Each Module has its own functionality. A module allows us to logically organize the code. Grouping related code into a module makes the code easier to understand and use.

And also, by using these two modules helps us to get movie recommendation more easily.

4.4. DEFINE THE MODULES

This application has two modules which are listed in the following.

1.Admin

2.User

4.5. MODULE FUNCTIONALITY

ADMIN:

In this module admin used to login, view all users and add sentiwords. Here Admin also uses Collaborative filtering and Content based filtering by using which he can recommend movies to the user based on either on twitter ratings from the early revies or from the user's movie history.

USER:

In this module user will be first get registered in this application, login with his details, search friends accordingly, requests, post, view all posts and Recommend Movies. User here after login will be getting movie recommendations from the admin based on the genre, directors or actors.

4. PROJECT SYSTEM DESIGN

5.1. SYSTEM ARCHITECTURE

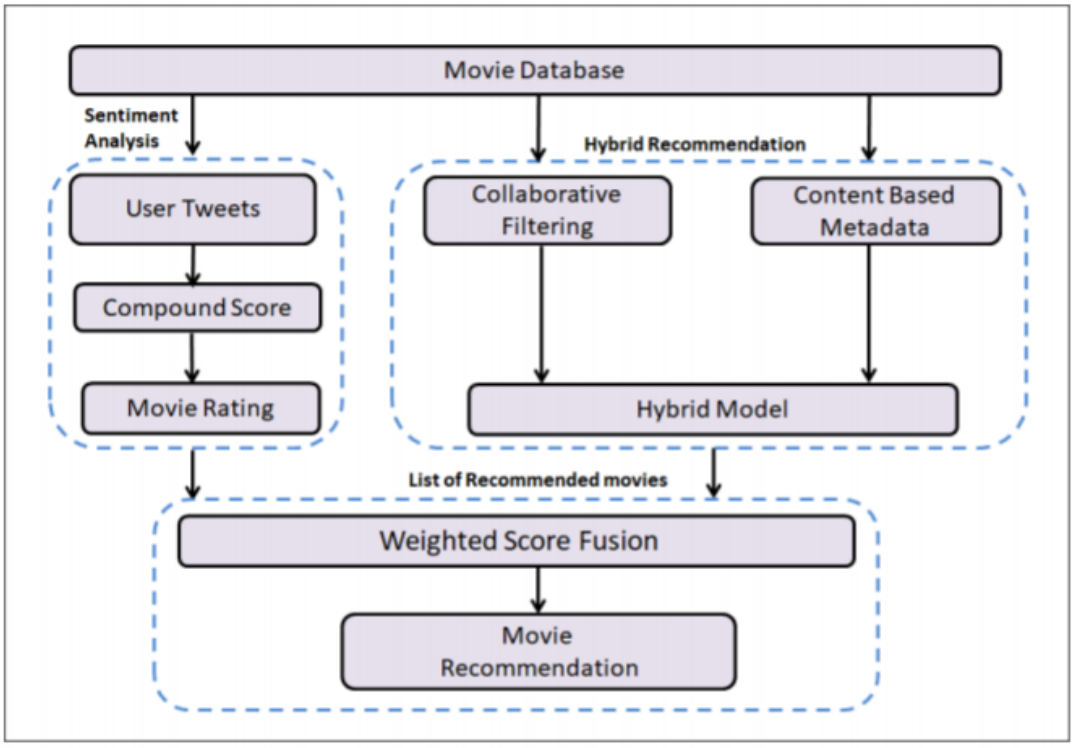


Fig.5.1: Movie Database Architecture

5.2. UML DIAGRAMS

UML stands for Unified Modeling Language. UML is a standardized general-purpose modeling language in the field of object-oriented software engineering. The standard is managed, and was created by, the Object Management Group.

The goal is for UML to become a common language for creating models of object oriented computer software. In its current form UML is comprised of two major components: a Meta-model and a notation. In the future, some form of method or process may also be added to; or associated with, UML.

The Unified Modeling Language is a standard language for specifying, Visualization, Constructing and documenting the artifacts of software system, as well as for business modeling and other non-software systems.

The UML represents a collection of best engineering practices that have proven successful in the modeling of large and complex systems.

The UML is a very important part of developing objects oriented software and the software development process. The UML uses mostly graphical notations to express the design of software projects.

CLASS DIAGRAM:

In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.

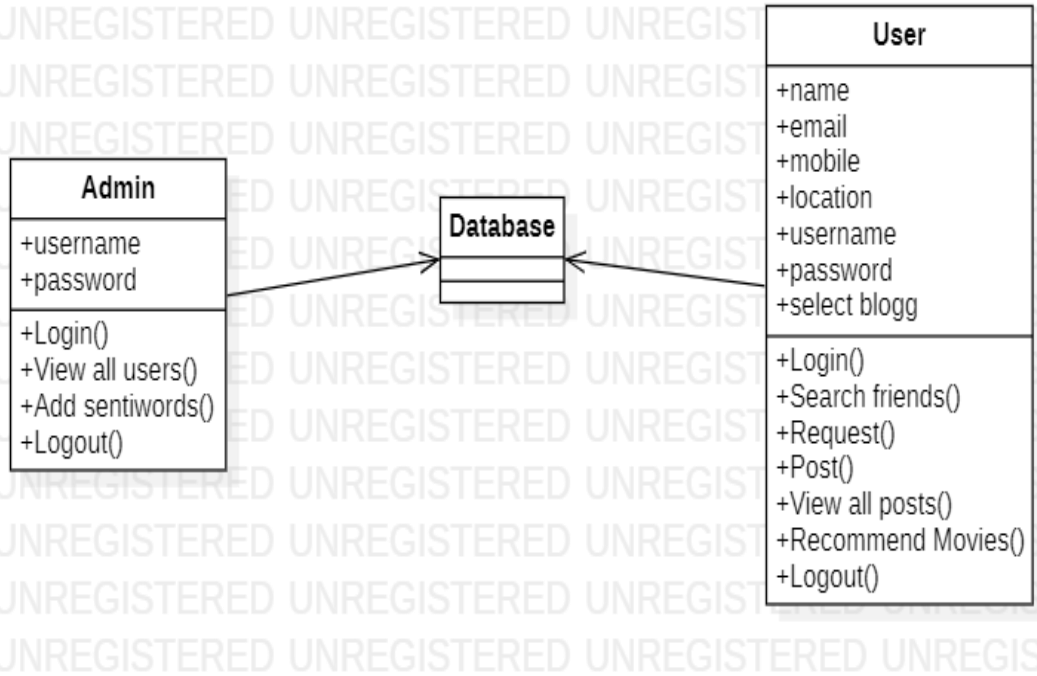


Fig. 5.2: Class Diagram

USE CASE DIAGRAM:

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.

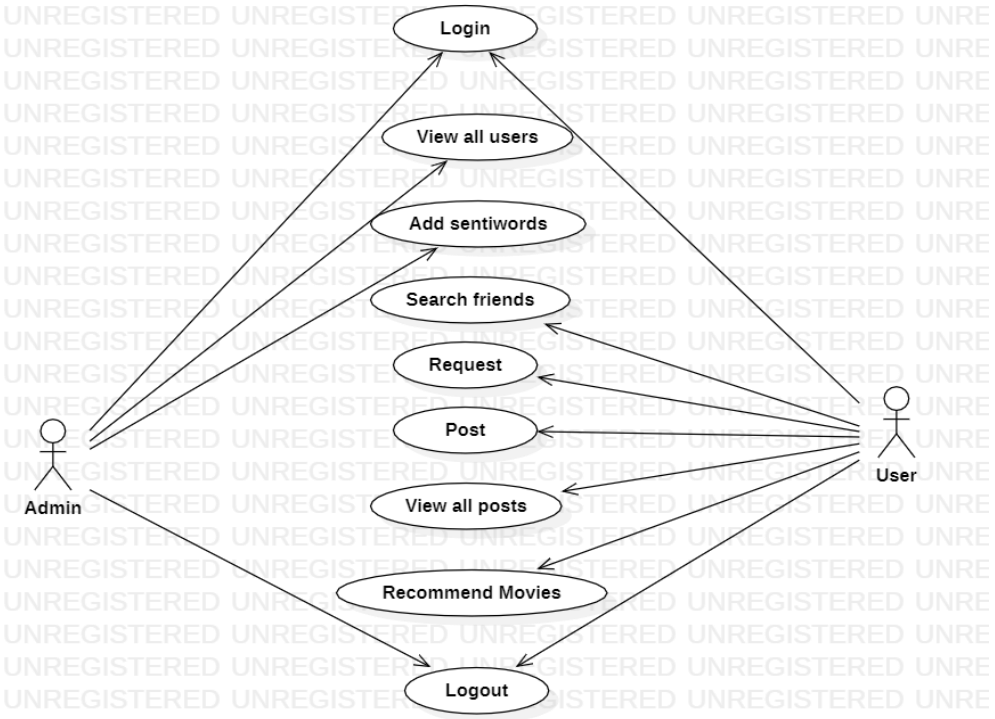


Fig. 5.3: Use Case Diagram

SEQUENCE DIAGRAM

A sequence diagram in Unified Modeling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.

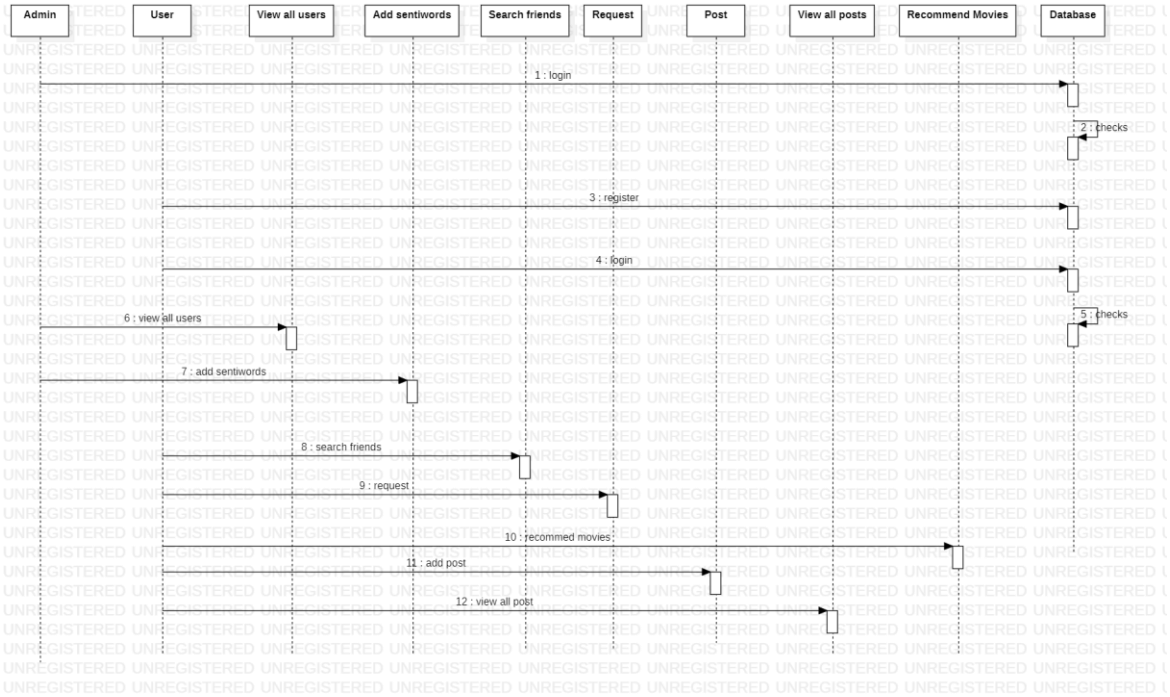


Fig.5.4: Sequence diagram

ACTIVITY DIAGRAM:

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modeling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.

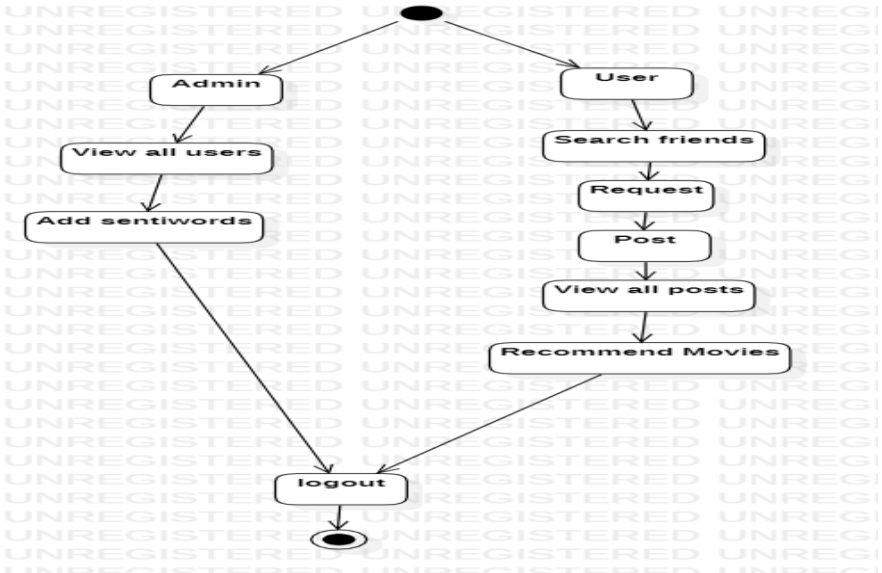


Fig.5.5: Activity Diagram

DEPLOYMENT DIAGRAM

The nodes appear as boxes, and the artifacts allocated to each node appear as rectangles within the boxes. Nodes may have subnodes, which appear as nested boxes. A single node in a deployment diagram may conceptually represent multiple physical nodes, such as a cluster of database servers.

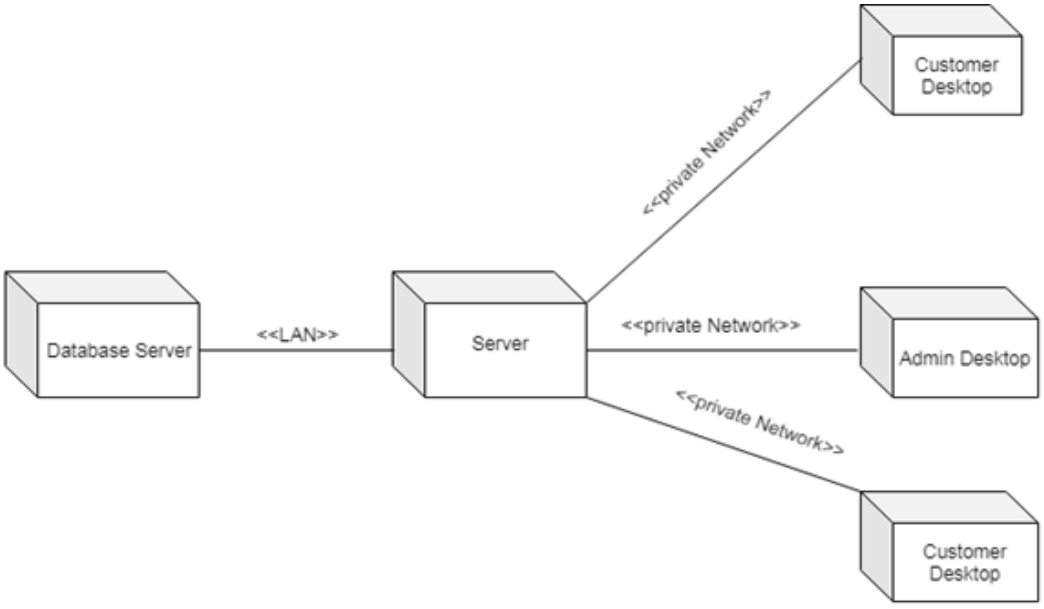


Fig.5.6: Deployment diagram

PACKAGE DIAGRAM

Package diagram is UML structure diagram which shows structure of the designed system at the level of packages. The following elements are typically drawn in a package diagram: package, packageable element, dependency, element import, package import, package merge.

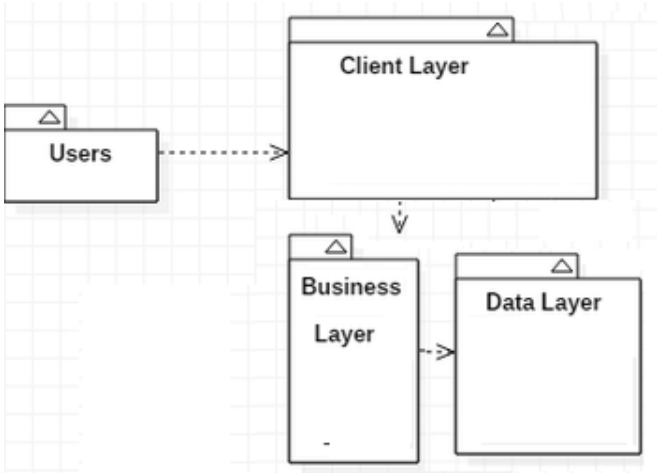


Fig.5.7: Package diagram

PROFILE DIAGRAM

A Profile diagram is any diagram created in a «profile» Package. Profiles provide a means of extending the UML. They are based on additional stereotypes and Tagged Values that are applied to UML elements, connectors and their components.

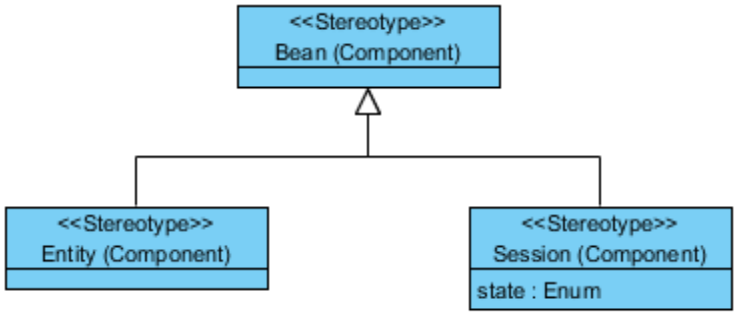


Fig:5.8: Profile diagram

6. PROJECT CODING

6.1. CODE TEMPLATES

TweetReader.py

Import modules:

1. Pandas
2. Numpy
3. Tweepy
4. Traceback

#Declare global variables

Def authentication handler ():

    # Insert OAuth token

Def PushTweetsCSV ():

    #Reads tweets and writes them to tweet.csv

MovieReccomendation.py

#import modules

Import:

1. Tkinter
2. Numpy
3. Mathplotlib
4. Pandas
5. vaderSentiment

#Declare global variables

Def upload ():

    # Import's csv files

Def readDataset ():

    #Reads Dataset from csv files

Def getSentiment():

    # Performs sentiment analysis on Tweets data

Def collaborativeFilter ():

```
    # Performs collabirative filtering on Tweets
Def contentFilter ():
    # Performs Content Based Filtering on Tweets
Def Recommendation ():
#Builds Reccomended Movies to the user

Def graph():
    #Builds graph to showcase how the sentiment is
```

6.2. OUTLINE FOR VARIOUS FILES

We used Python programming to implement our project. A single python file is used to implement our code. This file consists of various modules that we have used. Our project modules are – User Login, Admin Login, Add sentiwords, Recommend Movies, . We also used various python modules like tkinter, matplotlib, numpy.

6.3. METHODS INPUT AND OUTPUT PARAMETERS

In our project code, we implemented nine methods. They are:

1. Authenticaion Handler()
2. pushTweetsCSV()
3. upload()
4. readDataset()
5. getSentiment()
6. collabirativeFiltering()
7. contentFiltering()
8. recommendation()
9. graph()

Our TweetReader.py file script contains two methods Authintication Handler and PushTweetsCSV this methods perform fetching of tweets from twitter and writing them to .csv file.

Our MovieReccomendation.py file script contains all the other methods which build a Movie Recommendation and Graph to showcase the trend of upcoming movies.

7.PROJECT TESTING

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

7.1. VARIOUS TEST CASES

Unit testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

Integration testing

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

Functional test

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals. Functional testing is centered on the following items:

Valid Input : identified classes of valid input must be accepted.

Invalid Input : identified classes of invalid input must be rejected.

Functions : identified functions must be exercised.

Output : identified classes of application outputs must be exercised.

Systems/Procedures : interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

System Test

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

7.2. BLACK BOX

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box .you cannot “see” into it. The test provides inputs and responds to outputs without considering how the software works.

7.3. WHITE BOX TESTING

White Box Testing is a testing in which in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is purpose. It is used to test areas that cannot be reached from a black box level.

8. OUTPUT SCREENS

8.1. USER INTERFACES

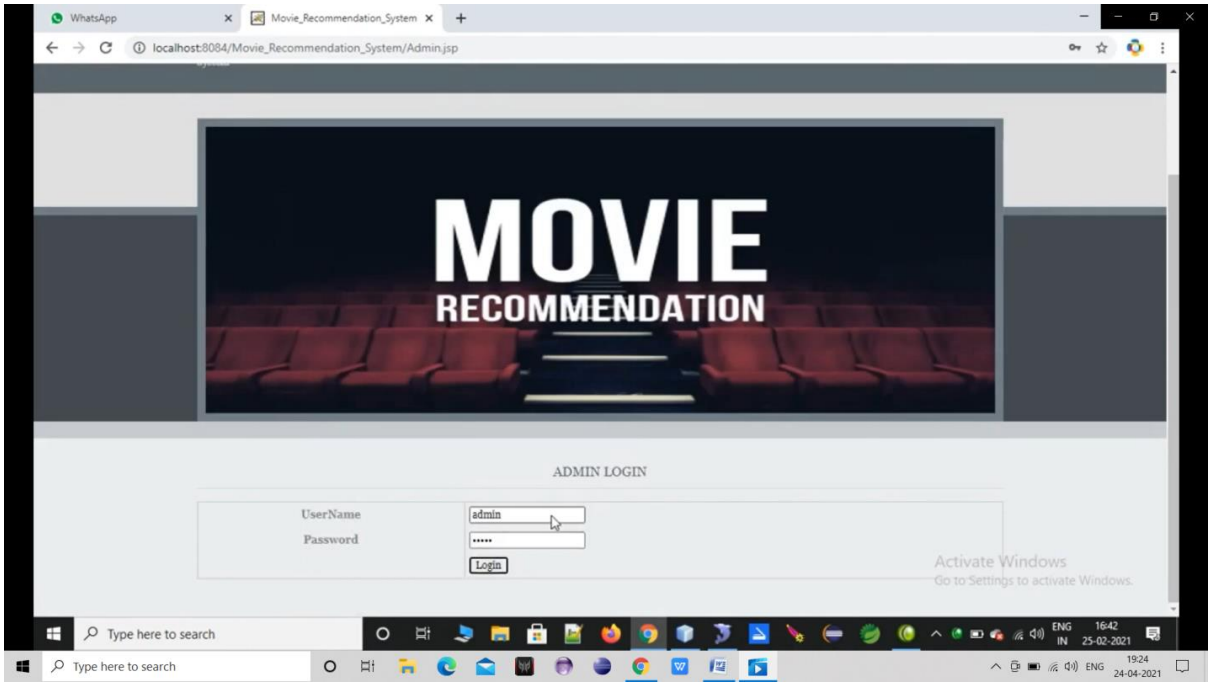


Fig.8.1: Admin login Interface

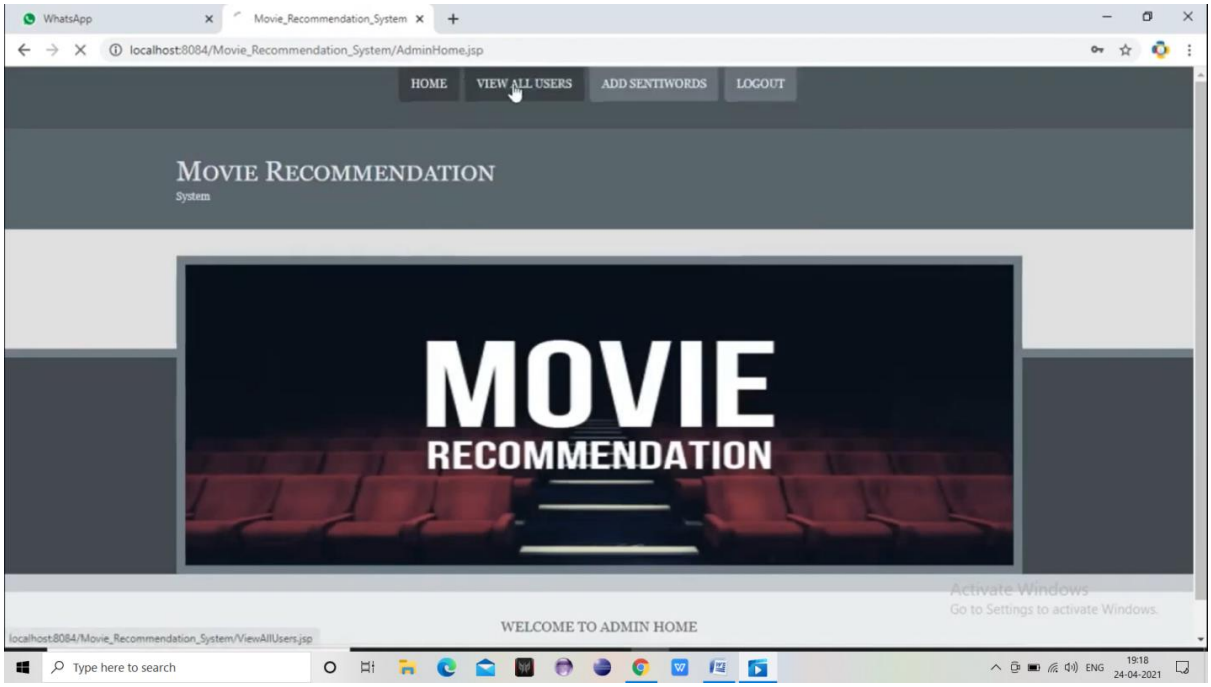


Fig.8.2: Admin page Interface

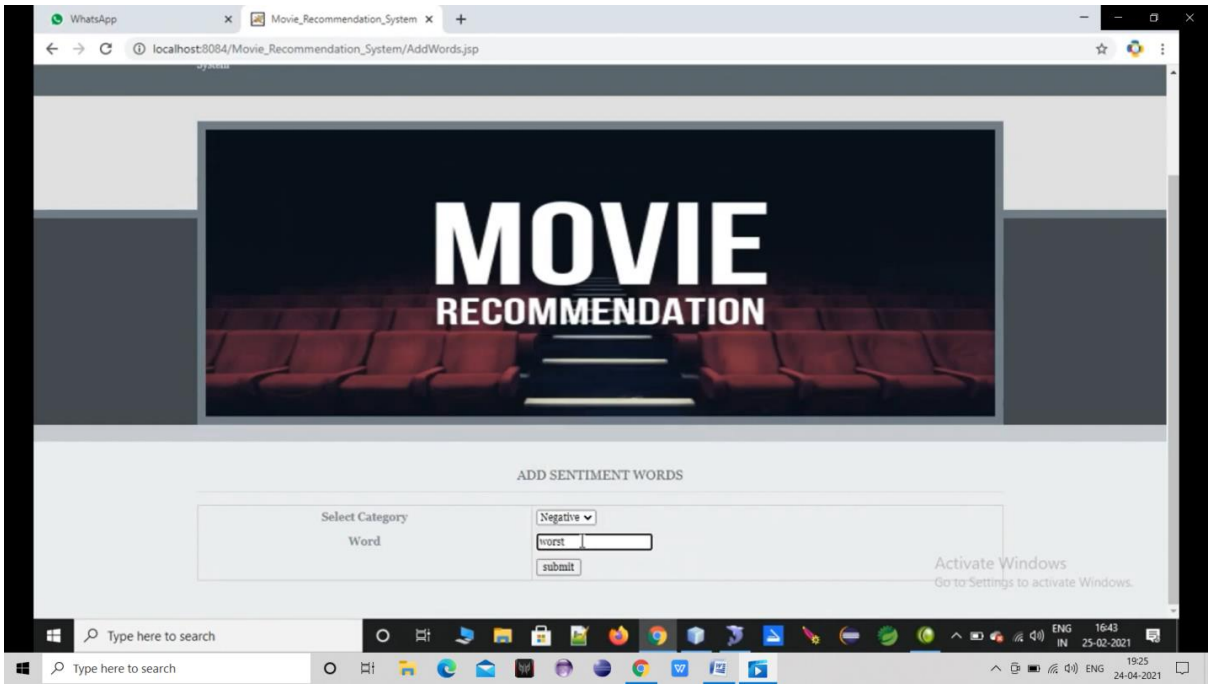


Fig:8.3: Add Senti-Words

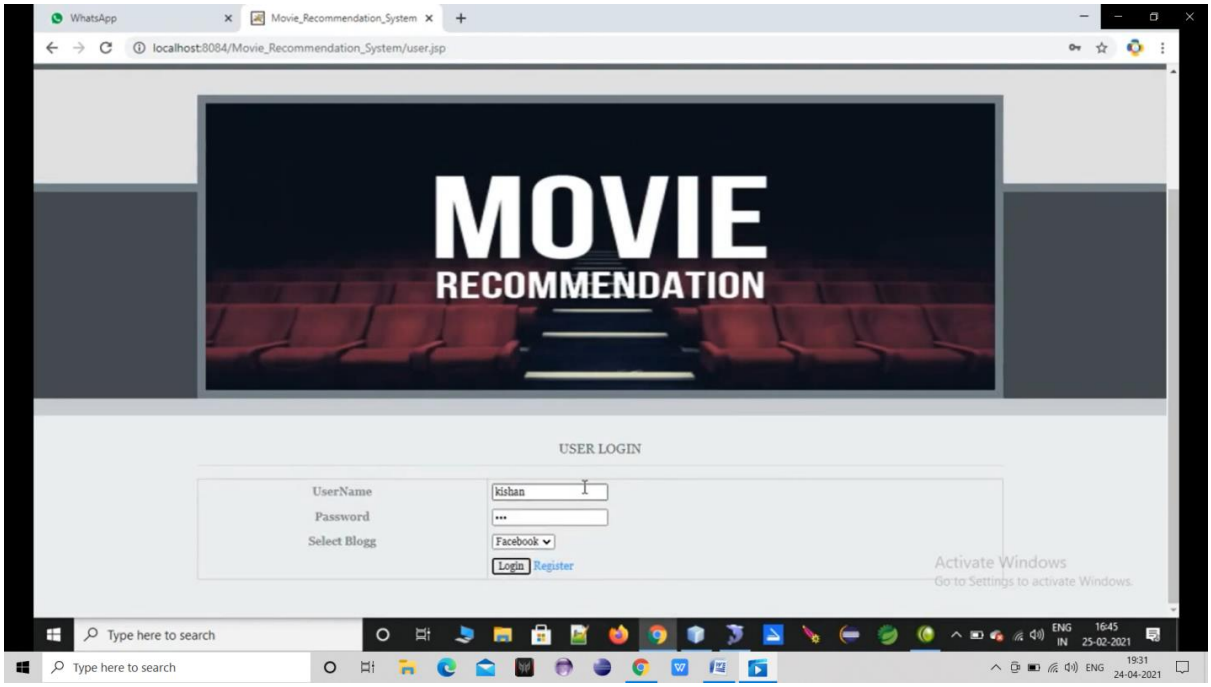


Fig:8.4: User login Interface

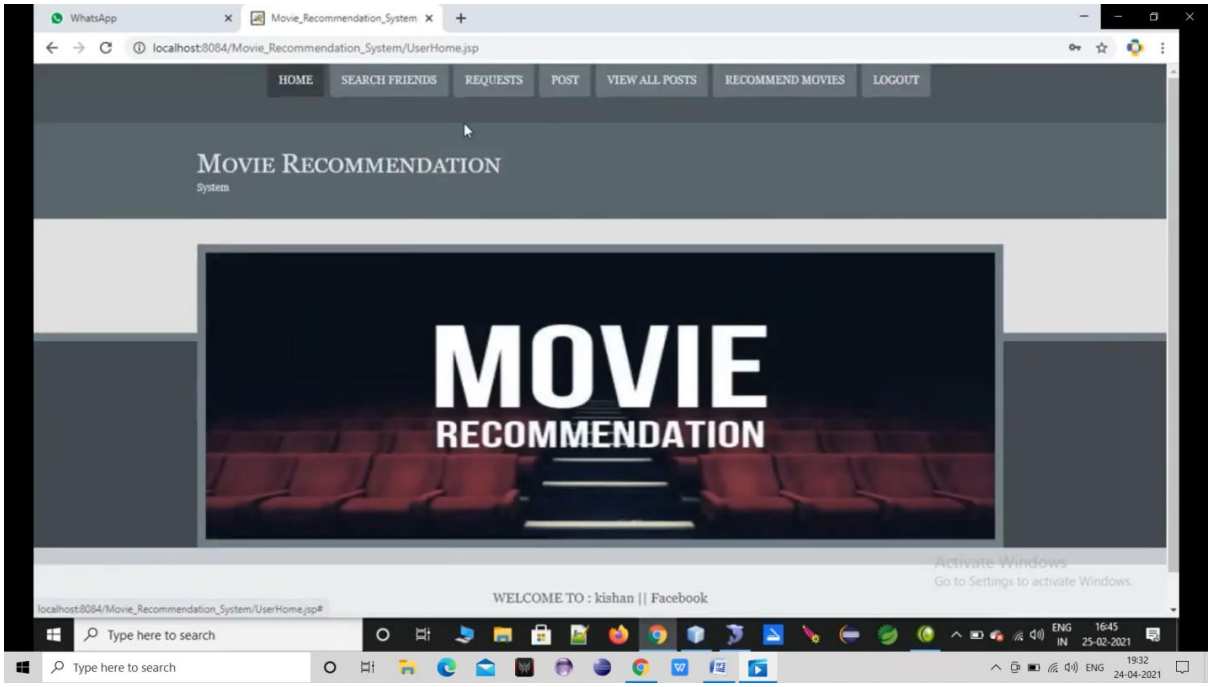


Fig:8.5: User home page Interface

8.2 OUTPUT SCREEN

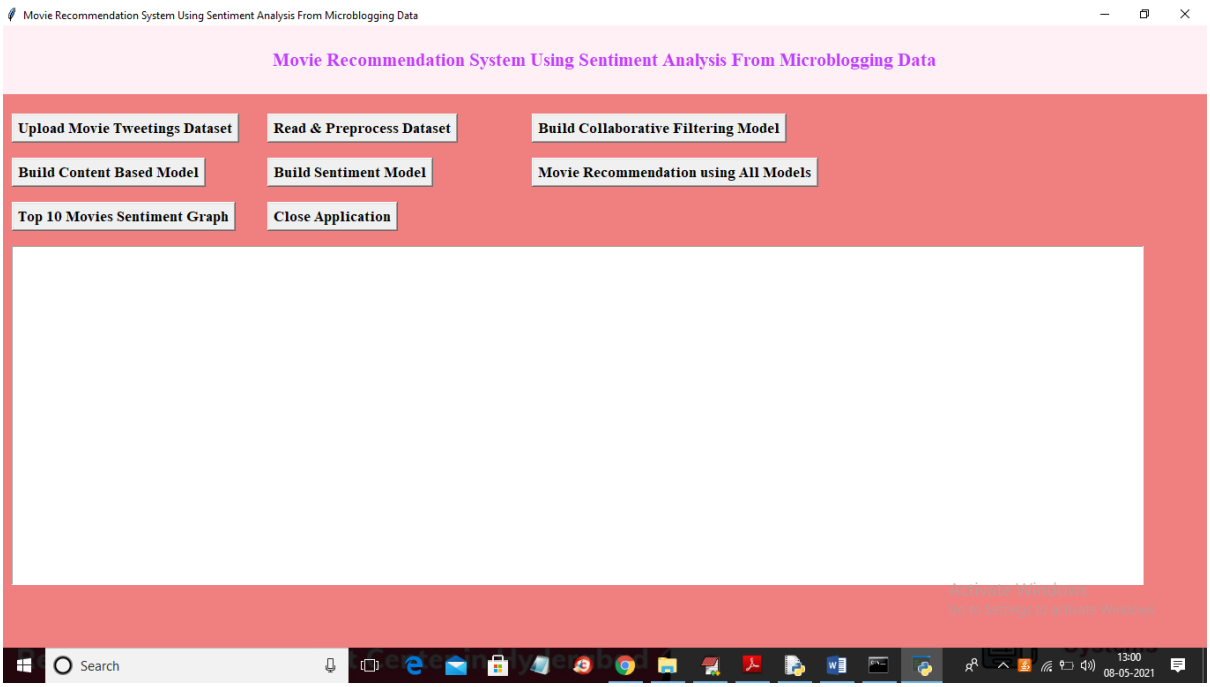


Fig.8.6: Home Screen

Fig 8.6 Shows the when we run the program the above Home Screen will display where there are various options present. In above screen click on 'Upload Movie Tweetings Dataset' button to load dataset.

9.EXPERIMENTAL RESULTS

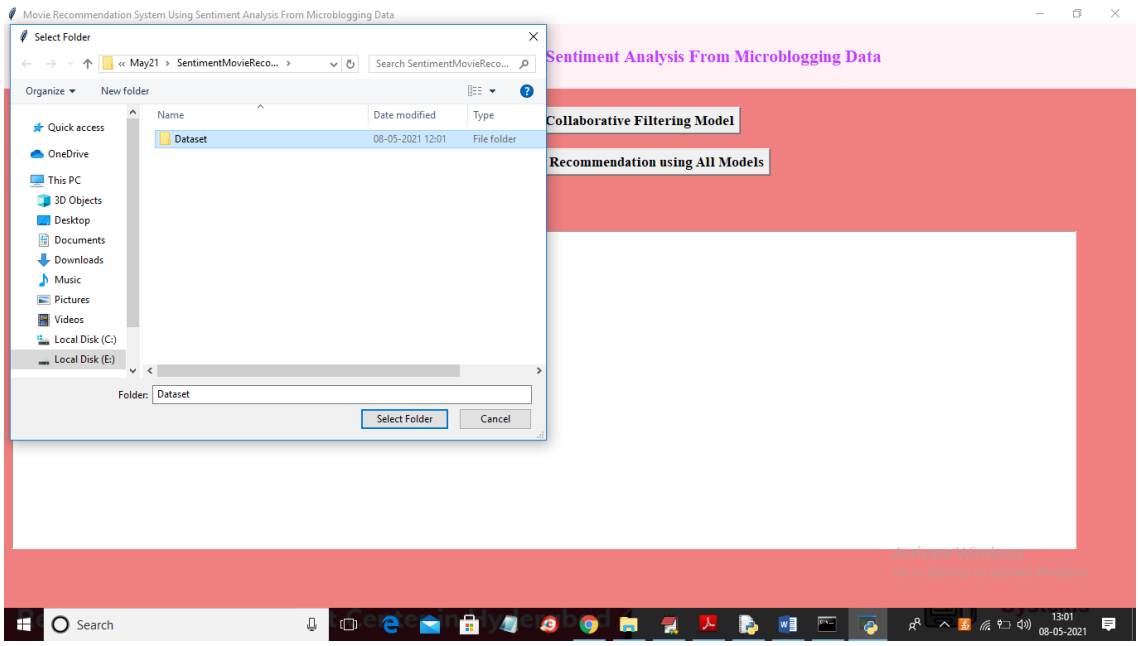


Fig.9.1: Uploading Dataset

Fig 9.1 shows firstly we should upload dataset which contains different genre of movies tweets, ratings. In above screen selecting and uploading entire "Dataset" folder and then click on 'Select Folder" button to load dataset and to get below screen.

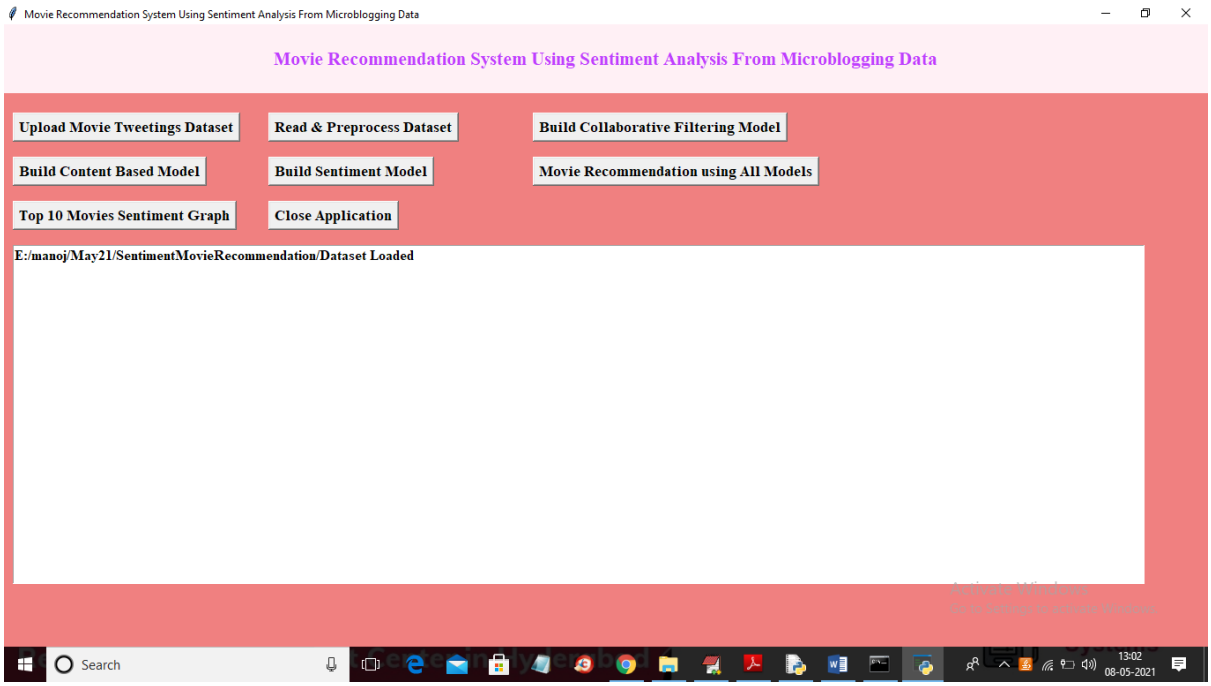


Fig.9.2: Dataset Uploaded

Fig 9.2 shows Dataset Uploaded and In above screen dataset loaded and now click on ‘Read & Preprocess Dataset’ button to read and clean data from special symbols available in movie names.

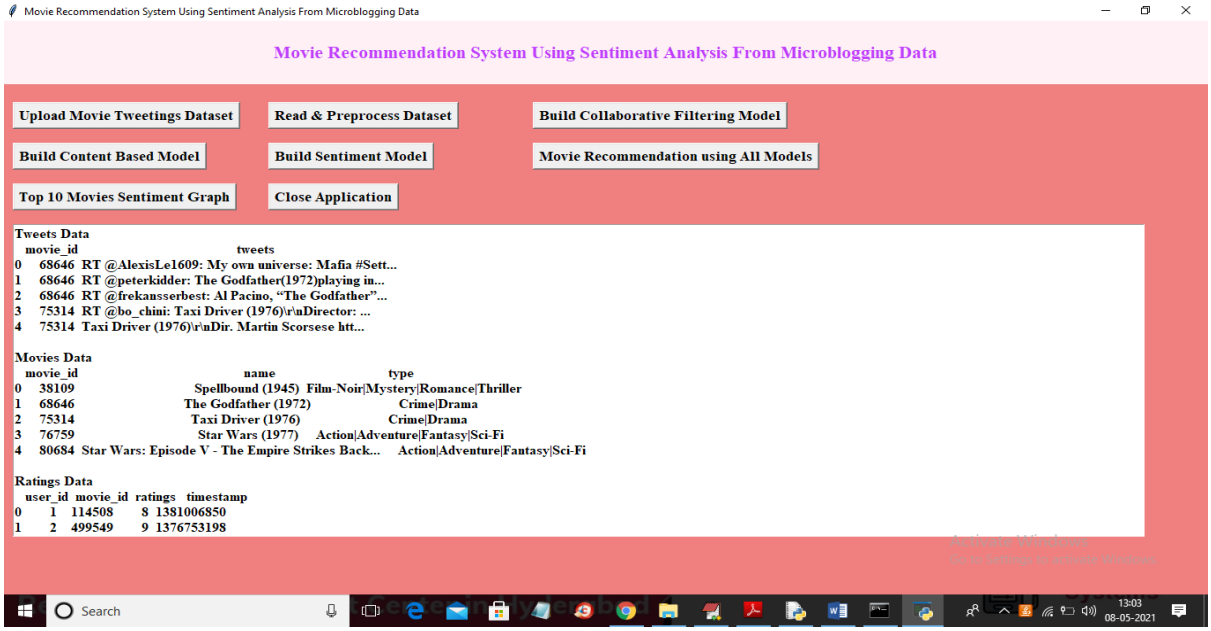


Fig.9.3: Collaborative Filtering

Fig 9.3 shows above screen shows application read data from tweets dataset, movies and rating dataset and now click on ‘Build Collaborative Filtering Model’ button to build model.

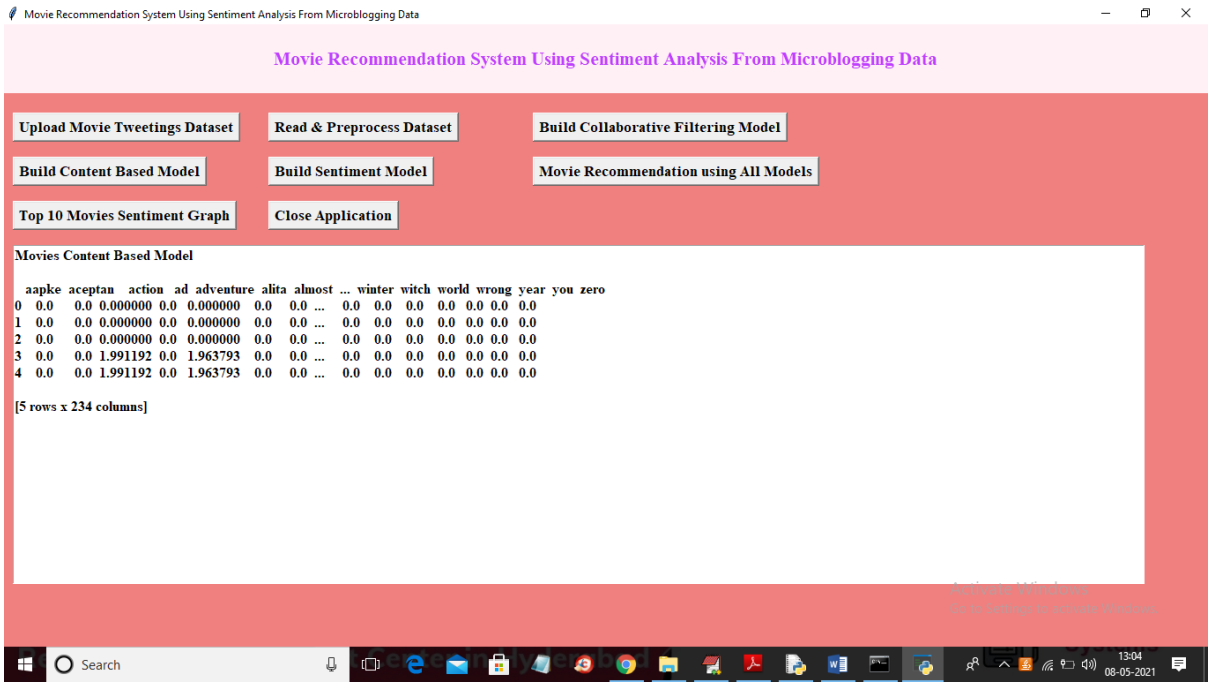


Fig.9.4: Collaborative Matrix

Fig 9.4: shows Collaborative Matrix where in above screen collaborative matrix is created and now click on ‘Build Content Based Model’ button to build content model.

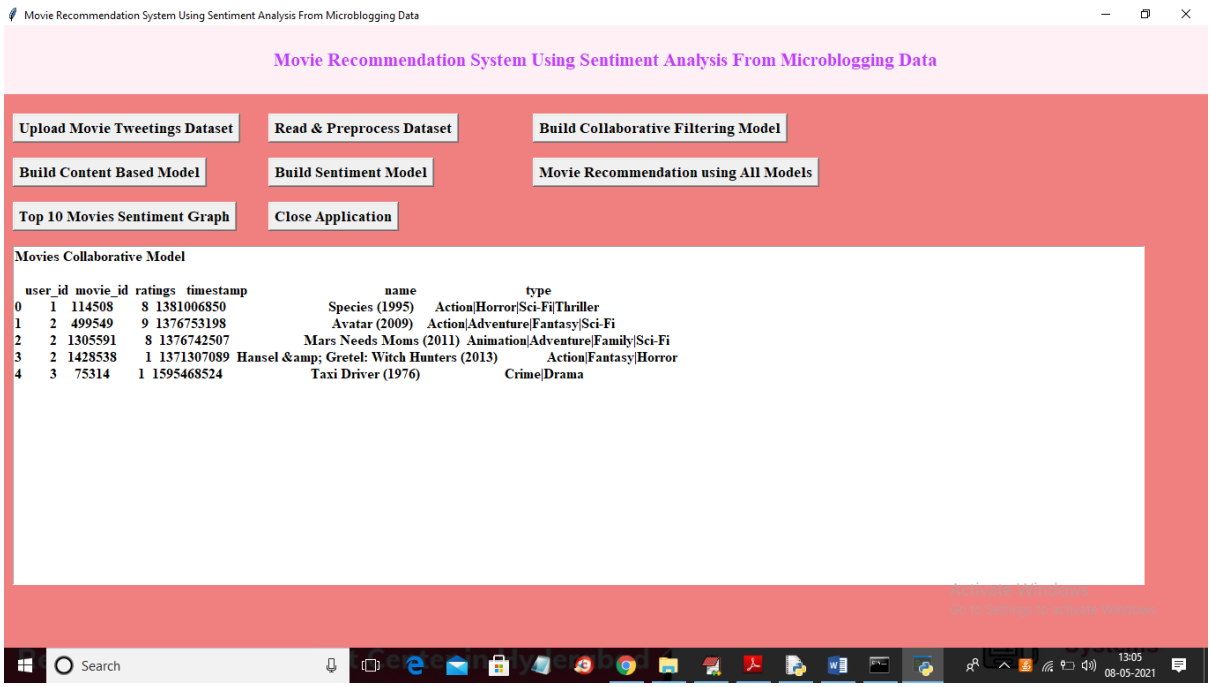


Fig.9.5: Content Based Filtering

Fig 9.5 explains Content Based Filtering where in above screen content matrix model is generated with ratings and movie details and now click on ‘Build Sentiment Model” button to build sentiment object.

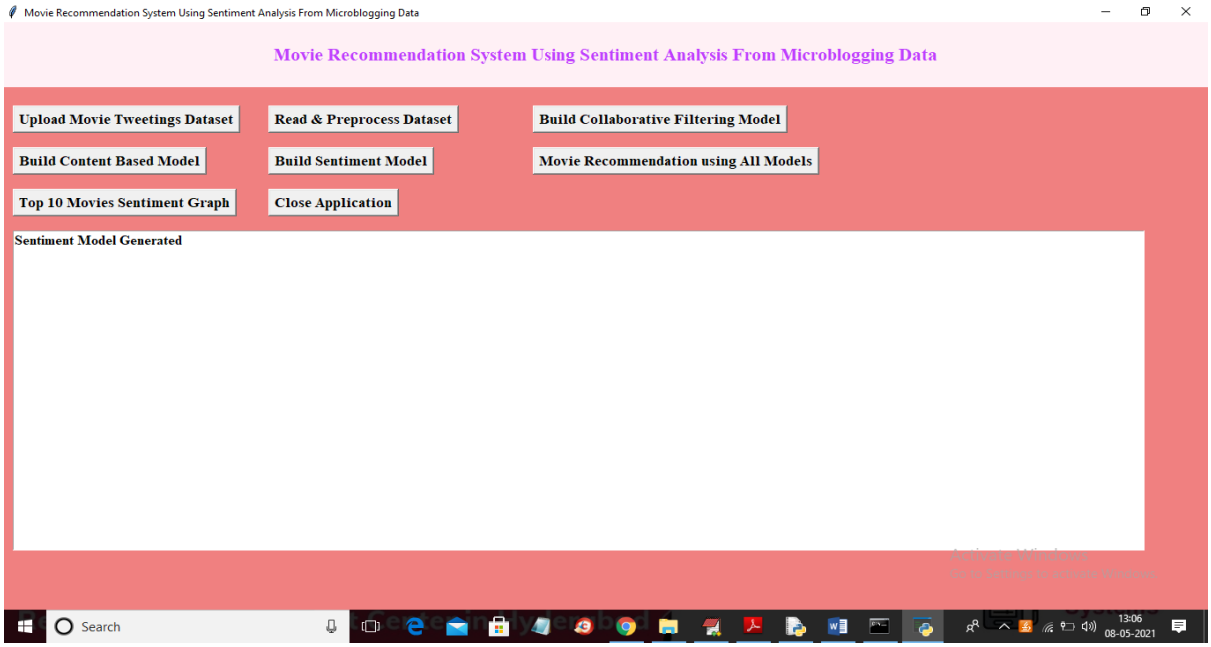


Fig.9.6: Sentiment Model

Fig 9.6 shows in above screen sentiment model generated and now click on 'Movie Recommendation using All Models" button to enter desire movie details and then application will recommend movies using all models.

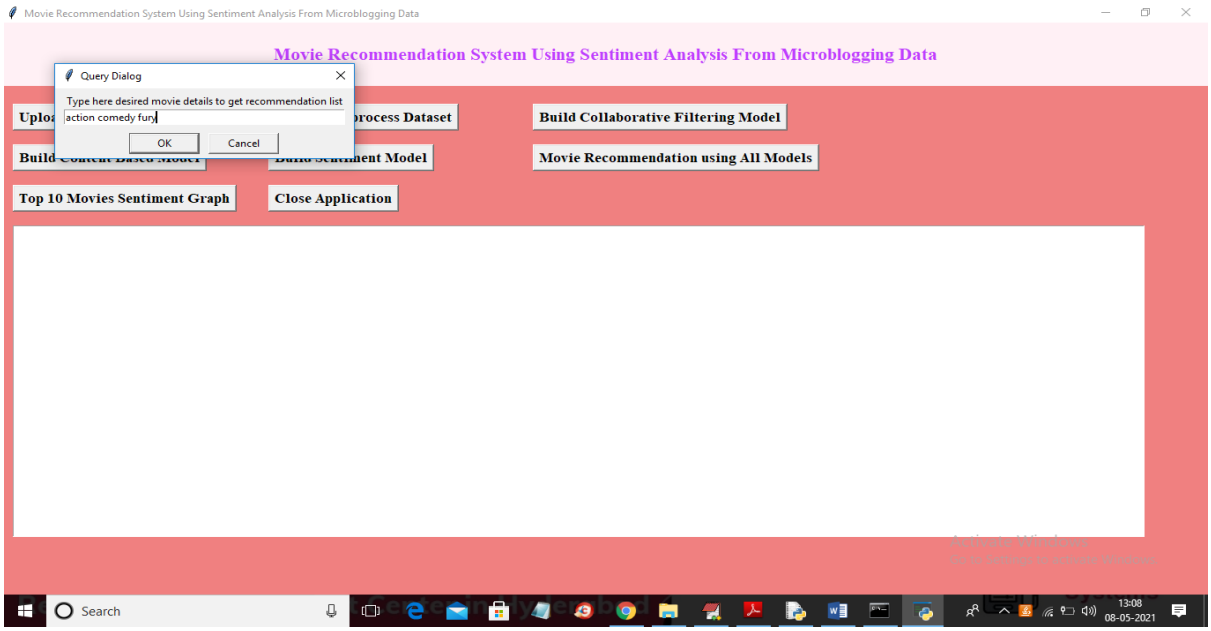


Fig.9.7: Movie Recommendation

Fig 9.7 shows in above screen I entered details like I want movie with action, comedy and title must contains word like fury and if dataset contains actor details then we can give actor names also but this dataset not contains actor or director details. Now click 'OK' button to get below result.

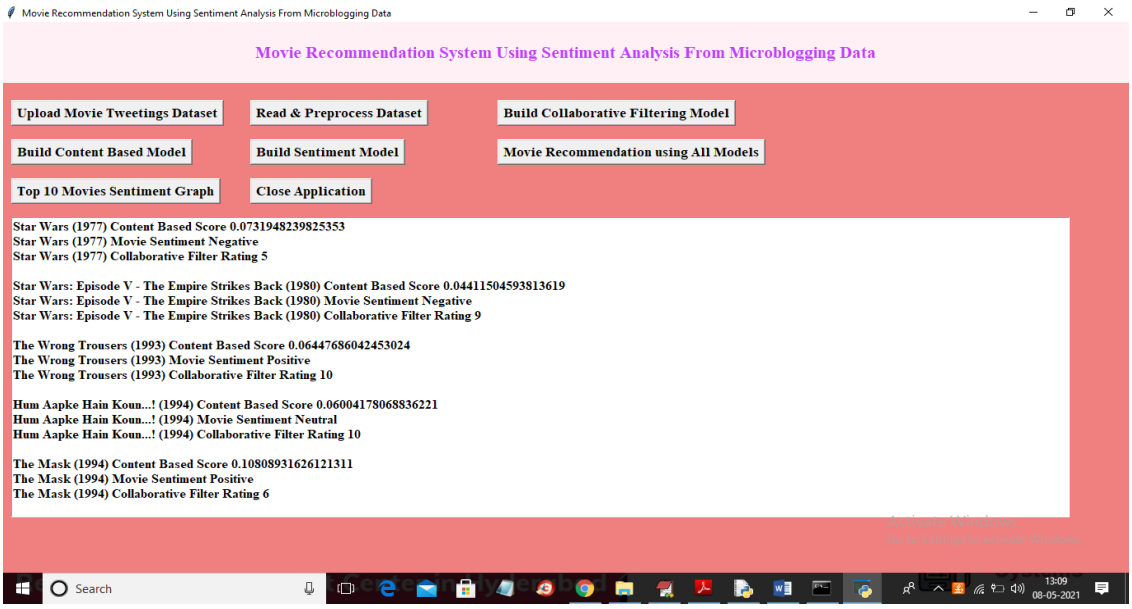


Fig.9.8: Recommending movies

Fig 9.8 shows In above screen we are recommending some movies and for each movie we are calculating Content filter, collaborative filter ratings and sentiment and by seeing this values user may select best movie for himself. Similarly you can give any query on movies then application will suggest top movies based on content, sentiment and collaborative. Now click on 'Top 10 Movie Sentiment Graph' button to get sentiments count for all recommended movies.

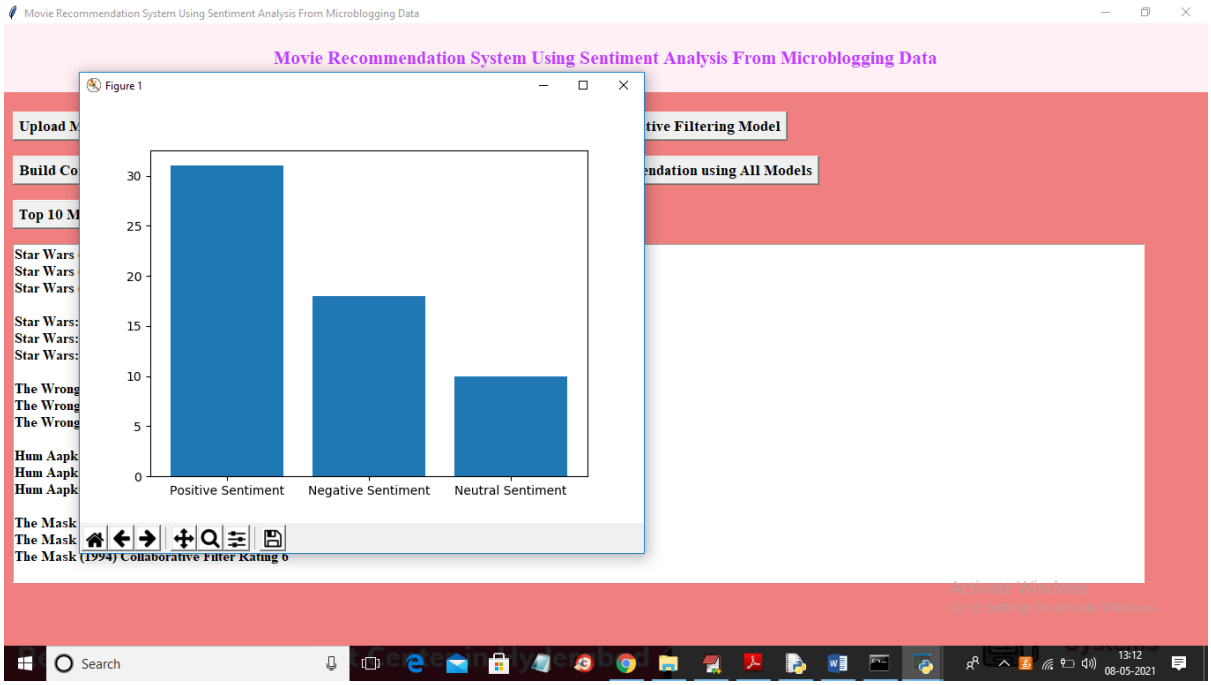


Fig.9.9: Sentiment Graph

Fig 9.9 shows in above graph x-axis represents sentiment type and y-axis represents count of recommended movies and in above graph nearly 30 movies recommended and most of them falls in positive sentiment.

CONCLUSION AND FUTURE ENHANCEMENT

RSs are an important medium of information filtering systems in the modern age, where the enormous amount of data is readily available. In this article, we have proposed a movie RS that uses sentiment analysis data from Twitter, along with movie metadata and a social graph to recommend movies. Sentiment analysis provides information about how the audience is respond to a particular movie and how this information is observed to be useful. The proposed system used weighted score fusion to improve the recommendations.

In the future, we plan to consider more information about the emotional tone of the user from different social media platforms and non-English languages to further improve the RS.

REFERENCES

1. F. Abel, Q. Gao, G.-J. Houben, and K. Tao, “Analyzing user modeling on Twitter for personalized news recommendations,” in Proc. 19th Int. Conf. Modeling, Adaption, Pers. (UMAP). Berlin, Germany: SpringerVerlag, 2011, pp. 1–12.

2.F. Abel, Q. Gao, G.-J. Houben, and K. Tao, “Twitter-based user modeling for news recommendations,” in Proc. Int. Joint Conf. Artif. Intell., vol. 13, 2013, pp. 2962–2966.

3.G. Adomavicius and A. Tuzhilin, “Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions,” IEEE Trans. Knowl. Data Eng., vol. 17, no. 6, pp. 734–749, Jun. 2005.

4.O. Araque, I. Corcuera-Platas, J. F. Sánchez-Rada, and C. A. Iglesias, “Enhancing deep learning sentiment analysis with ensemble techniques in social applications,” Expert Syst. Appl., vol. 77, pp. 236–246, Jul. 2017.

5.E. Aslanian, M. Radmanesh, and M. Jalili, “Hybrid recommender systems based on content feature relationship,” IEEE Trans. Ind. Informat., early access, Nov. 21, 2016, doi: 10.1109/TII.2016.2631138.

6.J. Bobadilla, F. Ortega, A. Hernando, and J. Alcalá, “Improving collaborative filtering recommender system results and performance using genetic algorithms,” Knowl.-Based Syst., vol. 24, no. 8, pp. 1310–1316, Dec. 2011.

7.R. Burke, “Hybrid recommender systems: Survey and experiments,” User Model. User-Adapted Interact., vol. 12, no. 4, pp. 331–370, 2002.

8. E. Cambria, “Affective computing and sentiment analysis,” IEEE Intell. Syst., vol. 31, no. 2, pp. 102–107, Mar./Apr. 2016.

9. I. Cantador, A. Bellogín, and D. Vallet, “Content-based recommendation in social tagging systems,” in Proc. 4th ACM Conf. Rec. Syst. (RecSys), 2010, pp. 237–240.

10.P. Cremonesi, Y. Koren, and R. Turrin, “Performance of recommender algorithms on top-N recommendation tasks,” in Proc. 4th ACM Conf. Rec. Syst. (RecSys), 2010, pp. 39–46.

PUBLICATION

MOVIE RECOMMENDATION SYSTEM USING SENTIMENT ANALYSIS FROM MICROBLOGGING DATA

Sai Chaitanya.R[1] , D.Varshitha[2], D.Charmitha[3], G.Vinay[4], V.L.Kartheek[5]
[1,2,3,4]UG Scholar, [5]Assistant Professor
Department of Computer Science & Engineering
St. Martin's Engineering College,
Near Forest Academy, Dhulapally, Secunderabad, Telangana, India-500014
Email-id: chaitanya.raparti@gmail.com[1], donthinenivarshitha@gmail.com[2], charmithar@gmail.com[3], vinayprasadgolkonda@gmail.com[4], kartheekv999@gmail.com[5]

ABSTRACT:

Recommendation systems (RSs) have garnered immense interest for applications in e-commerce and digital media. Traditional approaches in RSs include such as collaborative filtering (CF) and content-based filtering (CBF) through these approaches that have certain limitations, such as the necessity of prior user history and habits for performing the task of recommendation. To minimize the effect of such limitation, this article proposes a hybrid RS for the movies that leverage the best of concepts used from CF and CBF along with sentiment analysis of tweets from microblogging sites. The purpose to use movie tweets is to understand the current trends, public sentiment, and user response of the movie. Experiments conducted on the public database have yielded promising results.

Keywords: · Collaborative filtering , Content based filtering , Recommendation System, Sentiment Analysis , Twitter

I. INTRODUCTION:

Traditional approaches in RSs include such as collaborative filtering (CF) and content-based filtering (CBF) through these approaches that have certain limitations, such as the necessity of prior user history and habits for performing the task of recommendation. Users often face the problem of excessive available information. Recommendation systems (RSs) are deployed to help users cope up with the information explosion. RS is mostly used in digital entertainment, such as Netflix, Prime Video, and IMDB, and e-commerce portals such as Amazon, Flipkart, and eBay. In this article, we focus on RS for movies, which is an important source of recreation and entertainment in our life. Movie suggestions for users depend on Web-based portals.

Movies can be easily differentiated through their genres, such as comedy, thriller, animation, and action. Another possible way to categorize the movies based on its metadata, such as release year, language, director, or cast. Most online video-streaming services , provide personalized user experience by utilizing the user's historical data, such as previously viewed or rated history.The purpose to use movie tweets is to understand the current trends, public sentiment, and user response of the movie. Experiments conducted on the public database have yielded promising results.

II. LITERATURE SURVEY:

How can micro-blogging activities on Twitter be leveraged for user modeling and personalization? In this paper we investigate this question and introduce a framework for user modeling on Twitter which enriches the semantics of Twitter messages (tweets) and identifies topics and entities (e.g. persons, events, products) mentioned in tweets. We analyze how strategies for constructing hashtag-based, entity-based or topic-based user profiles benefit from semantic enrichment and explore the temporal dynamics of those profiles. We further measure and compare the performance of the user modeling strategies in context of a personalized news recommendation system. Our results reveal how semantic enrichment enhances the variety and quality of the generated user profiles. Further, we see how the different user modeling strategies impact personalization and discover that the consideration of temporal profile patterns can improve recommendation quality[1].

How can micro-blogging activities on Twitter be leveraged for user modeling and personalization? In this paper we investigate this question and introduce a framework for user modeling on Twitter which enriches the semantics of Twitter messages (tweets) and identifies topics and entities (e.g. persons, events, products) mentioned in tweets. We analyze how strategies for constructing hashtag-based, entity-based or topic-based user profiles benefit from semantic enrichment and explore the temporal dynamics of those profiles. We further measure and compare the performance of the user modeling strategies in context of a personalized news recommendation system. Our results reveal how semantic enrichment enhances the variety and quality of the generated user profiles. Further, we see how the different user modeling strategies impact personalization and discover that the consideration of temporal profile patterns can improve recommendation quality[2].

This paper presents an overview of the field of recommender systems and describes the current generation of recommendation methods that are usually classified into the

following three main categories: content-based, collaborative, and hybrid recommendation approaches. This paper also describes various limitations of current recommendation methods and discusses possible extensions that can improve recommendation capabilities and make recommender systems applicable to an even broader range of applications. These extensions include, among others, an improvement of understanding of users and items, incorporation of the contextual information into the recommendation process, support for multicriteria ratings, and a provision of more flexible and less intrusive types of recommendations [3].

Deep learning techniques for Sentiment Analysis have become very popular. They provide automatic feature extraction and both richer representation capabilities and better performance than traditional feature based techniques (i.e., surface methods). Traditional surface approaches are based on complex manually extracted features, and this extraction process is a fundamental question in feature driven methods. These long-established approaches can yield strong baselines, and their predictive capabilities can be used in conjunction with the arising deep learning methods. In this paper we seek to improve the performance of deep learning techniques integrating them with traditional surface approaches based on manually extracted features. The contributions of this paper are sixfold. First, we develop a deep learning based sentiment classifier using a word embeddings model and a linear machine learning algorithm. This classifier serves as a baseline to compare to subsequent results. Second, we propose two ensemble techniques which aggregate our baseline classifier with other surface classifiers widely used in Sentiment Analysis. Third, we also propose two models for combining both surface and deep features to merge information from several sources. Fourth, we introduce a taxonomy for classifying the different models found in the literature, as well as the ones we propose. Fifth, we conduct several experiments to compare the performance of these models with the deep learning baseline. For this, we use seven public datasets that were extracted from the microblogging and movie reviews domain. Finally, as a result, a statistical study confirms that the performance of these proposed models surpasses that of our original baseline on F1-Score[4].

Recommendation systems get ever-increasing importance due to their applications in both academia and industry. The most popular type of these systems, known as collaborative filtering algorithms, employ user-item interactions to perform the recommendation tasks. With growth of additional information sources other than the rating (or purchase) history of users on items, such as item descriptions and social media information, further extensions of these systems have been proposed, known as hybrid recommendation algorithms[5]. Hybrid

recommenders use both user-item interaction data and their contextual information. In this work, we propose new hybrid recommender algorithms by considering the relationship between content features. This relationship is embedded into the hybrid recommenders to improve their accuracy. We first introduce a novel method to extract the content feature relationship matrix, and then the collaborative filtering recommender is modified such that this relationship matrix can be effectively integrated within the algorithm. The proposed algorithm can better deal with the cold-start problem than the state-of-art algorithms. We also propose a novel content-based hybrid recommender system. Our experiments on a benchmark movie dataset show that the proposed approach significantly improves theaccuracy of the system, while resulting in satisfactory performance in terms of novelty and diversity of the recommendation lists[5].

III. PROPOSED METHODOLOGY:

The proposed sentiment-based RS is shown in Fig. 1. In this section, we describe various components of the proposed RS. A. Data Set Description The proposed system needs two types of databases. One is a user-rated movie database, where ratings for relevant movies are present, and another is the user tweets from Twitter.

Public Databases:

There are many popular public databases available, which have been widely used to recommend the movies and other entertainment media. To incorporate the sentiment analysis in the proposed framework, the tweets of movies were extracted from Twitter against the movies that were available in the database. Experiments conducted using various public databases, such as the Movielens 100K,2 Movielens 20M,3 Internet Movie Database (IMDb,4) and Netflix database,5 that were not found suitable for our work due to the absence of microblogging data. After a thorough assessment of the abovementioned databases, the MovieTweetings database [12] was finally selected for the proposed system. MovieTweetings is widely considered as a modern version of the MovieLens database. The purpose of this database is to provide an up-to-date movie rating so that it contains more realistic data for sentiment analysis. Table I displays the relevant details of the MovieTweetings database.

Modified MovieTweetings Database

In the proposed work, the MovieTweetings database is modified to implement the RS. The primary objective to modify the database was to use sentiment analysis of tweets by the users, in the prediction of the movie RS. The MovieTweetings database contains the movies with published years from 1894 to 2017. Due to the scarcity of tweets for old movies, we only

considered the movies that were released in or after the year 2014 and extracted a subset of the database which complied with our objective.

IV. RESULT AND DISCUSSION:

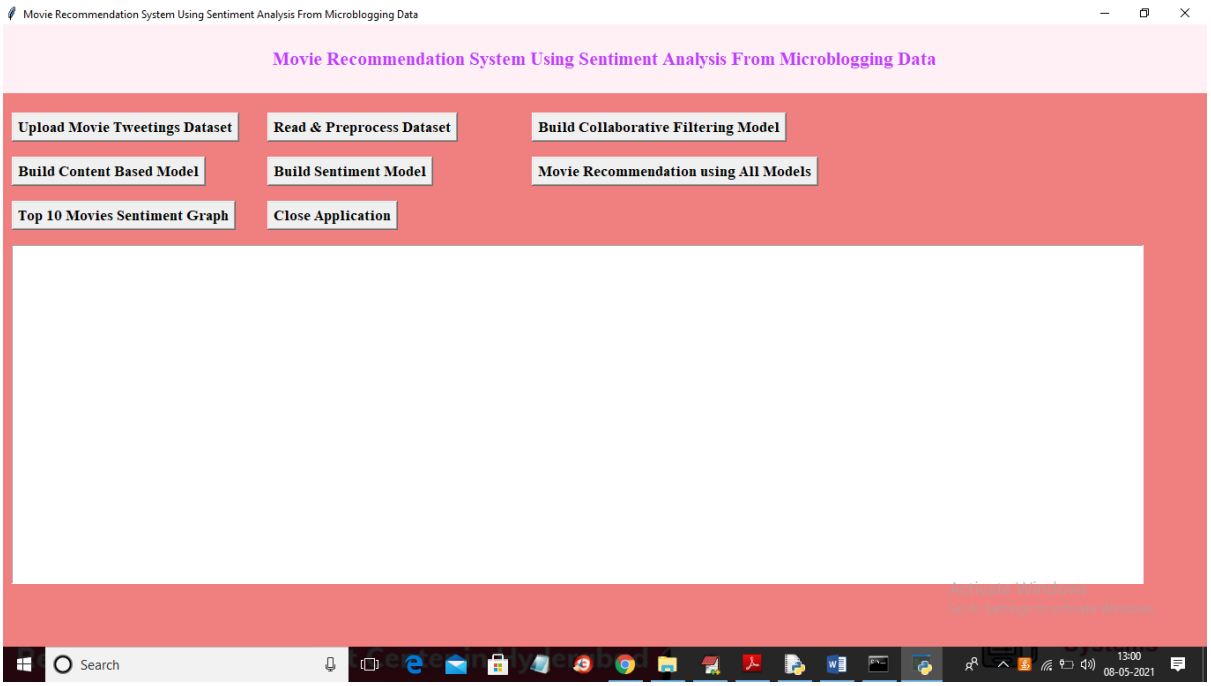


Fig 1: Home Screen

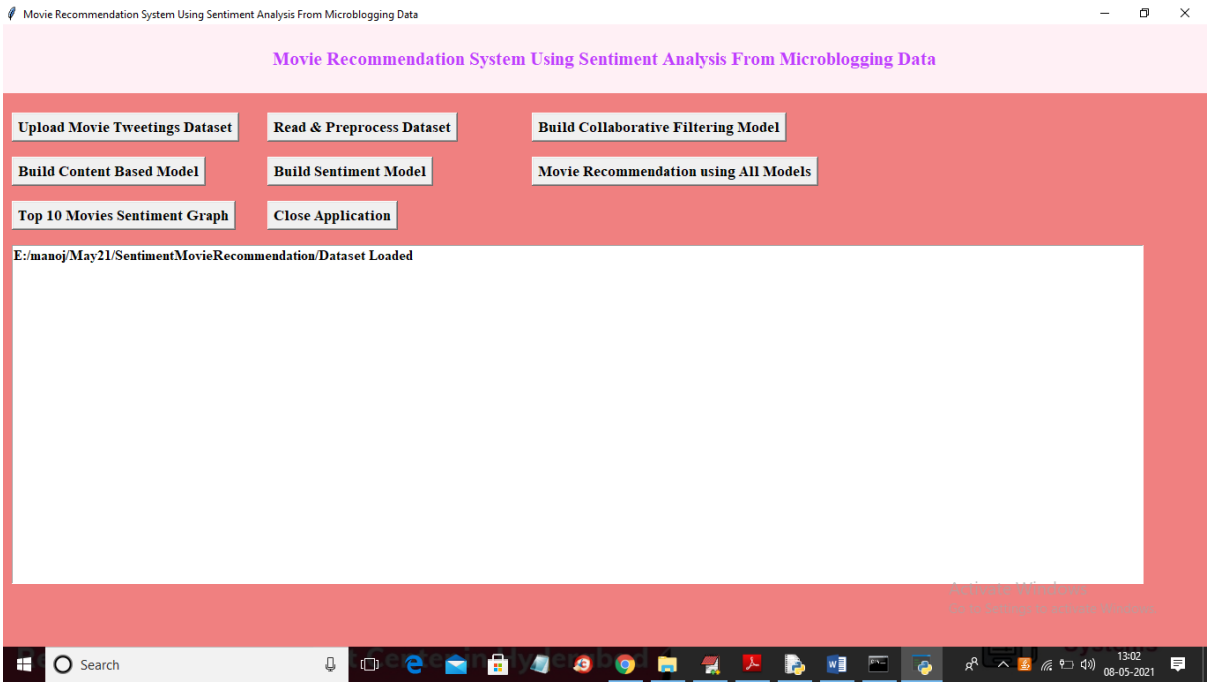


Fig 2: Dataset Uploaded

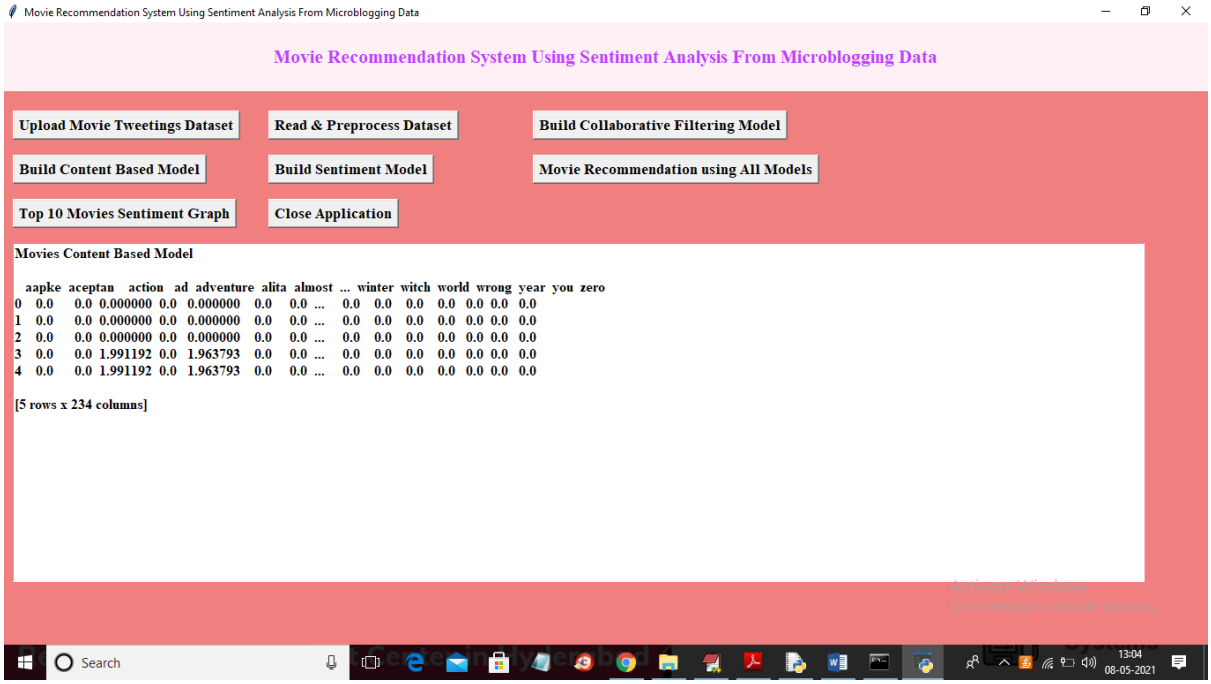


Fig 3: Collaborative Filtering

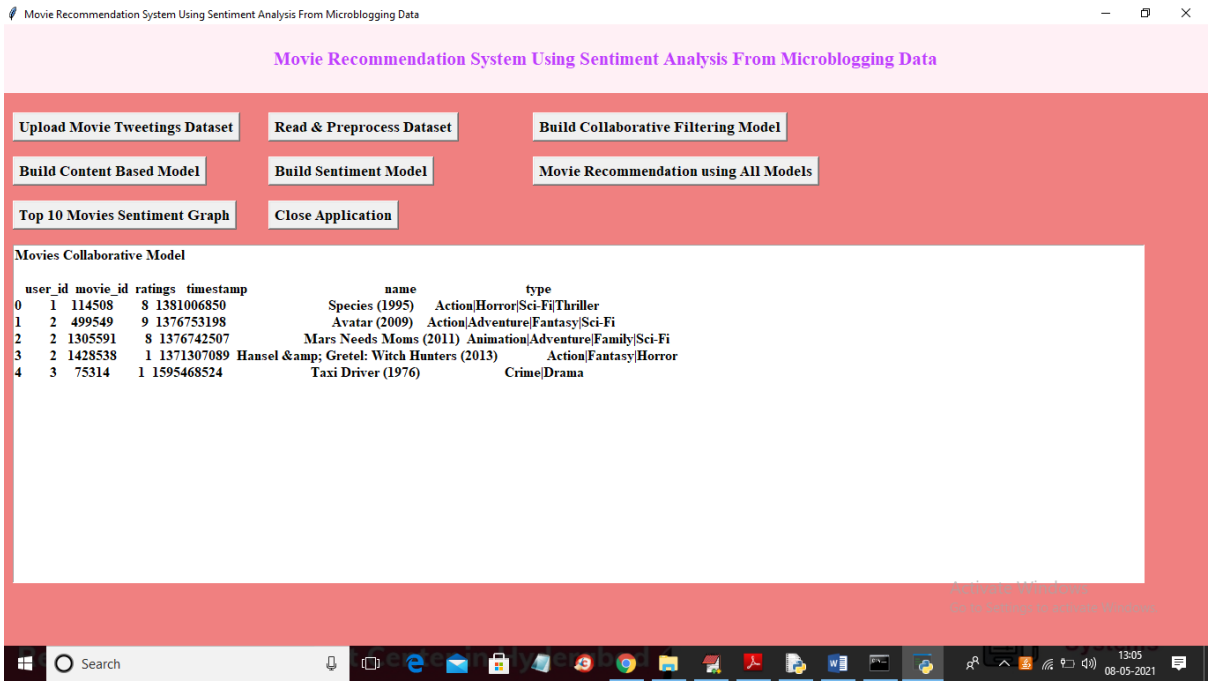


Fig 4: Content Based Filtering

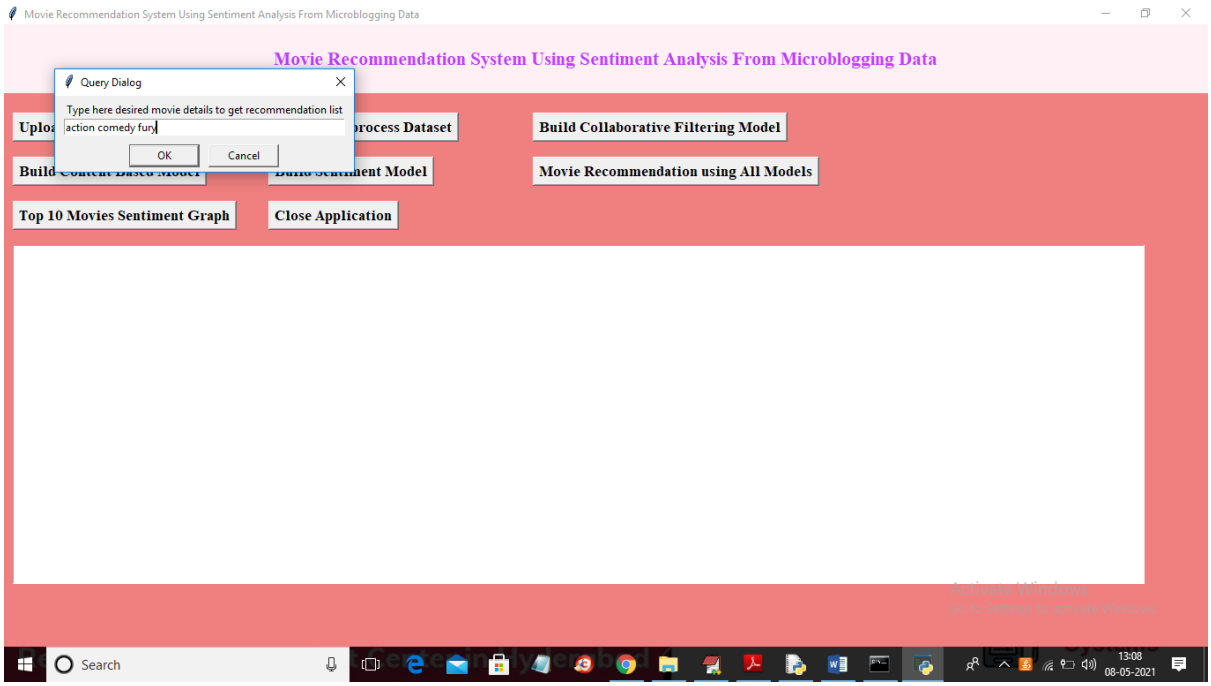


Fig 5: Movie Recommendation

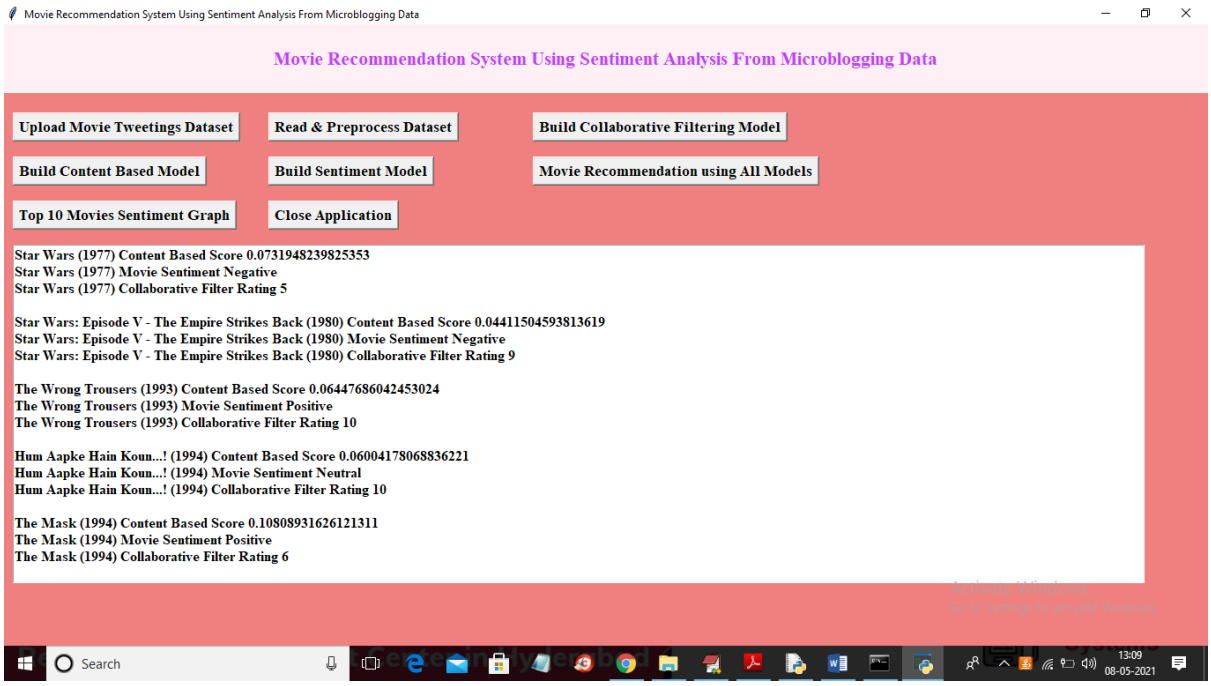


Fig 6: Recommending Movies

V. CONCLUSION:

Recommendation Systems are an important medium of information filtering systems in the modern age, where the enormous amount of data is readily available. In this article, we have proposed a movie Recommendation System that uses sentiment analysis data from Twitter, along with movie metadata and a social graph to recommend movies. Sentiment analysis provides information about how the audience is respond to a particular movie and how this information is observed to be useful. The proposed system used weighted score fusion to improve the recommendations.

VI. REFERENCES:

1. F. Abel, Q. Gao, G.-J. Houben, and K. Tao, "Analyzing user modeling on Twitter for personalized news recommendations," in Proc. 19th Int. Conf. Modeling, Adaption, Pers. (UMAP). Berlin, Germany: SpringerVerlag, 2011, pp. 1–12.

2.F. Abel, Q. Gao, G.-J. Houben, and K. Tao, "Twitter-based user modeling for news recommendations," in Proc. Int. Joint Conf. Artif. Intell., vol. 13, 2013, pp. 2962–2966.

3.G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," IEEE Trans. Knowl. Data Eng., vol. 17, no. 6, pp. 734–749, Jun. 2005.

4.O. Araque, I. Corcuera-Platas, J. F. Sánchez-Rada, and C. A. Iglesias, "Enhancing deep learning sentiment analysis with ensemble techniques in social applications," Expert Syst. Appl., vol. 77, pp. 236–246, Jul. 2017.

5.E. Aslanian, M. Radmanesh, and M. Jalili, "Hybrid recommender systems based on content feature relationship," IEEE Trans. Ind. Informat., early access, Nov. 21, 2016, doi: 10.1109/TII.2016.2631138.

6.J. Bobadilla, F. Ortega, A. Hernando, and J. Alcalá, "Improving collaborative filtering recommender system results and performance using genetic algorithms," Knowl.-Based Syst., vol. 24, no. 8, pp. 1310–1316, Dec. 2011.

7.R. Burke, "Hybrid recommender systems: Survey and experiments," User Model. User-Adapted Interact., vol. 12, no. 4, pp. 331–370, 2002.

8. E. Cambria, “Affective computing and sentiment analysis,” IEEE Intell. Syst., vol. 31, no. 2, pp. 102–107, Mar./Apr. 2016.

9. I. Cantador, A. Bellogín, and D. Vallet, “Content-based recommendation in social tagging systems,” in Proc. 4th ACM Conf. Rec. Syst. (RecSys), 2010, pp. 237–240.

10.P. Cremonesi, Y. Koren, and R. Turrin, “Performance of recommender algorithms on top-N recommendation tasks,” in Proc. 4th ACM Conf. Rec. Syst. (RecSys), 2010, pp. 39–46.

ONE PAGE PROFILE

1. D. VARSHITHA

D. Varshitha is currently pursuing her Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. She completed her intermediate from Page Junior College and 10th class from St. Michael's School. Her technical skills include C, Python and Java. She also has a basic understanding of C++. Her participations include: Workshop on "Ethical Hacking" which was conducted by college on 31st January to 1st February 2020, National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020, Women online workshop on "Women in Cyber Security and Privacy in 2020" a five day workshop which was conducted from 6th to 10th July 2020, the Global Webinar on Cyber Threats and Defense Techniques conducted by GECF on 22nd July 2020. Apart from that she also attended few online sessions conducted by Two-Day National Level Seminar On "Recent Trends in Cloud Computing, Fog and Edge Computing" scheduled on 18th June to 19th June 2021. She spends her free time taking online certification courses related to her field of study as well as personal interests from platform such as Coursera and CursaApp. Her areas of interests are Python, Machine learning and Deep learning.

2. G. VINAY PRASAD

G. Vinay Prasad is currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. He completed his intermediate from FIITJEE Junior College and 10th class from Delhi Public School. His technical skills include Java , Python. He also has a basic understanding of C. He is one of the student of Smart Interviews and participated in few tests conducted by them. His participations include: National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020,and he has completed few courses on coursera in the year 2020 and have also participated in Two-Day National Level Seminar On "Recent Trends in Cloud Computing, Fog and Edge Computing" scheduled on 18th June to 19th June 2021 April to 22nd May 2020. His areas of interest are Python, Java ,Cloud Computing, Cyber Security, Machine Learning . He completed few certification courses from online platforms like Coursera, Data Camp.

3. R. SAI CHAITANYA

R. Sai Chaitanya is currently pursuing her Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. He is also working as Software Engineer Trainee for Medplus. He completed his intermediate from Sri Chaitanya Junior College and 10th Standard from Secunderabad Public School. He has done courses on National Programme on Technology Enhanced Learning on Programming, Data Structures and Algorithms using Python and Problem solving through Programming in C. He actively takes part in online Competitive Programming on platforms like Codechef, Codeforces. He is Full Stack Java Developer containing development experience in Back-end: Java, Spring Framework and Front-end: ReactJs Library, Javascript. He is confident in Data Structures and Algorithms in C++. He has know how in Python, C++, C, Java, Javascript and MySQL Database. He has projects on Employee Detail and Payroll System using Java Spring Framework and Spotify Playlist generator using Python. He in future wants to study Astrophysics and Space Science.

4. D. CHARMITHA RAO

D. Charmitha Rao is currently pursuing her Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. She completed her intermediate from Sri Chaitanya Junior College and 10th class from Vijay High School. Her technical skills include C,C++, Python and Java. She also has a basic understanding of C. She is one of the student of Smart Interviews and participated in few tests conducted by them. Her participations include: National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020, "Know More - Teach More Women online workshop on "Women in Cyber Security and Privacy in 2020" which was conducted from 6th to 10th July 2020, "Know More - Teach More ",and she has completed few courses on coursera in the year 2020 and have also participated in Two-Day National Level Seminar On "Recent Trends in Cloud Computing, Fog and Edge Computing" scheduled on 18th June to 19th June 2021 April to 22nd May 2020. Her areas of interest are Python, Artificial Intelligence, Machine Learning and Deep Learning. She completed few certification courses from online platforms like Coursera, CursaApp and SoloLearn.

APENDENCIES

```
from tkinter import *

import tkinter
from tkinter import filedialog
import numpy as np
from tkinter.filedialog import askopenfilename
from tkinter import simpledialog
import matplotlib.pyplot as plt
import pandas as pd
import re
from sklearn.feature_extraction.text import TfidfVectorizer
from numpy import dot
from numpy.linalg import norm
from vaderSentiment.vaderSentiment import
    SentimentIntensityAnalyzer

main = tkinter.Tk()
main.title("Movie Recommendation System Using Sentiment
    Analysis From Microblogging Data")
main.geometry("1000x650")

global filename
global movies_df, ratings_df, tweets_df, users_df
textArray = []
movieArray = []
movieNames = []
global tfidf_vectorizer
global X
global sid
global pos
global neg
global neu

def getSentiment(movie,tweets_df):
   global sid
   result = "Unable to detect sentiment"
   tweet_data = ''
```

```
    for i in range(len(tweets_df)):
        mid = tweets_df[i,0]
        tweet = tweets_df[i,1]
        if mid == movie:
            tweet = re.sub('[^A-Za-z]+', ' ', tweet)
            tweet_data+=tweet+" "
    sentiment_dict = sid.polarity_scores(tweet_data.strip())
    negative = sentiment_dict['neg']
    positive = sentiment_dict['pos']
    neutral = sentiment_dict['neu']
    compound = sentiment_dict['compound']
    if compound >= 0.05 :
        result = 'Positive'
    elif compound <= - 0.05 :
        result = 'Negative'
    else :
        result = 'Neutral'
    return result


def upload():
    global filename
    filename = filedialog.askdirectory(initialdir = ".")
    text.delete('1.0', END)
    text.insert(END,filename+' Loaded')

def readDataset():
    global movies_df, ratings_df, tweets_df, users_df
    tweets_df = pd.read_csv("Dataset/tweets.csv", encoding='utf-
     8')
    movies_df =  pd.read_csv("Dataset/movies.csv")
    ratings_df = pd.read_csv("Dataset/ratings.csv")
    text.delete('1.0', END)
    text.insert(END,"Tweets Data\n")
    text.insert(END,str(tweets_df.head())+"\n\n")
    text.insert(END,"Movies Data\n")
    text.insert(END,str(movies_df.head())+"\n\n")
    text.insert(END,"Ratings Data\n")
    text.insert(END,str(ratings_df.head())+"\n\n")
    tweets_df = tweets_df.values

def collaborativeFilter():
    global X
    global tfidf_vectorizer
```

```
    textArray.clear()
    movieArray.clear()
    movieNames.clear()
    global movies_df
    movies_frame = movies_df.values
    for i in range(len(movies_df)):
        movie_id = movies_frame[i,0]
        movie_name = movies_frame[i,1]
        movie_type = movies_frame[i,2]
        movie_type = movie_type.replace("|"," ")
        data = movie_name+" "+movie_type
        data = data.lower()
        data = re.sub('[^A-Za-z]+', ' ', data)
        data = data.strip("\n").strip()
        textArray.append(data)
        movieArray.append(movie_id)
        movieNames.append(movie_name)

    tfidf_vectorizer = TfidfVectorizer(use_idf=True,
      smooth_idf=False, norm=None, decode_error='replace')
    tfidf = tfidf_vectorizer.fit_transform(textArray).toarray()
    df = pd.DataFrame(tfidf,
      columns=tfidf_vectorizer.get_feature_names())
    text.delete('1.0', END)
    text.insert(END,"Movies Content Based Model\n\n")
    text.insert(END,str(df.head()))
    df = df.values
    X = df[:, 0:df.shape[1]]

def contentFilter():
    global movies_df, ratings_df
    text.delete('1.0', END)
    ratings_df = pd.merge(ratings_df, movies_df, on='movie_id')
    text.insert(END,"Movies Collaborative Model\n\n")
    text.insert(END,str(ratings_df.head()))



def sentimentModel():
    text.delete('1.0', END)
    global sid
    sid = SentimentIntensityAnalyzer()
    text.insert(END,"Sentiment Model Generated\n\n")

def getCollaborative(name,ratings_df):
```

```
    ratings_value = 0
    temp = ratings_df.values
    for i in range(len(temp)):
        if temp[i,4] == name:
            ratings_value = temp[i,2]
            break
    return ratings_value

def recommendation():
    text.delete('1.0', END)
    global pos
    global neg
    global neu
    pos = 0
    neg = 0
    neu = 0
    global X
    global tfidf_vectorizer
    query = simpledialog.askstring("Query Dialog", "Type here
     desired movie details to get recommendation list",
     parent=main)
    if len(query) > 0:
        testData = query.lower()
        testData = testData.strip()
        testData = re.sub('[^A-Za-z]+', ' ', testData)
        testArray = []
        testArray.append(testData)
        testData = tfidf_vectorizer.transform(testArray).toarray()
        testData = testData[0]
        for i in range(len(X)):
            content_recom = dot(X[i],
      testData)/(norm(X[i])*norm(testData))
            if content_recom > 0:
                sentiment = getSentiment(movieArray[i],tweets_df)
                if sentiment == 'Positive':
                    pos = pos + 1
                if sentiment == "Negative":
                    neg = neg + 1
                if sentiment == 'Neutral':
                    neu = neu + 1
                text.insert(END,movieNames[i]+" Content Based
      Score "+str(content_recom)+"\n")
                text.insert(END,movieNames[i]+" Movie Sentiment
      "+sentiment+"\n")
```

```
            collaborative_filter =
    getCollaborative(movieNames[i],ratings_df)
            text.insert(END,movieNames[i]+" Collaborative Filter
    Rating "+str(collaborative_filter)+"\n\n")

def graph():
    global pos
    global neg
    global neu
    height = [pos,neg,neu]
    bars = ('Positive Sentiment','Negative Sentiment','Neutral
     Sentiment')
    y_pos = np.arange(len(bars))
    plt.bar(y_pos, height)
    plt.xticks(y_pos, bars)
    plt.show()

def close():
  main.destroy()

font = ('times', 16, 'bold')
title = Label(main, text='Movie Recommendation System Using
    Sentiment Analysis From Microblogging Data',
    justify=LEFT)
title.config(bg='lavender blush', fg='DarkOrchid1')
title.config(font=font)
title.config(height=3, width=120)
title.place(x=100,y=5)
title.pack()

font1 = ('times', 13, 'bold')
uploadButton = Button(main, text="Upload Movie Tweetings
    Dataset", command=upload)
uploadButton.place(x=10,y=100)
uploadButton.config(font=font1)

readButton = Button(main, text="Read & Preprocess Dataset",
    command=readDataset)
readButton.place(x=300,y=100)
readButton.config(font=font1)

cfButton = Button(main, text="Build Collaborative Filtering
    Model", command=collaborativeFilter)
cfButton.place(x=600,y=100)
```

```
cfButton.config(font=font1)

cbButton = Button(main, text="Build Content Based Model",
    command=contentFilter)
cbButton.place(x=10,y=150)
cbButton.config(font=font1)

smButton = Button(main, text="Build Sentiment Model",
    command=sentimentModel)
smButton.place(x=300,y=150)
smButton.config(font=font1)

recommendationButton = Button(main, text="Movie
    Recommendation using All Models",
    command=recommendation)
recommendationButton.place(x=600,y=150)
recommendationButton.config(font=font1)

graphButton = Button(main, text="Top 10 Movies Sentiment
    Graph", command=graph)
graphButton.place(x=10,y=200)
graphButton.config(font=font1)

closeButton = Button(main, text="Close Application",
    command=close)
closeButton.place(x=300,y=200)
closeButton.config(font=font1)

font1 = ('times', 12, 'bold')
text=Text(main,height=20,width=160)
scroll=Scrollbar(text)
text.configure(yscrollcommand=scroll.set)
text.place(x=10,y=250)
text.config(font=font1)

main.config(bg='light coral')
    main.mainloop()
```

A

PROJECT REPORT

On

SMART CONTRACT BASED ACCESS CONTROL FOR HEALTH CARE DATA

Submitted by

Ms.Chitra M (17K81A0570) **Mr.G. Saidatta (17K81A0582)**

Mr.O.Chakradhar (17K81A05A2) **Ms.V.Sandhya (17K81A05B8)**

in partial fulfillment for the award of the degreeof

BACHELOR OF TECHNOLOGY

IN

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Under the Guidance of

Mrs. Manu Hajari

Assistant Professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ST.MARTIN'S ENGINEERING COLLEGE
An Autonomous Institute

Dhulapally, Secunderabad – 500100

JUNE 2021

BONAFIDE CERTIFICATE

This is to certify that the project entitled Smart Contract Based Access Control For Health Care Data , is being submitted **Ms. Chitra M 17K81A0570, Mr. Gudala Saidatta 17K81A0582, Mr. Odnam Chakradhar 17K81A05A2, Ms. Vutukotu Sandhya 17K81A05B8** in partial fulfillment of the requirement for the award of the degree of **BACHELOR OF TECHNOLOGY IN Computer Science And Engineering** is recorded of bonafide work carried out by them. The result embodied in this report have been verified and found satisfactory.

Guide

Manu Hajari

Department of CSE

Head of the Department

Dr.M.NARAYANAN

Department of CSE

Internal Examiner

External Examiner

Place:

Date:

DECLARATION

We, the student of **Bachelor of Technology** in Department of Computer Science and Engineering', session: 2017 – 2021, St. Martin's Engineering College, Dhulapally, Kompally, Secunderabad, hereby declare that work presented in this Project Work entitled Smart Contract Based Access Control For Health Care Data is the outcome of our own bonafide work and is correct to the best of our knowledge and this work has been undertaken taking care of Engineering Ethics. This result embodied in this project report has not been submitted in any university for award of any degree.

Chitra M	17K81A0570
Gudala Saidatta	17K81A0582
Odnam Chakradhar	17K81A05A2
Vutukotu Sandhya	17K81A05B8

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible and whose encouragements and guidance have crowded effects with success.

We extended our deep sense of gratitude to Principal**, Dr. P. SANTOSH KUMAR PATRA**, St. Martin's Engineering College, Dhulapally, for permitting us to undertake this project.

We are also thankful to **Dr. M.NARAYANAN**, Head of the Department, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

We would like to thank **Dr. T. POONGOTHAI,** Department of Computer Science and Engineering (AI & ML) for her encouragement and insightful comments and as well as our project coordinators **Dr. B.RAJALINGAM**, Associate Professor and **Mr. J.SUDHAKAR**, Assistant Professor, in Department of Computer Science and Engineering for their valuable support.

We would like to express our sincere gratitude and indebtedness to our project supervisor Mrs. Manu Hajari, Assistant Professor, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for her support and guidance throughout our project.

Finally, we express thanks to all those who have helped us successfully to completing this project. Furthermore, we would like to thank our family and friends for their moral support and encouragement.

We express thanks to all those who have helped us in successfully completing the project.

Chitra M	17K81A0570
Gudala Saidatta	17K81A0582
Odnam Chakradhar	17K81A05A2
Vutukotu Sandhya	17K81A05B8

ABSTRACT

The main objective of this project is to securely store and maintain the patient records in the healthcare data. In the existing system, the collection and accessing of medical records are done only by storing which misleads and corrupts the data. Here search delay of the scheme is proportional to the size of the database and it is not suitable for the large scale databases.In this system, the large amount of data is created, disseminated, stored and accessed over the cloud using block chain technology.The block chain technology is used to protect the healthcare data hosted within the cloud. The blocks contain medical data and the timestamp. We achieve data security, data integrity and confidentiality by a layered approach that includes data encryption using AES algorithm, key management, strong access controls, and security intelligence.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF OUTPUT SCREENS

LIST OF ACRONYMS

<AVI>	Audio Video Interlace
<CSP>	Cloud Service Process
<CPU>	Central Processing Unit
<GB>	Giga Bytes
<GUI>	Graphical User Interface
<EHR>	Electronic Health Records

1. INTRODUCTION

Cloud computing offers an opportunity for individuals and companies to offload to powerful servers the burden of managing large amounts of data and performing computationally demanding operations. Due to the increasing popularity of cloud computing, more and more Data owners are motivated to outsource their data to cloud servers for great convenience and reduced cost in data management. Data owners offer services to a large number of businesses and companies, they stick to high security standards to improve data security by following a layered approach that includes data encryption, key management, strong access controls, and security intelligence.

Healthcare is a data-intensive domain where a large amount of data is created, disseminated, stored, and accessed daily. It is clear that technology can play a significant role in enhancing the quality of care for patients and potentially reduce costs by more efficiently allocating resources in terms of personnel, equipment, etc.

Generally, Electronic Medical Records (EMRs) contain medical and clinical data related to a given patient and stored by the responsible healthcare provider. This facilitates the retrieval and analysis of healthcare data. To better support the management of EMRs, early generations of Health Information Systems (HIS) are designed with the capability to create new EMR instances, store them, and query and retrieve stored EMRs of interest. HIS can be relatively simple solutions, which can be schematically described as a graphical user interface or a web service. These are generally the front-end with a database at the back-end, in a centralized or distributed implementation. With patient mobility being increasingly the norm in today's society, it became evident that multiple stand-alone EMR solutions must be made interoperable to facilitate sharing of healthcare data among different providers, even across national borders, as needed. For example, in medical tourism hubs such as Singapore, the need for real-time healthcare data sharing between different providers and across nations becomes more pronounced.

To facilitate data sharing or even patient data portability, there is a need for EMRs to formalize their data structure and the design of HIS. Electronic Health Records (EHRs), for example, are designed to allow patient medical history to move with the patient or be made available to multiple healthcare. EHRs have a richer data structure than EMRs. There have also been initiatives to develop HIS and infrastructures that are able to scale and support future needs, as evidenced by the various national and international initiatives such as the Fascicolo Sanitario Elettronico (FSE) project in Italy.

Recently, the pervasiveness of smart devices has also resulted in a paradigm shift within the healthcare industry. Such devices can be user-owned or installed by the healthcare provider to measure the

well-being of the users and inform/facilitate medical treatment and monitoring of patients. For example, there is a wide range of mobile applications (apps) in health, fitness, weight-loss, and other healthcare related categories. These apps mainly function as a tracking tool, such as registering user exercises/workouts, keeping the count of consumed calories, and other statistics, and so on.

There are also devices with embedded sensors for more advanced medical tasks, such as bracelets to measure heartbeat during workouts, or devices for self-testing of glucose. The data can be continuously gathered and sent in real-time to a smart device, before being sent to a remote healthcare cloud for further analysis.

Blockchain was originally designed to record transaction data, which is relatively small in size and linear. In other words, one only concerns itself about whether the current transaction can be traced backwards to the original "deal".

1.1 PROJECT OVERVIEW

Electronic Medical Records contain medical and clinical data related to a given patient and stored by the responsible healthcare provider. The collection and accessing of medical records are done only by storing which misleads and corrupts the data. The main objective of this project is to securely store and maintain the patient records in the healthcare data. The blockchain technology is used to protect the healthcare data hosted within the cloud. This facilitates the retrieval and analysis of healthcare data. To better support the management of EMRs.

1.2 PROJECT OBJECTIVE

Healthcare is a data-intensive domain where a large amount of data is created, disseminated, stored, and accessed daily. The main objective is to ensure the data security which enable access control that is cryptographically enhanced, and also provides the protection of cloud database. This leads to secure and efficient processing.

1.3 ORGANIZING THE CHAPTERS

This documentation consists of 10 different chapter and they are:

1. Introduction – This chapter covers the overview of our project and its objectives.

2. Literature Survey – This includes the details of our survey.

3. Software and Hardware Requirements – We specify our software and hardware requirements here.

4. Software Development Analysis – This section includes the problem definition and details of the modules we used in our project.

5. Project System Design – This chapter includes the design part of our project which includes uml diagrams.

6. Project Coding – This section contains the details of our project code.

7. Project Testing – The details of test cases and testing are included in this chapter.

8. Output Screens – This contains the screenshots of how our project looks like when executed.

9. Experimental Results – This chapter contains the screenshots of our results.

10. Conclusion and Future Enhancements – This covers the conclusion of our project and the possible future developments.

2. LITERATURE SURVEY

2.1 SURVEY ON BACKGROUND

1. BaDS: Blockchain-Based Architecture for Data Sharing with ABS and CP-ABE in IoT.
AUTHOURS: Yunru Zhang, Debiao He, and Kim-Kwang Raymond Choo

Internet of Things (IoT) and cloud computing are increasingly integrated, in the sense that data collected from IoT devices (generally with limited computational and storage resources) are being sent to the cloud for processing, etc., We proposed a novel blockchain-based architecture for data sharing with attribute-based cryptosystem (BaDS) in this paper. The architecture can achieve privacy-preserving, user-self-controlled data sharing, and decentralization by using blockchain and several attribute-based cryptosystems. Specifically, ABS and CP-ABE provide the capability for fine-grained access control. We introduced the security requirements of the proposed BaDS architecture and then explained how the proposed BaDS architecture satisfies the security requirement. We also implement the BaDS architecture and analyze its computation cost.

Advantages

- Implementing digital signatures.
- Cryptographic protocols with different security and privacy features.
- Supporting various signature schemes without adding additional hardware complexity compared to a hardware implementation of a conventional signature scheme.

Disadvantages

- Encryption keys aren't simple strings of text like passwords
- Damage is massive when you lost your symmetric key

2. Blockchain for Secure and Efficient Data Sharing in Vehicular Edge Computing and Networks.
AUTHORS: Jiawen Kang, Rong Yu, Xumin Huang, Maoqiang Wu, Sabita Maharjan, Shengli Xie, and Yan Zhang

The drastically increasing volume and the growing trend on the types of data have brought in the possibility of realizing advanced applications such as enhanced driving safety, and have enriched existing vehicular services through data sharing among vehicles and data analysis. We exploit consortium blockchain and smart contract technologies to achieve secure data storage and sharing in vehicular edge networks. These technologies efficiently prevent data sharing without authorization. In addition, we

propose a reputation-based data sharing scheme to ensure high-quality data sharing among vehicles. A three-weight subjective logic model is utilized for precisely managing reputation of the vehicles. Numerical results based on a real dataset show that our schemes achieve reasonable efficiency and high-level of security for data sharing in VECONs.

Advantages

- Security against adaptive chosen-keyword attacks.
- Compact indexes.
- Ability to add and delete files efficiently.

Disadvantages

- Every means of electronic communication is insecure as it is impossible to guarantee that no one will be able to tap communication channels. So the only secure way of exchanging keys would be exchanging them personally.

3. Blockchain Meets IoT: An Architecture for Scalable Access Management in IoT.

AUTHOR: Oscar Novo.

The Internet of Things (IoT) is stepping out of its infancy into full maturity and establishing itself as a part of the future Internet. One of the technical challenges of having billions of devices deployed worldwide is the ability to manage them. Although access management technologies exist in IoT, they are based on centralized models which introduce a new variety of technical limitations to manage them globally. In this paper, we propose a new architecture for arbitrating roles and permissions in IoT. The new architecture is a fully distributed access control system for IoT based on blockchain technology. The architecture is backed by a proof of concept implementation and evaluated in realistic IoT scenarios. The results show that the blockchain technology could be used as access management technology in specific scalable IoT scenarios.

Advantages

- Providing performance results of a prototype applied to several large representative data sets, including encrypted search over the whole English Wikipedia.

Disadvantages

- Exact matching may retrieve too few or too many documents.

4. Blockchain distributed ledger technologies for biomedical and health care applications.

AUTHOR: Kuo TT, Kim HE, and Ohno-Machado L

Blockchain is a distributed, immutable ledger technology introduced as the enabling mechanism to support cryptocurrencies. Blockchain solutions are currently being proposed to address diverse problems in different domains. This paper presents a scoping review of the scientific literature to map the current research area of blockchain applications in the biomedical domain. The goal is to identify biomedical problems treated with blockchain technology, the level of maturity of respective approaches, types of biomedical data considered, blockchain features and functionalities exploited and blockchain technology frameworks used. Our findings show that the field is still in its infancy, with the majority of studies in the conceptual or architectural design phase; only one study reports real world demonstration and evaluation. Research is greatly focused on integration, integrity and access control of health records and related patient data. However, other diverse and interesting applications are emerging, addressing medical research, clinical trials, medicines supply chain, and medical insurance.

Advantages

- Complete expressiveness for any identifiable subset of collection.
- A symmetric cryptosystem uses password authentication to prove the receiver's identity

Disadvantages

- Cannot provide digital signatures that cannot be repudiated

5. Towards Using Blockchain Technology for eHealth Data Access Management.

AUTHORS: Nabil Rifi, Elie Rachkidi, Nazim Agoulmine, and Nada Chendeb Taher.

ehealth is a technology that is growing in importance over time, varying from remote access to Medical Records, such as Electronic Health Records (EHR), or Electronic Medical Records (EMR), to real-time data exchange from different on-body sensors coming from different patients. With this huge amount of critical data being exchanged, problems and challenges arise. Privacy and confidentiality of this critical medical data are of high concern to the patients and authorized persons to use this data. On the other hand, scalability and interoperability are also important problems that should be considered in the final solution. This paper illustrates the specific problems and highlights the benefits of the blockchain technology for the deployment of a secure and a scalable solution for medical data exchange in order to have the best performance possible.

Advantage:

- Efficient data search

Disadvantage:

- Data Integrity Problem

6. A Standards-Based Architecture Proposal for Integrating PatientHealth Apps to Electronic Health Record Systems.

AUTHOR: S. Marceglia

Mobile health Applications (mHealth Apps) are opening the way to patients' responsible and active involvement with their own healthcare management. However, apart from Apps allowing patient's access to their electronic health records (EHRs), mHealth Apps are currently developed as dedicated "island systems". Although much work has been done on patient's access to EHRs, transfer of information from mHealth Apps to EHR systems is still low. This study proposes a standards-based architecture that can be adopted by mHealth Apps to exchange information with EHRs to support better quality of care.

7. Trustworthy Processing of Healthcare Big Data in Hybrid Clouds.

AUTHOR: S. Nepal.

Managing large, heterogeneous, and rapidly increasing volumes of data, and extracting value out of such data, has long been a challenge. In the past, this was partially mitigated by fast processing technologies that exploited Moore's law. However, with a fundamental shift toward big data applications, data volumes are growing faster than they can be analyzed, regardless of increased CPU speeds or other performance improvements. Efforts thus need to focus on the development of security and privacy techniques that can deal with changing volume, velocity, and variety of heterogeneous dataflow, be ported to diverse big data programming frameworks, deal with variable computational complexity due to heterogeneous VM, storage, and network configurations across multiple clouds, and be seamlessly implemented in multicloud orchestration APIs such as jclouds.

8. Healthcare-Related Data in the Cloud: Challenges and Opportunities.

AUTHOR: V. Casola

A key issue in electronic health systems is the underlying security and privacy risk. For example, confidential patient information or medical records ending up in the hands of a person not privy to the information could have far-reaching consequences. With the trend toward cloud computing use in the healthcare industry continuing to grow (for example, using cloud platforms to digitally manage health-related data including electronic health records), security and privacy concerns must be adequately addressed, and regulations on data protections made to be in compliance. This column examines several

key issues and requirements underpinning the use of cloud computing for managing healthcare-related data as well as potential solutions.

9. Opportunities and Challenges of Cloud Computing to Improve Health Care Services.

AUTHOR: Mu-Hsing Kuo,

Cloud computing is a new way of delivering computing resources and services. Many managers and experts believe that it can improve health care services, benefit health care research, and change the face of health information technology. However, as with any innovation, cloud computing should be rigorously evaluated before its widespread adoption. This paper discusses the concept and its current place in health care, and uses 4 aspects (management, technology, security, and legal) to evaluate the opportunities and challenges of this computing model. Strategic planning that could be used by a health organization to determine its direction, strategy, and resource allocation when it has decided to migrate from traditional to cloud-based health services.

10. Personal Health Records: Definitions, Benefits, and Strategies for Overcoming Barriers to Adoption.

AUTHOR: P.C. Tang

Recently there has been a remarkable upsurge in activity surrounding the adoption of personal health record (PHR) systems for patients and consumers. The biomedical literature does not yet adequately describe the potential capabilities and utility of PHR systems. In addition, the lack of a proven business case for widespread deployment hinders PHR adoption. In a 2005 working symposium, the American Medical Informatics Association's College of Medical Informatics discussed the issues surrounding personal health record systems and developed recommendations for PHR-promoting activities. Personal health record systems are more than just static repositories for patient data; they combine data, knowledge, and software tools, which help patients to become active participants in their own care. When PHRs are integrated with electronic health record systems, they provide greater benefits than would stand-alone systems for consumers. This paper summarizes the College Symposium discussions on PHR systems and provides definitions, system characteristics, technical architectures, benefits, barriers to adoption, and strategies for increasing adoption.

2.2. CONCLUSION ON SURVEY

Cloud computing is the delivery of computing and storage space as a service to a distributed community of end users. The model of Cloud computing is, all the servers, networks, applications and other elements related to data centers are made available to end users. Cloud computing is growing now-a-days in the interest of technical and business organizations but this can also be beneficial for solving social issues. Cloud computing refers to manipulating, configuring, and accessing the applications online. It offers online data storage, infrastructure and application. Implementing the system using the cloud makes the data decentralized that prevents data loss. Data sharing is also possible without any interruption and also enhances data integrity. With the help of blockchain data can be more secured.

3. SOFTWARE AND HARDWARE REQUIREMENTS

3.1 SOFTWARE REQUIREMENTS

- O/S : Windows 10.
- Language : Java.
- IDE : Net Beans 8.2
- Data Base : MySQL

3.2 HARDWARE REQUIREMENTS

- Processor : i3
- Hard Disk : 500 GB
- Mouse : Standard Mouse
- Keyboard : Standard Windows Keyboard
- Ram : 4GB

4. SOFTWARE DEVELOPMENT ANALYSIS

4.1 OVERVIEW OF PROBLEM

In current healthcare systems, electronic medical records (EMRs) are always located in different hospitals. However, it leads to single point of failure as healthcare providers being the real owner lose track of their private and sensitive EMRs, and which is not suitable for the large-scale databases.

4.2 DEFINE THE PROBLEM

The collection and accessing of medical records are done only by storing which misleads and corrupts the data. This is only permitted to the small-scale databases

4.3 MODULES OVERVIEW

We developed this system with the help of building three modules they are Registration, Healthcare Provider, Cloud Service Provider. Here a health care provider can register and upload the patients records to the cloud and encrypt the data for security using key generation, and also blocks are created to enhance the security. CSP module is to view the records by the health care providers. Now while downloading at other end provider can download the record.

4.4 DEFINE THE MODULES

- Registration
- Healthcare Provider
  - Load patient Records
  - Key Generation
  - Encrypt patient Records
  - Block Creation
  - Upload and Download Patient Records
- Cloud Service Provider
  - View Patient Records
  - Grant or Revoke Permission

4.5 MODULES FUNCTIONALITY

Registration

It is a process of enrolling or being enrolled into the cloud. To utilize the cloud documents, every healthcare provider should enroll. During this process your basic information like email, contacts etc., are collected and stored in the Cloud. The cloud id for a particular user will get automatically generated during the registration.

Cloud ID

Every user should create a Cloud ID and use it to identify something with near certainty that the identifier does not duplicate one that has already been, or will be, created to identify something else. Information labelled with Cloud ID by independent parties can therefore be later combined into a single database, or transmitted on the same channel, without needing to resolve conflicts between identifiers

Healthcare Provider

Data Selection and Loading

In this process, the health provider choose patient healthcare records for uploading and maintaining the dataset in the cloud

Key Generation

The secret key is generated using cryptographic algorithm. This key is used for encrypting the dataset.

Encrypt Patient Records

The data is encrypted for secure maintenance. So that the unauthorized person cannot be able to access the data that are presented in the cloud.

Block Creation

Each block contains patient record and it’s timestamp. A blockchain, originally block is a growing list of records called blocks.

Upload and Download Patient Records

After creating the block, the healthcare provider will upload the records into the cloud. Suppose, if they want to retrieve an record from cloud, first the healthcare provider search the record. Based on the search it will show the results. After getting an approval and key from the cloud service provider the healthcare provider can download the data.

Cloud Service Provider

The cloud service provider maintain all the patient records and also they can provide a permission to the user to access the data.The Cloud Service Provider can view all the uploaded and downloaded documents in the Cloud. The CSP receives the document request from the Data User, verifies the authentication before granting permission. Then the CSP executes the query and returns the encrypted document according to the search token. And also returns an additional proof with the document, to verify the search result.

Public Verification Key

Public verification key is a security measure designed to make sure that your document outsourced in cloud doesn't get hacked. By verifying public key, the Data Owner and the Data User adding another layer of protection to the documents or files in the cloud by confirming each other's identities.

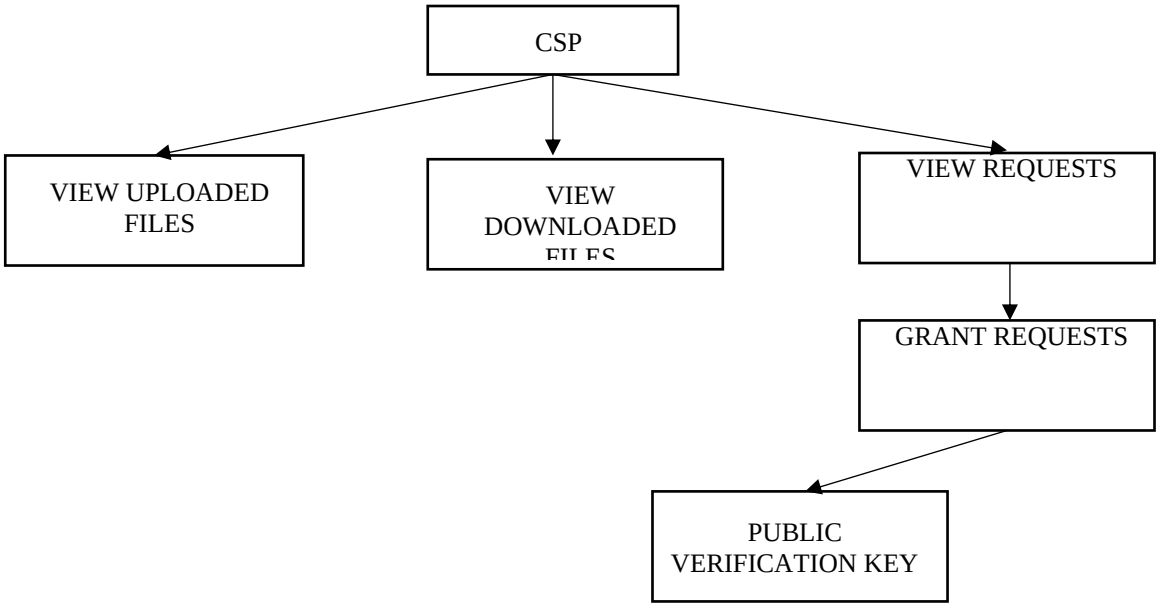


FIG 4.1: PUBLIC KEY VERIFICATION

5.PROJECT SYSTEM DESIGN

5.1 DFD'S IN CASE OF DATABASE PROJECT

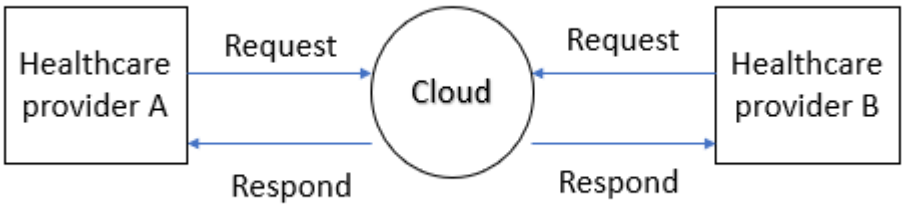


FIG 5.1 DFD

5.2 E-R DIAGRAM

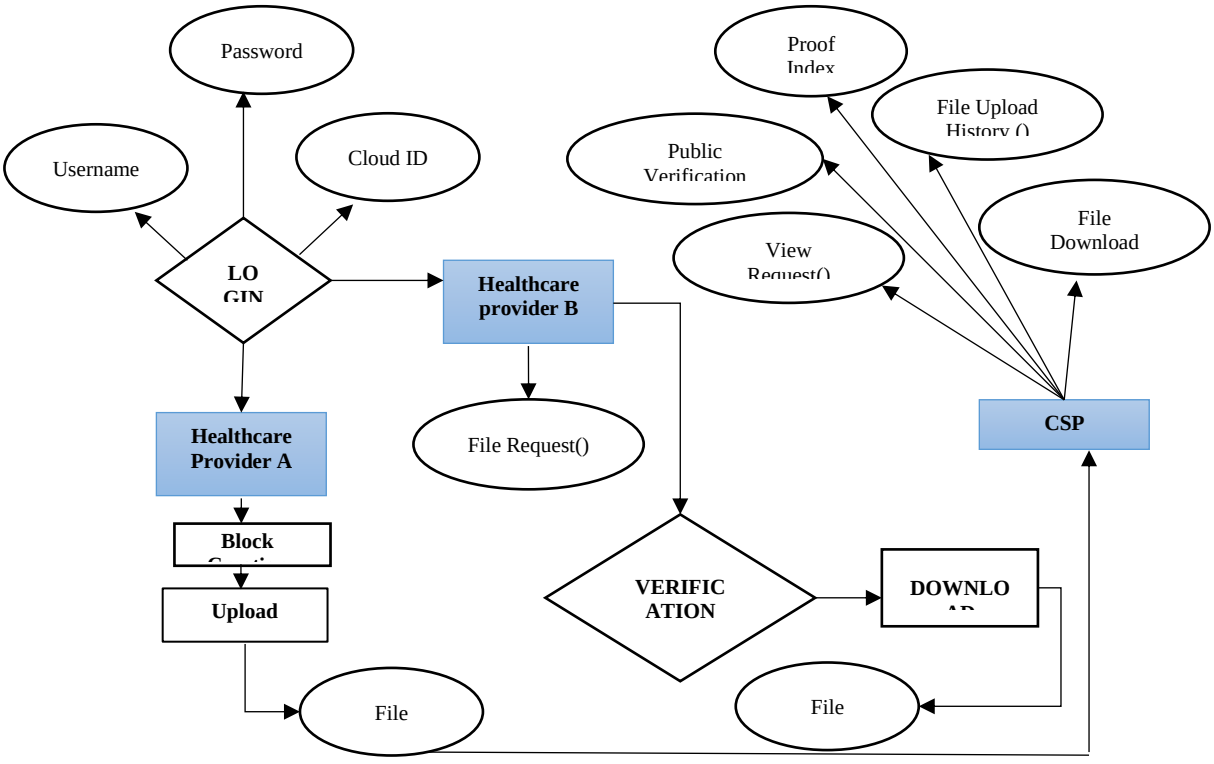


FIG 5.2 : E-R DIAGRAM

5.3 UML DIAGRAMS

USE CASE DIAGRAM: A use case diagram in the Unified Modelling Language (UML) is a type of behavioural diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases.

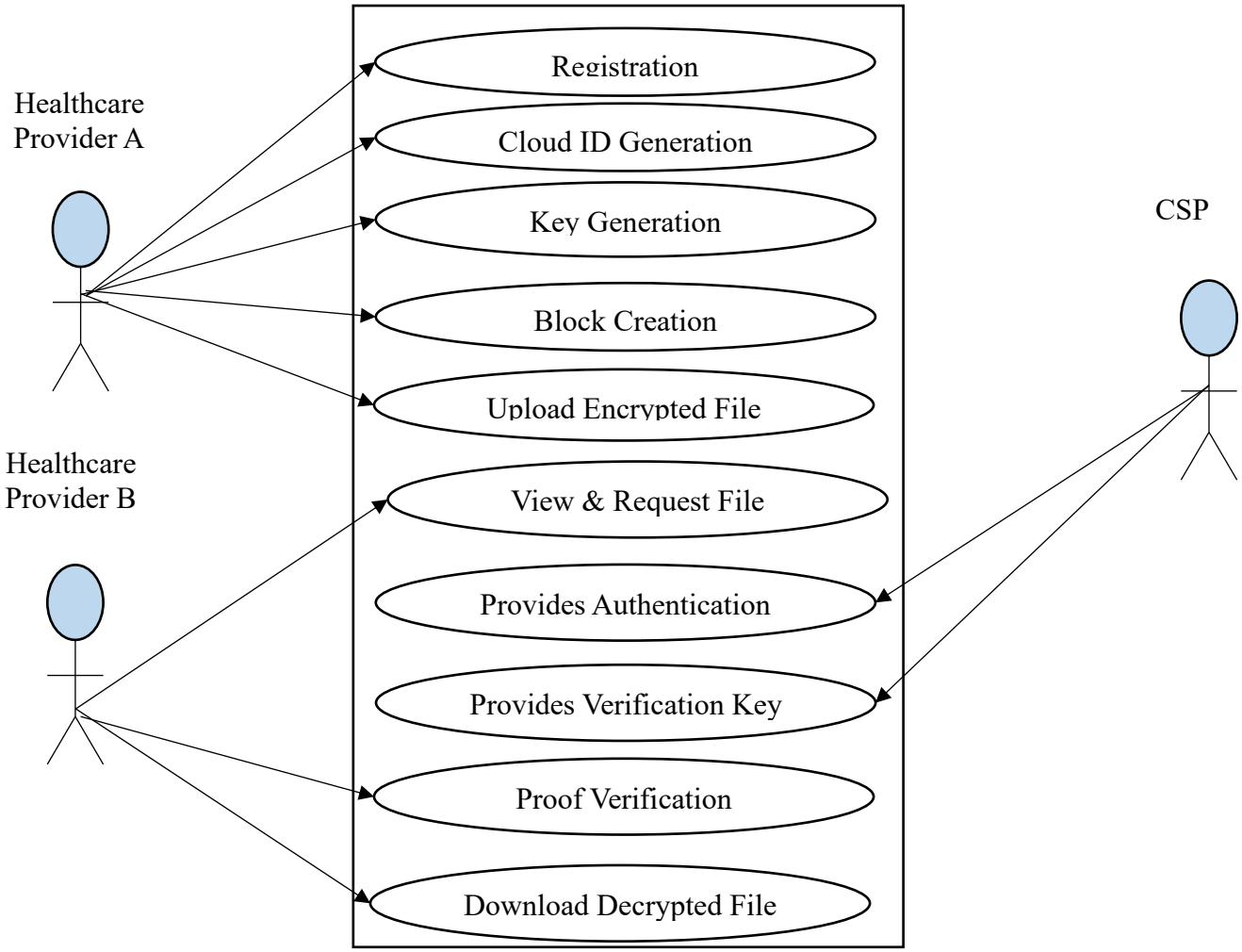


FIG 5.3: USECASE DIAGRAM

COMPONENT DIAGRAM: Component diagrams are used in modelling the physical aspects of object-oriented systems that are used for visualizing, specifying, and documenting component-based systems and also for constructing executable systems through forward and reverse engineering

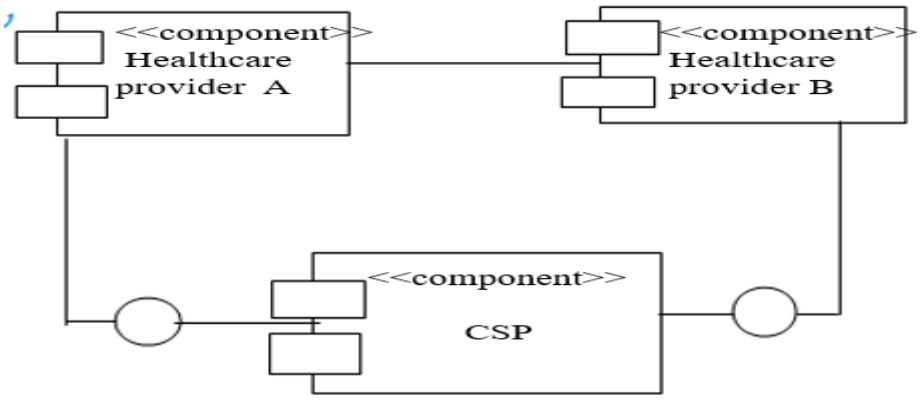


FIG 5.4 : COMPONENT DIAGRAM

CLASS DIAGRAM: A class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes.

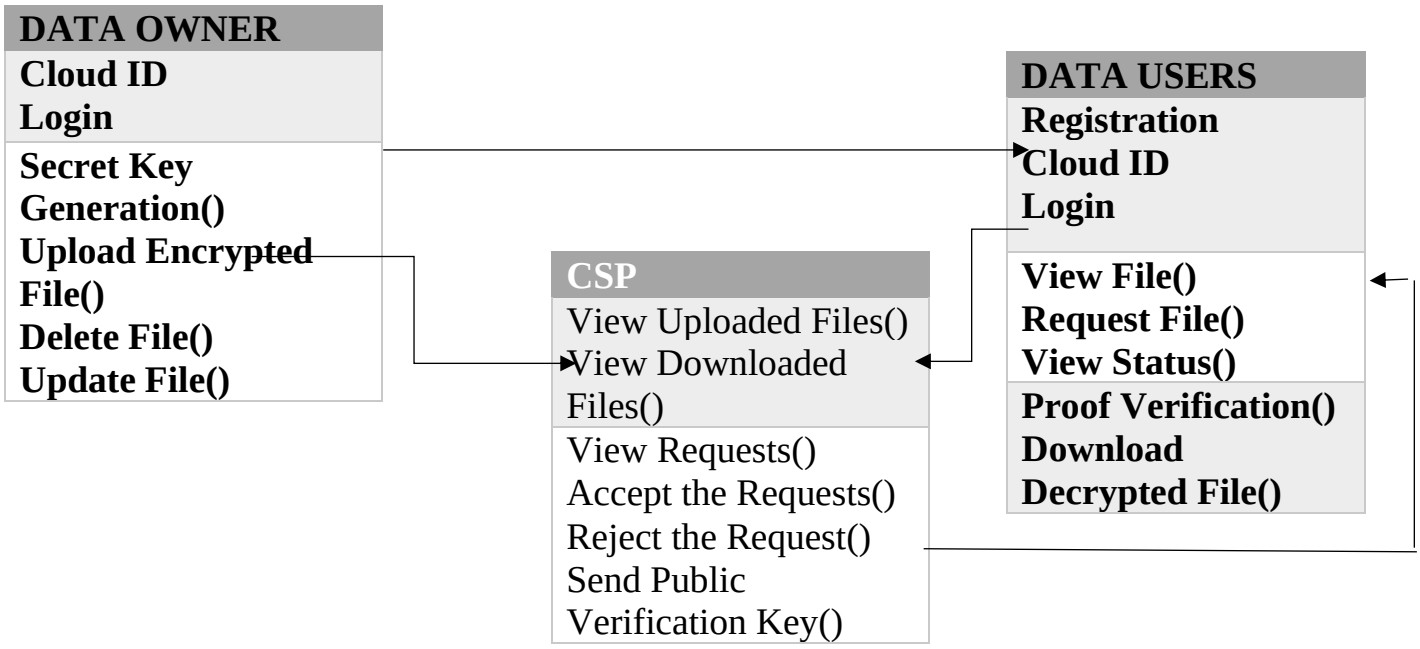


FIG 5.5: CLASS DIAGRAM

COLLABORATION DIAGRAM: Collaboration diagram is used to show how objects interact to perform the behaviour of a particular use case, or a part of a use case. Along with sequence diagrams,

collaboration are used by designers to define and clarify the roles of the objects that perform a particular flow of events of a use case.

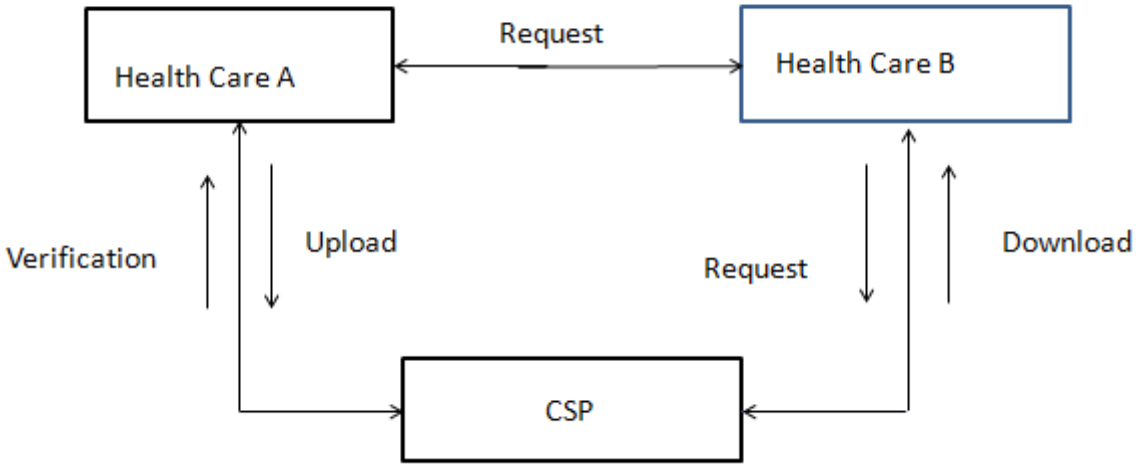


FIG 5.6: COLLABORATION COMPONENT

ACTIVITY DIAGRAM: Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modeling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system

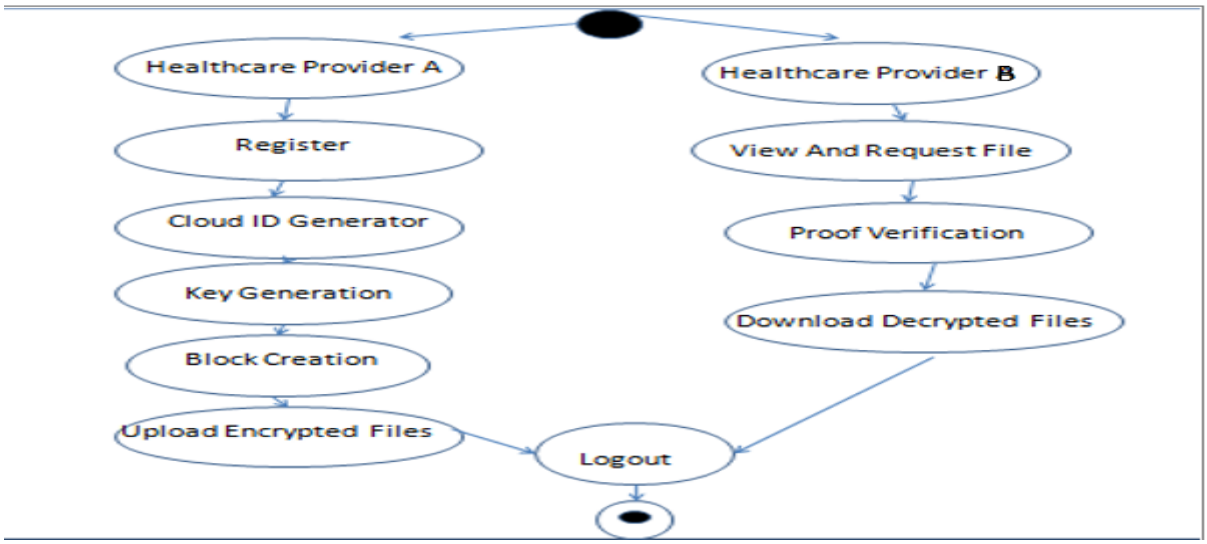


FIG 5.7: ACTIVITY DIAGRAM

COMMUNICATION DIAGRAM: UML Communication Diagrams, previously known as collaboration diagrams are a type of behavioural diagram that shows the interactions that take place between objects in a piece of software or system.

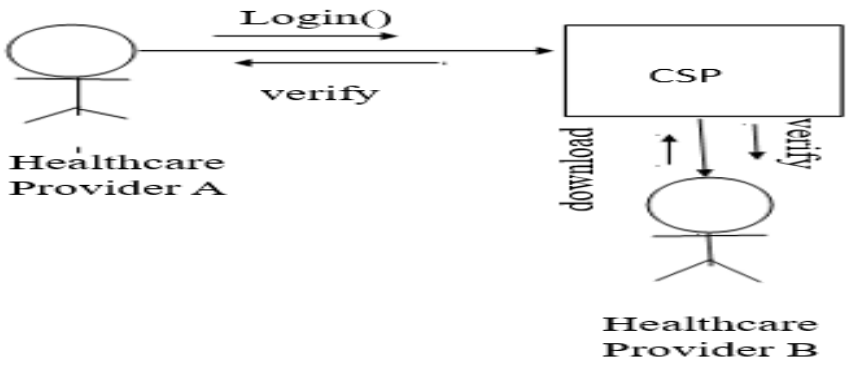


FIG 5.8: COMMUNICATION DIAGRAM

SEQUENCE DIAGRAM: A sequence diagram in Unified Modeling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams

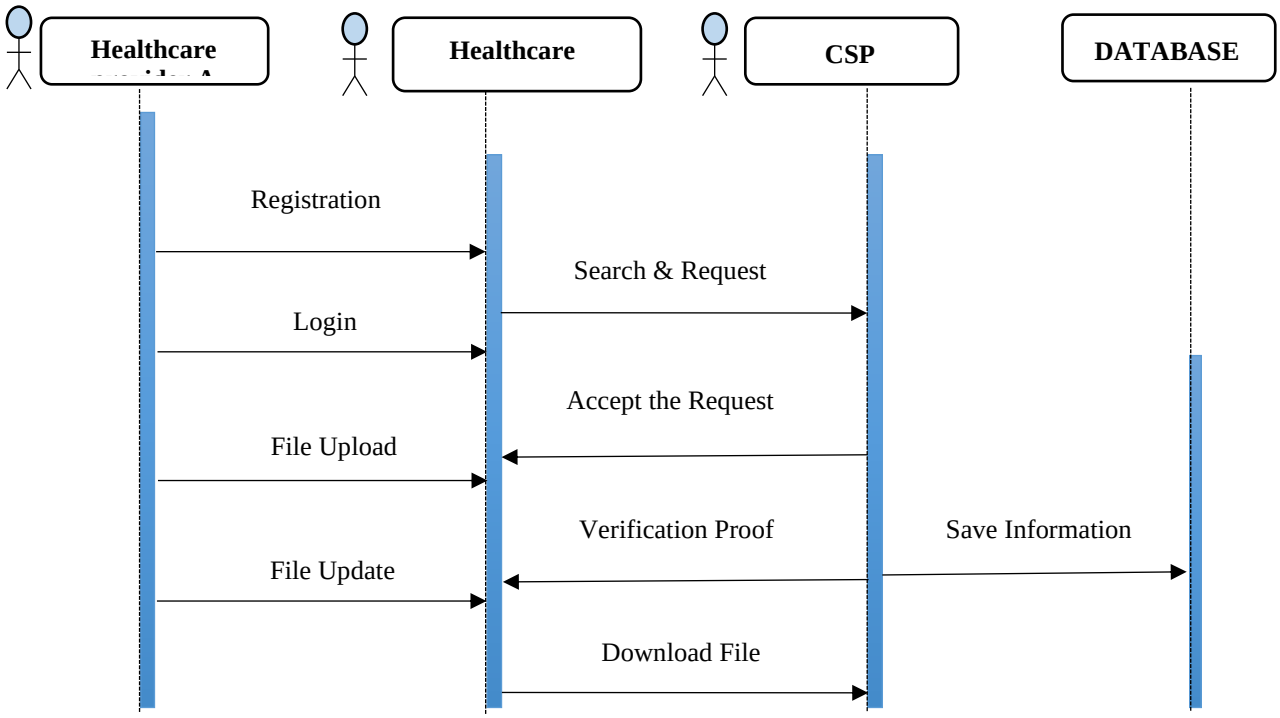


FIG 5.9: SEQUENCE DIAGRAM

6.PROJECT CODE DESIGN

6.1 CODE TEMPLATE:

```
public Blockcreation() {
        initComponents();
 }/**
     * This method is called from within the constructor to initialize the form.
     * regenerated by the Form Editor.
     */
public Cloud() {
        initComponents();
}/**
     * This method is called from within the constructor to initialize the form.
     * regenerated by the Form Editor.
     */
public Encryptrecord() {
        initComponents();
} /**
              * gets the AES encryption key. In your actual programs, this should be safely
              * stored.
              * @return
              * @throws Exception
              */
public Index() {
        initComponents();
}/**
```

```
    * This method is called from within the constructor to initialize the form.
    * regenerated by the Form Editor.
    */
public Mainpage() {
        initComponents();
 }/**
    * This method is called from within the constructor to initialize the form.
    * regenerated by the Form Editor.
    */
public Upload() {
        initComponents();
}/**
    * This method is called from within the constructor to initialize the form.
    * regenerated by the Form Editor.
    */
public Viewrecords() {
        initComponents();
}/**
    * This method is called from within the constructor to initialize the form.
    * regenerated by the Form Editor.
    */
public Download() {
        initComponents();
}/**
    * This method is called from within the constructor to initialize the form.
```

```
    * regenerated by the Form Editor.
    */
```

6.2 OUTLINE FOR VARIOUS FILES

- Blockcreation.java : Creates blocks of medical records
- Cloud.java : Stores the medical records of patients
- Download.java : we can download the EMR's from the cloud
- Encryptrecord.java : encrypts the medical records using the algorithm
- Keysize.java : encrypts record using the symmetric key of fixed size
- Upload.java : Uploads the medical records into the cloud
- Viewrecords.java : we can view the uploaded records from the cloud

6.3 CLASS WITH FUNCTIONALITY

```
public class Blockcreation extends javax.swing.JFrame {
        public Blockcreation()
}
public class Cloud extends javax.swing.JFrame {
        public Cloud()
}
public class Download extends javax.swing.JFrame {
        public Download()
}
public class Encryptrecord extends javax.swing.JFrame {
        public Encryptrecord()
```

```
}
public class Keysize extends ApplicationFrame {
        public Keysize(final String title)
}
public class Upload extends javax.swing.JFrame {
        public Upload()
}
public class Viewrecords extends javax.swing.JFrame {
        public Viewrecords()
}
public class Send_an_request extends javax.swing.JFrame {
        public Send_an_request()
```

6.4 METHODS INPUT AND OUTPUT PARAMETERS

public Blockcreation()

public Cloud()

public Download()

public Encryptrecord()

public Keysize(final String title)

public Mainpage()

public Send_an_request()

public Upload()

public Viewrecords()

7.PROJECT TESTING

7.1 VARIOUS TEST CASES

Unit testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

Integration testing

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

Functional test

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.
Functional testing is centered on the following items:

Valid Input : identified classes of valid input must be accepted.

Invalid Input : identified classes of invalid input must be rejected.

Functions : identified functions must be exercised.

Output : identified classes of application outputs must be exercised.

Systems/Procedures : interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined

processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

System Test

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

7.2 BLACK BOX

- Black box testing is done to find incorrect or missing function
- Interface error
- Errors in external database access
- Performance errors
- Initialization and termination errors

In ‘functional testing’, is performed to validate an application conforms to its specifications of correctly performs all its required functions. So this testing is also called ‘black box testing’. It tests the external behaviour of the system. Here the engineered product can be tested knowing the specified function that a product has been designed to perform, tests can be conducted to demonstrate that each function is fully operational.

7.3 WHITEBOX TESTING

White Box testing is a test case design method that uses the control structure of the procedural design to drive cases. Using the white box testing methods, we derived test cases that guarantee that all independent paths within a module have been exercised at least once.

8. OUTPUT SCREENS

8.1 USER INTERFACES

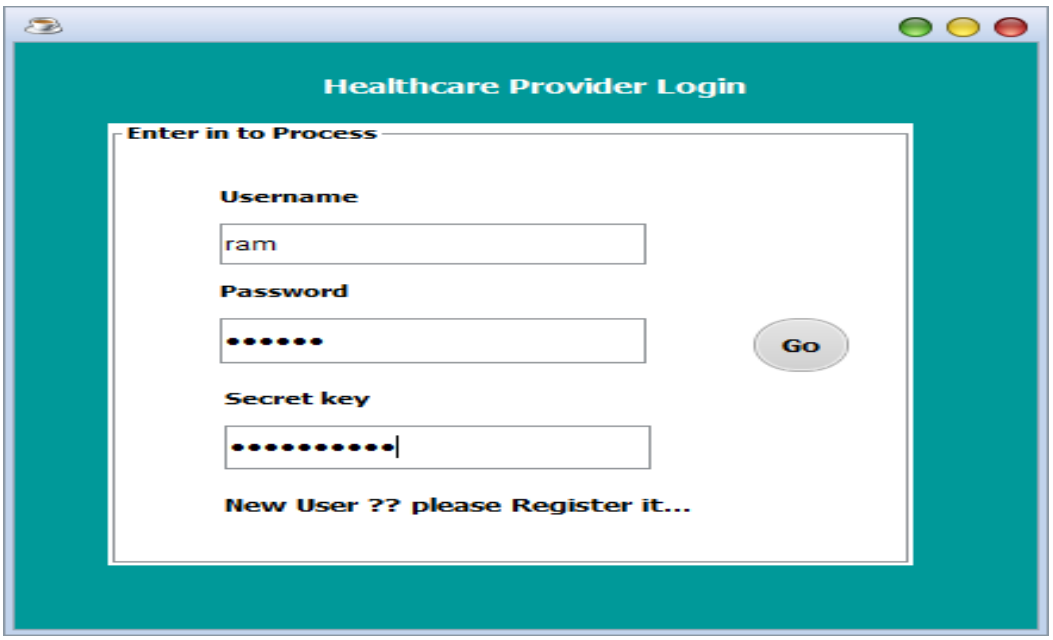


FIG 8.1: LOGIN PAGE

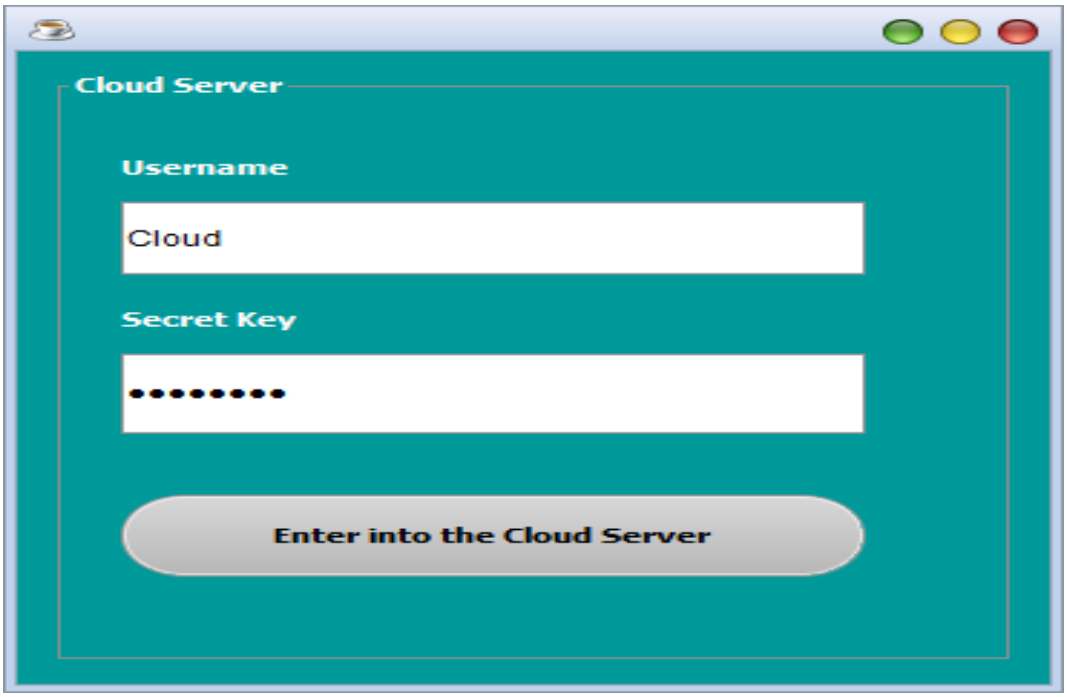


FIG 8.2: CLOUDE SERVER

8.2 OUTPUT SCREENS

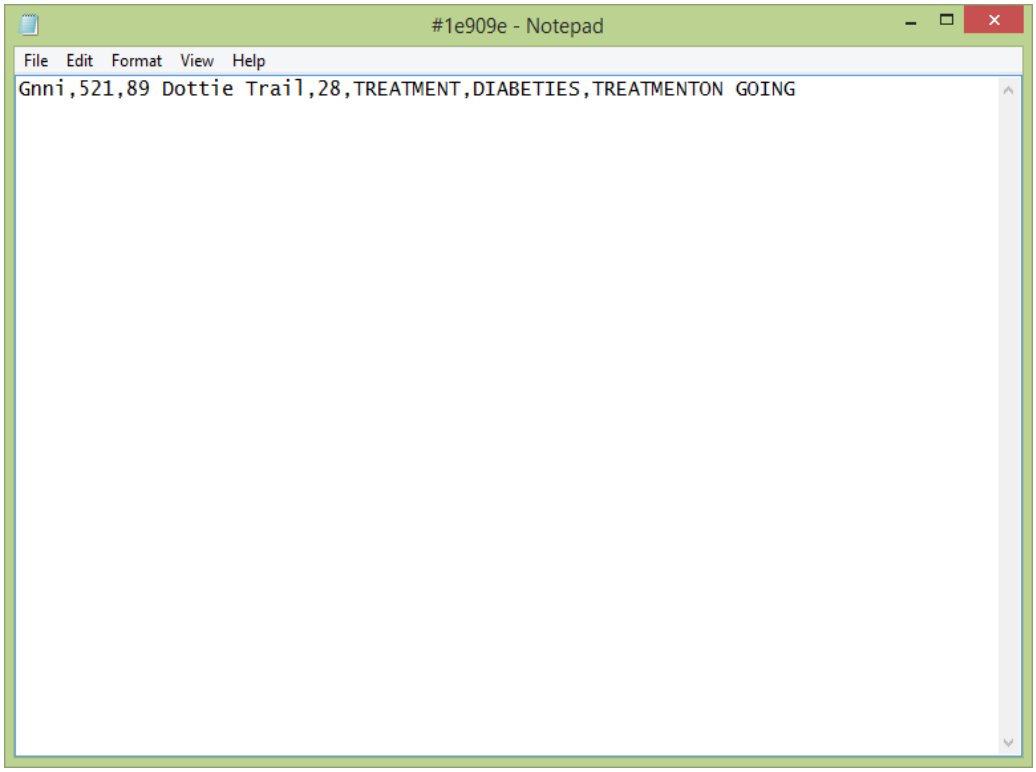


FIG 8.3: DOWNLOADED RECORD

9.EXPERIMENTAL RESULTS

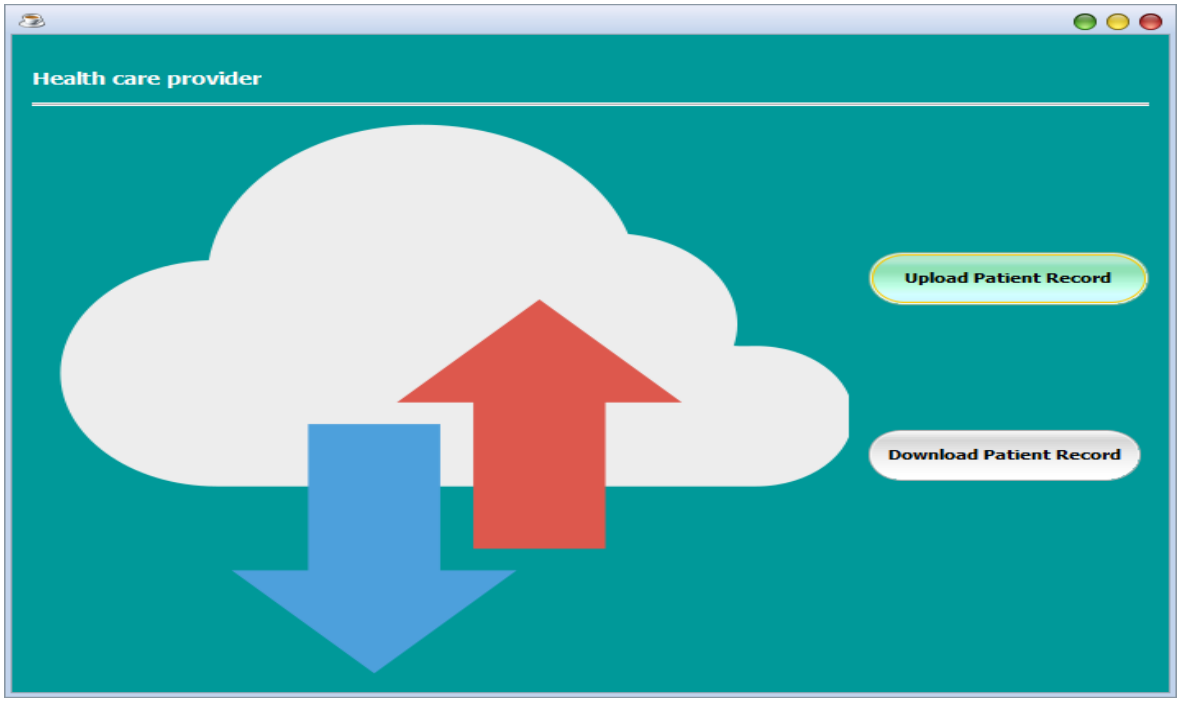


FIG 9.1: UPLOAD PATIENT RECORD

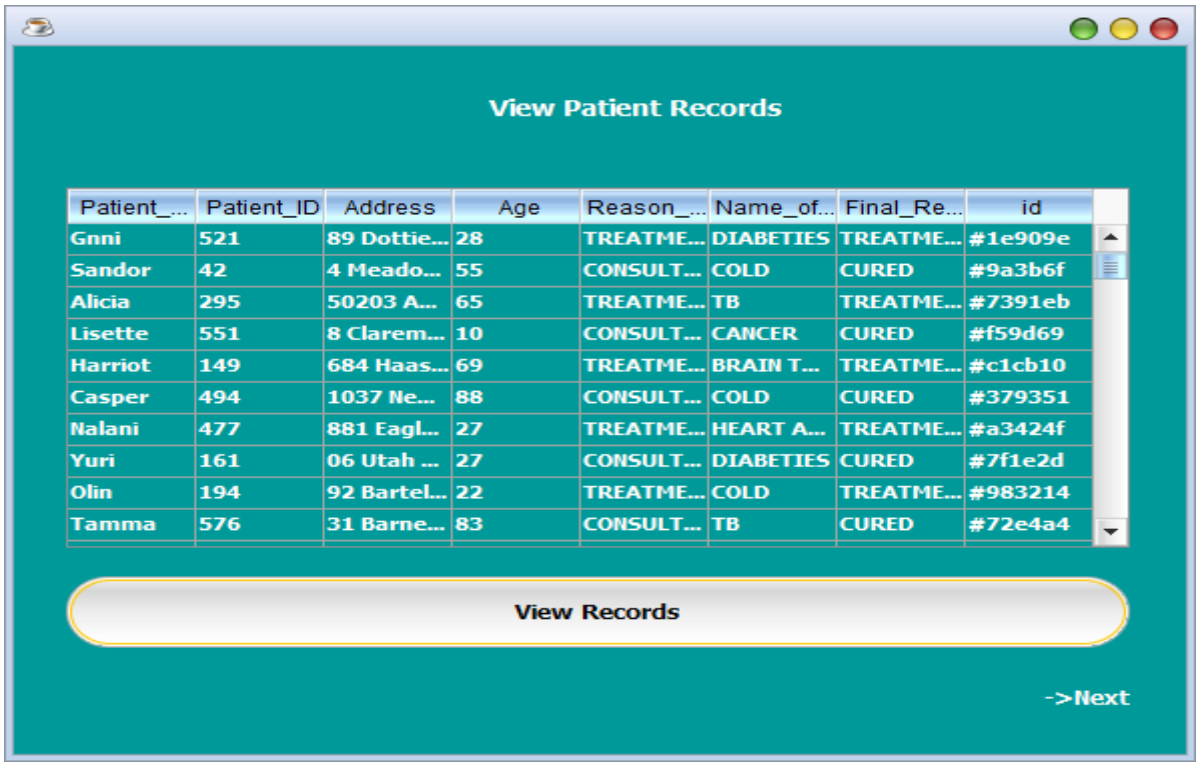


FIG 9.2: PATIENT RECORD

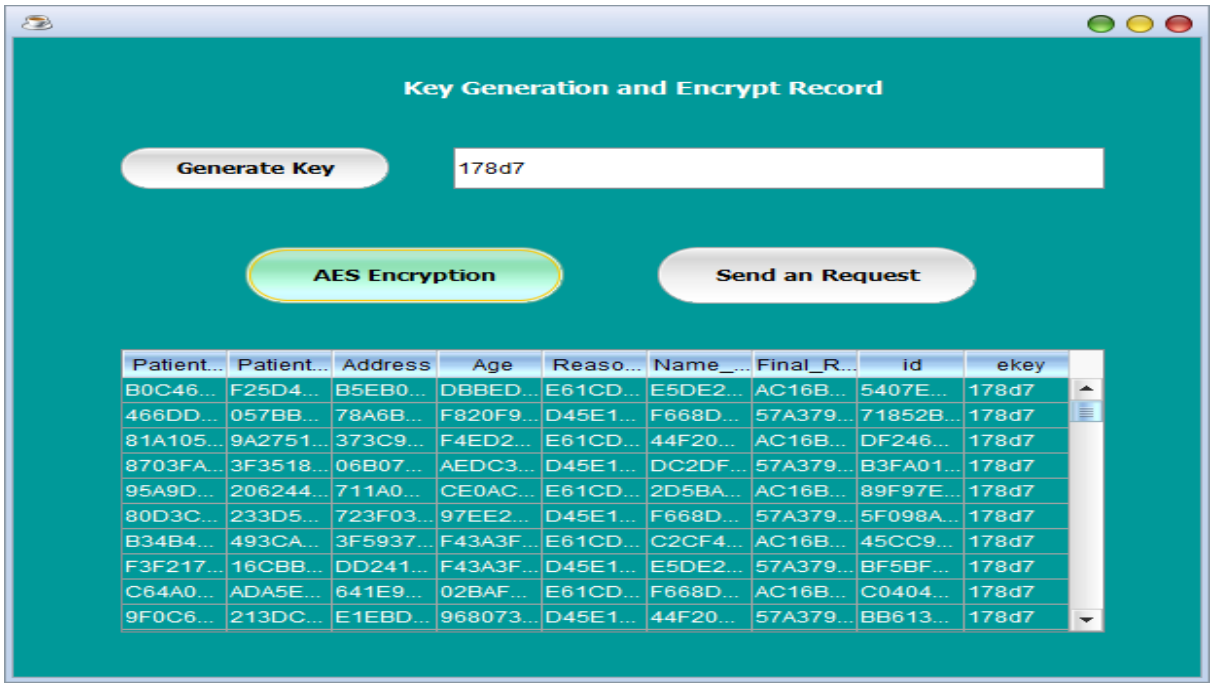


FIG 9.3: KEY GENERATION AND ENCRYPT RECORD

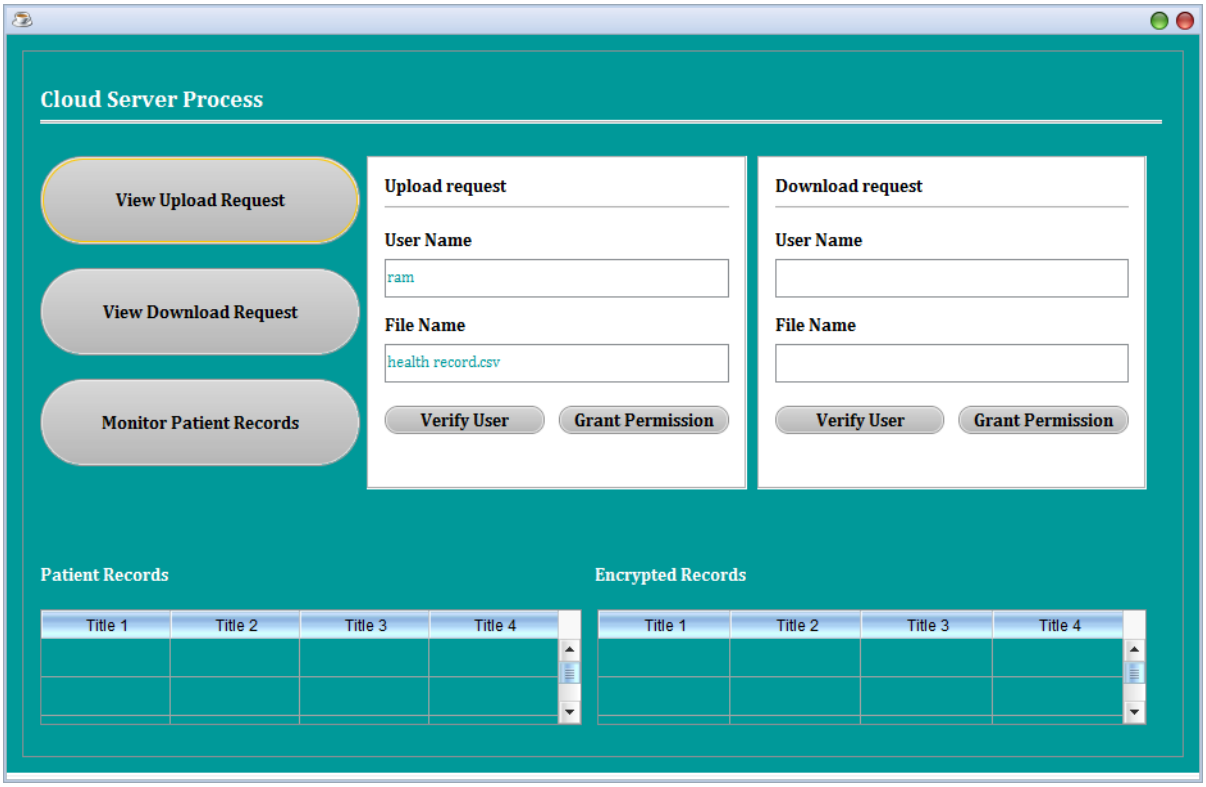


FIG 9.4: CLOUD SERVER PROCESS

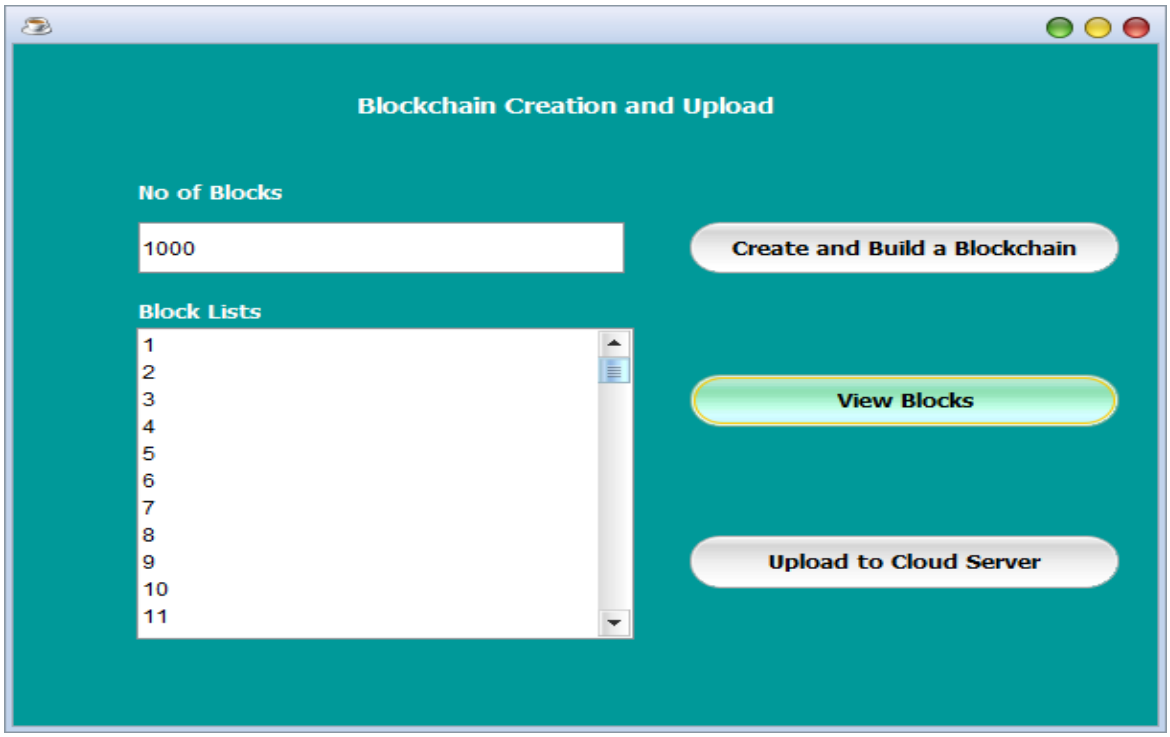


FIG 9.5: BLOCKCHAIN CREATION

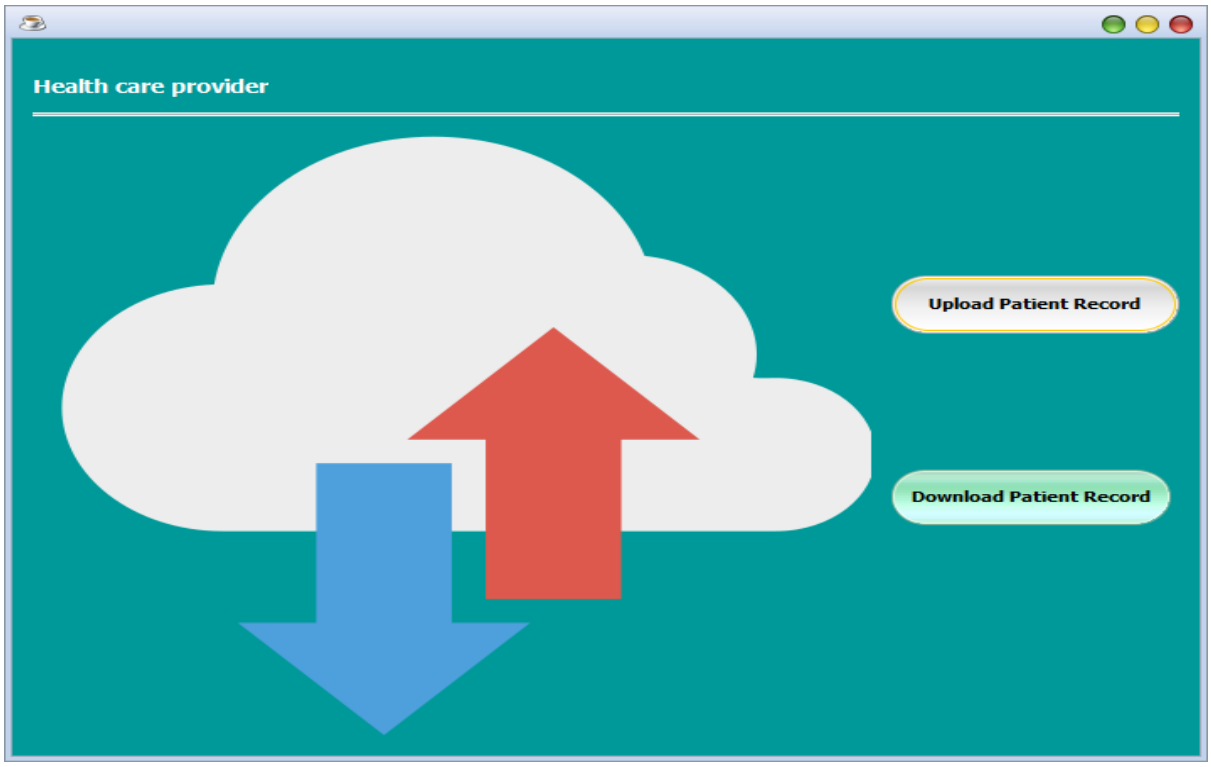


FIG 9.6: DOWNLOAD PATIENT RECORD

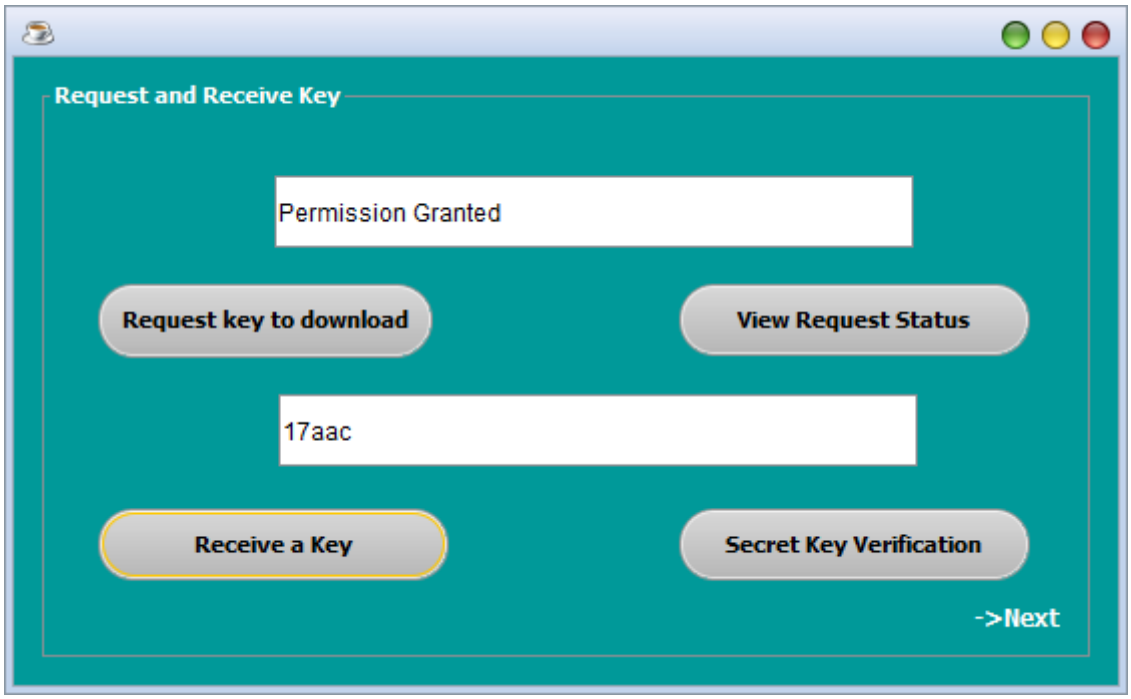


FIG 9.7: SECRET KEY VERIFICATION

FIG 9.8: SIGNING COST

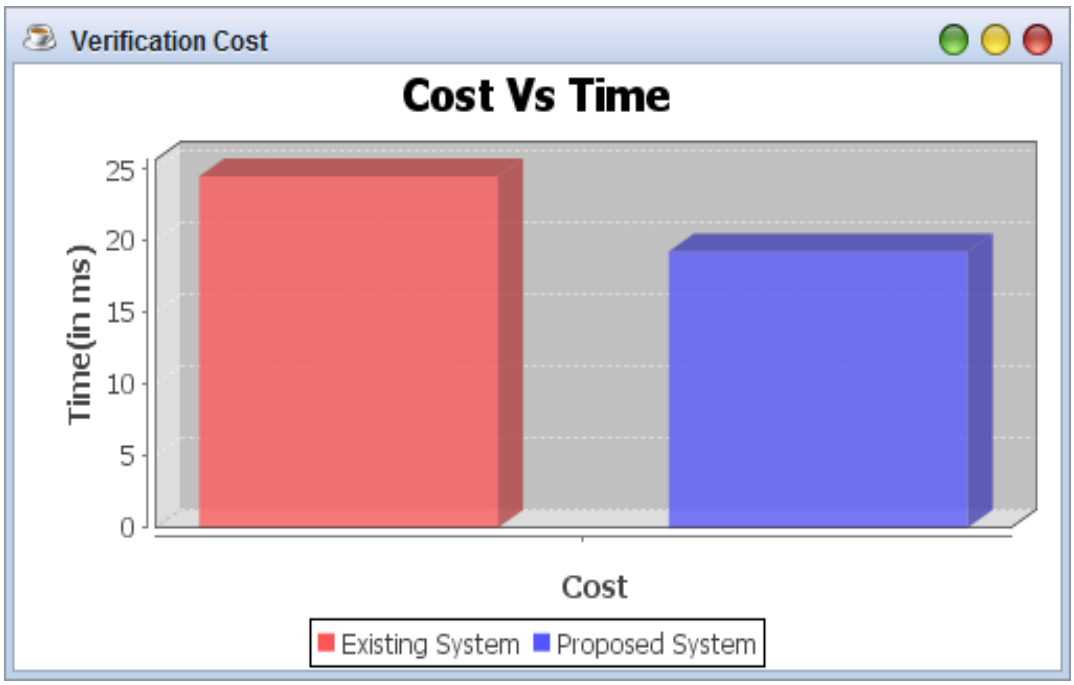


FIG 9.9:VERIFICATION COST

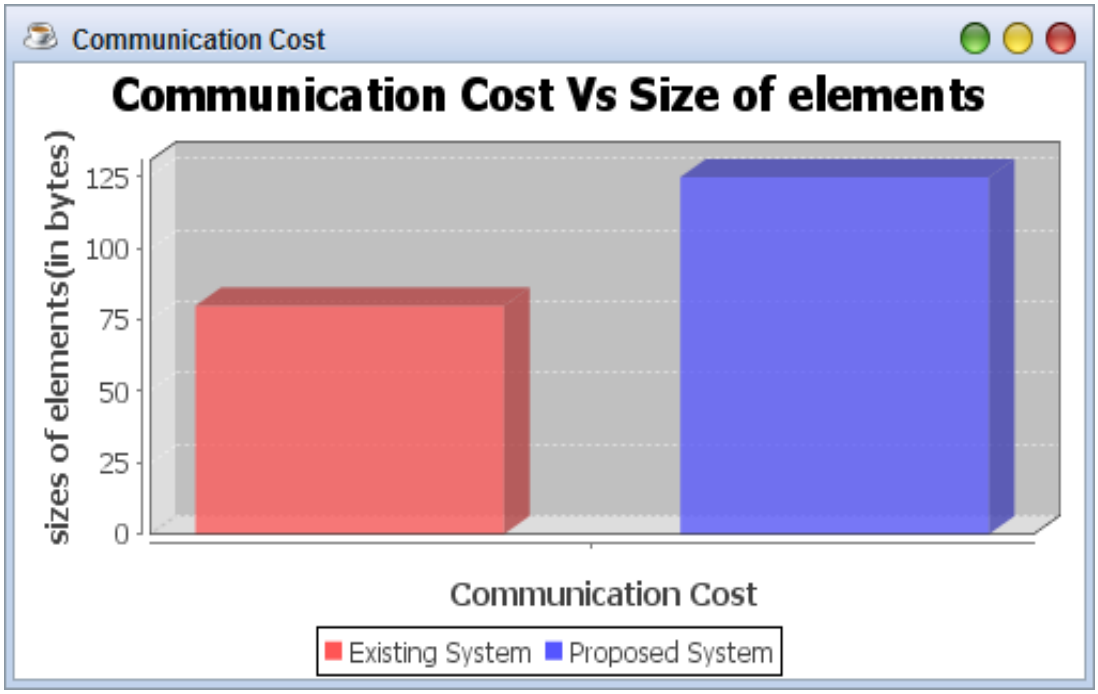


FIG 9.10:COMMUNICATION COST

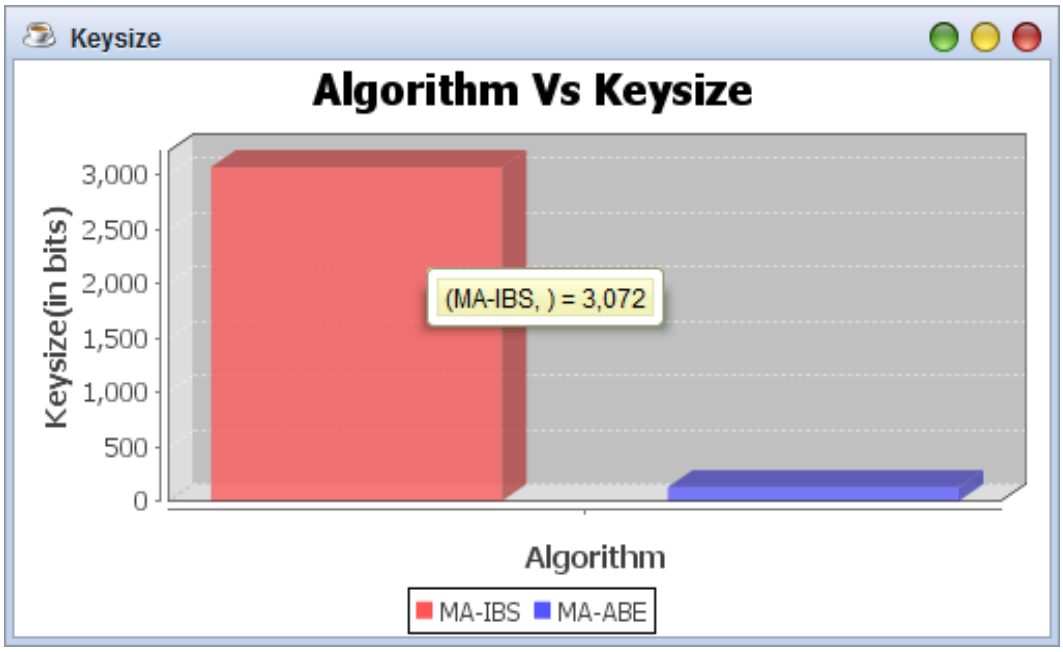


FIG 9.11:KEYSIZE

10. CONCLUSION AND FUTURE ENHANCEMENT

It’s more securely maintain all the patient records and it will be easily accessible by any healthcare providers. By building block chain, it provides efficient search result verification, while preventing data freshness attacks and data integrity attacks. We achieved the results with the help of cloud service process (CSP). The CSP receives the document request from the Data User, verifies the authentication before granting permission.

Here, there is no possibility for the user to view their electronic medical records (EMR) as they cannot access the healthcare data. In addition to proposed system, we can extend this project by enabling the users to view their own medical records with their specific access token generated by the healthcare provider.

REFERENCES

[1] Yunru Zhang, Debiao He, and Kim-Kwang Raymond Choo, "BaDS: Blockchain-Based Architecture for Data Sharing with ABS and CP-ABE in IoT," Wireless Commu. and Mobile Comput., 2018.

[2] Jiawen Kang, Rong Yu, Xumin Huang, Maoqiang Wu, Sabita Maharjan, Shengli Xie, and Yan Zhang "Blockchain for Secure and Efficient Data Sharing in Vehicular Edge Computing and Networks," IEEE Internet of Things J., 2018.

[3] Oscar Novo, "Blockchain Meets IoT: An Architecture for Scalable Access Management in IoT," IEEE Internet of Things J., vol. 5, pp. 1184-1195, 2018.

[4] Kuo TT, Kim HE, and Ohno-Machado L, "Blockchain distributed ledger technologies for biomedical and health care applications," Ame. Medi. Infor. Assoc. J., vol. 6, pp. 1211-1220, 2020.

[5] Nabil Rifi, Elie Rachkidi, Nazim Agoulmine, and Nada Chendeb Taher, "Towards Using Blockchain Technology for eHealth Data Access Management,"in Proc. IEEE on Advances in Bio. Engi., Oct. 2019.

[6] S.H. Han et al., “Implementation of Medical Information Exchange System Based on EHR Standard” 2020.

[7] D. He et al., “A Provably-Secure Cross-Domain Handshake Scheme with Symptoms-Matching for Mobile Healthcare Social Network,” IEEE Transactions on Dependable and Secure Computing, 2017

[8] F.Y. Leu et al., “A Smartphone-Based Wearable Sensors for Monitoring Real-Time Physiological Data,” Computers and Electrical Engineering, 2018.

[9] M. Memon et al., “Ambient Assisted Living Healthcare Frameworks, Platforms, Standards, and Quality Attributes”, 2018.

[10] P.C. Tang et al., “Personal Health Records: Definitions, Benefits, and Strategies for Overcoming Barriers to Adoption” 2019.

[11] S. Marceglia et al., “A Standards-Based Architecture Proposal for Integrating Patient Health Apps to Electronic Health Record Systems” Applied Clinical Informatics, 2019.

[12] A. Mu-Hsing Kuo, “Opportunities and Challenges of Cloud Computing to Improve Health Care Services” Journal of Medical Internet Research, 2017.

[13] V. Casola et al., “Healthcare-Related Data in the Cloud: Challenges and Opportunities” IEEE Cloud Computing, 2019.

[14] S. Nepal et al., “Trustworthy Processing of Healthcare Big Data in Hybrid Clouds” IEEE Cloud Computing, 2020.

[15] G.S. Poh et al., “Searchable Symmetric Encryption: Designs and Challenges” 2020.

PUBLICATIONS

SMART CONTRACT BASED ACCESS CONTROL FOR HEALTH CARE DATA

Chitra M[1] , Gudala SaiDatta[2], Odnam Chakradhar [3], Vutukotu Sandhya[4], Manu Hajari[5]

[1,2,3,4,] UG Scholar, [5] Assisstant Professor

Department of Cpmputer Science And Engineering

St.Martin's Engineering College, Near Forest Academy, Dhulapally , Secunderabad , Telangana, India-500100, India

Email-chitra.medipally@gmail.com[1] , chakra0500@gmail.com[2] , saidatta8055@gmail.com[3] , sandhya.v531@gmail.com[4] ,manuhajari @gmail.com[5]

Abstract:

The main objective of this project is securely store and maintain the patient records in the healthcare. Healthcare is a data-intensive domain where a large amount of data is created, disseminated, stored, and accessed daily. The blockchain technology is used to protect the healthcare data hosted within the cloud. The block that contain the medical data and the timestamp. Cloud computing will connect different healthcare providers. It allows healthcare provider to access the patient details more securely from anywhere. It preserve data from attackers. The data is encrypted prior to outsourcing to the cloud. The healthcare provider have to decrypt the data prior to download.

Keywords: Blockchain; Cloud computing; Data sharing ;Health record; Security;Timestamp

I. Introduction:

Cloud computing offers an opportunity for individuals and companies to offload to powerful servers the burden of managing large amounts of data and performing computationally demanding operations. Due to the increasing popularity of cloud computing, more and more Data owners are motivated to outsource their data to cloud servers for great convenience and reduced cost in data management. Data owners offer services to a large number of businesses and companies, they stick to high security standards to improve data security by following a layered approach that includes data encryption, key management, strong access controls, and security intelligence. To facilitate data sharing or even patient data portability, there is a need for EMRs to formalize their data structure and the design of HIS. Electronic Health Records (EHRs), for example, are designed to allow patient medical history to move with the patient or be made available to

multiple healthcare providers EHRs have a richer data structure than EMRs. Recently, the pervasiveness of smart devices has also resulted in a paradigm shift within the healthcare industry. Such devices can be user-owned or installed by the healthcare provider to measure the well-being of the users (e.g.patients) and inform/facilitate medical treatment and monitoring of patients. For example, there is a wide range of mobile applications in health, fitness, weight-loss, and other healthcare related categories. These apps mainly function as a tracking tool, such as registering user exercises/workouts, keeping the count of consumed calories, and other statistics and so on.There are also devices with embedded sensors for more advanced medical tasks, such as bracelets to measure heartbeat during workouts, or devices for self-testing of glucose. The data can be continuously gathered and sent in real-time to a smart device, before being sent to a remote healthcare cloud for further analysis. These developments have paved the way for Personal Health Records (PHR), where patients are more involved in their data collection, monitoring of their health conditions, etc, using their smart phones or wearable devices.Blockchain was originally designed to record transaction data, which is relatively small in size and linear.

II. Literature Survey

1. BaDS: Blockchain-Based Architecture for Data Sharing with ABS and CP-ABE in IoT

Author: Yunru Zhang, Debiao He, and Kim-Kwang Raymond Choo

Internet of Things (IoT) and cloud computing are increasingly integrated, in the sense that data collected from IoT devices (generally with limited computational and storage resources) are being sent to the cloud for processing, etc., We proposed a novel blockchain-based architecture for data sharing with attribute-based cryptosystem (BaDS) in this paper. The architecture can achieve privacy-preserving, user-self-controlled data sharing, and decentralization by using blockchain and several attribute-based cryptosystems. Specifically, ABS and CP-ABE provide the capability for fine-grained access control. We introduced the security requirements of the proposed BaDS architecture and then explained how the proposed BaDS architecture satisfies the security requirement. We also implement the BaDS architecture and analyze its computation cost.

Advantages

- Implementing digital signatures.
- Cryptographic protocols with different security and privacy features.
- Supporting various signature schemes without adding additional hardware complexity compared to a hardware implementation of a conventional signature scheme.

Disadvantages

- Encryption keys aren't simple strings of text like passwords
- **Damage** is massive when you lost your symmetric key

2. Blockchain for Secure and Efficient Data Sharing in Vehicular Edge Computing and Networks

Author: Jiawen Kang, Rong Yu, Xumin Huang, Maoqiang Wu, Sabita Maharjan, Shengli Xie, and Yan Zhang

The drastically increasing volume and the growing trend on the types of data have brought in the possibility of realizing advanced applications such as enhanced driving safety, and have enriched existing vehicular services through data sharing among vehicles and data analysis. We exploit consortium blockchain and smart contract technologies to achieve secure data storage and sharing in vehicular edge networks. These technologies efficiently prevent data sharing without authorization. In addition, we propose a reputation-based data sharing scheme to ensure high-quality data sharing among vehicles. A three-weight subjective logic model is utilized for precisely managing reputation of the vehicles. Numerical results based on a real dataset show that our schemes achieve reasonable efficiency and high-level of security for data sharing in VECONs.

Advantages

- Security against adaptive chosen-keyword attacks.
- Compact indexes.
- Ability to add and delete files efficiently.

Disadvantages

- Every means of electronic communication is insecure as it is impossible to guarantee that no one will be able to tap communication channels. So the only secure way of exchanging keys would be exchanging them personally.

3. Blockchain Meets IoT: An Architecture for Scalable Access Management in IoT

Author: Oscar Novo

The Internet of Things (IoT) is stepping out of its infancy into full maturity and establishing itself as a part of the future Internet. One of the technical challenges of having billions of devices deployed worldwide is the ability to manage them. Although access management technologies exist in IoT, they are based on centralized models which introduce a new variety of technical limitations to manage them globally. In this

paper, we propose a new architecture for arbitrating roles and permissions in IoT. The new architecture is a fully distributed access control system for IoT based on blockchain technology. The architecture is backed by a proof of concept implementation and evaluated in realistic IoT scenarios. The results show that the blockchain technology could be used as access management technology in specific scalable IoT scenarios.

Advantages

- Providing performance results of a prototype applied to several large representative data sets, including encrypted search over the whole English Wikipedia.

Disadvantages

- Exact matching may retrieve too few or too many documents.

4**.**Blockchain distributed ledger technologies for biomedical and health care applications

Author: Kuo TT, Kim HE, and Ohno-Machado L

Blockchain is a distributed, immutable ledger technology introduced as the enabling mechanism to support cryptocurrencies. Blockchain solutions are currently being proposed to address diverse problems in different domains. This paper presents a scoping review of the scientific literature to map the current research area of blockchain applications in the biomedical domain. The goal is to identify biomedical problems treated with blockchain technology, the level of maturity of respective approaches, types of biomedical data considered, blockchain features and functionalities exploited and blockchain technology frameworks used. Our findings show that the field is still in its infancy, with the majority of studies in the conceptual or architectural design phase; only one study reports real world demonstration and evaluation. Research is greatly focused on integration, integrity and access control of health records and related patient data. However, other diverse and interesting applications are emerging, addressing medical research, clinical trials, medicines supply chain, and medical insurance.

Advantages

- Complete expressiveness for any identifiable subset of collection.
- A symmetric cryptosystem uses password authentication to prove the receiver's identity

Disadvantages

- Cannot provide digital signatures that cannot be repudiated

5. Towards Using Blockchain Technology for ehealth Data Access Management

Author: Nabil Rifi, Elie Rachkidi, Nazim Agoulmine, and Nada Chendeb Taher

ehealth is a technology that is growing in importance over time, varying from remote access to Medical Records, such as Electronic Health Records (EHR), or Electronic Medical Records (EMR), to real-time data exchange from different on-body sensors coming from different patients. With this huge amount of critical data being exchanged, problems and challenges arise. Privacy and confidentiality of this critical medical data are of high concern to the patients and authorized persons to use this data. On the other hand, scalability and interoperability are also important problems that should be considered in the final solution. This paper illustrates the specific problems and highlights the benefits of the blockchain technology for the deployment of a secure and a scalable solution for medical data exchange in order to have the best performance possible.

Advantage:

- Efficient data search

Disadvantage:

- Data Integrity Problem

III. Proposed Methodology:

To overcome the security problems that are occurred in the existing system and effectively store the data over the cloud we introduce this system.The data user outsources the encrypted documents to the cloud.The Data user get the each result, the proof and the public verification key, they itself or others can verify the freshness, authenticity, and completeness of the search result even without decrypting them.

ADVANTAGES

- Efficient Search Result.
- Prevents data freshness attacks and data integrity attacks.
- It provides High Security.
- Files can be easily updated.

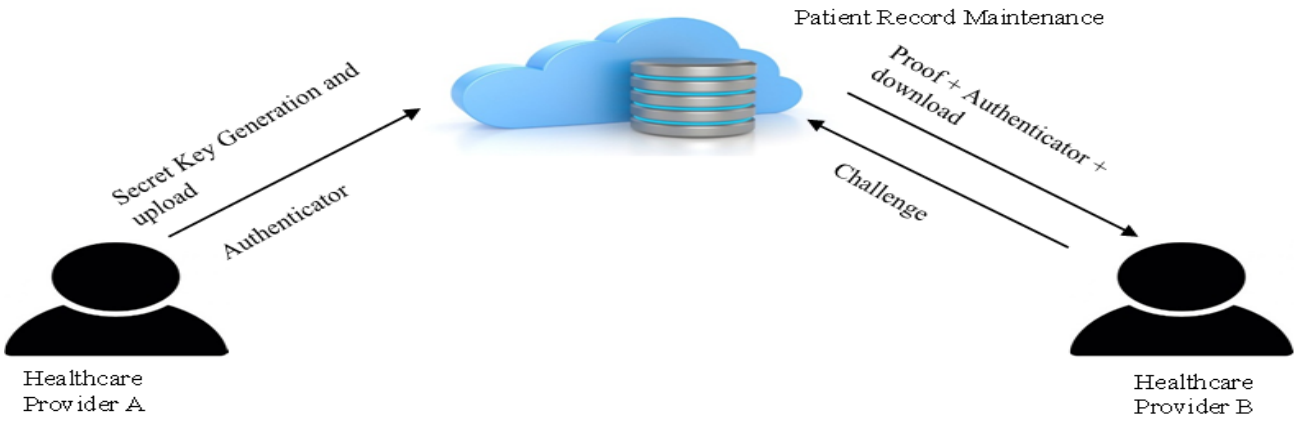


Fig 1: The architecture of proposed method.

MODULES

Registration

Healthcare Provider

- Load patient Records
- Key Generation
- Encrypt patient Records
- Block Creation
- Upload and Download Patient Records

Cloud Service Provider

- View Patient Records
- Grant or Revoke Permission

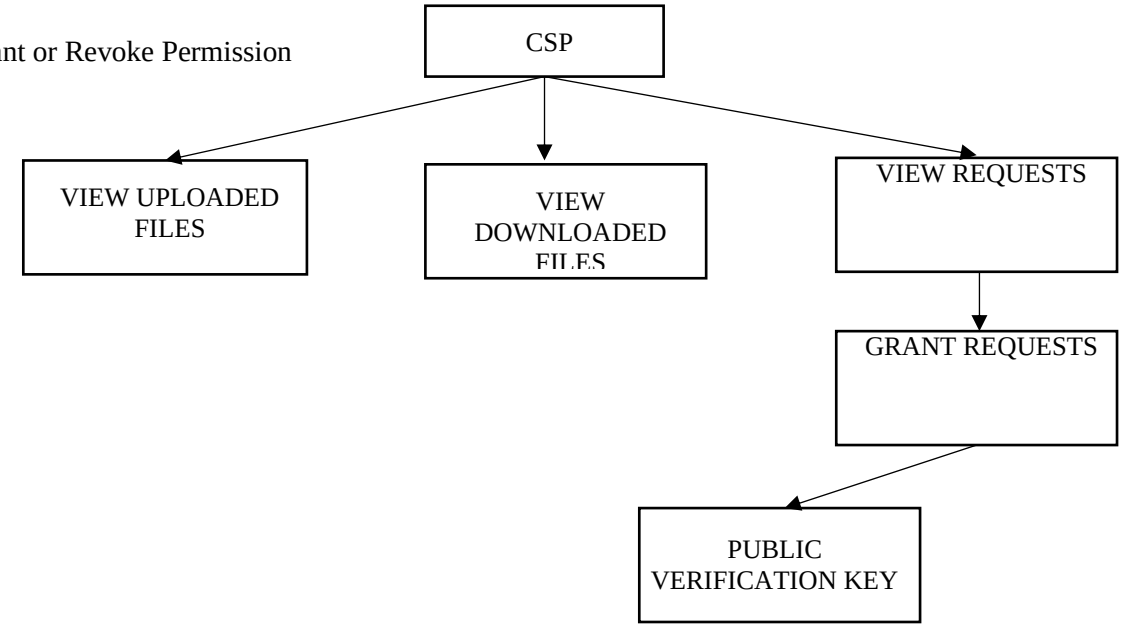


Fig.2: Public key Verification

IV Result And Discussion

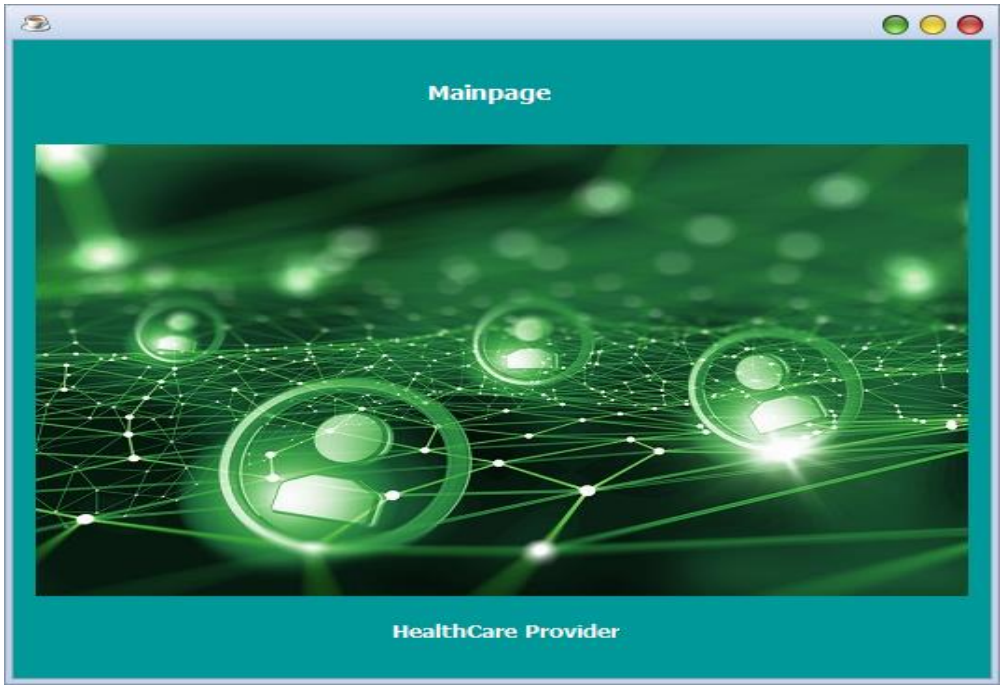


Fig 3:Home Page

Fig 4: Login Page

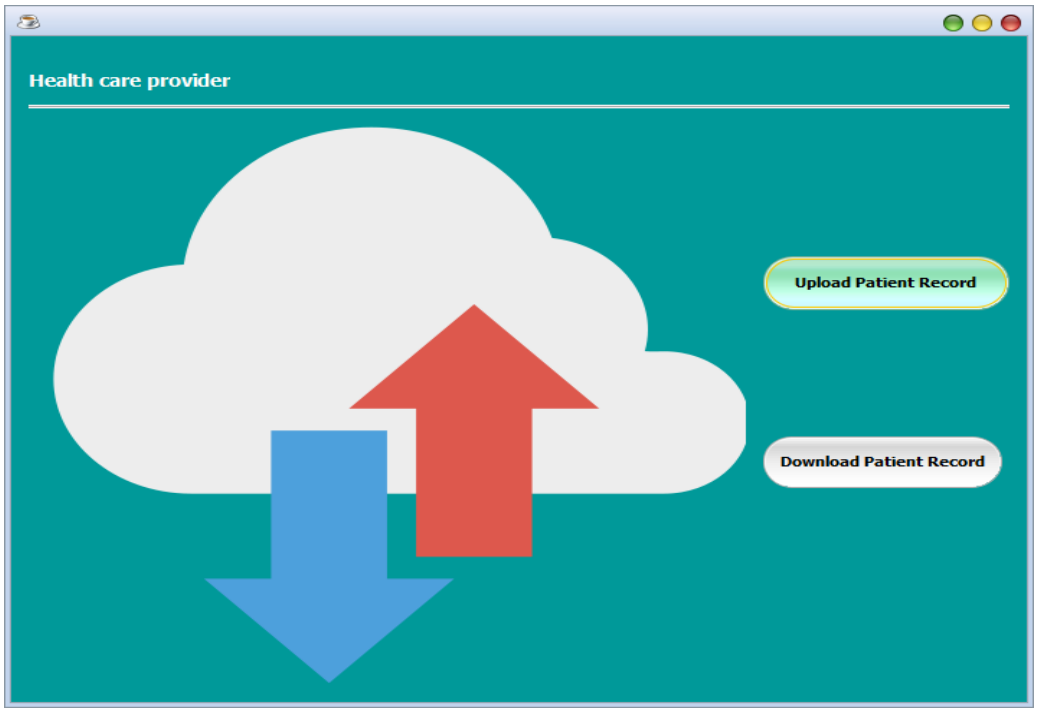


Fig :5 Upload and Download page

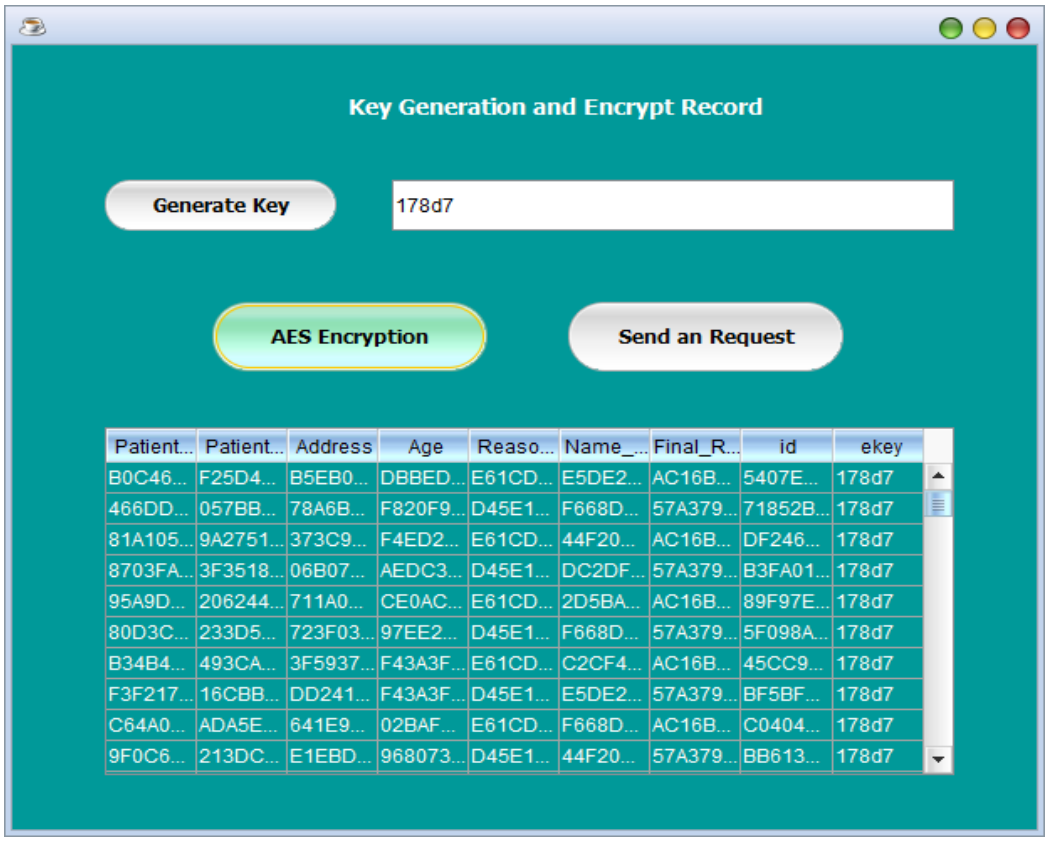


Fig :6 Encryption Page

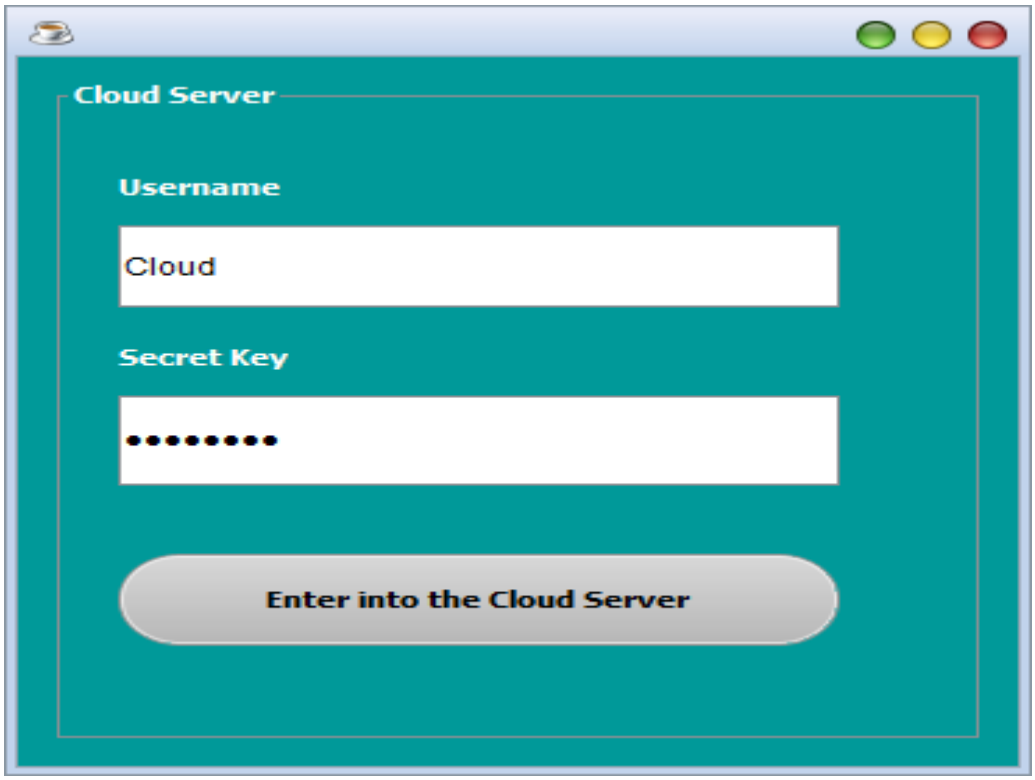


Fig 7: Cloud Login

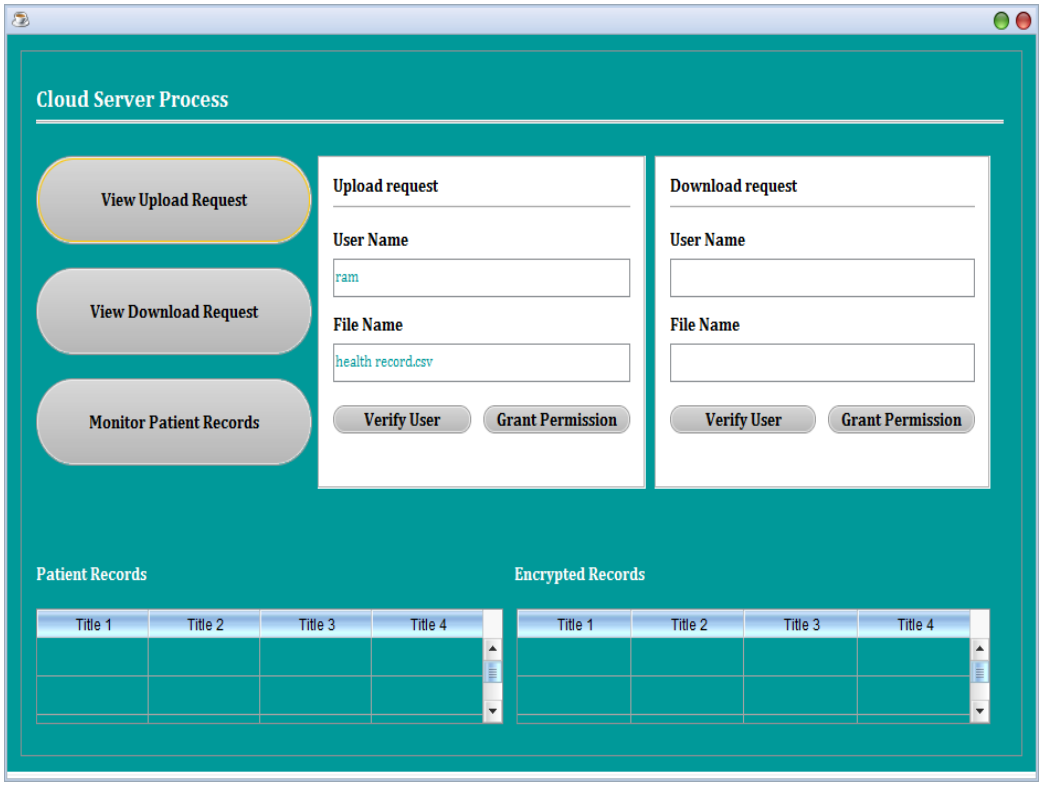


Fig 8: Cloud Server Process

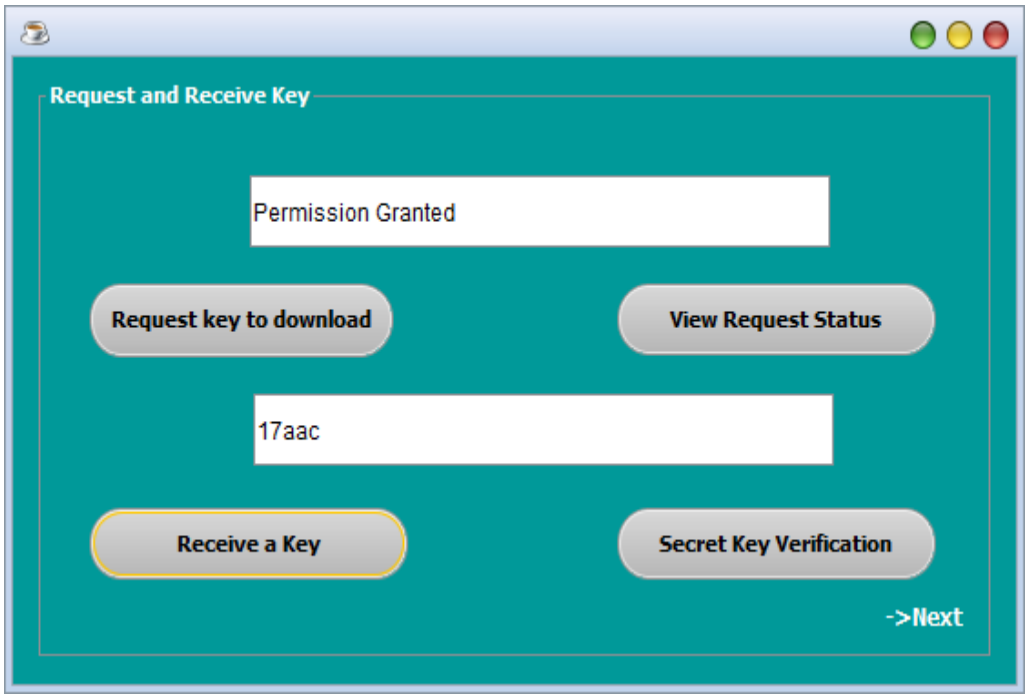


Fig 9: SecretKey Verification

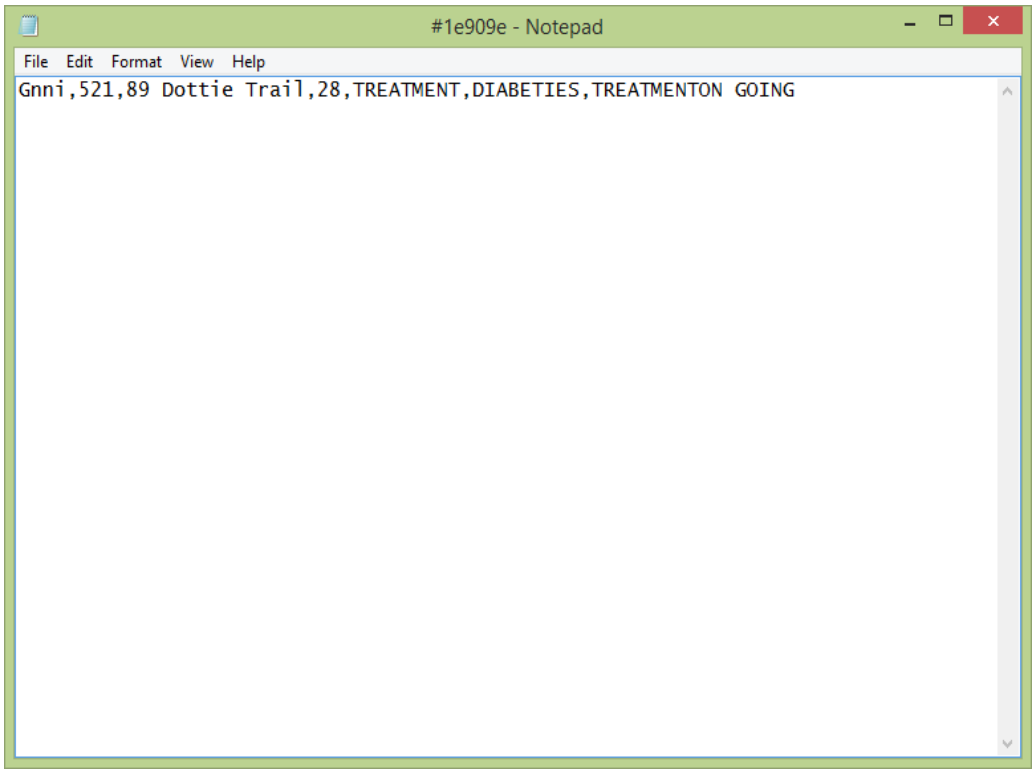


Fig 10: Download Page

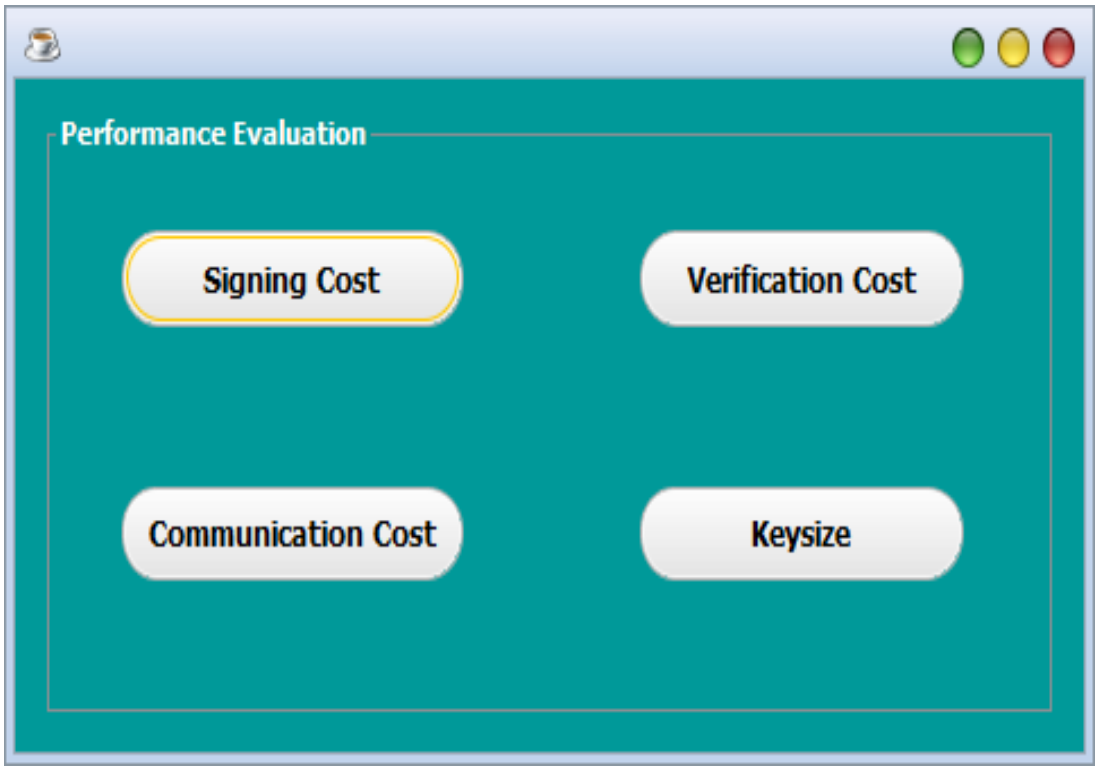


Fig 11: Performance Evaluation

IV. Conclusion:

It's more securely maintain all the patient records and it will be easily accessible by any healthcare providers.By building block chain, it provides efficient search result verification, while preventing data freshness attacks and data integrity attacks in SSE. Data Classification based on Security: A cloud computing data center can store data from various users. To provide the level of security based on the importance of data, classification of data can be done. This classification scheme should consider various aspects like access frequency, update frequency and access by various entities etc. based on the type of data. Once the data is classified and tagged, then level of security associated with this specific tagged data element can be applied. Level of security includes confidentiality, encryption, integrity and storage etc. that are selected based on the type of data.

V. References

[1] Yunru Zhang, Debiao He, and Kim-Kwang Raymond Choo, "BaDS: Blockchain-Based Architecture for Data Sharing with ABS and CP-ABE in IoT," Wireless Commu. and Mobile Comput., 2018.

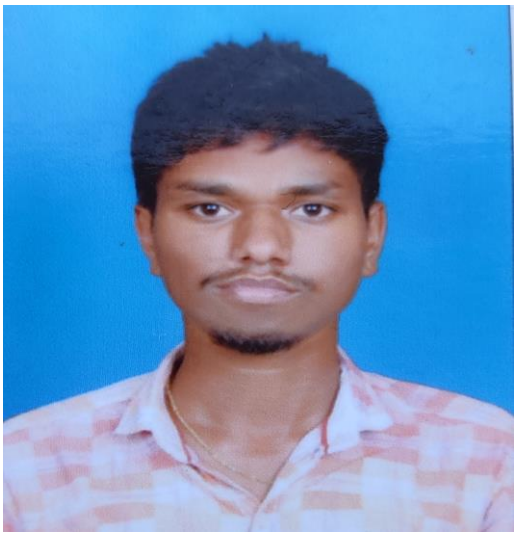
[2] Jiawen Kang, Rong Yu, Xumin Huang, Maoqiang Wu, Sabita Maharjan, Shengli Xie, and Yan Zhang "Blockchain for Secure and Efficient Data Sharing in Vehicular Edge Computing and Networks," IEEE Internet of Things J., 2018.

[3] Oscar Novo, "Blockchain Meets IoT: An Architecture for Scalable Access Management in IoT," IEEE Internet

of Things J., vol. 5, pp. 1184-1195, 2018.
[4] Kuo TT, Kim HE, and Ohno-Machado L, "Blockchain distributed ledger technologies for biomedical and health care applications," Ame. Medi. Infor. Assoc. J., vol. 6, pp. 1211-1220, 2020.
[5] Nabil Rifi, Elie Rachkidi, Nazim Agoulmine, and Nada Chendeb Taher, "Towards Using Blockchain Technology for eHealth Data Access Management,"in Proc. IEEE on Advances in Bio. Engi., Oct. 2019.
[6] S.H. Han et al., “Implementation of Medical Information Exchange System Based on EHR Standard” 2020.
[7] D. He et al., “A Provably-Secure Cross-Domain Handshake Scheme with Symptoms-Matching for Mobile Healthcare Social Network,” IEEE Transactions on Dependable and Secure Computing, 2017
[8] F.Y. Leu et al., “A Smartphone-Based Wearable Sensors for Monitoring Real-Time Physiological Data,” Computers and Electrical Engineering, 2018.
[9] M. Memon et al., “Ambient Assisted Living Healthcare Frameworks, Platforms, Standards, and Quality Attributes”, 2018.
[10] P.C. Tang et al., “Personal Health Records: Definitions, Benefits, and Strategies for Overcoming Barriers to Adoption” 2019.
[11] S. Marceglia et al., “A Standards-Based Architecture Proposal for Integrating PatientHealth Apps to Electronic Health Record Systems” Applied Clinical Informatics, 2019.
[12] A. Mu-Hsing Kuo, “Opportunities and Challenges of Cloud Computing to Improve Health Care Services” Journal of Medical Internet Research, 2017.
[13] V. Casola et al., “Healthcare-Related Data in the Cloud: Challenges and Opportunities” IEEE Cloud Computing, 2019.
[14] S. Nepal et al., “Trustworthy Processing of Healthcare Big Data in Hybrid Clouds” IEEE Cloud Computing, 2020.
[15] G.S. Poh et al., “Searchable Symmetric Encryption: Designs and Challenges” 2020.

ALL FOUR STUDENTS' ONE PAGE PROFILE

Chitra M is currently pursuing her Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. She completed her intermediate from Kasturba Gandhi Junior College and 10th class from Holy Family High School. Her technical skills include C, C++ and Java. She also has a basic understanding of Python. She is also a student of Smart Interviews where they have provided some different coding platforms like Hacker rank , Codechef ,Code Forces. Her participations include: Workshop on "HTML & CSS" of Technical Awareness Month event which was conducted on 5th January 2018 to 3rd February 2018, Technical Treasure hunt and Paper Presentation event during "SYMPO AAGNYA 2020"-A Two Day National Level Technical Symposium which was conducted on 30th and 31st January, National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020. Her areas of interest are Python, Artificial Intelligence, Machine Learning. She completed few certification courses from online platforms like Coursera, CursaApp and SoloLearn.



Gudala Saidatta is currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin’s Engineering College. He completed his intermediate from Narayana Junior College and 10th class from Jeevadan High School.His technical skills include C++ , Java and basic understanding of Python and C. His participations include:Two day National level seminar on "Recent trends in Cloud Computing, Fog and Edge computing" which was conducted on 18th and 19th June 2021,National Level Three Day Online Workshop on “AI & ML in Speech and Audio Processing” which was conducted from 10th to 12th December 2020, "Guinness World Record Event - Most users to take an online computer programming lesson in 24 hours " which was conducted by GUVI on 24th and 25th April 2021, IIC Online Sessions conducted by Institution's Innovation Council (IIC) of MHRD's Innovation Cell, New Delhi to promote Innovation, IPR, Entrepreneurship, and Start-ups among HEIs from 28th April to 22nd May 2020 and "Technovation-2018" Which was conducted by St.Martin's Engineering college in the year 2018. His areas of interest are Python and Web development. He completed few certification courses from online platforms like Udemy,coursera, CursaApp and SoloLearn.

Odnam Chakradhar is currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. He completed his intermediate from Narayana Junior College and 10th class from St Peter's High School .His technical skills include C++ , Java and basic understanding of Python and C. He took part in Employability Skill development Program conducted by Zensar. He is also a student of Smart Interviews where they have given some targets on coding in different platforms like Hacker rank , Codechef ,Code Forces. His participations include :Two day National level seminar on "Recent trends in Cloud Computing, Fog and Edge computing" which was conducted on 18thand 19thJune 2021,National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020 ,He had completed one month internship at Lasya IT Solution Pvt.Ltd , IIC Online Sessions conducted by Institution's Innovation Council (IIC) of MHRD's Innovation Cell, New Delhi to promote Innovation, IPR, Entrepreneurship, and Start-ups among HEIs from 28thApril to 22nd May 2020 . His areas of interest are Python ,Data Science and Cloud Computing. He completed few certification courses from online platforms like Udemy , Coursera and CursaApp .

Vutukotu Sandhya is currently pursuing her Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. She completed her intermediate from Narayana Junior College and 10th class from Tagore's Home High School. Her technical skills include C, C++ and Java. She also has a basic understanding of Python. She is also a student of Smart Interviews where they also provided some platforms to participate at codeforces, codechef . Her participations include: Workshop on "HTML & CSS" which was conducted on 5th January 2018, Technical Treasure hunt and Paper Presentation in "Sympo aagnya 2020-A Two Day National Level Technical Symposisum" Which was conducted on 30th and 31st January, National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020, Online Sessions conducted by Institution's Innovation Council (IIC) of MHRD's Innovation Cell, New Delhi to promote Innovation, IPR, Entrepreneurship, and Start-ups among HEIs from 28th April to 22nd May 2020. Her areas of interest are Python, Artificial Intelligence, Machine Learning. She completed few certification courses from online platforms like Coursera, CursaApp and SoloLearn.

APPENDICES

```
package block_chain;

import static block_chain.upload.filepath;

import static block_chain.upload.nor;

import java.io.BufferedReader;

import java.io.File;

import java.io.FileReader;

import java.io.FileWriter;

import java.util.Date;

import java.sql.Timestamp;

import javax.swing.JOptionPane;

import java.awt.*;

import java.applet.*;

import java.io.IOException;

import java.util.logging.Level;

import java.util.logging.Logger;

import javax.swing.DefaultListModel;

/**

*

* @author egc

*/

public class block_Creation extends javax.swing.JFrame {

/**

* Creates new form block_Creation

*/
```

```
public block_Creation() {
initComponents();
this.setResizable(false);
this.setLocationRelativeTo(null);
}
/**
* This method is called from within the constructor to initialize the form.
* WARNING: Do NOT modify this code. The content of this method is always
* regenerated by the Form Editor.
*/
@SuppressWarnings("unchecked")
// <editor-fold defaultstate="collapsed" desc="Generated Code">
private void initComponents() {
jPanel13 = new javax.swing.JPanel();
jPanel14 = new javax.swing.JPanel();
jLabel10 = new javax.swing.JLabel();
jPanel15 = new javax.swing.JPanel();
jLabel11 = new javax.swing.JLabel();
jSeparator5 = new javax.swing.JSeparator();
jButton6 = new javax.swing.JButton();
jButton7 = new javax.swing.JButton();
jButton8 = new javax.swing.JButton();
jPanel1 = new javax.swing.JPanel();
jLabel13 = new javax.swing.JLabel();
jTextField1 = new javax.swing.JTextField();
```

```
jLabel14 = new javax.swing.JLabel();
jScrollPane1 = new javax.swing.JScrollPane();
jList1 = new javax.swing.JList<>();
jLabel12 = new javax.swing.JLabel();

setDefaultCloseOperation(javax.swing.WindowConstants.EXIT_ON_CLOSE);
jPanel13.setBackground(new java.awt.Color(153, 102, 255));
jPanel14.setBackground(new java.awt.Color(204, 204, 255));
jLabel10.setFont(new java.awt.Font("Cambria", 1, 24)); // NOI18N
jLabel10.setText("Blockchain: A Panacea for Healthcare Cloud-Based Data");
jPanel15.setBackground(new java.awt.Color(255, 255, 255));
jLabel11.setFont(new java.awt.Font("Cambria", 1, 18)); // NOI18N
jLabel11.setForeground(new java.awt.Color(102, 0, 255));
jLabel11.setText("Request Status and Upload");
jButton6.setBackground(new java.awt.Color(153, 102, 255));
jButton6.setFont(new java.awt.Font("Cambria", 1, 14)); // NOI18N
jButton6.setText("Upload");
jButton6.addActionListener(new java.awt.event.ActionListener() {
public void actionPerformed(java.awt.event.ActionEvent evt) {
jButton6ActionPerformed(evt);
}
});
jButton7.setBackground(new java.awt.Color(153, 102, 255));
jButton7.setFont(new java.awt.Font("Cambria", 1, 14)); // NOI18N
jButton7.setText("Create and Bulid a Block Chain");
```

```
jButton7.addActionListener(new java.awt.event.ActionListener() {
public void actionPerformed(java.awt.event.ActionEvent evt) {
jButton7ActionPerformed(evt);
}
});
jButton8.setBackground(new java.awt.Color(153, 102, 255));
jButton8.setFont(new java.awt.Font("Cambria", 1, 14)); // NOI18N
jButton8.setText("View Blocks");
jButton8.addActionListener(new java.awt.event.ActionListener() {
public void actionPerformed(java.awt.event.ActionEvent evt) {
jButton8ActionPerformed(evt);
}
});
jPanel1.setBackground(new java.awt.Color(255, 255, 255));
jPanel1.setBorder(javax.swing.BorderFactory.createEtchedBorder());
jLabel13.setFont(new java.awt.Font("Cambria", 1, 14)); // NOI18N
jLabel13.setText("Block List");
jTextField1.setFont(new java.awt.Font("Cambria", 0, 14)); // NOI18N
jLabel14.setFont(new java.awt.Font("Cambria", 1, 14)); // NOI18N
jLabel14.setText("Number of Blocks");
jList1.setFont(new java.awt.Font("Cambria", 0, 14)); // NOI18N
jList1.addListSelectionListener(new javax.swing.event.ListSelectionListener() {
public void valueChanged(javax.swing.event.ListSelectionEvent evt) {
jList1ValueChanged(evt);
}
```

```
});
jScrollPane1.setViewportView(jList1);
javax.swing.GroupLayout jPanel1Layout = new javax.swing.GroupLayout(jPanel1);
jPanel1.setLayout(jPanel1Layout);
jPanel1Layout.setHorizontalGroup(
jPanel1Layout.createParallelGroup(javax.swing.GroupLayout.Alignment.LEADING)
.addGroup(jPanel1Layout.createSequentialGroup()
.addGap(18, 18, 18)
.addGroup(jPanel1Layout.createParallelGroup(javax.swing.GroupLayout.Alignment.LEADING, false)
.addComponent(jLabel14)
.addComponent(jLabel13)
.addComponent(jTextField1)
.addComponent(jScrollPane1, javax.swing.GroupLayout.DEFAULT_SIZE, 504, Short.MAX_VALUE))
.addContainerGap(javax.swing.GroupLayout.DEFAULT_SIZE, Short.MAX_VALUE))
);
jPanel1Layout.setVerticalGroup(
jPanel1Layout.createParallelGroup(javax.swing.GroupLayout.Alignment.LEADING)
.addGroup(jPanel1Layout.createSequentialGroup()
.addContainerGap()
.addComponent(jLabel14)
.addPreferredGap(javax.swing.LayoutStyle.ComponentPlacement.RELATED)
.addComponent(jTextField1, javax.swing.GroupLayout.PREFERRED_SIZE, 35,
javax.swing.GroupLayout.PREFERRED_SIZE)
.addGap(18, 18, 18)
.addComponent(jLabel13)
```

```
.addPreferredGap(javax.swing.LayoutStyle.ComponentPlacement.RELATED)
.addComponent(jScrollPane1, javax.swing.GroupLayout.PREFERRED_SIZE, javax.swing.GroupLayout.DEFAULT_SIZE, javax.swing.GroupLayout.PREFERRED_SIZE)
.addContainerGap(14, Short.MAX_VALUE))
);
javax.swing.GroupLayout jPanel15Layout = new javax.swing.GroupLayout(jPanel15);
jPanel15.setLayout(jPanel15Layout);
jPanel15Layout.setHorizontalGroup(
jPanel15Layout.createParallelGroup(javax.swing.GroupLayout.Alignment.LEADING)
.addGroup(jPanel15Layout.createSequentialGroup()
.addGap(26, 26, 26)
.addGroup(jPanel15Layout.createParallelGroup(javax.swing.GroupLayout.Alignment.LEADING, false)
.addComponent(jButton6, javax.swing.GroupLayout.DEFAULT_SIZE, 547, Short.MAX_VALUE)
.addComponent(jPanel1, javax.swing.GroupLayout.DEFAULT_SIZE, javax.swing.GroupLayout.DEFAULT_SIZE, Short.MAX_VALUE)
.addComponent(jLabel11)
.addComponent(jSeparator5, javax.swing.GroupLayout.DEFAULT_SIZE, 547, Short.MAX_VALUE)
.addComponent(jButton8, javax.swing.GroupLayout.DEFAULT_SIZE, javax.swing.GroupLayout.DEFAULT_SIZE, Short.MAX_VALUE)
.addComponent(jButton7, javax.swing.GroupLayout.DEFAULT_SIZE, 547, Short.MAX_VALUE))
.addContainerGap(30, Short.MAX_VALUE))
);
jPanel15Layout.setVerticalGroup(
jPanel15Layout.createParallelGroup(javax.swing.GroupLayout.Alignment.LEADING)
.addGroup(jPanel15Layout.createSequentialGroup()
.addGap(25, 25, 25)
```

```
.addComponent(jLabel11)

.addPreferredGap(javax.swing.LayoutStyle.ComponentPlacement.RELATED)

.addComponent(jSeparator5, javax.swing.GroupLayout.PREFERRED_SIZE, 10, javax.swing.GroupLayout.PREFERRED_SIZE)

.addGap(18, 18, 18)

.addComponent(jButton7, javax.swing.GroupLayout.PREFERRED_SIZE, 36, javax.swing.GroupLayout.PREFERRED_SIZE)

.addGap(18, 18, 18)

.addComponent(jButton8, javax.swing.GroupLayout.PREFERRED_SIZE, 36, javax.swing.GroupLayout.PREFERRED_SIZE)

.addGap(18, 18, 18)

.addComponent(jPanel1, javax.swing.GroupLayout.PREFERRED_SIZE, javax.swing.GroupLayout.DEFAULT_SIZE, javax.swing.GroupLayout.PREFERRED_SIZE)

.addGap(18, 18, Short.MAX_VALUE)

.addComponent(jButton6, javax.swing.GroupLayout.PREFERRED_SIZE, 36, javax.swing.GroupLayout.PREFERRED_SIZE)

.addGap(21, 21, 21))

);

jLabel12.setFont(new java.awt.Font("Cambria", 1, 24)); // NOI18N

jLabel12.setText(" Security and Privacy?");

javax.swing.GroupLayout jPanel14Layout = new javax.swing.GroupLayout(jPanel14);

jPanel14.setLayout(jPanel14Layout);

jPanel14Layout.setHorizontalGroup(

jPanel14Layout.createParallelGroup(javax.swing.GroupLayout.Alignment.LEADING)

.addGroup(jPanel14Layout.createSequentialGroup()

.addGroup(jPanel14Layout.createParallelGroup(javax.swing.GroupLayout.Alignment.TRAILING)
```

```
.addComponent(jPanel15, javax.swing.GroupLayout.PREFERRED_SIZE, javax.swing.GroupLayout.DEFAULT_SIZE, javax.swing.GroupLayout.PREFERRED_SIZE)
.addGroup(jPanel14Layout.createParallelGroup(javax.swing.GroupLayout.Alignment.TRAILING)
.addGroup(jPanel14Layout.createSequentialGroup()
.addContainerGap()
.addComponent(jLabel12, javax.swing.GroupLayout.PREFERRED_SIZE, 248, javax.swing.GroupLayout.PREFERRED_SIZE))
.addGroup(javax.swing.GroupLayout.Alignment.LEADING, jPanel14Layout.createSequentialGroup()
.addGap(43, 43, 43)
.addComponent(jLabel10))))
.addContainerGap(43, Short.MAX_VALUE))
);
jPanel14Layout.setVerticalGroup(
jPanel14Layout.createParallelGroup(javax.swing.GroupLayout.Alignment.LEADING)
.addGroup(jPanel14Layout.createSequentialGroup()
.addGap(26, 26, 26)
.addComponent(jLabel10)
.addGap(21, 21, 21)
.addComponent(jLabel12)
.addPreferredGap(javax.swing.LayoutStyle.ComponentPlacement.RELATED, 20, Short.MAX_VALUE)
.addComponent(jPanel15, javax.swing.GroupLayout.PREFERRED_SIZE, javax.swing.GroupLayout.DEFAULT_SIZE, javax.swing.GroupLayout.PREFERRED_SIZE)
.addGap(49, 49, 49))
);
javax.swing.GroupLayout jPanel13Layout = new javax.swing.GroupLayout(jPanel13);
jPanel13.setLayout(jPanel13Layout);
```

```
jPanel13Layout.setHorizontalGroup(
jPanel13Layout.createParallelGroup(javax.swing.GroupLayout.Alignment.LEADING)
.addGroup(jPanel13Layout.createSequentialGroup()
.addGap(45, 45, 45)
.addComponent(jPanel14, javax.swing.GroupLayout.PREFERRED_SIZE, javax.swing.GroupLayout.DEFAULT_SIZE, javax.swing.GroupLayout.PREFERRED_SIZE)
.addContainerGap(43, Short.MAX_VALUE))
);
jPanel13Layout.setVerticalGroup(
jPanel13Layout.createParallelGroup(javax.swing.GroupLayout.Alignment.LEADING)
.addGroup(jPanel13Layout.createSequentialGroup()
.addGap(49, 49, 49)
.addComponent(jPanel14, javax.swing.GroupLayout.PREFERRED_SIZE, javax.swing.GroupLayout.DEFAULT_SIZE, javax.swing.GroupLayout.PREFERRED_SIZE)
.addContainerGap(44, Short.MAX_VALUE))
);
javax.swing.GroupLayout layout = new javax.swing.GroupLayout(getContentPane());
getContentPane().setLayout(layout);
layout.setHorizontalGroup(
layout.createParallelGroup(javax.swing.GroupLayout.Alignment.LEADING)
.addComponent(jPanel13, javax.swing.GroupLayout.DEFAULT_SIZE, javax.swing.GroupLayout.DEFAULT_SIZE, Short.MAX_VALUE)
);
layout.setVerticalGroup(
layout.createParallelGroup(javax.swing.GroupLayout.Alignment.LEADING)
```

```
.addComponent(jPanel13, javax.swing.GroupLayout.DEFAULT_SIZE, javax.swing.GroupLayout.DEFAULT_SIZE, Short.MAX_VALUE)
);
pack();
}// </editor-fold>
private void jButton6ActionPerformed(java.awt.event.ActionEvent evt) {
// TODO add your handling code here:
JOptionPane.showMessageDialog(null, "Uploaded Successfully");
}
private void jButton7ActionPerformed(java.awt.event.ActionEvent evt) {
// TODO add your handling code here:
int i=1;
try{
BufferedReader br=new BufferedReader(new FileReader(filepath));
String s1="",s2="";
br.readLine();
while((s1=br.readLine())!=null){
System.out.println(s1+"\n");
s2+=s1+"\n\n";
File f=new File("./Cloud Me/Blocks/"+i);
f.mkdir();
File ff=new File("./Cloud Me/Blocks/"+i+"/"+i+".txt");
ff.delete();
FileWriter fw1=new FileWriter("./Cloud Me/Blocks/"+i+"/"+i+".txt",true);
fw1.write(s1);
```

```
fw1.write("\r\n");
fw1.close();
File ff1=new File("./Cloud Me/Blocks/"+i+"/"+"Time Stamp"+".txt");
ff1.delete();
FileWriter fw2=new FileWriter("./Cloud Me/Blocks/"+i+"/"+"Time Stamp"+".txt",true);
Date date= new Date();
long time = date.getTime();
System.out.println("Time in Milliseconds: " + time);
Timestamp ts = new Timestamp(time);
System.out.println("Current Time Stamp: " + ts);
fw2.write("Time in Milliseconds: " + time +"\n" + "Current Time Stamp: " + ts);
fw2.write("\r\n");
fw2.close();
i++;
}
br.close();
JOptionPane.showMessageDialog(null, "Blocks Created Successfully");
}
catch (Exception ex)
{}
}
private void jButton8ActionPerformed(java.awt.event.ActionEvent evt) {
// TODO add your handling code here:
jTextField1.setText(""+nor);
DefaultListModel lmodel = new DefaultListModel();
```

```
for(int i=1;i<=nor;i++)
{
lmodel.addElement(""+i);
}
jList1.setModel(lmodel);
}
private void jList1ValueChanged(javax.swing.event.ListSelectionEvent evt) {
// TODO add your handling code here:
Desktop desktop = Desktop.getDesktop();
File dirToOpen = null;
try {
dirToOpen = new File("Cloud Me\\Blocks\\"+jList1.getSelectedValue());
desktop.open(dirToOpen);
} catch (IllegalArgumentException iae) {
System.out.println("File Not Found");
} catch (IOException ex) {
}
}
/**
* @param args the command line arguments
*/
public static void main(String args[]) {
/* Set the Nimbus look and feel */
//<editor-fold defaultstate="collapsed" desc=" Look and feel setting code (optional) ">
/* If Nimbus (introduced in Java SE 6) is not available, stay with the default look and feel.
```

```
* For details see http://download.oracle.com/javase/tutorial/uiswing/lookandfeel/plaf.html
*/
try {
for (javax.swing.UIManager.LookAndFeelInfo info : javax.swing.UIManager.getInstalledLookAndFeels()) {
if ("Nimbus".equals(info.getName())) {
javax.swing.UIManager.setLookAndFeel(info.getClassName());
break;
}
}
} catch (ClassNotFoundException ex) {
java.util.logging.Logger.getLogger(block_Creation.class.getName()).log(java.util.logging.Level.SEVERE, null, ex);
} catch (InstantiationException ex) {
java.util.logging.Logger.getLogger(block_Creation.class.getName()).log(java.util.logging.Level.SEVERE, null, ex);
} catch (IllegalAccessException ex) {
java.util.logging.Logger.getLogger(block_Creation.class.getName()).log(java.util.logging.Level.SEVERE, null, ex);
} catch (javax.swing.UnsupportedLookAndFeelException ex) {
java.util.logging.Logger.getLogger(block_Creation.class.getName()).log(java.util.logging.Level.SEVERE, null, ex);
}
//</editor-fold>
/* Create and display the form */
java.awt.EventQueue.invokeLater(new Runnable() {
public void run() {
```

```
new block_Creation().setVisible(true);
}
});
}
// Variables declaration - do not modify
private javax.swing.JButton jButton6;
private javax.swing.JButton jButton7;
private javax.swing.JButton jButton8;
private javax.swing.JLabel jLabel10;
private javax.swing.JLabel jLabel11;
private javax.swing.JLabel jLabel12;
private javax.swing.JLabel jLabel13;
private javax.swing.JLabel jLabel14;
private javax.swing.JList<String> jList1;
private javax.swing.JPanel jPanel1;
private javax.swing.JPanel jPanel13;
private javax.swing.JPanel jPanel14;
private javax.swing.JPanel jPanel15;
private javax.swing.JScrollPane jScrollPane1;
private javax.swing.JSeparator jSeparator5;
private javax.swing.JTextField jTextField1;
// End of variables declaration
}
```

A

PROJECT REPORT

on

QUANTIFYING COVID-19 CONTENT ONLINE IN THE ONLINE HEALTH OPINION WAR USING MACHINE LEARNING

Submitted by

1) Nishna Nayana Reddy.N (17K81A05A0)

2) Nuchu Anurudh Yadav (17K81A05A1)

3) Parvatam Pavan Kumar (17K81A05A4)

4) Shikakolli Sandeep (17K81A05B1)

in partial fulfillment for the award of the

degree of

BACHELOR OF TECHNOLOGY

in

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Under the Guidance of

Dr.G.Govinda Rajulu

Associate Professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ST.MARTIN'S ENGINEERING COLLEGE

An Autonomous Institute

Dhulapally, Secunderabad – 500100

JUNE 2021

BONAFIDE CERTIFICATE

This is to certify that the project entitled "**Quantifying COVID-19 Content in the online Health Opinion war using Machine Learning",** is being submitted by **1. NISHNA NAYANA REDDY.N** (**17K81A05A0), 2. NUCHU ANURUDH YADAV (17K81A05A1) 3. PARVATAM PAVAN KUMAR (17K81A05A4) 4. SHIKAKOLLI SANDEEP (17K81A05B1)** in partial fulfillment of the requirement for the award of the degree of **BACHELOR OF TECHNOLOGY IN COMPUTER SCIENCE AND ENGINEERING** is recorded of bonafide work carried out by them. The result embodied in this report have been verified and found satisfactory.

Dr.G.Govinda Rajulu
Department of CSE

Head of the Department
Dr.M.NARAYANAN
Department of CSE

Internal Examiner

External Examiner

Place:

Date:

DECLARATION

We, the student of **Bachelor of Technology** in Department of Computer Science and Engineering', session: 2017 – 2021, St. Martin's Engineering College, Dhulapally, Kompally, Secunderabad, hereby declare that work presented in this Project Work entitled "**Quantifying COVID-19 Content in the online Health Opinion war using Machine Learning"** is the outcome of our own bonafide work and is correct to the best of our knowledge and this work has been undertaken taking care of Engineering Ethics. This result embodied in this project report has not been submitted in any university for award of any degree.

Nishna Nayana Reddy.N [17K81A05A0]
Nuchu Anurudh Yadav [17K81A05A1]
Parvatam Pavan Kumar [17K81A05A4]
Shikakolli Sandeep [17K81A05B1]

ABSTRACT

A huge amount of false content regarding this dangerous virus is shared online. In this project we use machine learning to quantify COVID-19 content, which is falsely appearing, online, which leads to establishment of health guidance, particularly about vaccinations. We found that the anti-vax community is developing a less focused debate around COVID-19 than its counterpart, the pro-vaccination community. However, the anti-vax community exhibits a broader range of topics related to COVID-19, and hence the information can appeal to a broader cross-section of individuals seeking COVID-19 guidance online, for example individuals wary of a mandatory fast-tracked COVID-19 vaccine or those seeking alternative remedies. Hence, the anti-vax community looks better positioned to attract fresh support going forward when compared to pro-vax community. The popularity of anti-vax community leads widespread lack of adoption of a COVID-19 vaccine, which means the world falls short of providing herd immunity, leaving countries open to future COVID-19 resurgences. We provide a mechanistic model that interprets these results and could help in assessing the likely efficacy of intervention strategies. Our approach is scalable and hence tackles the urgent problem facing social media platforms of having to analyse huge volumes of online health misinformation.

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible and whose encouragements and guidance have crowded effects with success.

We extended our deep sense of gratitude to Principal**, Dr. P. SANTOSH KUMAR PATRA**, St. Martin's Engineering College, Dhulapally, for permitting us to undertake this project.

We are also thankful to **Dr. M.NARAYANAN**, Head of the Department, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

We would like to thank **Dr. T. POONGOTHAI,** Department of Computer Science and Engineering (AI & ML) for her encouragement and insightful comments and as well as our project coordinators **Dr. B.RAJALINGAM**, Associate Professor and **Dr. G. GOVINDARAJULU,** Associate Professor, in Department of Computer Science and Engineering for their valuable support.

We would like to express our sincere gratitude and indebtedness to our project supervisor **Dr. G. GOVINDARAJULU,** Associate Professor, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

Finally, we express thanks to all those who have helped us successfully to completing this project. Furthermore, we would like to thank our family and friends for their moral support and encouragement.

We express thanks to all those who have helped us in successfully completing the project.

Nishna Nayana Reddy.N [17K81A05A0]
Nuchu Anurudh Yadav [17K81A05A1]
Parvatam Pavan Kumar [17K81A05A4]
Shikakolli Sandeep [17K81A05B1]

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF OUTPUT SCREENS

LIST OF ABBREVATIONS

AVI	Audio Video Interlace
BMP	Bitmap
CPU	Central Processing Unit
GB	Giga Bytes
GUI	Graphical User Interface
LDA	Latent Dirichlet Allocation
NLTK	National Language Toolkit

CHAPTER-I

1.INTRODUCTION

Scientific experts agree that defeating COVID-19 will depend on developing a vaccine. However, this assumes that a sufficiently large proportion of people would receive a vaccine so that herd immunity is achieved. Because vaccines tend to be less effective in older people, this will require younger generations to have very high COVID-19 vaccination rates in order to guarantee herd immunity. Yet there is already significant opposition to existing vaccinations, e.g. against measles, with some parents already refusing to vaccinate their children. Such vaccine opposition increased the number of cases in the 2019 measles outbreak in the U.S. and beyond. Any future COVID-19 vaccine will likely face similar opposition.

Online social media platforms, and in particular the built-in communities that platforms like Facebook (FB) feature, have become popular fora for vaccine opponents (anti-vax) to congregate and share health (mis)information. Such misinformation can endanger public health and individual safety. Likewise, vaccine supporters (pro-vax) also congregate in such online communities to discuss and advocate for professional public health guidance. Well before COVID-19, there was already an intense online conflict featuring anti-vax communities and pro-vax communities. Within anti-vax communities, the narratives typically draw on and generate misinformation about establishment medical guidance and distrust of the government, pharmaceutical industry, and new technologies such as 5G communications [1][5]. Adding fuel to this fire, the January 2020 birth of the COVID-19 “info Demic” has led to a plethora of misinformation in social media surrounding COVID-19 that directly threatens lives [6]. For example, harmful “cures” are being proposed such as drinking fish tank additives, bleach, or cow urine, along with coordinated threats against public health officials like Dr.Anthony Fauci, director of the U.S. National Institute of Allergic and Infectious Diseases. Moreover, false rumors have been circulating that individual with dark skin are immune to COVID-19. These may have contributed to more relaxed social distancing among some minorities and hence their over-representation as victims.

Unfortunately, the sheer volume of new online content and the speed with which it spreads, means that social media companies are struggling to contain such health misinformation [14][15]. Making matters worse, people around the world are spending more time on social media due to social distancing imposed during the COVID-19 pandemic.

1.1 PROJECT OVERVIEW

The need for a deeper understanding of this intersection between online vaccination opposition and the online conversation surrounding COVID-19 and the need for an automated approach since the sheer volume of new online material every day makes manual analysis a non-viable option going forward. We pursue an automated, machine learning approach that avoids the scalability limitations of manual content analysis.

1.2 PROJECT OBJECTIVES

The anti-vax community exhibits a broader range of "flavors" of COVID-19 topics, and hence can appeal to a broader cross-section of individuals seeking COVID-19 guidance online, e.g. individuals wary of a mandatory fast-tracked COVID-19 vaccine or those seeking alternative remedies. Hence the anti-vax community looks better positioned to attract fresh support going forward than the pro-vax community. This is concerning since a widespread lack of adoption of a COVID-19 vaccine will mean the world falls short of providing herd immunity, leaving countries open to future COVID-19 resurgences.

1.3 ORGANISATION OF CHAPTERS:

CHAPTER-1:

In this chapter of Introduction, the information is about the preface of our project which gives details related to overview and objectives of the project.

CHAPTER-2:

The chapter of Literature survey gives the matter related to some existing articles in detail, which are related to our project research and we also conclude the survey about related references.

CHAPTER-3:

Requirements of the project are given in this chapter. There are both software and hardware requirements were provided.

CHAPTER-4:

In the chapter of software development analysis, we discuss about overview and definition of problem and also overview, definition and functionality of modules.

CHAPTER-5:

The pictorial representation of our project is in this chapter. The UML diagrams are used to give system design of our project.

CHAPTER-6:

In this chapter we gather information about coding of the project. The templates of code, files used in code, functionalities of class, methods, inputs and output parameters are gathered.

CHAPTER-7:

The testing phase of our project is documented in this chapter, different types of testing like Black box and white box are defined along with test cases of project**.**

CHAPTER-8:

This chapter gives output screens of our project. In which the user interface and outputs of project are understood clearly.

CHAPTER-9:

The experimental results of the project are represented in this chapter.

CHAPTER-10:

The documentation of the project is concluded in this chapter along with future enhancement, references, paper publication, student profile and appendices.

CHAPTER-II

2. LITERATURE SURVEY

2.1 SURVEY ON BACKGROUND

1. The Internet plays a large role in disseminating anti-vaccination information. This paper builds upon previous research by analyzing the arguments proffered on anti-vaccination websites, determining the extent of misinformation present, and examining discourses used to support vaccine objections. Arguments around the themes of safety and effectiveness, alternative medicine, civil liberties, conspiracy theories, and morality were found on the majority of websites analyzed; misinformation was also prevalent.

2. The most commonly proposed method of combating this misinformation is through better education, although this has proven ineffective. Education does not consider the discourses supporting vaccine rejection, such as those involving alternative explanatory models of health, interpretations of parental responsibility, and distrust of expertise. Anti-vaccination protestors make postmodern arguments that reject biomedical and scientific "facts" in favour of their own interpretations. Pro-vaccination advocates who focus on correcting misinformation reduce the controversy to merely an "educational" problem; rather, these postmodern discourses must be acknowledged in order to begin a dialogue.

3. The number of stay-at-home dads (SAHDs) in the U.S. has risen dramatically over the past 30 years. Despite gaining social acceptability, SAHDs still experience isolation and judgment in their offline environments. This research explores how SAHDs use the Internet and social media related to their roles as fathers. We conducted interviews with 18 SAHDs about their families, their identities, and their social experiences. We find that they turn to social media to gain social support and overcome isolation they experience offline. However, they engage in strategic self-disclosure on particular platforms to avoid judgment related to being SAHDs. They rely on online platforms to give off both traditionally feminine and masculine impression’s as loving caregivers of their children while simultaneously as do-it-yourself men who make things around the house. Through creating Facebook groups and using anonymous social media sites, SAHDs create multidimensional social networks that allow them to cope better with the role change.

4. The World Health Organization lists vaccine hesitancy as one of 10 threats to global health. The anti-vaccine movement uses Facebook to promote messages on the alleged dangers and consequences of vaccinating, leading to a reluctance to immunize against preventable communicable diseases.

5. Worldwide measles cases surged 17 percent so far this year amid stagnating immunization rates, the World Health Organization said on Friday. "The fact that any child dies from a vaccine-preventable disease like measles is frankly an outrage and a collective failure to protect the world’s most vulnerable children," said Dr. Tedros Adhanom Ghebreysus, director-general of the WHO.

6. Many of those sharing the post are pushing a conspiracy theory falsely claiming that 5G - which is used in mobile phone networks and relies on signals carried by radio waves - is somehow responsible for coronavirus.
These theories appear to have first emerged via Facebook posts in late January, around the same time the first cases were recorded in the US.

7. We show that malicious COVID-19 content, including racism, disinformation, and misinformation, exploits the multiverse of online hate to spread quickly beyond the control of any individual social media platform. We provide a first mapping of the online hate network across six major social media platforms. We demonstrate how malicious content can travel across this network in ways that subvert platform moderation efforts. Machine learning topic analysis shows quantitatively how online hate communities are sharpening COVID-19 as a weapon, with topics evolving rapidly and content becoming increasingly coherent. Based on mathematical modeling, we provide predictions of how changes to content moderation policies can slow the spread of malicious content.

8. Online hate and extremist narratives have been linked to abhorrent real-world events, including a current surge in hate crimes[1-6] and an alarming increase in youth suicides that result from social media vitriol[7]; inciting mass shootings such as the 2019 attack in Christchurch, stabbings and bombings[8-11]; recruitment of extremists[12-16], including entrapment and sex-trafficking of girls as fighter brides[17]; threats against public figures, including the 2019 verbal attack against an anti-Brexit politician, and hybrid (racist-anti-women-anti-immigrant) hate threats against a US member of the British royal family[18]; and renewed anti-western hate in the 2019 post-ISIS landscape associated with support for Osama Bin Laden's son and Al Qaeda. Social media platforms seem to be losing the battle against online hate[19,20] and urgently need new insights. Here we show that the key to understanding the resilience of online hate lies in its global network-of-network dynamics. Interconnected hate clusters form global 'hate highways' that-assisted by collective online adaptations-cross social media platforms, sometimes using 'back doors' even after being banned, as well as jumping between countries, continents and languages. Our mathematical model predicts that policing within a single platform (such as Facebook) can make matters worse, and will eventually generate global 'dark pools' in which online hate will flourish. We observe the current hate network rapidly rewiring and self-repairing at the micro level when attacked, in a way that mimics the formation of covalent bonds in chemistry. This understanding enables us to propose a policy matrix that can help to defeat online hate, classified by the preferred (or legally allowed) granularity of the intervention and top-down versus bottom-up nature. We provide quantitative assessments for the effects of each intervention. This policy matrix also offers a tool for tackling a broader class of illicit online behaviours, such as financial fraud.

9. Support for an extremist entity such as Islamic State (ISIS) somehow manages to survive globally online despite considerable external pressure and may ultimately inspire acts by individuals having no history of extremism, membership in a terrorist faction, or direct links to leadership. Examining longitudinal records of online activity, we uncovered an ecology evolving on a daily time scale that drives online support, and we provide a mathematical theory that describes it. The ecology features self-organized aggregates (ad hoc groups formed via linkage to a Facebook page or analog) that proliferate preceding the onset of recent real-world campaigns and adopt novel adaptive mechanisms to enhance their survival. One of the predictions is that

development of large, potentially potent pro-ISIS aggregates can be thwarted by targeting smaller ones.

10. This paper assesses topic coherence and human topic ranking of uncovered latent topics from scientific publications when utilizing the topic model latent Dirichlet allocation (LDA) on abstract and full-text data. The coherence of a topic, used as a proxy for topic quality, is based on the distributional hypothesis that states that words with similar meaning tend to co-occur within a similar context. Although LDA has gained much attention from machine-learning researchers, most notably with its adaptations and extensions, little is known about the effects of different types of textual data on generated topics. Our research is the first to explore these practical effects and shows that document frequency, document word length, and vocabulary size have mixed practical effects on topic coherence and human topic ranking of LDA topics. We furthermore show that large document collections are less affected by incorrect or noise terms being part of the topic-word distributions, causing topics to be more coherent and ranked higher. Differences between abstract and full-text data are more apparent within small document collections, with differences as large as 90% high-quality topics for full-text data, compared to 50% high-quality topics for abstract data.

11. Quantifying the coherence of a set of statements is a long standing problem with many potential applications that has attracted researchers from different sciences. The special case of measuring coherence of topics has been recently studied to remedy the problem that topic models give no guaranty on the interpretablity of their output. Several benchmark datasets were produced that record human judgements of the interpretability of topics. We are the first to propose a framework that allows to construct existing word based coherence measures as well as new ones by combining elementary components. We conduct a systematic search of the space of coherence measures using all publicly available topic relevance data for the evaluation. Our results show that new combinations of components outperform existing measures with respect to correlation to human ratings. Finally, we outline how our results can be transferred to further applications in the context of text mining, information retrieval and the world wide web.

12. We present LDAvis, a web-based interactive visualization of topics estimated using Latent Dirichlet Allocation that is built using a combination of R and D3. Our visualization provides a global view of the topics (and how they differ from each other), while at the same time allowing for a deep inspection of the terms most highly associated with each individual topic. First, we propose a novel method for choosing which terms to present to a user to aid in the task of topic interpretation, in which we define the relevance of a term to a topic. Second, we present results from a user study that suggest that ranking terms purely by their probability under a topic is suboptimal for topic interpretation. Last, we describe LDAvis, our visualization system that allows users to flexibly explore topic-term relationships using relevance to better understand a fitted LDA model.

2.2 CONCLUSION ON SURVEY

Some Facebook messages encourage prevailing myths about the safety and consequences of vaccines and likely contribute to parents' vaccine hesitancy. Deeply concerning is the mistrust social media has the potential to cast upon the relationship between health care providers and the public. A grasp of common misconceptions can help support health care provider practice many diseases have been almost, or completely, eradicated due to immunization. Immunization against disease prevents 2-3 million deaths per year internationally and could prevent even more with global vaccination improvements. Immunization has vastly decreased mortality due to preventable communicable diseases. For example, before the introduction of the measles vaccine, 300,000-400,000 Canadians were infected every year, with some recoveries and many deaths. Since the elimination of measles in 1998 due to vaccines, there have been very few cases in Canada. Similarly, once the polio vaccine was introduced in Canada in the 1950s, cases reduced dramatically, and the current risk to the Canadian population is extremely low.

The World Health Organization (WHO) has declared vaccine hesitancy as one of the top 10 threats to global health. Social media has helped fuel the growth of the anti-vaccine movement, with Facebook being identified as a key disseminator of misinformation surrounding the campaign. Facebook is the largest social media platform, with more than 2 billion active monthly users. There have been serious efforts to reduce the amount of misinformation spread on the social media site by lowering the ranking of Groups and Pages making false claims. Social media administrators have been urged to remove these Pages and Groups altogether; however, counterarguments cite a violation of human rights to access uncensored information. This paper exposes the messages of the anti-vaccine movement online and how individuals perceive immunization. We aimed to uncover the myths and truths that users of Facebook Pages observe and partake in. Health care consumers and health care providers may find themselves on opposite ends of the debate. Lack of immunization places the public at risk and decreases public health efforts to curb measles and polio and prevent outbreaks of influenza (flu) along with other communicable diseases. The shift in power between doctors and patients due to easy access to information online has led to the questioning of health care providers and increased shared decision making. .

As most of the world awaits a vaccine to put an end to the COVID-19 (also known as the 2019 novel coronavirus) pandemic, "followers" of anti-vaccine Facebook Pages seem to fear the vaccine more than the virus itself. Amid the COVID-19 pandemic, social media sites such as Facebook are unable to control the health misinformation that is spread on its Pages. Anti-vaccine Pages have been providing conspiracy theories, safety concerns, and alternative health medication that grasp the attention of "undecided" individuals surfing the web for information on vaccines. The WHO is fighting to stop the spread of misinformation online by collaborating with social media giants to find a way to regulate false claims. Some examples of such claims include that COVID-19 is a bioweapon funded by the Bill & Melinda Gates Foundation or that it can simply be cured by consuming homemade concoctions (some include drinking bleach). Our aim was to know more about the messages that can influence readers and consumers that these websites are sharing via social media

CHAPTER-III

3.HARDWARE AND SOFTWARE REQUIREMENTS

3.1 SOFTWARE REQUIREMENTS

The functional requirements or the overall description documents include the product perspective and features, operating system and operating environment, graphics requirements, design constraints and user documentation.

The appropriation of requirements and implementation constraints gives the general overview of the project in regards to what the areas of strength and deficit are and how to tackle them.

- **Python version 3.7**
- **Latest version of Anaconda**
- **Jupiter**
- **Google co-lab**

3.2 HARDWARE REQUIREMENTS

Minimum hardware requirements are very dependent on the particular software being developed by a given Bethought Python / Canopy / VS Code user. Applications that need to store large arrays/objects in memory will require more RAM, whereas applications that need to perform numerous calculations or tasks more quickly will require a faster processor.

- **Operating system : Windows or Linux**
- **Processor : minimum Intel i3**
- **Ram : minimum 4 Gb**
- **Hard disk : minimum 250 Gb**

CHAPTER-IV

4. SOFTWARE DEVELOPMENT ANALYSIS

4.1 OVERVIEW OF THE PROBLEM:

The existing methods using Most models assume a standard SEIR structure. Fraser and colleagues to estimate size but make different changes on the nature of the different compartments and their respective residence times

4.2 DEFINING THE PROBLEM

There are of course many limitations of this study. There are other social media platforms, apart from Facebook, that should be explored _ but Facebook is the largest. Similar behaviors should arise in any platform where communities can form..

4.3 MODULES OVERVIEW

The functional requirements or the overall description documents include the product perspective and features, operating system and operating environment, graphics requirements, design constraints and user documentation.

4.4 DEFINE THE MODULE

The appropriation of requirements and implementation constraints gives the general overview of the project in regards to what the areas of strength and deficit are and how to tackle them

4.5 MODULE FUNCTIONALITY

Here we use machine learning to quantify COVID-19 content among online opponents of establishment health guidance, in particular vaccinations (``anti-vax''). We find that the anti-vax community is developing a less focused debate around COVID-19 than its counterpart, the pro-vaccination (``pro-vax'') community.hence can appeal to a broader cross-section of individuals seeking COVID-19 guidance online, e.g. individuals wary of a mandatory fast-tracked COVID-19 vaccine or those seeking alternative remedies. Hence the anti-vax community looks better positioned to attract fresh support going forward than the pro-vax community. This is concerning since a widespread lack of adoption of a COVID-19 vaccine will mean the world falls short of providing herd immunity, leaving countries open to future COVID-19 resurgences. We provide a mechanistic model that interprets these results and could help in assessing the likely efficacy of intervention strategies

CHAPTER-V

5. PROJECT SYSTEM DESIGN

The System Design Document describes the system requirements, operating environment, system and subsystem architecture, files and database design, input formats, output layouts, human-machine interfaces, detailed design, processing logic, and external interfaces.

It is divided into two types like GUI Designing, UML Designing with avails in development of project in facile way with different actor and its utilizer case by utilizer case diagram, flow of the project utilizing sequence, Class diagram gives information about different class in the project with methods that have to be utilized in the project if comes to our project our UML diagram is utilizable in this way.

5.1 UML DIAGRAMS

UML stands for Unified Modeling Language. UML is a standardized general-purpose modeling language in the field of object-oriented software engineering. The standard is managed, and was created by, the Object Management Group.

The goal is for UML to become a common language for creating models of object-oriented computer software. In its current form UML is comprised of two major components: a Meta-model and a notation. In the future, some form of method or process may also be added to; or associated with, UML.

The Unified Modeling Language is a standard language for specifying, Visualization, Constructing and documenting the artifacts of software system, as well as for business modeling and other non-software systems.

The UML represents a collection of best engineering practices that have proven successful in the modeling of large and complex systems.

The UML is a very important part of developing objects-oriented software and the software development process. The UML uses mostly graphical notations to express the design of software projects.

USE CASE DIAGRAM:

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.

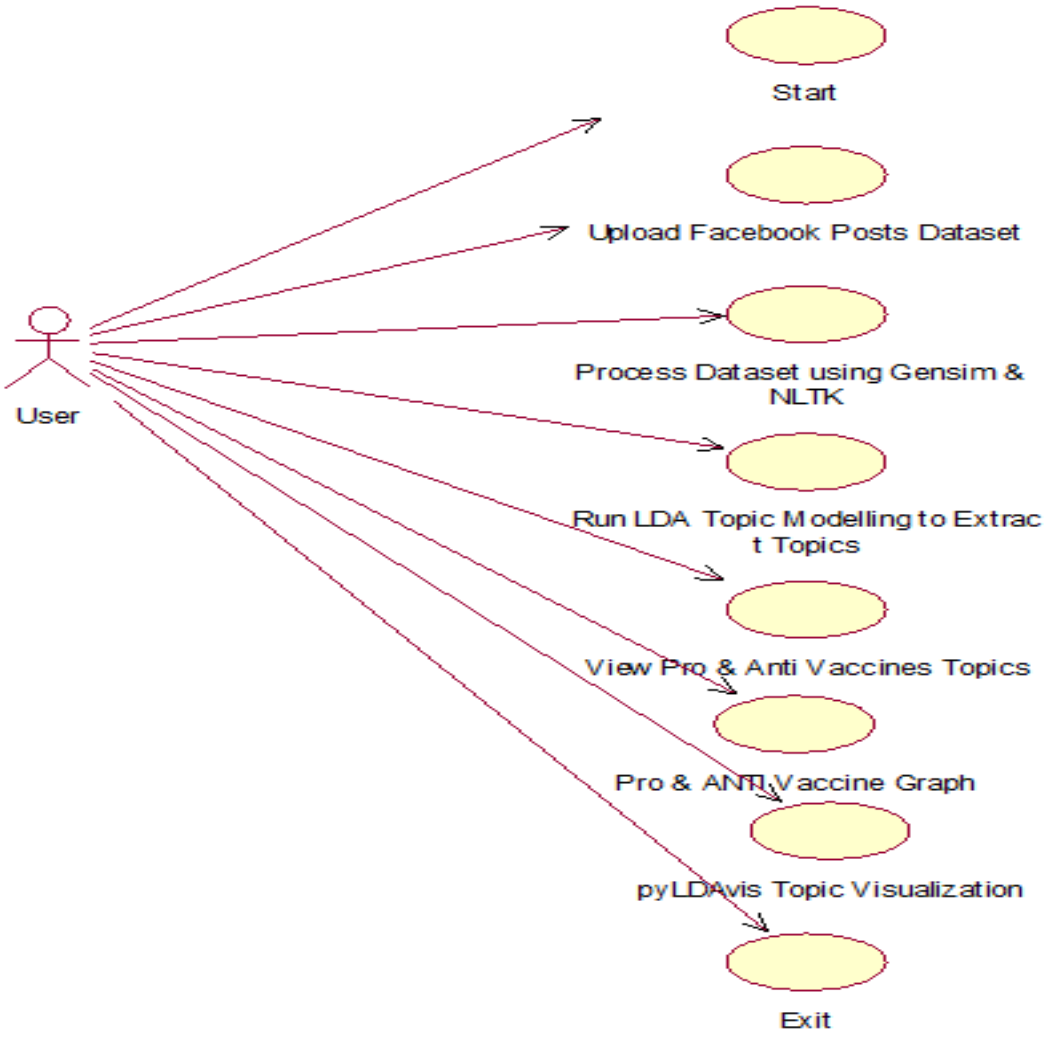


Fig5.1: Use Case Diagram

CLASS DIAGRAM:

In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.

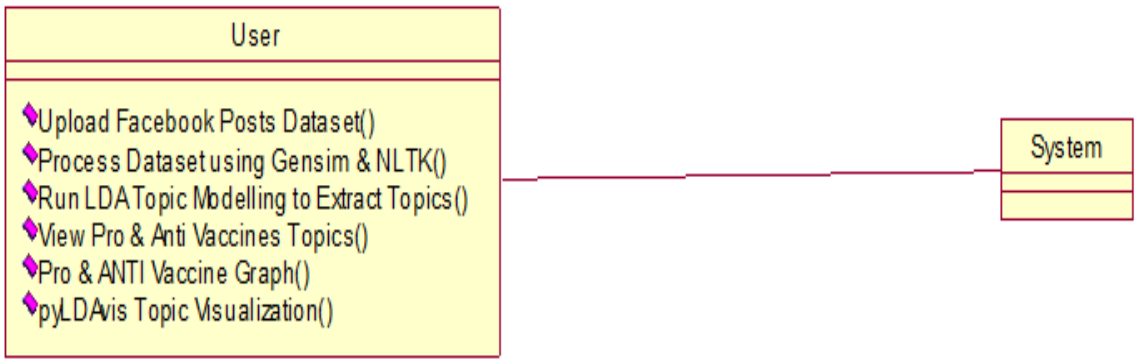


Fig:5.2 Class Diagram

SEQUENCE DIAGRAM:

A sequence diagram in Unified Modeling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.

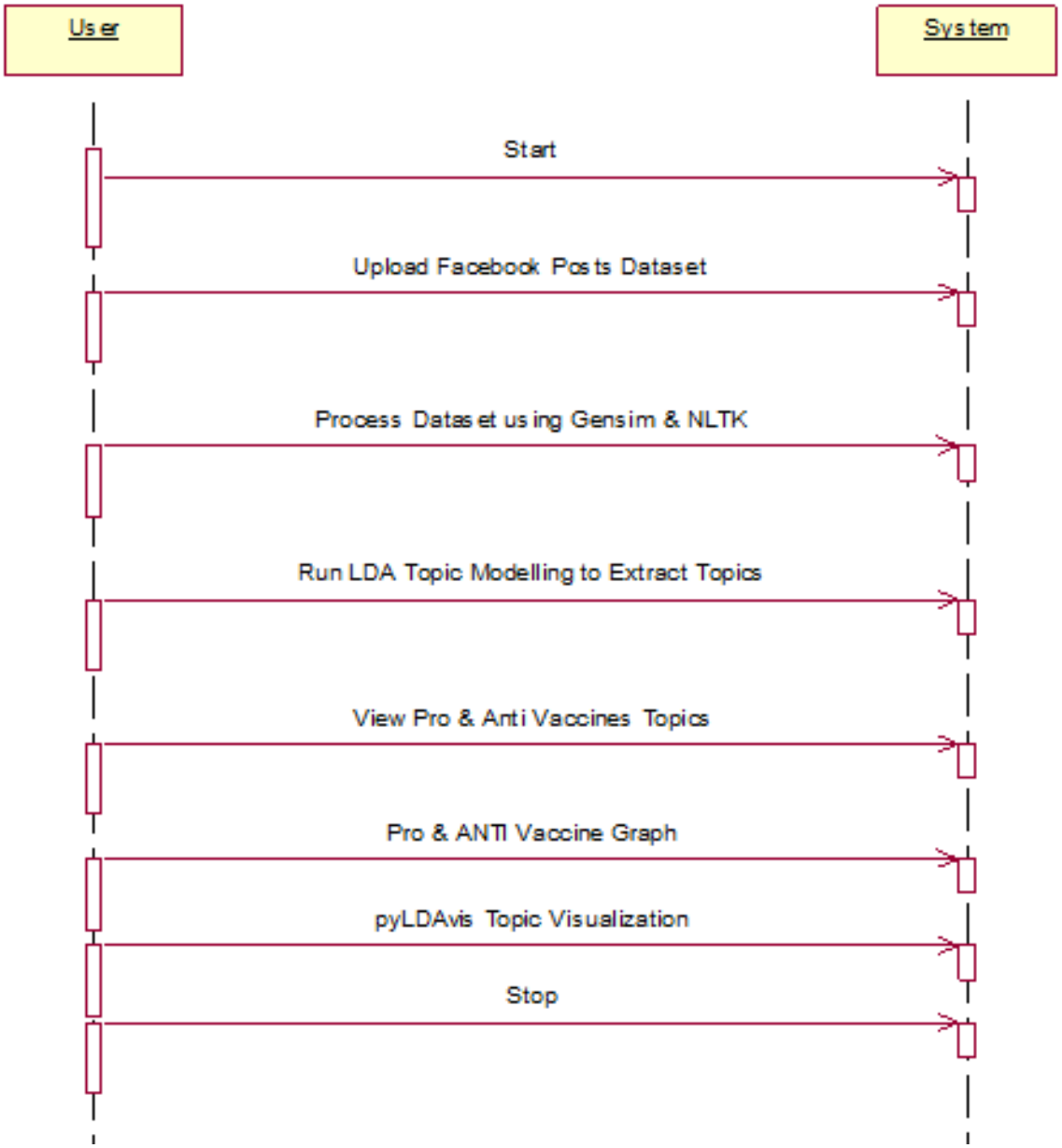


Fig:5.3 Sequence Diagram

ACTIVITY DIAGRAM:

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modeling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.

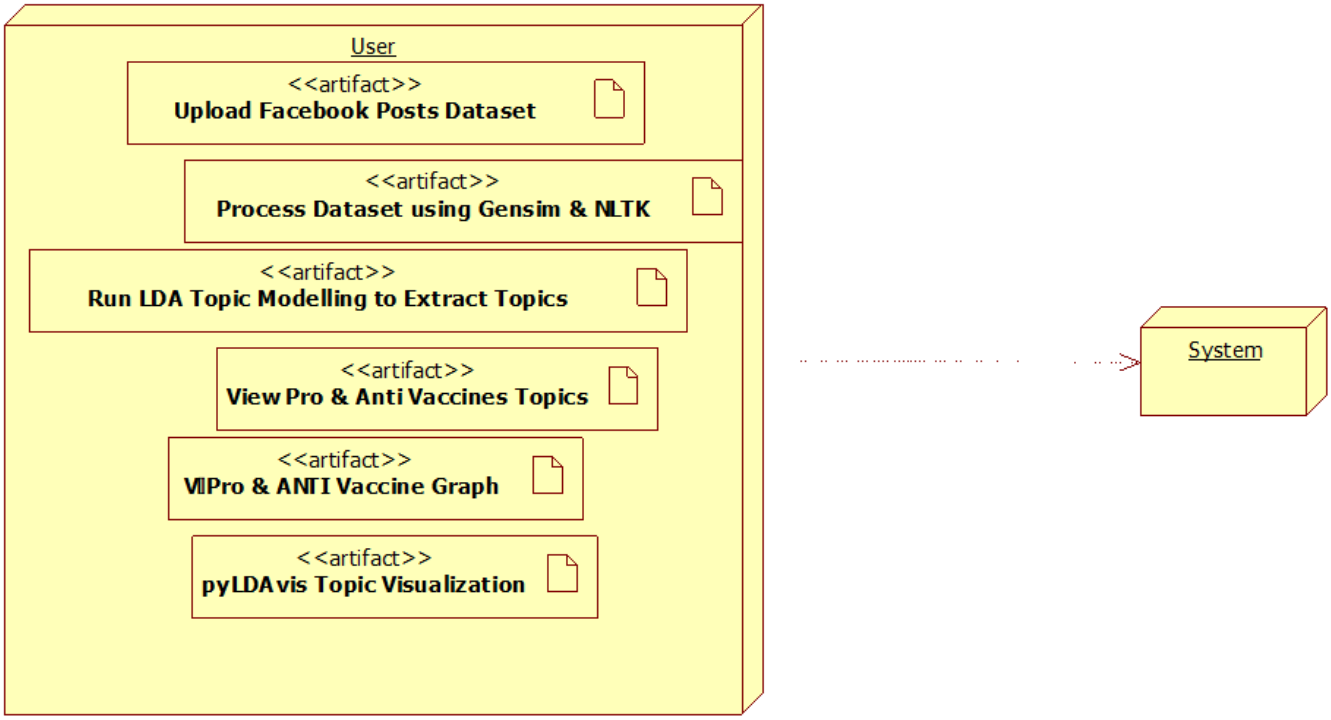


Fig:5.4 Activity Diagram

PACKAGE DIAGRAM:

Package diagram, a kind of structural diagram, shows the arrangement and organization of model elements in middle to large scale project. Package diagram can show both structure and dependencies between sub-systems or modules, showing different views of a system, for example, as multi-layered (aka multi-tiered) application - multi-layered application model.

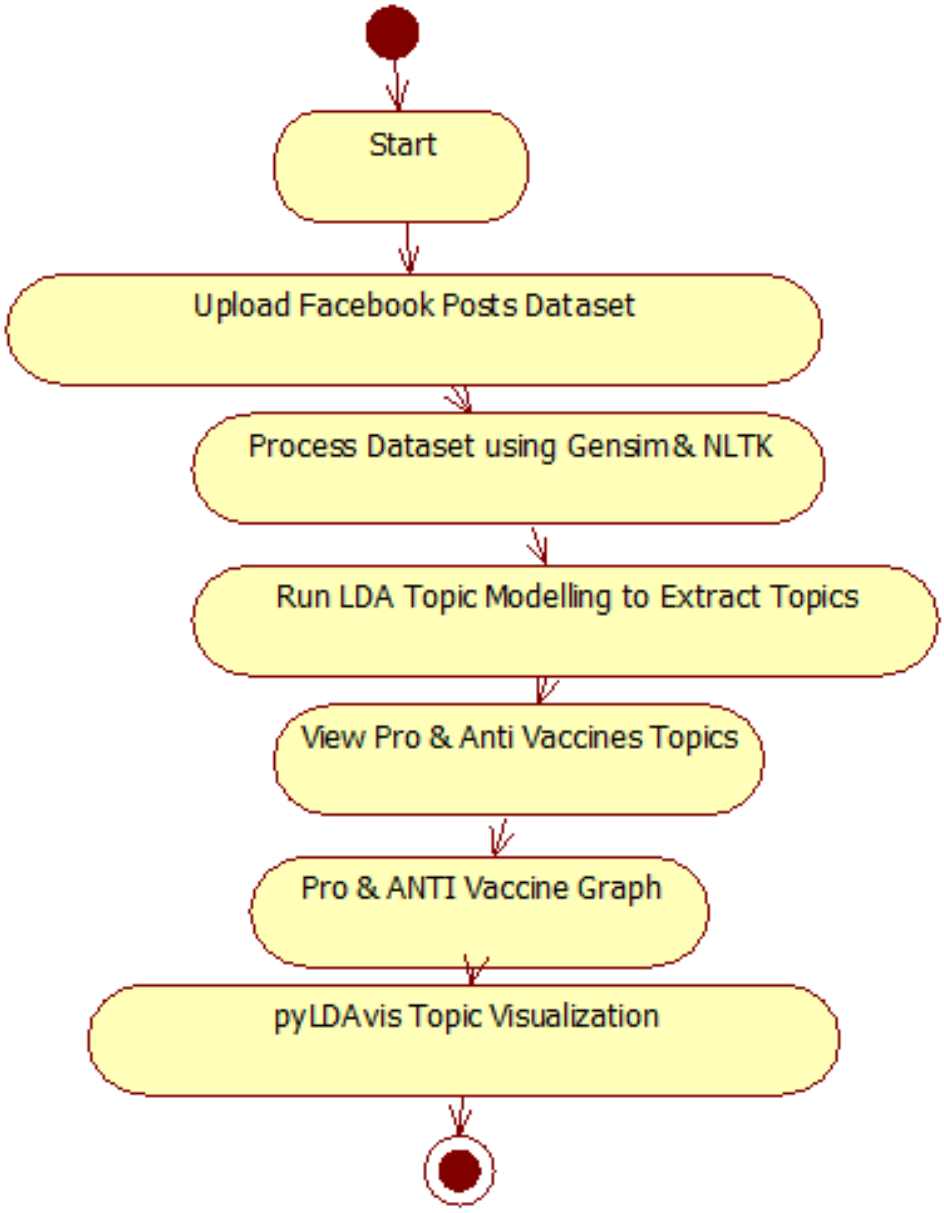


Fig:5.5 Package Diagram

DEPLOYMENT DIAGRAM:

A UML deployment diagram is a diagram that shows the configuration of run time processing nodes and the components that live on them. Deployment diagrams is a kind of structure diagram used in modeling the physical aspects of an object-oriented system. They are often be used to model the static deployment view of a system (topology of the hardware).

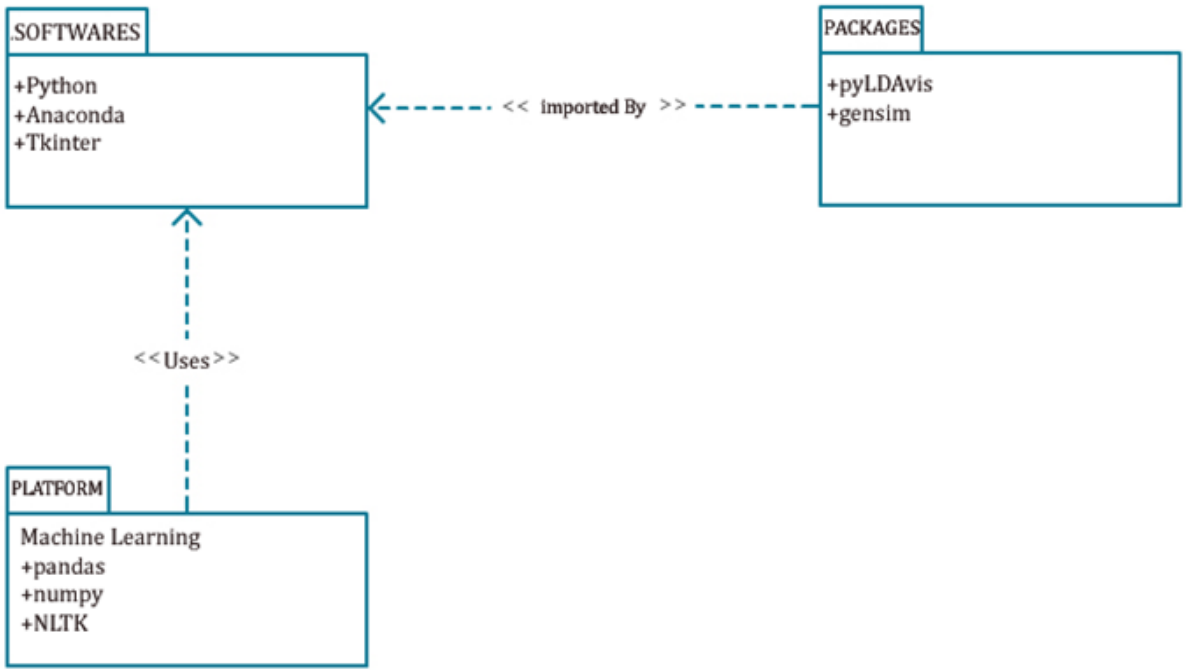


Fig:5.6 Deployment Diagram

COMPONENT DIAGRAM:

UML Component diagrams are used in modeling the physical aspects of object-oriented systems that are used for visualizing, specifying, and documenting component-based systems and also for constructing executable systems through forward and reverse engineering. Component diagrams are essentially class diagrams that focus on a system's components that often used to model the static implementation view of a system.

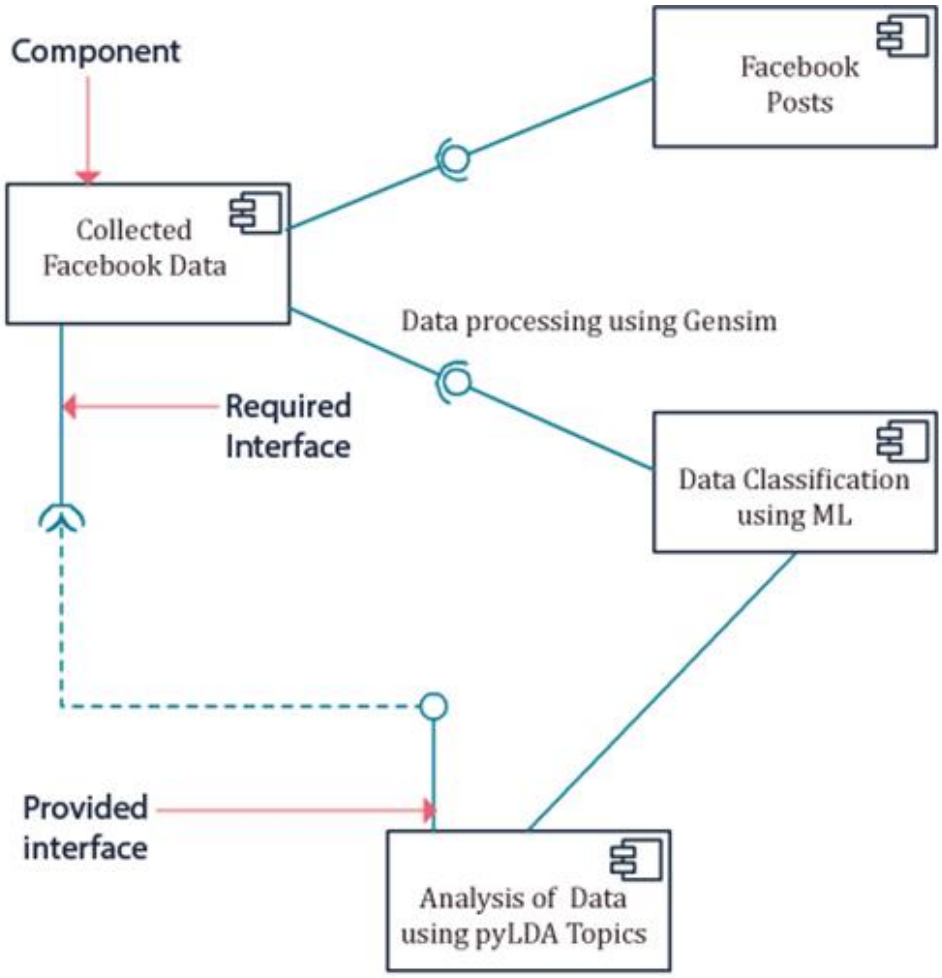


Fig:5.7 Component Diagram

CHAPTER-VI

6.PROJECT CODING

6.1 CODE TEMPLATES

```
from tkinter import messagebox
from tkinter import *
from tkinter import simpledialog
import tkinter
from tkinter import filedialog
import matplotlib.pyplot as plt
import numpy as np
from tkinter.filedialog import askopenfilename
import pandas as pd
from string import punctuation
from nltk.corpus import stopwords
import nltk
from nltk.corpus import wordnet as wn
from nltk.stem.wordnet import WordNetLemmatizer
from gensim import corpora
import pickle
import gensim
import pyLDAvis.gensim
import matplotlib.pyplot as plt
from gensim.models.coherencemodel import CoherenceModel
from gensim.corpora.dictionary import Dictionary
from collections import defaultdict

main = tkinter.Tk()
main.title("Quantifying COVID-19 Content in the Online Health Opinion War Using
main.geometry("1300x1200")

global filename
en_stop = set(nltk.corpus.stopwords.words('english'))
global text_data
global dictionary
global corpus
global ldamodel
global pro,anti

def cleanPost(doc):
    tokens = doc.split()
    table = str.maketrans('', '', punctuation)
    tokens = [w.translate(table) for w in tokens]
    tokens = [word for word in tokens if word.isalpha()]
    stop_words = set(stopwords.words('english'))
    tokens = [w for w in tokens if not w in stop_words]
    tokens = [word for word in tokens if len(word) > 1]
    tokens = ' '.join(tokens)
    #print(tokens)
    return tokens
```

Fig 6.1 code part

6.2 OUTLINES FOR VARIOUS FILES

D:\QuantifyCovid\QuantifyCovid\FacebookPost\Books.csv

6.3 CLASS WITH FUNCTIONALITY

from tkinter import messagebox

from tkinter import *

from tkinter import simpledialog

import tkinter

```
from tkinter import filedialog
import matplotlib.pyplot as plt
import numpy as np
from tkinter.filedialog import askopenfilename
import pandas as pd
from string import punctuation
from nltk.corpus import stopwords
import nltk
from nltk.corpus import wordnet as wn
from nltk.stem.wordnet import WordNetLemmatizer
from gensim import corpora
import pickle
import gensim
import pyLDAvis.gensim
import matplotlib.pyplot as plt
from gensim.models.coherencemodel import CoherenceModel
from gensim.corpora.dictionary import Dictionary
from collections import defaultdict

main = tkinter.Tk()
main.title("Quantifying COVID-19 Content in the Online Health Opinion War Using Machine Learning") #designing main screen
main.geometry("1300x1200")

global filename
en_stop = set(nltk.corpus.stopwords.words('english'))
global text_data
global dictionary
global corpus
```

```
global ldamodel
global pro,anti

def cleanPost(doc):
    tokens = doc.split()
    table = str.maketrans('', '', punctuation)
    tokens = [w.translate(table) for w in tokens]
    tokens = [word for word in tokens if word.isalpha()]
    stop_words = set(stopwords.words('english'))
    tokens = [w for w in tokens if not w in stop_words]
    tokens = [word for word in tokens if len(word) > 1]
    tokens = ' '.join(tokens)
    #print(tokens)
    return tokens

def get_lemma(word):
    lemma = wn.morphy(word)
    if lemma is None:
        return word
    else:
        return lemma

def get_lemma2(word):
    return WordNetLemmatizer().lemmatize(word)

def prepare_text_for_lda(text):
    tokens = text.split(" ")
    tokens = [token for token in tokens if len(token) > 4]
    tokens = [token for token in tokens if token not in en_stop]
```

```
    tokens = [get_lemma(token) for token in tokens]
    return tokens

def upload(): #function to upload tweeter profile
    global filename
    filename = filedialog.askopenfilename(initialdir="FacebookPost")
    text.delete('1.0', END)
    text.insert(END,filename+" loaded\n");

def processDataset():
    text.delete('1.0', END)
    global text_data
    text_data = []
    dataset = pd.read_csv(filename,encoding="ISO-8859-1")
    for i in range(len(dataset)):
        msg = dataset._get_value(i, 'Posts')
        clean = cleanPost(msg.strip('\n').strip().lower())
        clean = prepare_text_for_lda(clean)
        text_data.append(clean)
    text.insert(END,'Posts after processing\n\n')
    text.insert(END,str(text_data)+"\n\n")


def LDA():
    global dictionary
    global corpus
    global ldamodel
    text.delete('1.0', END)
    dictionary = corpora.Dictionary(text_data)
```

```
    corpus = [dictionary.doc2bow(text) for text in text_data]
    pickle.dump(corpus, open('corpus.pkl', 'wb'))
    dictionary.save('dictionary.gensim')
    NUM_TOPICS = 30
    ldamodel = gensim.models.ldamodel.LdaModel(corpus, num_topics = NUM_TOPICS,
id2word=dictionary, passes=15)
    ldamodel.save('model5.gensim')
    topics = ldamodel.print_topics(num_words=6)
    text.insert(END,'LDA Extracted Topics\n\n')
    for topic in topics:
        text.insert(END,str(topic)+"\n")


def viewTopics():
    global pro,anti
    anti_topics =
['shot','burder','protest','avoid','flu','fake','stop','afraid','never','test','spread','poison']
    pro_topics =
['maskwearing','protect','healthcare','trust','ailment','mask','wash','distancing','distance','soap','prev
ent','mandatory']
    pro =  {}
    anti = {}
    combine = {}
    for i in range(len(text_data)):
        data = text_data[i]
        for j in range(len(data)):
            if data[j] in anti_topics:
                if data[j] in anti:
                    anti[data[j]] = anti.get(data[j]) + 1
                else:
                    anti[data[j]] = 1
```

```
            if data[j] in pro_topics:
                if data[j] in pro:
                    pro[data[j]] = pro.get(data[j]) + 1
                else:
                    pro[data[j]] = 1
    text.delete('1.0', END)
    text.insert(END,'Pro vaccines topics details\n\n')
    text.insert(END,str(pro)+"\n\n")
    text.insert(END,'Anti vaccines topics details\n\n')
    text.insert(END,str(anti))


def scoreGraph():
    pro_graph = []
    anti_graph = []
    for key in pro:
        pro_graph.append(pro[key])
    for key in anti:
        anti_graph.append(anti[key])
    plt.figure(figsize=(10,6))
    plt.grid(True)
    plt.xlabel('Total Topics')
    plt.ylabel('Coherence scores')
    plt.plot(pro_graph, 'ro-', color = 'maroon')
    plt.plot(anti_graph, 'ro-', color = 'green')
    plt.legend(['Pro-Vax', 'Anti-Vax'], loc='upper left')
    plt.title('Coherence Topic Scores Graph')
    plt.show()
```

```
def graph():
    lda_display = pyLDAvis.gensim.prepare(ldamodel, corpus, dictionary, mds='mmds')
    #pyLDAvis.enable_notebook(local=True)
    pyLDAvis.show(lda_display)


font = ('times', 16, 'bold')
title = Label(main, text='Quantifying COVID-19 Content in the Online Health Opinion War Using Machine Learning')
title.config(bg='darkblue', fg='white')
title.config(font=font)
title.config(height=4, width=120)
title.place(x=0,y=0)

font1 = ('times', 12, 'bold')
text=Text(main,height=20,width=150)
scroll=Scrollbar(text)
text.configure(yscrollcommand=scroll.set)
text.place(x=50,y=120)
text.config(font=font1)


font1 = ('times', 14, 'bold')
uploadButton = Button(main, text="Upload Facebook Posts Dataset", command=upload, bg='yellow', fg='black')
uploadButton.place(x=50,y=550)
uploadButton.config(font=font1)

processButton = Button(main, text="Process Dataset using Gensim & NLTK", command=processDataset, bg='yellow', fg='black')
```

```
processButton.place(x=380,y=550)

processButton.config(font=font1)


LDAforest = Button(main, text="Run LDA Topic Modelling to Extract Topics",
command=LDA, bg='yellow', fg='black')

LDAforest.place(x=750,y=550)

LDAforest.config(font=font1)


topicButton = Button(main, text="View Pro & Anti Vaccines Topics",
command=viewTopics,bg='green', fg='white')

topicButton.place(x=50,y=600)

topicButton.config(font=font1)


vaccine = Button(main, text="Pro & Anti Vaccine Graph", command=scoreGraph, bg='green',
fg='white')

vaccine.place(x=380,y=600)

vaccine.config(font=font1)


graph = Button(main, text="pyLDAvis Topic Visualization", command=graph, bg='green',
fg='white')

graph.place(x=750,y=600)

graph.config(font=font1)


main.config(bg='lightblue')

main.mainloop()
```

6.4 METHOD INPUT AND OUTPUT PARAMETERS

Files given to the input

1. Posts.csv
2. Posts2.csv
3. Books.csv

CHAPTER-VII

7. PROJECT TESTING

Testing is a process of executing a program with the aim of finding error. To make our software perform well it should be error free. If testing is done in 7.1 successfully, it will remove all the errors from the software.

7.1 TYPES OF TEST CASES

Test Case Id	Test Case Name	Test Case Description	Test Steps			Test Case Status	Test Priority
			Step	Expected	Actual		
01	Start the Application	Host the application and test if it starts making sure the required software is available	If it doesn't Start	We cannot run the Application	The application hosts success.	High	High
02	Home Page	Check the deployment Environment for Properly loading the application.	If it doesn't load.	We cannot access the Application.	The application is running successfully	High	High
03	User Mode	Verify the working of the application	If it doesn't Respond	We cannot use the Freestyle	The application displays the Freestyle	High	High

		in freestyle mode		mode.	Page		
04	Data Input	Verify if the application takes input and updates	If it fails to take the input or store in The Database	We cannot proceed further	The application updates the input to application	High	High

7.2 WHITE BOX TESTING

Testing technique based on knowledge of the internal logic of an application's code and includes tests like coverage of code statements, branches, paths, conditions. It is performed by software developers

7.3 BLACK BOX TESTING

A method of software testing that verifies the functionality of an application without having specific knowledge of the application's code/internal structure. Tests are based on requirements and functionality.

Black box testing is testing the functionality of an application without knowing the details of its implementation including internal program structure, data structures etc. Test cases for black box testing are created based on the requirement specifications. Therefore, it is also called as specification-based testing. Fig.7.1 represents the black box testing:

Fig.: Black Box Testing

When applied to machine learning models, black box testing would mean testing machine-learning models without knowing the internal details such as features of the machine learning Model, the algorithm used to create the model etc. The challenge, however, is to verify the test outcome against the expected values that are known beforehand.

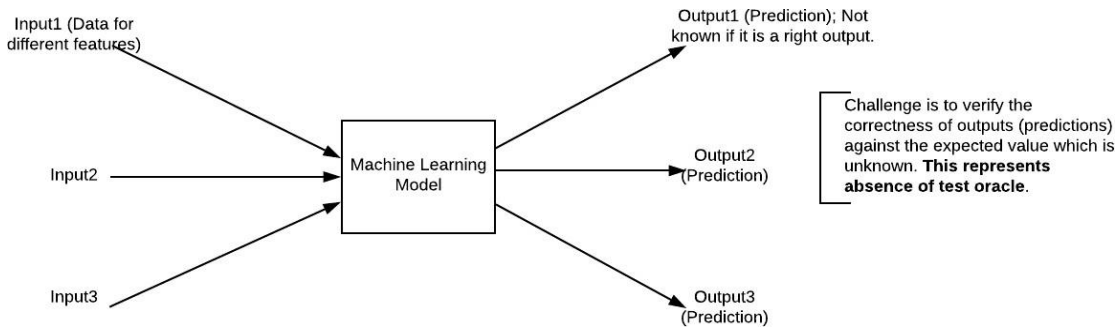


Fig.: Black Box Testing for Machine Learning algorithms

The above Fig.7.2 represents the black box testing procedure for machine learning algorithms.

Table.7.2: Black box Testing

Input	**Actual Output**	**Predicted Output**
[16,6,324,0,0,0,22,0,0,0,0,0,0]	0	0
[16,7,263,7,0,2,700,9,10,1153,832,9,2]	1	1

CHAPTER-VIII

8.OUTPUT SCREENS

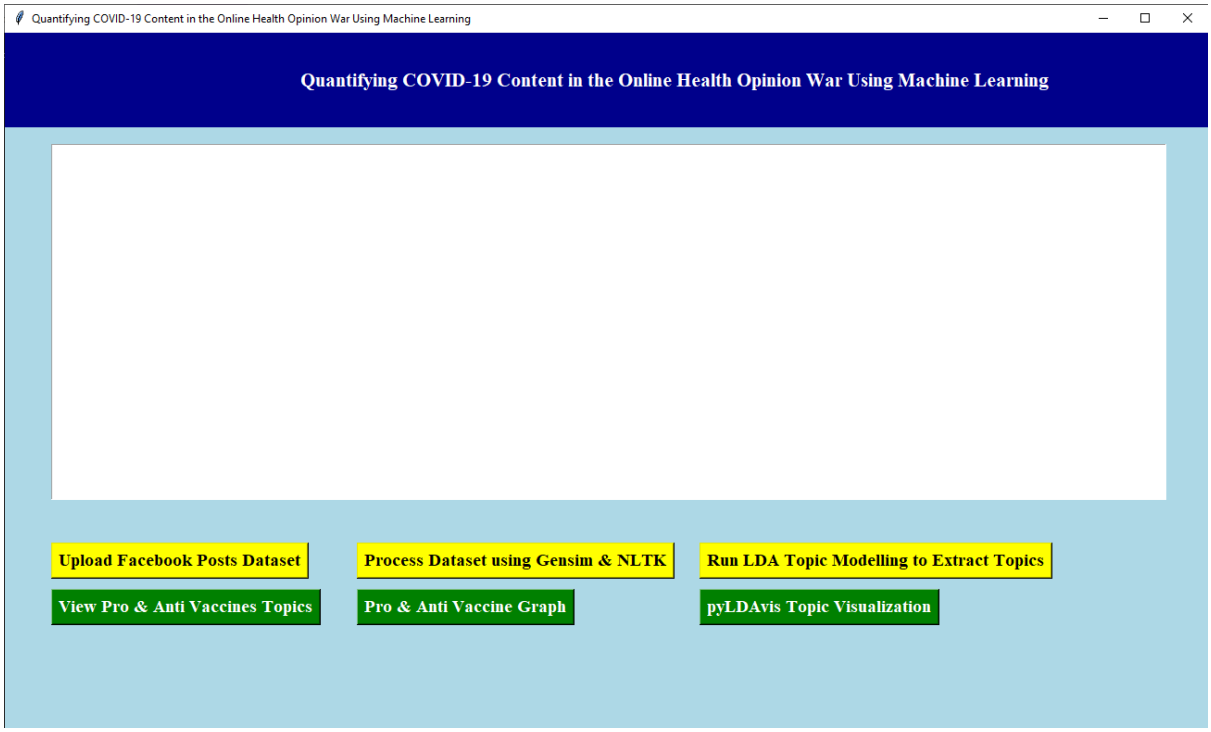


Fig: 8.1 User Interfaces

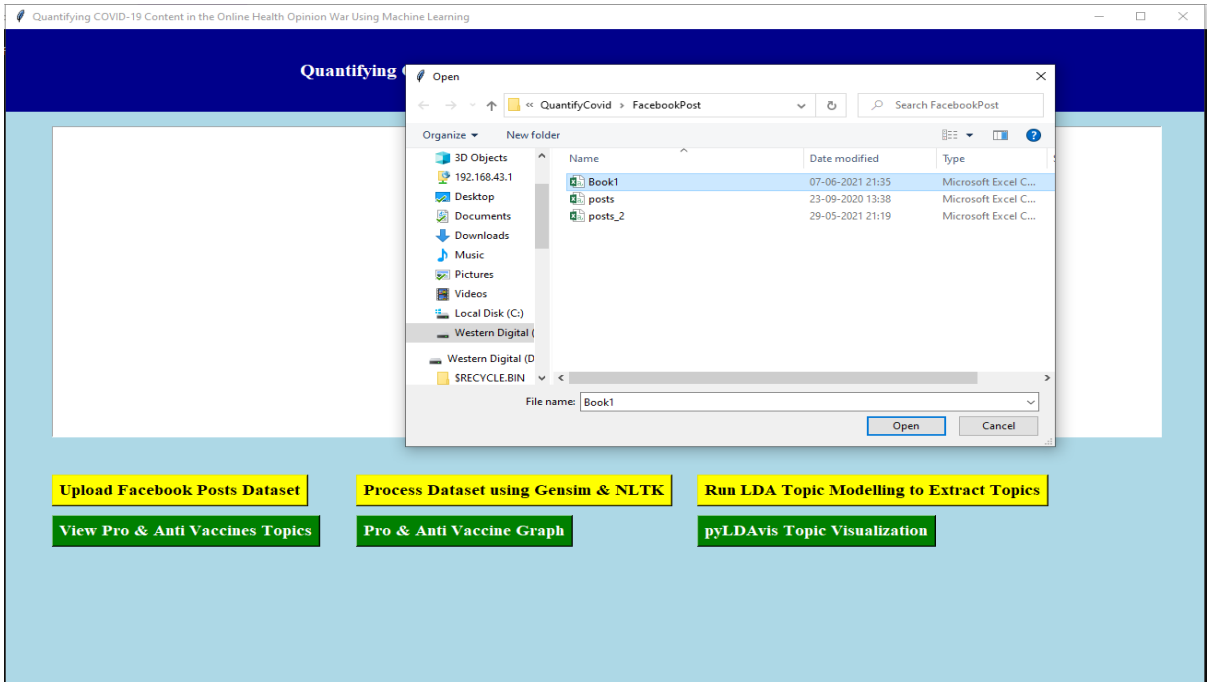


Fig:8.2 Uploading Posts

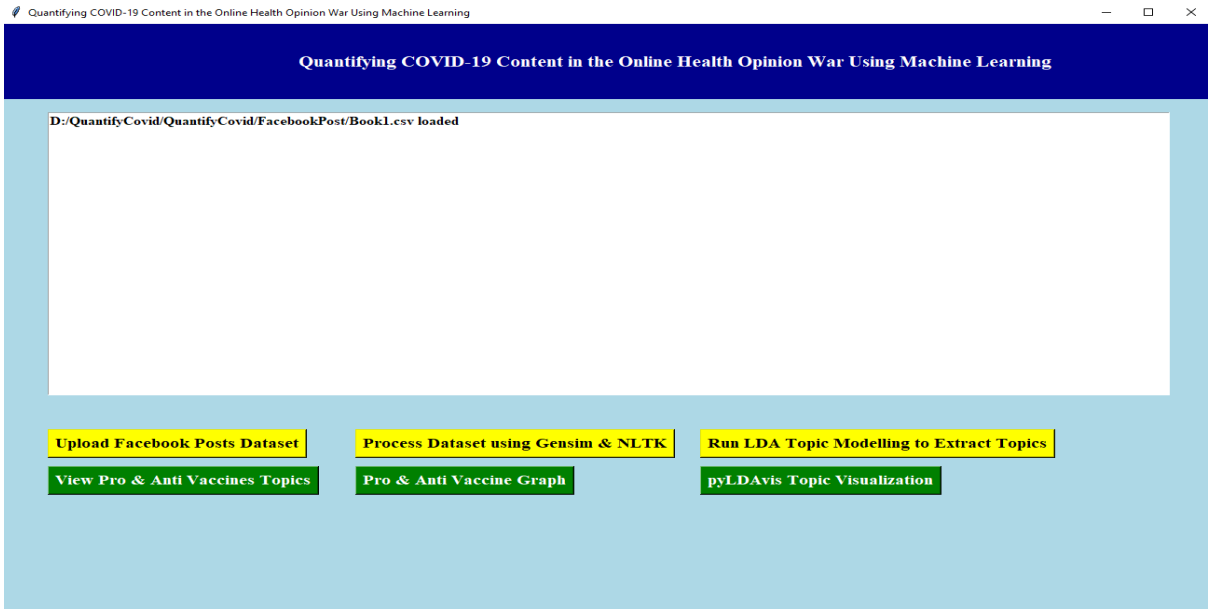


Fig:8.3 Uploaded Posts

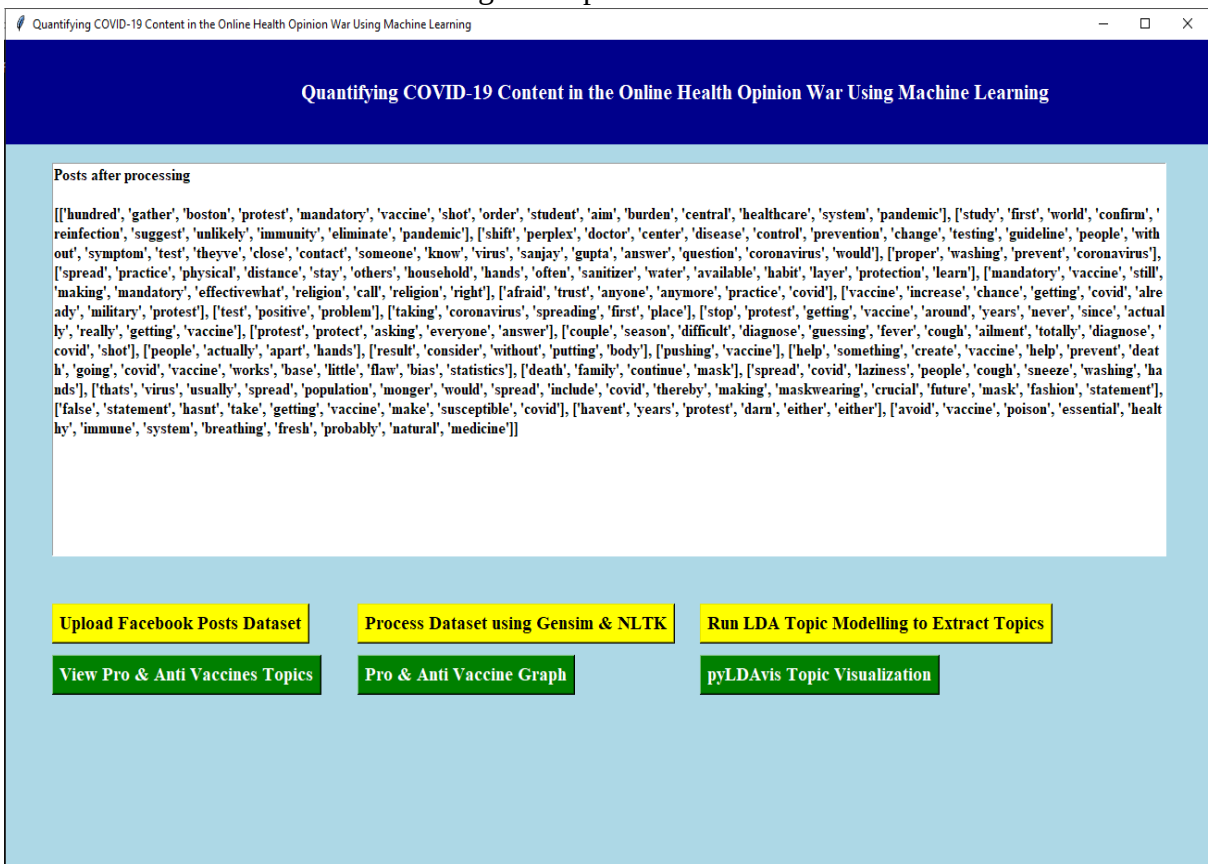


Fig:8.4Collected Data From Posts

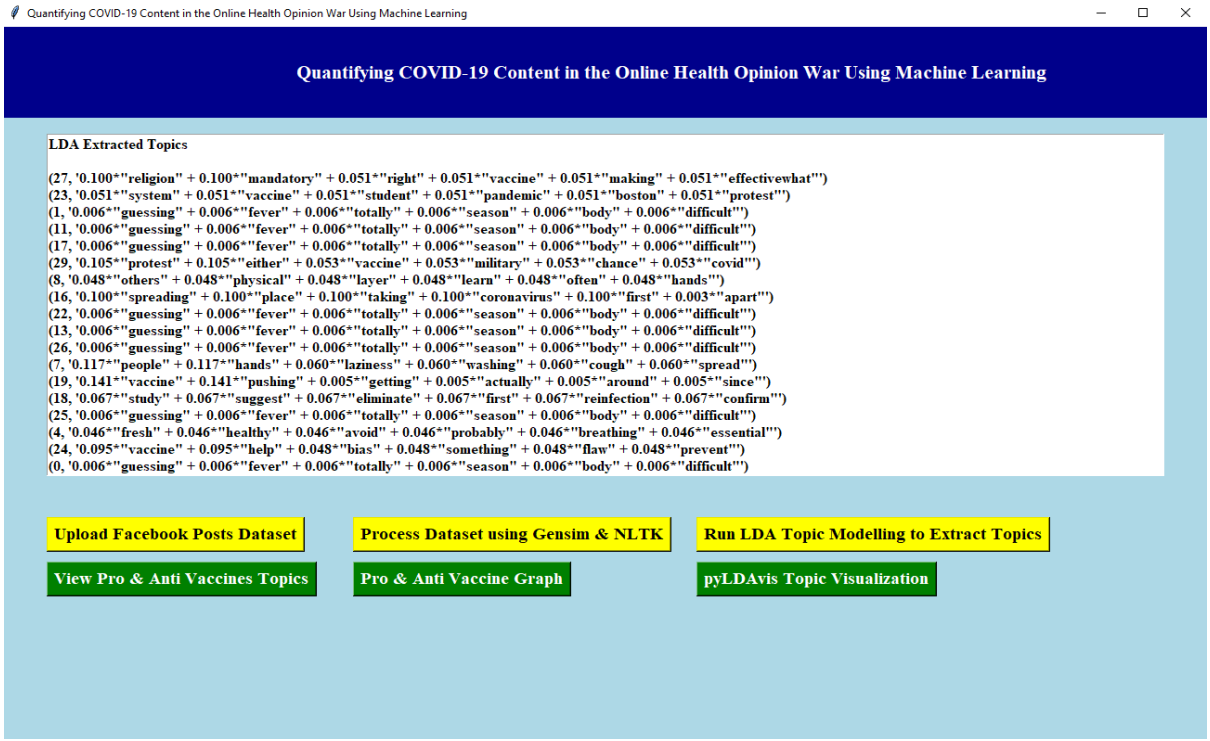


Fig:8.5 Processed Data in ML

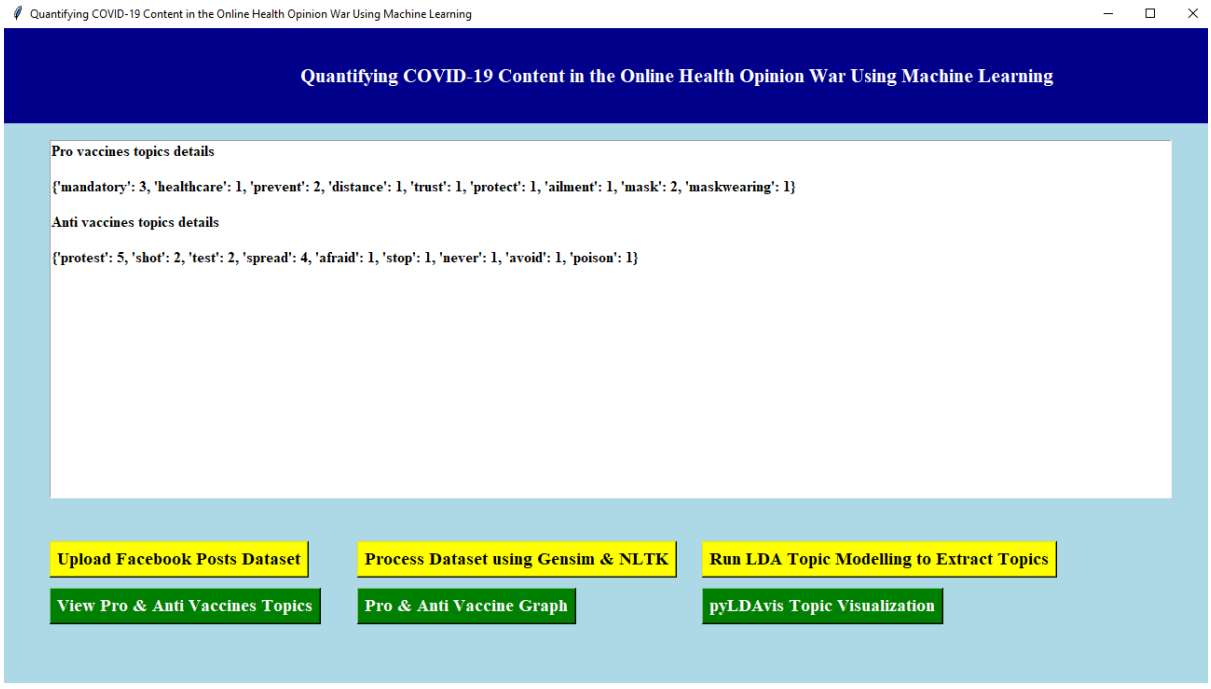


Fig:8.6 Extrated Topic Details

CHAPTER-IX

9.EXPERIMENTAL RESULTS

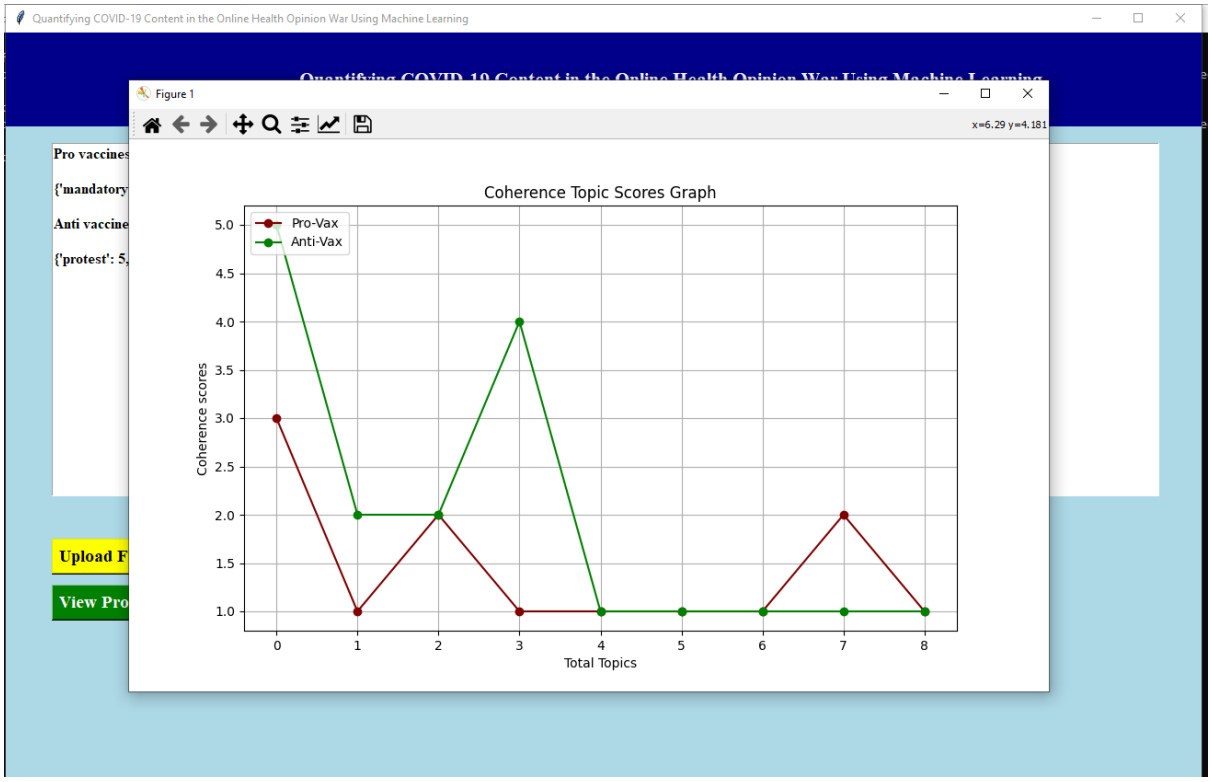


Fig:9.1 Coherence Graph

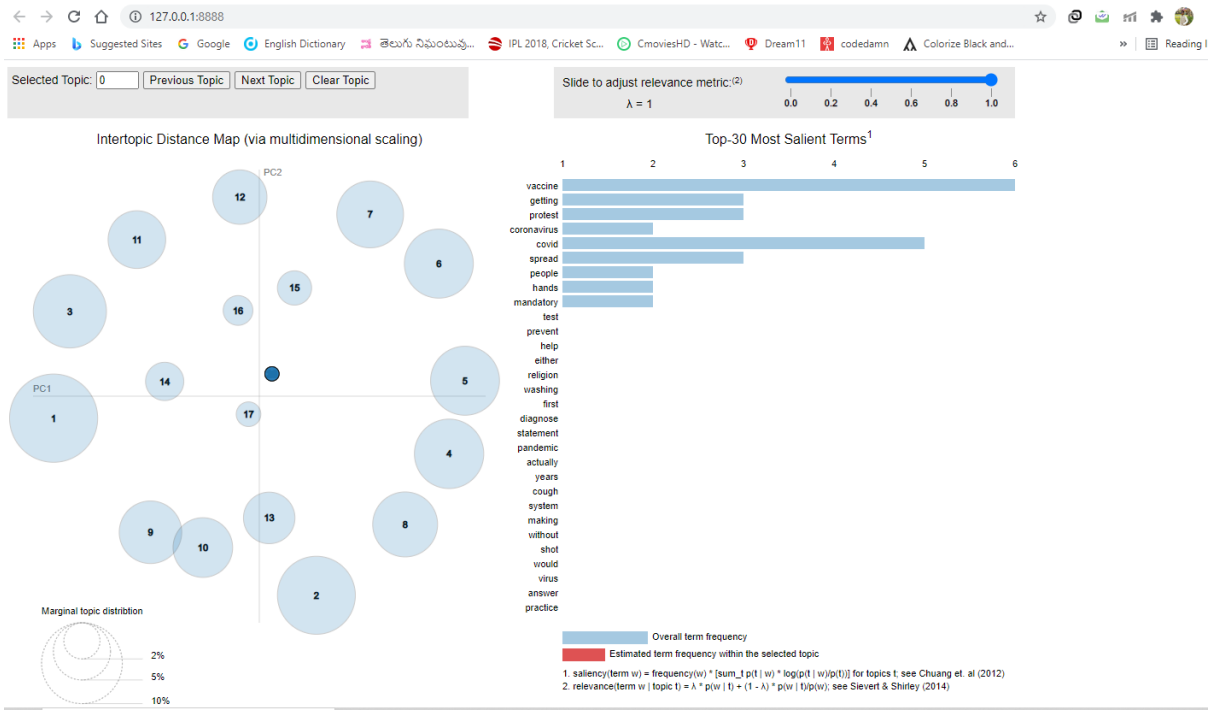


Fig:9.2 Pro-vax and Anti-vax Classification

CHAPTER-X

10. CONCLUSION AND FUTURE ENHANCEMENT

1. These findings suggest that the online anti-vax community is developing a more diverse and hence more broadly accommodating discussion around COVID-19 than the pro-vax community.

2. We pursue an automated, machine learning approach that avoids the scalability limitations of manual content analysis.

3. An popular approach of Elbow method is used to find out K value in K-Means Clustering, as a part of Un-Supervised Machine learning Algorithm.

4. At last our project is achieved, using different GUI based application like tKinter, Python library like Genism and NLTK platform for building code in Python.

5. The Machine can't analyze Human Emotions and Sarcasm. So, In our project the machine can't exactly predict the Sarcastic posts that are praising the community but in action they might be against the community.

6. Advancement in technology of machine learning can overcome this problem, which results in complete accuracy in successive days at a larger extent.

CHAPTER-XI

11. REFERENCES

1. A. Kata, ‘‘A postmodern Pandora’s box: Anti-vaccination misinformation on the Internet,’’ Vaccine, vol. 28, no. 7, pp. 1709–1716, Feb. 2010, doi: 10.1016/j.vaccine.2 009.12.022.

2. L. Givetash, Global measles cases surge amid stagnating vaccinations. New York, NY, USA: NBC News, 2019. Accessed: Apr. 13, 2020. [Online]. Available: https://www.nbcnews.com/news/world/globalmeasles-cases-surge-amid-decline-vaccinations-n1096921

3. B. Martin, Texas Anti-Vaxxers Fear Mandatory COVID-19 Vaccines More Than the Virus Itself. Austin, TX, USA: Texas Monthly, 2020. [Online]. Available: https://www.texasmonthly.com/news/texasanti-vaxxers-fear-mandatory-coronavirus-vaccines

4. H. J. Larson, ‘‘Blocking information on COVID-19 can fuel the spread of misinformation,’’ Nature, vol. 580, no. 7803, p. 306, Apr. 2020, doi: 10.1038/d41586-020-00920-w

5. R. Schraer and E. Lawrie, Coronavirus: Scientists Brand 5G Claims https://www.bbc.com/news/52168096

6. Coronavirus Disease (COVID-19) Advice for the Public: Myth Busters, W. H. Organization, Geneva, Switzerland, 2020. Accessed: Apr. 13, 2020. [Online]. Available: https://www.who.int/emergencies/diseases/novelcoronavirus-2019/advice-for-public

7. K. Benner and M. Shear, After Threats, Anthony Fauci to Receive Enhanced PersonalSecurity.Available:https://www.nytimes .com/2020/04/01/us/politics/coronavirus-faucisecurity.html

8. S. Almasy, H. Yan, and M. Holcombe, Coronavirus Pandemic Hitting Some African-American Communities Extremely Hard. New York, NY, USA: CNN Health, 2020. Accessed: Apr. 13, 2020. [Online]. Available: https://www.cnn.com/2020/04/06/health/us-coronavirus-updatesmonday/index.html

9. A. Maqbool, Coronavirus: Why Has The Virus Hit African Americans so Hard. London, U.K.: BBC News, 2020. Accessed: Apr. 13, 2020. [Online]. Available: https://www.bbc.com/news/world-us-canada-52245690

10. J. Guy. (2020). East Asian Student Assaulted in ’Racist’ Coronavirus Attack in London. Accessed: Apr. 13, 2020. [Online]. Available: https://www.cnn.com/2020/03/03/uk/coronavirus-assault-student-londonscli-intl-gbr/index.html

11. H. Yan, N. Chen, and D. Naresh. (2020). What’s Spreading Faster Than Coronavirus in the US? Racist Assaults and Ignorant Attacks Against Asians. Accessed: Apr. 13, 2020. [Online].

Available: https://www.cnn.com/2020/02/20/us/coronavirus-racist-attacks-againstasian-americans/index.html.

12. M. Rajagopalan, Korean Interpreter Says Men Yelling 'Chinese' Tried to Punch Her Off Her Bike. New York, NY, USA: BuzzFeed News, 2020. Accessed: Apr. 13, 2020. [Online]. Available: https://www.buzzfeednews.com/article/meghara/coronavirus-racismeurope-covid-19

13. K. Schaeffer, Nearly Three-in-ten Americans Believe COVID-19 was Made in a Lab, Washington, DC, USA: Pew Fact Tank, 2020. Accessed: Apr. 10, 2020. [Online]. Available: https://www.pewresearch.org/facttank/2020/04/08/nearly-three-in-ten-americans-believe-covid-19-wasmade-in-a-lab/

14. R. Iyengar, The Coronavirus is Stretching Facebook to its Limits. New York, NY, USA: CNN Business, 2020. Accessed: Apr. 13, 2020. [Online]. Available: https://www.cnn.com/2020/03/18/tech/zuckerbergfacebook-coronavirus-response/index.html

15. S. Frenkel, D. Alba, and R. Zhong. (2020). Surge of Virus Misinformation Stumps Facebook and Twitter. Accessed: Apr. 13, 2020. [Online]. Available: https://www.nytimes.com /2020/03/08 /technology/coronavirusmisinformation-social-media.html

CHAPTER-XII

12. PUBLICATIONS

INTERNATIONAL CONFERENCE ON INNOVATIONS IN COMPUTER NETWORKS, COMPUTATIONAL INTELLIGENCE AND IOT [ICICCI-2021].(Advanced self-assessment of Global pandemics like COVID-19 for healthy race using Machine Learning).

Advanced Self-assessment of Global pandemics like COVID-19 for healthy race using Machine Learning

Nishna Nayana Reddy Nimmala[1], Nuchu Anurudh Yadav[2], Parvatam Pavan Kumar[3], Shikakolli Sandeep[4] Dr.G.GovindaRajulu[5]

[1,2,3,4] UG Scholar [5]Professor
Department of Computer Science and Engineering,
St. Martin's Engineering College,
Near Forest Academy, Dulapally, Kompally, Secunderabad, Telangana 500 014, India
E-Mail:nishnareddy99@gmail.com[1], anurudhyadav010@gmail.com[2], pavankumar41299@gmail.com[3], shikakollisandeep645@gmail.com[4], drgovindacse@gmail.com[5]

1.ABSTRACT

A huge amount of false content regarding this dangerous virus is shared online. In this project we use machine learning to quantify COVID-19 content, which is falsely appearing, online, which leads to establishment of health guidance, particularly about vaccinations. We found that the anti-vax community is developing a less focused debate around COVID-19 than its counterpart, the pro-vaccination community. However, the anti-vax community exhibits a broader range of topics related to COVID-19, and hence the information can appeal to a broader cross-section of individuals seeking COVID-19 guidance online, For example individuals wary of a mandatory fast-tracked COVID-19 vaccine or those seeking alternative remedies. Hence the anti-vax community looks better positioned to attract fresh support going forward when compared to pro-vax community. The popularity of anti-vax community leads widespread lack of adoption of a COVID-19 vaccine, which means the world falls short of providing herd immunity, leaving countries open to future COVID-19 resurgences. We provide a mechanistic model that interprets these results and could help in assessing the likely efficacy of intervention strategies. Our approach is scalable and hence tackles the urgent problem facing social media platforms of having to analyse huge volumes of online health misinformation.

2.KEYWORDS

COVID-19: social computing: machine learning: mechanistic model: topic modelling: Pandemics

3.INTRODUCTION

As the vaccination is the cure for many dreadful viruses. The experts in Science agree that defeating COVID-19 will depend on developing a vaccine. However, The vaccination is best way to get herd immunity in large proportion of people. As we know vaccines tend to be less effective in older people, which leads to increase in rate of vaccination in younger generations which guarantees the higher immunities .Earlier we observed so many people opposing vaccinations due to some false information regarding vaccines side effects. For example some parents are against measles vaccination and even refused to vaccinate their children, this lead to increase in no of cases in 2019 measles outbreak in the U.S and beyond . Therefore any future COVID-19 vaccine will likely face similar opposition.

Online social media platforms, and in particular some built in community platforms like Facebook and Twitter, We have two communities anti-vax and pro-vax debating regarding COVID-19 vaccination. So this helps vaccine opponents (anti-vax) to create and share health misinformation such misinformation can endanger public health and individual safety. Likewise, vaccine supporters (pro-vax) also congregate in such online communities to discuss and advocate for professional public health guidance. Well before COVID-19, there was already an intense online conflict featuring anti-vax communities and pro-vax communities. Within anti-vax communities, the narrators typically draw on and generate false information about establishment medical guidance and distrust of the government, pharmaceutical industry, and new technologies such as 5G communications.

4.LITERATURE SURVEY

A. Kata, "A postmodern Pandora's box: Anti-vaccination misinformation on the Internet," Vaccine, vol. 28, no. 7, pp. 1709–1716,Feb.2010,doi:10.1016/j.vaccine.2009.12.022.

The Internet plays a large role in disseminating anti-vaccination information. This paper builds upon previous research by analysing the arguments proffered on anti-vaccination websites, determining the extent of misinformation present, and examining discourses used to support vaccine objections. Arguments around the themes of safety and effectiveness, alternative medicine, civil liberties, conspiracy theories, and morality were found on the majority of websites analysed; misinformation was also prevalent. The most commonly proposed method of combating this misinformation is through better education, although this has proven ineffective. Education does not consider the discourses supporting vaccine rejection, such as those involving alternative explanatory models of health, interpretations of parental responsibility, and distrust of expertise. Anti-vaccination protestors make postmodern arguments that reject biomedical and scientific "facts" in favour of their own interpretations. Pro-vaccination advocates who focus on correcting misinformation reduce the controversy to merely an "educational" problem rather,

CHAPTER-XIII

Nimmala Nishna Nayana Reddy is currently pursuing her Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. She completed her intermediate from Sri Chaitanya Junior College and studied 10th Standard at KKR's Gowtham School. Her technical skills include C and Python. She also has a basic understanding of C++. Her participations include: National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020, "India's First Leadership Talk" by MHRD'S Innovation Cell which was conducted on 16th May 2020, "What After College" Machine Learning workshop conducted by Kyrion Technologies Pvt. Ltd. on 14th and 15th February 2020 at Indian Institute of Technology (IIT) Hyderabad, "Talent Awareness Month" (TAM) conducted Machine Learning Workshop on 8th and 9th February 2020 in St. Martin's Engineering college and "The Entrepreneurship Summit" conducted at MLRIT, Hyderabad on 21st and 22nd August 2017 jointly organized by Nucleus Tech and SUMVN. Her areas of interest are Python, Machine Learning, Cloud computing and Kubernetes. She completed few certification courses from online platform like Coursera.

Nuchu Anurudh Yadav is currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. He completed his intermediate from Sri Gayatri Junior College and 10th class from St.Peter's High School. His technical skills include C, Python and Java. He also has a basic understanding of C++. He took part in Employability Skill development Program conducted by Zensar. He is also a student of Smart Interviews Where they have given some targets on coding in different platforms like HackerRank,Codechef,Codeforces. His participations include: Two day National level seminar on "Recent trends in Cloud Computing, Fog and Edge computing" which was conducted on 18th and 19th June 2021, National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020,He had completed one month Internship program at Lasya IT Solution Pvt.Ltd,and IIC Online Sessions conducted by Institution's Innovation Council (IIC) of MHRD's Innovation Cell, New Delhi to promote Innovation, IPR, Entrepreneurship, and Start-ups among HEIs from 28th April to 22nd May 2020,He had participated in HTML and CSS Workshop of TAM event held from 5th January 2018 to 3rd February 2018. His areas of interest are Python, Data Science, Machine Learning and Cloud Computing. He completed few certification courses from online platforms like Coursera and CursaApp

Parvatam Pavan Kumar is currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. He completed his Intermediate from Sri Chaitanya Junior College, KPHB and SSC from Siddhartha School of Excellence, Vikarabad. He Resides in Kompally. Being a CSE Student, he was well versed in Programming Languages like C, C++, C#, Java and Web development. He also has a basic understanding of Python and JavaScript. He is taking an active part in of Smart Interviews Club that target on coding in different platforms. He also took part in Employability Skill development Program conducted by Zensar and he was a part of Cyber Security club.

His participations include: Two day National level seminar on "Recent trends in Cloud Computing, Fog and Edge computing",National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing", "Webinar on ML for real world applications", He had completed one month Internship program at Lasya IT Solution Pvt.Ltd, IIC Online Sessions conducted by Institution's Innovation Council (IIC) of MHRD's Innovation Cell, New Delhi to promote, Innovation, IPR, Entrepreneurship, Start-ups among HEIs and "India's First Leadership Talk". "What After College" Machine Learning workshop conducted by Kyrion Technologies Pvt. Ltd. At Indian Institute of Technology (IIT) Hyderabad, C graphics Workshop at Siva Sivani Group of Institutions. He has successfully finished Internships on "Data Analytics", "Machine Learning", "Leader Ship and soft Skill development" in Internshala sponsored by AAM AADMI PARTY. His areas of interest are Data Science, Machine Learning and Cloud Computing. He completed a few certification courses from various online platforms.

He is placed in **TATA CONSULTANCY SOLUTIONS** (TCS) in college placement drive.

Shikakolli Sandeep is currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. He completed his intermediate from Sri Chaitanya Junior College and 10th from Marvel High School. His technical skills include C, C++, Python and Java. Web technologies like HTML5, CSS3 and BOOTSTRAP. He is also a student of Smart Interviews. He took part in Employability Skill development Program conducted by Zensar Technologies. His participations include:" Machine Learning workshop" at IIT Hyderabad which was conducted by Kyrion technologies Pvt Ltd from 14th t0 15th February 2020, National Level Three Day Online Workshop on "AI & ML in speech and Audio Processing" which was conducted from 10th to 12th December 2020, a Workshop on "C-Graphics" which was conducted by Siva Sivani Institute of Management from 6th to 7th January 2020, attended "Leadership Talk" by MHRD's Innovation Cell on 16th May 2020, National Level "project Expo and Competition" which was conducted on 28th March 2018 and attended a Workshop on HTML & CSS which was conducted by St. Martin's Engineering College on 5th January 2018. His area of interest are Web technologies, C++ and Creating Web pages. He completed few certification courses from online platform like Coursera.

CHAPTER-XIV

14. APPENDICES

As mentioned in the main text, the methodology starts with a seed of manually identified Facebook Pages discussing either vaccines, public policies about vaccination, or the provs-anti vaccination debate. Then their connections to other fan pages are indexed. At each step, new findings are vetted through a combination of human coding and computer assisted filters. This snowball process is continued, noting that new links can often lead back to members already in the list and hence some form of closure can in principle be achieved. This process leads to a set containing many hundreds of pages for both the anti-vax and pro-vax communities. Before training the LDA models, several steps are employed to clean the content of these pages in a similar way to other LDA analyses in the literature:

Step 1: Mentions of URL shorteners are removed, such as "bit.ly" since these are fragments output by Facebook's CrowdTangle API.

Step 2: Many of the posts link to external websites. The fact that these specific websites were mentioned could itself be an interesting component of the COVID-19 conversation. Hence instead of removing them completely, the pieces ".gov", ".com", and ".org" were replaced with "__gov", "__com", and "__org", respectively. This operation effectively concatenates domains into a form that will not be filtered out by the later preprocessing steps.

Step 3: The posts are then run through Gensim's simple_preprocess function, which tokenizes the post on spaces and removes tokens that are only 1 or 2 characters long. This step also removes numeric and punctuation characters.

Step 4: Tokens that are in Gensim's list of stopwords, are removed. For example, "the" is not a good indication of a topic.

Step 5: Tokens are lemmatized using the WordNetLemmatizer from the Natural Language Toolkit NLTK, which converts all words to singular form and/or present tense.

Step 6: Tokens are stemmed using the SnowballStemmer from NLTK, which removes affixes on words.

Step 7: Any remaining fragments of URLs (other than domain) that are left over after stemming, such as "http" and "www", are removed.

Steps 5 and 6 help ensure that words are compared fairly during the training process, and that if a particular word is a strong indicator of a topic, its signal is not lost just because it is used in many different forms. These steps rely on words existing in NLTK's pretrained vocabulary. Any word not in this vocabulary is left unchanged. After this preprocessing, we then train the LDA models on the cleaned data. Specifically, 10 separate LDA models were trained with the "number of topics" parameter ranging from 3-20, for a total of 180 models in each of the two time intervals T1 and T2. The CV coherence algorithm was then run over each of these models and the coherence scores were then averaged for each number of topics. To produce the results, multiple trials were run for each number of topics to ensure that the coherence for a particular number of topics is representative of what LDA models tend to find (and by extension a better fit for the data) and is not the result of unaccounted noise swaying the model to overfit in one way or another. These trials are independent because the random number generator for each LDA model was initialized with a different seed, ensuring that statistical inferences were not be repeated. The GitHub link is: https://github.com/searri/social-clusteringresearch/wiki/Coronavirus-Vax.

The following illustrates the topic output, focusing on anti-vax in time interval T2 in Fig. 2B. In this, 9 of the 10 topics had the word 'coronavirus' among the 5 highest weighted words in the topic; 4 were focused around 'coronavirus' and 'vaccine' co-occuring together. Others had 'vitamin', 'fear' and 'ddt' in relation to alternative treatments, and 'weapon' related to conspiracy theories of COVID-19's origin. Within one of the topics, which is focused around alternative health explanations and cures with words like 'vitamin' etc., illustrative posts include the following from Feb 8, 2020 in one of the 'Coalition for Vaccine Choice' pages, with spelling mistakes left as is: "The story of this FAKE "epidemic" with the "corona virus" from China is a cover-up story for the grim reality of the health problems due to 5G technology exposure coroborated with a lot of other factors: vaccination, poor alimentation in vitamins, bad water, air pollution, lack of sleep, etc.... scientists have shown that low level microwave EMF exposure can result in VGCC activation and elevated intracellular calcium". Meanwhile, for a topic focused on conspiracy theories with words such as 'weapon' and 'fear', an example phrase from a posting is "..keeping the world under the thumb of tyrants! You are soldiers, and that means that you are expendable by your trained nature. You are being micro-managed by people that give not one caring thought of you, five thousand miles away, that know little of the true nature of the battle". This illustrates the type of detailed analyses that we carried out to check our automated approach, and which underlie our claim that the groupings do correspond to reasonably distinct conversation topics.

A

PROJECT REPORT

On

Blockchain for Secure EHRs Sharing of Mobile Cloud Based E-Health Systems

Submitted by

1) Ms.Kothapu Lakshmi Narayana Chandana (17K81A0592)

2) Mr.Pitla Prem Kumar (17K81A05A5)

3) Ms.Vallandas Ramya(17K81A05B4)

4) Mr.Vanam Thrishul (17K81A05B5)

in partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Under the Guidance of

Mr.Mruthyunjayam Allakonda

Assistant Professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ST.MARTIN'S ENGINEERING COLLEGE
An Autonomous Institute

Dhulapally, Secunderabad – 500 100

JUNE 2021

BONAFIDE CERTIFICATE

This is to certify that the project entitled **Blockchain for Secure EHRs Sharing of Mobile Cloud Based E-Health Systems**, is being submitted by **Ms.Kothapu Lakshmi Narayana Chandana 17K81A0592, Mr.Pitla Prem Kumar 17K81A05A5, Ms.Vallandas Ramya 17K81A05B4, Mr.Vanam Thrishul 17K81A05B5** in partial fulfillment of the requirement for the award of the degree of **BACHELOR OF TECHNOLOGY IN COMPUTER SCIENCE AND ENGINEERING** is recorded of bonafide work carried out by them. The results embodied in this report have been verified and found satisfactory.

Mr.Mruthyunjayam Allakonda	**Dr.M.NARAYANAN**
Assistant Professor	Head of the Department
Department of CSE	Department of CSE

Internal Examiner	External Examiner

Place:

Date:

DECLARATION

We, the student of **Bachelor of Technology** in Department of Computer Science and Engineering', session: 2017 – 2021, St. Martin's Engineering College, Dhulapally, Kompally, Secunderabad, hereby declare that work presented in this Project Work entitled **Blockchain for Secure EHRs Sharing of Mobile Cloud Based E-Health Systems** is the outcome of our own bonafide work and is correct to the best of our knowledge and this work has been undertaken taking care of Engineering Ethics. This result embodied in this project report has not been submitted in any university for award of any degree.

Kothapu Lakshmi Narayana Chandana 17K81A0592

Pitla Prem Kumar 17K81A05A5

Vallandas Ramya 17K81A05B4

Vanam Thrishul 17K81A05B5

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible and whose encouragement and guidance have crowded effects with success.

We extended our deep sense of gratitude to Principal**, Dr. P. SANTOSH KUMAR PATRA**, St. Martin's Engineering College, Dhulapally, for permitting us to undertake this project.

We are also thankful to **Dr. M.NARAYANAN**, Head of the Department, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

We would like to thank **Dr. T. POONGOTHAI,** Department of Computer Science and Engineering (AI & ML) for her encouragement and insightful comments and as well as our project coordinators **DR.B.RAJALINGAM**, Associate Professor and **Dr. G.GOVINDARAJULU**, Assistant Professor, in the Department of Computer Science and Engineering for their valuable support.

We would like to express our sincere gratitude and indebtedness to our project supervisor Mr. Mruthyunjayam Allakonda, Assistant Professor, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

Finally, we express thanks to all those who have helped us successfully to complete this project. Furthermore, we would like to thank our family and friends for their moral support and encouragement.

We express thanks to all those who have helped us in successfully completing the project.

Kothapu Lakshmi Narayana Chandana	17K81A0592
Pitla Prem Kumar	17K81A05A5
Vallandas Ramya	17K81A05B4
Vanam Thrishul	17K81A05B5

ABSTRACT

Late years we have seen a change in perspective of Electronic Health Records (EHRs) on versatile cloud conditions where cell phones are coordinated with distributed computing to work with clinical information trades among patients and medical services suppliers. This high level model empowers medical care administrations with low operational expense, high adaptability and EHRs accessibility. Notwithstanding, this new worldview likewise raises worries about information protection and organization security for e-wellbeing frameworks. Instructions to dependably divide EHRs between versatile clients while ensuring high security levels in portable cloud is a difficult issue. In this paper, we propose a novel EHRs sharing structure that joins blockchain and the decentralized interplanetary record framework (IPFS) on a versatile cloud stage. In Particular, we plan a dependable access control component utilizing savvy agreements to accomplish secure EHRs dividing between various patients and clinical suppliers. We present a model execution utilizing Ethereum blockchain in a genuine information sharing situation on a portable application with Amazon distributed computing. Observational outcomes show that our proposition gives a successful answer for dependable information trades on portable mists while safeguarding delicate wellbeing data against likely dangers. The framework assessment and security examination additionally show execution upgrades in lightweight access control plan, least organization inertness with high security and information protection levels, contrasted with existing information sharing models.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF OUTPUT SCREENS

LIST OF ACRONYMS

<EHR>	Electronic Health Records
<MCC>	Mobile Cloud Computing
<IoMT>	Internet of Medical Things
<IPFS>	Interplanetary File System
<AWS>	Amazon Web Services
<IoV>	Internet of Vehicles
<NC>	Network Coded
<DS>	Distributed Storage
<DSB>	Distributed Storage Blockchain
<LSS>	Local Secret Sharing
<CSP>	Cloud Service Provider
<DFDS>	Data Flow Diagrams

1.INTRODUCTION

As of late, there has been a developing interest in utilizing blockchain innovation to advance clinical and e-wellbeing administrations [1]-[3]. Blockchain with its decentralized and trustworthy nature has shown tremendous possibilities in different e-wellbeing areas like secure sharing of Electronic Health Records (EHRs) and information access to the executives among numerous clinical substances [4]-[6]. In this manner, the reception of blockchain can give promising answers for work with medical services conveyance and subsequently alter the medical services industry .With the development of creative advances, including Mobile Cloud Computing (MCC) and Internet of Medical Things (IoMT), the medical care industry has seen huge changes in e-wellbeing tasks [7], [8]. Patients presently can gather their own wellbeing data locally established on cell phones, (for example, cell phones and wearable sensors) and offer cloud conditions where medical care suppliers can get to immediately to investigate clinical records and give convenient clinical backings. This brilliant e-wellbeing administration permits medical care suppliers distantly screen patients and offer wandering consideration at home, which works with medical services conveyance as well as carries monetary advantages to patients. Further, the accessibility of complete EHRs on mists likewise helps medical care suppliers track patient wellbeing and offers appropriate clinical benefits during determination and therapy measures [9]. Other than every one of these extraordinary benefits, notwithstanding, the pattern of EHRs stockpiling on mists additionally presents security challenges which ruin the organization of e-wellbeing applications on mists [10], [11]. Among such security issues is secure EHRs dividing among patients and medical services suppliers on versatile cloud conditions. Unapproved substances may acquire malignant admittance to EHRs without the assent of patients, which inconveniently affects information honesty, protection and security of cloud e-wellbeing frameworks [12]. Also, patients may think that it's hard to follow and deal with their wellbeing records divided between medical care suppliers on mists. It therefore is important to propose effective access control arrangements for mobile cloud EHRs sharing systems.Traditional access control approaches [14]-[16] for EHR sharing accept that cloud workers are completely trusted by

data owners and empower the workers to play out all entrance control and confirmation rights on information utilization. Be that as it may, this supposition no longer holds in portable mists since the cloud worker is straightforward however inquisitive. The cloud cut off will sincerely play out the information demands, however in the interim will get individual data without assent of clients, which prompts genuine data spillage issues and organization security, as needs be. All the more critically, ordinary access control frameworks chiefly depend on a predefined 2 point of access, for example a brought together cloud worker, and this can prompt the main issue of disappointment for e-wellbeing networks [25]. In the meantime, blockchain-based admittance control gives different new security highlights to e-wellbeing with incredible benefits over customary access control arrangements. To start with, the blockchain develops permanent records of exchanges for information sharing framework [5]. This implies that exchanges recorded in the blockchain can't be adjusted or modified by any elements and exchanges are simply composed to blockchain while recuperation activities are not allowed. This ensures high framework dependability and uprightness. Second, access control utilizing blockchain can accomplish the straightforwardness property with the capacity of addressing successfully the issue of information spillage which can be brought about by inquisitive workers. Any unlawful access of workers and different substances to information stockpiling will be thought about the blockchain and broadcast to all organized members. Along these lines, any blockchain client can handle information access and identify such malignant exchanges for preventive activities. Third, the utilization of blockchain-based keen agreement [21] can accomplish the confirmation and client check property. By implementing exacting access control approaches, shrewd agreements can approve viably client admittance to wellbeing information stockpiling just as distinguish and forestall successfully possible dangers to wellbeing networks in an appropriate way. Last, blockchain combined with the brilliant agreement innovation disposes of the dependence on focal workers to guarantee reasonableness among exchange parties. As the keen agreements are public on the blockchain, every one of the associated elements on the blockchain organization will have a duplicate of them, which gives an equivalent option to

control all agreement tasks. Exceptionally, blockchain based access control with its circulated nature can function admirably when any gathering falls flat without loss of information, dangers and trust concerns [5]. Persuaded by such benefits of blockchain, in this paper, we propose another EHRs sharing model on a versatile cloud stage dependent on the blockchain method. The establishment of our proposed framework is a client access control structure to oversee information access from network substances. Access control components are able to do viably limiting illicit admittance to EHRs assets, while guaranteeing quick information recovery for approved entities.The paper makes the accompanying commitments:

• We propose a novel EHRs sharing engineering dependent on blockchain and decentralized capacity interplanetary document framework (IPFS) for an e-wellbeing framework on a portable cloud stage. To improve security of EHRs sharing, we foster a reliable access control instrument utilizing shrewd agreements. Further, an information sharing convention is intended to oversee client admittance to our EHRs framework.

• We explore the exhibition of the proposed EHRs sharing model through ease of use testson a portable Android application and distributed computing given by Amazon Web Services (AWS). The assessment results show that the proposed strategy is achievable to different e-wellbeing situations.

• We give a security investigation and broad assessment in different execution measurements to feature the benefit of the proposed structure over current EHRs sharing arrangements. The rest of our paper is coordinated as follows. The Section II presents cutting edge concentrates towards the utilization of access control arrangements and blockchain for EHRs sharing on mists. Section III presents blockchain ideas and its principle parts that will be used in this paper. In the Section IV, we propose a framework model for the EHRs sharing plan utilizing blockchain and savvy agreements, and plan goals are likewise introduced. Then, we present a model execution of decentralized admittance control for EHRs partaking in a versatile cloud blockchain network in the Section V. Brilliant agreement plan and information sharing convention are likewise

dissected. The trial results are talked about in the Section VI, while security examination and framework assessments are given in the Section VII. At last, our decisions are attracted to Section VIII.

1.1 PROJECT OVERVIEW

This paper proposes a novel EHRs sharing scheme enabled by mobile cloud computing and blockchain. We identify critical challenges of current EHRs sharing systems and propose efficient solutions to address these issues through a real prototype implementation. In this work, our focus is on designing a trustworthy access control mechanism based on a single smart contract to manage user access for ensuring efficient and secure EHRs sharing. To investigate the performance of the proposed approach, we deploy an Ethereum blockchain on the Amazon cloud, where medical entities can interact with the EHRs sharing system via a developed mobile Android application. We also integrate the peer-to-peer IPFS storage system with blockchain to achieve decentralized data storage and data sharing. The implementation results show that our framework can allow medical users to share medical data over mobile cloud environments in a reliable and quick manner, in comparison to conventional schemes. In particular, our access control can identify and prevent unauthorized access to the e-health system, aiming for achieving a desired level of patient privacy and network security. We also provide security analysis and extensive evaluations on various technical aspects of the proposed system, showing advantages of our proposal over existing solutions. Based on the merits of our model, we believe that our blockchain enabled solution is a step towards efficient management of e-health records on mobile clouds, which is promising in many healthcare applications.

1.2 PROJECT OBJECTIVE

Recent years have witnessed a paradigm shift in the storage of Electronic Health Records(EHRs) on mobile cloud environments, where mobile devices are integrated with cloud computing to facilitate medical data exchanges among patients and healthcare providers. This advanced model enables healthcare services with low operational cost,

high flexibility, and EHRs availability. However, this new paradigm also raises concerns about data privacy and network security for e-health systems. How to reliably share EHRs among mobile users while guaranteeing high-security levels in the mobile cloud is a challenging issue. In this paper, we propose a novel EHRs sharing framework that combines blockchain and the decentralized interplanetary file system (IPFS) on a mobile cloud platform. Particularly, we design a trustworthy access control mechanism using smart contracts to achieve secure EHRs sharing among different patients and medical providers. We present a prototype implementation using Ethereum blockchain in a real data sharing scenario on a mobile app with Amazon cloud computing. The empirical results show that our proposal provides an effective solution for reliable data exchanges on mobile clouds while preserving sensitive health information against potential threats. The system evaluation and security analysis also demonstrate the performance improvements in lightweight access control design, minimum network latency with high security and data privacy levels, compared to the existing data sharing models.

1.3 ORGANIZATION OF CHAPTERS

CHAPTER-1:

In this chapter of Introduction, the information is about the preface of our project which gives details related to overview and objectives of the project.

CHAPTER-2:

The chapter of Literature survey gives the matter related to some existing articles in detail, which are related to our project research and we also conclude the survey about related references.

CHAPTER-3:

Requirements of the project are given in this chapter. Both software and hardware requirements were provided.

CHAPTER-4:

In the chapter of software development analysis, we discuss overview and definition of problems and also overview, definition and functionality of modules.

CHAPTER-5:

The pictorial representation of our project is in this chapter. The UML diagrams are used to give system design of our project.

CHAPTER-6:

In this chapter we gather information about coding of the project. The templates of code, files used in code, functionalities of class, methods, inputs and output parameters are gathered.

CHAPTER-7:

The testing phase of our project is documented in this chapter, different types of testing like Black box and white box are defined along with test cases of the project.

CHAPTER-8:

This chapter gives output screens of our project. In which the user interface and outputs of the project are understood clearly.

CHAPTER-9:

The experimental results of the project are represented in this chapter.

CHAPTER-10:

The documentation of the project is concluded in this chapter along with future enhancement, references, paper publication, student profile and appendices.

2.LITERATURE SURVEY

“An energy-efficient transaction model for the blockchain-enabled Internet of Vehicles (IoV),”

The blockchain is a safe, reliable and innovative mechanism for managing numerous vehicles seeking connectivity. However, following the principles of the blockchain, the number of transactions required to update ledgers pose serious issues for vehicles as these may consume the maximum available energy. To resolve this, an efficient model is presented in this letter which is capable of handling the energy demands of the blockchain enabled Internet of Vehicles (IoV) by optimally controlling the number of transactions through distributed clustering. Numerical results suggest that the proposed approach is 40.16% better in terms of energy conservation and 82.06% better in terms of the number of transactions required to share the entire blockchain data compared with the traditional blockchain

“On scaling decentralized blockchains,”

The increasing popularity of blockchain-based cryptocurrencies has made scalability a primary and urgent concern. We analyze how fundamental and circumstantial bottlenecks in Bitcoin limit the ability of its current peer-to-peer overlay network to support substantially higher throughputs and lower latencies. Our results suggest that reparameterization of block size and intervals should be viewed only as a first increment toward achieving next-generation, high-load blockchain protocols, and major advances will additionally require a basic rethinking of technical approaches. We offer a structured perspective on the design space for such approaches. Within this perspective, we enumerate and briefly discuss a number of recently proposed protocol ideas and offer several new ideas and open challenges.

“A low storage room requirement framework for distributed ledger in blockchain,”

Traditional centralized commerce on the Internet relies on trusted third parties to process electronic payments. It suffers from the weakness of the trust-based model. A pure decentralized mechanism called blockchain tackles the above problem and has become a

hot research area. However, since each node in a blockchain system needs to store all transactions of the other nodes, as time continues, the storage room required to store the entire blockchain will be huge. Therefore, the current storage mechanism needs to be revised to cater to the rapidly increasing need for storage. Network coded (NC) distributed storage (DS) can significantly reduce the required storage room. This paper proposes a NC-DS framework to store the blockchain and proposes corresponding solutions to apply the NC-DS to the blockchain systems. Analysis shows that the proposed scheme achieves significant improvement in saving storage room.

"Distributed storage meets secret sharing on the blockchain,"

Blockchain systems establish a cryptographically secure data structure for storing data in the form of a hash chain. We use a novel combination of distributed storage, private key encryption, and Shamir's secret sharing scheme to distribute transaction data, without significant loss in data integrity. Additionally, using Shamir's secret sharing scheme on the hash values and dynamic zone allocation, we further enhance the integrity. We highlight the tradeoff in storage cost and data loss probability with varying zone size choices. We also study the tradeoff between recovery cost and security from adversarial corruption with varying recovery mechanisms. Then, we formulate code design, given a probability of data recovery and targeted corruption, as an integer program. Using the coding scheme we establish a mechanism to insure data, for instance in blockchain-based cloud storage systems, based on the value of the data, by understanding the costs involved for the service provider.

"Efficient local secret sharing for distributed blockchain systems,"

Blockchain systems store transaction data in the form of a distributed ledger where each peer is to maintain an identical copy. Blockchain systems resemble repetition codes, incurring high storage cost. Recently, distributed storage blockchain (DSB) systems have been proposed to improve storage efficiency by incorporating secret sharing, private key encryption, and information dispersal algorithms. However, the DSB results in significant communication costs when peer failures occur due to denial of service attacks. In this

letter, we propose a new DSB approach based on a local secret sharing (LSS) scheme with a hierarchical secret structure of one global secret and several local secrets. The proposed DSB approach with LSS improves the storage and recovery communication costs.

2.1 SURVEY ON BACKGROUND

Scenario 1:Primary Patient Care

Using blockchain technology for primary patient care can help to address the following problems of the current healthcare systems:

A patient often visits multiple disconnected hospitals. He has to keep the history of all his data and maintain the updates. This leads to the situation when required information may not be available.

Due to the unavailability of the data, patients may have to repeat some tests for laboratory results. This is common when the results are stored in another hospital and can not be immediately accessed.

The healthcare data are sensitive and their management is cumbersome. Yet, there is no privacy-preserving system in clinical practice that allows patients to maintain access control policy in an efficient manner.

Sharing data between different healthcare providers may require major effort and could be time consuming.

Next, we propose two approaches that can be implemented separately or combined to improve patient care.

Institution-based: The network would be formed by the trusted peers: healthcare institutions or general practitioners (caregivers). The peers will run consensus protocol and maintain a distributed ledger. The patient (or his relatives) will be able to access and manage his data through an application at any node where his information is stored. If a peer is off-line, a patient could access the data through any other online node. The key

management process and the access control policy will be encoded in a chaincode, thus, ensuring data security and patient's privacy.

Case specific (serious medical conditions, examination, elderly care): During a patient's stay in a hospital for treatment, rehabilitation, examination, or surgery, a case-specific ledger could be created. The network would connect doctors, nurses, and family to achieve efficiency and transparency of the treatment. This will help to eliminate human-made mistakes, to ensure consensus in case of a debate about a certain stage of the treatment.

Scenario 2: Data Aggregation for Research Purposes.

It is highly important to ensure that the sources of the data are trusted medical institutions and, therefore, the data are authentic. Using shared distributed ledger will provide traceability and will guarantee patients' privacy as well as the transparency of the data aggregation process. Due to the current lack of appropriate mechanisms, patients are often unwilling to participate in data sharing. Using blockchain technology within a network of researchers, biobanks, and healthcare institutions will facilitate the process of collecting patients' data for research purposes.

Scenario 3: Connecting Different Healthcare Players for Better Patient Care.

Connected health is a model for healthcare delivery that aims to maximize healthcare resources and provide opportunities for consumers to engage with caregivers and improve self-management of a health condition23. Sharing the ledger (using the permission-based approach) among entities (such as insurance companies and pharmacies) will facilitate medication and cost management for a patient, especially in case of chronic disease management. Providing pharmacies with accurately updated data about prescriptions will improve the logistics. Access to a common ledger would allow transparency in the whole process of the treatment, from monitoring if a patient follows the prescribed treatment, to facilitating communication with an insurance company regarding the costs of the treatment and medications[4].

2.2 CONCLUSION ON SURVEY

The proposed EHRs sharing system is discussed and evaluated under various performance metrics to demonstrate the feasibility of our model for real usability scenarios.

1) **Flexibility**

Since our design is deployed on a mobile platform, any users with smartphones can easily work on our system while allowing the freedom of users with high flexibility. Our system can work well with different mobile platforms, including Android and iOS versions, increasing the usability of our design in different healthcare systems.

2) **Availability**

Our system allows authorized mobile users to access ehealthcare records anytime and anywhere with a mobile application. The use of mobile app allows users to interact with our system in a real time and dynamic manner, with highly available medical data on cloud.

3) **Avoid single point of failure**

Our design employed the decentralized storage system IPFS that solves effectively the single point of failure problem. Besides, access control enabled by the blockchain technique is running in a peer-to-peer manner among decentralized entities that can also contribute to overcome this challenge.

4) **Integrity**

Integrity guarantees that patient information is shared between authorized users without any change. Medical records collected from mobile gateways are always encrypted to avoid any alterations. Meanwhile, for EHRs sharing, mobile users are unable to modify the signed transactions to smart contracts and no entities can tamper and change content of recorded transactions. Importantly, mobile users cannot have rights to change or alter the agreement in the smart contract and access policies in our scenario.

3.SOFTWARE AND HARDWARE REQUIREMENTS

3.1 SOFTWARE REQUIREMENTS

- Operating System : Windows XP/7.
- Coding Language : JAVA/J2EE.
- Data Base : MYSQL.
- IDE : NetBeans 8.2.

3.2 HARDWARE REQUIREMENTS

- System : Pentium IV 2.4 GHz.
- Hard Disk : 40 GB.
- Floppy Drive : 1.44 Mb.
- Monitor : 15 VGA Colour.
- Mouse : Logitech.
- Ram : 512 Mb.

4.SOFTWARE DEVELOPMENT ANALYSIS

4.1 OVERVIEW OF PROBLEM

Before the presentation of shrewd contacts on the blockchain, the principle conversations on Electronic Health Record (EHR) Management zeroed in on whether to utilize cloud foundations [39][41] or neighborhood incorporated frameworks for putting away and sharing EHRs. These concentrated frameworks suggested that every clinic and medical services organization would need to keep information on premise in privately oversaw structures and information bases. Notwithstanding, incorporated EHRs the executives frameworks present a few issues as depicted underneath:

• No quiet control: The patients don't claim the information and have no control over it. The patients ought to possess and control their information.

• Scattered records: As patients look for medicines in changed constructions, the records are repeated. The data gets dissipated.

• Limited framework interoperability: Different clinics and wellbeing offices have various frameworks. Mix and interoperability issues are the results.

• Inconvenient secure sharing: Oftentimes, the way toward sharing wellbeing records is unpredictable and tedious. In the U.S. a safe email standard called Direct is utilized to give scrambled transmission between the sender (for instance, an E.R. doctor) and beneficiary. A piece of the writing on EHRs the board resolves these issues by proposing unified structures and frameworks for sharing EHRs on cloud foundations [39][41]. Albeit these structures carried answers for a significant number of the difficulties recorded above, they actually experienced impediments particularly as identified with straightforwardness, information proprietorship, and security. Besides, the brought together model for EHRs the board battles in emergency and debacle situations in light of the fact that the reaction to the crisis is regularly sloppy and decentralized. Although catastrophic events are

uncommon occasions, they present new difficulties as the medical services area ought to be ready and ready to react to the emergency speedily [19]. Truth be told, flooding, waves, and seismic tremors might actually disturb offices and frameworks, in this way restricting the admittance to records and patient data. This is one of the contentions that show how decentralizing the the executives of EHRs and repeating and circulating the data can guarantee better execution and accessibility in debacle circumstances, contrasted with incorporated models [29]. A decentralized framework is a dispersed organization where no gathering has the full power over the information and the tasks, yet the choices are made on the whole through an agreement interaction. The gatherings shaping the organization are called hubs also, conveyed through message passing [10]. For the most part talking, by sharing also, reproducing the data, gives the organization accessibility and strength particularly if there should arise an occurrence of broad disappointments. Additionally, Peer-to-peer frameworks (P2P) can indeed, even give information proprietorship as the private data can be put away and mentioned distinctly to the restrictive hub. Be that as it may, arriving at agreement while protecting obscurity, security, and rightness regardless of disappointments has been a difficult issue concentrated in the writing [8][10][28][27]. The acquaintance of blockchain made it conceivable to accomplish it while protecting secrecy and giving security and recognizability [49]. A blockchain is an information structure where the records are put away in a connected arrangement of squares. This succession frames a disseminated record, which implies it is duplicated in various machines, called hubs, that speak with each other. The hubs structure a distributed organization where each update to the record should be acknowledged by the organization utilizing an agreement convention. The agreement convention guarantees that everyone has a similar view on the situation with the framework [49]. A blockchain can be carried out in two principle models: the authorization less, or public model, and the permissioned model. In the public model, any member can join and leave voluntarily in light of the fact that no standard confines access and cooperation. Thus, the information put away in a public blockchain (i.e., Bitcoin [37] or Ethereum [52]) is available by anybody except if encryption and keen agreement rationale are utilized.

4.2 DEFINE THE PROBLEM

Other than the public model, blockchain can likewise be utilized in a confined organization where the members' characters are known. This confined model is normally alluded to as permissioned or consortium. The model of investment has a critical impact on how the agreement is reached by the organization [50].

With the presentation of blockchain and the likelihood to make circulated furthermore, decentralized applications, new plans and arrangements embracing the decentralized way of thinking have begun to draw in interests by the scholarly community and the business [3][12][22][29][39][53][32]. These arrangements could be more impervious to disappointment than the focal one in crisis circumstances. The writing on blockchain for EHRs the executives has proposed frameworks dependent on both public and permissioned blockchain executions. Among the public arrangements, the first and generally significant for culmination and importance is MedRec [3]: an EHR the executives framework dependent on the public Ehtereum execution intended for information sharing and mix with current frameworks. In any case, MedRec and the biggest piece of the ensuing writing neglects to give definite examinations and tests. In reality, the greater part of the works center around the plan and execution of blockchain arrangements without examines in regards to execution and versatility. In addition, nearly each proposed arrangement doesn't consider the specific instance of a mass crisis circumstance produced by catastrophic events. Along these lines, it is imperative to dissect if and how a blockchain framework planned for wellbeing records would act in such situations. This suggests that an exact evaluation of safety and versatility, just as exchange throughput, should be done to be certain that the framework would meet the security and execution prerequisites. Among the necessities, it is feasible to list: the capacity to facilitate and participate through a protected means of correspondence; the capacity to ceaselessly arrive at life-saving data; and the capacity to let NGOs and rescuers join and access clinical records without bargaining security and classification. In this setting, blockchain is possibly the best arrangement on the grounds that the tasks during a calamity

are intrinsically decentralized and the assistance should be facilitated and on time: this is fundamental when the life or prosperity of somebody is in question.

4.3 MODULES OVERVIEW

Registration

Healthcare Provider

- Load patient Records
- Key Generation
- Encrypt patient Records
- Block Creation
- Upload and Download Patient Records

Cloud Service Provider

- View Patient Records
- Grant or Revoke Permission

4.4 DEFINE THE MODULES

Registration

It is a process of enrolling or being enrolled into the cloud. To utilize the cloud documents, every healthcare provider should enroll. During this process your basic information like email, contacts etc., are collected and stored in the Cloud. The cloud id for a particular user will get automatically generated during the registration.

Cloud ID

Every user should create a Cloud ID and use it to identify something with near certainty that the identifier does not duplicate one that has already been, or will be, created to identify something else. Information labelled with Cloud ID by independent parties can therefore be later combined into a single database, or transmitted on the same channel, without needing to resolve conflicts between identifiers

Healthcare Provider

• Load patient Records

• Key Generation

• Encrypt patient Records

• Block Creation

• Upload and Download Patient Records

Data Selection and Loading

In this process, the health provider chooses patient healthcare records for uploading and maintaining the dataset in the cloud.

Select Data

Load Data

Preview Data

Key Generation

The secret key is generated using a cryptographic algorithm. This key is used for encrypting the dataset.

Encrypt Patient Records

The data is encrypted for secure maintenance. So that the unauthorized person cannot be able to access the data that are presented in the cloud.

Block Creation

➢ Each block contain patient record and it's timestamp.

➢ A blockchain, originally block is a growing list of records called blocks.

Upload and Download Patient Records

After creating the block, the healthcare provider will upload the records into the cloud.Suppose, if they want to retrieve a record from the cloud, first the healthcare provider searches the record. Based on the search it will show the results. After getting an

approval and key from the cloud service provider the healthcare provider can download the data.

Cloud Service Provider

The cloud service provider maintains all the patient records and also they can provide a permission to the user to access the data.

The Cloud Service Provider can view all the uploaded and downloaded documents in the Cloud. The CSP receives the document request from the Data User, verifies the authentication before granting permission. Then the CSP executes the query and returns the encrypted document according to the search token. And also returns an additional proof with the document, to verify the search result.

Public Verification Key

Public verification key is a security measure designed to make sure that your document outsourced in the cloud doesn't get hacked. By verifying the public key, the Data Owner and the Data User add another layer of protection to the documents or files in the cloud by confirming each other's identities.

4.5 MODULE FUNCTIONALITY

INPUT DESIGN

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps necessary to put transaction data into a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining privacy. Input Design considered the following things:

What data should be given as input?

How should the data be arranged or coded?

The dialog to guide the operating personnel in providing input.

Methods for preparing input validations and steps to follow when errors occur.

OBJECTIVES

1. Input Design is the process of converting a user-oriented description of the input into a computer-based system. This design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerized system.

2. It is achieved by creating user-friendly screens for the data entry to handle large volumes of data. The goal of designing input is to make data entry easier and to be free from errors. The data entry screen is designed in such a way that all the data can be performed. It also provides record viewing facilities.

3. When the data is entered it will check for its validity. Data can be entered with the help of screens. Appropriate messages are provided as when needed so that the user will not be in maize instant. Thus the objective of input design is to create an input layout that is easy to follow

OUTPUT DESIGN

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other systems through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the

most important and direct source of information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making.

1. Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is designed so that people will find the system can be used easily and effectively. When analysis design computer output, they should Identify the specific output that is needed to meet the requirements.

2. Select methods for presenting information.

3. Create documents, reports, or other formats that contain information produced by the system.

The output form of an information system should accomplish one or more of the following objectives.

- ❖ Convey information about past activities, current status or projections of the Future.
- ❖ Signal important events, opportunities, problems, or warnings.
- ❖ Trigger an action.
- ❖ Confirm an action.

5.PROJECT SYSTEM DESIGN

The System Design Document describes the system requirements, operating environment, system and subsystem architecture, files and database design, input formats, output layouts, human-machine interfaces, detailed design, processing logic, and external interfaces.

It is divided into two types like GUI Designing, UML Designing with avails in development of project in facile way with different actor and its utilizer case by utilizer case diagram, flow of the project utilizing sequence, Class diagram gives information about different class in the project with methods that have to be utilized in the project if comes to our project our UML diagram is utilizable in this way.

5.1 E-R DIAGRAMS

ER Diagram stands for Entity Relationship Diagram, also known as ERD is a diagram that displays the relationship of entity sets stored in a database. In other words, ER diagrams help to explain the logical structure of databases. ER diagrams are created based on three basic concepts: entities, attributes and relationships.

ER Diagrams contain different symbols that use rectangles to represent entities, ovals to define attributes and diamond shapes to represent relationships.

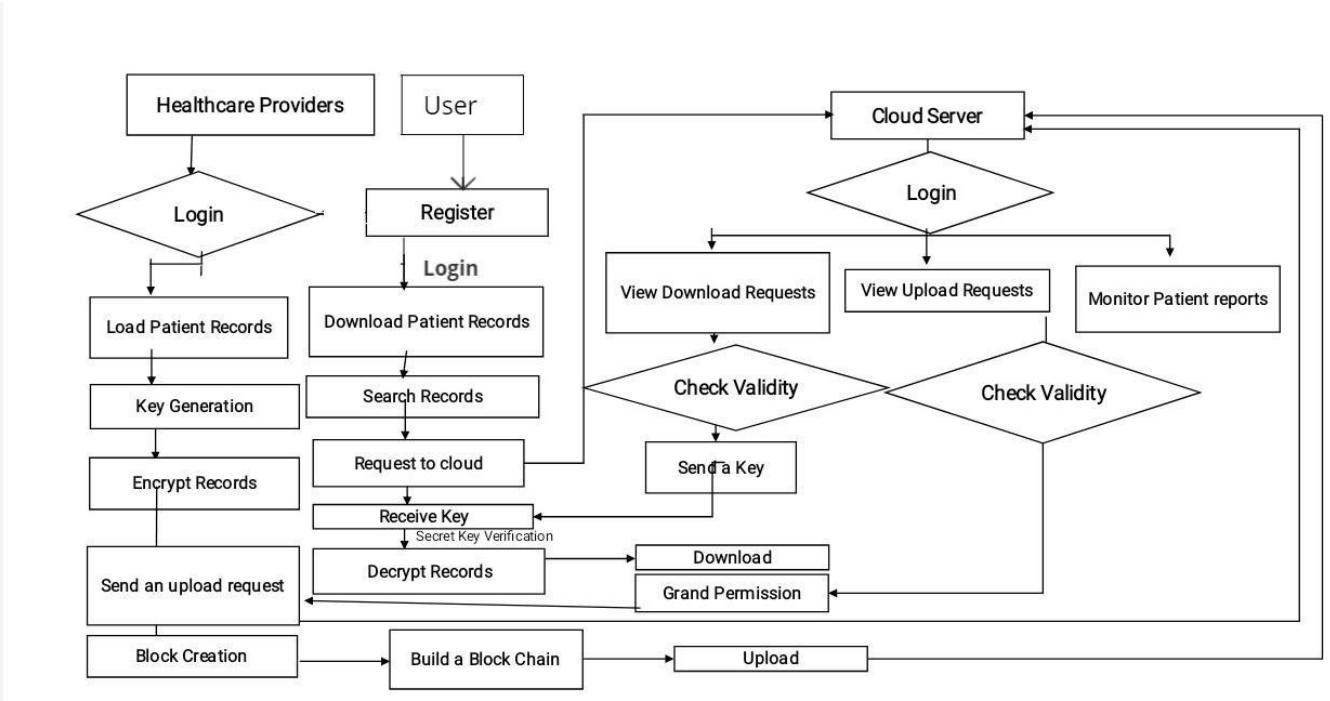


Figure 1: E-R Diagram

5.2 UML DIAGRAMS

UML stands for Unified Modeling Language. UML is a standardized general-purpose modeling language in the field of object-oriented software engineering. The standard is managed, and was created by, the Object Management Group.

The goal is for UML to become a common language for creating models of object-oriented computer software. In its current form UML comprises two major components: a Meta-model and a notation. In the future, some form of method or process may also be added to; or associated with, UML.

The Unified Modeling Language is a standard language for specifying, Visualization, Constructing and documenting the artifacts of software systems, as well as for business modeling and other non-software systems.

The UML represents a collection of best engineering practices that have proven successful in the modeling of large and complex systems.

The UML is a very important part of developing objects-oriented software and the software development process. The UML uses mostly graphical notations to express the design of software projects.

Use Case Diagram

A use case diagram is a way to summarize details of a system and the users within that system. It is generally shown as a graphic depiction of interactions among different elements in a system. Use case diagrams will specify the events in a system and how those events flow, however, the use case diagram does not describe how those events are implemented.

A use case is a methodology used in system analysis to identify, clarify, and organize system requirements. In this context, the term "system" refers to something being developed or operated, such as a mail-order product sales and service Web site. Use case diagrams employed in UML (Unified Modeling Language), a standard notation for modeling real-world objects and systems. There are many benefits to having a use case diagram over similar diagrams such as flowcharts.

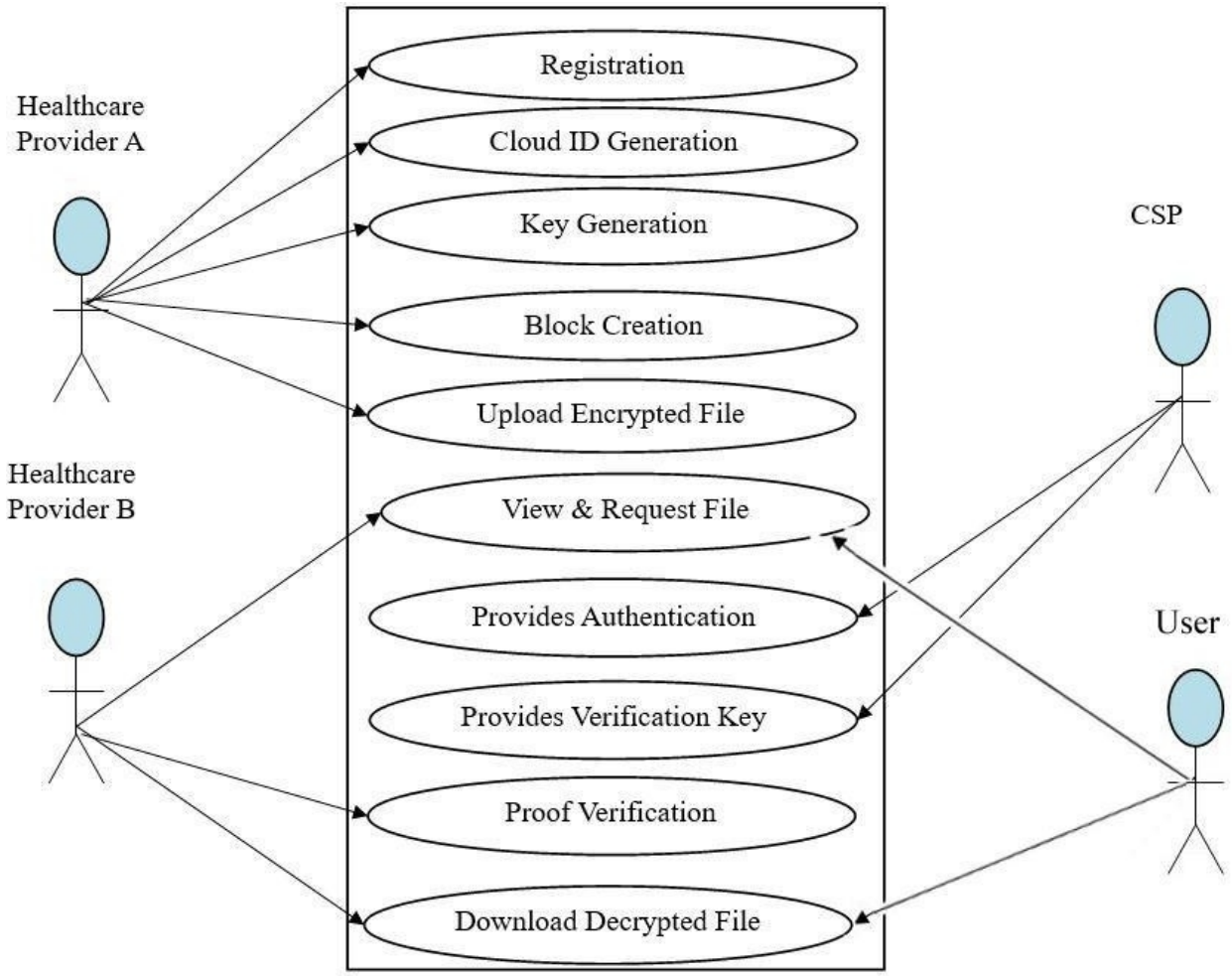


Figure 2: Use Case Diagram

Class Diagram

Class diagrams are the main building block in object-oriented modeling. They are used to show the different objects in a system, their attributes, their operations and the relationships among them.The purpose of class diagrams is to model the static view of an application. Class diagrams are the only diagrams which can be directly mapped with object-oriented languages and thus widely used at the time of construction. Analysis and design of the static view of an application.

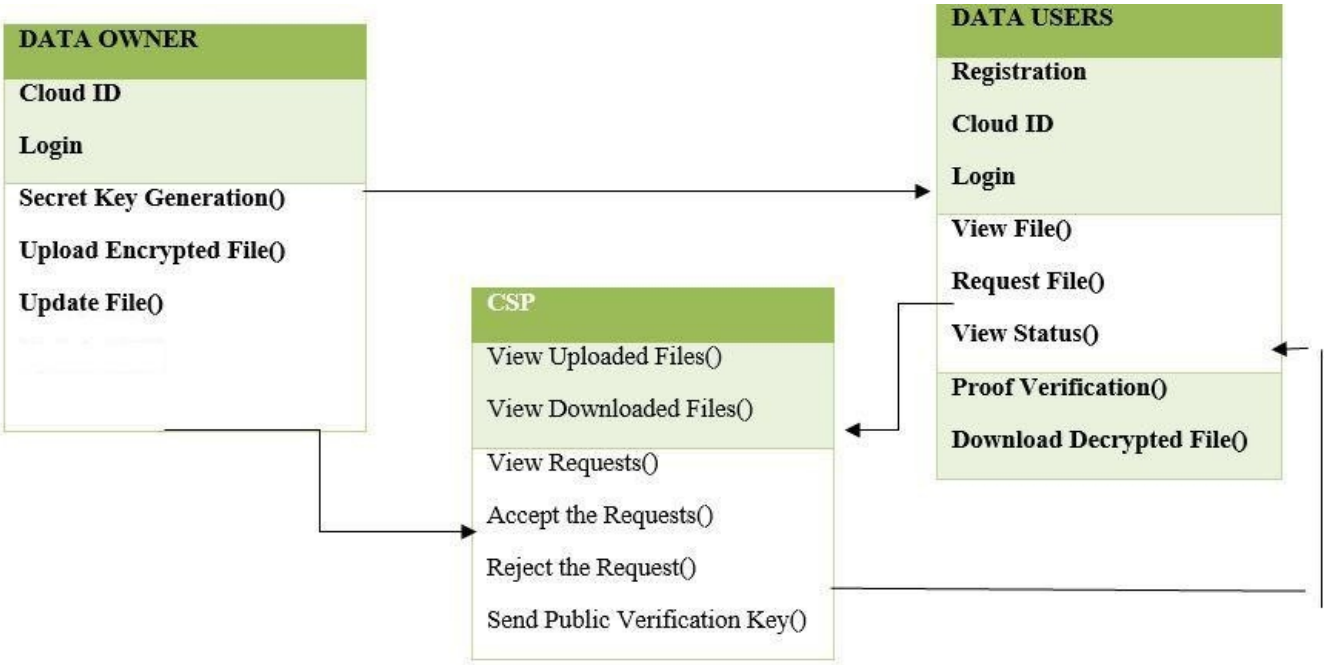


Figure 3: Class Diagram

Object Diagram

Object is an instance of a class at a particular moment in runtime that can have its own state and data values. Likewise a static UML object diagram is an instance of a class diagram; it shows a snapshot of the detailed state of a system at a point in time, thus an object diagram encompasses objects and their relationships which may be considered a special case of a class diagram or a communication diagram.

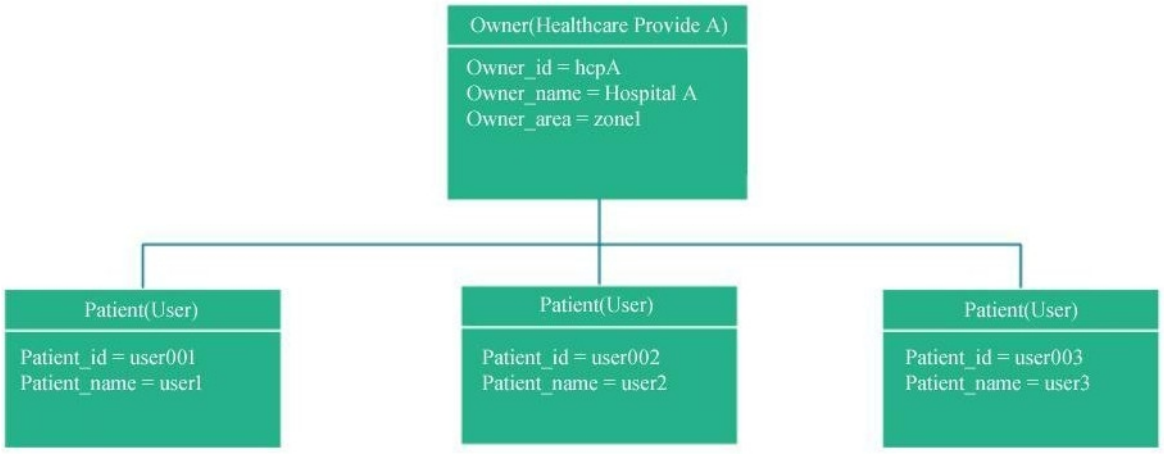


Figure 4: Object Diagram

Sequence Diagram

Sequence diagrams, commonly used by developers, model the interactions between objects in a single use case. They illustrate how the different parts of a system interact with each other to carry out a function, and the order in which the interactions occur when a particular use case is executed.

In simpler words, a sequence diagram shows different parts of a system work in a ‘sequence’ to get something done.

This sequence diagram tutorial is to help you understand sequence diagrams better; to explain everything you need to know, from how to draw a sequence diagram to the common mistakes you should avoid when drawing one.

There are 3 types of Interaction diagrams; Sequence diagrams, communication diagrams, and timing diagrams. These diagrams are used to illustrate interactions between parts within a system. Among the three, sequence diagrams are preferred by both developers and readers alike for their simplicity.

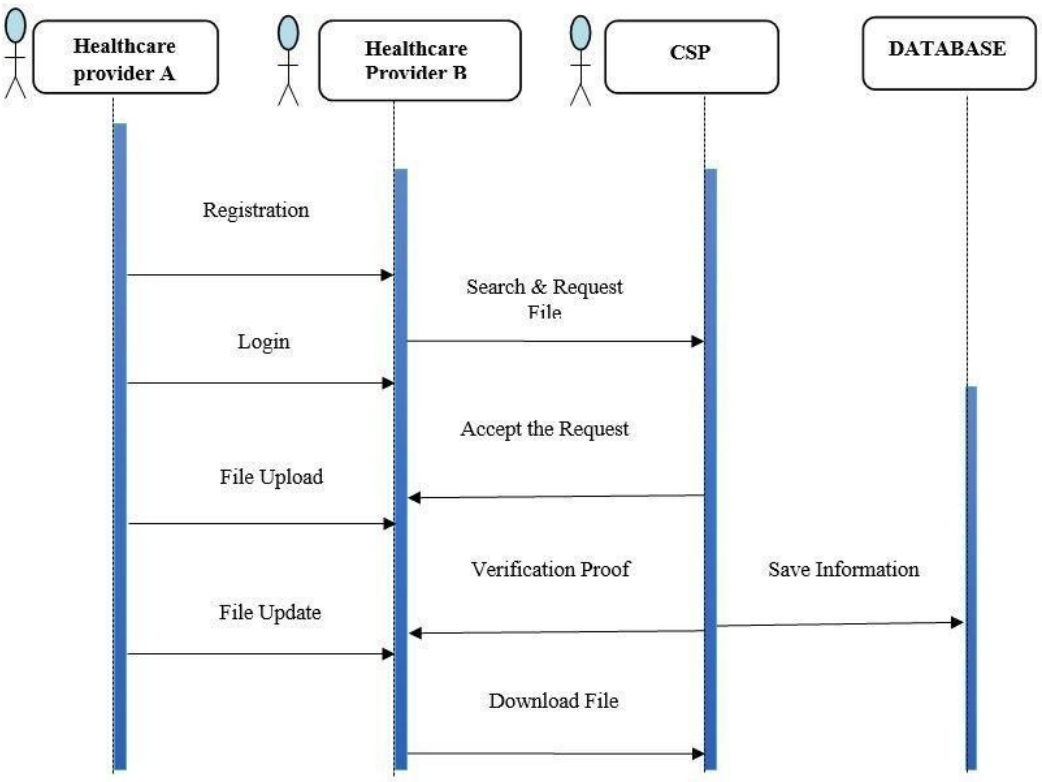


Figure 5: Sequence Diagram

6.PROJECT CODING

6.1 CODE TEMPLATE

Login Page(login.jsp)

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">
<html xmlns="http://www.w3.org/1999/xhtml">
  <head>
    <meta http-equiv="content-type" content="text/html; charset=utf-8" />
    <title>block chain for secure ehrs sharing of mobile</title>
    <meta name="keywords" content="" />
    <meta name="description" content="" />
    <link href="styles.css" rel="stylesheet" type="text/css" media="screen" />
        <link rel="stylesheet" href="nivo-slider.css" type="text/css" media="screen" />
  </head>
  <body>
    <div id="back">
      <div id="footer_bg">
        <div id="content">
          <!-- header begins -->
          <div id="header">
            <div id="logo">
                <h1><a href="#">block chain for secure ehrs sharing of mobile</a></h1>
            </div>
            <div id="menu">
              <ul>
                <li><a href="#" title="" class="action">Home</a></li>
                <li><a href="register.jsp" title="">Register</a></li>
```

```
                    <li><a href="login.jsp" title="">Login</a></li>
                    <li><a href="about.jsp" title="">About</a></li>
                    <li><a href="contact.jsp" title="">Contact</a></li>
                  </ul>
                </div>
              </div>

                          <div class="header_slider"> <div id="slider-wrapper">
                            <div id="slider" class="nivoSlider">
                                <img src="images/header1.jpg"alt="" title="" />
                                 <img src="images/header2.jpg" alt="" title="" />
                                <img src="images/header3.jpg" alt="" title="" />
                                 </div>
                                 </div>

<script type="text/javascript" src="lib/jquery-1.4.3.min.js"></script>
    <script type="text/javascript" src="lib/jquery.nivo.slider.pack.js"></script>
    <script type="text/javascript">
    $(window).load(function() {
        $('#slider').nivoSlider();
    });
    </script>  </div>

                <!-- header ends -->
                <!-- content begins -->
                <div class="main">


                <div class="main2">
```

```
                <div id="blog_col2">

                    <div class="col2_1">
                        <h1>Login Page</h1>
    <%
    String va=request.getParameter("va");
    if(va!=null)
    {
    if(va.equals("1"))
    {%>
    <h2 style="color:#ff99cc"> REGISTRATION SUCCESSFUL </h2>
    <%}else  if(va.equals("2"))
    {%>

    <h2 style="color:#ff99cc">UserName and Password Invalid </h2>
    <%}
    }
%>
                        <pre style="color:#ffcccc">

 <form action="logindb.jsp" method="post">
        user id :::: <input type="text" name="uid" >

        password ::: <input type="password" name="pwd">

           Type :: <select name="type"> <option value=""> Select </option>
                                <option value="owner">Owner</option>
                                <option value="user">User </option>
                                <option value="csp">Csp </option>
```

```
            </select>
            <input type="submit" value="Login"> <input type="reset" value="reset">
        </form>
</pre>
                    </div>
                    <div class="col2_2">
                       <h1>MAIN OBJECTIVES</h1>
                       <ul>
                          <li><a href="#">set up and file operation</a></li>
                                    <li><a href="#">dynamic operation on outsourced data</a></li>
                           <li><a href="#">data accessing and cheating detection</a></li>
                          <li><a href="#">broadcasting encryption algorithm</a></li>
                          <li><a href="#">data owner</a></li>
                                <li><a href="#">cloud service provider</a></li>
                       </ul>
                    </div>
                    <div style="clear: both"></div>
                 </div>
              </div>
              <!--content ends -->
              <!--footer begins -->
              <div id="footer">
                                <p>Copyright 2015. <a href="#">Privacy Policy</a> | <a href="#">Terms of Use</a> | <a href="http://validator.w3.org/check/referer" title="This page validates as XHTML 1.0 Transitional"><abbr title="eXtensible HyperText Markup Language">Trylogic</abbr></a></p>
                    </div>
            </div>
```

```
        </div>
      </div>
      <!-- footer ends-->
   </body>
</html>
```

Uploading Files in Cloud(updb.jsp)

```
<%@page import="java.sql.PreparedStatement"%>
<%@page import="java.sql.ResultSet"%>
<%@page import="java.sql.Statement"%>
<%@page import="java.io.PrintStream"%>
<%@page import="java.io.BufferedReader"%>
<%@page import="java.io.LineNumberReader"%>
<%@page import="java.io.FileReader"%>
<%@page import="java.util.Scanner"%>
<%@page import="java.io.FileInputStream"%>
<%@page import="java.io.File"%>
 <%@page  import="java.io.IOException" %>
<%@page import="java.util.List"%> z
<%@include file="db.jsp" %>

<%

try{
 String owner=session.getAttribute("owner").toString();
session.setAttribute("owner", owner);

String id=request.getParameter("id");  System.out.println("file id    ::"+id);
```

```
String key=request.getParameter("key");  System.out.println("file key  ::"+key);

String name=request.getParameter("name");  System.out.println("file name  ::"+name);

String fpath=request.getParameter("fpath"); System.out.println("file fpath ::"+fpath);

String com=request.getParameter("com"); System.out.println("file com ::"+com);

String com1=request.getParameter("com1"); System.out.println("file com1 ::"+com1);

String com2=request.getParameter("com2"); System.out.println("file com2 ::"+com2);

String dis=request.getParameter("dis");  System.out.println("file fdata ::"+dis);



String[] end=fpath.split("\\.");

String fir=end[0];System.out.println("file fir ::"+fir);

String sec=end[1]; System.out.println("file sec::"+sec);

String[] send=fir.split("\\/"); String fn=send[1]; System.out.println("file sec::"+fn);


 String path=request.getRealPath("fpath");
 String comm=fn+"0";  String fnn=comm+"."+sec;
```

```
 File f= new File(path+"\\"+fnn);    String pa="fpath"+"//"+fnn; System.out.println("file
::"+pa);
 PrintStream psm=new PrintStream(f);
 byte b[]=com.getBytes();
 psm.write(b);
String comm1=fn+"1";  String fnn1=comm1+"."+sec;
    File    f1=    new    File(path+"\\"+fnn1);          String    pa1="fpath"+"//"+fnn1;
System.out.println("file 1::"+pa1);
 PrintStream psm1=new PrintStream(f1);
 byte b1[]=com1.getBytes();
 psm1.write(b1);
 String comm2=fn+"2"; String fnn2=comm2+"."+sec;
    File    f2=    new    File(path+"\\"+fnn2);          String    pa2="fpath"+"//"+fnn2;
System.out.println("file 2::"+pa2);
 PrintStream psm2=new PrintStream(f2);
 byte b2[]=com2.getBytes();
 psm2.write(b2);
int ci=0;
Statement st=con.createStatement();
ResultSet rs,rs1,rs2,rs3;
PreparedStatement  ps,ps1,ps2,ps3;
rs=st.executeQuery("select count(FILEID) from UFILES");    while(rs.next())
   { ci=rs.getInt(1)+1; }
String pf1="Fileid_"+ci;   System.out.println("passportid ::"+pf1);
ps=con.prepareStatement("insert into UFILES values(?,?,?,?,?,?,?,?)");
ps.setString(1, pf1);
ps.setString(2, name);
ps.setString(3,  fpath);
ps.setString(4, dis);
```

```
ps.setString(5, fpath);
ps.setString(6, key);
ps.setString(7, owner);
ps.setString(8, "uploading");
int io1= ps.executeUpdate();
rs=st.executeQuery("select count(FILEID) from FILES");    while(rs.next())
    { ci=rs.getInt(1)+1; }
String pd="Fileid_"+ci;    System.out.println("passportid----  111 ::"+pd);
ps=con.prepareStatement("insert into FILES values(?,?,?,?,?,?,?)");
ps.setString(1, pd);
ps.setString(2, name);
ps.setString(3, pa);
ps.setString(4, dis);
ps.setString(5, fpath);
ps.setString(6, key);
ps.setString(7,  owner);
int i= ps.executeUpdate();
if(i>0)
{
int count=0;   st=con.createStatement();
rs=st.executeQuery("select count(FILEID) from FILES");    while(rs.next())
    { count=rs.getInt(1)+1; }
String pid="Fileid_"+count;    System.out.println("passportid...22 ::"+pid);
ps1=con.prepareStatement("insert into FILES values(?,?,?,?,?,?,?)");
ps1.setString(1, pid);
ps1.setString(2, name);
ps1.setString(3, pa1);
ps1.setString(4, dis);
ps1.setString(5, fpath);
```

```
ps1.setString(6, key);
ps1.setString(7, owner);
int i1= ps1.executeUpdate();
if(i1>0)
    {
    int coun=0;   st=con.createStatement();
rs=st.executeQuery("select count(FILEID) from FILES");    while(rs.next())
    { coun=rs.getInt(1)+1; }
String pid1="Fileid_"+coun;    System.out.println("passportid---- 33::"+pid1);
ps2=con.prepareStatement("insert into FILES values(?,?,?,?,?,?,?)");
ps2.setString(1, pid1);
ps2.setString(2, name);
ps2.setString(3, pa2);
ps2.setString(4, dis);
ps2.setString(5, fpath);
ps2.setString(6, key);
ps2.setString(7, owner);
int i2= ps2.executeUpdate();
if(i2>0)
    {
 response.sendRedirect("Oupload.jsp?s=1");
 }
}
 }
 else
{
    response.sendRedirect("Oupload.jsp?s=2");
  }
}
```

```
catch(Exception e)
        {
    e.printStackTrace();
out.println(e);
}
%>
```

6.2 OUTLINE FOR VARIOUS FILES

Java Files

EncDec.java -Encryption and decryption of uploaded data files using AES algorithm.

encryptionclass.java - String encryption

WebContent Files

home.jsp - Home page of the project.

register.jsp - Register page where the owner,user,csp can login.

login.jsp - Login page for owner,user,csp.

search.jsp - Search page for the user .

uploadb.jsp - Upload page where Owners can upload their data files.

verify.jsp - Verify page where owners can verify the request sent by the users.

profile.jsp - Profile page to see and to make changes if required.

viewfile.jsp - View files in the Owner page .

viewreq.jsp - View required page where the owners can verify the request sent by the users.

Cviewfiles.jsp - View files page in the csp.

6.3 CLASS WITH FUNCTIONALITY

Encryption class

```
package Encryption;
import java.util.StringTokenizer;
public class encryptionclass
{
	//enc="";
	public static String encrypt(String s)
	{
	  String enc="";
	  String sa=s;
	    System.out.println("value is "+sa);
	for(int i=0;i<sa.length();i++)
	{
	//String enc="";
	char ch=sa.charAt(i);
	int j=ch;
	int k=j*16;
	enc+=k+"#";
	}
	return enc;
	}
	public static String decrypt(String enc)
	{
	String data,actdata="";
	data=enc;
	StringTokenizer str=new StringTokenizer(data,"#");
	while(str.hasMoreTokens())
	{
```

```
            String s=str.nextToken();
            int i=Integer.parseInt(s);
            int k=i/16;
            actdata+=(char)k;
        }
        return actdata;
        }
}
```

Encryption and Decryption using AES Algorithm

```
package Encryption;
import java.security.*;
import java.security.spec.InvalidKeySpecException;
import javax.crypto.*;
import javax.crypto.spec.SecretKeySpec;
import sun.misc.*;
public class EncDec {
    private static final String ALGO = "AES";
   private static final byte[] keyValue =
       new byte[] { 'T', 'h', 'e', 'B', 'e', 's', 't','S', 'e', 'c', 'r','e', 't', 'K', 'e', 'y' };
public static String encrypt(String Data) throws Exception
       {Key key = generateKey();
       System.out.println("ecncry.....key...::"+key);
       Cipher c = Cipher.getInstance(ALGO);
       c.init(Cipher.ENCRYPT_MODE, key); byte[]
       encVal = c.doFinal(Data.getBytes());
       String encryptedValue = new BASE64Encoder().encode(encVal);
       return encryptedValue;
   }
```

```
  public static String decrypt(String encryptedData) throws Exception
      {Key key = generateKey();
      System.out.println("decry....key...::"+key);
      Cipher c = Cipher.getInstance(ALGO);
      c.init(Cipher.DECRYPT_MODE, key);
      byte[] decordedValue = new BASE64Decoder().decodeBuffer(encryptedData);
      byte[] decValue = c.doFinal(decordedValue);
      String decryptedValue = new String(decValue);
      return decryptedValue;
   }
   public static Key generateKey() throws Exception
      { Key key = new SecretKeySpec(keyValue, ALGO);
      return key;
}
}
```

6.4 METHODS INPUTS AND OUTPUTS PARAMETERS

public static String encrypt(String Data)

public static String decrypt(String encryptedData)

public static Key generateKey()

7.PROJECT TESTING

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of tests. Each test type addresses a specific testing requirement.

7.1 VARIOUS TEST CASES

Unit testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .It is done after the completion of an individual unit before integration. This is a structural testing that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results. Unit testing is usually conducted as part of a combined code and unit test phase of the software lifecycle, although it is not uncommon for coding and unit testing to be conducted as two distinct phases.

Test strategy and approach

Field testing will be performed manually and functional tests will be written in detail.

Test objectives

- All field entries must work properly.

- Pages must be activated from the identified link.
- The entry screen, messages and responses must not be delayed.

Features to be tested

- Verify that the entries are of the correct format
- No duplicate entries should be allowed
- All links should take the user to the correct page.

Integration testing

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.Software integration testing is the incremental integration testing of two or more integrated software components on a single platform to produce failures caused by interface defects.

The task of the integration test is to check that components or software applications, e.g. components in a software system or – one step up – software applications at the company level – interact without error.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

Acceptance Testing

User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

Functional test

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items:

Valid Input : identified classes of valid input must be accepted.

Invalid Input : identified classes of invalid input must be rejected.

Functions : identified functions must be exercised.

Output : identified classes of application outputs must be exercised.

Systems/Procedures: interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identifying Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

System Test

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

7.2 BLACK BOX TESTING

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as

specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated as a black box .you cannot “see” into it. The test provides inputs and responds to outputs without considering how the software works.

7.3 WHITE BOX TESTING

White Box Testing is a testing in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It has a purpose. It is used to test areas that cannot be reached from a black box level.

8.OUTPUT SCREENS

8.1 USER INTERFACES

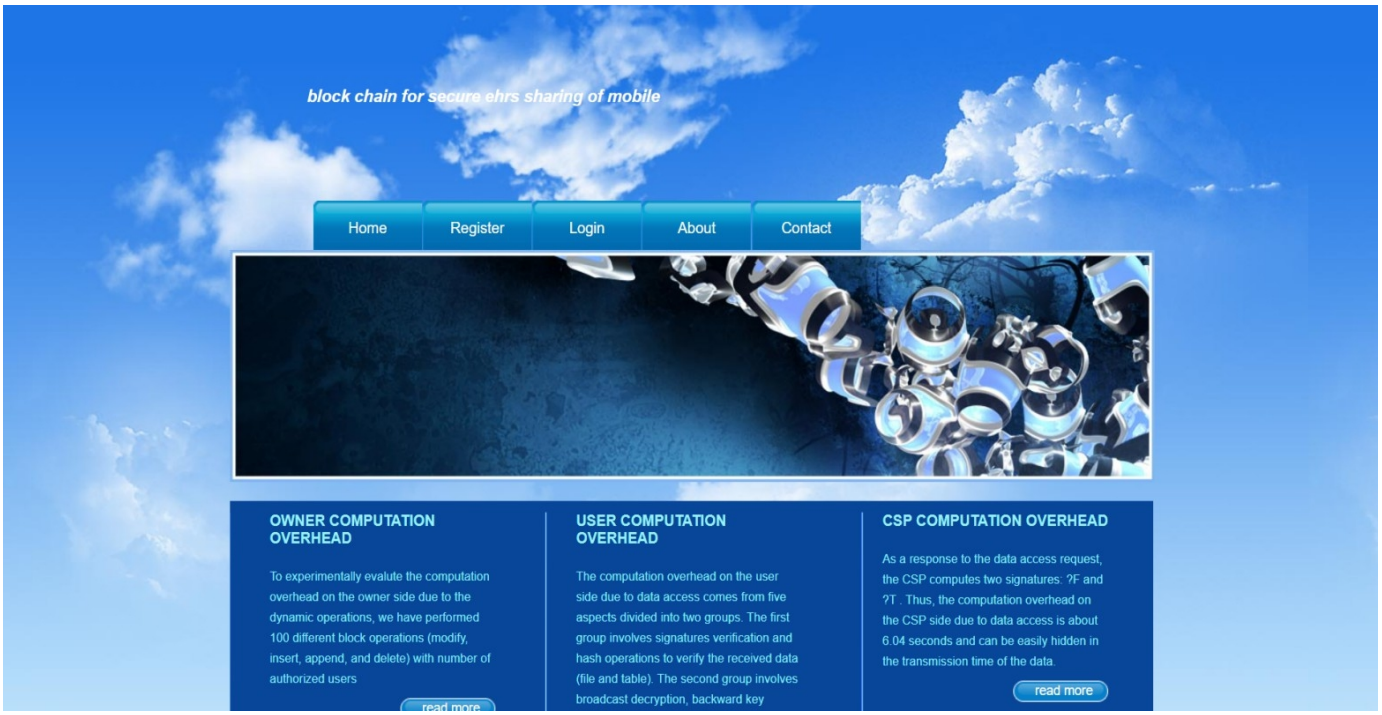


Figure 1:Home Page

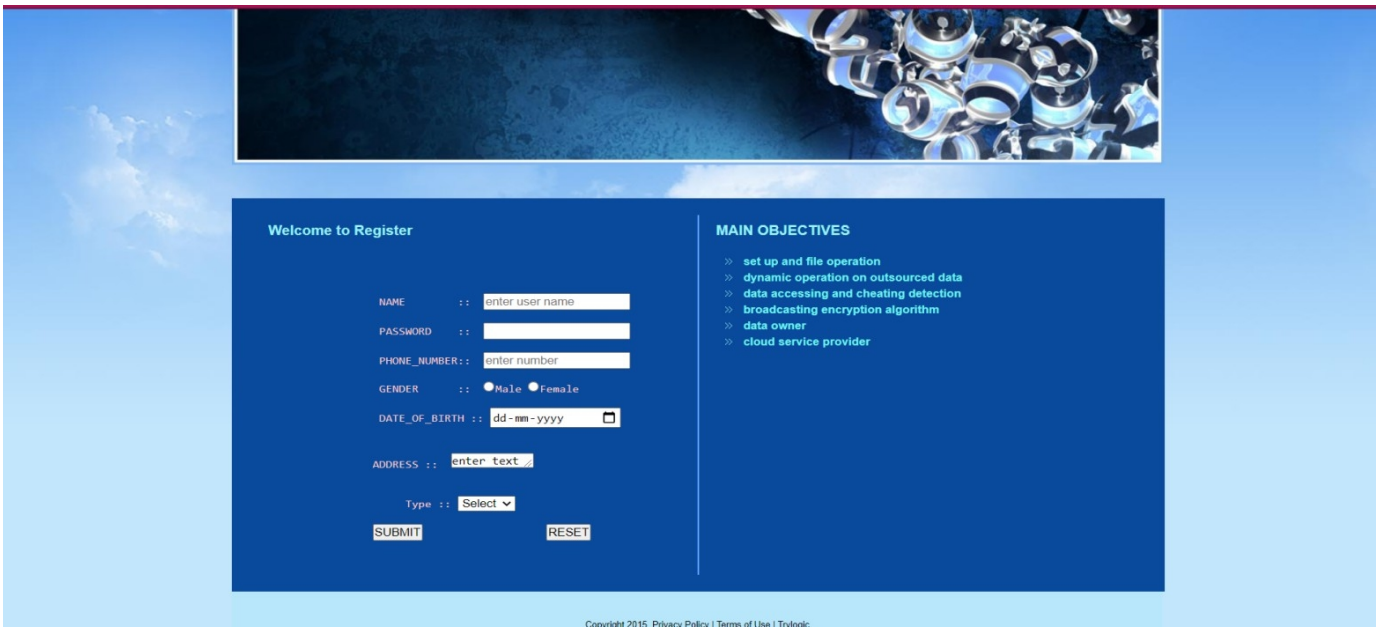


Figure 2:Register Page for User and Owner

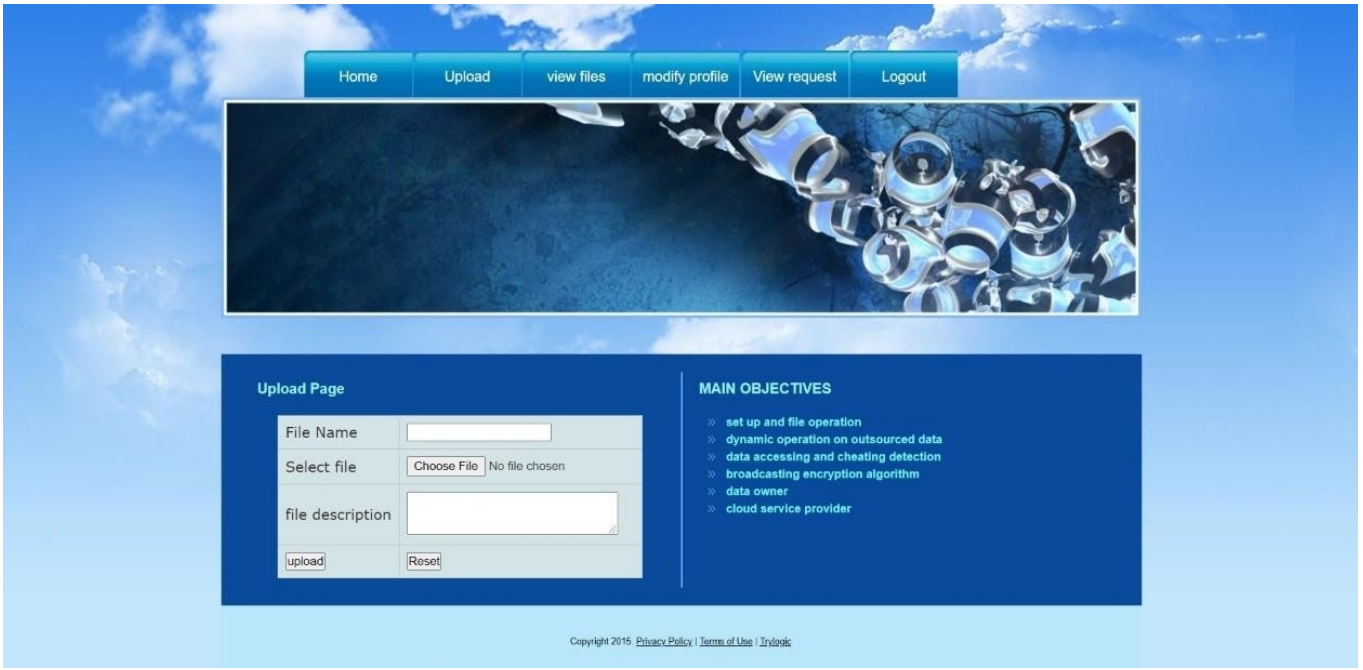


Figure 3:Owner Upload Page

8.2 OUTPUT SCREENS

Fileid	Name	Key	Descfile	Owner name	Uploading Status
Fileid_7	user002	1H1H	this user002 file	prem	uploaded
Fileid_14	user008	8Y41	user data entry	prem	uploaded
Fileid_18	user011	1DQQ	user 11 file	prem	uploaded
Fileid_19	user012	8884	user 12 health data 09 06	prem	uploaded
Fileid_6	user001	6YHF	this is my user001 file	prem	uploading
Fileid_16	user009	6TY1	data	prem	uploaded
Fileid_8	user003	EQEQ	des info	prem	uploaded
Fileid_9	user004	1F45	this is user004 desc file	prem	uploaded
Fileid_10	abc	6D1F	abc screenshort 20/05/12	prem	uploaded
Fileid_12	user006	AEH6	user 6 file	prem	uploading
Fileid_15	user008	644E	user health datafrom 08-06-2021	prem	uploaded

Figure 4:Owner uploaded file

Figure 5:CSP View Files

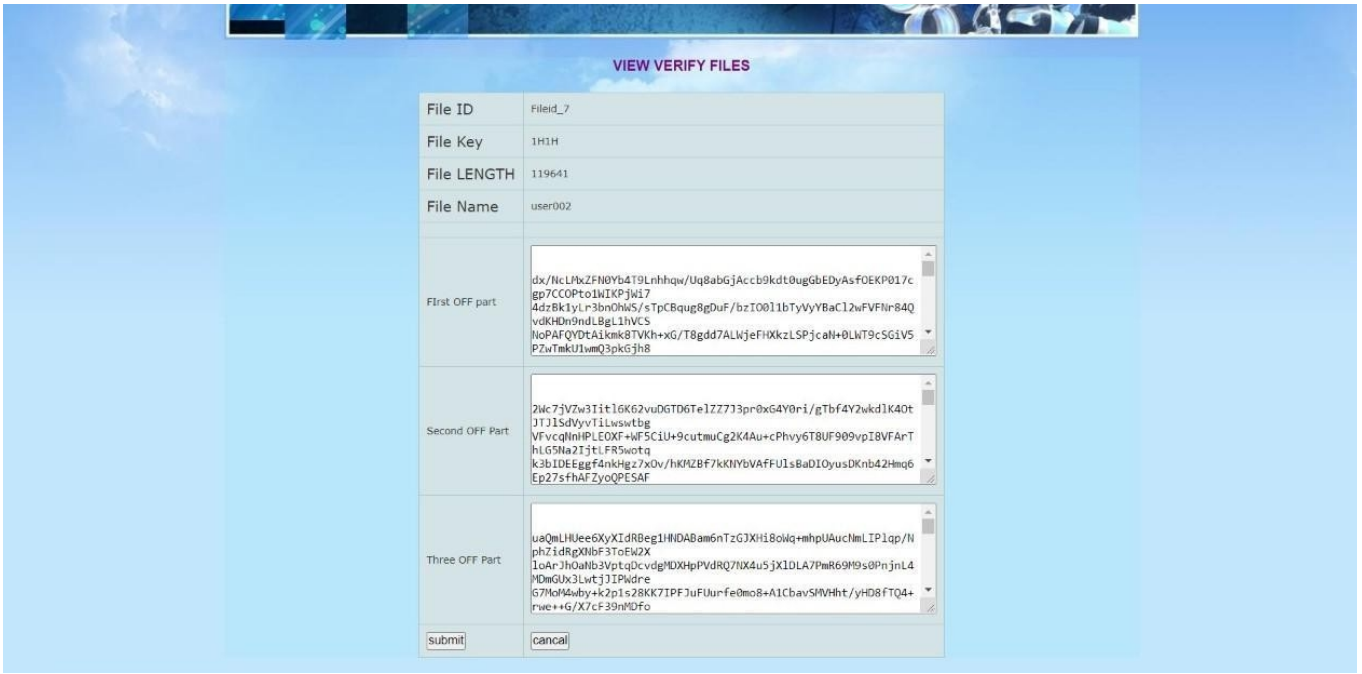


Figure 6:CSP verify upload

Figure 7:CSP verify successful

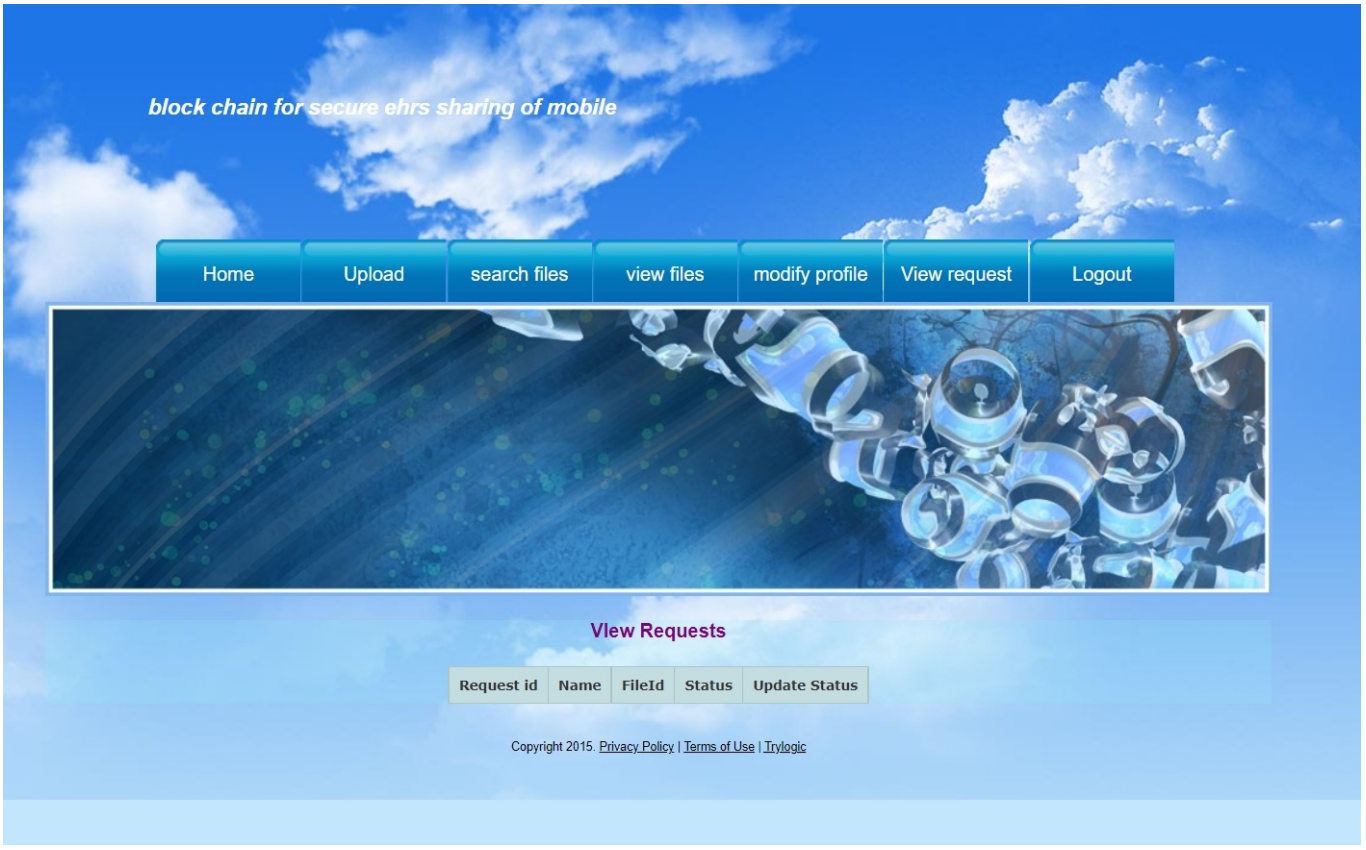


Figure 8:Owner View Request Page

9.EXPERIMENTAL RESULTS

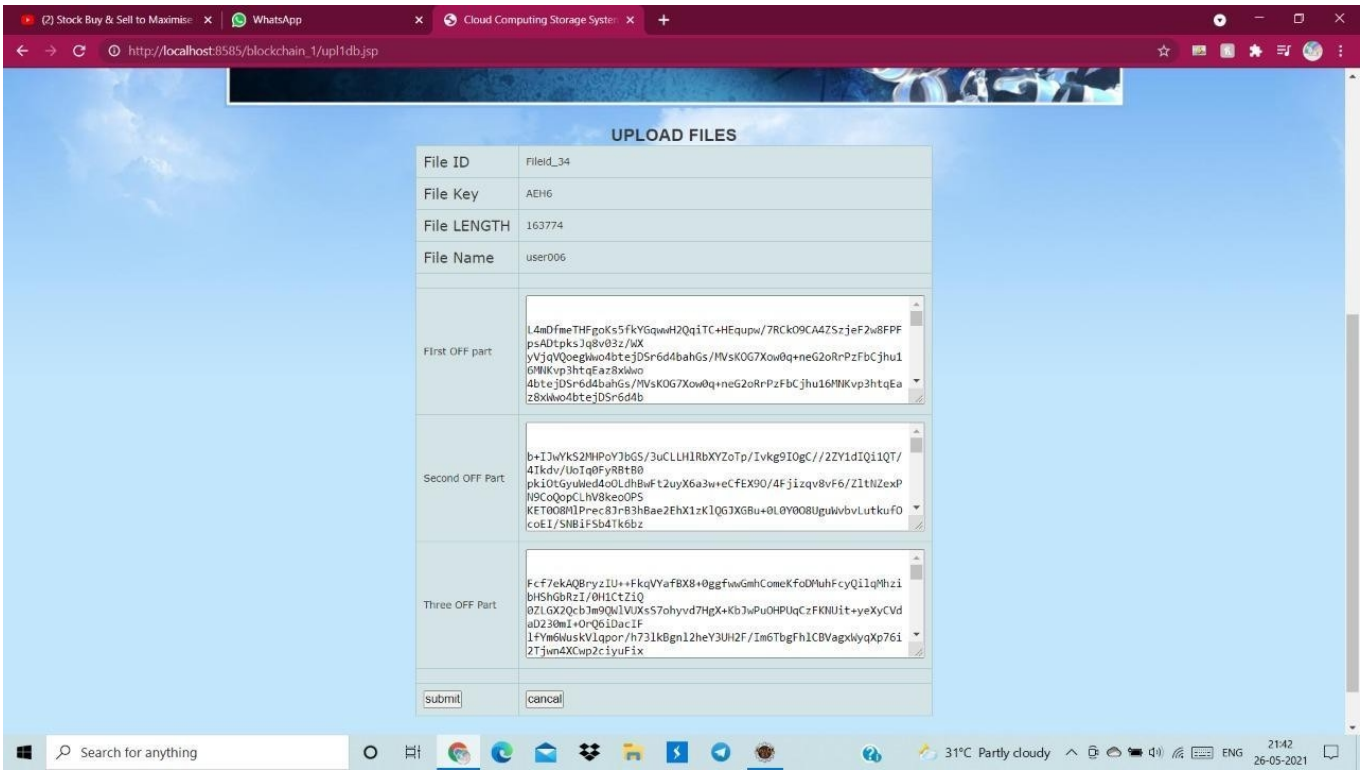


Figure 9:Owner Upload with Encrypted Key

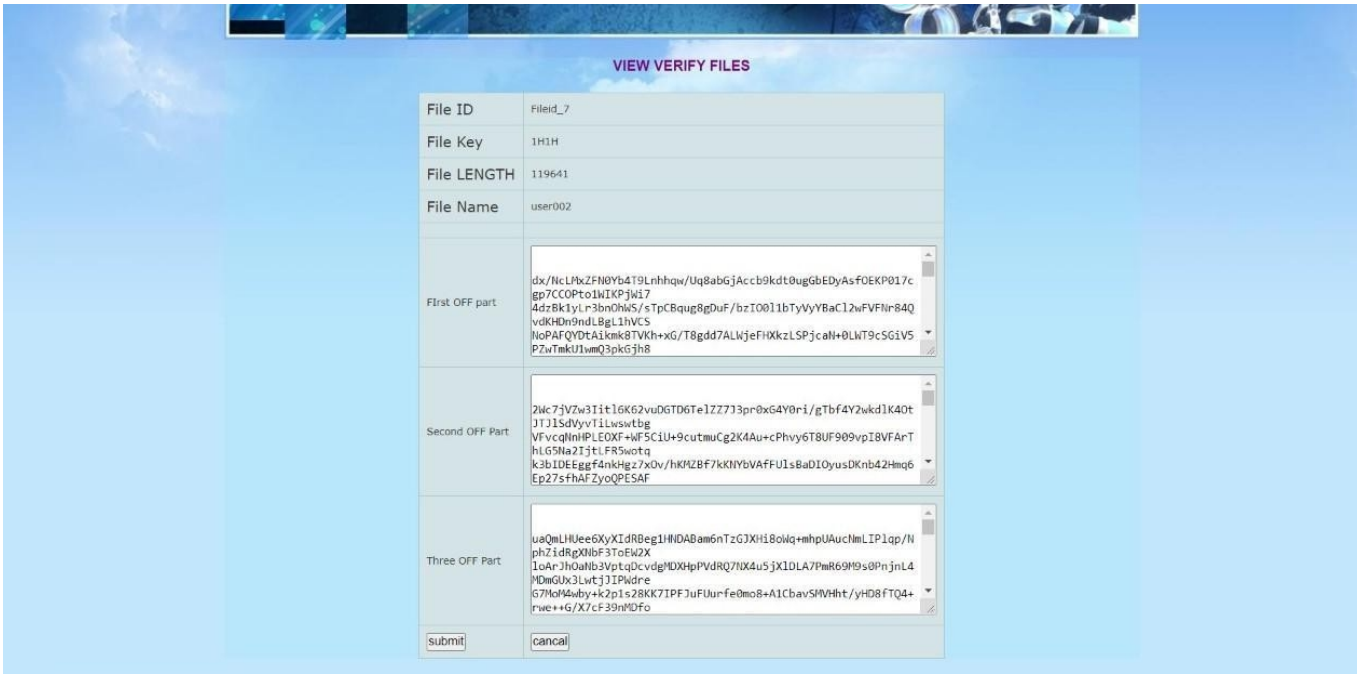


Figure 10:CSP verifying Uploaded File Page

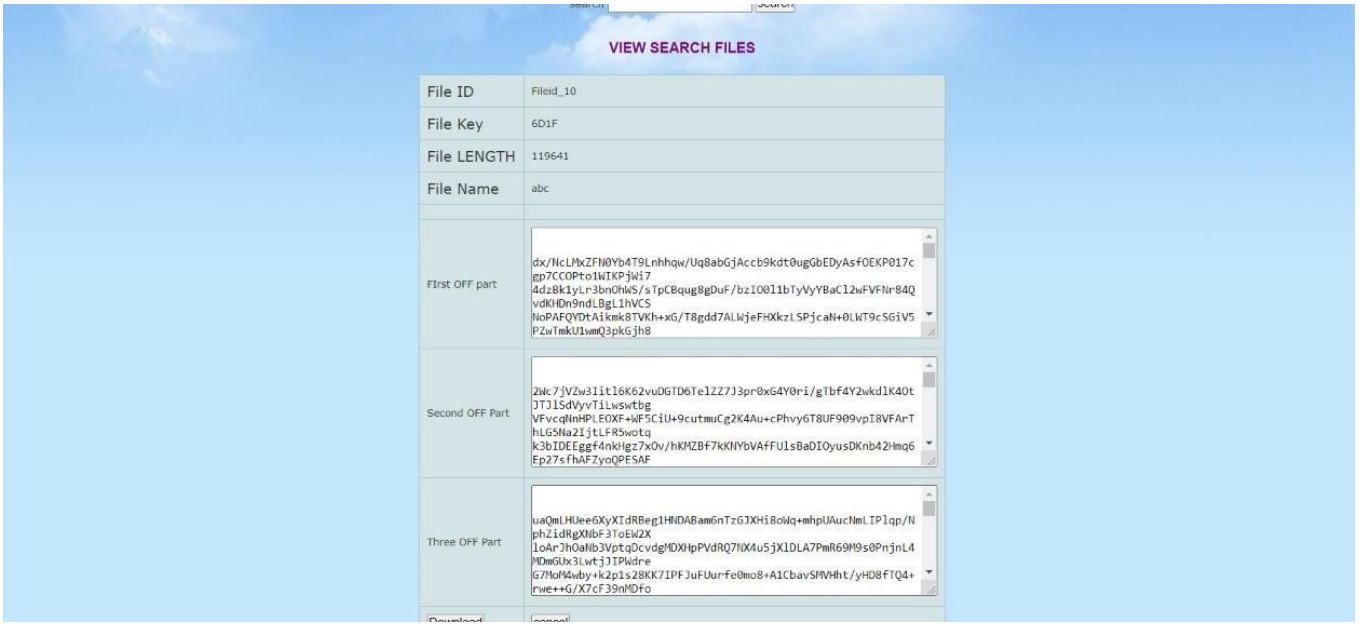
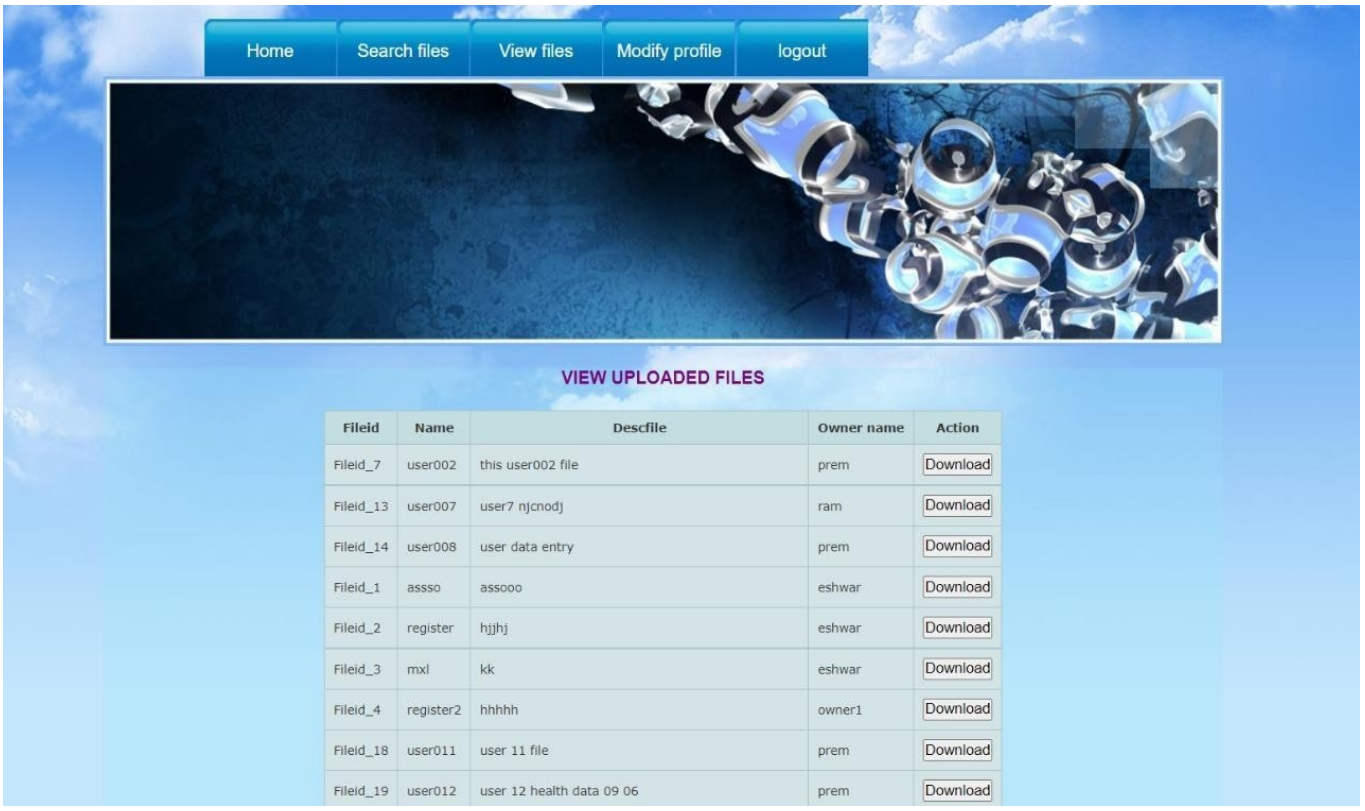


Figure 11:User Download Page



Fileid	Name	Descfile	Owner name	Action
Fileid_7	user002	this user002 file	prem	Download
Fileid_13	user007	user7 njcnodj	ram	Download
Fileid_14	user008	user data entry	prem	Download
Fileid_1	assso	assooo	eshwar	Download
Fileid_2	register	hjjhj	eshwar	Download
Fileid_3	mxl	kk	eshwar	Download
Fileid_4	register2	hhhhh	owner1	Download
Fileid_18	user011	user 11 file	prem	Download
Fileid_19	user012	user 12 health data 09 06	prem	Download

Figure 12:User View Files Page

10.CONCLUSION AND FUTURE ENHANCEMENT

This paper proposes a novel EHRs sharing plan empowered by portable distributed computing and blockchain. We recognize basic difficulties of current EHRs sharing frameworks and propose productive answers for addressing these issues through a genuine model execution. In this work, our attention is on planning a reliable access control system dependent on a solitary shrewd agreement to oversee client access for guaranteeing productive and secure EHRs sharing. To examine the perfor-mance of the proposed approach, we convey an Ethereum blockchain on the Amazon cloud, where clinical elements can connect with the EHRs sharing framework by means of a created versatile Android application. We additionally incorporate the distributed IPFS stockpiling framework with blockchain to accomplish a decen-tralized information stockpiling and information sharing. The execution results show that our structure can permit clinical clients to share clinical information over versatile cloud conditions in a dependable and speedy way, in contrast with regular plans. Specifically, our entrance control can recognize and forestall successfully unapproved admittance to the e-wellbeing sys-tem, targeting accomplishing an ideal degree of patient protection and organization security. We likewise give security examinations and broad assessments on different specialized parts of the proposed framework, showing

benefits of our proposition over existing arrangements. In light of the benefits of our model, we accept that our blockchain empowered arrangement is a stage towards effective administration of e-wellbeing records on versatile mists, which is promising in numerous medical services applications.

We identify critical challenges of current EHRs sharing systems and propose efficient solutions to address these issues through a real prototype implementation.

•We also integrate the peer-to-peer Interplanetary File System(IPFS) storage system with blockchain to achieve decentralized data storage and data sharing.

•By using Advanced Technology like IPFS and Blockchain we have come up with a new solution to give security to the records of the Patients.

•As we are using Local Server it’s not useful in real world databases so, to overcome this we thought of future enhancement.

•To overcome the drawback we can use AWS,microsoft azure to make best use of the available cloud service

REFERENCES

[1] Kuo TT, Kim HE, and Ohno-Machado L, "Blockchain distributed ledger technologies for biomedical and health care applications," Ame. Medi . Infor. Assoc. J., vol. 6, pp. 1211-1220, 2020.

[2] Nabil Rifi, El Rachkidi, NazimAgoulmine, and Nada ChendebTaher, “Towards Using Blockchain Technology for eHealth Data Access Management”, in Proc. IEEE on Advances in Bio.Engi., Oct. 2020.

[3] Gordon W and Catalini, "Blockchain Technology for Healthcare: Facilitating the Transition to Patient-Driven Interoperability," Comput Struct Biotechnol J., pp. 224-230, 2018.

[4] Marko Holbl et al., "A Systematic Review of the Use of Blockchain in Healthcare,"Symmetry,2018.

[5] F.Y. Leu et al., “A Smartphone-Based Wearable Sensors for Monitoring Real-Time Physiological Data,” Computers and Electrical Engineering, 2020.

[6] M. Memon et al., “Ambient Assisted Living Healthcare Frameworks, Platforms, Standards, and Quality Attributes”, 2021.

[7] Zuobin Ying, Lu Wei, Qi Li, Ximeng Liu and Jie Cui, "A Lightweight Policy Preserving EHRs Sharing Scheme in the Cloud," IEEE Access, vol. 6, pp. 53698-53708, 2018.

[8] Vidhya Ramani, Tanesh Kumar, An Braeken, and Mika Ylianttila, "Secure and Efficient Data Accessibility in Blockchain based Healthcare Systems,"in GLOBECOM, Dec. 2018.

[9] Amazon Web Services (AWS) - Cloud Computing Services. [Online].

[10] Mathis Steichen, Robert Norvill, Beltran Borja Fiz Pontiveros, and Wazen Shbair, "Blockchain-Based, Decentralized Access Control for IPFS," in Proc. IEEE on Blockchain,2018.

[11] M. Min et al., "Learning-Based Privacy-Aware Offloading for Healthcare IoT with Energy Harvesting," IEEE Internet of Things Journal, 2019.

[12] Yuanyu Zhang, Shoji Kasahara, Yulong Shen, Xiaohong Jiang, and Jianxiong Wan, "Smart Contract-Based Access Control for the Internet of Things," IEEE Internet of J.,2018.

[13] M. Min et al., "Learning-Based Privacy-Aware Offloading for Healthcare IoT with Energy Harvesting," IEEE Internet of Things Journal, 2019.

[14] G. Wood, "Ethereum: A secure decentralised generalised transaction ledger," Yellow Paper. [Online].

[15]V. Casola et al., “Healthcare-Related Data in the Cloud: Challenges and Opportunities” IEEE Cloud Computing, 2020.

PUBLICATIONS

INTERNATIONAL CONFERENCE ON INNOVATIONS IN COMPUTER NETWORKS, COMPUTATIONAL INTELLIGENCE AND IOT [ICICCI-2021].(Cloud Based Solution to Ensure the Security of E -Health Records using Blockchain)

ALL FOUR STUDENTS ONE PAPER PROFILE

Kothapu Lakshmi Narayana Chandana is currently pursuing her Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. She completed her intermediate from Sri Chaitanya Junior College and 10^{th} class from Sri Chaitanya High School.She is one of the members of Placement Coordinators in our college. Her responsibilities in that group include Organizing campus placement and motivating students to get placed as a serious goal. She is also one of the members of MG Technology Solutions. Her responsibilities in that group include Developing Websites and taking coding as a serious hobby. Her technical skills include HTML5,CSS3,JavaScript, C, Python and Java. She also has a basic understanding of C++. She took part in the Employability Skill development Program conducted by Zensar. She is also a student of Smart Interviews. Her participations include: WAC Workshop on "AI & ML" which was conducted from 14^{th} to 15^{th} February 2020 conducted by Kyrion Technologies at Indian Institute of Technology, "Know More - Teach More ", the Global Webinar on Cloud & Big Data – Changing the way we work which was conducted Global Education & Careers Forum (GECF) on 12^{th} August 2020, Women online workshop on "Women in Cyber Security and Privacy in 2020" which was conducted from 6^{th} to 10^{th} July 2020, "Know More - Teach More ", the Global Webinar on Cyber Threats and Defense Techniques conducted by GECF on 22^{nd} July 2020, "Quiz on Ethical Hacking" conducted by Itronix Solutions on 26^{st} May 2021. Her areas of interest are HTML5,CSS3,JavaScript,AngularJS,Python, Artificial Intelligence, Machine Learning and Deep Learning. She completed a few certification courses from online platforms like Coursera, Cursa App, and SoloLearn.

PITLA PREM KUMAR is currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering from St. Martin's Engineering College.He completed his Intermediate from Srigayatri Junior College and Secondary Schooling from Krishnaveni High School.He took part in the Employability Skill Development Program Conducted by Zensar.His Participation includes National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020, IIC Online Sessions conducted by Institution's Innovation Council (IIC) of MHRD's Innovation Cell, New Delhi to promote Innovation, IPR, Entrepreneurship, and Start-ups among HEIs from 28 April 2020 to 22 May 2020.His areas of Interest includes Python,JAVA,Artificial Intelligence,Machine Learning,HTML,CSS,C++,C.Attended WAC Workshop on "AI & ML" which was conducted from 14th to 15th February 2020 conducted by Kyrion Technologies at Indian Institute of Technology Hyderabad(IITH) and Technology Awareness Month(TAM) Conducted in our College.A part from his Academics he had done some of the courses which are related to CIVILS and participated in many Group Discussions,Quizzes hosted by UPSC GUIDE, OPEDEMY related to UPSC.He Completed a few Certification courses from Online Platforms like Coursera,SoloLearn and CurseApp.

Vallandas Ramya is currently pursuing her Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin’s Engineering College. She completed her intermediate from Narayana Junior College and 10th class from Brilliant Grammar High School. She took part in the Employability Skill development Program conducted by Zensar. She is also a student of Smart Interviews. Her participation includes the National Level Three Day Online Workshop on “AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020, IIC Online Sessions conducted by Institution's Innovation Council (IIC) of MHRD's Innovation Cell, New Delhi to promote Innovation, IPR, Entrepreneurship, and Start-ups among HEIs from 28th April to 22thnd May 2020. Her areas of interest are Python, Artificial Intelligence, Machine Learning, and Deep Learning. Attended workshop on ML at IIT Hyderabad and TAM college program. She completed the ML course by The Smart Bridge with a skill index of 9 out of 10 and did a project on vehicle resale value prediction. She completed a few certification courses from online platforms like Coursera, CursaApp, and SoloLearn.

Vanam Thrishul is currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. He completed his intermediate from Sri Chaitanya Junior College and 10th class from Kotwal's School. He took part in the Employability Skill development Program conducted by Zensa. He is one of the members of Placement Coordinators in our college. His responsibilities in that group include Organizing campus placement and motivating students to get placed as a serious goal. His technical skills include Python,Html5,Css3,JavaScript,C++ and Java. He also has a basic understanding of C. He is also a student of Smart Interviews. His participations include National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020, "Know More - Teach More ", the Global Webinar on Cloud & Big Data – Changing the way we work which was conducted by Global Education & Careers Forum (GECF) on 12th August 2020, National Level project Expo and Competition "TECHNOVATION-2018" on 28th March 2018, IIC Online Sessions conducted by Institution's Innovation Council (IIC) of MHRD's Innovation Cell, New Delhi to promote Innovation, IPR, Entrepreneurship, and Start-ups among HEIs, 23rd SAMANVAY event at Siva Sivani Institute of Management held on 6th and 7th January 2020. His areas of interest are Python, Web Designing, Web Development, Artificial Intelligence, and Machine Learning. Attended TAM college program, workshop on ML at IIT Hyderabad, and C-graphics at Siva Sivani Institute of Management. He completed a few certification courses from online platforms like Coursera, Cursa App, and SoloLearn.

APPENDICES

Late years we have seen an adjustment in the context of Electronic Health Records (EHRs) on flexible cloud conditions where phones are composed with dispersed figuring to work with clinical data exchanges among patients and clinical benefits providers. This general model engages clinical consideration organizations with low operational cost, high versatility and EHRs availability. Regardless, this new perspective in like manner raises stresses over data insurance and association security for e-prosperity systems. Directions to reliably split EHRs between adaptable customers while guaranteeing high security levels in convenient clouds is a troublesome issue. In this paper, we propose a novel EHRs sharing design that joins blockchain and the decentralized interplanetary record system (IPFS) on a flexible cloud stage. In Particular, we plan a trustworthy access control segment using insightful arrangements to achieve secure EHRs splitting between different patients and clinical providers. We present a model execution using Ethereum blockchain in a certifiable data offering circumstance on a convenient application to Amazon appropriated registering. Observational results show that our suggestion offers a fruitful response for reliable data exchanges on versatile dogs while protecting fragile prosperity information against likely threats. The structure evaluation and security assessment also show execution updates in lightweight access control plan, least association dormancy with high security and data assurance levels, which stood out from existing data sharing models.

This paper proposes a novel EHRs sharing arrangement enabled by convenient circulated registering and blockchain. We perceive fundamental troubles of current EHRs sharing structures and propose useful responses for resolving these issues through a veritable model execution. In this work, our consideration is on arranging a solid access control framework reliant upon a singular wise consent to regulate customer access for ensuring useful and secure EHRs sharing. To inspect the performance of the proposed approach, we pass on an Ethereum blockchain on the Amazon cloud, where clinical components can associate with the EHRs sharing system through a made flexible Android application. We moreover join the circulated IPFS amassing structure with blockchain to achieve

decentralized data storing and data sharing. The execution results show that our design can allow clinical customers to share clinical data over adaptable cloud conditions in a reliable and fast manner, conversely with normal plans. In particular, our passageway control can perceive and hinder effectively unapproved induction to the e-prosperity system, focusing on achieving an optimal level of patient insurance and association security. We similarly give security assessments and wide appraisals on various specific pieces of the proposed system.

A

PROJECT REPORT

On

FACIAL EXPRESSION RECOGNITION AND THEIR TEMPORAL SEGMENTS FROM FACE PROFILE IMAGE SEQUENCES USING YOLO OBJECT DETECTION ALGORITHM

Submitted by

1)Ms. V. Akanksha Rao (17K81A05B6)

2)Ms. Gogineni Darvika (17K81A0581)

3)Mr. Abhishek Harish Thumar (17K81A0561)

4)Mr. Bundey Mohtih (17K81A0566)

in partial fulfillment for the award of the degreeof

BACHELOR OF TECHNOLOGY
IN
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Under the Guidance of

Ms. E. Soumya
Assistant Professor
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ST.MARTIN'S ENGINEERING COLLEGE
An Autonomous Institute

Dhulapally, Secunderabad – 500 100

JUNE 2021

BONAFIDE CERTIFICATE

This is to certify that the project entitled Facial Expression Recognition on their Temporal Segments from face profile image sequences using YOLO object detection Algorithm, is being submitted by **1.Ms. V. Akanksha Rao**, **17K81A05B6, 2. Gogineni Darvika, 17K81A0581, 3. Abhishek Harish Thumar, 17K81A0561, 4. Bundey Mohith, 17K81A0566** in partial fulfillment of the requirement for the award of the degree of **BACHELOR OF TECHNOLOGY IN Computer Science and Engineering** is recorded of bonafide work carried out by them. The result embodied in this report have been verified and found satisfactory.

E. Soumya
Department of CSE

Head of the Department
Dr.M.NARAYANAN
Department of CSE

Internal Examiner

External Examiner

Place:
Date:

DECLARATION

We, the student of **Bachelor of Technology** in Department of Computer Science and Engineering', session: <2017 – 2021>, St. Martin's Engineering College, Dhulapally, Kompally, Secunderabad, hereby declare that work presented in this Project Work entitled "Facial Expression Recognition and their Temporal Segments From Face Profile image Sequences Using YOLO Object Detection Algorithm" is the outcome of our own bonafide work and is correct to the best of our knowledge and this work has been undertaken taking care of Engineering Ethics. This result embodied in this project report has not been submitted in any university for award of any degree.

V. Akanksha Rao 17K81A05B6

Gogineni Darvika 17K81A0581

Abhishek Harish Thumar 17K81A0561

Bundey Mohith 17K81A0566

ABSTRACT

Automatic face expression analysis is a difficult problem with various applications. The majority of currently available automated facial expression analysis algorithms aim to recognise a few prototypical emotional expressions, such as anger and happiness. The method provided, rather than being another approach to machine analysis of prototypic facial expressions of emotion, seeks to manage a wide range of human facial behaviour by identifying facial muscle motions that form expressions. The method provided, will use three algorithms: You only look Once (YOLO), Convolution Neural Networks (CNN) and Recurrent Neural Network (RNN) to identify facial expressions. With a combination of these three, this project can recognize the expressions in both frontal view images and profile view images.

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible and whose encouragements and guidance have crowded effects with success.

We extended our deep sense of gratitude to Principal**, Dr. P. SANTOSH KUMAR PATRA**, St. Martin's Engineering College, Dhulapally, for permitting us to undertake this project.

We are also thankful to **Dr. M.NARAYANAN**, Head of the Department, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

We would like to thank **Dr. T. POONGOTHAI,** Department of Computer Science and Engineering (AI & ML) for her encouragement and insightful comments and as well as our project coordinators **Dr. B.RAJALINGAM**, Associate Professor and **Dr.GOVINDA RAJULU.G** Associate Professor, in Department of Computer Science and Engineering for their valuable support.

We would like to express our sincere gratitude and indebtedness to our project supervisor E. Soumya, Assistant Professor, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for her support and guidance throughout our project.

Finally, we express thanks to all those who have helped us successfully to completing this project. Furthermore, we would like to thank our family and friends for their moral support and encouragement.

We express thanks to all those who have helped us in successfully completing the project.

V. Akanksha Rao 17K81A05B6

Gogineni Darvika 17K81A0581

Abhishek Harish Thumar 17K81A0561

Bundey Mohith 17K81A0566

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ACRONYMS

YOLO	You Only Look Once
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
GPU	Graphical Processing Unit
CPU	Central Processing Unit
UML	Unified Modeling Language
FER	Facial Expression Reecognition
Iot	Internet of Things
AR	Augmented Reality
FOL	Floor of log
HMM	Hidden Model Markov
SVM	Support Vector Machine
RAM	Random Access Memory

1. INTRODUCTION

In recent years, Artificial Intelligence and Machine Learning are among the most remarkable technological developments. The technology shift seems almost magical but an inevitable development given its fast paced growth. Machine learning is a subset of Artificial Intelligence, while Deep Learning is a subset of Machine Learning. Image recognition and Facial Expression recognition are the applications of Deep Learning. Based on human brain structure and allows us to train the computer to learn through examples. Face detection entails splitting an image into two classes: one with faces (targets) and another with background (clutter) that must be deleted. Faces have commonalities, yet they differ in terms of age, skin colour, and facial expression, making it difficult to distinguish between them. Different lighting conditions, picture quality, and geometries complicate the situation even further. In every backdrop environment, a face detector should be able to identify the presence of any face under various lighting situations. This can be achieved by breaking down the face detection analysis phase into two phases. The first phase is a classification task, which takes some random picture areas as input and returns a binary result of yes or no, indicating whether or not there are any faces in the image. The face localization job is the second phase of the face detection analysis which requires an image to be input and the position of any face or faces inside that picture to be produced as some bounding box/boxes with (x, y, width, height).

Once the faces are detected from the picture and separated from the background, different Deep Learning algorithms can be applied to the dataset to analyze and detect the facial expressions. According to Ekman, there are six universal expressions: fear, disgust, surprise, anger, sorrow, and happiness. To recognise these expressions, the task is to look at the variations in the face. For instance, we may say a person is happy if they make a grin motion with their clenched eyelids and lifted mouth corners. Facial expressions indicate a person's psychological moods, social communication, and intentions; therefore this has a wide range of applications in mental health, psychological, and behavioural study. Face Recognition using Histogram of Oriented Gradients Using CNN Detection has been a hot topic in the technology community, as human beings have discovered that facial expressions are one of the most natural and powerful ways to convey their intents and feelings. Feature learning, feature selection and classifier construction are the three main processes in the training process for facial expression recognition systems. After the feature learning step, only learned face expression variances among all characteristics are retrieved. The finest characteristics

are then picked through feature selection to depict facial emotion. The aim is to decrease intra-class variations of expressions as well as to maximise inter-class variety. Because the identical expressions of various individuals in an image are widely apart in pixel space, reducing intra-class variance in expressions is a challenge.

1.1 PROJECT OVERVIEW

This project helps identify faces in images and helps classify these faces into one of seven emotions. The main steps in the project are:

1. Image pre-processing
2. Training the Algorithms
3. Classifying the expression

1.2 PROJECT OBJECTIVES

The project objective is to make the detection of facial expressions more accurate and specific. Through this project we observed that a combination of 3 algorithms i.e. YOLO, CNN and RNN resulted in a more accurate prediction of facial expression. And unlike existing systems these combined algorithms were able to predict one of seven expressions even in profile view images.

1.3 ORGANIZATION OF CHAPTERS

Chapter 1: Introduction - This chapter contains a brief introduction to the project and to what it entails

Chapter 2: Literature Survey - Chapter 2 consists of learnings and research from other research papers and how they affected this project.

Chapter 3: Software and Hardware Requirements - This is simply the Software and Hardware that would be needed for this project to be executed.

Chapter 4: Software Development Analysis - Software Development Analysis is the chapter in which the problem which we want to address is defined along with the project modules explanation

Chapter 5: Project System Design - Here our projects is depicted visually through UML Diagrams

Chapter 6: Project Coding - Sample codes, files used, classes and their functions are all mentioned in this chapter.

Chapter 7: Project Testing - The testing techniques that were used in this project are all mentioned here.

Chapter 8: Output Screens - Screenshots of the output

Chapter 9: Experimental Result - A brief explanation of how we obtained the output.

Chapter 10: Conclusion and Future Enhancement - Here we conclude and share ideas of any future enhancements.

2. LITERATURE SURVEY

Review on Facial Expression Recognition System Using Machine Learning Techniques

This research examined the recognition system of facial expression. FER is implemented in a variety of applications, including medical, lie detection, cognitive activity, robotic, forensic, automated training, safety, identification of the intellectual state, mood music, tiredness monitoring for operators, etc. This article explains the publicly accessible FER datasets. Different approaches of FER extraction and categorization are compared in this paper.

Systematic review of 3D facial expression recognition methods

Three main features of the 3D FER research published between 2013 and 2018 are detailed in the results of the systematically documented review process outlined in this work: facial representation approaches, kinds of necessary preprocessing, and lastly the nature of the classification tests. This paper compares the traditional approaches for face recognition and the deep learning approaches which have been increasingly adopted in 3D FER.

A fast face detection method via convolutional neural network

This study presented a quick-face recognition approach based on discriminative complete features derived by CNN. Experimental findings have demonstrated that various popular face recognition data sets are capable of achieving promising performance using the suggested DCF facial detection approach.

YOLO v3-Tiny: Object Detection and Recognition using one stage improved model

This paper presents the comparison of several methods in the input image identification and localization of objects on the basis of the precision, time and parameter values. The results of the comparison demonstrate that YOLO v3-Tiny enhances the detection speed and guarantees precision.

Comparison of Different Convolutional Neural Network Structures Based on Keras

This paper explains the categorization of images using Keras libraries for profound learning. This article uses the python for binary image classification and is a deep learning neural network founded on Keras. Deep learning using keras was effective as simulation, train and classification can be carried out with accuracy up to 90%.

Facial Expression Recognition Using Deep Convolutional Neural Networks

In this document, they offer effective CNN architectures to address the challenge of recognition of facial expression. There are a set of convolutionary blocks inside the suggested networks. Each block has a couple of three to three convolution layers, followed by a max pool. Although our suggested networks have considerably lower parameters, they exceed the winning Kaggle competition model. The findings show the strength of tiny filter and highly deep cluster network.

Advances in Computer, Communication and Computational Sciences. Advances in Intelligent Systems and Computing

This paper highlights the downsides of the traditional methods that comprise digital negative, thresholding, and bit-plane slicing are way too simple to enhance the efficiency. But, combining the traditional methods with the advanced algorithms helps us save computational time and more accurate and faster results.

An Emotion Recognition Model Based on Facial Recognition in Virtual Learning Environment

This paper provided a proposed model to solve the problems of emotion recognition based on facial recognition in virtual learning environments, and the efficiency and accuracy are considered at the same time. The application of emotion recognition in virtual learning environments is a much-researched topic.

Face Detection and Recognition Using OpenCV
This paper gave us insight on how the traditional algorithms works with OpenCV and Cascade method and all its downsides on how Cascade method is one dimensional due to its binary pattern recognition.

Facial Recognition, Expression Recognition, and Gender Identification
This paper helped us tackle one of the barriers while detecting facial features of different genders as there is a vast difference between the features of a man and a woman. This was possible by training the CNN algorithm on the data with an initial accuracy of 70% and after training the highest accuracy of 96% was achieved.

Where am I from? –East Asian Ethnicity Classification from Facial Recognition
This paper helped us detect and identify faces of different ethnicities with the help of CNN algorithm. People from different ethnicities have different facial features and the expression may vary slightly. So it was important for the algorithm to detect all ethnicities which improves the learning rate and the efficiency of the algorithm.

A high-efficiency energy and storage approach for IoT applications of facial recognition
One of the barriers we faced was about the storage of the data. Facial recognition in general requires a large dataset. This paper helped us store and compress this dataset more efficiently. This further enhances the efficiency and opens the door for the IoT applications to use the features in real time. To prove the compression capacity of FoL(Floor of Log algorithm), four datasets were used, in which the method achieved a reduction of 88.4%, 88.5%, 91.3%, and 86.3%, in the AR, LFW, ExtendedYaleB and CelebA bases, respectively, in relation to the original size of the file containing the uncompressed features.

Human Emotions Recognition from Thermal Images using Yolo Algorithm
This paper describes how prominent is YOLOv3 technique for detecting objects and human emotion recognition based on thermal imaging, the model shows that YOLO and Darknet framework can ultimately be used to predict facial expression.It endorse a temperature space method to sort out eyeglasses utilizing the thermal facial information . The background of the thermal image is eliminated by using a background elimination algorithm.

Attention Mechanism-based CNN for Facial Expression Recognition

This paper explains how convolutional neural network with attention mechanism is used for facial expression recognition. It combines LBP features and CNN features with attention mechanism to improve performance of the network.

Exploiting multi-CNN features in CNN-RNN based Dimensional Emotion Recognition on the OMG in-the-wild Dataset.

This paper presents the development of novel architectures for predicting facial expressions. The system combines CNN and RNN algorithms. Low CNN layers contain rich, complete and time varying information, whilst high-level features are highly specific and characteristic of the specific problem studied. Taking this into account, they have developed and used CNN plus Multi-RNN networks these networks extract low-, mid- and high- level features from different layers of the CNN and pass them through RNNs.

2.1 SURVEY ON BACKGROUND

The automated facial expression literature has increased significantly in recent years by the use of sophisticated image processing and analysis techniques. Most automated facial recognition experiments concentrate on fewer facial expressions, including happiness, sorrow, anger. Ekman's detailed research identified the presence and acknowledgement of these key facial features universal in various cultures. Computer-aided facial recognition experiments were initiated in the 1990s. Several people later explored various facial expression recognition methods and models such as optical flow, radial function network, the hidden Model Markov (HMM), a neural-network-based classifier etc. A cascade-based approach for identifying faces in a certain image has recently been proposed. By taking the alignment phase in a cascade structure, this approach increased detection efficiency. CNN performs well with various vision-related activities by acquiring more than 80 percent of the features extracted. In comparison to the best previous performance of the 2012 VOC data collection, R-CNN increases average precision of over 30% and achieves a 53.3% map. Though CNN has demonstrated good object detection ability, its computer productivity has been insufficient. With the sliding window technique, OverFeat gets multi-scale dense functionality. Because OverFeat detects all possible positions and scales using the classifier

and regression network, this process in OverFeat takes so much time. To enhance this effectiveness, R-CNN first produces a number of class independent regions, then features are extracted on multi-scale images from regions with qualified CNN. Each proposal window would then be allocated a result by using a linear SVM (Support Vector Machines) for the functions. A detector is run on a 500 x 375 picture in approximately 15 seconds for R-CNN. The pictures are read and collected using OpenCV and the framework also proposes a web service for recording images through images or live camera. Currently, this approach achieves best image and live camera results.

2.2 CONCLUSIONS ON SURVEY

After referring to the various papers, we were able to tackle most of our obstacles that included the efficiency of the algorithms to facial features detection. One factor that was clearly evident was that integrating traditional methods with the advanced algorithms helped us increase the efficiency of the whole detection process. With the help of YOLO, CNN and RNN, we were able to divide the process into 3 stages, face detection and preprocessing, feature selection and lastly detection of facial expression. By completely excluding Cascade methods, we were able to achieve multiple dimensions to the detection process. And lastly, we were able to manage our dataset effectively even after compressing the images in the dataset without losing any important data and features of the images.

3. SOFTWARE AND HARDWARE REQUIREMENT

The project involved analysing the design of few applications so as to make the application more users friendly. To do so, it was really important to keep the navigations from one screen to the other well-ordered and at the same time reducing the amount of typing the user needs to do. In order to make the application more accessible, the browser version had to be chosen so that it is compatible with most of the Browsers.

3.1 SOFTWARE REQUIREMENTS

For developing the application the following are the Software Requirements:

1. Python

Operating Systems supported

1. Windows

Technologies and Languages used to Develop

1. Python

Debugger and Emulator

- Any Browser (Particularly Chrome)

3.2 HARDWARE REQUIREMENTS

For developing the application the following are the Hardware Requirements:

- Processor: Pentium IV or higher
- RAM: 256 MB or higher
- Space on Hard Disk: minimum 512MB
- Preferably a GPU

4. SOFTWARE DEVELOPMENT ANALYSIS

4.1 OVERVIEW OF PROBLEM

Facial expressions play an important role in the recognition of human emotions. It is a known fact that human emotions play an important role in our day to day work.

Facial expression detection helps in some of the major field of works, some of them are:

Psychology

Scientific research

Security and lie detection:

Music for mood

Facial animation

4.2 DEFINE THE PROBLEM

Psychology: Facial expressions have a major role in detecting human emotions. Psychologists deal with the mental and emotional state of a person. By using Facial expression detection we

can detect the emotion and can derive the mental state of the individual even without being in person.

Scientific research: Describing in terms of science , detecting a subject's emotional and mental state can be very vital for the research which can be achieved by facial expression detection.

Facial Animation: Facial animation is an area of computer graphics that consists of methods and techniques for generating and animating models of a human, an animal, or a fantasy character face.

Lie detection and Security: During interrogation detecting a lie can be crucial , it is a known fact that facial expression can give out a lie to the trained eye. In prisons it is very important to avoid any brawls by detecting the mental state of inmates through expressions.

Music for mood: Music is many people's go to for relieving stress and it is also helpful for mental health , through facial expression detection we can tell the mood of the individual and based on their mood play the music .

4.3 MODULES OVERVIEW

The project is divided into five modules:

Module 1: Data collection

Module 2: Pre-Processing the data

Module 3: Training the algorithms

Module 4: Model accuracy

Module 5: Output

4.4 DEFINE THE MODULES

Module 1: Data Collection- In this module, various datasets of images are collected and loaded into the Model. These datasets have various facial expressions.

Module 2: Pre-Processing the data- The data collected in the previous module has too much information, we get rid of it by Pre-Processing the data.

Module 3: Training the Algorithms- The Pre-Processed data is used to train the algorithms.

Module 4: Model Accuracy- Once the algorithms are trained, it gives us an accuracy graph.

Module 5: Output- We can upload an image and predict the facial expression of the uploaded image.

4.5 MODULE FUNCTIONALITY

Module 1: Data Collection- The data set for this project was taken from Kaggle. The dataset consists of 28709 images. This helped train the algorithms and played a big part.

Module 2: Pre-Processing the data- All the images in the data set might not be able to be used to train the algorithms. For example all the images do not have the minimal number of facial points. So by pre-processing the data we are removing the images that we can not use.

Module 3: Training the Algorithms- YOLO is the first algorithm that identifies the faces in the image. Then the CNN and RNN algorithms get trained with the filters and the facial points on features like lips, eyes and morning. The algorithms learn that with different places for the facial points the expression is different.

Module 4: Model Accuracy- The algorithm learns and shows us how accurate it is.

Module 5: Output- After uploading an image, the algorithms try and predict the expression in the face in the image. The combination of YOLO, CNN and RNN is the combination which gives the best accuracy that we observed.

5. PROJECT SYSTEM DESIGN

5.1 UML DIAGRAMS

UML stands for Unified Modeling Language. UML is a standardized general-purpose modeling language in the field of object-oriented software engineering. The standard is managed, and was created by, the Object Management Group.

The goal is for UML to become a common language for creating models of object oriented computer software. In its current form UML is comprised of two major components: a Meta-model and a notation. In the future, some form of method or process may also be

added to; or associated with, UML.

The Unified Modeling Language is a standard language for specifying, Visualization, Constructing and documenting the artifacts of software system, as well as for business modeling and other non-software systems.

The UML represents a collection of best engineering practices that have proven successful in the modeling of large and complex systems.

The UML is a very important part of developing objects oriented software and the software development process. The UML uses mostly graphical notations to express the design of software projects.

Goals:

The Primary goals in the design of the UML are as follows:

1. Provide users a ready-to-use, expressive visual modeling Language so that they can develop and exchange meaningful models.
2. Provide extendibility and specialization mechanisms to extend the core concepts.

USE CASE DIAGRAM:

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.

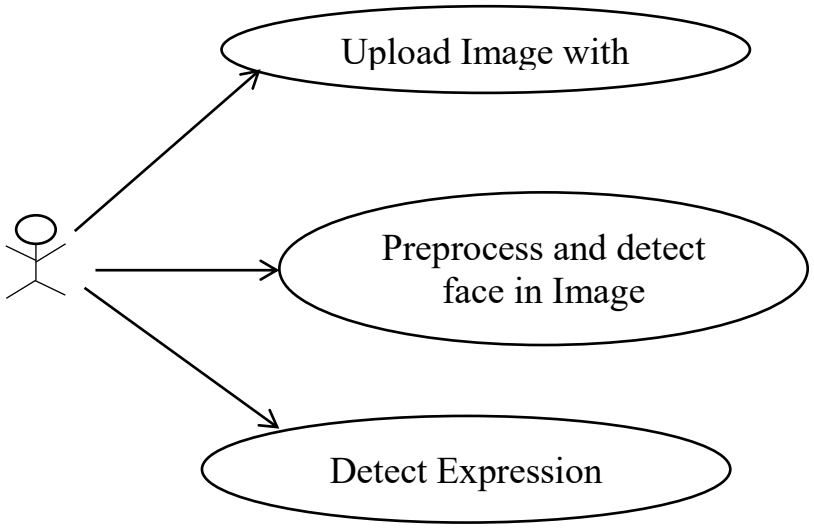


Figure 5.1

CLASS DIAGRAM:

In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.

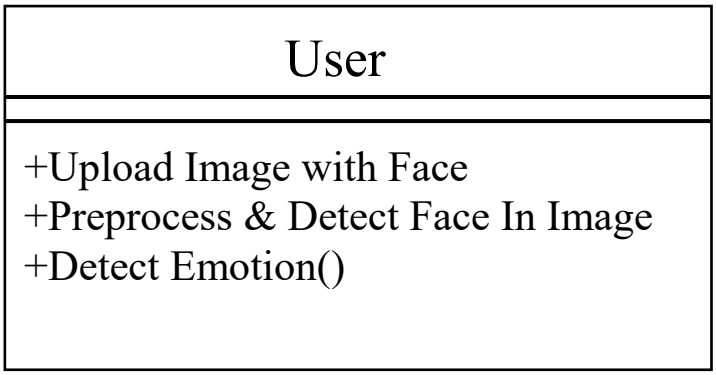


Figure 5.2

SEQUENCE DIAGRAM:

A sequence diagram in Unified Modeling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.

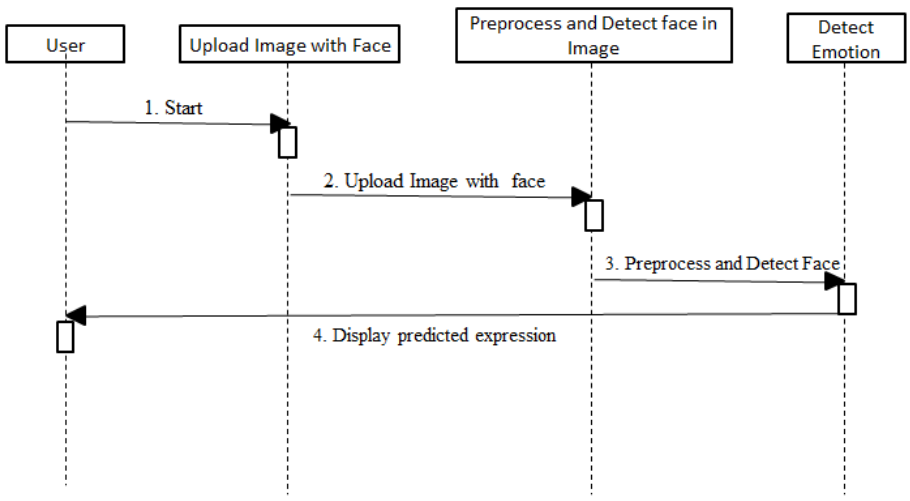


Figure 5.3

ACTIVITY DIAGRAM:

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modeling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.

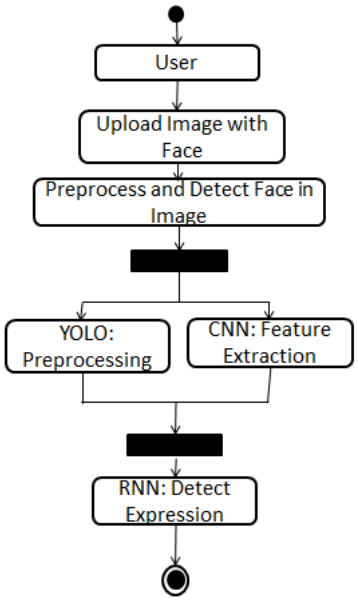


Figure 5.4

STATE CHART DIAGRAM:

A state diagram is a type of diagram used in computer science and related fields to describe the behavior of systems. State diagrams require that the system described is composed of a finite number of states; sometimes, this is indeed the case, while at other times this is a reasonable abstraction.

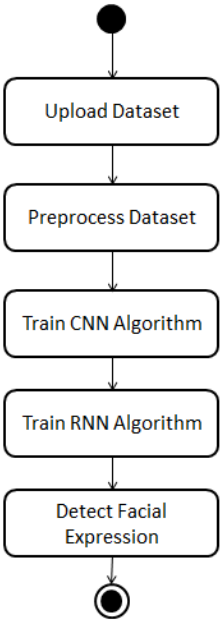


Figure 5.5

DATA FLOW DIAGRAM:

A data flow diagram shows the way information flows through a process or system. It includes data inputs and outputs, data stores, and the various subprocesses the data moves through.

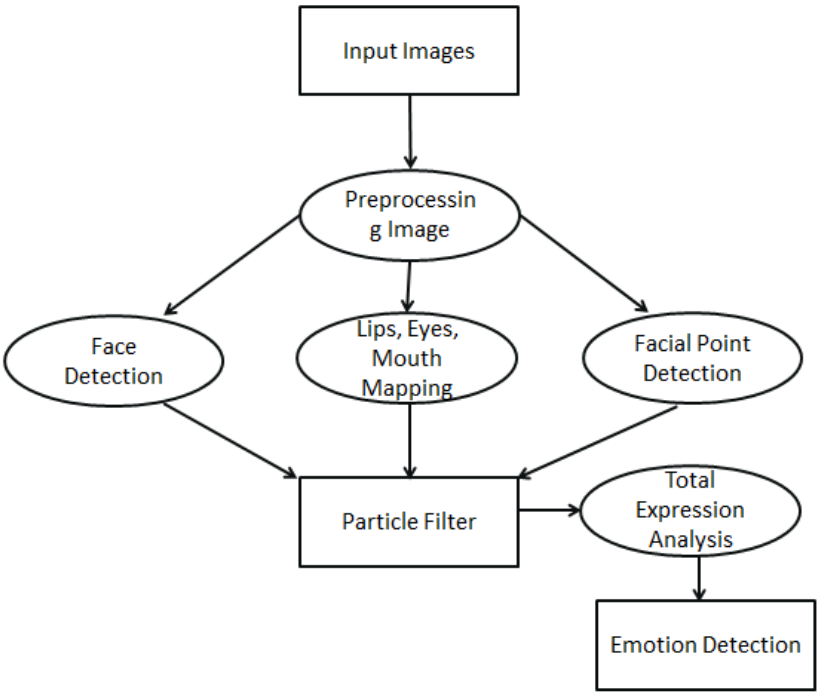


Figure 5.6

6. PROJECT CODING

6.1 CODE TEMPLATES

```
font = ('Helvetica', 13,'bold')

title = Label(main, text='Automatic Facial Expression Recognition using Features Extraction Based on Spatial & Temporal Sequences using CNN & RNN Algorithm')

title.config(bg='#3500D3', fg='#FFFFFF')

title.config(font=font)

title.config(height=3, width=130)

title.place(x=5,y=5)


font1 = ('Helvetica', 12,'bold')

text=Text(main,height=11.5,width=77)

scroll=Scrollbar(text)

text.configure(yscrollcommand=scroll.set)

text.place(x=320,y=100)

text.config(font=font1)


uploadButton = Button(main,bg='#3500D3', fg='#FFFFFF',text="Upload Facial Emotion Dataset", command=upload)

uploadButton.place(x=50,y=100)

uploadButton.config(font=font1)


processButton = Button(main,bg='#3500D3', fg='#FFFFFF', text="Preprocess Dataset", command=processDataset)

processButton.place(x=50,y=200)

processButton.config(font=font1)


cnnButton = Button(main,bg='#3500D3', fg='#FFFFFF', text="Train CNN Algorithm with YOLO Faces", command=trainCNN)

cnnButton.place(x=1050,y=100)

cnnButton.config(font=font1)
```

```
graphButton = Button(main,bg='#3500D3', fg='#FFFFFF', text="Accuracy Comparison Graph", command=graph)

graphButton.place(x=50,y=300)

graphButton.config(font=font1)


predictButton = Button(main,bg='#3500D3', fg='#FFFFFF', text="Predict Facial Expression", command=predict)

predictButton.place(x=1050,y=200)

predictButton.config(font=font1)


exitButton = Button(main,bg='#3500D3', fg='#FFFFFF', text="Exit", command=exit)

exitButton.place(x=1050,y=300)

exitButton.config(font=font1)
```

6.2 OUTLINE FOR VARIOUS FILES

Tensorflow: It is an open source artificial intelligence library, using data flow graphs to build models. It allows developers to create large-scale neural networks with many layers. TensorFlow is mainly used for: Classification, Perception, Understanding, Discovering, Prediction and Creation.

numpy: NumPy is a Python library used for working with arrays.It also has functions for working in domain of linear algebra, fourier transform, and matrices. Files are saved as .npy

Pandas: Pandas is mainly used for data analysis. Pandas allows importing data from various file formats such as comma-separated values, JSON, SQL, Microsoft Excel. Pandas allows various data manipulation operations such as merging, reshaping, selecting, as well as data cleaning, and data wrangling features.

Matplotlib: Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib makes easy things easy and hard things possible.

Keras: Keras is a powerful and easy-to-use free open source Python library for developing and evaluating deep learning models. It wraps the efficient numerical computation libraries

Theano and TensorFlow and allows you to define and train neural network models in just a few lines of code.

Opencv: OpenCV is the huge open-source library for the computer vision, machine learning, and image processing and now it plays a major role in real-time operation which is very important in today's systems. By using it, one can process images and videos to identify objects, faces, or even handwriting of a human.

6.3 CLASS WITH FUNCTIONALITY

Class upload(): This class is called when we upload a dataset into the model. This class lets us upload the file from our drive.

Class processdataset(): This class uses YOLO algorithm to pre-process the dataset and remove all the useless information and only detect the images with faces in it. It gives a number count and classes of the images detected from the dataset.

Class trainCNN(): This class loads the pre-trained CNN algorithm and the pre-processed dataset is used to further train the algorithm.

Class trainRNN(): This class loads the pre-trained RNN algorithm and the dataset is used to further train the algorithm and increase its accuracy.

Class graph(): This class is called to view the accuracy of the learning model in graph format.

Class predict(): When this class is called, the user can upload an image into the model and the model will predict the expression of the image

Class exit(): This class is used to exit the application.

6.4 METHODS INPUT AND OUTPUT PARAMETERS

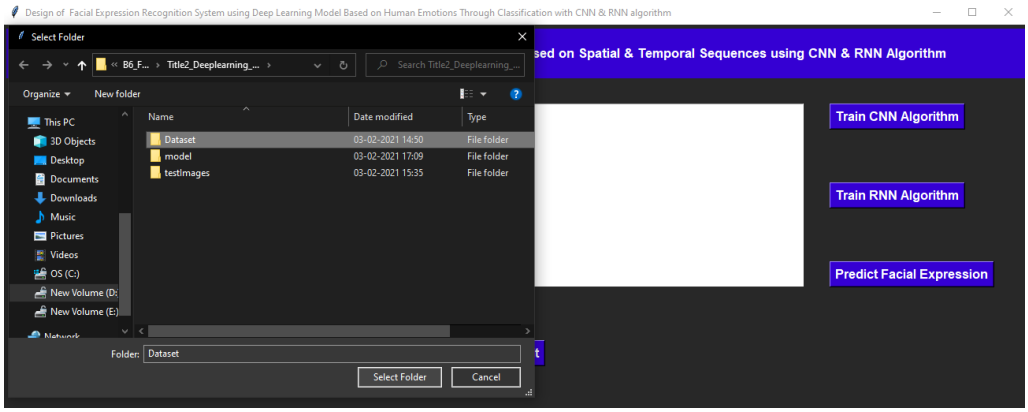


Figure 6.1

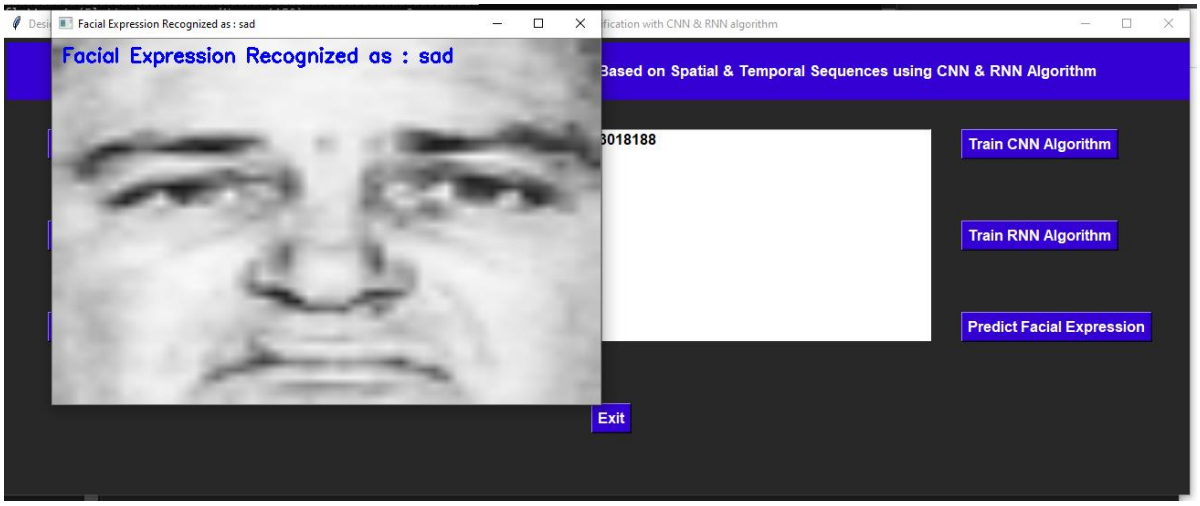


Figure 6.2

7. PROJECT TESTING

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub-assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

7.1 VARIOUS TEST CASES

Unit testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

Integration testing

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

Functional test

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items:

Valid Input : identified classes of valid input must be accepted.

Invalid Input : identified classes of invalid input must be rejected.

Functions : identified functions must be exercised.

Output : identified classes of application outputs must be exercised.

Systems/Procedures : interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

System Test

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

7.2 BLACK BOX TESTING

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box .you cannot “see” into it. The test provides inputs and responds to outputs without considering how the software works.

7.3 WHITE BOX TESTING

White Box Testing is a testing in which in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is purpose. It is used to test areas that cannot be reached from a black box level.

8. OUTPUT SCREENS

8.1 USER INTERFACE

Figure 8.1

8.2 OUTPUT SCREENS

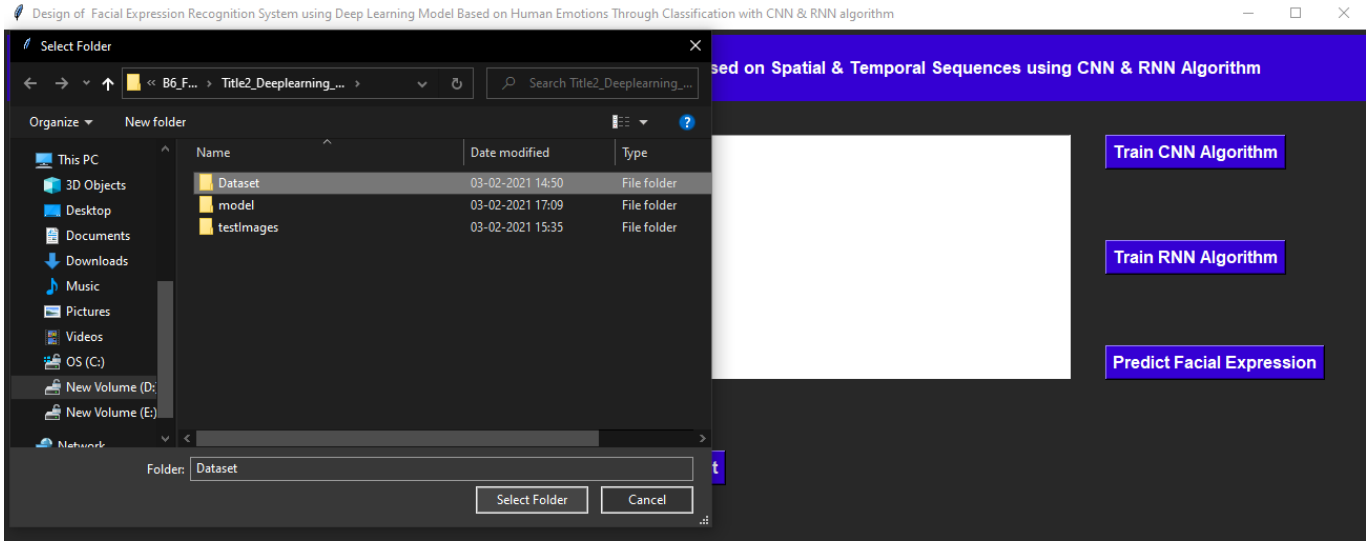


Figure 8.2

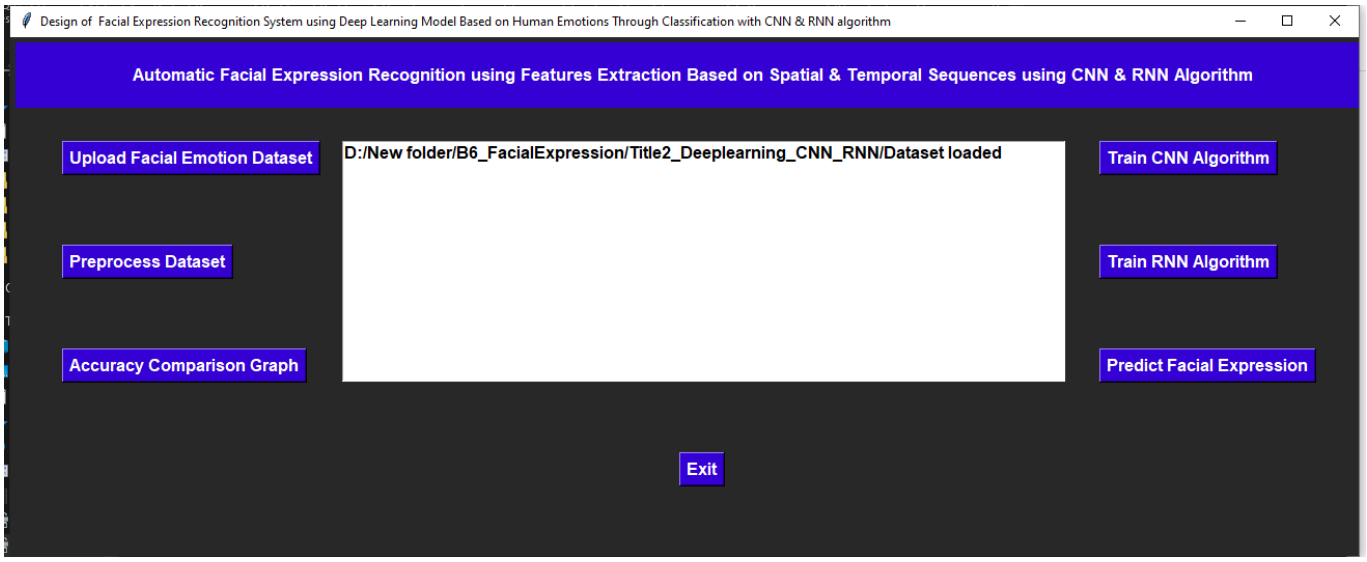


Figure 8.3

Figure 8.4

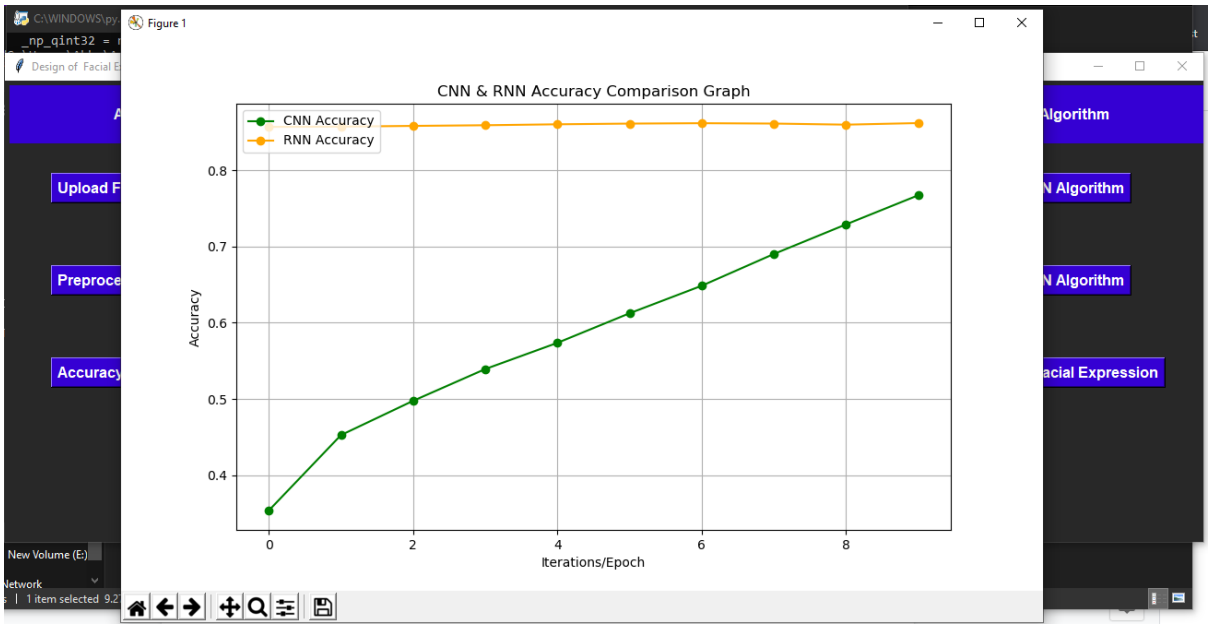


Figure 8.5

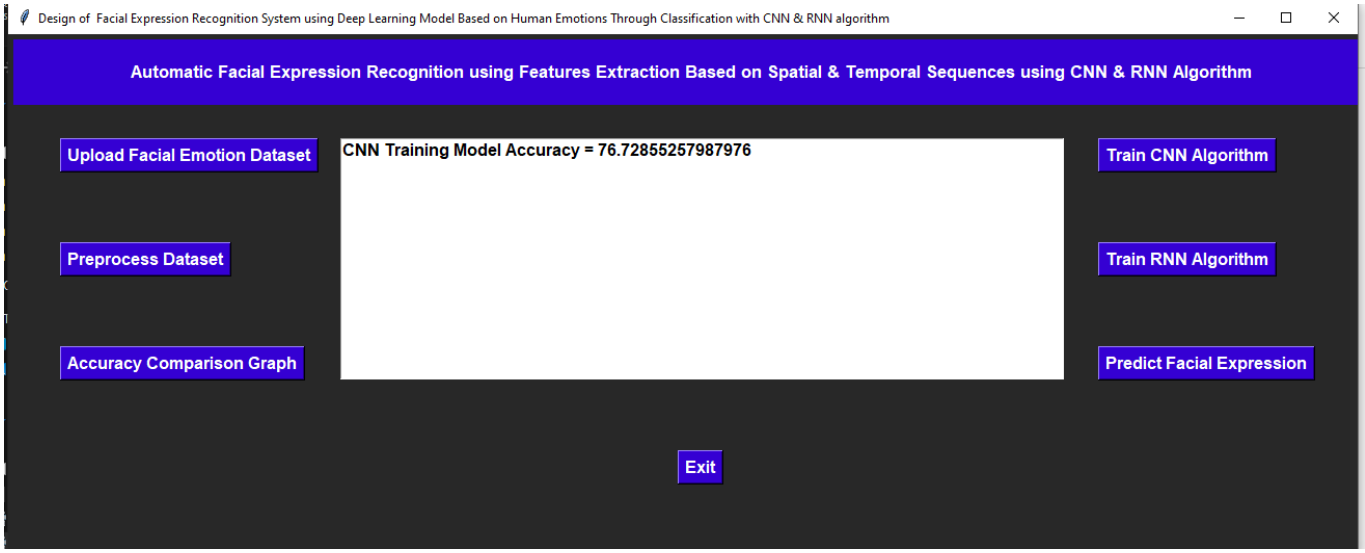


Figure 8.6

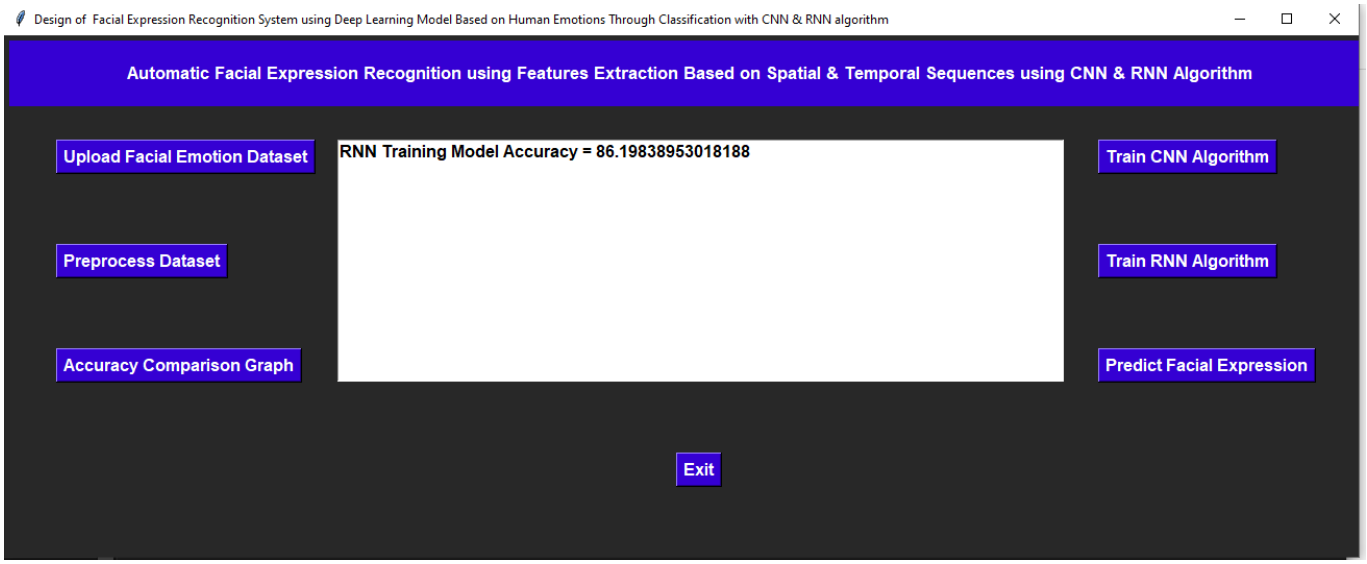


Figure 8.7

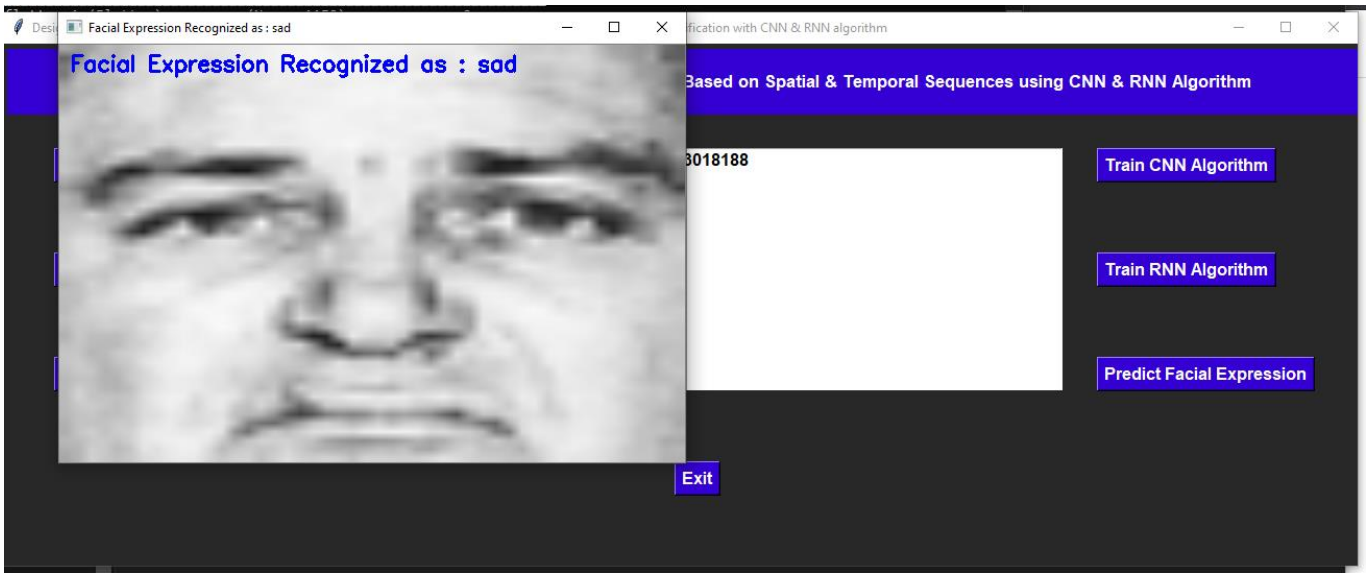


Figure 8.8

9. EXPERIMENTAL RESULTS

The experimental result of this project can be described in sections. Initially in order for the algorithms to learn and go off of a dataset we must upload one. The dataset that we used is from Kaggle and it has approximately 2500 images. This dataset needed to be preprocessed to find all the usable images. After preprocessing that is when we can start using the algorithms. We noticed that a combination of three algorithms i.e. YOLO, CNN and RNN give higher accuracy. YOLO algorithm detects the faces in the images and removes the background. Later CNN and RNN are trained with the help of filters and facial points and how the position of the facial points help us classify the expression on the face into one of seven expressions. The accuracy of both is mentioned on the output screen. We can also observe the graph of accuracy. Lastly we choose an image to upload for the expression to be predicted. RNN helps in predicting the expression in the image uploaded.

10. CONCLUSIONS AND FUTURE ENHANCEMENTS

On numerous databases such as FIA, Chokepoint, and others, we investigated numerous algorithms for face identification and pattern recognition techniques and came to the conclusion that the YOLO method has one of the most presentable architectures, is easy to comprehend, and gives one of the most exact and efficient results. Because it just requires one pass of the convolution network to produce the desired output of the specified input image. We noticed that when YOLO is specifically paired with CNN and RNN we got the best results. In our system, we use a combination of algorithms: YOLO, CNN and RNN. The proposed system can detect expressions in both frontal and profile view images. It classifies the expressions into one of seven categories. Initially, YOLO algorithm is used to preprocess and detect faces in the images. Then CNN helps extract facial features and adds nodes on the face. Lastly, RNN is applied to detect the facial expression. One drawback is that in system requirements we require high GPU. Another place of improvement is that it works only on images not videos. This only works on previous version of Python. In the future we would like it to be compatible with systems which don’t have higher specifications. We would like to improve it to work on videos as well.

REFERENCES

1. Bhatia, S. K., Tiwari, S., Ruidan, S., Trivedi, M. C., & Mishra, K. K. (Eds.). (2021). Advances in Computer, Communication and Computational Sciences. Advances in Intelligent Systems and Computing.
2. Peixoto, S. A., Vasconcelos, F. F. X., Guimarães, M. T., Medeiros, A. G., Rego, P. A. L., Lira Neto, A. V., … Rebouças Filho, P. P. (2020). A high-efficiency energy and storage approach for IoT applications of facial recognition. Image and Vision Computing, 103899.
3. Jing Li, Kan Jin, Dalin Zhou, Naoyuki Kubota, Zhaojie Ju(2020).Attention Mechanism-based CNN for Facial Expression Recognition.
4. Chaitanya, Sarath S, Malavika, Prasanna and Karthik (2020).Human Emotions Recognition from Thermal Images using Yolo Algorithm
5. Pooja.V.Magdum, & Mahadev.S.Patil. (2020). Comparison of Different Convolutional Neural Network Structures Based on Keras. Journal of Electronics and Communication Systems, 5(1), 1–11. http://doi.org/10.5281/zenodo.36128 01
6. Kollias, D., & Zafeiriou, S. P. (2020). Exploiting multi-CNN features in CNN-RNN based Dimensional Emotion Recognition on the OMG in-the-wild Dataset
7. Adarsh, P., Rathi, P., & Kumar, M. (2020). YOLO v3-Tiny: Object Detection and Recognition using one stage improved model. 2020 6th International
8. Fathima, A., & Vaidehi, K. (2019). Review on Facial Expression Recognition System Using Machine Learning Techniques. Advances in Decision Sciences, Image Processing, Security and Computer Vision, 608–618.
9. Alexandre, G. R., Soares, J. M., & Pereira Thé, G. A. (2019). Systematic review of 3D facial expression recognition methods. Pattern Recognition, 107108.
10. Guo, G., Wang, H., Yan, Y., Zheng, J., & Li, B. (2019). A Fast Face Detection Method via Convolutional Neural Network. Neurocomputing.
11. Khan, M., Chakraborty, S., Astya, R., & Khepra, S. (2019). Face Detection and Recognition Using OpenCV. 2019 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS).
12. Conference on Advanced Computing and Communication Systems (ICACCS).
13. Yang, D., Alsadoon, A., Prasad, P. W. C., Singh, A. K., & Elchouemi, A. (2018). An Emotion Recognition Model Based on Facial Recognition in Virtual Learning Environment. Procedia Computer Science, 125, 2–10.

14. Mane, S., & Shah, G. (2018). Facial Recognition, Expression Recognition, and Gender Identification. Advances in Intelligent Systems and Computing, 275–290.
15. Sang, D. V., Van Dat, N., & Thuan, D. P. (2017). Facial expression recognition using deep convolutional neural networks. 2017 9th International Conference on Knowledge and Systems Engineering (KSE).
16. Haoxuan Chen, Yiran Deng and Shuying Zhang (2016). Where am I from? –East Asian Ethnicity Classification from Facial Recognition. Stanford

PUBLICATIONS

Facial Expression Recognition Using YOLO Object Detection Algorithm

V. Akanksha Rao[1], Gogineni Darvika[2], Abhishek Harish Thumar[3], Bundey Mohith[4], E. Soumya[5]

[1,2,3,4] UG Students [5]Asst. Professor, Department of Computer Science and Engineering,

St. Martin's Engineering College, Near Forest Academy, Dulapally, Kompally, Secunderabad, Telangana 500014, India

Email: v.akankshacs@gmail.com[1], goginenidarvika@gmail.com, abhishek.thumar@gmail.com[3], mohithbundey@gmail.com[4], esoumyait@smec.ac.in[5]

Abstract- Automatic face expression analysis is a difficult problem with various applications. The majority of currently available automated facial expression analysis algorithms aim to recognise a few prototypical emotional expressions, such as anger and happiness. The method provided, will use three algorithms: You only look Once (YOLO), Convolution Neural Networks (CNN) and Recurrent Neural Network (RNN) to identify facial expressions. With a combination of these three, this project can recognize the expressions in both frontal view images and profile view images

Keywords: Faces, CNN, YOLO, RNN, Detection, Recognition, Expressions

I. INTRODUCTION

In recent years, Artificial Intelligence and Machine Learning are among the most remarkable technological developments. The technology shift seems almost magical but an inevitable development given its fast paced growth. Machine learning is a subset of Artificial Intelligence, while Deep Learning is a subset of Machine Learning. Image recognition and Facial Expression recognition are the applications of Deep Learning. Based on human brain structure and allows us to train the computer to learn through examples. Face detection entails splitting an image into two classes: one with faces (targets) and another with background (clutter) that must be deleted. Faces have commonalities, yet they differ in terms of age, skin colour, and facial expression, making it

difficult to distinguish between them. Different lighting conditions, picture quality, and geometries complicate the situation even further. In every backdrop environment, a face detector should be able to identify the presence of any face under various lighting situations. This can be achieved by breaking down the face detection analysis phase into two phases.[1] The first phase is a classification task, which takes some random picture areas as input and returns a binary result of yes or no, indicating whether or not there are any faces in the image. The face localization job is the second phase of the face detection analysis which requires an image to be input and the position of any face or faces inside that picture to be produced as some bounding box/boxes with (x, y, width, height).[2]
Once the faces are detected from the picture and separated from the background, different Deep Learning algorithms can be applied to the dataset to analyze and detect the facial expressions. According to Ekman, there are six universal expressions: fear, disgust, surprise, anger, sorrow, and happiness.[3] To recognise these expressions, the task is to look at the variations in the face. For instance, we may say a person is happy if they make a grin motion with their clenched eyelids and lifted mouth corners. Facial expressions indicate a person's psychological moods, social communication, and intentions; therefore this has a wide range of applications in mental health, psychological, and behavioural study. Face Recognition using Histogram of Oriented Gradients Using CNN Detection has been a hot topic in the technology community, as human beings have discovered that facial expressions are one of the most natural and powerful ways to convey their intents and feelings.[4] Feature learning, feature selection and classifier construction are the three main processes in the training process for facial expression recognition systems. After the feature learning step, only learned face expression variances among all characteristics are retrieved. The finest characteristics are then picked through feature selection to depict facial emotion. The aim is to decrease intra-class variations of expressions as well as to maximise inter-class variety.[5] Because the identical expressions of various individuals in an image are widely apart in pixel space, reducing intra-class variance in expressions is a challenge.
YOLO, SDD, RCNN, and Faster RCNN are some of the techniques that may be implemented for facial expression recognition.

II. RELATED WORK

The automated facial expression literature has increased significantly in recent years by the use of sophisticated image processing and analysis techniques. Most automated facial recognition experiments concentrate on fewer facial expressions, including happiness, sorrow, anger. Ekman's detailed research identified the presence and acknowledgement of these key facial features universal in various cultures. Computer-aided facial recognition experiments were initiated in the 1990s. Several people later explored various facial expression recognition methods and models such as optical flow, radial function network, the hidden Model Markov (HMM), a neural-network-based classifier etc. A cascade-based approach for identifying faces in a certain image has recently been proposed. By taking the alignment phase in a cascade structure, this approach increased detection efficiency. CNN performs well with various vision-related activities by acquiring more than 80 percent of the features extracted. In comparison to the best previous performance of the 2012 VOC data collection, R-CNN increases average precision of over 30% and achieves a 53.3% map. Though CNN has demonstrated good object detection ability, its computer productivity has been insufficient. With the sliding window technique, OverFeat gets multi-scale dense functionality. Because OverFeat detects all possible positions and scales using the classifier and regression network, this process in OverFeat takes so much time. To enhance this effectiveness, R-CNN first produces a number of class independent regions, then features are extracted on multi-scale images from regions with qualified

CNN. Each proposal window would then be allocated a result by using a linear SVM (Support Vector Machines) for the functions. A detector is run on a 500 x 375 picture in approximately 15 seconds for R-CNN. The pictures are read and collected using OpenCV and the framework also proposes a web service for recording images through images or live camera. Currently, this approach achieves best image and live camera results.

III. METHADOLOGY

For the project to work it is intended a coalition of algorithms are used. The algorithms include CNN, RNN, OpenCV and YOLO. CNN and RNN are used for image segmentation to segregate the face from the background of the image. OpenCV and YOLO are used to detect the face and its expression.

A. CNN (Convolution Neural Network):

Advancements in Computer Vision using Deep Learning have been built and developed through time, mostly through the use of a single algorithm – the Convolutional Neural Network. A Convolutional Neural Network (ConvNet/CNN) is a Deep Learning system that can take an input image, assign value (learnable weights and biases) to multiple parts in the image, and distinguish between them.

B. RNN (Recurrent Neural Network):

Recurrent Neural Network algorithm is one of the best algorithms for sequential data. RNNs can recall critical details about the input they receive, allowing them to anticipate what will happen next with great accuracy. RNNs have a "memory" that stores all information about the calculations. It employs the same settings for each input since it produces the same outcome by executing commands on all inputs or hidden layers. The complexity of the parameters is reduced as a result.

C. OpenCV (Open Computer Vision):

OpenCV is a library that we can use in computer vision applications. It can also be used to detect items, faces, and even human handwriting in images and recordings.

D. YOLO (You only look once):

This is an algorithm for detecting and recognising different items in a photograph (in real-time). Object detection in YOLO is done as a regression problem, and the identified images' class probabilities are provided. To detect objects, the approach just takes a single forward propagation through a neural network, as the name suggests. This indicates that a single algorithm run is used to forecast the entire image.

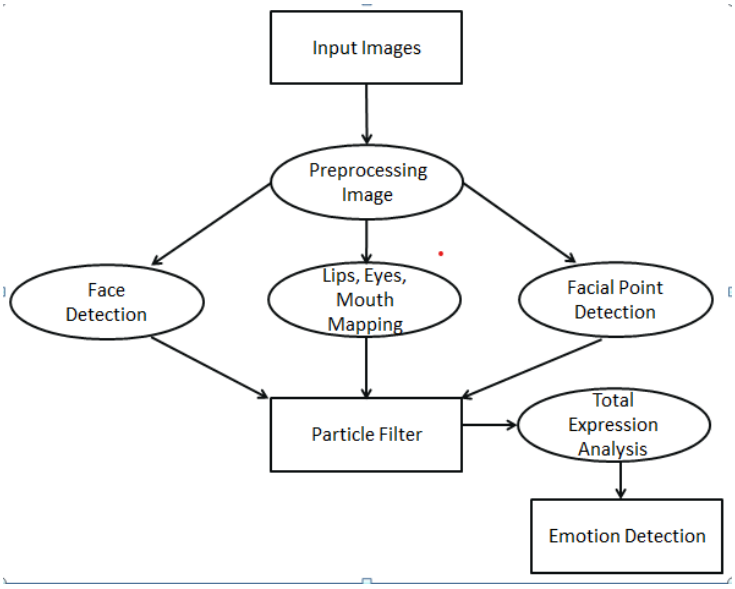


Fig 1: Architecture

In the above figure 1 after giving input images from the database taken from Kaggle, the process starts. First, YOLO removes the unnecessary part in the image in the pre-processing stage. Then YOLO helps identify the faces in the image. CNN learns with the help of the database given and starts mapping the nodes on the lips, eyes and mouth. Other nodes can be added elsewhere on the face like the ends of eyebrows. This is considered facial point detection. The particle filter is basically a filter to detect all these facial points. Then RNN comes in to analyse the expression and detect the apt expression of the face in the image.

IV. EXPERIMENTAL ANALYSIS

We studied how these three algorithms ie. YOLO, CNN and RNN worked together. We analysed that this is one of the most efficient combinations because with YOLO it detects the faces within one pass and then CNN and RNN follow. This method if faster when compared to previous documented methods. This strategy presently delivers the best picture and lives camera performance. In this proposed system, the facial expression detection application undergoes three different steps. In the first step, the YOLO algorithm divides the image into 13 x 13 blocks and the objects are detected in the image. In the second step, the CNN algorithm detects all the faces from the different sets of objects. And lastly, the RNN detects facial expression. This architecture is implemented by integrating three different algorithms. With the help of this architecture we can apply this system to various applications in the field of Automobiles, Interview process, mental health, psychological, and behavioural study, and also for animations and graphics design testing for Video games and animated movies. These are just a few applications to look into. In Automobiles, if the driver is feeling drowsy or sleepy, with the help of this system we can easily detect his facial expressions and warn the driver. A bot can be used to study the facial expressions of the interviewee and detect their traits based on their expression which can be very valuable during the interview process. When it comes to mental health, psychological and behavioural studies, we can use this system to find patterns among patients who are diagnosed under the same condition, which can even help the doctors to examine these conditions better. And when it comes to Graphics testing models, it’s pretty straightforward. With powerful GPUs in the market, the game developers are pushing through the barriers of Graphics and designs. With the help of this system, it’ll help the game developers to capture the expressions better and replicate a highly accurate graphical or animated model.

V. RESULTS AND DISCUSSION

With a coalition of several algorithms we are able to achieve a unique facial expression recognizer that is distinct as it not only works on the frontal view of an image but the profile view as well. Another distinction in this system is that it classifies the expressions into one of seven - Anger, Disgust, Fear, Happiness, Neutral, Sorrow and Surprise. The accuracy of each algorithm is displayed once the algorithm learns. The accuracy and speed of this system are improved by combining these methods. This system has made it easy to upload a database from anywhere for it to learn and then based on the learning we can upload an image for it to be segmented, detected and then classified.

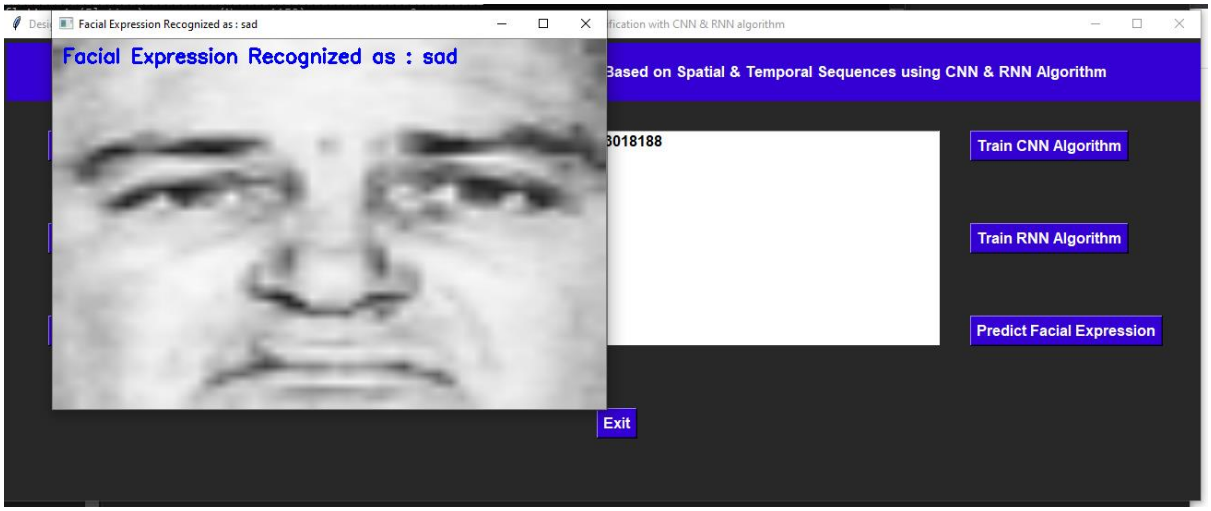


Fig 2: Expression Detection

In the above figure we can see that the expression is detected after the algorithm learns.

VI. CONCLUSION

On numerous databases such as FIA, Chokepoint, and others, we investigated numerous algorithms for face identification and pattern recognition techniques and came to the conclusion that the YOLO method has one of the most presentable architectures, is easy to comprehend, and gives one of the most exact and efficient results. Because it just requires one pass of the convolution network to produce the desired output of the specified input image. We noticed that when YOLO is specifically paired with CNN and RNN we got the best results.

VII. REFERENCES

[1] Jumani, S.Z., Ali, F., Guriro, S., Kandhro, I.A., Khan, A. and Zaidi, A., 2019. Facial Expression Recognition with Histogram of Oriented Gradients using CNN. Indian Journal of Science and Technology, 12, p.24.

[2] Khan, M., Chakraborty, S., Astya, R., & Khepra, S. (2019). Face Detection and Recognition Using OpenCV. 2019 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS).

[3] Garg, D., Goel, P., Pandya, S., Ganatra, A., & Kotecha, K. (2018). A Deep Learning Approach for Face Detection using YOLO. 2018 IEEE Punecon.

[4] Du, J. (2018). Understanding of Object Detection Based on CNN Family and YOLO. Journal of Physics: Conference Series, 1004, 012029. doi:10.1088/1742-6596/1004/1/012029

[5] Revina, I. M., & Emmanuel, W. R. S. (2018). A survey on human face expression recognition techniques. Journal of King Saud University - Computer and Information Sciences.

[6] Chen, X., Yang, X., Wang, M., & Zou, J. (2017). Convolution neural network for automatic facial expression recognition. 2017 International Conference on Applied System Innovation (ICASI)

[7] Kumar, G. A. R., Kumar, R. K., & Sanyal, G. (2017). Facial emotion analysis using deep convolution neural network. 2017 International Conference on Signal Processing and Communication (ICSPC).

[8] Dang, K., & Sharma, S. (2017). Review and comparison of face detection algorithms. 2017 7th International Conference on Cloud Computing, Data Science & Engineering - Confluence.

[9] Shan, K., Guo, J., You, W., Lu, D., & Bie, R. (2017). Automatic facial expression recognition based on a deep convolutional-neural-network structure. 2017 IEEE 15th

[10]International Conference on Software Engineering Research, Management and Applications (SERA).

[11] Ren, S., He, K., Girshick, R. and Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In Advances in neural information processing systems (pp. 91-99).

[12] J. R. Barr, L. A. Cament, K. W. Bowyer, and P. J. Flynn. Active clustering with ensembles for social structure extraction. In Winter Conference on Applications of Computer Vision, pages 969–976, 2014

[13]L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In Computer Vision and Pattern Recognition, pages 529–534,

ALL FOUR STUDENTS ONE PAGE PROFILE

V. Akanksha Rao

Velanky Akanksha Rao is a final year CSE, B.Tech student at St. Martin's Engineering College. She completed her intermediate studies in Narayana Junior College and 10th class in Naryana Concept School. Her technical skills include C and Python. She has a basic understanding of C++, Java, HTML and CSS as well. She was one of the few who were selected to participate in Zensar's Employability Skill Development Program where there were several workshops ranging from technical skills like SQL to soft skills. She has attended seminars and workshops like National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020, Digital Transformation in Education Sector Post- Covid era conducted by Collegedunia on 11th June 2021, Two-Day National Level Seminar on Recent Trends in Cloud Computing ,Fog and Edge Computing from 18th June 2021 to 19th June 2021. She likes doing courses on a wide range of topics including Python and intercultural management on Coursera and Udemy platforms. She interned at Campk12 last year for 3 months teaching students how to program and discovered that teaching is an interest of hers. She also worked as a Social Media Marketing at BornInFlight a talent transformation company. She is a Neuro Linguistic Programming Business Practitioner. Apart from these she has been a member of Rotary International from 2018 as a Rotaractor, a social organization/ NGO. Last year she was the Community Service Director of her Rotaract Club named Rotaract Club of New Age Engineers. This year she is the professional service director in her club. She has always loved speaking on stage and hence whenever she finds the opportunity she takes it. Rotary International and college have been great places for her to explore her interests and talents.

G. Darvika

Gogineni Darvika is currently pursuing her Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. She completed her intermediate from Narayana Junior College and 10th class from Narayana Concept School. Her technical skills include C and Python. She also has a basic understanding of C++, Java, JavaScript, HTML and CSS. She took part in Employability Skill development Program conducted by Zensar. She is also a student of Smart Interviews. Her participations include: showcasing the project model in National Level Project Expo and Competition "Technovation – 2018" which was conducted on 28th March, 2018 , National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020, "Machine Learning Workshop" conducted by Kryion Technologies Pvt. Ltd. on 14th & 15th February, 2020, "HTML & CSS Workshop" of TAM event held from 5th January, 2018 to 3rd February, 2018, "Machine Learning Workshop" conducted on 8th & 9th February, 2019, "Leadership Talk" conducted by MHRD's Innovation Cell on 16th May, 2020 at 01.00 PM. Her areas of interest are Web Development, Artificial Intelligence and Python. She completed 5 certification courses from Coursera and 5 certification courses CursaApp. She secured 2nd position in Codevita event conducted by TAM from 17th January, 2019 to 15rd February, 2019. She is also a recognized as Joint Community Service Director for the R.I. Year 2020-21. Her Internships include: Trainer for C, Python at "GCS Institute of Computer Technologies" from May 2019 to August 2019, Web Application Development at "MANAC Infotech (P) Limited" from 2nd May 2020 to 23rd August 2020.

Abhishek Thumar

Abhishek Harish Thumar is currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. He completed his intermediate from Narayana Junior College and 10th class from St. Andrews School. His technical skills include Python, C, C++, HTML5, CSS and basic understanding of Java mySQL. He took part in various football tournaments from state level to even national level tournaments and achieved various accolades. He is a member of a community organization called RAC New Age Engineers. His participations include: National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020, Digital Transformation in Education Sector Post- Covid era conducted by Collegedunia on 11th June 2021,Two-Day National Level Seminar on Recent Trends in Cloud Computing ,Fog and Edge Computing from 18th June 2021 to 19th June 2021.His area of interest include web development, python and ML. He also has completed different certification courses on online platforms from Coursera and CursaApp.

Mohith Bundey

Bundey Mohith is currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. He completed his intermediate from Sri Chaitanya Junior Kalashala and 10th class from Sri Chaitanya Techno School. His technical skills include Python ,C and he has a basic understanding of Java ,C++ and SQL .He took part in Employability Skill development Program conducted by Zensar. He is a member of a community organization called RAC New Age Engineers. His participations include: National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020 ,Digital Transformation in Education Sector Post- Covid era conducted by Collegedunia on 11th June 2021,Two-Day National Level Seminar on Recent Trends in Cloud Computing ,Fog and Edge Computing from 18th June 2021 to 19th June 2021.His area of interests include Python,3D modelling and Video editing .He has completed few certification courses on online platforms from Coursera, Udemy and CursaApp.

APPENDICES

```
from tkinter import messagebox

from tkinter import *

from tkinter import simpledialog

import tkinter

import matplotlib.pyplot as plt

import numpy as np

import pandas as pd

from tkinter import simpledialog

from tkinter import filedialog

import os

import cv2

import numpy as np

from keras.utils.np_utils import to_categorical

from keras.layers import  MaxPooling2D

from keras.layers import Dense, Dropout, Activation, Flatten, LSTM

from keras.layers import Convolution2D

from keras.models import Sequential

from keras.models import model_from_json

import pickle

main = tkinter.Tk()
```

```
main.title("Design of  Facial Expression Recognition System using Deep Learning Model Based on Human Emotions Through Classification with CNN & RNN algorithm") #designing main screen

main.geometry("1300x500")

global filename

global X, Y

global classifier

names = ['angry','disgusted','fearful','happy','neutral','sad','surprised']

def getID(name):
    index = 0
    for i in range(len(names)):
        if names[i] == name:
            index = i
            break
    return index

def upload():
    global filename
    filename = filedialog.askdirectory(initialdir=".")
    text.delete('1.0', END)
    text.insert(END,filename+" loaded\n")

def processDataset():
    text.delete('1.0', END)
```

```
global X, Y

'''

X = []

Y = []

for root, dirs, directory in os.walk(filename):

    for j in range(len(directory)):

        name = os.path.basename(root)

        print(name+" "+root+"/"+directory[j])

        if 'Thumbs.db' not in directory[j]:

            img = cv2.imread(root+"/"+directory[j])

            img = cv2.resize(img, (32,32))

            im2arr = np.array(img)

            im2arr = im2arr.reshape(32,32,3)

            X.append(im2arr)

            Y.append(getID(name))

X = np.asarray(X)

Y = np.asarray(Y)

print(Y)

X = X.astype('float32')

X = X/255
```

```
    test = X[3]
    test = cv2.resize(test,(400,400))
    cv2.imshow("aa",test)
    cv2.waitKey(0)
    indices = np.arange(X.shape[0])
    np.random.shuffle(indices)
    X = X[indices]
    Y = Y[indices]
    Y = to_categorical(Y)
    np.save('model/X.txt',X)
    np.save('model/Y.txt',Y)
    X = np.load('model/X.txt.npy')
    Y = np.load('model/Y.txt.npy')
    text.insert(END,"Total number of images found in dataset is : "+str(len(X))+"\n")
    text.insert(END,"Total classes found in dataset is : "+str(names)+"\n")
def trainCNN():
    global classifier
    text.delete('1.0', END)
    if os.path.exists('model/cnnmodel.json'):
        with open('model/cnnmodel.json', "r") as json_file:
```

```
        loaded_model_json = json_file.read()
        classifier = model_from_json(loaded_model_json)
    classifier.load_weights("model/cnnmodel_weights.h5")
    classifier._make_predict_function()
    print(classifier.summary())
    f = open('model/cnnhistory.pckl', 'rb')
    data = pickle.load(f)
    f.close()
    acc = data['accuracy']
    accuracy = acc[9] * 100
    text.insert(END,"CNN Training Model Accuracy = "+str(accuracy))
else:
    classifier = Sequential()
    classifier.add(Convolution2D(32, 3, 3, input_shape = (32, 32, 3), activation = 'relu'))
    classifier.add(MaxPooling2D(pool_size = (2, 2)))
    classifier.add(Convolution2D(32, 3, 3, activation = 'relu'))
    classifier.add(MaxPooling2D(pool_size = (2, 2)))
    classifier.add(Flatten())
    classifier.add(Dense(output_dim = 256, activation = 'relu'))
    classifier.add(Dense(output_dim = 7, activation = 'softmax'))
```

```
    print(classifier.summary())
    classifier.compile(optimizer = 'adam', loss = 'categorical_crossentropy', metrics = ['accuracy'])
    hist = classifier.fit(X, Y, batch_size=16, epochs=10, shuffle=True, verbose=2)
    classifier.save_weights('model/cnnmodel_weights.h5')
    model_json = classifier.to_json()
    with open("model/cnnmodel.json", "w") as json_file:
        json_file.write(model_json)
    f = open('model/cnnhistory.pckl', 'wb')
    pickle.dump(hist.history, f)
    f.close()
    f = open('model/cnnhistory.pckl', 'rb')
    data = pickle.load(f)
    f.close()
    acc = data['accuracy']
    accuracy = acc[9] * 100
    text.insert(END,"CNN Training Model Accuracy = "+str(accuracy))
def trainRNN():
    global X
    text.delete('1.0', END)
    if os.path.exists('model/rnnmodel.json'):
```

```
    with open('model/rnnmodel.json', "r") as jsonFile:
        loadedModelJson = jsonFile.read()
        lstm_model = model_from_json(loadedModelJson)
    lstm_model.load_weights("model/rnnmodel_weights.h5")
    lstm_model._make_predict_function()
    print(lstm_model.summary())
    f = open('model/rnnhistory.pckl', 'rb')
    data = pickle.load(f)
    f.close()
    acc = data['accuracy']
    accuracy = acc[9] * 100
    text.insert(END,"RNN Training Model Accuracy = "+str(accuracy))
else:
    X = np.reshape(X, (X.shape[0],X.shape[1],(X.shape[2]*X.shape[3])))
    print(X.shape)
    print(Y.shape)
    lstm_model = Sequential()
    lstm_model.add(LSTM(100, input_shape=(32, 96), activation='relu'))
    lstm_model.add(Dropout(0.5))
    lstm_model.add(Dense(100, activation='relu'))
```

```
    lstm_model.add(Dense(7, activation='softmax'))
    lstm_model.compile(loss='binary_crossentropy', optimizer='adam', metrics=['accuracy'])
    hist = lstm_model.fit(X, Y, batch_size=16, epochs=10, shuffle=True, verbose=2)
    lstm_model.save_weights('model/rnnmodel_weights.h5')
    model_json = lstm_model.to_json()
    with open("model/rnnmodel.json", "w") as json_file:
        json_file.write(model_json)
    f = open('model/rnnhistory.pckl', 'wb')
    pickle.dump(hist.history, f)
    f.close()
    f = open('model/rnnhistory.pckl', 'rb')
    data = pickle.load(f)
    f.close()
    acc = data['accuracy']
    accuracy = acc[9] * 100
    text.insert(END,"RNN Training Model Accuracy = "+str(accuracy))
def predict():
    filename = filedialog.askopenfilename(initialdir="testImages")
    image = cv2.imread(filename)
    img = cv2.resize(image, (32,32))
```

```
    im2arr = np.array(img)
    im2arr = im2arr.reshape(1,32,32,3)
    img = np.asarray(im2arr)
    img = img.astype('float32')
    img = img/255
    preds = classifier.predict(img)
    predict = np.argmax(preds)
    img = cv2.imread(filename)
    img = cv2.resize(img, (600,400))
    cv2.putText(img, 'Facial Expression Recognized as : '+names[predict], (10, 25),
                cv2.FONT_HERSHEY_SIMPLEX,0.7, (255, 0, 0), 2)
    cv2.imshow('Facial Expression Recognized as : '+names[predict], img)
    cv2.waitKey(0)
def graph():
    f = open('model/cnnhistory.pckl', 'rb')
    cnn_data = pickle.load(f)
    f.close()
    cnn_accuracy = cnn_data['accuracy']
    f = open('model/rnnhistory.pckl', 'rb')
    rnn_data = pickle.load(f)
    f.close()
```

```
    rnn_accuracy = rnn_data['accuracy']
    plt.figure(figsize=(10,6))
    plt.grid(True)
    plt.xlabel('Iterations/Epoch')
    plt.ylabel('Accuracy')
    plt.plot(cnn_accuracy, 'ro-', color = 'green')
    plt.plot(rnn_accuracy, 'ro-', color = 'orange')
    plt.legend(['CNN Accuracy', 'RNN Accuracy'], loc='upper left')
    #plt.xticks(wordloss.index)
    plt.title('CNN & RNN Accuracy Comparison Graph')
    plt.show()
def exit():
    main.destroy()
font = ('Helvetica', 13,'bold')
title = Label(main, text='Automatic Facial Expression Recognition using Features Extraction Based on Spatial &
                Temporal Sequences using CNN & RNN Algorithm')
title.config(bg='#3500D3', fg='#FFFFFF')
title.config(font=font)
title.config(height=3, width=130)
title.place(x=5,y=5)
font1 = ('Helvetica', 12,'bold')
```

```
text=Text(main,height=11.5,width=77)

scroll=Scrollbar(text)

text.configure(yscrollcommand=scroll.set)

text.place(x=320,y=100)

text.config(font=font1)

uploadButton = Button(main,bg='#3500D3', fg='#FFFFFF',text="Upload Facial Emotion Dataset", command=upload)

uploadButton.place(x=50,y=100)

uploadButton.config(font=font1)

processButton = Button(main,bg='#3500D3', fg='#FFFFFF', text="Preprocess Dataset", command=processDataset)

processButton.place(x=50,y=200)

processButton.config(font=font1)

cnnButton = Button(main,bg='#3500D3', fg='#FFFFFF', text="Train CNN Algorithm", command=trainCNN)

cnnButton.place(x=1050,y=100)

cnnButton.config(font=font1)

rnnButton = Button(main,bg='#3500D3', fg='#FFFFFF', text="Train RNN Algorithm", command=trainRNN)

rnnButton.place(x=1050,y=200)

rnnButton.config(font=font1)

graphButton = Button(main,bg='#3500D3', fg='#FFFFFF', text="Accuracy Comparison Graph", command=graph)

graphButton.place(x=50,y=300)
```

```
graphButton.config(font=font1)

predictButton = Button(main,bg='#3500D3', fg='#FFFFFF', text="Predict Facial Expression", command=predict)

predictButton.place(x=1050,y=300)

predictButton.config(font=font1)

exitButton = Button(main,bg='#3500D3', fg='#FFFFFF', text="Exit", command=exit)

exitButton.place(x=645,y=400)

exitButton.config(font=font1)

main.config(bg='#282828')

main.mainloop()
```

A

PROJECT REPORT

On

A MACHINE LEARNING MODEL FOR AVERAGE FUEL CONSUMPTION IN HEAVY VEHICLES

Submitted by

1) Ms. Yukthi G(17K81A05C0) **2) Ms. S J N V L Sai Shravani(17K81A05A8)**

3) Mr. Rohan Reddy(17K81A0580) **4) Mr.Aila Tanay Reddy (17K81A0563)**

in partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Under the Guidance of

Dr. R Santosh Kumar

Associate Professor (CSE)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ST.MARTIN'S ENGINEERING COLLEGE
An Autonomous Institute

Dhulapally, Secunderabad – 500 100

JUNE 2021

BONAFIDE CERTIFICATE

This is to certify that the project entitled A machine learning model for average fuel consumption in heavy vehicles, is being submitted by **1.Yukthi G (17K81A05C0), 2.S J N V L Sai Shravani (17K81A05A8), 3.Rohan Reddy (17K81A0580), 4.Aila Tanay Reddy (17K81A0563)** in partial fulfillment of the requirement for the award of the degree of **BACHELOR OF TECHNOLOGY IN COMPUTER SCIENCE AND ENGINEERING** is recorded of bonafide work carried out by them. The result embodied in this report have been verified and found satisfactory.

<Signature>

Dr.R SANTOSH KUMAR

Department of CSE

Head of the Department

Dr.M.NARAYANAN

Department of CSE

Internal Examiner

External Examiner

Place:
Date:

DECLARATION

We, the student of **Bachelor of Technology** in Department of Computer Science and Engineering', session: <2017 – 2021>, St. Martin's Engineering College, Dhulapally, Kompally, Secunderabad, hereby declare that work presented in this Project Work entitled A machine learning model for average fuel consumption in heavy vehicles is the outcome of our own bonafide work and is correct to the best of our knowledge and this work has been undertaken taking care of Engineering Ethics. This result embodied in this project report has not been submitted in any university for award of any degree.

Yukthi G	17K81A05C0
S J N V L Sai Shravani	17K81A05A8
Rohan Reddy	17K81A0580
Aila Tanay Reddy	17K81A0563

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible and whose encouragements and guidance have crowded effects with success.

We extended our deep sense of gratitude to Principal**,** **Dr. P. SANTOSH KUMAR PATRA**, St. Martin's Engineering College, Dhulapally, for permitting us to undertake this project.

We are also thankful to **Dr. M.NARAYANAN**, Head of the Department, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

We would like to thank **Dr. T. POONGOTHAI,** Department of Computer Science and Engineering (AI & ML) for her encouragement and insightful comments and as well as our project coordinators **Dr. B.RAJALINGAM**, Associate Professor and **Dr. G. GOVINDARAJULU,** Associate Professor, in Department of Computer Science and Engineering for their valuable support.

We would like to express our sincere gratitude and indebtedness to our project supervisor Dr. R Santosh Kumar, Associate Professor, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

Finally, we express thanks to all those who have helped us successfully to completing this project. Furthermore, we would like to thank our family and friends for their moral support and encouragement.

We express thanks to all those who have helped us in successfully completing the project.

Yukthi G	17K81A05C0
S J N V L Sai Shravani	17K81A05A8
Rohan Reddy	17K81A0580
Aila Tanay Reddy	17K81A0563

ABSTRACT

We used vehicle travel distance rather than the traditional time period when developing individualized machine learning models for fuel consumption. This approach is used in conjunction with seven predictors derived from vehicle speed and road grade to produce a highly predictive neural network model for average fuel consumption in heavy vehicles. The proposed model can easily be developed and deployed for each individual vehicle in a fleet in order to optimize fuel consumption over the entire fleet. The predictors of the model are aggregated over fixed window sizes of distance travelled. Different window sizes are evaluated and the results show that a 1 km window is able to predict fuel consumption with a 0.91 coefficient of determination and mean absolute peak-to-peak percent error less than 4% for routes that include both city and highway duty cycle segments.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF OUTPUT SCREENS

LIST OF ACRONYMS

<UML>	Unified Modeling Language
<AI>	Artificial Intelligence
<GB>	Giga Bytes
<RAM>	Random Access Memory
<MB>	Mega Bytes
<ANN>	Artificial Neural Network
<SVM>	Support Vector Machine

1. INTRODUCTION

1.1 PROJECT OVERVIEW

In this paper, a model that can be easily developed for individual heavy vehicles in a large fleet is proposed. Relying on accurate models of all of the vehicles in a fleet, a fleet manager can optimize the route planning for all of the vehicles based on each unique vehicle predicted fuel consumption thereby ensuring the route assignments are aligned to minimize overall fleet fuel consumption. These types of fleets exist in various sectors including, road transportation of goods, public transportation, construction trucks and refuse trucks. For each fleet, the methodology must apply and adapt to many different vehicle technologies (including future ones) and configurations without detailed knowledge of the vehicles specific physical characteristics and measurements. These requirements make machine learning the technique of choice when taking into consideration the desired accuracy versus the cost of the development and adaptation of an individualized model for each vehicle in the fleet.

1.2 PROJECT OBJECTIVES

Existing model that can be easily developed for individual heavy vehicles in a large fleet is proposed. Relying on accurate models of all of the vehicles in a fleet, a fleet manager can optimize the route planning for all of the vehicles based on each unique vehicle predicted fuel consumption thereby ensuring the route assignments are aligned to minimize overall fleet fuel consumption.This approach is used in conjunction with seven predictors derived from vehicle speed and road grade to produce a highly predictive neural network model for average fuel consumption in heavy vehicles.Different window sizes are evaluated and the results show that a 1 km window is able to predict fuel consumption with a 0.91 coefficient of determination and mean absolute peak-to-peak percent error less than 4% for routes that include both city and highway duty cycle segments.

1.3 ORGANIZATION OF CHAPTERS

This documentation consists of 10 different chapters and they are: Introduction covers the overview of our project and its objectives. Literature Survey includes the details of our survey. Software and Hardware Requirements specify our software and hardware requirements here. Software Development Analysis section includes the problem definition and details of the modules we used in our project. Project System Design includes the design part of our project which includes uml diagrams. Project Coding section contains the details of our project code. Project Testing includes details of test cases and testing. Output Screens contains the screenshots of how our project looks like when executed.

Experimental Results chapter contains the screenshots of our results. Conclusion and Future Enhancements covers the conclusion of our project and the possible future developments.

2. LITERATURE SURVEY

2.1 SURVEY ON BACKGROUND

[5] F. Perrotta, T. Parry, and L. C. Neves, "Application of machine learning for fuel consumption modelling of trucks," in Big Data (Big Data), 2017 IEEE International Conference on. IEEE, 2017, pp. 3810–3815.

This paper presents the application of three Machine Learning techniques to fuel consumption modelling of articulated trucks for a large dataset. In particular, Support Vector Machine (SVM), Random Forest (RF), and Artificial Neural Network (ANN) models have been developed for the purpose and their performance compared. Fleet managers use telematic data to monitor the performance of their fleets and take decisions regarding maintenance of the vehicles and training of their drivers. The data, which include fuel consumption, are collected by standard sensors (SAE J1939) for modern vehicles. Data regarding the characteristics of the road come from the Highways Agency Pavement Management System (HAPMS) of Highways England, the manager of the strategic road network in the UK. Together, these data can be used to develop a new fuel consumption model, which may help fleet managers in reviewing the existing vehicle routing decisions, based on road geometry. The model would also be useful for road managers to better understand the fuel consumption of road vehicles and the influence of road geometry. Ten-fold cross-validation has been performed to train the SVM, RF, and ANN models. Results of the study shows the feasibility of using telematic data together with the information in HAPMS for the purpose of modelling fuel consumption. The study also shows that although all the three methods make it possible to develop models with good precision, the RF slightly outperforms SVM and ANN giving higher R 2 , and lower error.

[2] Chang Liu, Jian Rong, Yunlong Zhang, "Vehicle Fuel Consumption Prediction Method Based on Driving Behavior Data Collected from Smartphones", Journal of Advanced Transportation, 2020.

Transportation is an important factor that affects energy consumption, and driving behavior is one of the main factors affecting vehicle fuel consumption. The purpose of this paper is to improve fuel consumption monitoring databases based on mobile phone data. Based on the mobile phone terminals and on-board diagnostic system (OBD) installed in taxis, driving behavior data and fuel consumption data are extracted, respectively. By matching the driving behavior data collected by a mobile phone with the fuel consumption data collected by OBD, the correlation between driving behavior and fuel consumption is explored, so that vehicle fuel consumption could be predicted based on mobile phone data. The fuel consumption prediction models are built using back propagation (BP) neural network,

support vector regression (SVR), and random forests. The results show that the average speed, average speed except for idle (ASEI), average acceleration, average deceleration, acceleration time percentage, deceleration time percentage, and cruising time percentage are important indicators for fuel consumption evaluation. All three models could predict fuel consumption accurately, with an absolute relative error less than 10%. The random forest model is proved to have the highest accuracy and runs faster, making it suitable for wide application. This method lays a foundation for monitoring database improvement and fine management of urban transportation fuel consumption.

[6] S. Wickramanayake and H. D. Bandara, "Fuel consumption prediction of fleet vehicles using machine learning: A comparative study," in Moratuwa Engineering Research Conference (MERCon), 2016. IEEE, 2016, pp. 90–95.

Ability to model and predict the fuel consumption is vital in enhancing fuel economy of vehicles and preventing fraudulent activities in fleet management. Fuel consumption of a vehicle depends on several internal factors such as distance, load, vehicle characteristics, and driver behavior, as well as external factors such as road conditions, traffic, and weather. However, not all these factors may be measured or available for the fuel consumption analysis. We consider a case where only a subset of the aforementioned factors is available as a multi-variate time series from a long distance, public bus. Hence, the challenge is to model and/or predict the fuel consumption only with the available data, while still indirectly capturing as much as influences from other internal and external factors. Machine Learning (ML) is suitable in such analysis, as the model can be developed by learning the patterns in data. In this paper, we compare the predictive ability of three ML techniques in predicting the fuel consumption of the bus, given all available parameters as a time series. Based on the analysis, it can be concluded that the random forest technique produces a more accurate prediction compared to both the gradient boosting and neural networks.

[4] Qi Zhao, Qi Phen, Liwang, "Real-Time Prediction of Fuel Consumption Based on Digital Map API", Appl. Sci. 2019, 9, 1369.

At present, digital maps can estimate the travel time of each trip's route but cannot offer a fuel consumption estimation at the same time. In this paper, we develop a fuel consumption model based on the Vehicle Specific Power (VSP) distribution, which can connect the traffic condition prediction with the fuel consumption model to predict fuel consumption. First, the traffic condition forecasting and the trip time of each route can be obtained through the digital map Application Programming Interface (API). Secondly, the users need to provide the engine displacement of their vehicles to match the fuel consumption model. Then, the fuel consumption prediction application based on Android is developed

to forecast the fuel consumption by using traffic prediction data. Finally, the fuel consumption provided by the On-Board Diagnostic (OBD) data is used to verify the proposed application, and the forecasting error is less than 20%.

[15] H. Almer, "Machine learning and statistical analysis in fuel consumption ´prediction for heavy vehicles," 2015.

This study evaluates methods of machine learning (ml) and statistical analysis for predicting fuel consumption in heavy vehicles. The idea is to use historical data describing driving situations to predict a fuel consumption in liters per distance. The general problem description is to examine a large number of attributes describing a fuel consumption situation and to employ ml methods to find a regression from such attributes to a fuel consumption. Attributes included could be environmental conditions, vehicle configuration, driver behavior and weather conditions. Research has been made into how to do such predictions for aircraft, engines and passenger cars as well as heavy vehicles. The previous research makes suggestions about which ml methods are most successful in fuel consumption prediction as well as what kind of attributes are most influential in fuel consumption for road vehicles

[11] J. Zhao, W. Li, J. Wang, and X. Ban, "Dynamic traffic signal timing optimization strategy incorporating various vehicle fuel consumption characteristics," IEEE Transactions on Vehicular Technology, vol. 65,no. 6, pp. 3874–3887, June 2016.

This paper proposes a dynamic traffic signal timing optimization strategy (DTSTOS) based on various vehicle fuel consumption and dynamic characteristics to minimize the combined total energy consumption and traffic delay for vehicles passing through an intersection. With increasing penetration of new vehicle types and configurations, vehicle fuel consumption characteristics have become rather diversified and dynamic and need to be explicitly incorporated in the traffic light timing control to reduce total energy consumption and traffic delay. Through vehicle-to-infrastructure (V2I) communications, information and states of individual vehicles around an intersection can be made available to the traffic light controller to produce optimal traffic light timing. Unified and control-oriented speed-type fuel consumption models for various types of vehicles in conjunction with a simplified traffic model are employed to conduct real-time traffic light timing control optimization using an iterative grid search (IGS) method. The effectiveness of the DTSTOS was evaluated and demonstrated with a traffic simulator in VISSIM with various traffic flows and vehicle types. The proposed timing plan was compared with Synchro, and consistent results were obtained.

[14] W. Zeng, T. Miwa, and T. Morikawa, "Exploring trip fuel consumptionby machine learning from gps and can bus data," Journal of the EasternAsia Society for Transportation Studies, vol. 11, pp. 906–921, 2015.

This study aims to explore the trip fuel consumption from a large-scale dataset. To better understand how the multiple variables (e.g., average travel speed, trip distance) influence the trip fuel consumption, we propose the support vector machine (SVM) to learn the relationship between the trip fuel consumption and the corresponding factors. A large-scale GPS and CAN (Controller Area Network) bus data provided by 153 probe vehicles during one month are used. Elasticity analysis indicates that trip distance and coefficient of variance of link speed have relatively great importance on the SVM model. To demonstrate the performance of the proposed method, three other regression methods, i.e., the multiple linear regression model, artificial neural network (ANN), and the link fuel summation SVM model (LSSVM) are also adopted for performance comparisons. The results show that SVM model is much closer to the target than the other three models.

[7] A. A. Zaidi, B. Kulcsr, and H. Wymeersch, "Back-pressure traffic signal control with fixed and adaptive routing for urban vehicular networks," IEEE Transactions on Intelligent Transportation Systems, vol. 17, no. 8, pp. 2134–2143, Aug 2016.

City-wide control and coordination of traffic flow can improve efficiency, fuel consumption, and safety. We consider the problem of controlling traffic lights under fixed and adaptive routing of vehicles in urban road networks. Multicommodity back-pressure algorithms, originally developed for routing and scheduling in communication networks, are applied to road networks to control traffic lights and adaptively reroute vehicles. The performance of the algorithms is analyzed using a microscopic traffic simulator. The results demonstrate that the proposed multicommodity and adaptive routing algorithms provide significant improvement over a fixed schedule controller and a single-commodity back-pressure controller in terms of various performance metrics, including queue length, trips completed, travel times, and fair traffic distribution.

2.2 CONCLUSION ON SURVEY

After referring to various papers, we were able to tackle most of our obstacles. Our paper presents a machine learning model for each heavy vehicle in a fleet that can be easily built. The last two predictors i.e., change in kinetic energy and change in potential energy are added in this paper to aid in capturing the vehicle's average dynamic activity. The model's predictors are all based on vehicle speed

and road grade. Telematics systems, which are becoming an increasingly important part of connected cars, provide easy access to these variables. Furthermore, from these two variables, the predictors can be conveniently computed on-board. The cost of the model in terms of data collection and on-board computation should be considered when choosing an appropriate window size. Furthermore, the window size is likely to vary depending on the application. A 1 km window size is recommended for fleets with short trips (e.g., construction vehicles inside a site) or urban traffic routes. A 5km window size can be suitable for long-haul fleets. Since the service cycles in this study included both highway and city traffic, the 1 km window was more appropriate than the 5 km window.

3. SOFTWARE AND HARDWARE REQUIREMENTS

3.1 Software Requirements

- **Operating system** **:** Windows 7, Windows XP, Windows 8.
- **Platform** **:** Python Technology
- **Coding Language** **:** Python

3.2 Hardware Requirements

- **System** **:** INTEL i3
- **Hard Disk** **:** 512 GB
- **Ram** **:** 4 GB.

4. SOFTWARE DEVELOPMENT ANALYASIS

4.1 Overview of problem

In this paper, a model that can be easily developed for individual heavy vehicles in a large fleet is proposed. Relying on accurate models of all of the vehicles in a fleet, a fleet manager can optimize the route planning for all of the vehicles based on each unique vehicle predicted fuel consumption thereby ensuring the route assignments are aligned to minimize overall fleet fuel consumption. These types of fleets exist in various sectors including, road transportation of goods, public transportation, construction trucks and refuse trucks. For each fleet, the methodology must apply and adapt to many different vehicle technologies (including future ones) and configurations without detailed knowledge of the vehicles specific physical characteristics and measurements. These requirements make machine learning the technique of choice when taking into consideration the desired accuracy versus the cost of the development and adaptation of an individualized model for each vehicle in the fleet.

4.2 Define the problem

Fuel consumption models for vehicles are of interest to manufacturers, regulators, and consumers. They are needed across all the phases of the vehicle life-cycle. In this paper, we focus on modeling average fuel consumption for heavy vehicles during the operation and maintenance phase. In general, techniques used to develop models for fuel consumption fall under three main categories: Physics-based models, which are derived from an in-depth understanding of the physical system. These models describe the dynamics of the components of the vehicle at each time step using detailed mathematical equations. Machine learning models, which are data-driven and represent an abstract mapping from an input space consisting of a selected set of predictors to an output space that represents the target output, in this case average fuel consumption. Statistical models, which are also data-driven and establish a mapping between the probability distribution of a selected set of predictors and the target outcome.

4.3 Modules Overview

To predict fuel consumption author has extracted 7 predictor features from heavy vehicle dataset such as number of stops, time stopped, average moving speed, characteristic acceleration, aerodynamic speed squared, change in kinetic energy and change in potential energy. Above seven features are recorded from each vehicle travel up to 100 kilo meters

like number of times vehicle stop, total stopped time taken etc. All this values are collected from heavy vehicle and use as dataset to train ANN model

4.4 Define the modules

This project mainly consists of 5 modules. They are: -

- Upload Heavy Vehicles Fuel Dataset
- Read Dataset & Generate Model
- Run ANN Algorithm
- Predict Average Fuel Consumption
- Fuel Consumption Graph

4.5 Module functionality

1. Upload Heavy Vehicles Fuel Dataset: -

Using this module we can upload train dataset to application. Dataset contains comma separated values.

2. Read Dataset and Generate Model: -

Using this module we will parse comma separated dataset and the generate train and test model for ANN from that dataset values. Dataset will be divided into 80% and 20% format. *0% will be used to train ANN model and 20% will be used to test ANN model.

3. Run ANN Algorithm: -

Using this model we can create ANN object and then feed train and test data to build ANN model.

4. Predict Average Fuel Consumption: -

Using this model we will upload new test data and then ANN will apply train model on that test data to predict average fuel consumptions for that test records.

5. Fuel Consumption Graph: -

Using this module we will plot fuel consumption graph for each test record.

5. PROJECT SYSTEM DESIGN

5.1 Architecture Diagram

An architectural diagram is a diagram of a system that is used to abstract the overall outline of the software system and the relationships, constraints, and boundaries between components. It is an important tool as it provides an overall view of the physical deployment of the software system and its evolution roadmap.

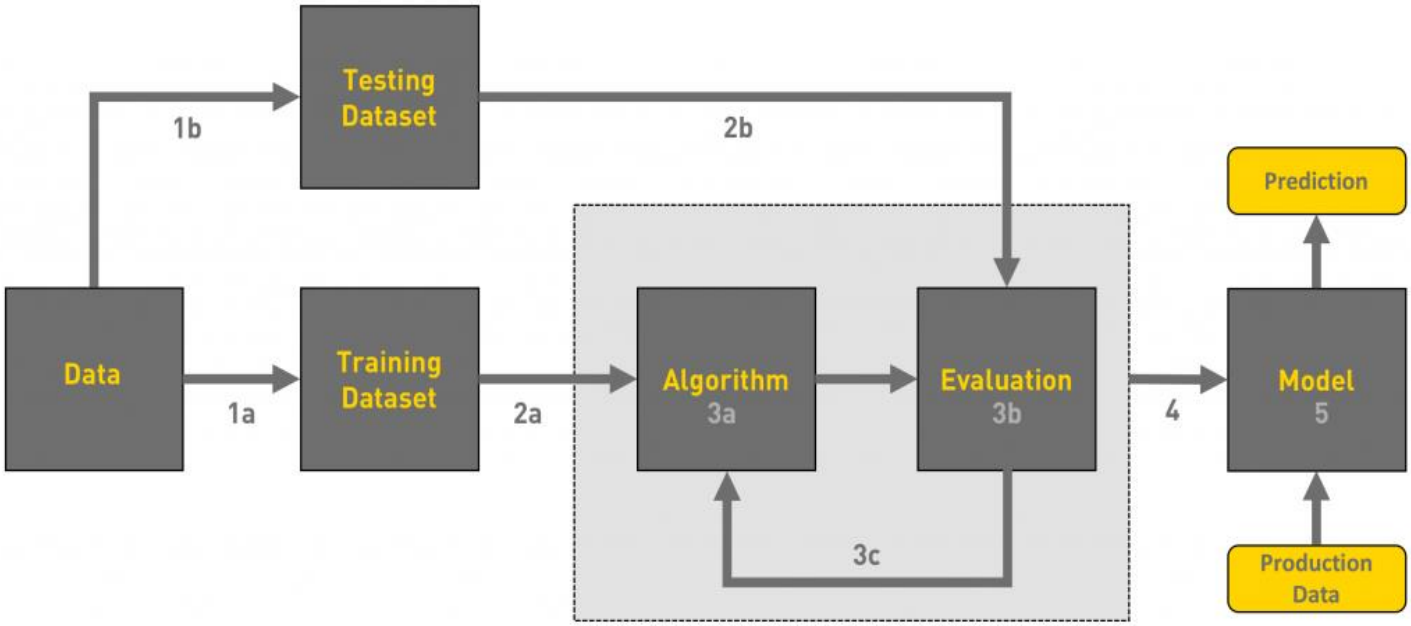


Figure 5.1

5.2 Data Flow Diagrams

A data flow diagram (DFD) maps out the flow of information for any process or system. It uses defined symbols like rectangles, circles and arrows, plus short text labels, to show data inputs, outputs, storage points and the routes between each destination.

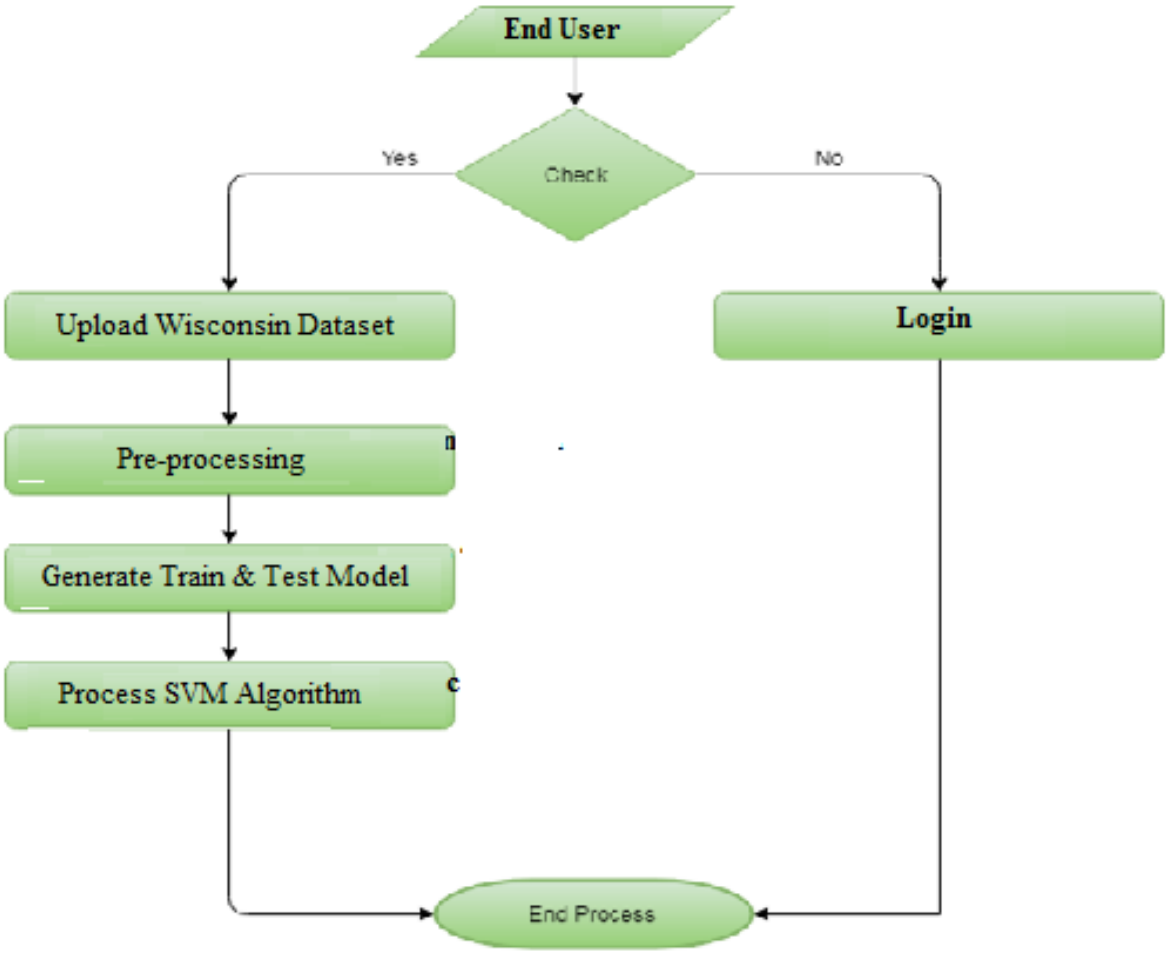


Figure 5.2

5.3 E-R Diagrams

An Entity Relationship (ER) Diagram is a type of flowchart that illustrates how “entities” such as people, objects or concepts relate to each other within a system. ER Diagrams are most often used to design or debug relational databases in the fields of software engineering, business information systems, education and research.

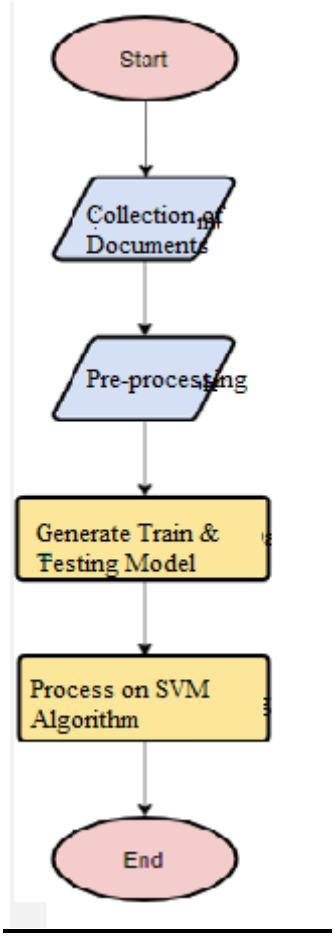


Figure 5.3

5.4 UML DIAGRAMS

UML stands for Unified Modeling Language. UML is a standardized general-purpose modeling language in the field of object-oriented software engineering. The standard is managed, and was created by, the Object Management Group.

The goal is for UML to become a common language for creating models of object-oriented computer software. In its current form UML is comprised of two major components: a Meta-model and a notation. In the future, some form of method or process may also be added to; or associated with, UML.

The Unified Modeling Language is a standard language for specifying, Visualization, Constructing and documenting the artifacts of software system, as well as for business modeling and other non-software systems.

The UML represents a collection of best engineering practices that have proven successful in the modeling of large and complex systems.

The UML is a very important part of developing objects-oriented software and the software development process. The UML uses mostly graphical notations to express the design of software projects.

GOALS:

The Primary goals in the design of the UML are as follows:

1. Provide users a ready-to-use, expressive visual modeling Language so that they can develop and exchange meaningful models.
2. Provide extendibility and specialization mechanisms to extend the core concepts.
3. Be independent of particular programming languages and development process.
4. Provide a formal basis for understanding the modeling language.
5. Encourage the growth of OO tools market.
6. Support higher level development concepts such as collaborations, frameworks, patterns and components.
7. Integrate best practices.

USE CASE DIAGRAM

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.

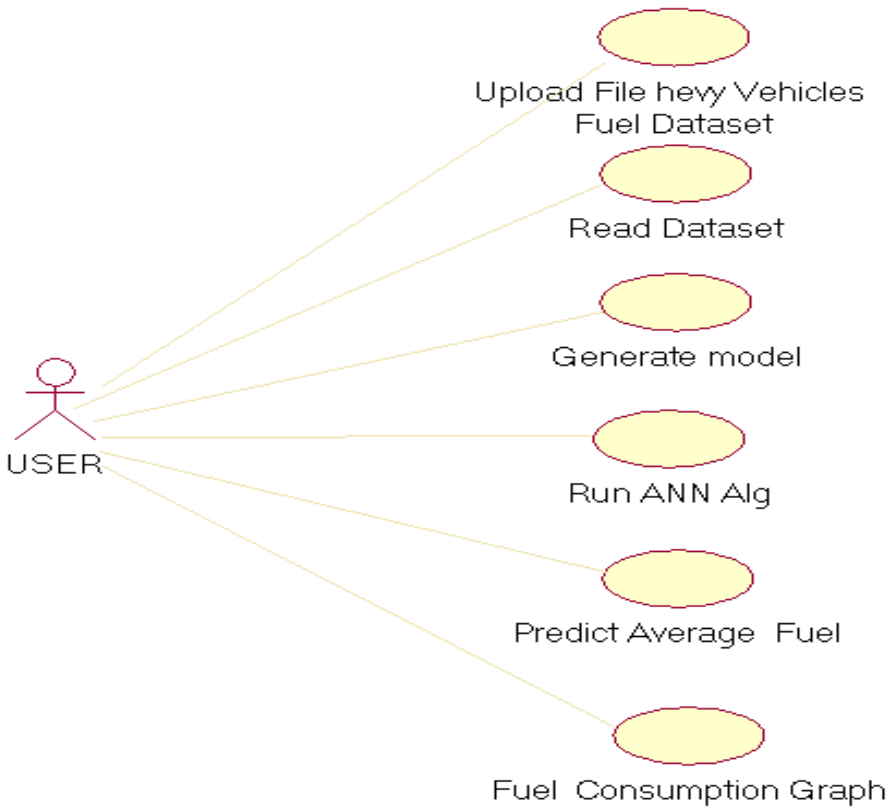


Figure 5.4

CLASS DIAGRAM

In software engineering, a class diagram in the Unified Modelling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.

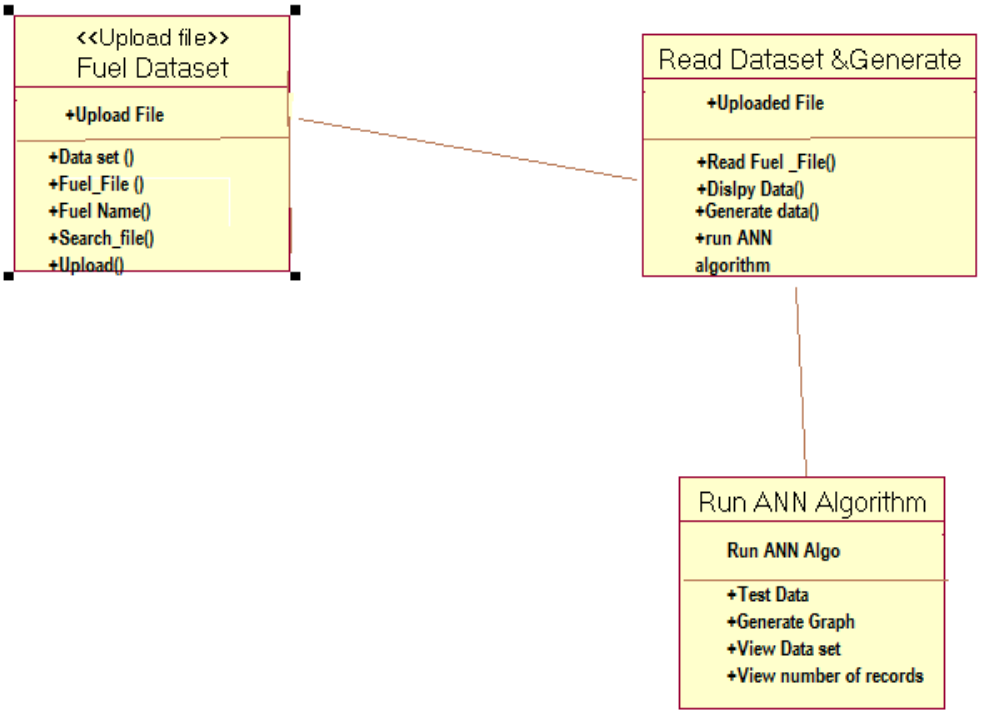


Figure 5.5

SEQUENCE DIAGRAM

A sequence diagram in Unified Modelling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.

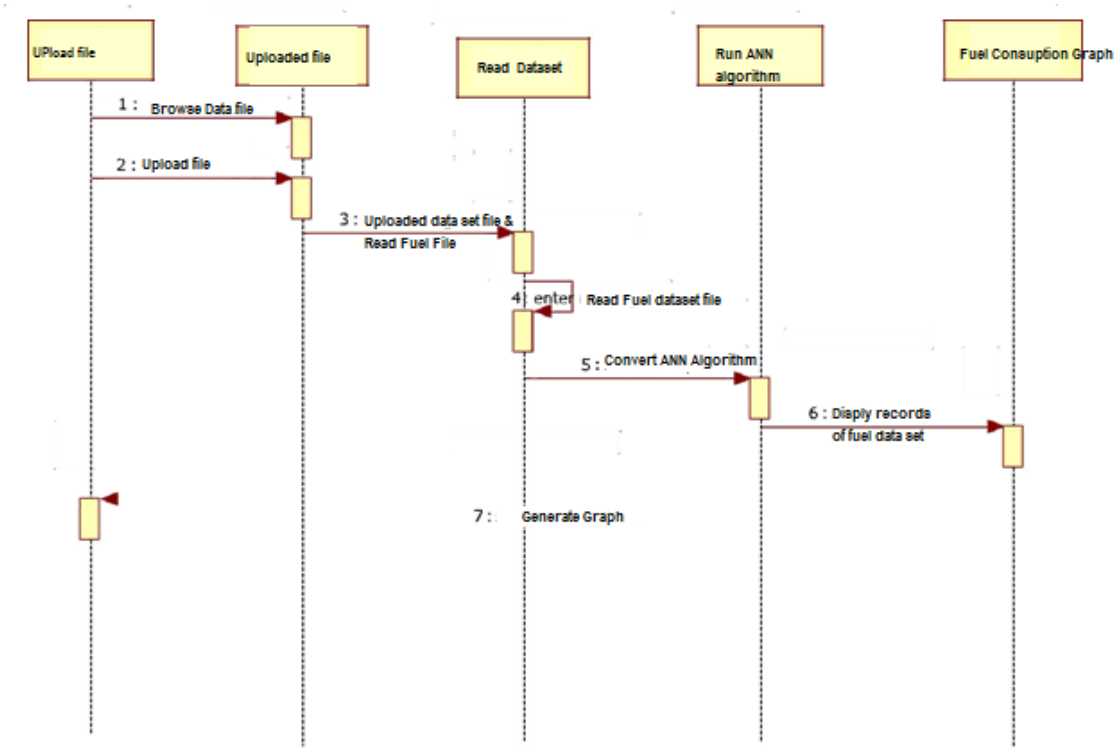


Figure 5.6

ACTIVITY DIAGRAM

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modelling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.

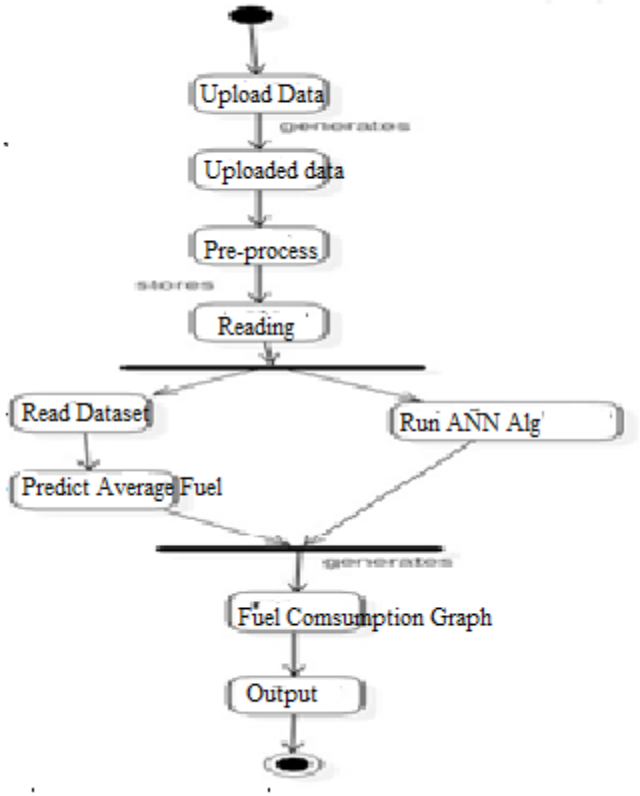


Figure 5.7

COMPONENT DIAGRAM

In Unified Modelling Language (UML), a component diagram depicts how components are wired together to form larger components or software systems. They are used to illustrate the structure of arbitrarily complex systems.

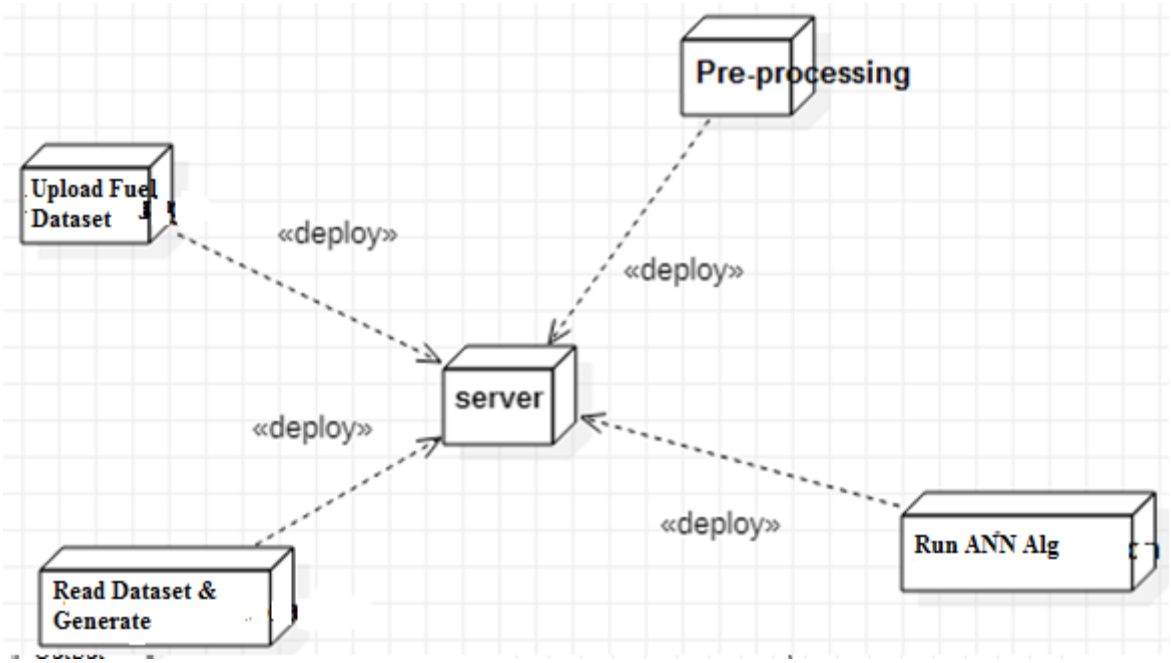


Figure 5.8

6. PROJECT CODING

6.1 CODE TEMPLATES

```
font = ('times', 16, 'bold')
title = Label(main, text='A Machine Learning Model for Average Fuel Consumption in Heavy Vehicles')
title.config(bg='greenyellow', fg='dodger blue')
title.config(font=font)
title.config(height=3, width=120)
title.place(x=0,y=5)

font1 = ('times', 12, 'bold')
text=Text(main,height=20,width=150)
scroll=Scrollbar(text)
text.configure(yscrollcommand=scroll.set)
text.place(x=50,y=120)
text.config(font=font1)



font1 = ('times', 14, 'bold')
uploadButton = Button(main, text="Upload Heavy Vehicles Fuel Dataset", command=upload)
uploadButton.place(x=50,y=550)
uploadButton.config(font=font1)

modelButton = Button(main, text="Read Dataset & Generate Model", command=generateModel)
modelButton.place(x=420,y=550)
modelButton.config(font=font1)

annButton = Button(main, text="Run ANN Algorithm", command=ann)
annButton.place(x=760,y=550)
annButton.config(font=font1)
```

```
predictButton = Button(main, text="Predict Average Fuel Consumption", command=predictFuel)
predictButton.place(x=50,y=600)
predictButton.config(font=font1)

graphButton = Button(main, text="Fuel Consumption Graph", command=graph)
graphButton.place(x=420,y=600)
graphButton.config(font=font1)

exitButton = Button(main, text="Exit", command=exit)
exitButton.place(x=760,y=600)
exitButton.config(font=font1)
```

6.2 OUTLINE FOR VARIOUS FILES

Tensorflow

It is an open source artificial intelligence library, using data flow graphs to build models. It allows developers to create large-scale neural networks with many layers. TensorFlow is mainly used for: Classification, Perception, Understanding, Discovering, Prediction and Creation.

Numpy

Numpy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays.

It is the fundamental package for scientific computing with Python. It contains various features including these important ones:

- A powerful N-dimensional array object
- Sophisticated (broadcasting) functions
- Tools for integrating C/C++ and Fortran code
- Useful linear algebra, Fourier transform, and random number capabilities

Besides its obvious scientific uses, Numpy can also be used as an efficient multi-dimensional container of generic data. Arbitrary data-types can be defined using Numpy which allows Numpy to seamlessly and speedily integrate with a wide variety of databases.

Pandas

Pandas is an open-source Python Library providing high-performance data manipulation and analysis tool using its powerful data structures. Python was majorly used for data munging and preparation. It had very little contribution towards data analysis. Pandas solved this problem. Using Pandas, we can accomplish five typical steps in the processing and analysis of data, regardless of the origin of data load, prepare, manipulate, model, and analyze. Python with Pandas is used in a wide range of fields including academic and commercial domains including finance, economics, Statistics, analytics, etc.

Matplotlib

Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms. Matplotlib can be used in Python scripts, the Python and IPython shells, the Jupyter Notebook, web application servers, and four graphical user interface toolkits. Matplotlib tries to make easy things easy and hard things possible. You can generate plots, histograms, power spectra, bar charts, error charts, scatter plots, etc., with just a few lines of code. For examples, see the sample plots and thumbnail gallery.

For simple plotting the pyplot module provides a MATLAB-like interface, particularly when combined with IPython. For the power user, you have full control of line styles, font properties, axes properties, etc, via an object oriented interface or via a set of functions familiar to MATLAB users.

Scikit – learn

Scikit-learn provides a range of supervised and unsupervised learning algorithms via a consistent interface in Python. It is licensed under a permissive simplified BSD license and is distributed under many Linux distributions, encouraging academic and commercial use. The library is built upon the SciPy (Scientific Python) that must be installed before you can use scikit-learn. This stack that includes:

- **NumPy**: Base n-dimensional array package
- **SciPy**: Fundamental library for scientific computing
- **Matplotlib**: Comprehensive 2D/3D plotting

- **IPython**: Enhanced interactive console
- **Sympy**: Symbolic mathematics
- **Pandas**: Data structures and analysis

Extensions or modules for SciPy care conventionally named SciKits

6.3 CLASS WITH FUNCTIONALITY

```
def importdata():
    global balance_data
    balance_data = pd.read_csv(filename)
    balance_data = balance_data.abs()
    return balance_data


def splitdataset(balance_data):
    global train_x, test_x, train_y, test_y
    X = balance_data.values[:, 0:7]
    y_ = balance_data.values[:, 7]
    print(y_)
    y_ = y_.reshape(-1, 1)
    encoder = OneHotEncoder(sparse=False)
    Y = encoder.fit_transform(y_)
    print(Y)
    train_x, test_x, train_y, test_y = train_test_split(X, Y, test_size=0.2)
    text.insert(END,"Dataset Length : "+str(len(X))+"\n");
    return train_x, test_x, train_y, test_y


def upload():
    global filename
```

```
    filename = filedialog.askopenfilename(initialdir="dataset")
    text.delete('1.0', END)
    text.insert(END,filename+" loaded\n\n");

def generateModel():
    global train_x, test_x, train_y, test_y
    data = importdata()
    train_x, test_x, train_y, test_y = splitdataset(data)
    text.insert(END,"Splitted Training Length : "+str(len(train_x))+"\n");
    text.insert(END,"Splitted Test Length : "+str(len(test_x))+"\n");

def ann():
    global model
    global ann_acc
    model = Sequential()
    model.add(Dense(200, input_shape=(7,), activation='relu', name='fc1'))
    model.add(Dense(200, activation='relu', name='fc2'))
    model.add(Dense(19, activation='softmax', name='output'))
    optimizer = Adam(lr=0.001)
    model.compile(optimizer, loss='categorical_crossentropy', metrics=['accuracy'])
    print('CNN Neural Network Model Summary: ')
    print(model.summary())
    model.fit(train_x, train_y, verbose=2, batch_size=5, epochs=200)
    results = model.evaluate(test_x, test_y)
    text.insert(END,"ANN Accuracy for dataset "+filename+"\n");
    text.insert(END,"Accuracy Score : "+str(results[1]*100)+"\n\n")
```

```
    ann_acc = results[1] * 100


def predictFuel():
    global testdata
    global predictdata
    text.delete('1.0', END)
    filename = filedialog.askopenfilename(initialdir="dataset")
    testdata = pd.read_csv(filename)
    testdata = testdata.values[:, 0:7]
    predictdata = model.predict_classes(testdata)
    print(predictdata)
    for i in range(len(testdata)):
        text.insert(END,str(testdata[i])+" Average Fuel Consumption : "+str(predictdata[i])+"\n");



def graph():
    x = []
    y = []
    for i in range(len(testdata)):
        x.append(i)
        y.append(predictdata[i])
    plt.plot(x, y)
    plt.xlabel('Vehicle ID')
    plt.ylabel('Fuel Consumption/10KM')
    plt.title('Average Fuel Consumption Graph')
    plt.show()
```

6.4 METHODS INPUT AND OUTPUT PARAMETERS

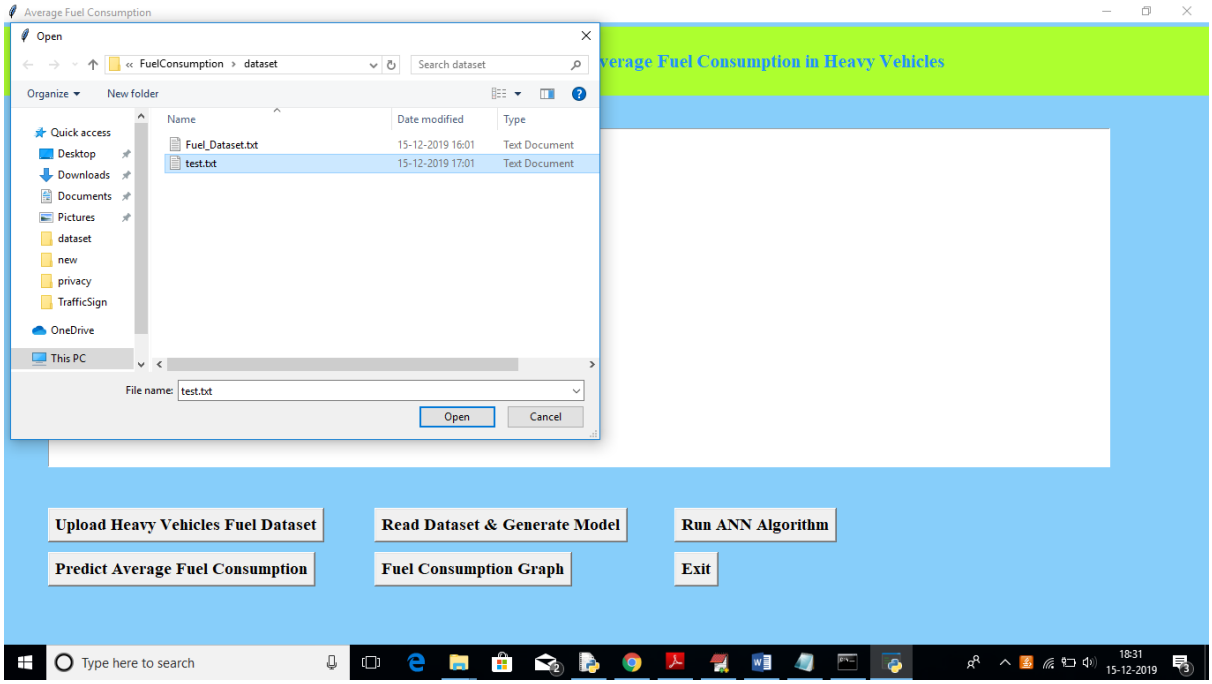


Figure 6.1

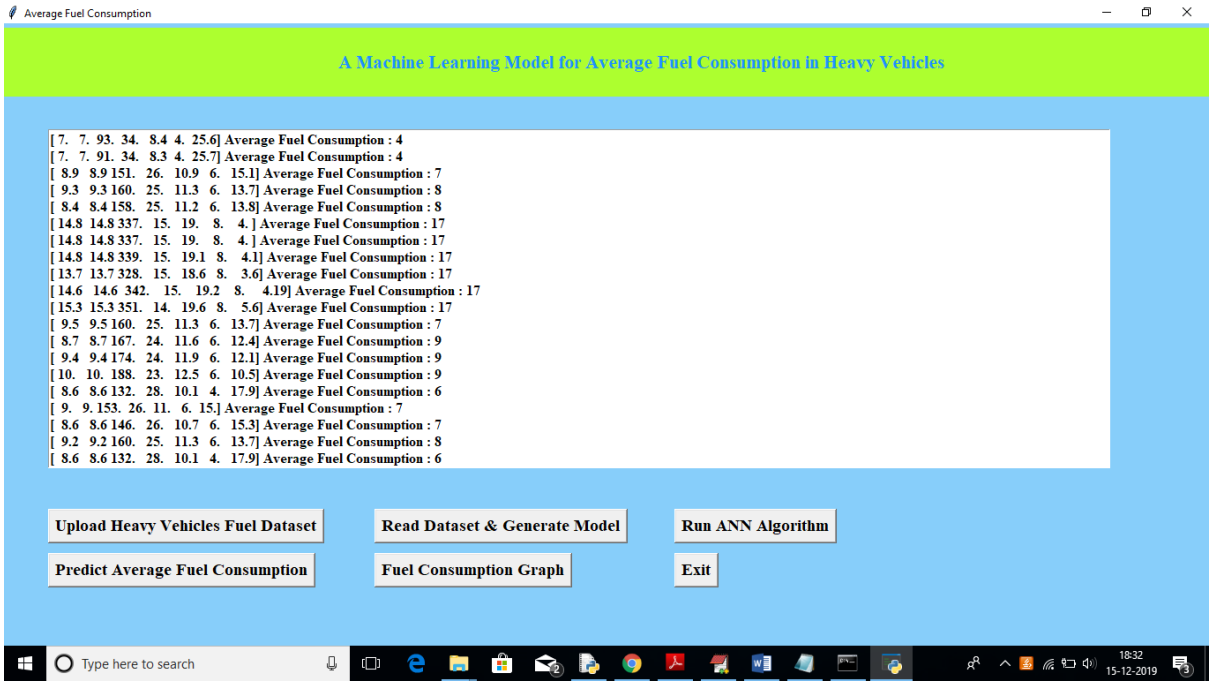


Figure 6.2

7. PROJECT TESTING

7.1 VARIOUS TEST CASES

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub-assemblies, assemblies and/or a finished product. It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of tests. Each test type addresses a specific testing requirement.

TYPES OF TESTS

Unit testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

Integration testing

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

Functional test

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centred on the following items:

Valid Input	: identified classes of valid input must be accepted.
Invalid Input	: identified classes of invalid input must be rejected.
Functions	: identified functions must be exercised.
Output	: identified classes of application outputs must be exercised.
Systems/Procedures	: interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

System Test

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration-oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

Unit Testing

Unit testing is usually conducted as part of a combined code and unit test phase of the software lifecycle, although it is not uncommon for coding and unit testing to be conducted as two distinct phases.

Test strategy and approach

Field testing will be performed manually and functional tests will be written in detail.

Test objectives

- All field entries must work properly.
- Pages must be activated from the identified link.
- The entry screen, messages and responses must not be delayed.

Features to be tested

- Verify that the entries are of the correct format
- No duplicate entries should be allowed
- All links should take the user to the correct page.

Integration Testing

Software integration testing is the incremental integration testing of two or more integrated software components on a single platform to produce failures caused by interface defects.

The task of the integration test is to check that components or software applications, e.g. components in a software system or – one step up – software applications at the company level – interact without error.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

Acceptance Testing

User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

7.2 BLACK BOX TESTING

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box .you cannot “see” into it. The test provides inputs and responds to outputs without considering how the software works.

7.3 WHITE BOX TESTING

White Box Testing is a testing in which in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is purpose. It is used to test areas that cannot be reached from a black box level.

OUTPUT SCREENS

8.1 User Interface

Figure 8.1

8.2 Output Screens

In above screen click on ‘Upload Heavy Vehicles Fuel Dataset’ button to upload train dataset.

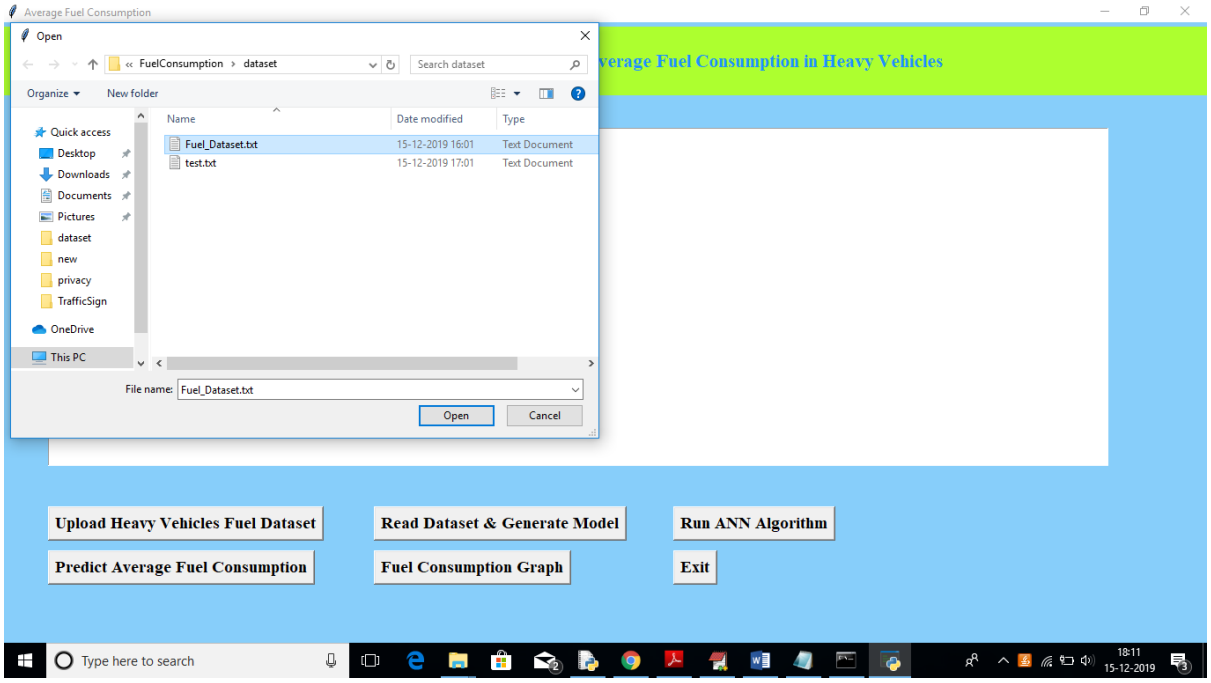


Figure 8.2

In above screen uploading ‘Fuel_Dataset.txt’ which can be used to train model. After uploading dataset will get below screen

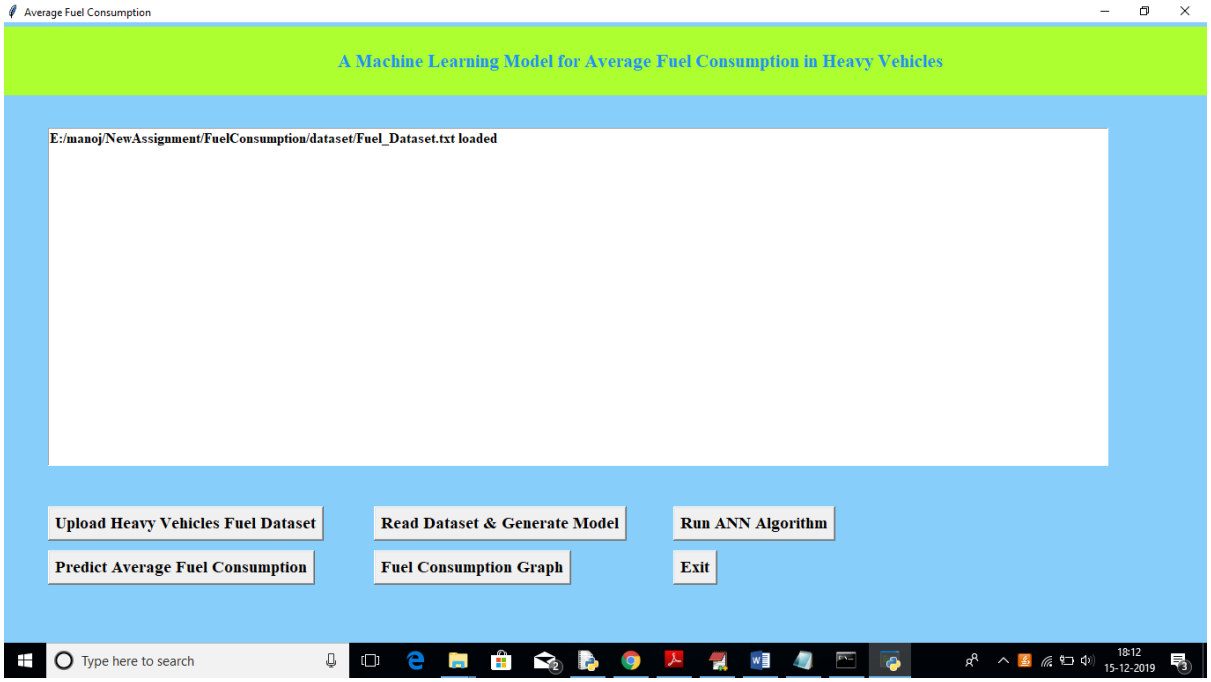


Figure 8.3

Now in above screen click on ‘Read Dataset & Generate Model’ button to read uploaded dataset and to generate train and test data

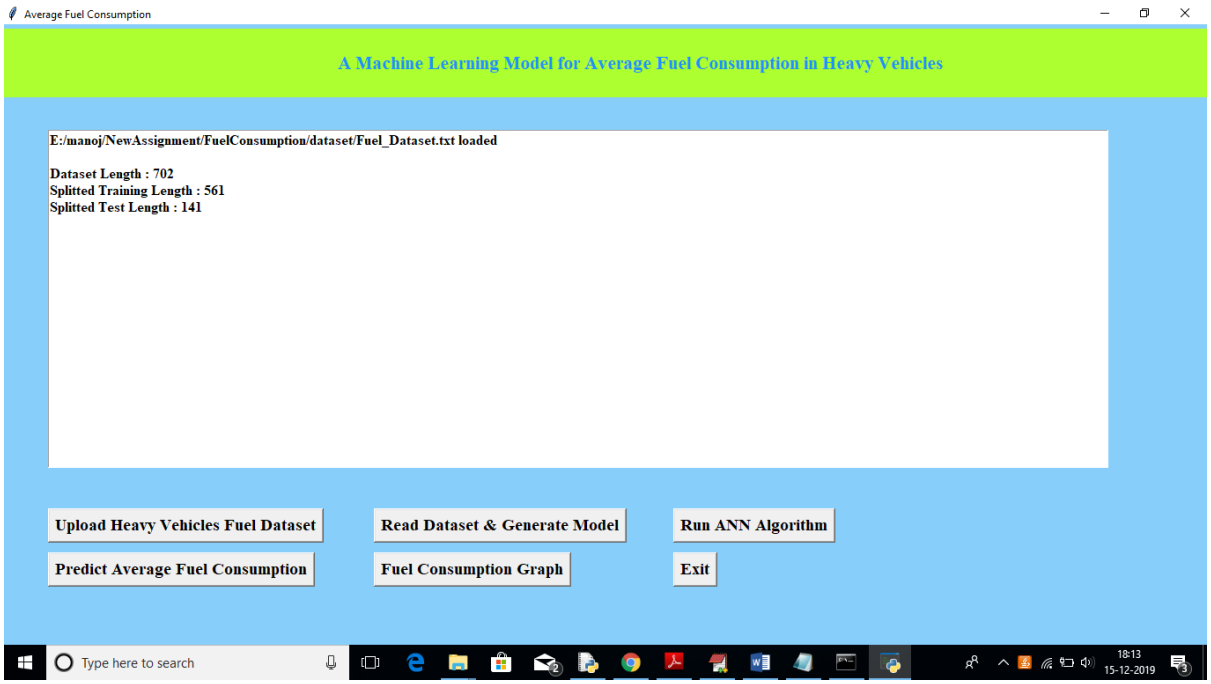


Figure 8.4

In above screen we can see total number of records in dataset, number of records used for training and number for records used for testing. Now click on 'Run ANN Algorithm' button to input train and test data to ANN to build ANN model.

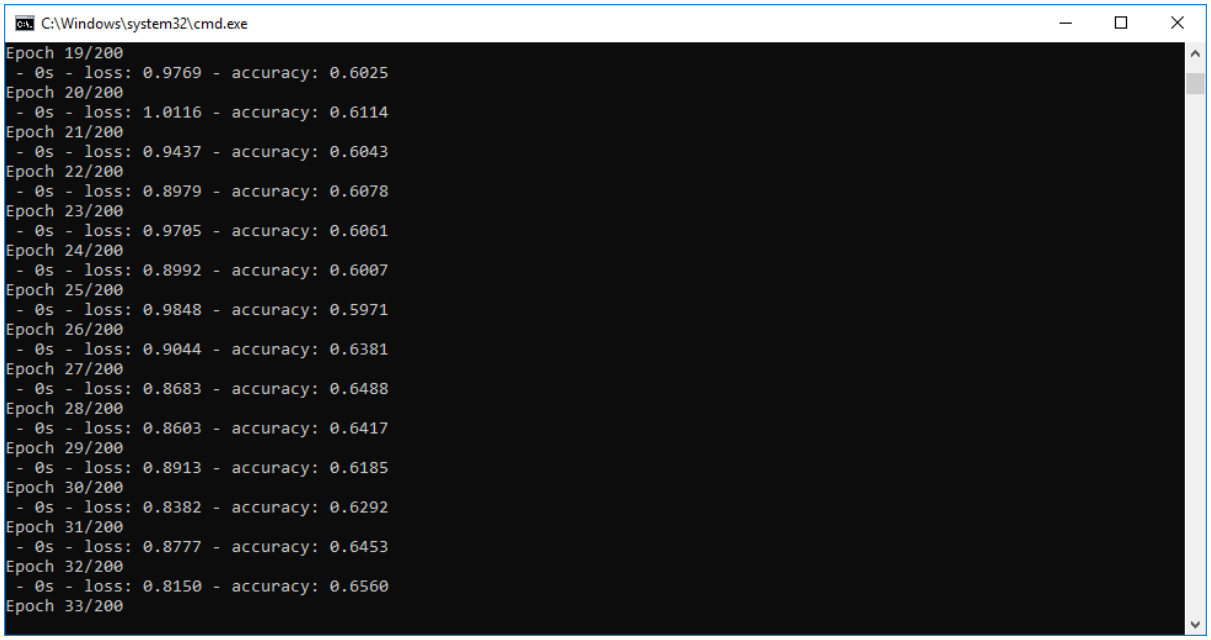


Figure 8.5

In above black console we can see all ANN processing details, After building model will get below screen

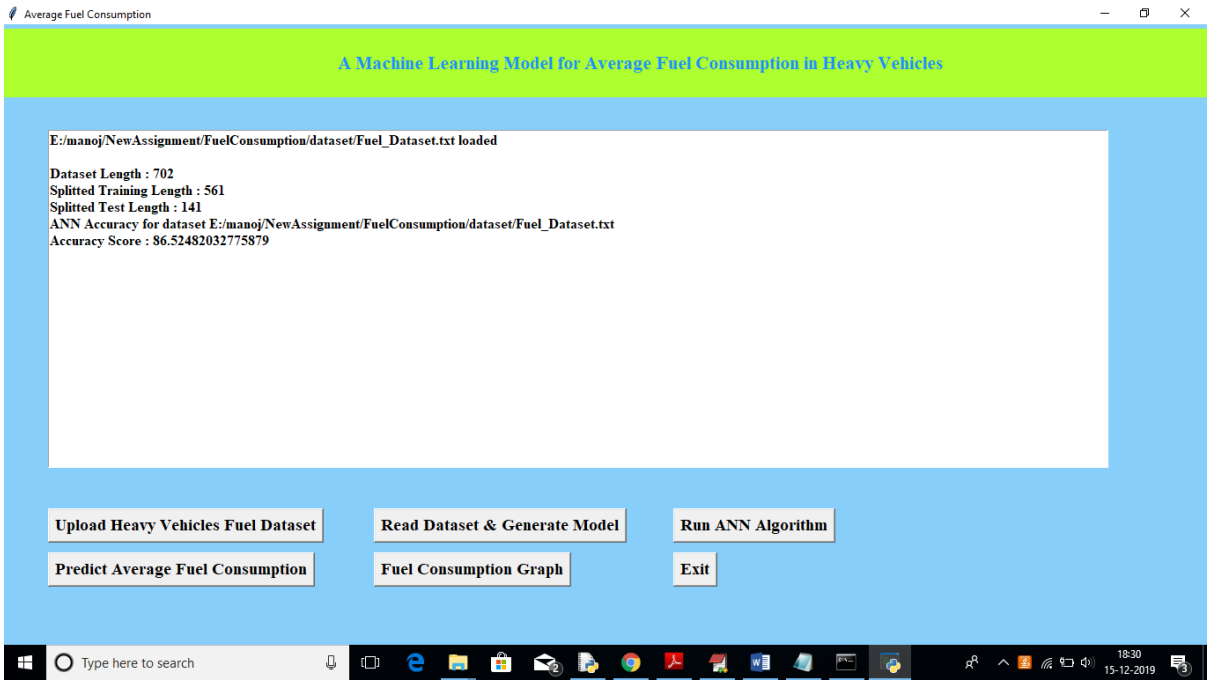


Figure 8.6

In above screen we got ANN prediction accuracy upto 86%. Now click on 'Predict Average Fuel Consumption' button to upload test data and to predict consumption for test data

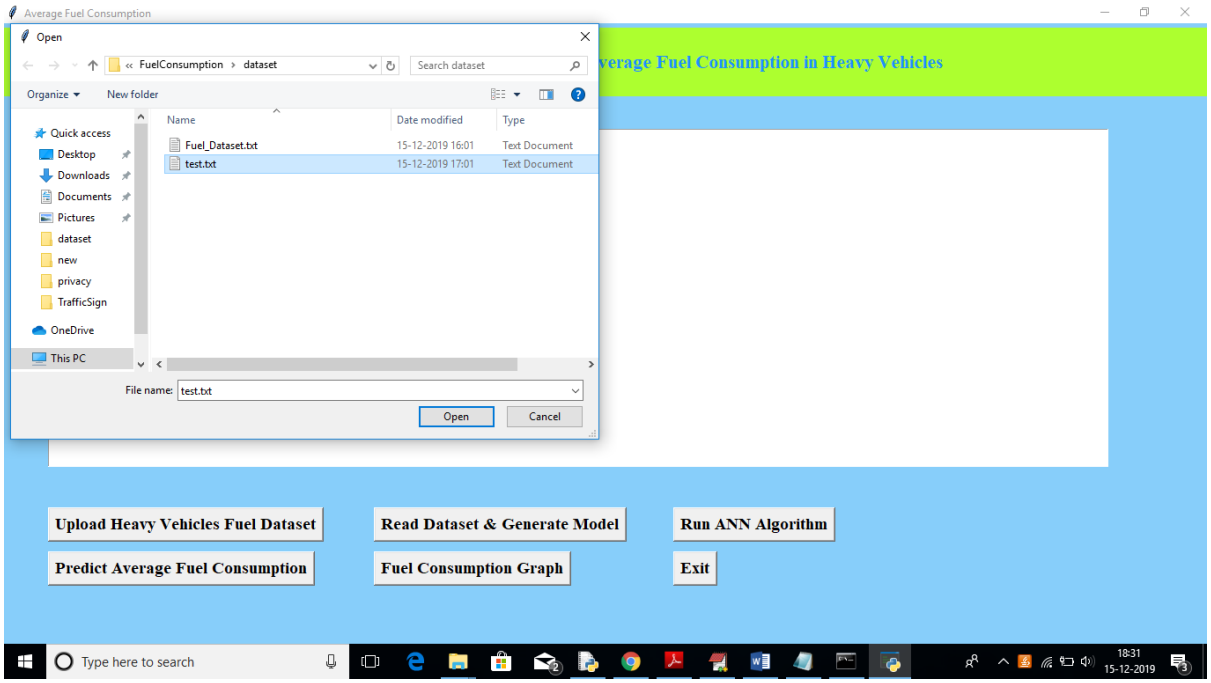


Figure 8.7

After uploading test data will get fuel consumption prediction result in below screen

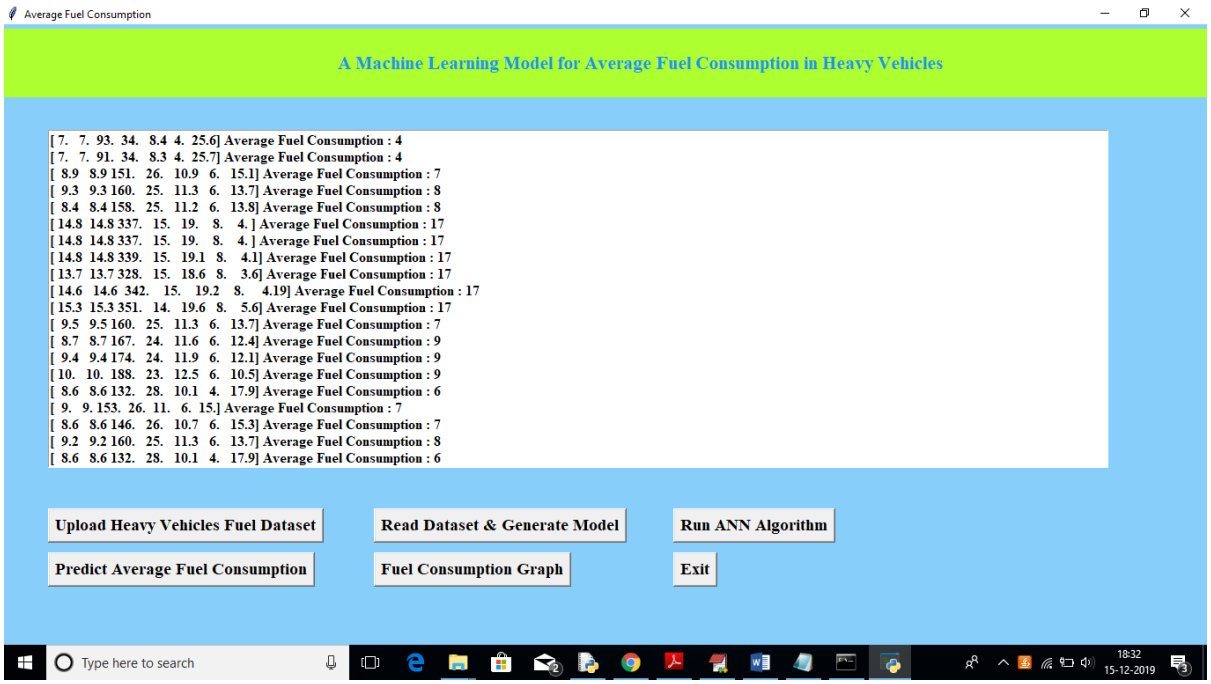


Figure 8.8

In above screen we got average fuel consumption for each test record per 100 kilo meter. Now click on ‘Fuel Consumption Graph’ to view below graph

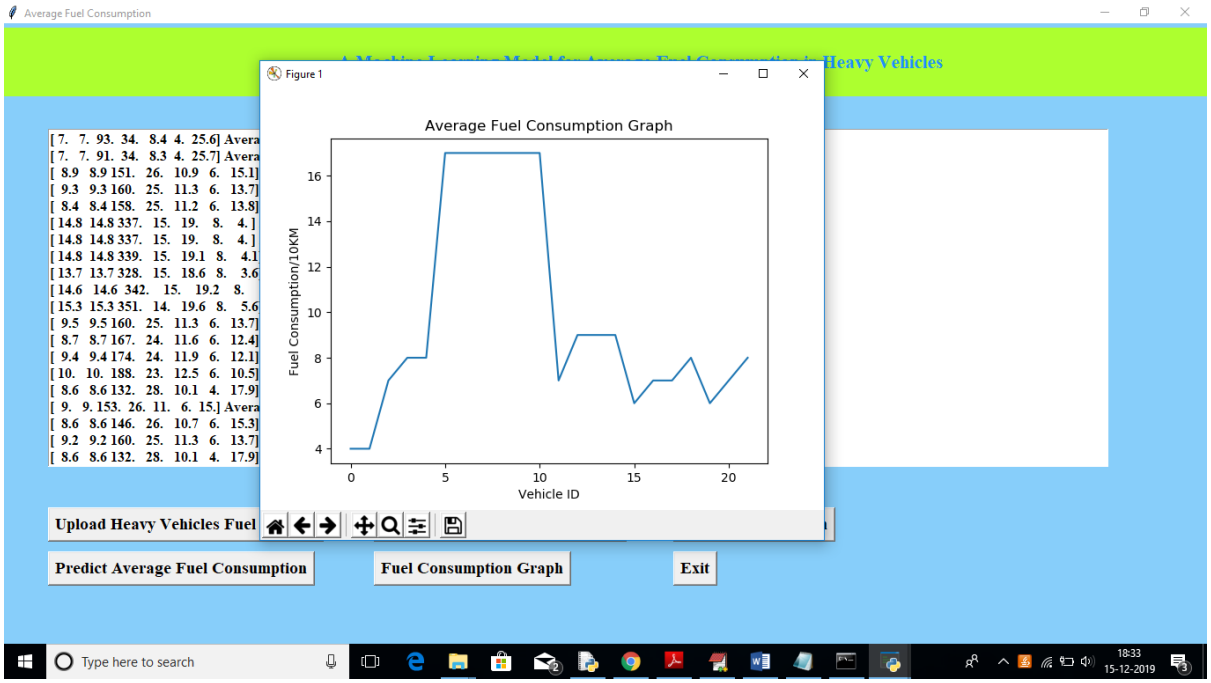


Figure 8.9

In above graph x-axis represents test record number as vehicle id and y-axis represents fuel consumption for that record.

CONCLUSION AND FUTURE ENHANCEMENT

Machine learning model that can be conveniently developed for each heavy vehicle in a fleet. The model relies on seven predictors: number of stops, stop time, average moving speed, characteristic acceleration, aerodynamic speed squared, change in kinetic energy and change in potential energy. The last two predictors are introduced in this paper to help capture the average dynamic behaviour of the vehicle. All of the predictors of the model are derived from vehicle speed and road grade. These variables are readily available from telematics devices that are becoming an integral part of connected vehicles. Moreover, the predictors can be easily computed on-board from these two variables. Future work includes understanding these differentiating factors and the selection of the appropriate window size. Expanding the model to other vehicles with different characteristics such as varying masses and aging vehicles is being studied. Predictors for these characteristics will be added in order to allow for the same model to capture the impact on fuel consumption due to changes in vehicle mass and wear.

REFERENCES

[1] Young-Rong Kim, Min Jung, Jun-Pum Park, “Development of a Fuel Consumption Prediction Model Based on Machine Learning Using Ship In-Service Data”, Journal of Marine Science and Engineering 9(2), 2021

[2] Chang Liu, Jian Rong, Yunlong Zhang, “Vehicle Fuel Consumption Prediction Method Based on Driving Behavior Data Collected from Smartphones”, Journal of Advanced Transportation, 2020.

[3] Zhihui Hu, Yongxin Jin, Qinyou Hu, “Prediction of Fuel Consumption for Enroute Ship Based on Machine Learning”, IEEE Access (Volume:7), 2019.

[4] Qi Zhao, Qi Phen, Liwang, “Real-Time Prediction of Fuel Consumption Based on Digital Map API”, Appl. Sci. 2019, 9, 1369.

[5] F. Perrotta, T. Parry, and L. C. Neves, “Application of machine learning for fuel consumption modelling of trucks,” in Big Data (Big Data), 2017 IEEE International Conference on. IEEE, 2017, pp. 3810–3815.

[6] S. Wickramanayake and H. D. Bandara, “Fuel consumption prediction of fleet vehicles using machine learning: A comparative study,” in Moratuwa Engineering Research Conference (MERCon), 2016. IEEE, 2016, pp. 90–95.

[7] A. A. Zaidi, B. Kulcsr, and H. Wymeersch, “Back-pressure traffic signal control with fixed and adaptive routing for urban vehicular networks,” IEEE Transactions on Intelligent Transportation Systems, vol. 17, no. 8, pp. 2134–2143, Aug 2016.

[8] L. Wang, A. Duran, J. Gonder, and K. Kelly, “Modeling heavy/mediumduty fuel consumption based on drive cycle properties,” SAE Technical Paper, Tech. Rep., 2015.

[9] A. Ivanco, R. Johri, and Z. Filipi, “Assessing the regeneration potential for a refuse truck over a real-world duty cycle,” SAE International Journal of Commercial Vehicles, vol. 5, no. 2012-01-1030, pp. 364–370, 2012.

[10] S. F. Haggis, T. A. Hansen, K. D. Hicks, R. G. Richards, and R. Marx, “In-use evaluation of fuel economy and emissions from coal haul trucks using modified sae j1321 procedures and pems,” SAE International Journal of Commercial Vehicles, vol. 1, no. 2008-01-1302, pp. 210–221, 2008.

[11] J. Zhao, W. Li, J. Wang, and X. Ban, “Dynamic traffic signal timingoptimization strategy incorporating various vehicle fuel consumptioncharacteristics,” IEEE Transactions on Vehicular Technology, vol. 65,no. 6, pp. 3874–3887, June 2016.

[12] G. Ma, M. Ghasemi, and X. Song, “Integrated powertrain energymanagement and vehicle coordination for multiple connected hybridelectric vehicles,” IEEE Transactions on Vehicular Technology, vol. 67,no. 4, pp. 2893–2899, April 2018.9

[13] A. C. Brandt, E. A. Frame, and R. W. Warden, “Sae j1321 testing usingm1083a1 fmtvs,” SOUTHWEST RESEARCH INST SAN ANTONIOTX TARDEC FUELS AND LUBRICANTS RESEARCH FACILITY,Tech. Rep., 2010.

[14] W. Zeng, T. Miwa, and T. Morikawa, “Exploring trip fuel consumptionby machine learning from gps and can bus data,” Journal of the EasternAsia Society for Transportation Studies, vol. 11, pp. 906–921, 2015.

[15] H. Almer, “Machine learning and statistical analysis in fuel consumption ´prediction for heavy vehicles,” 2015.

PUBLICATIONS

Measuring of Average Fuel Consumption in Heavy Vehicles Using Machine Learning Algorithms

[1]Yukthi G, [2]S J N V L Sai Shravani, [3]Rohan Reddy, [4]Aila Tanay Reddy,

[5]Dr.R Santosh Kumar, [6]Dr.P. Santosh Kumar Patra, [7]Dr. G. Jawaharlal Nehru

[1234]UG Scholar, [5]Associate Professor, [6]Principal & Professor in CSE, [7]Assistant Professor

Department of Computer Science and Engineering

St. Martin's Engineering College, Secunderabad – 500 100, India

E-Mail: [1]yukthi99@gmail.com, [2]surekutchishravani@gmail.com, [3]rohan.badhri@gmail.com, [4]atanayreddy@gmail.com

Abstract

When designing individualized machine learning models for fuel consumption, this paper proposes a data summarization approach focused on distance rather than the standard time span. To develop a highly predictive neural network model for average fuel consumption in heavy vehicles, this method is combined with seven predictors derived from vehicle speed and road grade. To maximize fuel consumption across a fleet, the proposed model can be easily developed and deployed for each individual vehicle in the fleet. The model's predictors are averaged over predetermined distance window sizes. The results show that a 1 km window can predict fuel consumption with a 0.91 coefficient of determination and a mean absolute peak-to-peak percent error of less than 4% for routes that involve both city and highway duty cycle segments with a 0.91 coefficient of determination and a mean absolute peak-to-peak percent error of less than 4%.

Keywords: vehicle modelling, neural networks, average fuel consumption, data summarization, fleet management

1. INTRODUCTION

Manufacturers, regulators, and customers are all interested in vehicle fuel consumption models. They are needed at all stages of the vehicle's life cycle. The emphasis of this paper is on calculating average fuel consumption for heavy vehicles during service and maintenance. In general, there are three types of techniques for developing fuel consumption models: Models based on physics, which are derived from a thorough knowledge of the physical system. These

models use comprehensive mathematical equations to explain the dynamics of the vehicle's components at each time phase.

Data-driven machine learning models, which reflect an abstract mapping from an input space containing a selected set of predictors to an output space containing the target output, in this case average fuel consumption

Statistical models which are data-driven and provide a mapping between the probability distribution of a selected set of predictors and the target outcome.

The above techniques have trade-offs in terms of cost and precision, depending on the specifications of the intended application. This paper proposes a model for individual heavy vehicles in a large fleet that can be easily produced. A fleet management system relies on accurate representations of all of the vehicles in the fleet. Allison Transmission, Inc. contributed to this study. Manager should refine route planning for all vehicles based on each vehicle's expected fuel usage, ensuring that route assignments are matched to reduce total fleet fuel consumption. These fleets can be used in a variety of industries, including goods transportation, public transportation, construction trucks, and refuse trucks. Without comprehensive knowledge of the vehicles' physical characteristics and dimensions, the approach for each fleet must apply and adjust to a wide range of vehicle technologies (including potential ones) and configurations. When weighing the desired precision against the expense of developing and adapting an individualised model for each vehicle in the fleet, machine learning emerges as the method of choice.

There have been some previous models suggested for both instantaneous and average fuel consumption. Since they can capture the dynamics of the system's actions at various time steps, physics-based models are ideally suited for predicting instantaneous fuel consumption. Due to the difficulty of recognising trends in real-time data, machine learning models are unable to predict instantaneous fuel consumption with a high degree of accuracy. These models, on the other hand, are capable of accurately identifying and learning patterns in average fuel usage. Previously proposed machine learning models for average fuel consumption use a series of predictors gathered over time to predict fuel consumption in gallons per mile or litres per kilometre. Although our proposed method is still based on average fuel consumption, it varies from previous models in that the predictors' input space is quantized with respect to a fixed distance rather than a fixed time span.

All predictors in the proposed model are aggregated with respect to a fixed window that reflects the vehicle's distance travelled, resulting in a better mapping from the input space to the model's output space. Previous machine learning models, on the other hand, had to not only learn the patterns in the input data, but also convert from the input domain's time-based scale to the output domain's distance-based scale (i.e., average fuel consumption). Using the same scale for the model's input and output spaces has many advantages:

Data is obtained at a proportional rate to its influence on the result. The amount of data obtained from a vehicle at a stop is the same as the amount of data2 collected while the vehicle is moving when the input space is sampled with respect to time.
The model's predictors will account for the effect of both the service cycle and the climate on the vehicle's average fuel consumption (for example, the number of stops in city traffic over a given distance).

Data from raw sensors can be consolidated on-board into a small number of predictors, requiring less storage and transmission bandwidth. Data summarization is best done on-board near the source of the data, given the increased computational capacities of modern vehicles.
New technologies such as vehicle-to-infrastructure (V2I) and dynamic traffic management (DTM) can be used to improve fuel efficiency at the vehicle, road, and time of day levels.
The following is how the rest of the paper is organised: Section II reviews previous related work, Section III introduces the proposed machine learning model, Section IV explains the data collection and data summarization method, Section V discusses the effects of using the proposed model in various settings, and Section VI summarises the main findings of this study and suggests potential research directions.

2. RELATED WORK

To model average fuel consumption, physics-based, machine learning, and statistical models have all been used. For heavy-duty vehicles, the EPA and the European Commission developed physics-based, full-vehicle simulation models. As compared to actual measurements obtained from a flow meter, these models are capable of predicting average fuel consumption with an accuracy of 3%. This degree of precision comes at the expense of a significant amount of production time. Statistical methods are at the other end of the modelling continuum, and they are used under strict testing conditions to ensure that the recorded findings are standardised and repeatable. For example, the Code of Federal Regulations (CFR) proposes a model that estimates fuel consumption for new vehicles based on well-defined statistical methods for particular duty cycles derived from segments of real-world trips.

Similarly, the SAE J1321 standard is used to estimate fuel consumption for trucks and buses after market changes or under different operating conditions. Using real data obtained from the field, this standard compares similar vehicles travelling the same route under similar operating conditions. The norm was used into equate the fuel consumption of a control vehicle to that of two test vehicles after the engine, transmission, and axle lubrication fluids were changed. The standard was also used in to compare the efficiency of three different fuel technologies in two coal-mining vehicles. In several studies, the generalizability of machine learning models to various vehicles and operating conditions made this modelling approach appealing for fuel consumption prediction. The remainder of this section discusses

these models in terms of the underlying machine learning technique, input space representation, and output space representation.

For the purpose of modelling fuel consumption, various forms of machine learning techniques have been used and compared. For example, compares gradiant boosting, neural networks, and random forest; compares neural networks and multivariate regression splines; and compares support vector machine, neural networks, and random forest. These studies determine a preferred methodology based on the findings. However, the differences between these techniques are minor, and the techniques are comparable, as described. Different data collection and data summarization methodologies, we conclude, are to blame for the discrepancies. We used neural networks in this paper because they are ideally suited for models with continuous input and output variables. Furthermore, neural networks are less vulnerable to data that is noisy.

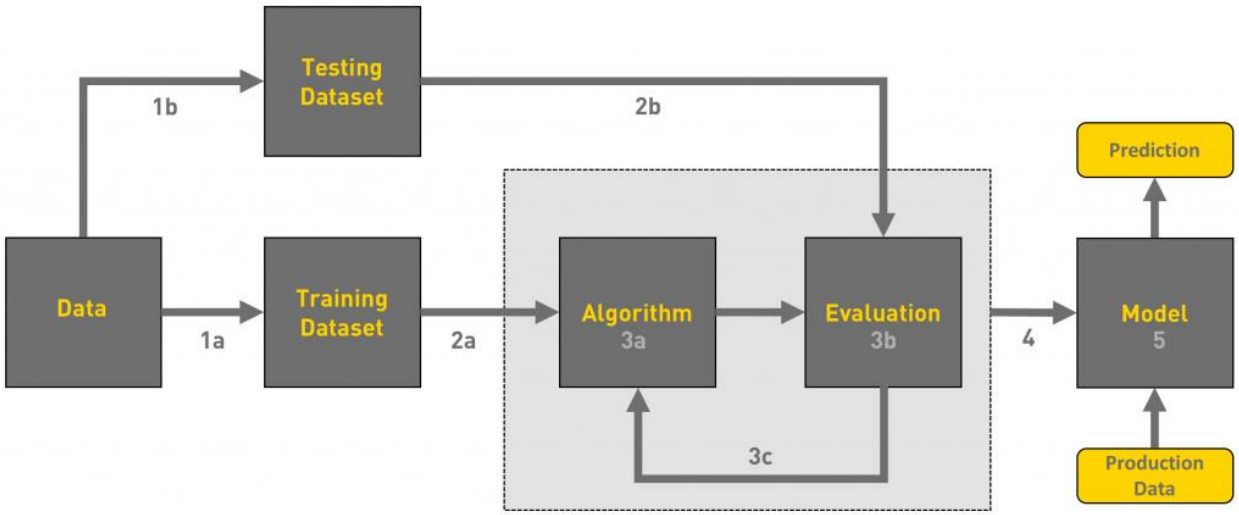
Similarly, the input to previously proposed fuel consumption models varies greatly. A holistic model could try to capture driver action, vehicle dynamics, and the vehicle's effect on the environment. Combinations of first, second, third, and fourth orders of vehicle acceleration and speed are used as predictors in the models implemented. Vehicle speed, distance travelled, elevation, longitude, latitude, and day of the week are all predictors. It employs road-related predictors (e.g., slope, curvature, and roughness) as well as vehicle-related predictors (e.g., vehicle speed, acceleration, gear, and percent torque). Acceleration, percent torque, and gradient were found to be the most powerful predictors in a previous analysis. Since the vehicle speed was kept nearly constant during the data collection, it was unimportant. If we looked at over 30 predictors, including wind speed, platooning, engine power, and breaking rate, with road grade, vehicle speed, and vehicle weight emerging as the most critical. Since a standard sensor for vehicle weight is not commonly available, the weight was calculated using the suspension. We also use vehicle speed and road grade to derive predictors for the proposed model in this paper. These variables can be accessed directly from telematics devices that are non-invasive, inexpensive, and widely accessible. The predictors of the models are usually extracted from various sensor values sampled at fixed time intervals. The accuracy of the proposed fuel consumption models is compared with input data obtained at 1 minute and 10 minute intervals, with the author concluding that the 10 minute interval yields more accurate models. Measurements are taken every 1 minute or 1 mile, whichever is shorter. Given that the vehicles in this study were travelling at a constant speed, this equates to collecting input data over a fixed distance of approximately one mile. Both models seem to suggest that gathering input data over a longer distance is better for fuel consumption modelling. The values of the predictors are aggregated in this paper over fixed windows of travelled time. We also look at how the length of the window affects the model's precision for real-world service cycles of different vehicle speeds.

Fuel consumption models can produce either a fuel rate (liters/hour) or an average fuel consumption (liters/100km) as a production. The aim of fuel rate models is to forecast3 immediate fuel consumption.

These models have a poor point-wise accuracy (i.e., accuracy at the sample level). The models are able to provide reasonably accurate average fuel consumption by averaging the expected fuel rates over a long time period or distance. [4] used this technique, with precision ranging from 3.73 percent to 6.83 percent from the measured average fuel rate over an entire real-world service cycle. The method was also used in [3], and after the expected fuel rates were converted into fuel consumption, the model was able to predict total fuel consumption with a 2 percent accuracy over a 365 km journey. However, when the study's point-wise expected fuel rates were compared to observed fuel rates, the coefficient of determination (0.3) was found to be poor. In both of the above experiments, the models' input and output are in the time domain. The contribution of the models implemented in [7] is the ratio of consumed fuel to distance travelled. As previously mentioned, the vehicles' speed in this study was constant, resulting in the sampling of both the model's input and output in the distance domain. The estimated fuel consumption had a 3 percent mean absolute error. In contrast to the two studies discussed above, the model presented in [7] has a high R2 value (> 0.8), indicating that it can make strong point-wise predictions. The models proposed in [15] illustrate some of the challenges that machine learning models face when the input and output domains are not the same. The input is aggregated in the time domain over 10 minute intervals in this analysis, and the output is fuel consumption over the distance travelled over that time span. For a mean fuel consumption of 30 l/100km, the root mean square error of the expected fuel consumption over the entire duty cycle was 7.4 l/100km. As compared to models that calculate the error over the entire trip [3], [4], this error, which is calculated point-wise for each 10 minute time interval, is comparatively high.

3. METHODOLOGY OF THE RESEARCH WORK

The model is based on service cycles obtained from a single truck with an estimated mass of 8,700kg that was exposed to a variety of transients in the Indianapolis area, including both urban and highway traffic. The SAE J1939 standard for serial control and communications in heavy-duty vehicle networks [24] was used to collect data.



Along two separate routes, twelve drivers were asked to demonstrate good or poor conduct. Drivers who were behaving well expected braking and, where possible, enabled the vehicle to coast. Since some drivers participated rather than others, the distribution of drivers and routes across the data set is not standardised. The vehicle CAN bus produced 3,302,890 data points sampled at 50Hz during this field test, covering a total distance of 778.89 km over 56 trips of varying lengths. The majority of the trips were between 10 and 15 kilometres long. Synthetic duty cycles over a long distance were created by assembling segments from field duty cycles chosen at random to maximise the number of data points. Furthermore, a different group of drivers is allocated to the training segments than to the testing segments, ensuring that the training (Ftr) and testing (Fts) data sets extracted from the respective segments are entirely separate.

A. Model Predictors

To generate the model's predictors, several processing steps were required. Two parameters, namely road grade and transmission output speed, are used to create these predictors. Down sampling the road grade and calculating the vehicle speed from the transmission output speed was the first processing stage. An on-board inclinometer was used to calculate the road grade, which was then down-sampled to 1Hz. An analysis of the data also revealed that the vehicle speed and transmission output speed have a linear relationship, as shown by the equation:

$$\text{VehicleSpeed} \approx 59.3 \times \text{TransmissionOutputSpeed} \quad (15)$$

A moving average low pass filter was added to the vehicle speed obtained using (15) in order to minimise noise in the variable, and the variable was down-sampled from 50Hz to 1Hz..The synthetic duty cycles were calculated in the second processing stage. The duty cycles in the real data were divided into segments identified by intervals between consecutive vehicle stops in order to achieve this goal (Figure 1). Both twelve drivers in the study contributed a total of 455 real data segments. The training data set (Ftr) was made up of 358 segments obtained from nine drivers, while the testing data set (Fts) was made up of 97 segments obtained from the remaining three drivers in the sample.

One synthetic duty cycle is generated by sampling real data segments without replacement and concatenating the selected segments until a total distance of 15 kilometres is reached. The total distance of 15 kilometres was chosen to mimic the actual field data collection routes. It was discovered that 15km of data needs an average of five segments. Figure 1 depicts a synthetic duty cycle generated using this method. The above method of combining segments resulted in a constant vehicle speed. However, as shown in Figure 2, there were discontinuities in road grade from one section to the next. These service cycles are then averaged over a predetermined distance depending on the desired window size (x). For each data set and window size considered in this paper, Table I shows the total number of data points (i.e., windows) as well as the total distance.

The third step in the input data processing consists of generating the predictors for the proposed model. As previously mentioned, these predictors are calculated for each window and derived from vehicle speed and road grade. The selected predictors consist of:

- number of stops,
- time stopped,
- average moving speed,
- characteristic acceleration,
- aerodynamic speed squared,
- change in kinetic energy and
- change in potential energy.

The above predictors were selected because they are believed to capture the vehicle dynamics as well as the driver’s behavior and the impact of the route on the target output of the model (i.e., fuel consumption).

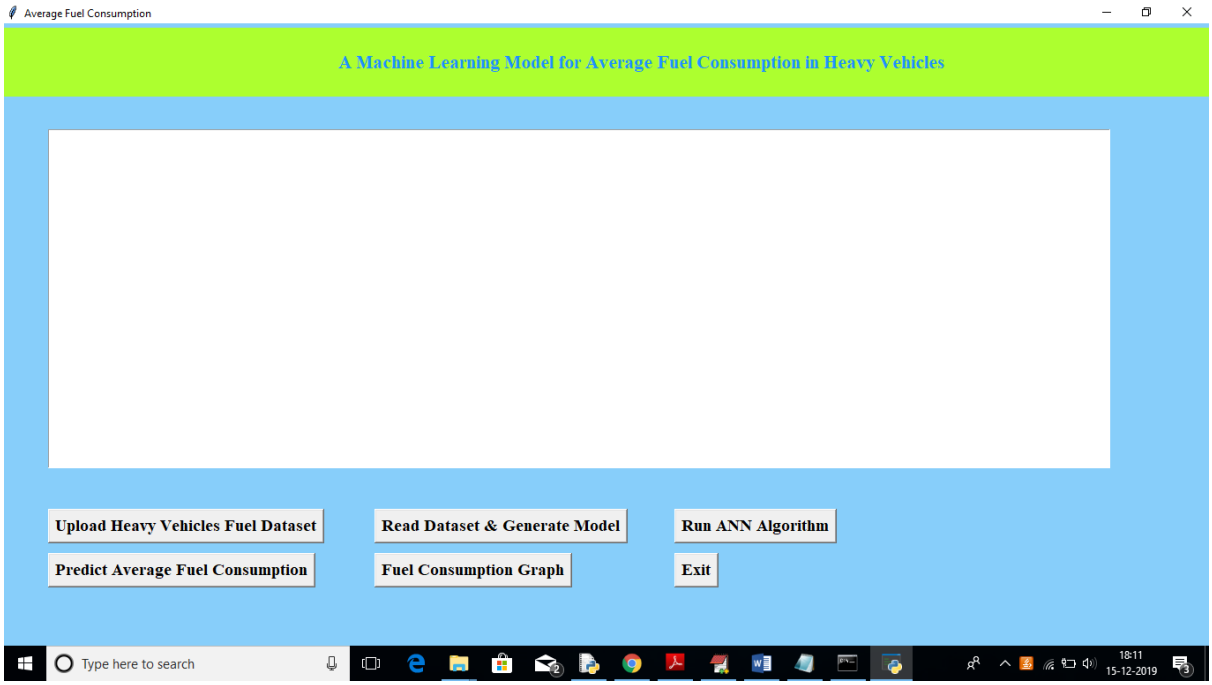
EXPERIMENTAL ANALYSIS

The seven predictors are used as input to the neural network model. The first layer of the network is made up of this. The first layer then feeds five neurons into a secret layer. The secret layer then feeds a single neuron into an output player. Figure 3 shows the RMSE (11) for three models with window sizes of 1, 2, and 5 kilometres during preparation. Each data point in the top plot corresponds to the RMSE values after training the model with 500 windows. All models converge to an RMSE of less than 0.2 l/100km, as shown in this graph. The models' convergence rates, on the other hand, differ. In reality, after 500 training windows, the 5km has an RMSE value of 0.16 l/100km, which drops to 0.08 l/100km when the model converges.

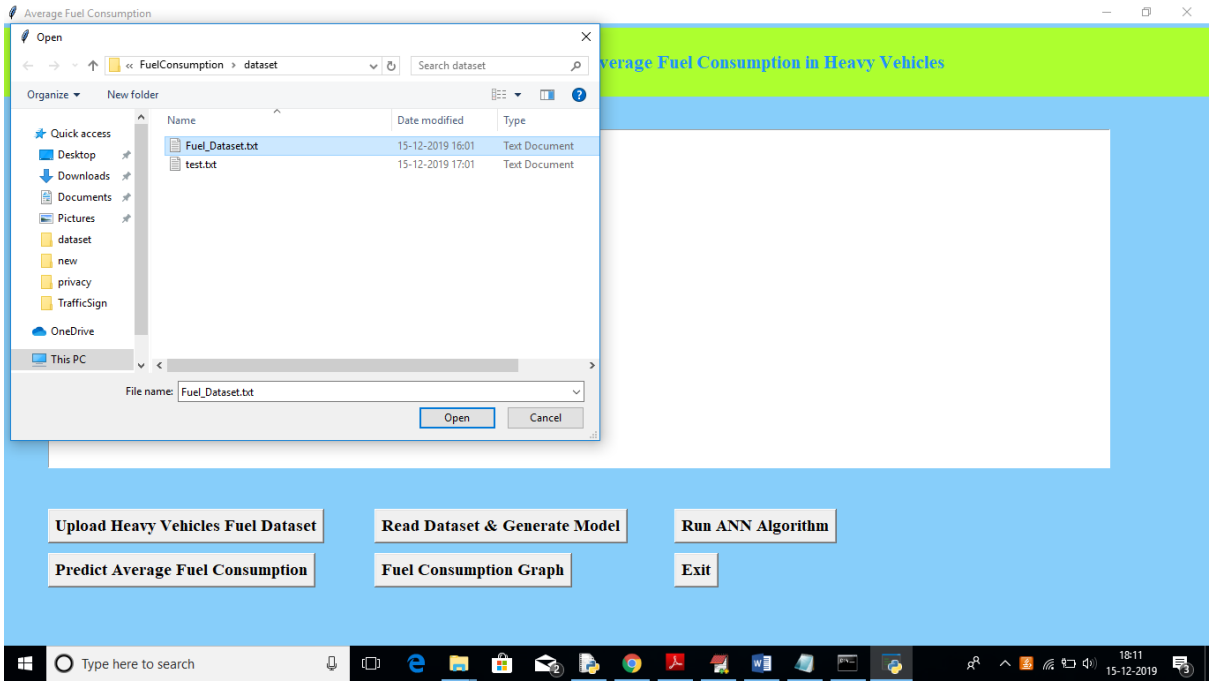
For the 1km model, the corresponding values are 0.34 l/100km and 0.14 l/100km, respectively. This pattern, when combined with the difference in standard deviation of average fuel consumption for the 1km and 5km windows, suggests that aggregating the input and output data over 5km provides a consistent profile for the vehicle's fuel consumption over the routes, which does not require comprehensive learning.

RESULTS AND ANALYSIS

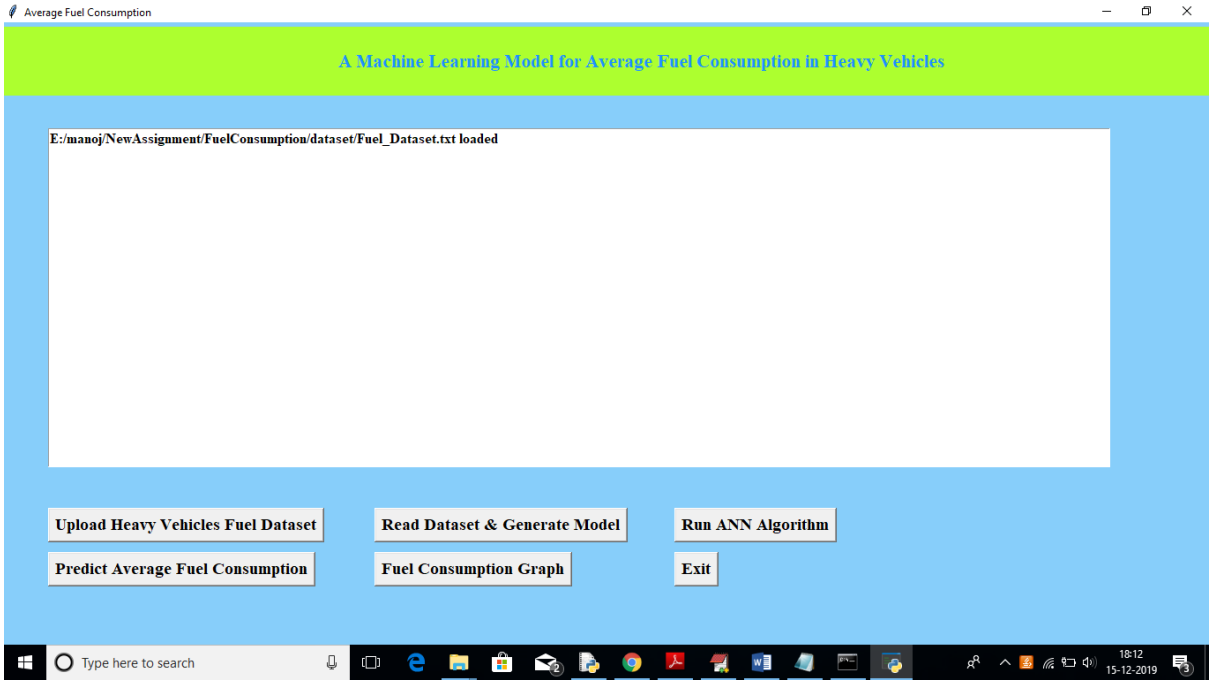
To run this project double click on 'run.bat' file to get below screen



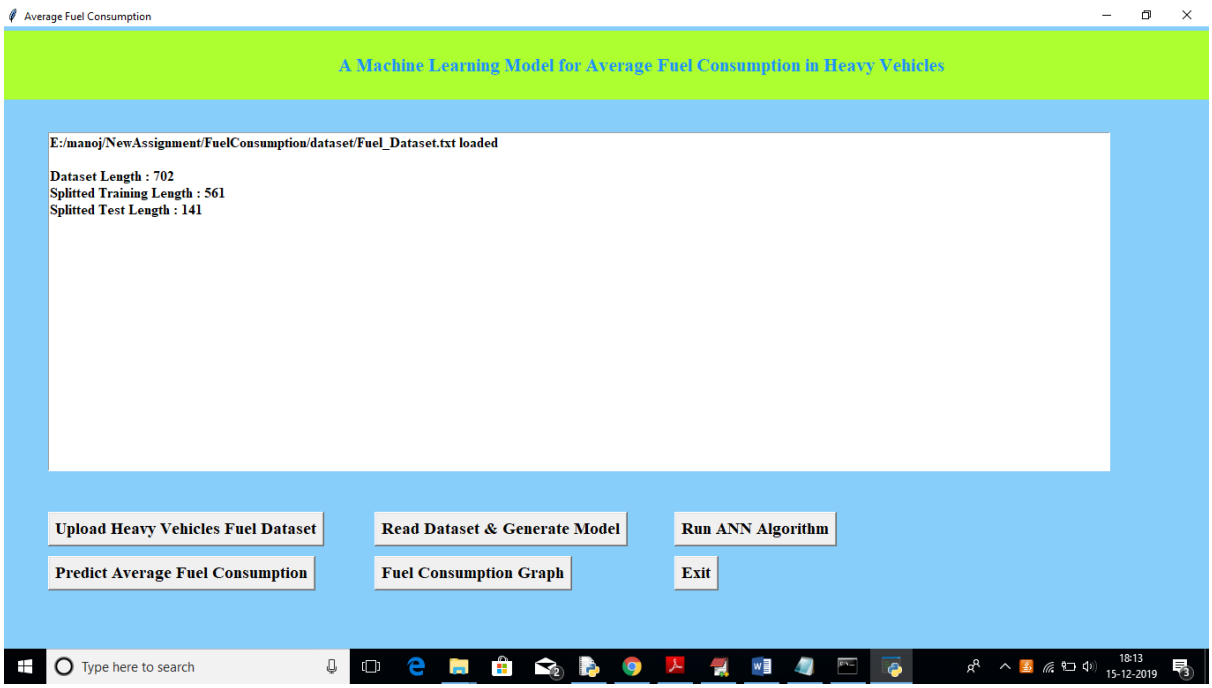
In above screen click on 'Upload Heavy Vehicles Fuel Dataset' button to upload train dataset



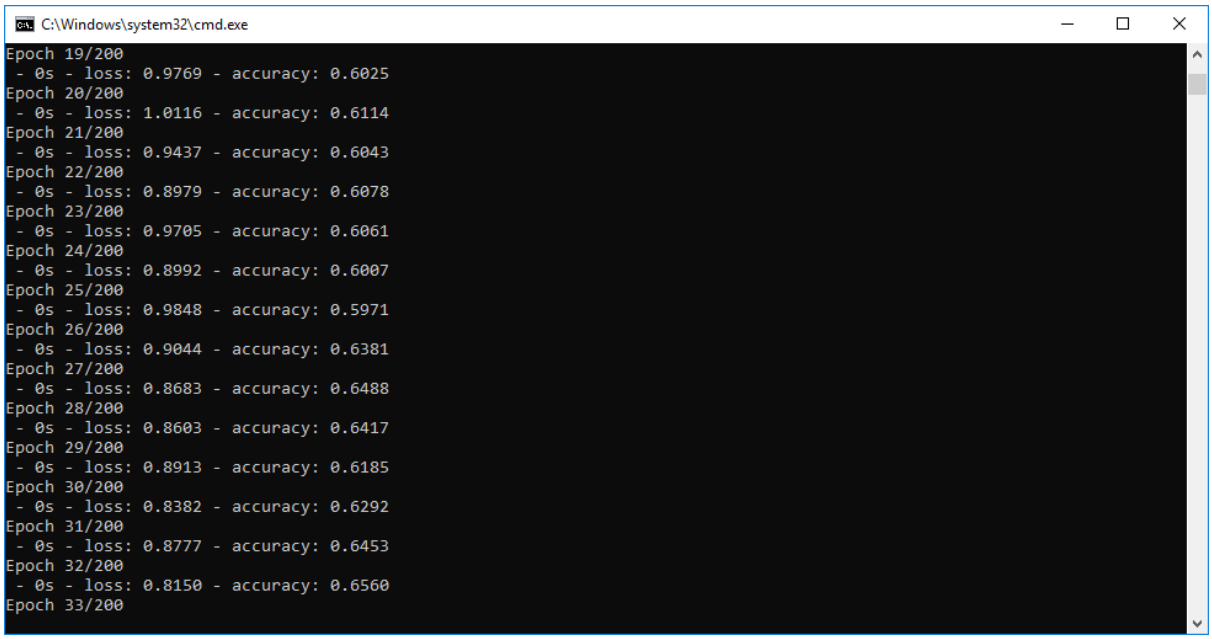
In above screen uploading ‘Fuel_Dataset.txt’ which can be used to train model. After uploading dataset will get below screen



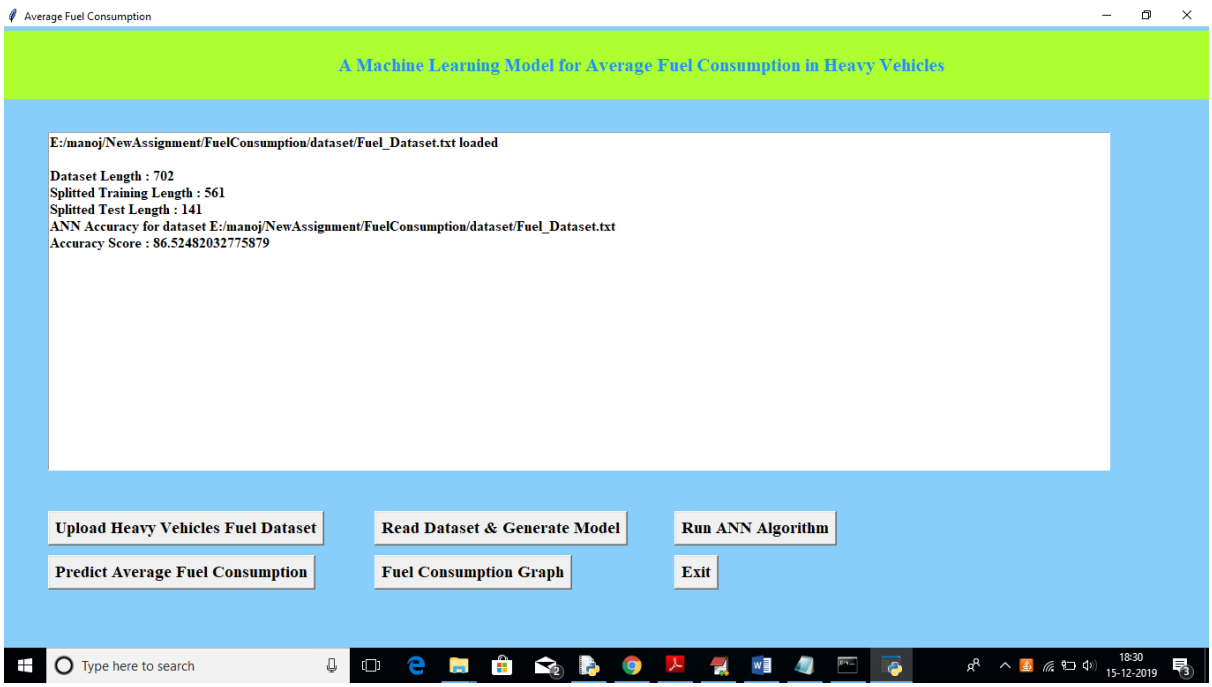
Now in above screen click on ‘Read Dataset & Generate Model’ button to read uploaded dataset and to generate train and test data



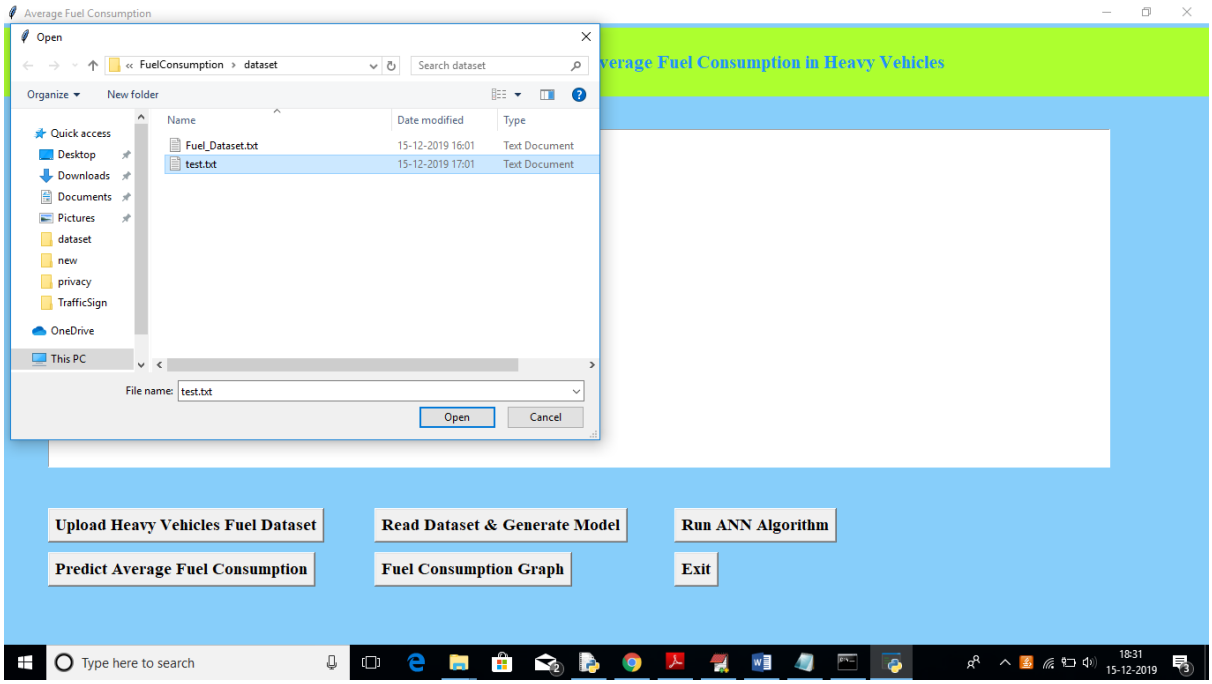
In above screen we can see total number of records in dataset, number of records used for training and number for records used for testing. Now click on ‘Run ANN Algorithm’ button to input train and test data to ANN to build ANN model.



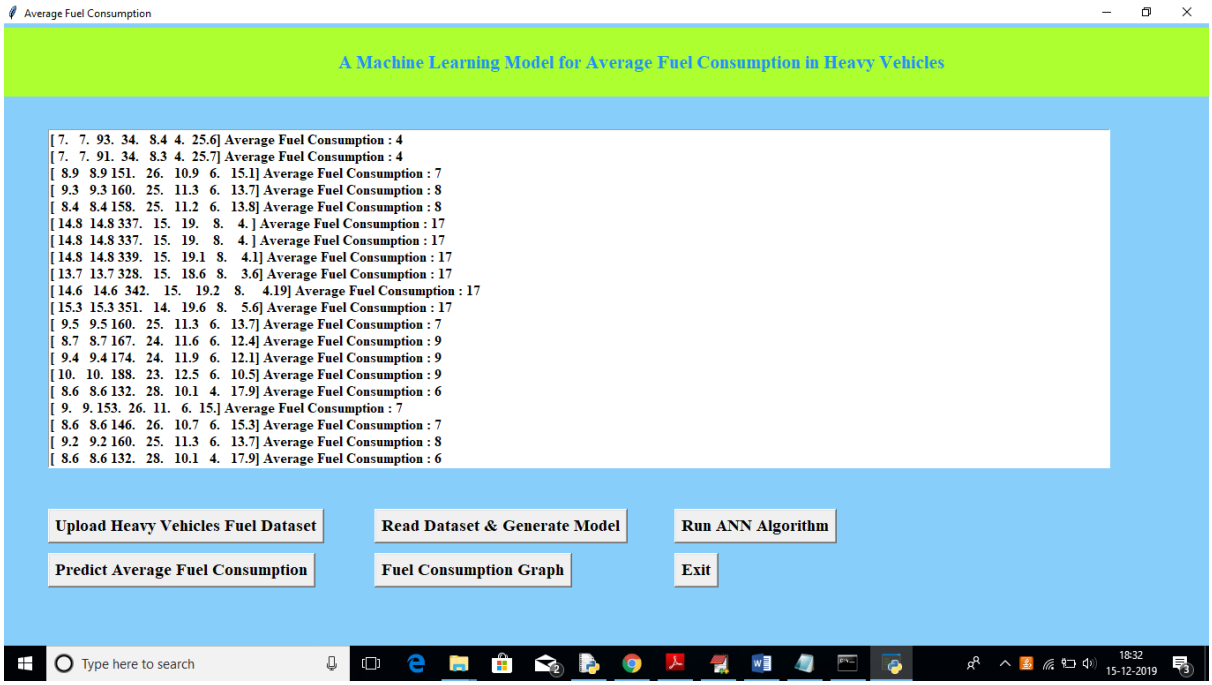
In above black console we can see all ANN processing details, After building model will get below screen



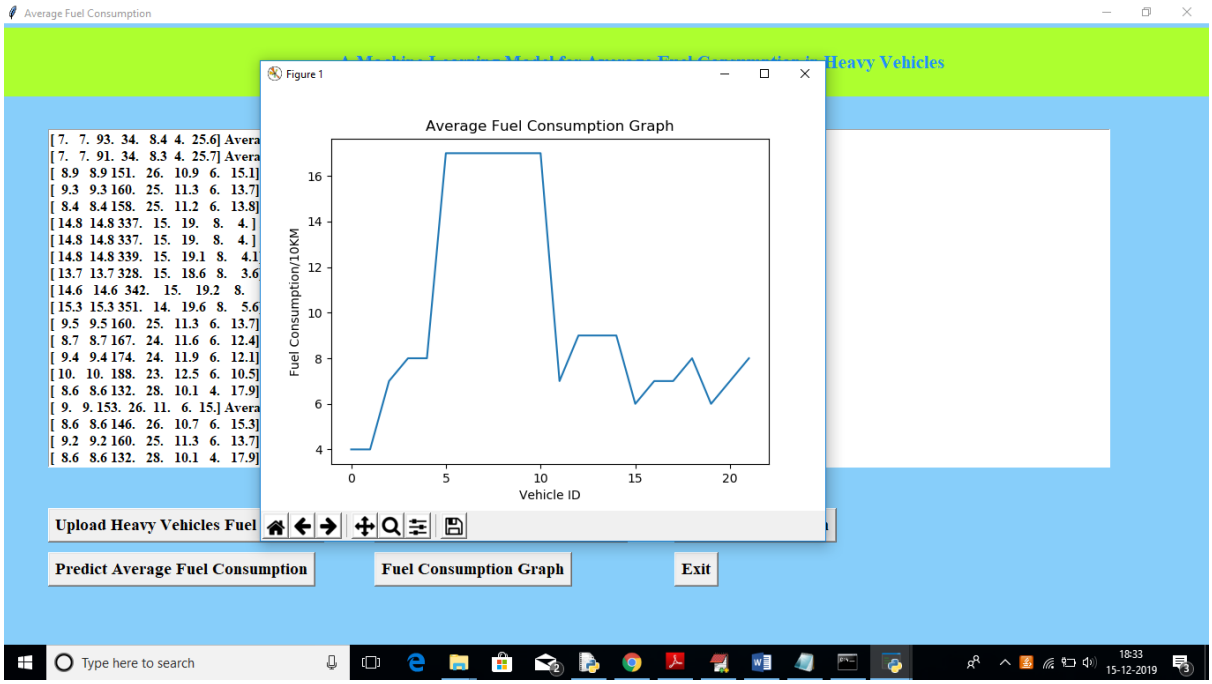
In above screen we got ANN prediction accuracy upto 86%. Now click on 'Predict Average Fuel Consumption' button to upload test data and to predict consumption for test data



After uploading test data will get fuel consumption prediction result in below screen



In above screen we got average fuel consumption for each test record per 100 kilo meter. Now click on 'Fuel Consumption Graph' to view below graph



In above graph x-axis represents test record number as vehicle id and y-axis represents fuel consumption for that record.

RELATED WORK

The model's performance is the average fuel consumption in litres per 100 kilometres for each window. Fuel rates from the CAN bus are used to calculate the average consumption. Since synthetic duty cycles are extracted from a random collection of real duty cycle segments, discontinuities in the fuel rate are observed from one segment to the next, just as they are with road grade (Figure 2). Since the fuel prices are averaged over the entire window in order to measure the model's performance, the effect of these discontinuities is minimal (i.e., average fuel consumption). An examination of the segments in the field data reveals a variation in average fuel consumption across all trips. Over the course of a journey, for example, a 20% difference in fuel consumption was observed between good and poor driver action. Furthermore, differences in average fuel consumption are observed for various window sizes. Table II displays the average fuel consumption for the 1, 2, and 5km windows for all journeys, as well as the mean and standard deviation. The standard deviation decreases as the window size increases, while the mean fuel consumption remains relatively constant across all windows. In summary, all of the proposed model's input features are extracted using the above technique from 1Hz samples of vehicle speed and road grade. A telematics software may be used to acquire these variables. These variables were derived from sensor values transmitted on the CAN bus in this analysis. The model's accuracy can differ depending on the data source and sampling frequency. The model's accuracy is also dependent on the output feature's accuracy. As compared to real fuel use, fuel consumption obtained from the May bus can have a margin of error of up to 5% [26]. Flowmeters have a higher level of precision. Flowmeters, on the other hand, are more costly. Fuel

consumption data from the CAN bus is used in [7], [15], and this paper, as well as high-precision fuel sensors in [3]. In the future, aspects of the data sources' accuracy will be investigated.

CONCLUSION

This paper presented a machine learning model for each heavy vehicle in a fleet that can be easily built. Number of stops, stop time, average moving speed, characteristic acceleration, aerodynamic speed squared, change in kinetic energy, and change in potential energy are the seven predictors used in the model. The last two predictors are added in this paper to aid in capturing the vehicle's average dynamic activity. The model's predictors are all based on vehicle speed and road grade. Telematics systems, which are becoming an increasingly important part of connected cars, provide easy access to these variables. Furthermore, from these two variables, the predictors can be conveniently computed on-board. Instead of a fixed time interval, the model predictors are aggregated over a fixed distance travelled (i.e., window). This mapping of the input space to the distance domain aligns with the domain of the target output, resulting in a fuel consumption machine learning model with an RMSE of 0.015 l/100km. The effectiveness of various model configurations with 1, 2, and 5km window sizes was tested. The 1km window has the best accuracy, according to the data. With a CD of 0.91, this model can estimate real fuel consumption on a per 1km basis. This performance is comparable to that of physics-based models, and the proposed model outperforms previous machine learning models that were only comparable for entire long-distance trips.

The cost of the model in terms of data collection and on-board computation should be considered when choosing an appropriate window size. Furthermore, the window size is likely to vary depending on the application. A 1 km window size is recommended for fleets with short trips (e.g., construction vehicles inside a site) or urban traffic routes. A 5km window size can be suitable for long-haul fleets. Since the service cycles in this study included both highway and city traffic, the 1 km window was more appropriate than the 5 km window. Understanding these differentiating variables and selecting the optimal window size are two areas of future research. The model is being expanded to include other vehicles with various features, such as varying masses and ageing vehicles. Predictors for these characteristics will be introduced so that the same model will capture the effect of changes in vehicle mass and wear on fuel consumption.

Future work may involve determining the shortest distance needed for training and model and determining how often a model must be synchronised with the physical system in operation using online training to maintain the model's prediction accuracy.

REFERENCES

[1] B. Lee, L. Quinones, and J. Sanchez, “Development of greenhouse gas emissions model for 2014-2017 heavy-and medium-duty vehicle compliance,” SAE Technical Paper, Tech. Rep., 2011.

[2] G. Fontaras, R. Luz, K. Anagnostopoulus, D. Savvidis, S. Hausberger, and M. Rexeis, “Monitoring co2 emissions from hdv in europe-an experimental proof of concept of the proposed methodolgical approach,” in *20th International Transport and Air Pollution Conference*, 2014.

[3] S. Wickramanayake and H. D. Bandara, “Fuel consumption prediction of fleet vehicles using machine learning: A comparative study,” in *Moratuwa Engineering Research Conference (MERCon), 2016*. IEEE, 2016, pp. 90–95.

[4] L. Wang, A. Duran, J. Gonder, and K. Kelly, “Modeling heavy/mediumduty fuel consumption based on drive cycle properties,” SAE Technical Paper, Tech. Rep., 2015.

[5] *Fuel Economy and Greenhouse gas exhaust emissions of motor vehicles Subpart B - Fuel Economy and Carbon-Related Exhaust Emission Test Procedures*, Code of Federal Regulations Std. 600.111-08, Apr 2014.

[6] *SAE International Surface Vehicle Recommended Practice, Fuel Consumption Test Procedure - Type II*, Society of Automotive Engineers Std., 2012.

[7] F. Perrotta, T. Parry, and L. C. Neves, “Application of machine learning for fuel consumption modelling of trucks,” in *Big Data (Big Data), 2017 IEEE International Conference on*. IEEE, 2017, pp. 3810–3815.

[8] S. F. Haggis, T. A. Hansen, K. D. Hicks, R. G. Richards, and R. Marx, “In-use evaluation of fuel economy and emissions from coal haul trucks using modified sae j1321 procedures and pems,” *SAE International Journal of Commercial Vehicles*, vol. 1, no. 2008-01-1302, pp. 210–221 , 2008.

[9] A. Ivanco, R. Johri, and Z. Filipi, “Assessing the regeneration potential for a refuse truck over a real-world duty cycle,” *SAE International Journal of Commercial Vehicles*, vol. 5, no. 2012-01-1030, pp. 364–370 , 2012.

[10] A. A. Zaidi, B. Kulcsr, and H. Wymeersch, “Back-pressure traffic signal control with fixed and adaptive routing for urban vehicular networks,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 8 , pp. 2134–2143, Aug 2016.

[11] J. Zhao, W. Li, J. Wang, and X. Ban, “Dynamic traffic signal timing optimization strategy incorporating various vehicle fuel consumption characteristics,” *IEEE Transactions on Vehicular Technology*, vol. 65 , no. 6, pp. 3874–3887, June 2016.

[12] G. Ma, M. Ghasemi, and X. Song, “Integrated powertrain energy management and vehicle coordination for multiple connected hybrid electric vehicles,” *IEEE Transactions on Vehicular Technology*, vol. 67 , no. 4, pp. 2893–2899, April 2018.

[13] A. C. Brandt, E. A. Frame, and R. W. Warden, “Sae j1321 testing using m1083a1 fmtvs,” SOUTHWEST RESEARCH INST SAN ANTONIO TX TARDEC FUELS AND LUBRICANTS RESEARCH FACILITY,
Tech. Rep., 2010.

[14] W. Zeng, T. Miwa, and T. Morikawa, “Exploring trip fuel consumption by machine learning from gps and can bus data,” *Journal of the Eastern Asia Society for Transportation Studies*, vol. 11, pp. 906–921, 2015.

[15] H. Almer,′ “Machine learning and statistical analysis in fuel consumption prediction for heavy vehicles,” 2015.

[16] S. McBride, C. Sandu, A. Alatorre, and A. Victorino, “Estimation of vehicle tire-road contact forces: A comparison between artificial neural network and observed theory approaches,” *SAE Technical Paper*, pp. 01–0562, 2018.

[17] G. N. Bifulco, F. Galante, L. Pariota, and M. R. Spena, “A linear model for the estimation of fuel consumption and the impact evaluation of advanced driving assistance systems,” *Sustainability*, vol. 7, no. 10, pp. 14326–14343, 2015.

[18] C. M. Atkinson, S. Petreanu, N. N. Clark, R. J. Atkinson, T. I. McDaniel, S. Nandkumar, and P. Famouri, “Numerical simulation of a two-stroke linear engine-alternator combination,” SAE Technical Paper, Tech. Rep., 1999.

[19] C.-M. Vong, P.-K. Wong, and Y.-P. Li, “Prediction of automotive engine power and torque using least squares support vector machines and bayesian inference,” *Engineering Applications of Artificial Intelligence*, vol. 19, no. 3, pp. 277–287, 2006.

[20] Y. Shi and R. Eberhart, “A modified particle swarm optimizer,” in *1998 IEEE International Conference on Evolutionary Computation Proceedings. IEEE World Congress on Computational Intelligence (Cat. No.98TH8360)*, May 1998, pp. 69–73.

[21] P. J. Werbos, “Backpropagation through time: what it does and how to do it,” *Proceedings of the IEEE*, vol. 78, no. 10, pp. 1550–1560, 1990.

[22] S. Shah, M. Hosseini, Z. B. Miled, R. Shafer, and S. Berube, “A water demand prediction model for central indiana,” in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[23] B. E. Flores, “A pragmatic view of accuracy measurement in forecasting,” *Omega*, vol. 14, no. 2, pp. 93–98, 1986.

[24] *SAE Serial Control and Communications Heavy Duty Vehicle Network - Top Level Document*, Society of Automotive Engineers Std., jun 2012. [Online]. Available: https://doi.org/10.4271/J1939-201206

[25] M. P. O’Keefe, A. Simpson, K. J. Kelly, and D. S. Pedersen, “Duty cycle characterization and evaluation towards heavy hybrid vehicle applications,” SAE Technical Paper, Tech. Rep., 2007.

Student Profiles

Yukthi G is currently pursuing her Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. She completed her intermediate from Narayana Junior College and 10th class from Pallavi Model School. She has been in the Marketing Department of Technology Awareness Month (TAM) for 2 years. Her responsibilities in that group included spawning marketing strategies laying hold on the aspects of TAM and aided in coordinating with the ideas and implementations of advertising and promotions of events in TAM. Her technical skills include C, Python and Java. She also has a basic understanding of C++, HTML, CSS. She took part in Employability Skill development Program conducted by Zensar. She is also an active student of Smart Interviews. Her participations include: National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020 , Women online workshop on "Women in Cyber Security and Privacy in 2020" which was conducted from 6th to 10th July 2020, "Know More - Teach More ", the Global Webinar on Cyber Threats and Defense Techniques conducted by GECF on 22nd July 2020 and IIC Online Sessions conducted by Institution's Innovation Council (IIC) of MHRD's Innovation Cell. She also took part in Workshop on "Arduino/Robotics" which was conducted in the college on 12th February 2019 and 13th February 2019, Workshop on "Ethical Hacking" which was conducted in the college on 31st January 2020 and 1st February 2020. She was also a student organizing member during two days "National Level Hackathon-2020" held on 7th and 8th February 2020 at the college. She has also done internships during her summer break and got practical hands on experience on Machine Learning(ML).She spends her free time taking online certification courses related to her field of study as well as personal interests from platform such as Coursera, Cursa and EdX. Her areas of interest are Python, Artificial Intelligence, Machine Learning and Deep Learning.

S J N V L Sai Shravani is currently pursuing her Bachelor of Technology in the stream of Computer Science and Engineering from St. Martin's Engineering College. She completed her Intermediate from Sri Chaitanya Junior Kalasala and 10th standard from Sri Chaitanya Techno School. Her technical skills include C, Java and Python and she has a basic understanding of C++. She is a student of Smart Interviews and participated in many events conducted by them through Hackerrank, Codechef, Codeforces and Interviewbit. Her other participations include: National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020, Women online workshop on "Women in Cyber Security and Privacy in 2020" which was conducted from 6 th to 10th July 2020, Workshop on "Ethical Hacking" which was conducted in St. Martin's Engineering College on 31st January 2020 and 1st February 2020. Apart from this, she was also a part of a student run NGO, Street cause during the year 2018-2019. She completed few certification courses from online platforms like Coursera, CursaApp and Unschool. Her areas of interest include Python, Data Analytics, Machine Learning and Artificial Intelligence.

Rohan Reddy is currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. He completed his intermediate from Narayana Junior College and 10th class from Delhi Public School. His technical skills include Java, Python. He also has a basic understanding of C. He is one of the student of Smart Interviews and participated in few tests conducted by them. His participations include: National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020, and he has completed few courses on coursera in the year 2020 and have also participated in Two-Day National Level Seminar On "Recent Trends in Cloud Computing, Fog and Edge Computing" scheduled on 18th June to 19th June 2021 April to 22nd May 2020. His areas of interest are Python, Java, Cloud Computing, Cyber Security, Machine Learning. He completed few certification courses from online platforms like Coursera, Data Camp.

Aila Tanay Reddy is currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. He completed his intermediate from Narayana Junior College and 10th class from Delhi Public School. His technical skills include Java , Python. He also has a basic understanding of C. He is one of the student of Smart Interviews and participated in few tests conducted by them. His participations include: National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020,and he has completed few courses on coursera in the year 2020 and have also participated in Two-Day National Level Seminar On "Recent Trends in Cloud Computing, Fog and Edge Computing" scheduled on 18th June to 19th June 2021 April to 22nd May 2020. His areas of interest are Python, Java ,Cloud Computing, Cyber Security, Machine Learning . He completed few certification courses from online platforms like Coursera, Data Camp.

APPENDICES

```
from tkinter import messagebox
from tkinter import *
from tkinter import simpledialog
import tkinter
from tkinter import filedialog
import matplotlib.pyplot as plt
import numpy as np
from tkinter.filedialog import askopenfilename
import pandas as pd
from sklearn import *
from sklearn.model_selection import train_test_split
from tensorflow.keras.models import Sequential

from tensorflow.keras.layers import Input, Dense



from tensorflow.keras.callbacks import EarlyStopping
from sklearn.preprocessing import OneHotEncoder
from tensorflow.keras.optimizers import Adam
import os



main = tkinter.Tk()
main.title("Average Fuel Consumption") #designing main screen
main.geometry("1300x1200")

global filename
global train_x, test_x, train_y, test_y
global balance_data
global model
```

```
global ann_acc
global testdata
global predictdata

def importdata():
    global balance_data
    balance_data = pd.read_csv(filename)
    balance_data = balance_data.abs()
    return balance_data

def splitdataset(balance_data):
    global train_x, test_x, train_y, test_y
    X = balance_data.values[:, 0:7]
    y_ = balance_data.values[:, 7]
    print(y_)
    y_ = y_.reshape(-1, 1)
    encoder = OneHotEncoder(sparse=False)
    Y = encoder.fit_transform(y_)
    print(Y)
    train_x, test_x, train_y, test_y = train_test_split(X, Y, test_size=0.2)
    text.insert(END,"Dataset Length : "+str(len(X))+"\n");
    return train_x, test_x, train_y, test_y

def upload(): #function to upload tweeter profile
    global filename
    filename = filedialog.askopenfilename(initialdir="dataset")
    text.delete('1.0', END)
    text.insert(END,filename+" loaded\n\n");

def generateModel():
    global train_x, test_x, train_y, test_y
```

```
    data = importdata()
    train_x, test_x, train_y, test_y = splitdataset(data)
    text.insert(END,"Splitted Training Length : "+str(len(train_x))+"\n");
    text.insert(END,"Splitted Test Length : "+str(len(test_x))+"\n");



def ann():
    global model
    global ann_acc
    model = Sequential()
    model.add(Dense(200, input_shape=(7,), activation='relu', name='fc1'))
    model.add(Dense(200, activation='relu', name='fc2'))
    model.add(Dense(19, activation='softmax', name='output'))
    optimizer = Adam(lr=0.001)
    model.compile(optimizer, loss='categorical_crossentropy', metrics=['accuracy'])
    print('CNN Neural Network Model Summary: ')
    print(model.summary())
    model.fit(train_x, train_y, verbose=2, batch_size=5, epochs=200)
    results = model.evaluate(test_x, test_y)
    text.insert(END,"ANN Accuracy for dataset "+filename+"\n");
    text.insert(END,"Accuracy Score : "+str(results[1]*100)+"\n\n")
    ann_acc = results[1] * 100


def predictFuel():
    global testdata
    global predictdata
    text.delete('1.0', END)
    filename = filedialog.askopenfilename(initialdir="dataset")
    testdata = pd.read_csv(filename)
```

```
    testdata = testdata.values[:, 0:7]
    predictdata = model.predict_classes(testdata)
    print(predictdata)
    for i in range(len(testdata)):
        text.insert(END,str(testdata[i])+" Average Fuel Consumption : "+str(predictdata[i])+"\n");


def graph():
    x = []
    y = []
    for i in range(len(testdata)):
        x.append(i)
        y.append(predictdata[i])
    plt.plot(x, y)
    plt.xlabel('Vehicle ID')
    plt.ylabel('Fuel Consumption/10KM')
    plt.title('Average Fuel Consumption Graph')
    plt.show()

font = ('times', 16, 'bold')
title = Label(main, text='A Machine Learning Model for Average Fuel Consumption in Heavy Vehicles')
title.config(bg='greenyellow', fg='dodger blue')
title.config(font=font)
title.config(height=3, width=120)
title.place(x=0,y=5)

font1 = ('times', 12, 'bold')
text=Text(main,height=20,width=150)
scroll=Scrollbar(text)
text.configure(yscrollcommand=scroll.set)
text.place(x=50,y=120)
```

```
text.config(font=font1)

font1 = ('times', 14, 'bold')
uploadButton = Button(main, text="Upload Heavy Vehicles Fuel Dataset", command=upload)
uploadButton.place(x=50,y=550)
uploadButton.config(font=font1)

modelButton = Button(main, text="Read Dataset & Generate Model", command=generateModel)
modelButton.place(x=420,y=550)
modelButton.config(font=font1)

annButton = Button(main, text="Run ANN Algorithm", command=ann)
annButton.place(x=760,y=550)
annButton.config(font=font1)

predictButton = Button(main, text="Predict Average Fuel Consumption", command=predictFuel)
predictButton.place(x=50,y=600)
predictButton.config(font=font1)

graphButton = Button(main, text="Fuel Consumption Graph", command=graph)
graphButton.place(x=420,y=600)
graphButton.config(font=font1)

exitButton = Button(main, text="Exit", command=exit)
exitButton.place(x=760,y=600)
exitButton.config(font=font1)


main.config(bg='LightSkyBlue')
main.mainloop()
```

A

PROJECT REPORT

On

E ASSESSMENT USING IMAGE PROCESSING IN EXAMS

Submitted by

Ms. A. Nageswari (17K81A0562) **Mr. B . Hemanth Aditya (17K81A0565)**

Mr. CH. Naveen (17K81A0568) **Ms. K. Madhurima (17K81A0584)**

in partial fulfillment for the award of the

degree of

BACHELOR OF TECHNOLOGY

IN

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Under the Guidance of

Dr.G.Govinda Rajulu

Associate Professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ST.MARTIN'S ENGINEERING COLLEGE
An Autonomous Institute

Dhulapally, Secunderabad – 500 100

JUNE 2021

BONAFIDE CERTIFICATE

This is to certify that the project entitled **E-Assessment using Image processing**, is being submitted by **Achanta Lakshmi Durga Nageswari 17K81A0562, Boddapu Hemanth Aditya 17K81A0565, Chatla Naveen 17K81A0568**, **K. Madhurima 17K81A0584,** in partial fulfillment of the requirement for the award of the degree of **BACHELOR OF TECHNOLOGY** in **Computer Science Of Engineering** is recorded of bonafide work carried out by them. The result embodied in this report have been verified and found satisfactory.

.

Guide

Dr. G. GOVINDA RAJULU

Department of CSE

Head of the Department

Dr. M. NARAYANAN

Department of CSE

Internal Examiner

External Examiner

Place:

Date:

DECLARATION

We, the student of **Bachelor of Technology** in Department of Computer Science and Engineering', session: 2017 – 2021, St. Martin's Engineering College, Dhulapally, Kompally, Secunderabad, hereby declare that work presented in this Project Work entitled E-Assessment Using Image Processing is the outcome of our own bonafide work and is correct to the best of our knowledge and this work has been undertaken taking care of Engineering Ethics. This result embodied in this project report has not been submitted in any university for award of any degree.

A. Nageswari	17K81A0562
B. Hemanth Aditya	17K81A0565
Ch. Naveen	17K81A0568
K. Madhurima	17K81A0584

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible and whose encouragements and guidance have crowded effects with success.

We extended our deep sense of gratitude to Principal**,** **Dr. P. SANTOSH KUMAR PATRA**, St. Martin's Engineering College, Dhulapally, for permitting us to undertake this project.

We are also thankful to **Dr. M.NARAYANAN**, Head of the Department, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

We would like to thank **Dr. T. POONGOTHAI,** Department of Computer Science and Engineering (AI & ML) for her encouragement and insightful comments and as well as our project coordinators **Dr. B.RAJALINGAM**, Associate Professor and **Dr. GOVINDA RAJULU. G** Associate Professor, in Department of Computer Science and Engineering for their valuable support.

We would like to express our sincere gratitude and indebtedness to our project supervisor **Dr. GOVINDA RAJULU. G,** Associate Professor, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

Finally, we express thanks to all those who have helped us successfully to completing this project. Furthermore, we would like to thank our family and friends for their moral support and encouragement.

We express thanks to all those who have helped us in successfully completing the project.

A. Nageswari 17K81A0562
B. Hemanth Aditya 17K81A0565
Ch. Naveen 17K81A0568
K. Madhurima 17K81A0584

ABSTRACT

The keyword “e-assessment” refers to electronic assessment as software is used to mark the exam papers filled by the students after the exam is completed. Here we are using image processing to accomplish the MCQ correction in very easy manner. It produces the great effort to deal to remove the barriers of multi choice assessment correction. We are here using Open Source Computer Vision Library (Open CV) to process and correct the answers. The usage of Multiple Choice Questions (MCQs) to test the knowledge of a person has been increased gradually. These tests can be evaluated either using OMR technology or manually. In real-time, it is quite difficult to have OMR machine under all circumstances and at the same time, manual correction is time consuming and error prone.These disadvantages have overcame in our proposed system by using digital image processing technique to correct the answers on OMR sheet.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF OUTPUT SCREENS

LIST OF ACRONYMS

CPU	Central Processing Unit
GB	Giga Bytes
GUI	Graphical User Interface
MCQ	Multiple Choices Question
CV	Computer Vision
OMR	Optical Mark Recognition
CDM	Computational Discrete Mathematics
UML	Unified Modeling Language

1.INTRODUCTION

There is a growing need for storing paper-based information digitalized nowadays. This problem concerns education as well but it does not always get enough attention, however using our technology accordingly many aspects of the educational process could be made a lot simpler, easier, faster, more comfortable and automatable.

Most of the educational institutions are using traditional teaching and examination methods in most of their subjects still. Though the digitalization of teaching got a little bit of attention in the previous years and began its growth since then. Alongside it there are also computer-based examination methods but it is not the main functionality of the e-learning systems. So mostly the traditional examination models are used concerning those subjects who require such a way to be examined accordingly. From now on the paper-based examination method will be discussed, since it is the main concern of this paper. The keyword “e-assessment” refers to electronic assessment as software is used to mark the exam papers filled by the students after the exam is completed.

Multiple choices Question (MCQ) are a form of an objective assessment in which respondents are asked to select only correct answers out of the choices from a list. The multiple choice format is most frequently used in educational testing, in market research, and in elections, when a person chooses between multiple candidates, parties, or policies.

In this project we are using image processing to accomplish the MCQ correction in very easy manner. It produces the great effort to deal to remove the barriers of multi choice assessment correction. In this we are using array format to correct the answer paper which is photo copier and uploaded by user. The main concept is to get the image and get the answer which is shadowed by user.

In Python Open CV library is available for image processing. In order to get the best effective output we use the django framework along with python. The Open CV is a library of programming functions mainly aimed at real-time computer vision. The Following topics are organized to explain the process of how to deal with this technique.

1.1. PROJECT OVERVIEW

Image processing is used to accomplish the MCQ correction in very easy manner. It produces the great effort to deal to remove the barriers of multi choice assessment correction. In this we are using array format to correct the answer paper which is photo copier and uploaded by user. The main concept is to get the image and get the answer which is shadowed by user. In Python, Open CV library is available for image processing.

1.2. PROJECT OBJECTIVE

The heart of the project relies on the most important element of the software which is the image processing flow. To image processing to accomplish the MCQ correction in very easy manner. It produces the great effort to deal to remove the barriers of multi choice assessment correction. The proposed system is taking the digital image of the answer sheet in the given pattern and uploads to the given system. This method avoids the machine dependency and people dependency in high manner.

1.3. ORGANIZATION OF CHAPTERS

This documentation consists of 10 different chapter and they are:

1. Introduction – This chapter covers the overview of our project and its objectives.

2. Literature Survey – This includes the details of our survey.

3. Software and Hardware Requirements – We specify our software and hardware requirements here.

4. Software Development Analysis – This section includes the problem definition and details of the modules we used in our project.

5. Project System Design – This chapter includes the design part of our project which includes uml diagrams.

6. Project Coding – This section contains the details of our project code.

7. Project Testing – The details of test cases and testing are included in this chapter.

8. Output Screens – This contains the screenshots of how our project looks like when executed.

9. Experimental Results – This chapter contains the screenshots of our results.

10. Conclusion and Future Enhancements – This covers the conclusion of our project and the possible future developments.

2.LITERATURE SURVEY

2.1 SURVEY ON BACKGROUND

1 Online Testing Suffers Setbacks in Multiple States.

AUTHORS: Davis, Michelle R.

In this paper we attempt to classify published thousands of students experienced slow loading times of test questions, students were closed out of testing in mid-answer, and some were unable to log in to the tests. Hundreds, if not thousands, of tests may be invalidated.The difficulties prompted all four states' education departments to extend testing windows, made some state lawmakers and policymakers reconsider the idea of online testing, and sent district officials into a tailspin.

The testing problems were "absolutely horrible, in terms of kids being anxious," said Eric F. Hileman, the executive director of information technology services for the 43,000-student Oklahoma City schools. Some high school students were taking Oklahoma's high-stakes tests, which require that students pass four out of seven end-of-instruction tests to graduate.

2. Intelligent Assessment Systems for e-Learning.

AUTHORS: Csink, L., Gyorgy, A., Raincsak, Z., Schmuck, B., Sima, D., Sziklai, Z., & Szoll″osi, S.

In this paper we attempt to classify published computer based assessment (CBA) systems using the design space approach. We think that the services of CBA systems need to be standardized and algorithms should be developed accordingly. The authors' pilot system EVITA[7] is also presented as well as the main issues regarding further development.

3. CDM: Teaching Discrete Mathematics to Computer Science Majors

AUTHORS: Klaus Sutner

In this paper we attempt to classify published CDM, for computational discrete mathematics, is a course that attempts to teach a number of topics in discrete mathematics to computer science majors. The course abandons the classical definition-theorem-proof model, and instead relies heavily on

computation as a source of motivation and also for experimentation and illustration. The emphasis on computational issues is particularly attractive to computer science majors and increases their involvement and participation. Categories and Subject Descriptors: K.3.1 [Computers and Education]: Computer Uses in Education; G.2.1 [Discrete Mathematics]: Combinatorics; 1.1.3 [Symbolic and Algebraic Manipulation]: Languages and Systems. CDM, for computational discrete mathematics, is a course that attempts to teach a number of topics in discrete mathematics to computer science majors. The course abandons the classical definition-theorem-proof model, and instead relies heavily on computation as a source of motivation and also for experimentation and illustration. The emphasis on computational issues is particularly attractive to computer science majors and increases their involvement and participation. Categories and Subject Descriptors: K.3.1 [Computers and Education]: Computer Uses in Education; G.2.1 [Discrete Mathematics]: Combinatorics;

[Symbolic and Algebraic Manipulation]: Languages and Systems.

4. Intelligent mathematics assessment in eMax

AUTHOR: György, A., & Vajda, I.

The emergence of the knowledge based society evoked a rapid expansion of student numbers enrolled in higher education. With increasing student numbers, however, the time necessary to assess the acquired knowledge in examinations approaches - or, with large student numbers, even exceeds- the time needed to deliver the lectures of a course. At the same time, distance learning systems have an urgent need to provide an integrated knowledge assessment component. These are the main reasons for the appearance of knowledge assessment systems (KAS's) in the 1970's as well as their steadily increasing importance. As shown in our paper, published KAS's do not support the evaluation of either freely formulated short answers or partially solved mathematical problems. However, the eMax KAS, developed at the Intelligent Knowledge Management Innovation Center of IBM Hungary and the John von Neumann Faculty of Informatics at Budapest Tech was conceived to fill this void. The short description of eMax and the method of quasi-automatic assessment of short texts can be found in a companion paper of the present conference. In this article we will demonstrate how eMax is capable of quasi-automatic assessment when answers are only partially correct, using questions of higher mathematics.

5. Intelligent short text assessment in eMax.

AUTHORS: Sima, D., Schmuck, B., Szöll˝osi, S., & Miklós, Á.

Rapidly increasing student numbers and spreading distance learning systems strengthen the urgent need for effective knowledge assessment systems (KAS's). Recent KAS's have however, the deficiency of not providing intelligent assessment modules for eg the evaluation of freely formulated short answers including a few sentences or partially solved mathematical problems. The eMax KAS, developed at the Intelligent Knowledge Management Innovation Center of IBM Hungary and the John von Neumann Faculty of Informatics at Budapest Tech, aims to provide these capabilities. Our paper gives an introduction to the intelligent assessment of short texts component of eMax by presenting the approach used for the formal description of the "answer space" defined as well as the methods chosen for the syntactic analysis, semantic analysis and scoring. The a version of eMax is now completed and is in testing.

2.2. CONCLUSION ON SURVEY

There are currently many commercial solutions for automated scoring of Multiple-Choice Tests, usually composed of a software and a scanner, but the widespread dissemination of laptops, tablets and smartphones with built in cameras offers new possibilities for doing the same job with no need for any extra hardware. This article presents a simple and innovative method to transform captured images of answer sheets into reduced binary matrices containing answers to the questions plus some control elements, using simple morphological operations for segmentation. This methodology is applied to the real problem of automatic correction of Multiple-Choice Tests. Initially, the user positions the answer key sheet in front of the camera in order to save the image to the disk, then the image is gauged to evaluate the test type. Subsequently, student answer sheets can be read using the camera, having the test score displayed on the screen and/or saved to a file.

3.SOFTWARE AND HARDWARE REQUIREMENT

The project involved analyzing the design of few applications so as to make the application more users friendly. To do so, it was really important to keep the navigations from one screen to the other well ordered and at the same time reducing the amount of typing the user needs to do.

3.1. SOFTWARE REQUIREMENTS

For developing the application the following are the Software Requirements:

Operating Systems supported: Windows 10

Technologies and Languages used to Develop: Python

3.2. HARDWARE REQUIREMENTS

For developing the application the following are the Hardware Requirements:

- Processor: I3 Processor
- RAM: 4 GB
- Space on Hard Disk: 500 GB

4.SOFTWARE DEVELOPMENT ANALYSIS

4.1. OVERVIEW OF PROBLEM

The Main disadvantage is the necessity of the OMR machine to correct the answer. On the other hand the manual work is heavy and problems to deal with accuracy, time delay and people management**.**And toUpload scanned images of a given exam, automatic image processing of these files and preparing them to be corrected by the teachers.

4.2. DEFINE THE PROBLEM

The digital image of the answer sheet is taken in the given pattern and uploads to the given system. In order to correct the answer digital image processing is used to get the answer sheet and proceed it to read the image. This method avoids the machine dependency and people dependency in high manner. This system brings out the effectiveness by image processing technique. The major impact is open CV library, which is available to access the image to get it as a matrix and make that very effective to deal with correcting answers in the image.

4.3. MODULES OVERVIEW

This project has two users Faculty and Student and five modules are designed for the interactions between users and application. Each Module has its own functionality. A module allows us to logically organize the code. Grouping related code into a module makes the code easier to understand and use.

4.4. DEFINE THE MODULES

This application has four modules which are listed in the following.

1. Student Management
2. Exam Assessment
3. Result Details
4. Graph Analysis

4.5. MODULE FUNCTIONALITY

In this project there are four modules to achieve our expected result. These are the major functionalities of the project. The registration and login process are important to access the project for both users. There are two users' admin (Teacher) and user (Student).

1. Student Management

The students are not directly registered. Faculty is uploading the bulk details of students with details of name, student id, class and so on. Students will receive manually student id from faculty manually. With the username and student id as password, student can authenticate to access the details. The details can be modified by students not by faculty at the same time student cannot modify their student id which given to them. The students are given papers to attempt the exam .Every paper contains some essential information about the given exam and about the one who took the exam on the given paper. All of this is represented in the header of the paper where a QR code can be found containing information about the exam and six cells as well where the students must put their identification codes. Also, if there are different groups of papers it is also represented in the QR code but the teachers are not expected to scan each and every one of these so instead it is represented as a big Latin letter beside the QR code

2. Evaluation using Image Processing

After logging in as a teacher the user can create the exam papers in some easy steps, during this process all essential information regarding the given exam will be recorded and using these, the special exam papers will be created for the teacher to print them and handle them to the students. After the exam is completed and the students filled their papers those can be scanned, it is recommended to use a document feeder to do so since it can speed up the process quite a lot. The quality of the scanned images has a minimum requirement which is low and exactly is as follows: 1448 × 2048 pixel grey scale image. Better quality can also be used but it will unnecessarily slow down the scanning and later the image processing processes as well. The next step is to upload the files which can be easily done by some clicks on the user interface. The Faculty will upload the students answer sheets as photos. Those photos can be evaluated with the help of Digital Image Processing technique. It can be achieved with the help of python's opencv library. The matrix form is created with answer key to identify and give the result as per the photos. The uploaded images will be processed right during their upload. Once all the selected images are uploaded and processed the teachers can begin the correction of these exams on the given user interface. When they are done, the results are ready to be published and the students can view and react to these of course.

.

3. Result Analysis

The results from the above module are handled by some math functions to put those values into calculations. Get the total marks accomplished by students and average of the student can be calculated by the auto functionalities and display to users.

4. Graph Analysis

The graph analysis is done by the values taken from the result analysis part and it can be analyzed by the graphical representations. Such as pie chart, pyramid chart and funnel chart here in this project.

5. PROJECT SYSTEM DESIGN

5.1. SYSTEM ARCHITECTURE

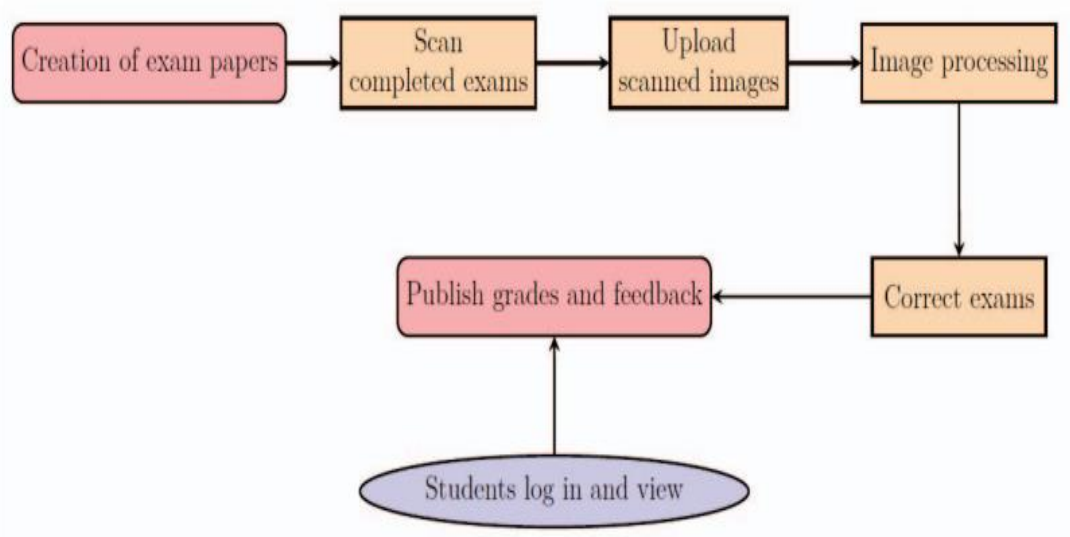


Fig.5.1: User Functional Model Image

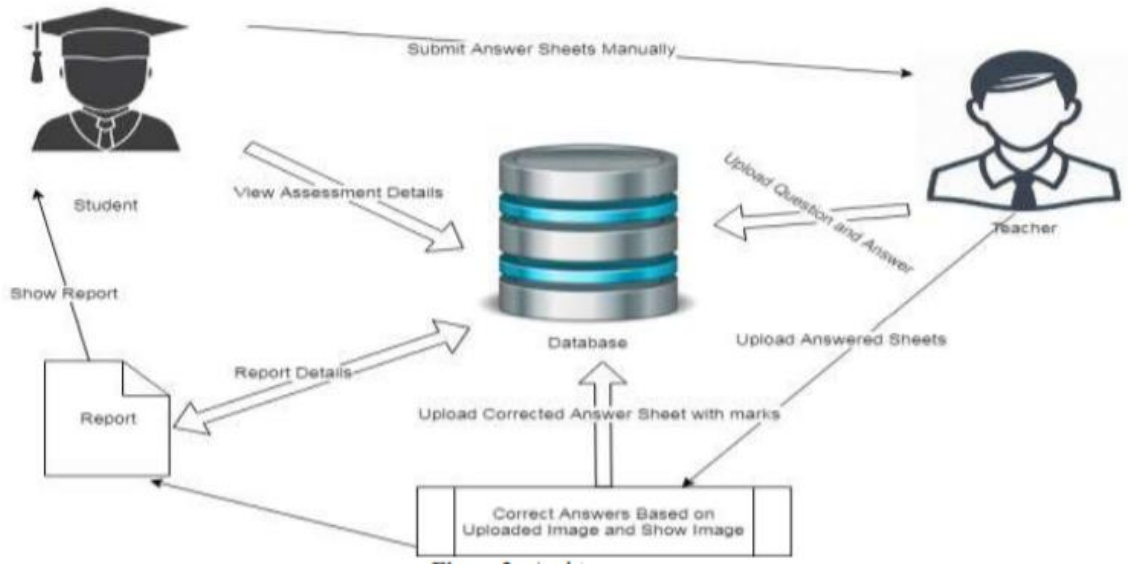


Fig.5.2 : The Main Steps Of Image Processing Flow

5.2. UML DIAGRAMS

UML stands for Unified Modeling Language. UML is a standardized general-purpose modeling language in the field of object-oriented software engineering. The standard is managed, and was created by, the Object Management Group.

The goal is for UML to become a common language for creating models of object oriented computer software. In its current form UML is comprised of two major components: a Meta-model and a notation. In the future, some form of method or process may also be added to; or associated with, UML.

The Unified Modeling Language is a standard language for specifying, Visualization, Constructing and documenting the artifacts of software system, as well as for business modeling and other non-software systems.

The UML represents a collection of best engineering practices that have proven successful in the modeling of large and complex systems.

The UML is a very important part of developing objects oriented software and the software development process. The UML uses mostly graphical notations to express the design of software projects.

USE CASE DIAGRAM:

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.

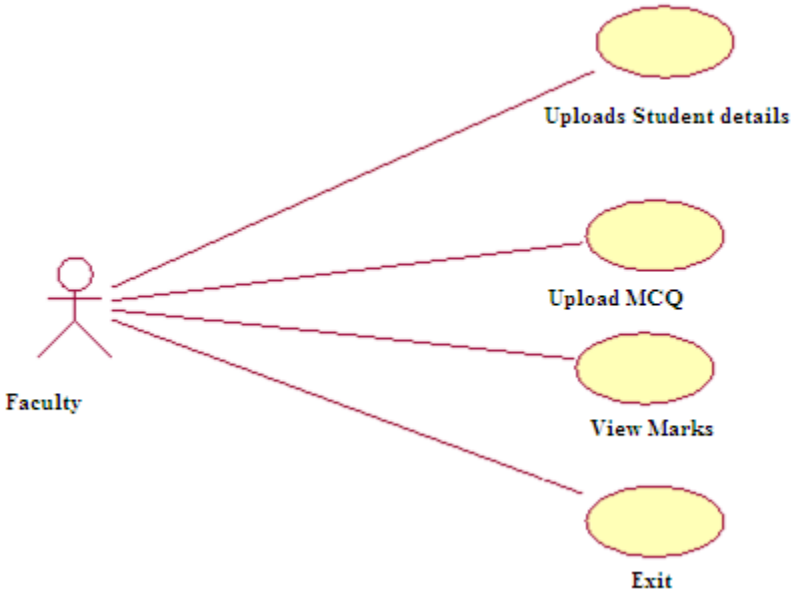


Fig. 5. 3: Use Case Diagram

CLASS DIAGRAM:

In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.

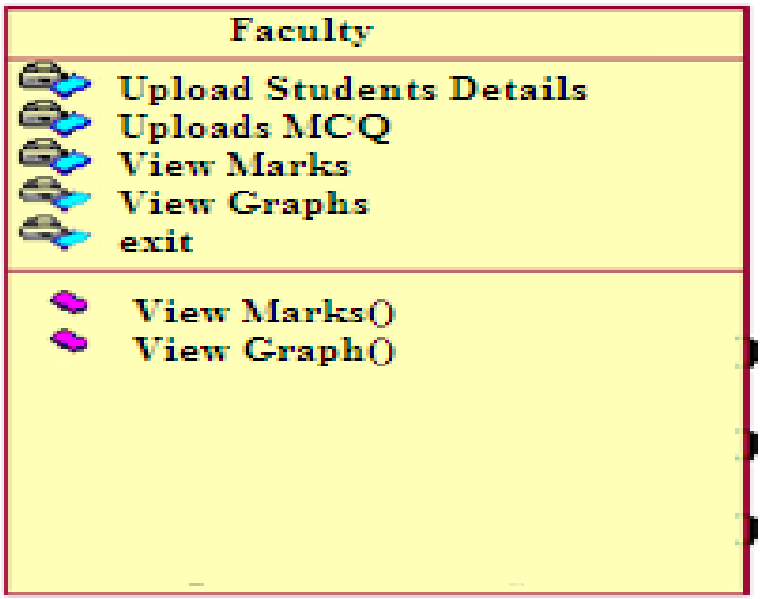


Fig. 5.4: Class Diagram

SEQUENCE DIAGRAM:

A sequence diagram in Unified Modeling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.

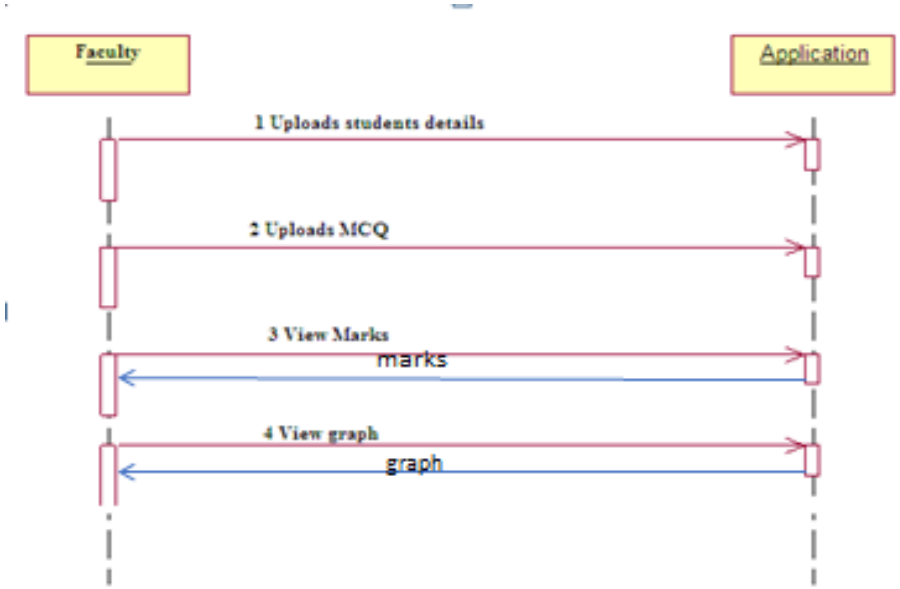


Fig. 5. 5: Sequence Diagram

ACTIVITY DIAGRAM:

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modeling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.

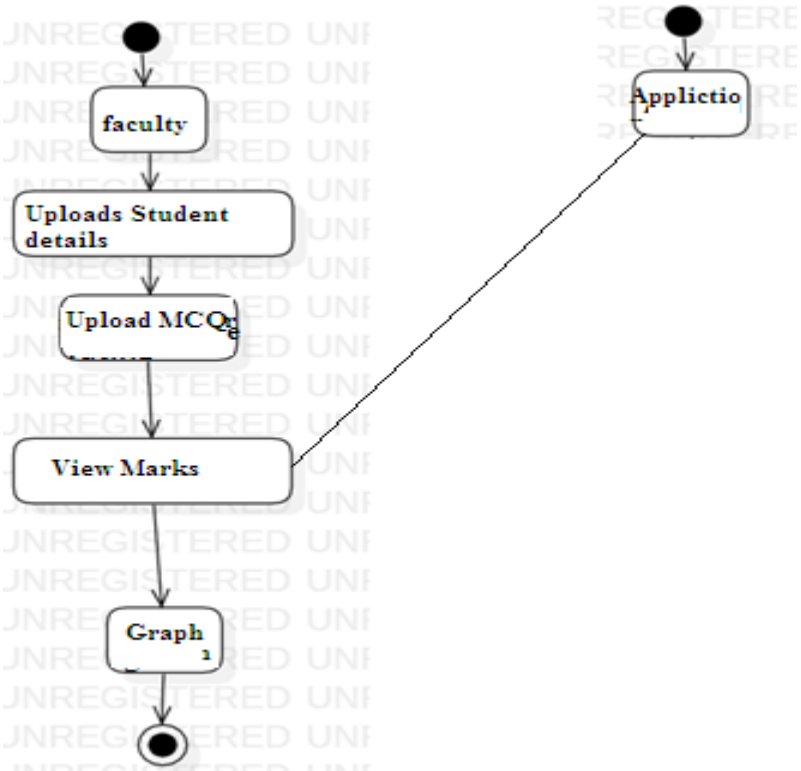


Fig. 5. 6: Activity Diagram

OBJECT DIAGRAM:

An object diagram is a UML structural diagram that shows the instances of the classifiers in models. Object diagrams use notation that is similar to that used in class diagrams. Class diagrams show the actual classifiers and their relationships in a system.

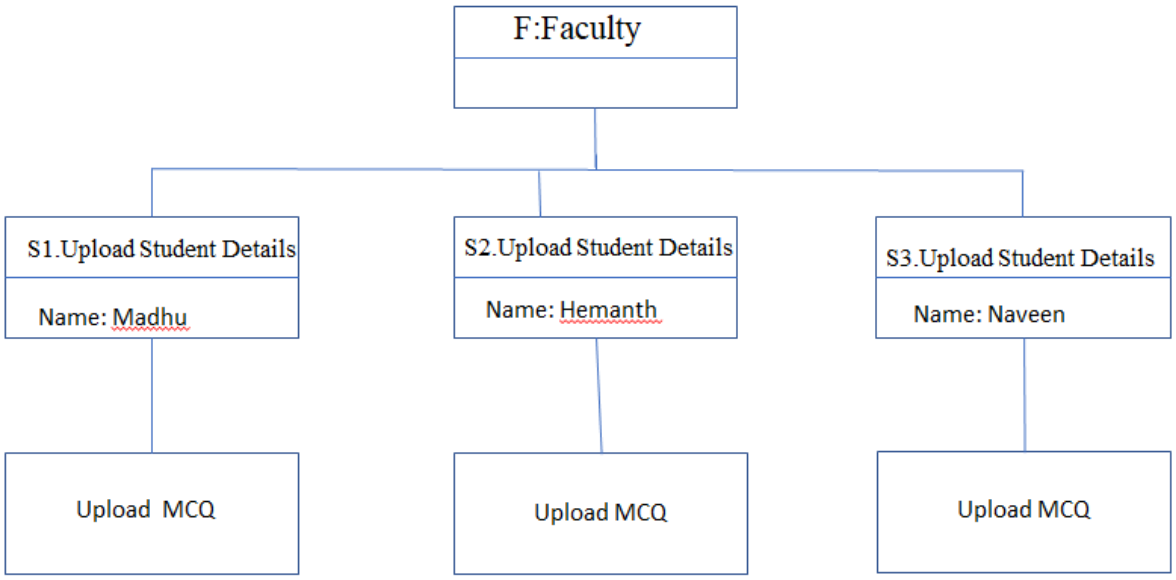


Fig. 5.7: Object Diagram

COMMUNICATION DIAGRAM

UML Communication Diagrams, previously known as collaboration diagrams are a type of behavioural diagram that shows the interactions that take place between objects in a piece of software or system. This type of diagram emphasizes the messages exchanged between objects.

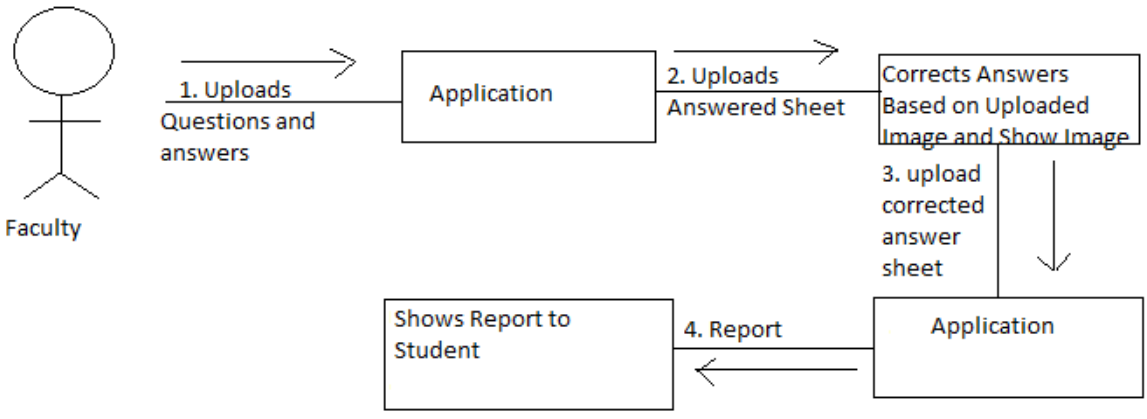


Fig. 5.8: Communication Diagram

6. PROJECT CODING

6.1. CODE TEMPLATES

```
# Import required modules

# Declare global variables

def uploadPaper():

    # Dialog box appears for entering the Student Details and uploading the OMR

def loadPaper():

    # Uploaded images will be processed.

def viewmarks():

    # Student can view their result

def marksGraph():

    # Graph representing Student results

def  login():

    # login into the faculty account

# Create buttons for all the methods used

main.config()

main.mainloop()
```

6.2. OUTLINE FOR VARIOUS FILES

We used Python programming to implement our project. A single python file is used to implement our code. This file consists of various modules that we have used. Our project modules are - Student Management, Exam Assessment, Result Details, Graph Analysis. We also used various python modules like tkinter, matplotlib, numpy, imutils, os, cv2..

6.3. METHODS INPUT AND OUTPUT PARAMETERS

In our project code, we implemented five different methods. They are:

1. uploadPaper()

2. loadPaper()

3. viewmarks()

4. marksGraph()

5. login()

Our first method uploadPaper() takes student details and OMR as an input". Second method loadPaper() takes input from uploadPaper module and does the image processing and gives the result. viewmarks() doesn't have any input parameters.marksgraph() also don't have any input parameters but it displays a graph showing result of the students. Login() takes username and password as parameters and shows the user interface for faculty as a result.

7.PROJECT TESTING

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

7.1. VARIOUS TEST CASES

Unit testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

Integration testing

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

Functional test

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items:

Valid Input : identified classes of valid input must be accepted.

Invalid Input : identified classes of invalid input must be rejected.

Functions : identified functions must be exercised.

Output : identified classes of application outputs must be exercised.

Systems/Procedures : interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

System Test

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

7.2. BLACK BOX

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box .you cannot “see” into it. The test provides inputs and responds to outputs without considering how the software works.

7.3. WHITE BOX TESTING

White Box Testing is a testing in which in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is purpose. It is used to test areas that cannot be reached from a black box level.

8. OUTPUT SCREENS

8.1. USER INTERFACES

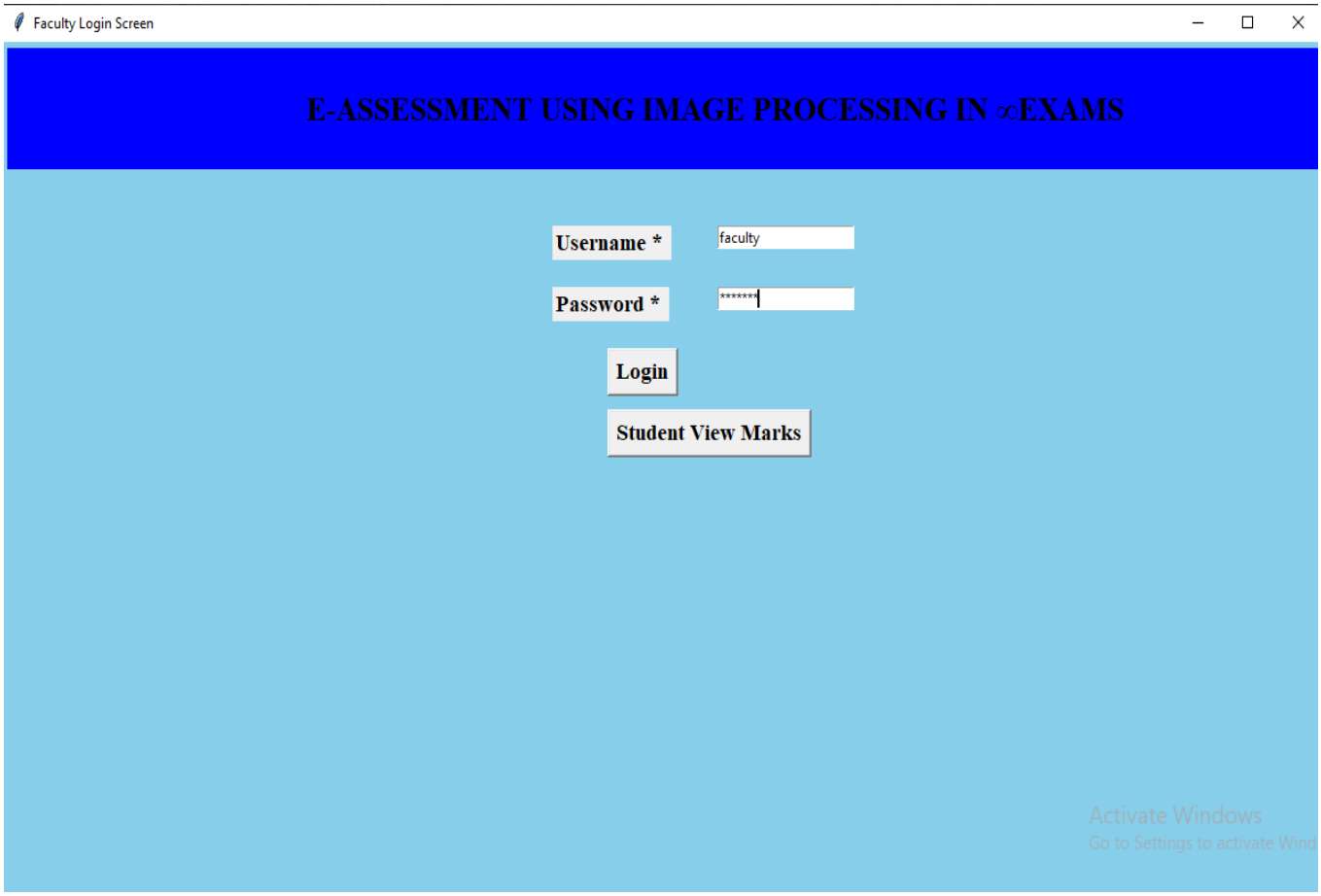


Fig. 8.1: Faculty User Interface

The above figure shows the faculty user interface where the faculty can login with their user name and password to perform the functions like upload paper and view marks graph

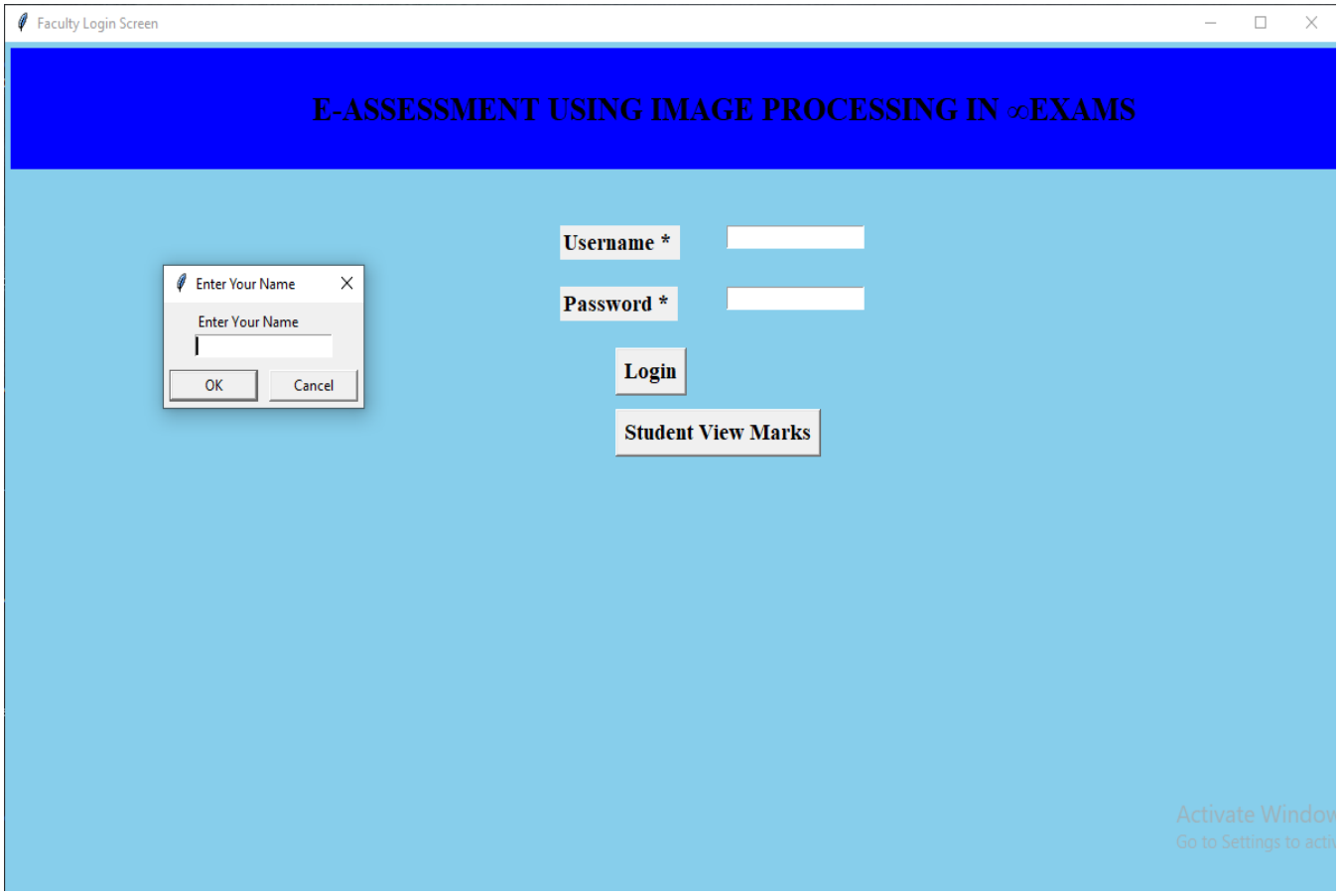


Fig. 8.2: Student Login Interface

In above screen click on 'Student View Marks' button to get dialog box ,In dialog box enter student name to get his marks.

8.2. OUTPUT SCREENS

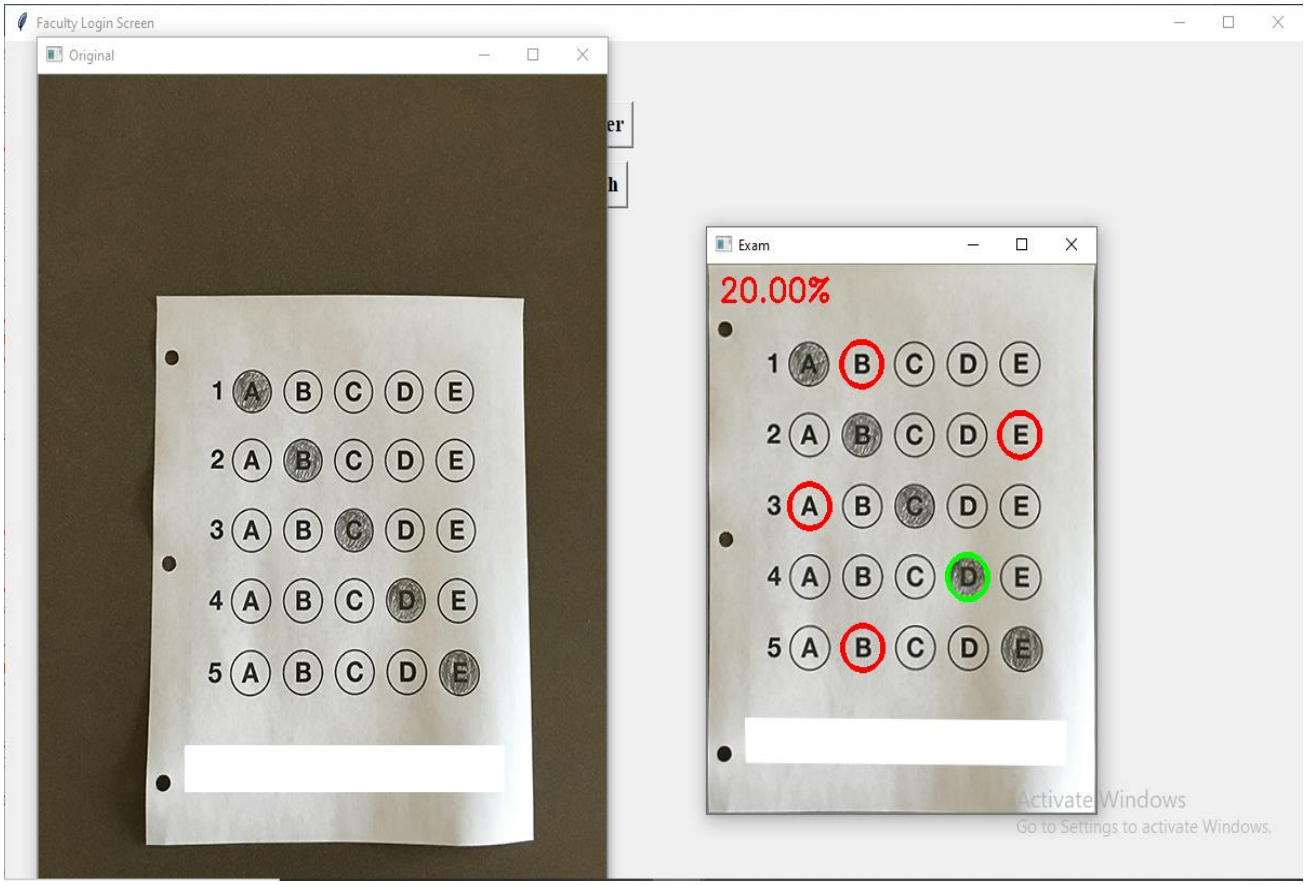


Fig. 8.3: Output Screen

In above screen we can see student answers at left side image and corrected image at right side where wrong answer mark with red colour. Similarly we can upload any number of images. Each image must have five questions and left side paper must have three dot symbols.

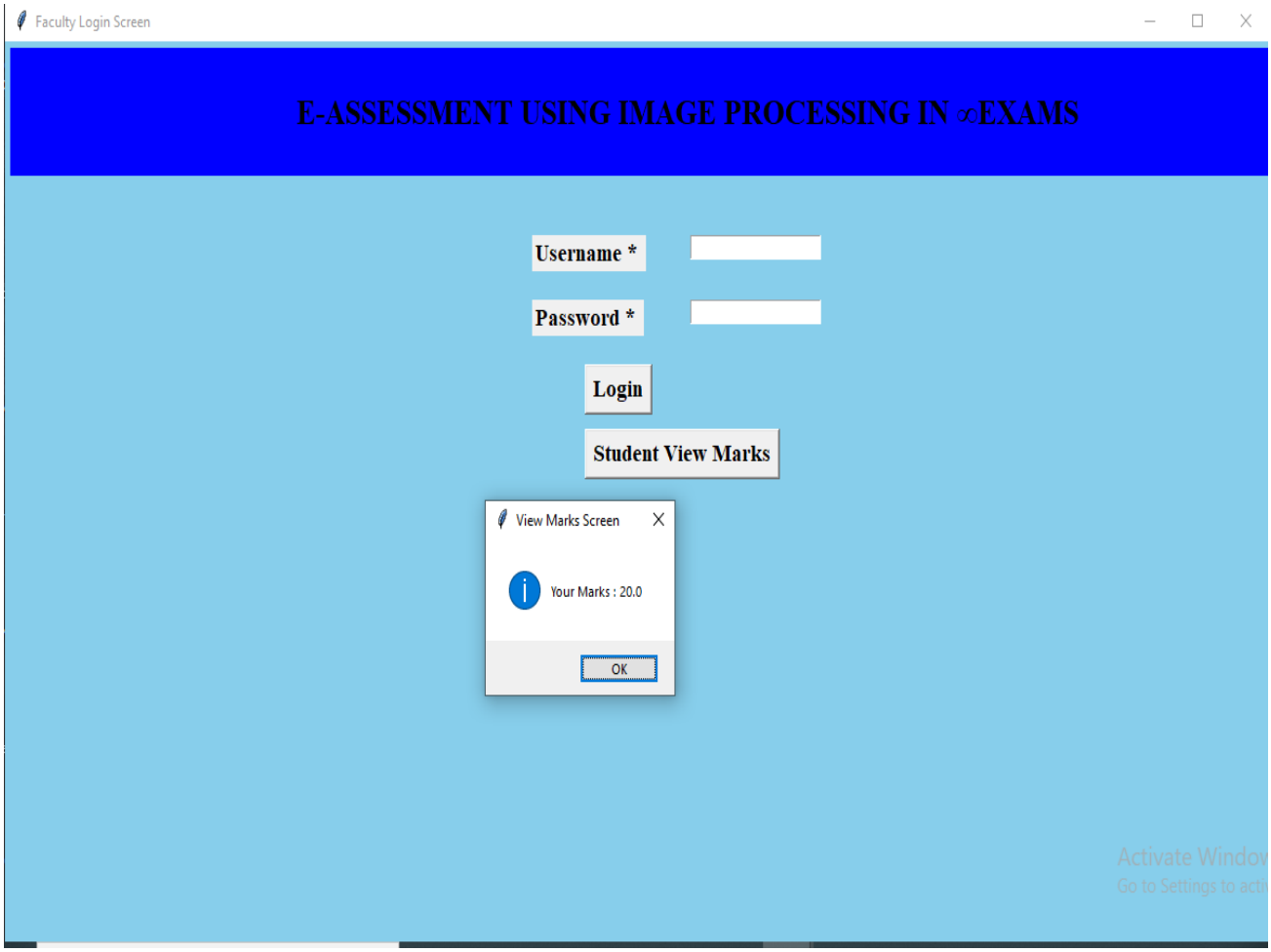


Fig. 8.4: Student Result

The Above Figure shows the result of the student,Student can view their marks by clicking on view Student marks and entering their unique ID given to them.

9.EXPERIMENTAL RESULTS

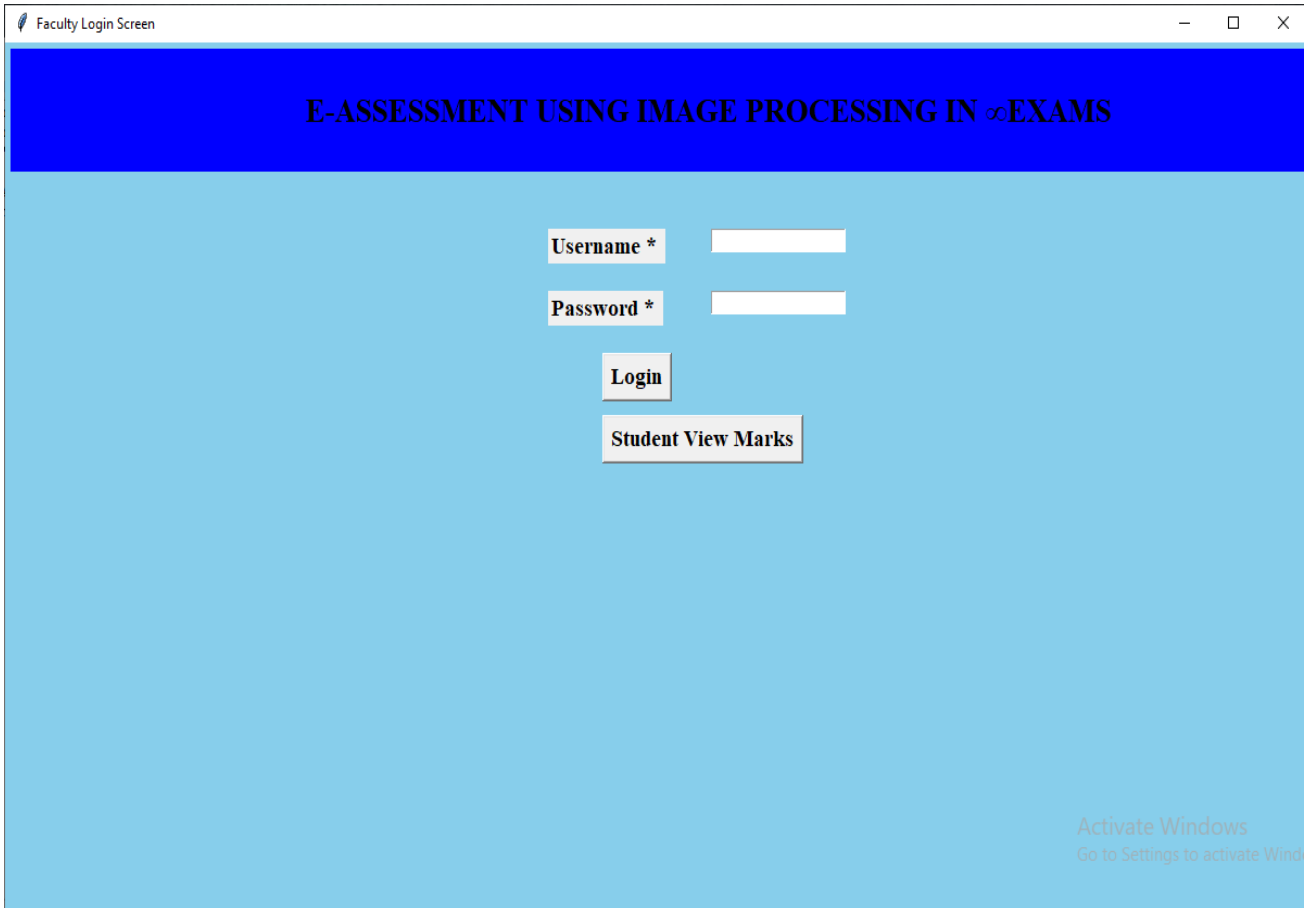


Fig. 9.1: Login Page

Fig.9.1 Shows the login page where there are two features for the faculty and the student respectively. Faculty can login by their username and password for uploading the test papers , whereas the students can view their marks by using their unique id which is given to them

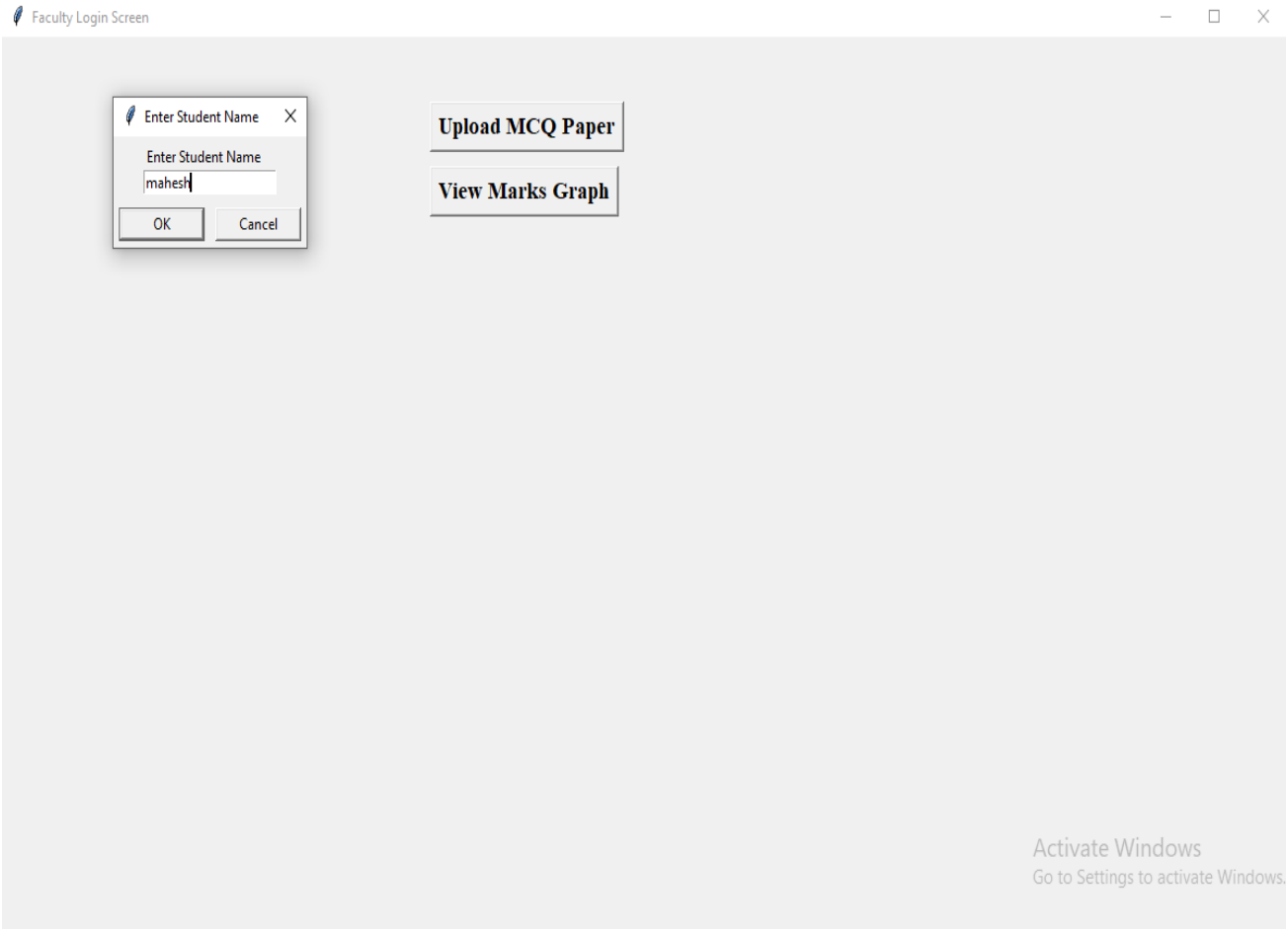


Fig. 9.2: Entering Student Details

Fig.9.2 is the image we get when faculty login and selects the option upload MCQ paper where the faculty has to enter the name of student or the ID and has to upload the image

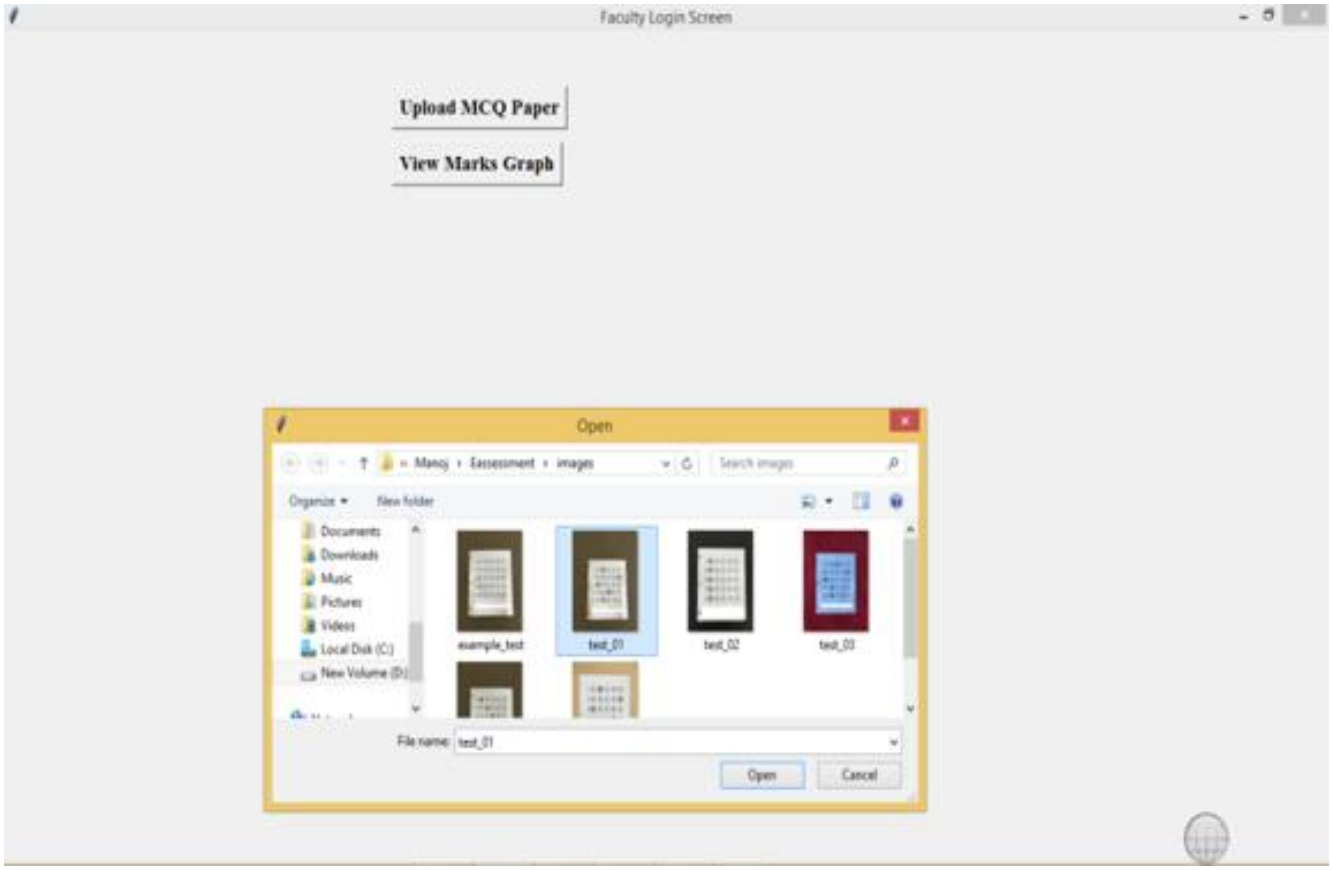


Fig. 9.3: Uplaoding OMR Sheet

Fig 9.3 shows the uploading of OMR sheet ,where the faculty need to choose the student and has to upload the OMR.

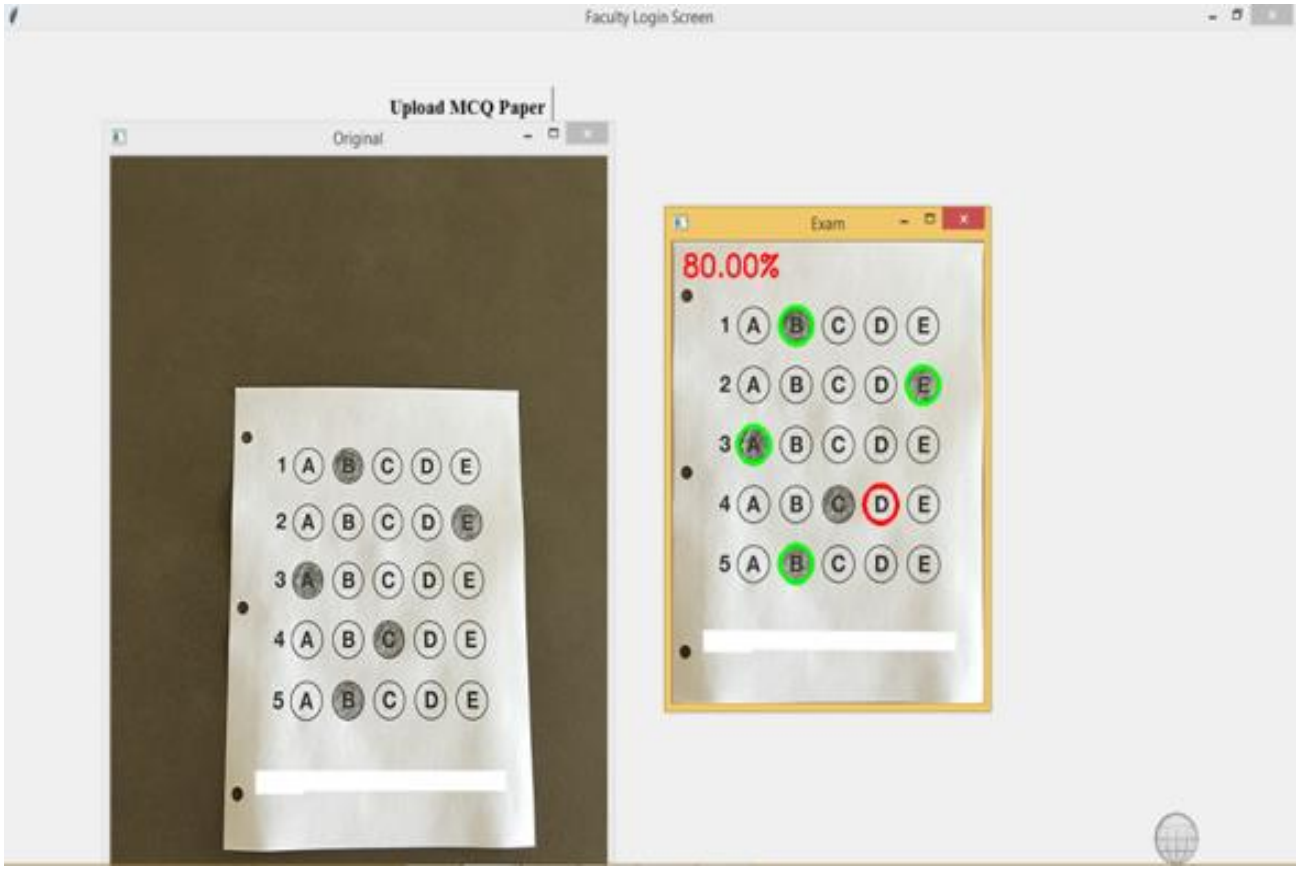


Fig. 9.4: Result Of the uploaded OMR

The above figure shows the Evaluation of OMR, where two images are showm.The first image is the original answer script which is uploaded by the faculty. The second image adjacent to it shows the score and answers for them

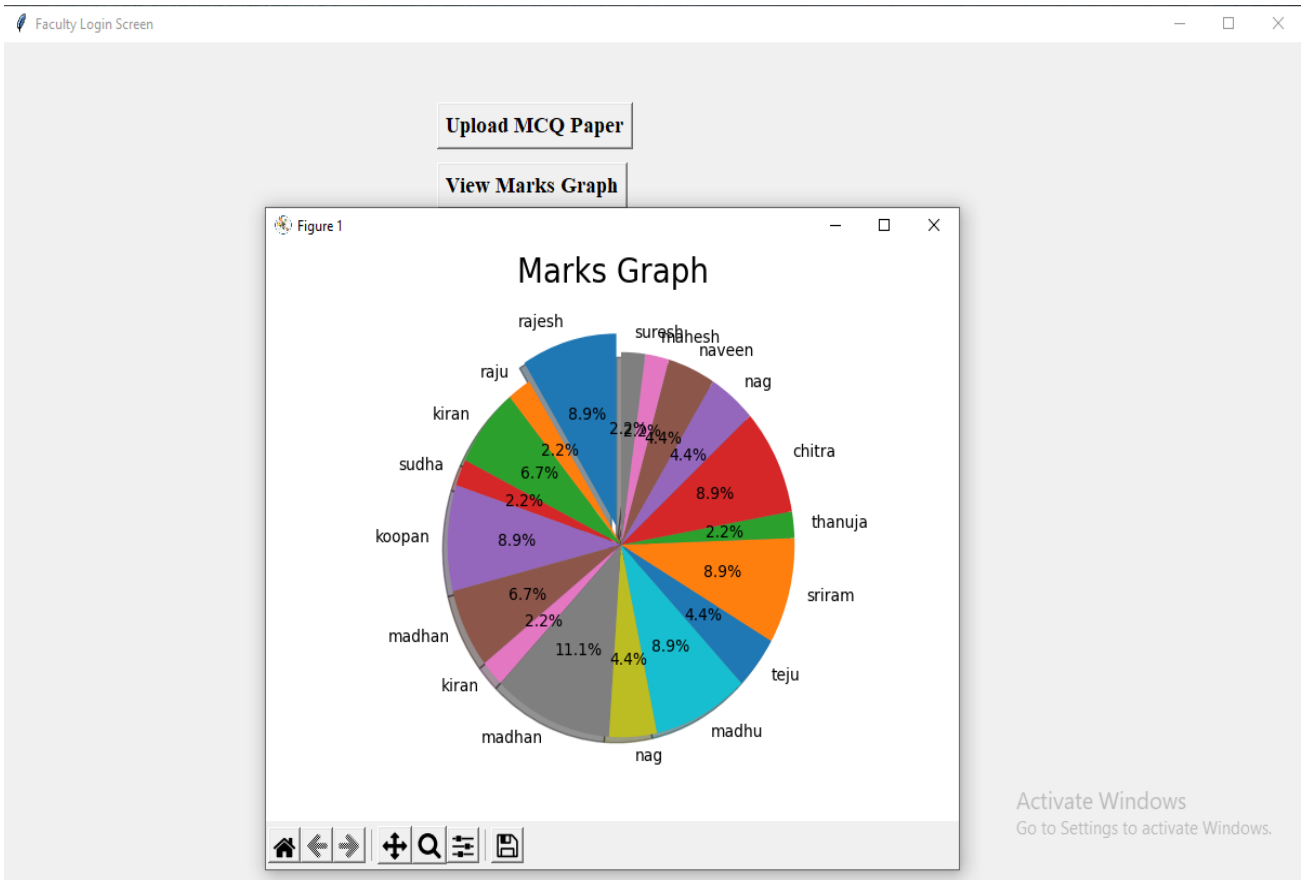


Fig. 9.5: View Marks Graph

Fig 9.5 Shows the graphical representation of the students result with the help of pie chart.. In the left bottom where we have many option's like to save the figure , Zoom the picture , and then you can the previous view and we can also forward to the next view , and at last there is option for the reset view.

10. CONCLUSION AND FUTURE ENHANCEMENT

The featured so-called E-Assessment is taking the digital image of the answer sheet in the given pattern and uploads to the given system. MCQ Test format have different way of correction and conducting the tests which is very difficult. The proposed system addresses the issue and solving this problem with the help of Image Processing.

The image processing part of the system has given satisfactory results as it seemed fast enough to process even a massive number of images at once without a single error. It is beyond doubt that the further development potential of the software system is great and by seizing this opportunity, when it will be completed and released, it could play a considerable role in the future of the revolution of the digitalization of education.

In future we can able to have many services to be included in this application.

- We will improve this project by giving ‘n’ student OMR sheets into a folder and upload the folder at once and try to get the result.
- Giving set of keys at a time(Set 1, Set 2,...)
- Giving scope for faculty for selecting number of choices (like 2,3, 4,...) for the specific paper.

REFERENCES

[1] Davis, Michelle R. "Online Testing Suffers Setbacks in Multiple States." Education Week 32.30 (2019): 1-18.

[2] István Vajda, "Computer Aided Teaching of Discrete Mathematics and Linear Algebra", University of Debrecen, PhD Thesis (2018).

[3] Csink, L., György, A., Raincsák, Z., Schmuck, B., Sima, D., Sziklai, Z., & Szöll″osi, S. "Intelligent assessment systems for e-learning." Proc. of the 4-th European Conference on E-Activities, ECOMM-LINE 2003. (2013).

[4] György, A., & Vajda, I. "Intelligent mathematics assessment in eMax." AFRICON 2007. IEEE (2017).

[5] Sima, D., Schmuck, B., Szöll″osi, S., & Miklós, Á. "Intelligent short text assessment in eMax." Towards intelligent engineering and information technology. Springer Berlin Heidelberg (2019): 435-445.

[6] Keady, G., Fitz-Gerald, G., Gamble, G., & Sangwin, C. "Computer-aided assessment in mathematical sciences." Proceedings of The Australian Conference on Science and Mathematics Education (formerly UniServe Science Conference). (2017).

[7] Hendriks, Remco. "Automatic exam correction." UVA Universiteit van Amsterdam (2019).

[8] de Assis Zampirolli, Francisco, José Artur Quilici Gonzalez, and Rogério Perino de Oliveira Neves. "Automatic Correction of Multiple-Choice Tests using Digital Cameras and Image Processing." Universidade Federal do ABC (2019).

[9] Llamas-Nistal, M., Fernández-Iglesias, M. J., GonzálezTato, J., & Mikic-Fonte, F. A. "Blended e-assessment: Migrating classical exams to the digital world." Computers & Education 62 (2018): 72-87.

[10] Duda, Richard O., and Peter E. Hart. "Use of the Hough transformation to detect lines and curves in pictures." Communications of the ACM 15.1 (2017): 11-15.

[11] Otsu, Nobuyuki. "A threshold selection method from gray-level histograms." Automatica 11.285-296 (2018): 23-27.

[12] Soille, Pierre. "On morphological operators based on rank filters." Pattern recognition 35.2 (2018): 527-535.

[13] Deodhare, Dipti, NNR Ranga Suri, and R. Amit. "Preprocessing and Image Enhancement Algorithms for a Form-based Intelligent Character Recognition System." IJCSA 2.2 (2015): 131-144.

[14] Eikvil, Line. "Optical character recognition." citeseer.ist.psu.edu/142042.html (2017).

PUBLICATION

CONFERENCE :

- Submitted Conference Paper to INTERNATIONAL CONFERENCE ON INNOVATIONS IN COMPUTER NETWORKS, COMPUTATIONAL INTELLIGENCE AND IOT [ICICCI-2021].
- Paper ID : ICICCI – 21 – 0078

ONE PAGE PROFILE

1. A. NAGESWARI

Achanta Lakshmi Durga Nageswari is currently pursuing her Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. She completed her intermediate from Sri Chaitanya Junior College and 10th class from St. Martin's High School. Her technical skills include C, Python and Java. She also has a basic understanding of C++. She took part in Employability Skill development Program conducted by Zensar. Her participations include: Workshop on "HTML & CSS" which was conducted on 5th January 2018, Technical Treasure hunt and Paper Presentation in "Sympo aagnya 2020-A Two Day National Level Technical Symposisum" Which was conducted on 30th and 31st January, "Know More - Teach More ", the Global Webinar on Cyber Threats and Defense Techniques conducted by GECF on 22nd July 2020, National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020, Women online workshop on "Women in Cyber Security and Privacy in 2020" which was conducted from 6th to 10th July 2020. Online Sessions conducted by Institution's Innovation Council (IIC) of MHRD's Innovation Cell, New Delhi to promote Innovation, IPR, Entrepreneurship, and Start-ups among HEIs from 28th April to 22nd May 2020. Two-Day National Level Seminar On "Recent Trends in Cloud Computing, Fog and Edge Computing" scheduled on 18th June to 19th June 2021. Her areas of interest are Python,Machine learning. She completed few certification courses from online platforms like Coursera and CursaApp

2. B. HEMANTH ADITYA

Boddapu Hemanth Aditya is currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. He completed his intermediate from Narayana Junior College and 10th class from Triveni Talent School. His technical skills include C and Python. He also has a basic understanding of C++ and Java. He took part in Employability Skill development Program conducted by Zensar. His participations include: Two-Day National Level Seminar On "Recent Trends in Cloud Computing, Fog and Edge Computing" scheduled on 18th June to 19th June 2021 , National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020 , "Two Day Entrepreneurship Summit" jointly organized by NucleusTech and SUMVN, "IIC Online Sessions" conducted by Institution's Innovation Council (IIC) of MHRD's Innovation Cell, New Delhi to promote Innovation, IPR, Entrepreneurship, and Start-ups among HEIs from 28th April to 22nd May 2020 , The Guinness World Record Event- Most users to take an online computer programming lesson in 24 hours from April 24th 2021, 6PM to 25th April 2021, 6PM conducted by GUVI, Build a Face Recognition Application using Python as part of AI-For-India Event conducted by GUVI. His areas of interest are Python, Cyber Security. He completed few certification courses from online platforms like Coursera, CursaApp, Udemy, SoloLearn and also successfully completed the course ICSI | CNSS Certified Network Security Specialist.

3. CH. NAVEEN

Chatla Naveen is currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin’s Engineering College,Secunderabad. He completed his intermediate from NarayanaJunior College,Hyderabad and 10thclass from St. Thomas High School,Nirmal. His technical skills include C andPython. His participations include: Two-Day National Level Seminar on " Recent Trends in Cloud Computing , Fog And Edge Computing conducted on 18th june to 19th june 2021. " Build a Face Recognition Application using Python as part of Al-For-India Event" on April 25, 2021. "The Guinness World Record Event - Most users to take an online computer programming lesson on April 24tg to 25th,2021". IICOnlineSessions conducted by Institution's Innovation Council (IIC) of MHRD's Innovation cell, New Delhi to promote Innovation, IPR, Entrepreneurship, and Start-ups among HEIs from 28thApril to 22nd May 2020. His areas of interest are Python, Web Development. He completed few certification courses from online platforms like Coursera and Guvi.

4. K. MADHURIMA

K Madhurima is currently pursuing her Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College, Dhulapally . She completed her intermediate from Narayana College and 10th class from St Anns's High School. She actively participates in all the events held in the college . Her technical skills include C, Python and Java. She also has a basic understanding of C++. She also took part in technical seminars, technovation . Her participations include: National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020, "Know More - Teach More ", the Global Webinar on Cloud & Big Data – Changing the way we work which was conducted Global Education & Careers Forum (GECF) on 12th August 2020, Women online workshop on "Women in Cyber Security and Privacy in 2020" which was conducted from 6th to 10th July 2020, "Know More - Teach More ", the Global Webinar on Cyber Threats and Defense Techniques conducted by GECF on 22nd July 2020, She took part in Employability Skill development Program conducted by Zensar. Her participations include: Workshop on "HTML & CSS" which was conducted on 5th January 2018, Technical Treasure hunt and Paper Presentation in "Sympo aagnya 2020-A Two Day National Level Technical Symposisum" Which was conducted on 30th and 31st January,

Her areas of interest are Python, Artificial Intelligence, Machine Learning and Deep Learning. She completed few certification courses from online platforms like Coursera, CursaApp and SoloLearn.

APPENDICES

```
from tkinter import messagebox

from tkinter import *

from tkinter import simpledialog

import tkinter

from tkinter.filedialog import askopenfilename

from tkinter import simpledialog

from imutils.perspective import four_point_transform

from imutils import contours

import numpy as np

import argparse

import imutils

import cv2

import matplotlib.pyplot as plt

import time

import numpy as np

import ntpath

import os

confid = 0.5

thresh = 0.5

mouse_pts = []

main = tkinter.Tk()

main.title("Faculty Login Screen")

main.geometry("1200x1200"

global paper
```

```
global newwin
ANSWER_KEY = {0: 1, 1: 4, 2: 0, 3: 3, 4: 1}
global password_login_entry
global username_login_entry
global username_verify
global password_verify
username_verify = StringVar()
password_verify = StringVar()
def uploadPaper():
    global newwin
    sname = simpledialog.askstring("Enter Student Name", "Enter Student Name",parent=newwin)
    global paper
    paper = askopenfilename(initialdir = "images")
    print("[INFO] accessing image ..paper .",paper)
    print(ntpath.basename("images"))
    head, tail = os.path.split(paper)
    print(head)
    print(tail)
    image = cv2.imread(tail,1)
    window_name = 'image'
    #cv2.imshow(window_name, image)
    #cv2.waitKey(0)
    gray = cv2.cvtColor(image, cv2.COLOR_RGB2GRAY)
    blurred = cv2.GaussianBlur(gray, (5, 5), 0)
    edged = cv2.Canny(blurred, 75, 200)
```

```
cnts = cv2.findContours(edged.copy(),cv2.RETR_EXTERNAL,cv2.CHAIN_APPROX_SIMPLE)
cnts = imutils.grab_contours(cnts)
docCnt = None
if len(cnts) > 0:
  cnts = sorted(cnts, key=cv2.contourArea, reverse=True)
  for c in cnts:
    peri = cv2.arcLength(c, True)
    approx = cv2.approxPolyDP(c, 0.02 * peri, True)
    if len(approx) == 4:
      docCnt = approx
      break
paper = four_point_transform(image, docCnt.reshape(4, 2))
warped = four_point_transform(gray, docCnt.reshape(4, 2))
thresh = cv2.threshold(warped, 0, 255,cv2.THRESH_BINARY_INV | cv2.THRESH_OTSU)[1]
cnts = cv2.findContours(thresh.copy(),cv2.RETR_EXTERNAL,cv2.CHAIN_APPROX_SIMPLE)
cnts = imutils.grab_contours(cnts)
questionCnts = []
for c in cnts:
  (x, y, w, h) = cv2.boundingRect(c)
  ar = w / float(h)
  if w >= 20 and h >= 20 and ar >= 0.9 and ar <= 1.1:
    questionCnts.append(c)
questionCnts = contours.sort_contours(questionCnts,method="top-to-bottom")[0]
correct = 0
for (q, i) in enumerate(np.arange(0, len(questionCnts), 5)):
```

```
        cnts = contours.sort_contours(questionCnts[i:i + 5])[0]
        bubbled = None
        for (j, c) in enumerate(cnts):
            mask = np.zeros(thresh.shape, dtype="uint8")
            cv2.drawContours(mask, [c], -1, 255, -1)
            mask = cv2.bitwise_and(thresh, thresh, mask=mask)
            total = cv2.countNonZero(mask)
            if bubbled is None or total > bubbled[0]:
                bubbled = (total, j)
        color = (0, 0, 255)
        k = ANSWER_KEY[q]
        if k == bubbled[1]:
            color = (0, 255, 0)
            correct += 1
        cv2.drawContours(paper, [cnts[k]], -1, color, 3)
    score = (correct / 5.0) * 100
    f = open("marks.txt", "a+")
    f.write(sname+","+str(score)+"\n")
    f.close()
    print("[INFO] score: {:.2f}%".format(score))
    cv2.putText(paper, "{:.2f}%".format(score), (10, 30),cv2.FONT_HERSHEY_SIMPLEX, 0.9, (0,
0, 255), 2)
    cv2.imshow("Original", image)
    cv2.imshow("Exam", paper)
    cv2.waitKey(0)
```

```
def marksGraph():
    names = []
    marks = []
    explode = []
    with open("marks.txt", "r") as file:
      for line in file:
        line = line.strip('\n')
        arr = line.split(",")
        names.append(arr[0])
        marks.append(float(arr[1]));
        explode.append(0)
    explode[0] = 0.1
    fig1, ax1 = plt.subplots()
    fig1.suptitle('Marks Graph', fontsize=20)
    ax1.pie(marks, explode=explode, labels=names, autopct='%1.1f%%',shadow=True, startangle=90)
    ax1.axis('equal')
    plt.show()
  def loadPaper():
    global main
    global newwin
    font1 = ('times', 14, 'bold')
    newwin = Toplevel(main)
    newwin.geometry("1200x1200")
    main.withdraw()
    uploadbutton = Button(newwin, text="Upload MCQ Paper", command=uploadPaper)
```

```
    uploadbutton.place(x=400,y=50)
    uploadbutton.config(font=font1)
    markschart = Button(newwin, text="View Marks Graph", command=marksGraph)
    markschart.place(x=400,y=100)
    markschart.config(font=font1)
def viewmarks():
    index = 0
    sname = simpledialog.askstring("Enter Your Name", "Enter Your Name",parent=main)
    with open("marks.txt", "r") as file:
      for line in file:
        line = line.strip('\n')
        arr = line.split(",")
        if arr[0] == sname:
          messagebox.showinfo("View Marks Screen","Your Marks : "+arr[1])
          index = 1
    if index == 0:
      messagebox.showinfo("View Marks Screen","Given name does not exists")
def login():
    global password_login_entry
    global username_login_entry
    username = username_verify.get()
    password = password_verify.get()
    username_login_entry.delete(0, END)
    password_login_entry.delete(0, END)
    if username == 'faculty' and password == 'faculty':
```

```
        loadPaper()
    else:
        messagebox.showinfo("Invalid Login Details","Invalid Login Details")
font = ('times', 20, 'bold')
title = Label(main, text='E-ASSESSMENT USING IMAGE PROCESSING IN ∞EXAMS')
title.config(bg='blue', fg='black')
title.config(font=font)
title.config(height=3, width=80)
title.place(x=5,y=5)
font1 = ('times', 14, 'bold')
l1 = Label(main, text="Username * ")
l1.place(x=500,y=150)
l1.config(font=font1)
username_login_entry = Entry(main, textvariable=username_verify)
username_login_entry.place(x=650,y=150)
l2 = Label(main, text="Password * ")
l2.place(x=500,y=200)
l2.config(font=font1)
password_login_entry = Entry(main, textvariable=password_verify, show= '*')
password_login_entry.place(x=650,y=200)
upload = Button(main, text="Login", command=login)
upload.place(x=550,y=250)
upload.config(font=font1)
viewmark = Button(main, text="Student View Marks", command=viewmarks)
viewmark.place(x=550,y=300)
```

```
viewmark.config(font=font1)

main.config(bg='skyblue')

main.mainloop()
```

A

PROJECT REPORT

On

Practical Privacy-Preserving User Profile Matching in Social Networks

Submitted by

Mr. D.G. Saiteja (17K81A0676) **Ms. G.M. Vaibhavi (17K81A0577)**
Mr. K. Mounik (17K81A0586) **Ms. K. Kiranmayi (17K81A0587)**

in partial fulfillment for the award of the

degreeof

BACHELOR OF TECHNOLOGY

IN

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Under the Guidance of Mr. V. Bhaskar

Assistant Professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ST. MARTIN'S ENGINEERING COLLEGE
An Autonomous Institute

Dhulapally, Secunderabad – 500 100

JUNE 2021

BONAFIDE CERTIFICATE

This is to certify that the project entitled **Practical Privacy-Preserving User Profile Matching in Social Networks**, is being submitted by **D.G.Sai teja 17K81A0576, G.M.Vaibhavi 17K81A0577, K.Mounik 17K81A0586 K.Kiranmayi 17K81A0587**, in partial fulfillment of the requirement for the award of the degree of **BACHELOR OF TECHNOLOGY** in **Computer Science Of Engineering** is recorded of bonafide work carried out by them. The result embodied in this report have been verified and found satisfactory.

Guide

Mr. V. Bhaskar

Department of CSE

Head of the Department

Dr.M.NARAYANAN

Department of CSE

Internal Examiner

External Examiner

Place:

Date:

DECLARATION

We, the student of **Bachelor of Technology** in Department of Computer Science and Engineering', session: 2017 – 2021 , St. Martin's Engineering College, Dhulapally, Kompally, Secunderabad, hereby declare that work presented in this Project Work entitled **Practical Privacy-Preserving User Profile Matching in Social Networks** is the outcome of our own bonafide work and is correct to the best of our knowledge and this work has been undertaken taking care of Engineering Ethics. This result embodied in this project report has not been submitted in any university for award of any degree.

D.G Sai Teja	17K81A0576
G.M.Vaibhavi	17K81A0577
K. Mounik	17K81A0586
K. Kiranmayi	17K81A0587

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible and whose encouragements and guidance have crowded effects with success.

We extended our deep sense of gratitude to Principal, Dr. P. SANTOSH KUMAR PATRA, St. Martin's Engineering College, Dhulapally, for permitting us to undertake this project.

We are also thankful to Dr. M.NARAYANAN, Head of the Department, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

We would like to thank Dr. T. POONGOTHAI, Department of Computer Science and Engineering (AI & ML) for her encouragement and insightful comments and as well as our project coordinators Dr. B.RAJALINGAM, Associate Professor and Mr. J.SUDHAKAR, Assistant Professor, in Department of Computer Science and Engineering for their valuable support.

We would like to express our sincere gratitude and indebtedness to our project supervisor Mr. V. Bhaskar Assistant Professor, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

Finally, we express thanks to all those who have helped us successfully to completing this project. Furthermore, we would like to thank our family and friends for their moral support and encouragement.

We express thanks to all those who have helped us in successfully completing the project.

D.G Sai Teja 17K81A0576

G.M.Vaibhavi 17K81A0577

K. Mounik 17K81A0586

K. Kiranmayi 17K81A0587

ABSTRACT

Our project is based on practical privacy preserving in social network. A user queries a user profile database, maintained by social networking service provider to find out other users whose profiles are similar to the profile specified by the quering user.

We give a privacy preserving solution for user profile matching in social networks by using multiple servers.We are using homomorphic encryption and allows a user to find out some matching users with the help of the multiple servers without revealing to anyone privacy of the query and the queried user profiles.In our proposed system we give three solutions for privacy-preserving user profile matching (a basic protocol, a privacy-enhanced protocol and a two-party protocol) for three different settings.If at least one of multiple servers is honest, our privacy-enhanced and two-party protocols achieve the user profile privacy and the user query privacy.Our solution achieves user profile privacy and user query privacy as long as at least one of the multiple servers is honest.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF OUTPUT SCREENS

LIST OF ACRONYMS

AVI	Audio Video Interlace
BMP	Bitmap
CPU	Central Processing Unit
GB	Giga Bytes
GUI	Graphical User Interface

1. Introduction

Online dating is a growing industry, increasing in popularity every year The proliferation of dating sites has become a cultural phenomenon as millions of users flock to find romantic partners online.

Online dating is attractive for several reasons the pool of eligible partners is large it offers an alternative to relying on family and friends as matchmakers; people live longer and are more likely to seek new relationships later in life ; and the increase in broadband access to the Internet has expanded the potential market. Online dating is a valuable innovation. It is now estimated that 1 in 5 marriages are a result of online dating.

When you sign up for an online dating service, you create a "profile" of yourself that others can browse. You may be asked to reveal your age, sex, education, profession, number of children, religion, geographic location, sexual proclivities, drinking behavior, hobbies, income, religion, ethnicity, drug use, where you live, where you work, and the places you go. Once an online dating service has your information, it has it for keeps. Even after you cancel your account (fall in love, get married, take a vow of celibacy, etc.), most dating sites retain your information

In the hope of attracting romantic interest, customers disclose sensitive personal information about themselves. This information may then be re-disclosed not only to prospective dates, but also to advertisers and, ultimately, to data aggregators who use the data for purposes unrelated to online dating and without customer consent. In addition, there are risks such as scammers, sexual predators, and reputational damage that come along with using online dating services.

Many online dating sites take shortcuts with respect to safeguarding the privacy and security of their customers. Often, they use counterintuitive "privacy" settings, and permit serious security flaws.

In July 2015, a group calling itself "The Impact Team" stole the user data of Ashley Madison, a commercial website billed as enabling extramarital affairs. The group copied personal information about the site's user base, and threatened to release users' names and personally identifying information if Ashley Madison was not immediately shut down

Madison was not immediately shut down. On 18 and 20 August 2015, the group leaked more than 25 gigabytes of company data, including user details. Because of the site's policy of not deleting users' personal information, including real names, home addresses, search history and credit card transaction records, many users feared being publicly shamed.

On 24 August 2015, Toronto police announced that two unconfirmed suicides had been linked to the data breach. In addition, a pastor and professor at the New Orleans Baptist Theological Seminary committed suicide citing the leak that had occurred six days before.

The serious data breach has raised growing concerns amongst users on the dangers of giving out too much personal information. Users of these services also need to be aware of data theft.

How can we protect privacy of user profiles in social networks? So far, the best solution is through encryption, i.e., users encrypt their profiles before uploading them onto social networks. When user profiles are encrypted, it is a challenging problem to match the users with the similar profiles.

In this paper, we consider a scenario where a user queries a user profile database, maintained by a social networking service provider, to find out some users whose profiles are similar to the profile specified by the querying user. A typical example of this application is online dating. We give a privacypreserving solution for user profile matching in social networks by using multiple servers.

Our basic idea can be summarized as follows. Before uploading user profile to a social network, each user encrypts his profile by a homomorphic encryption scheme with the common encryption key. Therefore, even if the user profile database falls into the hand of a hacker, he can only get the garbage encrypted data. When a user wishes to find people in the social network, he encrypts his preferred user profile and dissimilarity threshold and submits his query to the social networking service provider. Based on the query, multiple servers, which secretly share the decryption key, compare the preferred user profile with each record in the database. If dissimilarity is less than the threshold, the matching user' contact information is returned to the querying user.

Our main contributions include

1) We formally define the user profile matching model, and the user profile privacy and the user query privacy.

2) We give three solutions for privacy-preserving user profile matching (a basic protocol, a privacyenhanced protocol and a two-party protocol) for three different settings. If at least one of multiple servers is honest, our privacy-enhanced and two-party protocols achieve the user profile privacy and the user query privacy.

3) We implement our two-party protocol. Experiments show that our solution is practical and efficient. The rest of the paper is organized as: Section II gives a survey of related work. Section III introduces some building techniques; Section IV describes our model for privacy preserving user profile matching; Sections V to VII present our three solutions on the basis of the model. Sections VIII and IX perform the security and performance analysis; The last section concludes our work.

1.1 Project overview

we consider a scenario where a user queries a user profile database, maintained by a social networking service provider, to find out some users whose profiles are similar to the profile specified by the querying user. A typical example of this application is online dating. Most recently, an online data site, Ashley Madison, was hacked, which results in disclosure of a large number of dating user profiles. This serious data breach has urged researchers to explore practical privacy protection for user profiles in online dating. In this paper, we give a privacy preserving solution for user profile matching in social networks by using multiple servers. Our solution is built on homomorphic encryption and allows a user to find out some matching users with the help of the multiple servers without revealing to anyone privacy of the query and the queried user profiles. Our solution achieves user profile privacy and user query privacy as long as at least one of the multiple servers is honest. Our implementation and experiments demonstrate that our solution is practical. In today's digital age, the ever-increasing dependency on computer technology has left the average citizen vulnerable to crimes such as data breaches and possible identity theft. These attacks can occur without notice and often without notification to the victims of a data breach. At this time, there is little incentive for social networks to improve their data security. These breaches often target social media networks such as Facebook and Twitter. They can also target banks and other financial institutions. Malicious

users create fake profiles to phish login information from unsuspecting users. A fake profile will send friend requests to many users with public profiles. These counterfeit profiles bait unsuspecting users with pictures of people that are considered attractive. Once the user accepts the request, the owner of the phony profile will spam friend requests to anyone this user is a friend. In this paper using Artificial Neural Networks we are identifying whether given account details are from genuine or fake users. ANN algorithm will be trained with all previous users fake and genuine account data and then whenever we gave new test data then that ANN train model will be applied on new test data to identify whether given new account details are from genuine or fake users. Online social networks such as Facebook or Twitter contains users details and some malicious users will hack social network database to steal or breach users information, To protect users data we are using ANN Algorithm.

1.2 project objectives

1) We formally define the user profile matching model, the user profile privacy and the user query privacy.

2) We give a solution for privacy-preserving user profile matching for a single dissimilarity threshold and then extend it for multiple dissimilarity thresholds.

3) We perform security analysis on our protocols. If at least one of multiple servers is honest, our protocols achieve user profile privacy and user query privacy.

4) We conduct extensive experiments on a real dataset to evaluate the performance of our proposed protocols under different parameter settings. Experiments show that our solutions are practical and efficient.

1.3 organization of chapter's

This documentation consists of 10 different chapter and they are:

1. Introduction – This chapter covers the overview of our project and its objectives.

2. Literature Survey – This includes the details of our survey.

3. Software and Hardware Requirements – We specify our software and hardware requirements here.

4. Software Development Analysis – This section includes the problem definition and details of the modules we used in our project.

5. Project System Design – This chapter includes the design part of our project which includes uml diagrams.

6. Project Coding – This section contains the details of our project code.

7. Project Testing – The details of test cases and testing are included in this chapter.

8. Output Screens – This contains the screenshots of how our project looks like when executed.

9. Experimental Results – This chapter contains the screenshots of our results.

10. Conclusion and Future Enhancements – This covers the conclusion of our project and the possible future developments.

2. Literature Survey

In recent years, wireless sensor networks have been widely used in healthcare applications, such as hospital and home patient monitoring. Wireless medical sensor networks are more vulnerable to eavesdropping, modification, impersonation and replaying attacks than the wired networks. A lot of work has been done to secure wireless medical sensor networks. The existing solutions can protect the patient data during transmission, but cannot stop the inside attack where the administrator of the patient database reveals the sensitive patient data. In this paper, we propose a practical approach to prevent the inside attack by using multiple data servers to store patient data. The main contribution of this paper is securely distributing the patient data in multiple data servers and employing the Paillier and ElGamal cryptosystems to perform statistic analysis on the patient data without compromising the patients' privacy.

With the rapid growth in the development of smart devices equipped adopted across various applications. Among many biometric traits, fingerprint-based identification systems have been

extensively studied and deployed. However, to adopt biometric identification systems in practical applications, two main obstacles in terms of efficiency and client privacy must be resolved simultaneously. That is, identification should be performed at an acceptable time, and only a client should have access to his/her biometric traits, which are not revocable if leaked. Until now, multiple studies have demonstrated successful protection of client biometric data; however, such systems lack efficiency that leads to excessive time utilization for identification. The most recently researched scheme shows efficiency improvements but reveals client biometric traits to other entities such as biometric database server. This violates client privacy. In this paper, we propose an efficient and privacy-preserving fingerprint identification scheme by using cloud systems. The proposed scheme extensively exploits the computation power of a cloud so that most of the laborious computations are performed by the cloud service provider. According to our experimental results on an Amazon EC2 cloud, the proposed scheme is faster than the existing schemes and guarantees client privacy by exploiting symmetric homomorphic encryption. Our security analysis shows that during identification, the client fingerprint data is not disclosed to the cloud service provider or fingerprint database server.

Computing Set Intersection privately and efficiently between two mutually mistrusting parties is an important basic procedure in the area of private data mining. Assuring robustness, namely, coping with potentially arbitrarily misbehaving (i.e., malicious) parties, while retaining protocol efficiency (rather than employing costly generic techniques) is an open problem. In this work the first solution to this problem is presented.

A Distributed Key Generation (DKG) protocol is an essential component of threshold cryptosystems required to initialize the cryptosystem securely and generate its private and public keys. In the case of discrete-log-based (dlog-based) threshold signature schemes (ElGamal and its derivatives), the DKG protocol is further used in the distributed signature generation phase to generate one-time signature randomizers ($r = g^k$). In this paper we show that a widely used dlog-based DKG protocol suggested by Pedersen does not guarantee a uniformly random distribution of generated keys: we describe an efficient active attacker controlling a small number of parties which successfully biases the values of the generated

keys away from uniform. We then present a new DKG protocol for the setting of dlog-based cryptosystems which we prove to satisfy the security requirements from DKG protocols and, in particular, it ensures a uniform distribution of the generated keys. The new protocol can be used as a secure replacement for the many applications of Pedersen's protocol. Motivated by the fact that the new DKG protocol incurs additional communication cost relative to Pedersen's original protocol, we investigate whether the latter can be used in specific applications which require relaxed security properties from the DKG protocol. We answer this question affirmatively by showing that Pedersen's protocol suffices for the secure implementation of certain threshold cryptosystems whose security can be reduced to the hardness of the discrete logarithm problem. In particular, we show Pedersen's DKG to be sufficient for the construction of a threshold Schnorr signature scheme. Finally, we observe an interesting trade-off between security (reductions), computation, and communication that arises when comparing Pedersen's DKG protocol with ours.

Making new connections according to personal preferences is a crucial service in mobile social networking, where the initiating user can find matching users within physical proximity of him/her. In existing systems for such services, usually all the users directly publish their complete profiles for others to search. However, in many applications, the users' personal profiles may contain sensitive information that they do not want to make public. In this paper, we propose FindU, the first privacy-preserving personal profile matching schemes for mobile social networks. In FindU, an initiating user can find from a group of users the one whose profile best matches with his/her; to limit the risk of privacy exposure, only necessary and minimal information about the private attributes of the participating users is exchanged. Several increasing levels of user privacy are defined, with decreasing amounts of exchanged profile information. Leveraging secure multi-party computation (SMC) techniques, we propose novel protocols that realize two of the user privacy levels, which can also be personalized by the users. We provide thorough security analysis and performance evaluation on our schemes, and show their advantages in both security and efficiency over state-of-the-art schemes.

2.1 survey on back ground

How can we protect privacy of user profiles in social Networks? So far, the best solution is through encryption, i.e., Users encrypt their profiles before uploading them onto social Networks. When user profiles are encrypted, it is a challenging Problem to match the users with the similar profiles.

In this paper, we consider a scenario where a user queries A user profile database, maintained by social networking Service provider, to find out some users whose profiles are Similar to the profile specified by the querying user. A typical Example of this application is online dating. We give a privacy-Preserving solution for user profile matching in social neserver By using multiple servers.

2.2 conclusion on survey

Our basic idea can be summarized as follows. Before Uploading user profile to a social network, each user encrypts His profile by a homomorphic encryption scheme with the Common encryption key. Therefore, even if the user profile Database falls into the hand of a hacker, he can only get the Garbage encrypted data. When a user wishes to find people In the social network, he encrypts his preferred user profile And dissimilarity threshold and submits his query to the social Networking service provider.

Based on the query, multiple Servers, which secretly share the decryption key, compare the Preferred user profile with each record in the database. If the dissimilarity is less than the threshold, the matching user' contact information is returned to the querying user.

3.Software and hardware requirements

The project involved analyzing the design of few applications so as to make the application more users friendly. To do so, it was really important to keep the navigations from one screen to the other well ordered and at the same time reducing the amount of typing the user needs to do. In order to make the application more accessible, the browser version had to be chosen so that it is compatible with most of the Browsers.

REQUIREMENT SPECIFICATION

Functional Requirements

Graphical User interface with the User.

3.1 Software Requirements

For developing the application the following are the Software Requirements:

1.Python

2.Django

3.MySql

4.MySqlclien

5.WampServer 2.4

Operating Systems supported

1.Windows 7

2.Windows Xp

3.Windows 8

Technologies and Languages used to Develop

1.Python

Debugger and Emulator

Any Browser (Particularly Chrome)

3.2 Hardware Requirements

For developing the application the following are the Hardware Requirements:

Processor: Pentium IV or higher

RAM: 256

Space on Hard Disk: minimum 512MB

4.software development and analysis

4.1 overview of problem

Our model considers a social networking service environment with users and servers. Our model with one user, one Database (DB) server and n matching servers can be illustrated

In our model, all users stores their profiles in the DB Server in the social networking service provider. User profile Attributes are either sensitive or insensitive. We consider protection of sensitive attributes only. In addition, user profile at-Tributes are either numeric or categorical. We consider numeric Attributes only.

4.2 Define the problem

For a user who queries the user profile database, different Attributes may have different impacts on the dissimilarity.

Some attributes may be more important for the dissimilarity Than other attributes.

We introduce the concept of attributes Measure the importance of an attribute to the dissimilarity. The Weight for an attribute is an integer more than or equal to 0.

It is specified by a user who queries the database To protect the user profile privacy, the ElGamal encryption Is used. At first, each matching server Si generates its ElGamal Public/private key pair (pki, ski).

According to the common public key PK, a user encrypts Each attribute of his profile along with his contact information And sends the encrypted profile to the DB server, which database Them in a user profile database. The encrypted profile can be Decrypted only by the cooperation of the n matching servers.

All servers (either DB or matching servers) are assumed To be “semi-honest”, i.e., they follows protocols or algorithms Exactly, but may be curious about the profile privacy of the User.

In addition, at least one out of n matching servers is Assumed to be trusted not to collude with other matching Servers. In view of this, the n servers can cooperate to decrypt The encrypted data only required by defined protocols.

To find users with similar profiles from the social network, A user specifies a profile, encrypts and submits it along with A dissimilarity threshold to the social networking service Provider.

The n matching servers cooperate to find the matching users From the user profile database according to the user profile And the dissimilarity threshold and then return the contact Information of the matching users with dissimilarity less than The threshold to the querying user.

4.3 define module

1: a standard or unit of measurement

2: the size of some one part taken as a unit of measure by which the proportions of an architectural composition are regulated

3a: any in a series of standardized units for use together: such as

(1): a unit of furniture or architecture

(2): an educational unit which covers a single subject or topic

b: a usually packaged functional assembly of electronic components for use with other such assemblies

the subwoofer module

4: an independently operable unit that is a part of the total structure of a space vehicle

5a: a subset of an additive group that is also a group under addition

b: a mathematical set that is a commutative group under addition and that is closed under multiplication which is distributive from the left or right or both by elements of a ring and for which a(bx) = (ab)x or (xb)a = x(ba) or both where a and b are elements of the ring and x belongs to the set

4.4 We use two modules:

1.Admin
In this admin used to login , view all users.
2.User
In this module user will register, login, can search for other users.

4.5 modules functionality

Part of the admin set up is users management which allows users to be set up with definable access level/roles, access to a single or multiple branches

Admin module allows system administrator to set up back-end of the system and perform basic system configuration, mainly definition of predefined drop-down fields, definition of classes time schedule, etc. All the new packages and promo bundles as well as new prices and price types for classes, new subjects offered, etc. are defined here.Part of the admin set up is users management which allows users to be set up with definable access level/roles, access to a single or multiple branches. Admin can also set up overall system security settings such as required password strength, inactive session time out, inactive accounts lock out, password reset period, etc. Important part of security is audit log – any changes in the system are logged here – so it's easy to check who changed/removed what, at what time, what was the original value and what is the new value set.

The user module allows users to register, log in, and log out. Users benefit from being able to sign on because this associates content they create with their account and allows various permissions to be set for their roles.

The user module supports user roles, which can be set up with fine-grained permissions allowing each role to do only what the administrator permits. Each user is assigned one or more roles. By default there are three roles: anonymous (a user who has not logged in) and authenticated (a user who is registered), and administrator (a signed in user who will be assigned site administrator permissions).

Users can use their own name or handle and can fine tune some personal configuration settings through their individual my account page. Registered users need to authenticate by supplying their username and password, or alternately an OpenID login.

A visitor accessing your website is assigned a unique ID, the so-called session ID, which is stored in a cookie. For security's sake, the cookie does not contain personal information but acts as a key to retrieving the information stored on your server.

5. PROJECT SYSTEM DESGIN

5.1 SYSTEM ARCHITECTURE

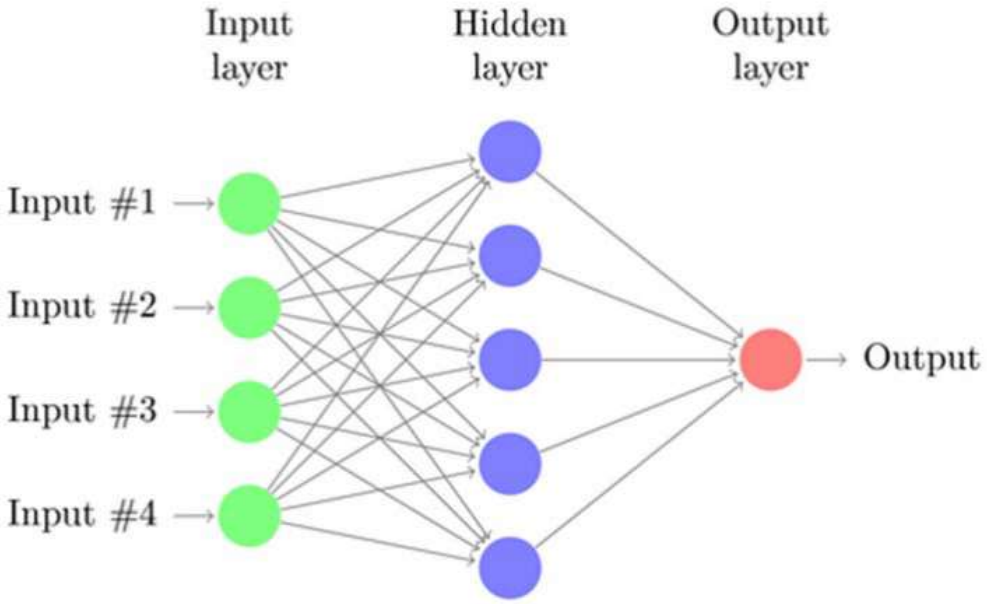


Fig 1: System architecture

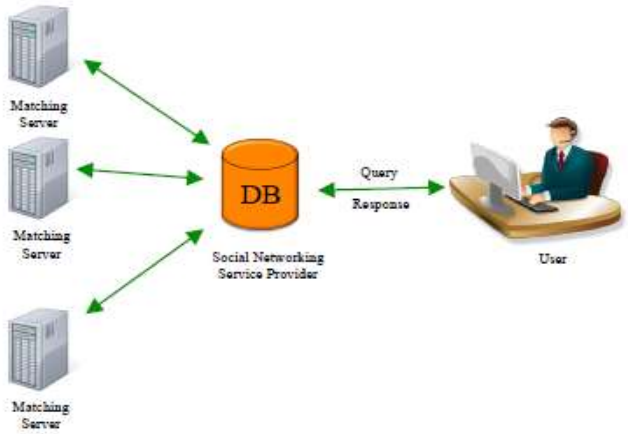


Fig 2: Our Model for Privacy-Preserving User Profile Matching

5.2 UML DIAGRAMS :

UML stands for Unified Modeling Language. UML is a standardized general-purpose modeling language in the field of object-oriented software engineering. The standard is managed, and was created by, the Object Management Group.

The goal is for UML to become a common language for creating models of object oriented computer software. In its current form UML is comprised of two major components: a Meta-model and a notation. In the future, some form of method or process may also be added to; or associated with, UML.

The Unified Modeling Language is a standard language for specifying, Visualization, Constructing and documenting the artifacts of software system, as well as for business modeling and other non-software systems.

The UML represents a collection of best engineering practices that have proven successful in the modeling of large and complex systems.

The UML is a very important part of developing objects oriented software and the software development process. The UML uses mostly graphical notations to express the design of software projects.

USE CASE DIAGRAM:

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.

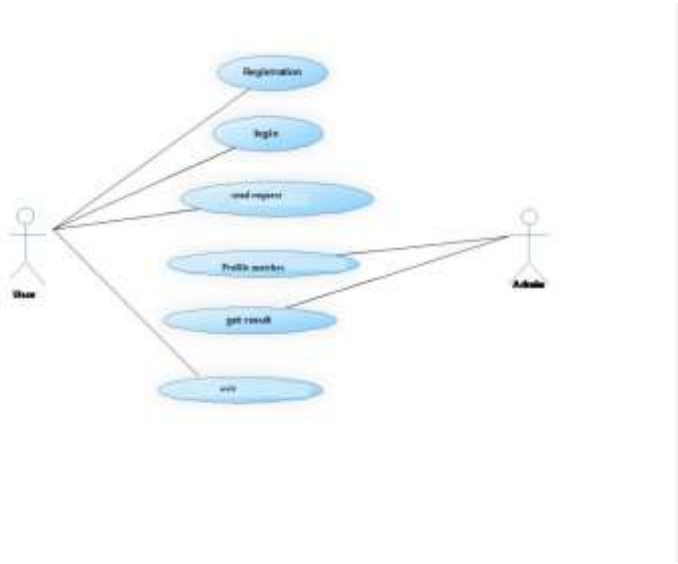


Fig 3: Use case diagram

CLASS DIAGRAM:

In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.

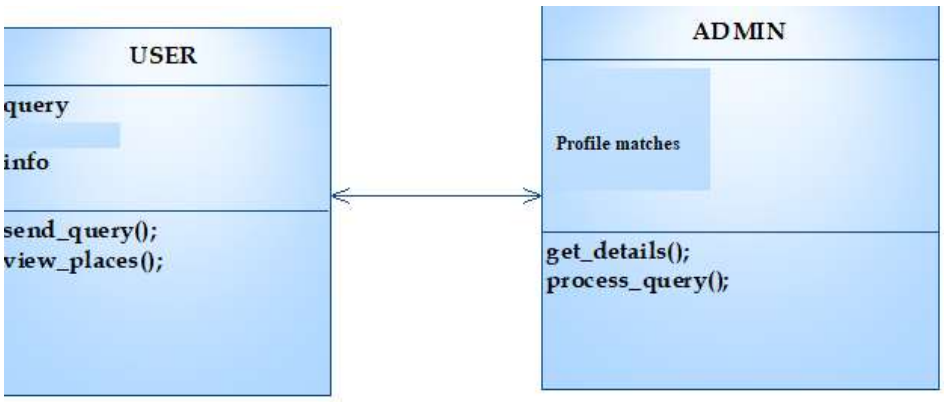


Fig 4: Class diagram

SEQUENCE DIAGRAM:

A sequence diagram in Unified Modeling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a

Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.

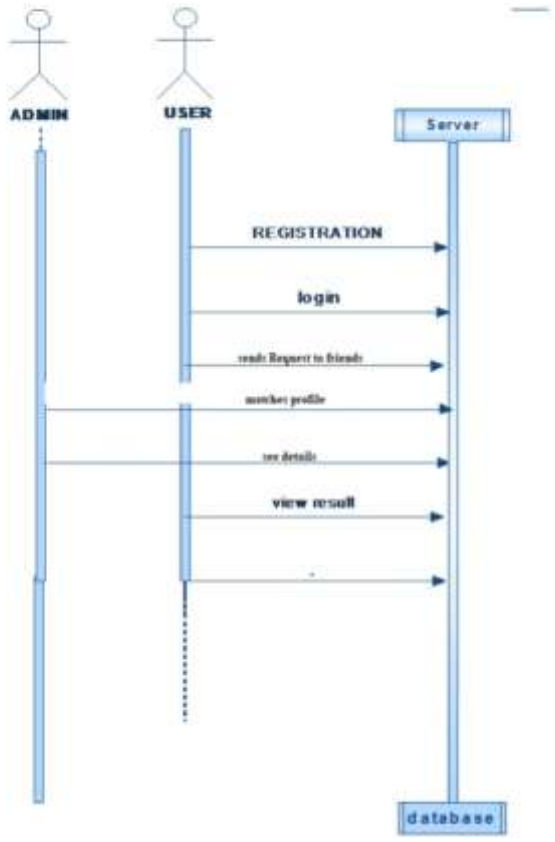


Fig 5: Sequence diagram

ACTIVITY DIAGRAM:

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modeling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.

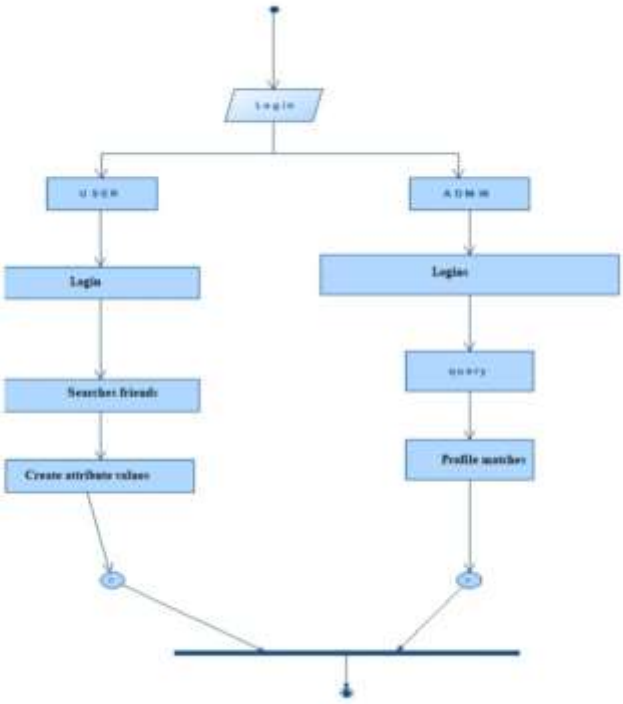


Fig 6: Activity diagram

6. PROJECT CODING

6.1. CODE TEMPLATES

```
# Import required modules #

#Declare global variables def

def Login page():

    #Dialog box appears entering user details for login

def register page():

    #Dialog box appears user has to register before login

def view location based matching():

    #In this user search for other user in their location

def generate message():

    #In this the user can send message to other users
```

6.2. OUTLINE FOR VARIOUS FILES

We used html, java, and css to design this project and code is different file in this we two types of modules – Admin and user admin can see the number of users using this profile matching site and give keys to decrypt the data the use can interact with other user but can't see other user information.

6.3 INPUT AND OUTPUT DESIGN

INPUT DESIGN

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the following things:

What data should be given as input?

How the data should be arranged or coded?

The dialog to guide the operating personnel in providing input.

Methods for preparing input validations and steps to follow when error occur.

OBJECTIVES

1.Input Design is the process of converting a user-oriented description of the input into a computer-based system. This design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerized system.

2. It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The goal of designing input is to make data entry easier and to be free from errors. The data entry screen is designed in such a way that all the data manipulates can be performed. It also provides record viewing facilities.

3.When the data is entered it will check for its validity. Data can be entered with the help of screens. Appropriate messages are provided as when needed so that the user will not be in maize of instant. Thus the objective of input design is to create an input layout that is easy to follow

OUTPUT DESIGN

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making.

1. Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is designed so that people will find the system can use easily and effectively. When analysis design computer output, they should Identify the specific output that is needed to meet the requirements.

2.Select methods for presenting information.

3.Create document, report, or other formats that contain information produced by the system.

The output form of an information system should accomplish one or more of the following objectives.

Convey information about past activities, current status or projections of the

Future.

Signal important events, opportunities, problems, or warnings.

Trigger an action.

Confirm an action.

7. PROJECT TESTING

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

7.1 various tests

Unit testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

Integration testing

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

Functional test

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items:

Valid Input : identified classes of valid input must be accepted.

Invalid Input : identified classes of invalid input must be rejected.

Functions : identified functions must be exercised.

Output : identified classes of application outputs must be exercised.

Systems/Procedures : interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

System Test

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

7.2 Black Box Testing

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box .you cannot "see" into it. The test provides inputs and responds to outputs without considering how the software works.

7.3 White Box Testing

White Box Testing is a testing in which in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is purpose. It is used to test areas that cannot be reached from a black box level.

Unit Testing

Unit testing is usually conducted as part of a combined code and unit test phase of the software lifecycle, although it is not uncommon for coding and unit testing to be conducted as two distinct phases.

Test strategy and approach

Field testing will be performed manually and functional tests will be written in detail.

Test objectives

All field entries must work properly.

Pages must be activated from the identified link.

The entry screen, messages and responses must not be delayed.

Features to be tested

Verify that the entries are of the correct format

No duplicate entries should be allowed

All links should take the user to the correct page.

Integration Testing

Software integration testing is the incremental integration testing of two or more integrated software components on a single platform to produce failures caused by interface defects.

The task of the integration test is to check that components or software applications, e.g. components in a software system or – one step up – software applications at the company level – interact without error.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

Acceptance Testing

User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

8. OUTPUT SCREENS

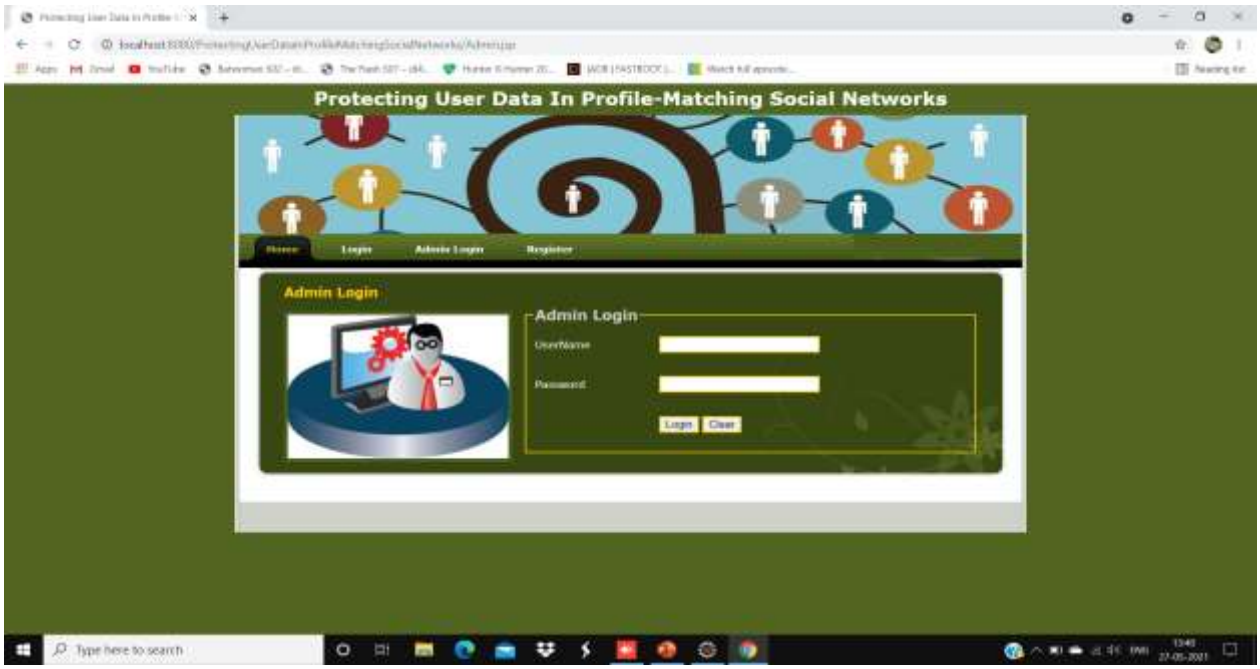


Fig 8.1. Admin interface

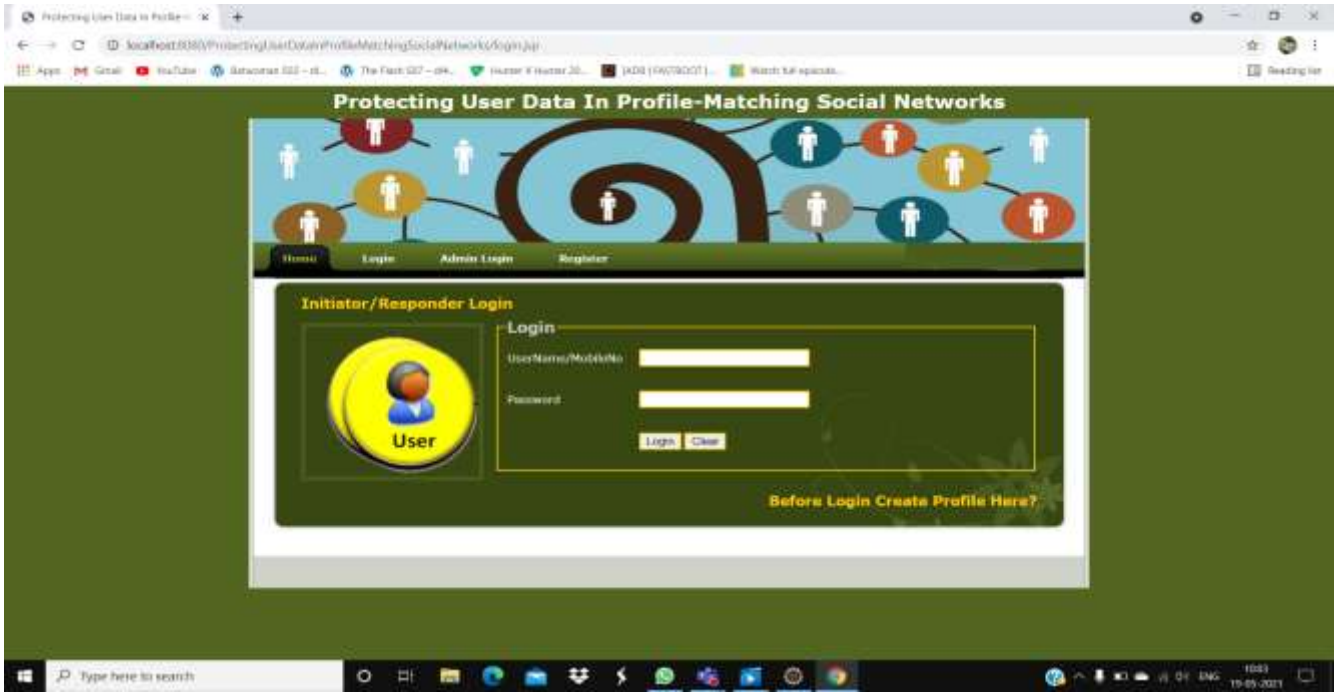


Fig 8.2. User interface

8.2 OUTPUT SCREENS

Fig 8.3. User profile match

9. EXPERMENTAL RESULTS

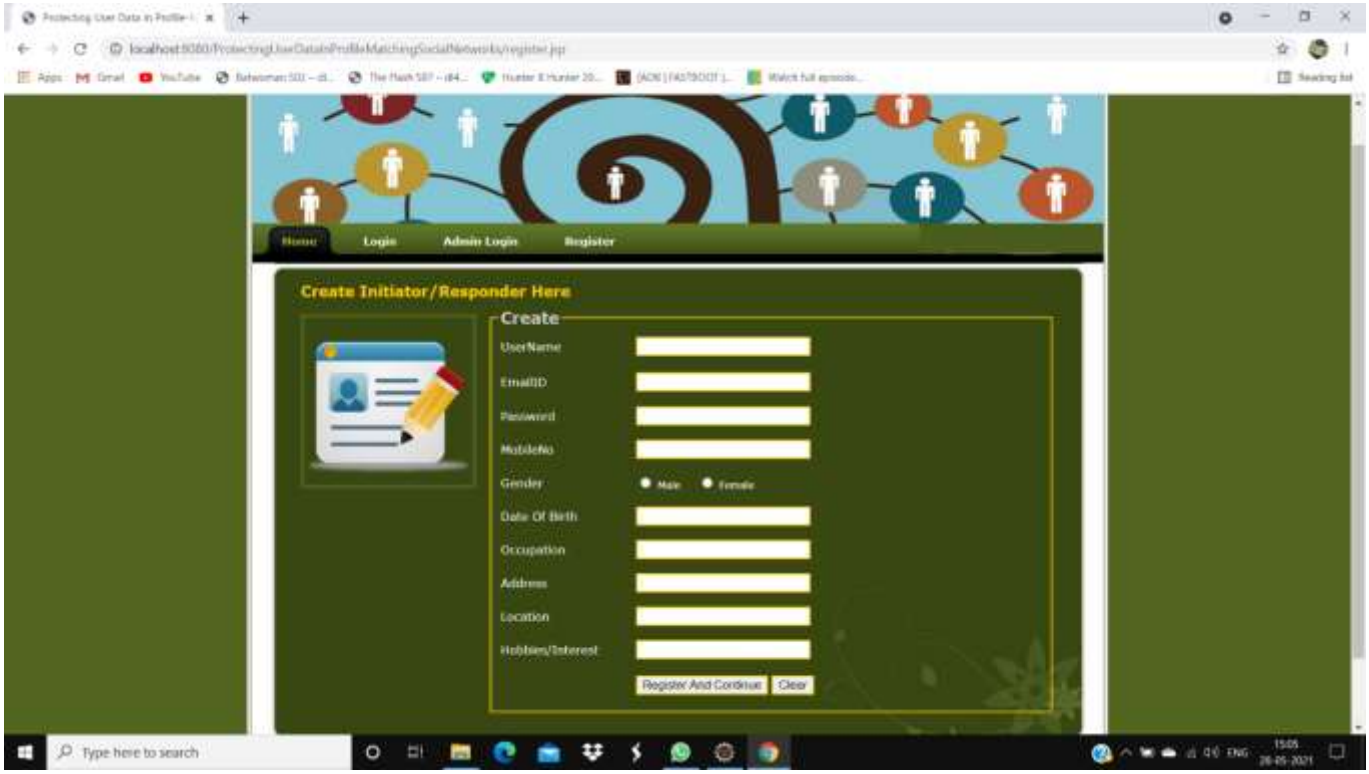


Fig 9.1 Register page

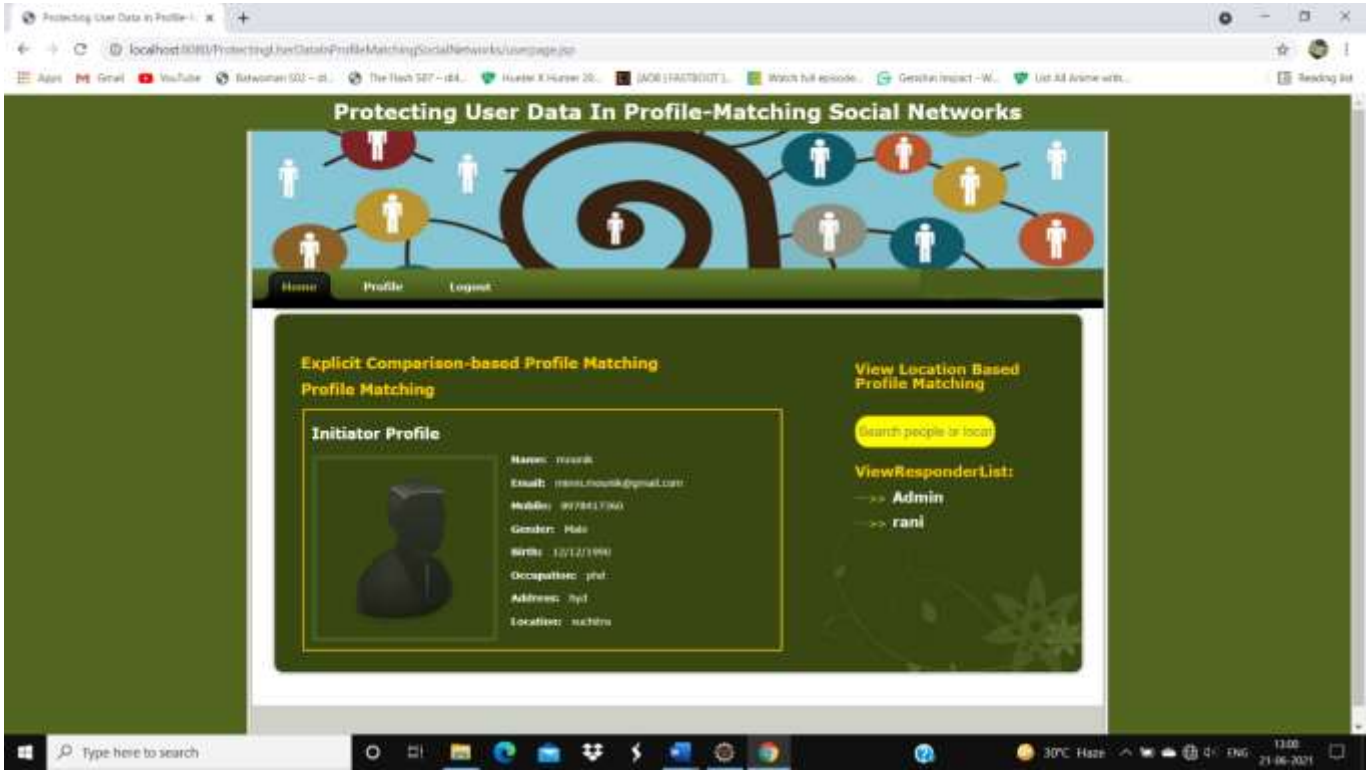


Fig 9.2 User search page

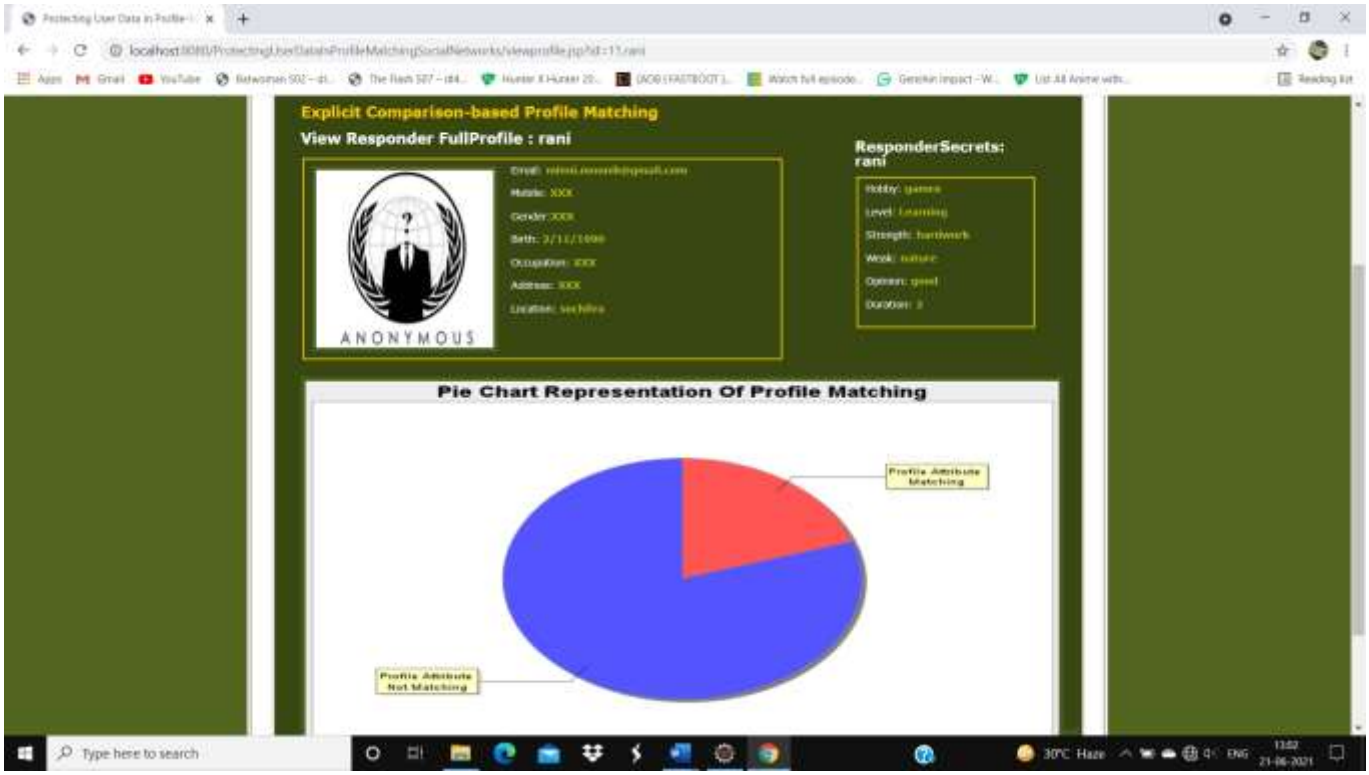


Fig 9.3 Profile of other user

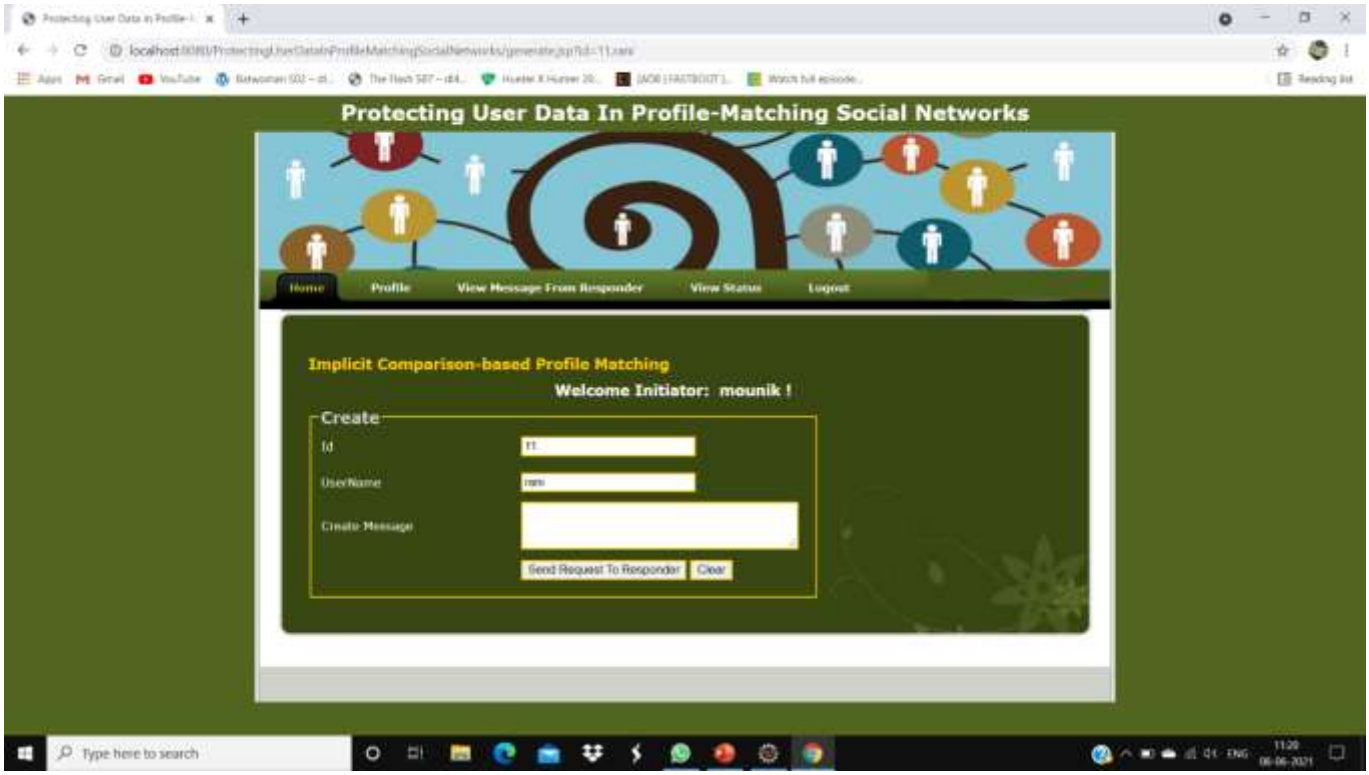


Fig 9.4 Request page

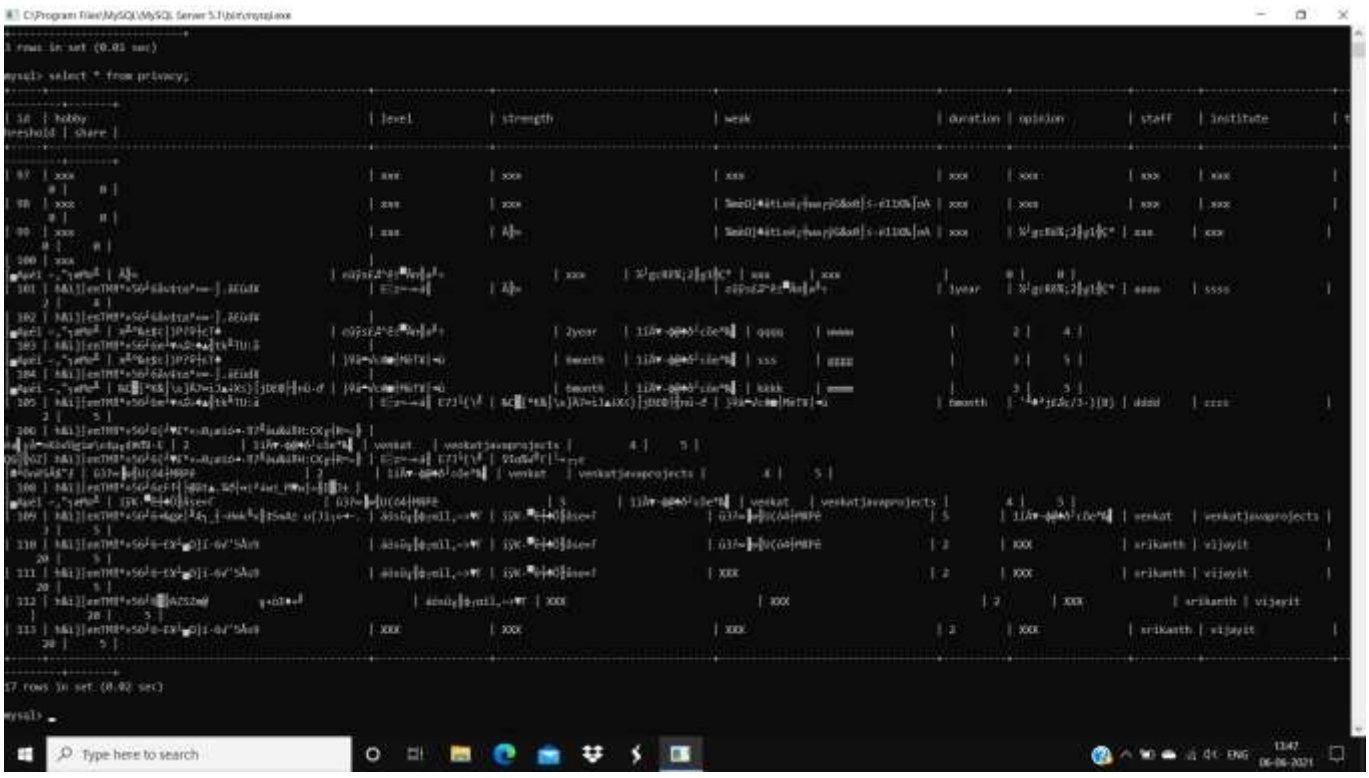


Fig 9.5 Encrypted database.

10.CONCLUSION AND FUTURE ENHANCEMENT

We present three solutions for privacy preserving user profile matching with homomorphic encryption technique and multiple servers.

Security analyses have showed that our privacy-enhanced protocol and optimized two-party protocol achieve user profile privacy and user query privacy.

FUTURE ENHANCEMENT - As the technology is developing, the web-based social networking has turned into the routine for every individual, groups are seen dependent with this technology consistently.

We are hopeful that the technical and social affordances of these sites may contribute to positive social outcomes by enabling individuals to talk, act, and connect with diverse strangers and friends.

REFERENCES

1. F. Rezaeibagha, Y. Mu, Ke Huang, Lanxiang Chenless Secure and Efficient Data Aggregation for IoT Monitoring Systems Published 2021.

2. Xun Yi Russell Paulet Elisa Bertino Fang-Yu Rao Practical Anonymous Subscription with Revocation Based on Broadcast Encryption Published in**:** 2020 IEEE 36th International Conference on Data Engineering (ICDE).

3. Xun Yi Elisa Bertino Fang-Yu Rao Kwok-Yan Lam Surya Nepal Athman Bouguettaya Privacy-Preserving User Profile Matching in Social Networks Published in**:** IEEE Transactions on Knowledge and Data Engineering (Volume: 32, Issue: 8, Aug. 1 2019).

4. Xun Yi Fang-Yu Rao Gabriel Ghinita Elisa Bertino Privacy-Preserving Spatial Crowdsourcing Based on Anonymous Credentials 19th IEEE International Conference on Mobile Data Management (MDM), 2018.

5. L. Sun, L. Zhang, X. Ye, Randomized bit vector: Privacy-preserving encoding mechanism, in CIKM 2018, pp. 1263-1272.

6. Fang-Yu Rao Gabriel Ghinita Elisa Bertino Hybrid Differentially-Private String Matching, IEEE 38th International Conference on Distributed Computing Systems (ICDCS), accepted in 2018.

7. F. Y. Rao, On the security of a variant of ElGamal encryption scheme, IEEE Trans. Dependable and Secure Computing, 2017.

8. X. Yi, E. Bertino, F. Y. Rao, A. Bouguettaya, Practical privacy-preserving user profile matching in social networks, in ICDE 2016, pp. 373-384.

9. X. Yi, A. Bouguettaya, D. Georgakopoulos, A. Song. Privacy protection for wireless medical sensor data, IEEE Transactions on Dependable and Secure Computing, accepted in 2015.

10. Chenyun Dai Fang-Yu RaoTraian Marius Truta Elisa Bertino Privacy-Preserving Assessment of Social Network Data Trustworthiness ,2014.

11. J Cao, FY Rao, M Kuzu, E Bertino, M Kantarcioglu Efficient tree pattern queries on encrypted XML documents Proceedings of the Joint EDBT/ICDT 2013 workshops, 2013.

12. C Dai, FY Rao, G Ghinita, E Bertino Privacy-preserving assessment of location data trustworthiness Proceedings of the 19th ACM SIGSPATIAL, 2011.

13. D. Cristofaro and G. Tsudik, Practical private set intersection protocols with linear complexity, in Financial Cryptography and Data Security'10, 2010.

14. Z. Yang, B. Zhang, J. Dai, A. Champion, D. Xuan, and D. Li, Esmalltalker: A distributed mobile system for social networking in physical proximity, in IEEE ICDCS, 2010.

15. D. Dachman-Soled, T. Malkin, M. Raykova, and M. Yung, Efficient robust private set intersection, in ACNS 2009, pp. 125-142.

PUBLICATION

PROTECTING USER DATA IN PROFILE MATCHING SOCIAL NETWORKS

D.G. Sai Teja [1], G.M. Vaibhavi [2], K. Mounik [3], K. Kiranmayi [4], V. Bhaskar [5]

[1,2,3,4,] UG Scholar,Assistant Professor

Department Of Computer Science And Engineering,

St. Martin's Engineering College,

Near Forest Academy, Dhulapally , Kompally , Secunderabad , Telangana 500014,India

Email-id: saitejadorbala2@gmail.com[1], vaibhaviganta1@gmail.com[2],

minni.mounik@gmail.com[3],kiranmayi0512@gmail.com[4],bhaskarmarikal@gmail.com[5]

Abstract:

In this paper, we consider a scenario where a user queries a user profile database, maintained by a social networking service provider, to find out some users whose profiles are similar to the profile specified by the querying user. A typical example of this application is online dating. Most recently, an online data site, Ashley Madison, was hacked, which results in disclosure of a large number of dating user profiles. This serious data breach has urged researchers to explore practical privacy protection for user profiles in online dating. In this paper, we give a privacy preserving solution for user profile matching in social networks by using multiple servers. Our solution is built on homomorphic encryption and allows a user to find out some matching users with the help of the multiple servers without revealing to anyone privacy of the query and the queried user profiles. Our solution achieves user profile privacy and user query privacy as long as at least one of the multiple servers is honest. Our implementation and experiments demonstrate that our solution is practical.

Keywords - User profile matching, data privacy protection, Elgamal encryption, Paillier encryption, homomorphic encryption

I . Introduction

Online dating is a growing industry, increasing in popularity every year. The proliferation of dating sites has become a cultural phenomenon as millions of users flock to find romantic partners online. Online dating is attractive for several reasons: the pool of eligible partners is large; it offers an alternative to relying on family and friends as matchmakers; people live longer and are more likely to seek new relationships later in life; and the increase in broadband access to the Internet has expanded the potential market. Online dating is a valuable innovation. It is now estimated that 1 in 5 marriages are a result of online dating. When you sign up for an online dating service, you create a "profile" of yourself that others can browse. You may be asked to reveal your age, sex, education, profession, number of children, religion, geographic location, sexual proclivities, drinking behavior, hobbies, income, religion, ethnicity, drug use, where you live, where you work, and the places you go.

Once an online dating service has your information, it has it for keeps. Even after you cancel your account (fall in love, get married, take a vow of celibacy, etc.), most dating sites retain your information. In the hope of attracting romantic interest, customers disclose sensitive personal information about themselves. This information may then be re-disclosed not only to prospective dates, but also to advertisers and, ultimately, to data aggregators who use the data for purposes unrelated to online dating and without customer consent. In addition, there are risks such as scammers, sexual predators, and reputational damage that come along with using online dating services. Many online dating sites take shortcuts with respect to safeguarding the privacy and security of their customers. Often, they use counterintuitive "privacy" settings, and permit serious security flaws. In July 2015, a group calling itself "The Impact Team" stole the user data of Ashley Madison, a commercial website billed as enabling extramarital affairs. The group copied personal information about the site's user base, and threatened to release users' names and personally identifying information if Ashley Madison was not immediately shut down. On 18 and 20 August 2015, the group leaked more than 25 gigabytes of company data, including user details. Because of the site's policy of not deleting users' personal information, including real names, home addresses, search history and credit card transaction records, many users feared being publicly shamed. On 24 August 2015, Toronto police announced that two unconfirmed suicides had been linked to the data breach. In addition, a pastor and professor at the New Orleans Baptist Theological Seminary committed suicide citing the leak that had occurred six days before. The serious data breach has raised growing concerns amongst users on the dangers of giving out too much personal information. Users of these services also need to be aware of data theft. How can we protect privacy of user profiles in social networks? So far, the best solution is through encryption, i.e., users encrypt their profiles before uploading them onto social networks. When user profiles are encrypted, it is a challenging problem to match the users with the similar profiles.

II . Literature Survey

In recent years, wireless sensor networks have been widely used in healthcare applications, such as hospital and home patient monitoring. Wireless medical sensor networks are more vulnerable to eavesdropping, modification, impersonation and replaying attacks than the wired networks. A lot of work has been done to secure wireless medical sensor networks. The existing solutions can protect the patient data during transmission, but cannot stop the inside attack where the administrator of the patient database reveals the sensitive patient data. In this paper, we propose a practical approach to prevent the inside attack by using multiple data servers to store patient data. The main contribution of this paper is securely distributing the patient data in multiple data servers and employing the Paillier and ElGamal cryptosystems to perform statistic analysis on the patient data without compromising the patients' privacy.

With the rapid growth in the development of smart devices equipped adopted across various applications. Among many biometric traits, fingerprint-based identification systems have been

extensively studied and deployed. However, to adopt biometric identification systems in practical applications, two main obstacles in terms of efficiency and client privacy must be resolved simultaneously. That is, identification should be performed at an acceptable time, and only a client should have access to his/her biometric traits, which are not revocable if leaked. Until now, multiple studies have demonstrated successful protection of client biometric data; however, such systems lack efficiency that leads to excessive time utilization for identification. The most recently researched scheme shows efficiency improvements but reveals client biometric traits to other entities such as biometric database server. This violates client privacy. In this paper, we propose an efficient and privacy-preserving fingerprint identification scheme by using cloud systems. The proposed scheme extensively exploits the computation power of a cloud so that most of the laborious computations are performed by the cloud service provider. According to our experimental results on an Amazon EC2 cloud, the proposed scheme is faster than the existing schemes and guarantees client privacy by exploiting symmetric homomorphic encryption. Our security analysis shows that during identification, the client fingerprint data is not disclosed to the cloud service provider or fingerprint database server.

Computing Set Intersection privately and efficiently between two mutually mistrusting parties is an important basic procedure in the area of private data mining. Assuring robustness, namely, coping with potentially arbitrarily misbehaving (i.e., malicious) parties, while retaining protocol efficiency (rather than employing costly generic techniques) is an open problem. In this work the first solution to this problem is presented.

A Distributed Key Generation (DKG) protocol is an essential component of threshold cryptosystems required to initialize the cryptosystem securely and generate its private and public keys. In the case of discrete-log-based (dlog-based) threshold signature schemes (ElGamal and its derivatives), the DKG protocol is further used in the distributed signature generation phase to generate one-time signature randomizers ($r = g^k$). In this paper we show that a widely used dlog-based DKG protocol suggested by Pedersen does not guarantee a uniformly random distribution of generated keys: we describe an efficient active attacker controlling a small number of parties which successfully biases the values of the generated keys away from uniform. We then present a new DKG protocol for the setting of dlog-based cryptosystems which we prove to satisfy the security requirements from DKG protocols and, in particular, it ensures a uniform distribution of the generated keys. The new protocol can be used as a secure replacement for the many applications of Pedersen's protocol. Motivated by the fact that the new DKG protocol incurs additional communication cost relative to Pedersen's original protocol, we investigate whether the latter can be used in specific applications which require relaxed security properties from the DKG protocol. We answer this question affirmatively by showing that Pedersen's protocol suffices for the secure implementation of certain threshold cryptosystems whose security can be reduced to the hardness of the discrete logarithm problem. In particular, we show Pedersen's DKG to be sufficient for the construction of a threshold Schnorr signature scheme. Finally, we observe an interesting trade-off between security (reductions), computation, and communication that arises when comparing Pedersen's DKG protocol with ours.

Making new connections according to personal preferences is a crucial service in mobile social networking, where the initiating user can find matching users within physical proximity of him/her. In existing systems for such services, usually all the users directly publish their complete profiles for others to search. However, in many applications, the users' personal profiles may contain sensitive information that they do not want to make public. In this paper, we propose FindU, the first privacy-preserving personal profile matching schemes for mobile social networks. In FindU, an initiating user can find from a group of users the one whose profile best matches with his/her; to limit the risk of privacy exposure, only necessary and minimal information about the private attributes of the participating users is exchanged. Several increasing levels of user privacy are defined, with decreasing amounts of exchanged profile information. Leveraging secure multi-party computation (SMC) techniques, we propose novel protocols that realize two of the user privacy levels, which can also be personalized by the users. We provide thorough security analysis and performance evaluation on our schemes, and show their advantages in both security and efficiency over state-of-the-art schemes.

III . Proposed Methodology

In this paper, we consider a scenario where a user queries a user profile database, maintained by a social networking service provider, to find out some users whose profiles are similar to the profile specified by the querying user. A typical example of this application is online dating. We give a privacy preserving solution for user profile matching in social networks by using multiple servers. Our basic idea can be summarized as follows. Before uploading user profile to a social network, each user encrypts his profile by a homomorphic encryption scheme with the common encryption key. Therefore, even if the user profile database falls into the hand of a hacker, he can only get the garbage encrypted data. When a user wishes to find people in the social network, he encrypts his preferred user profile and dissimilarity threshold and submits his query to the social networking service provider. Based on the query, multiple servers, which secretly share the decryption key, compare the preferred user profile with each record in the database. If the dissimilarity is less than the threshold, the matching user' contact information is returned to the querying user. Our main contributions include

1) We formally define the user profile matching model, and the user profile privacy and the user query privacy.

2) We give three solutions for privacy-preserving user profile matching (a basic protocol, a privacy enhanced protocol and a two-party protocol) for three different settings. If at least one of multiple servers is honest, our privacy-enhanced and two-party protocols achieve the user profile privacy and the user query privacy.

3) We implement our two-party protocol. Experiments show that our solution is practical and efficient.

MODULES:

This application has 2 modules.

a)Admin

b)User

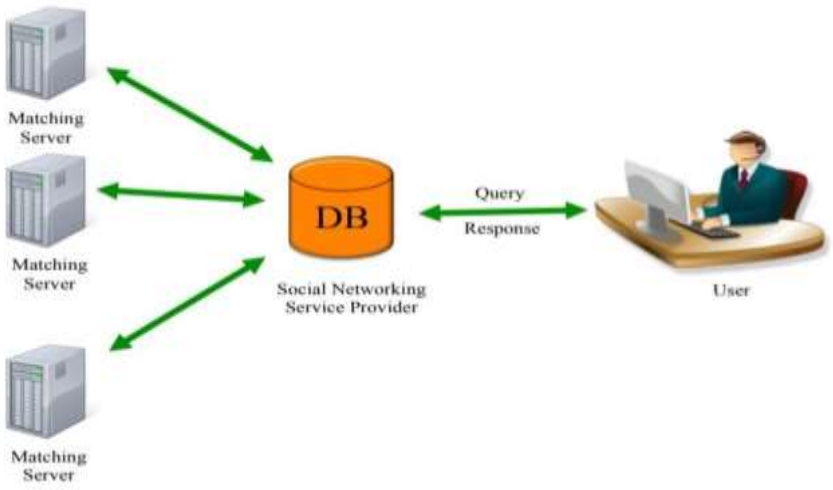
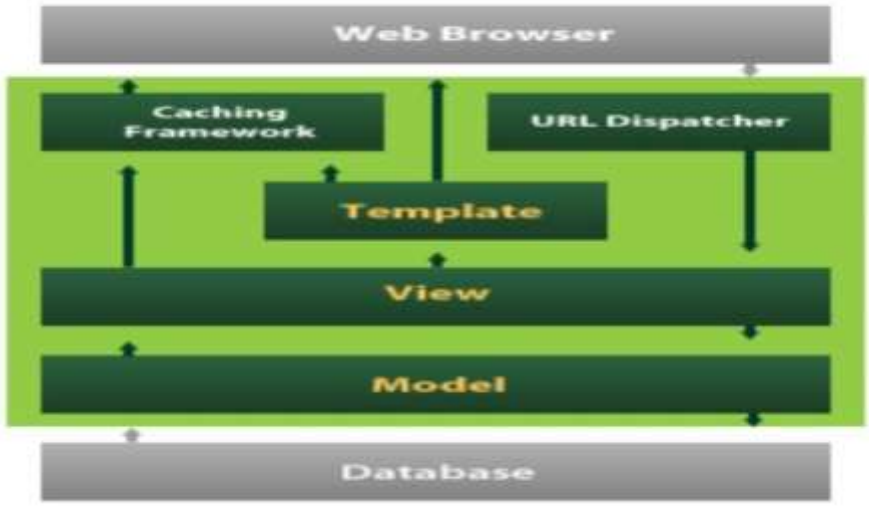


Fig 1: Our model for privacy preserving user profile matching

IV .Experimental And Results

This experiment uses Windows as its operating system. Python is the programming language, and Django is a high level python web framework.



Django also provides an optional administrative create, read, update and delete interface that is generated dynamically through introspection and configured via admin tools.

Implementation

This paper develops a framework based on Django to help users to protect their data in profile matching. Our solution achieves user profile privacy and user query privacy.

Table 1: Complexities per server

Protocols	Distance Computation	Private Comparison
Basic Protocol	$7m$ (Exp.)	1 (Exp.)
Privacy-Enhanced Protocol	$12m$ (Exp.)	$O(\lambda)$ (exp.)
Two-Party Protocol	$12m$ (Exp.)	4λ (exp.)

Our solution is built on homomorphic encryption and allows user to find out some matching users with the help of multiple servers without revealing to anyone privacy and user query and queried user profiles.

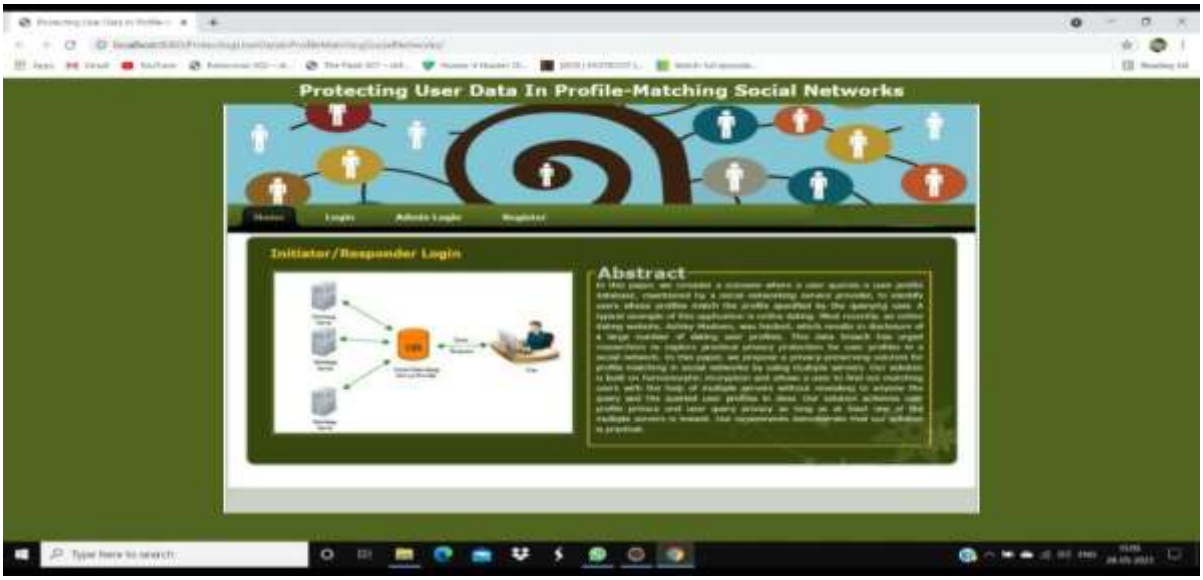


Fig 2: Home Page

This is the home page. Here we can see the abstract of our project. There are 3 portals here

1.login 2.admin login 3.register

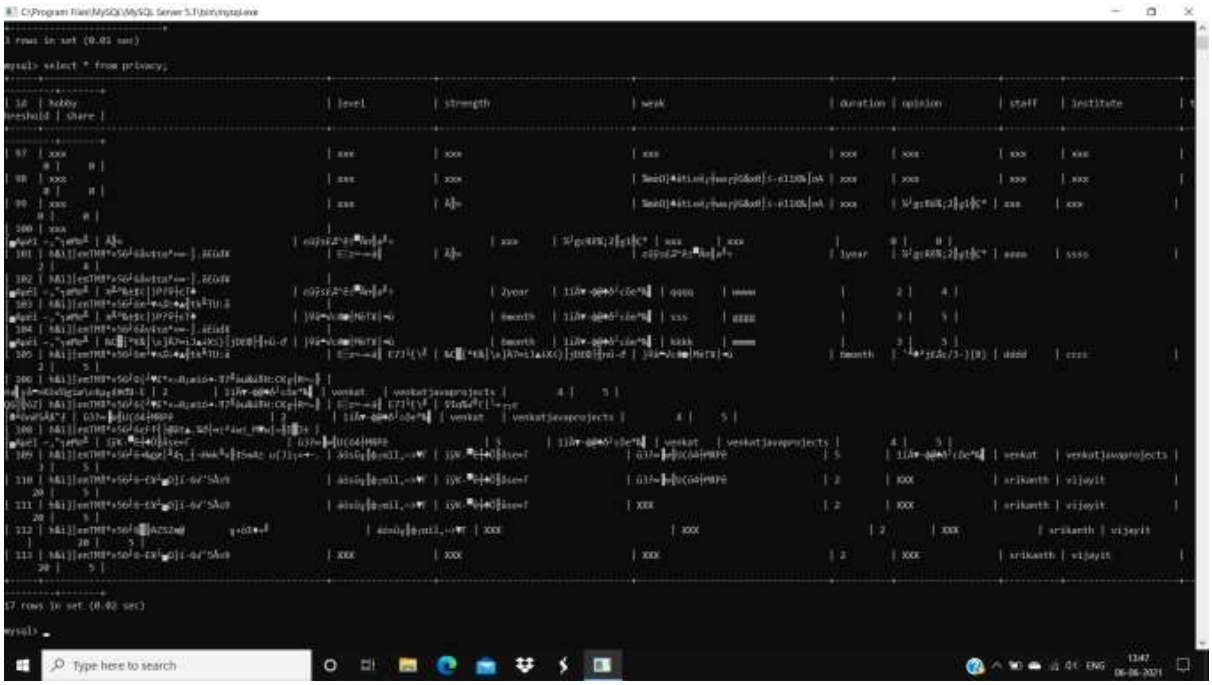


Fig 3: Database Encryption

This is the Database Encryption. In this all the User data that present in database is encrypted. the admin access the data any can modify the data as we can see the data is encrypted in AES Algorithm every time when user access his data it gets decrypted.

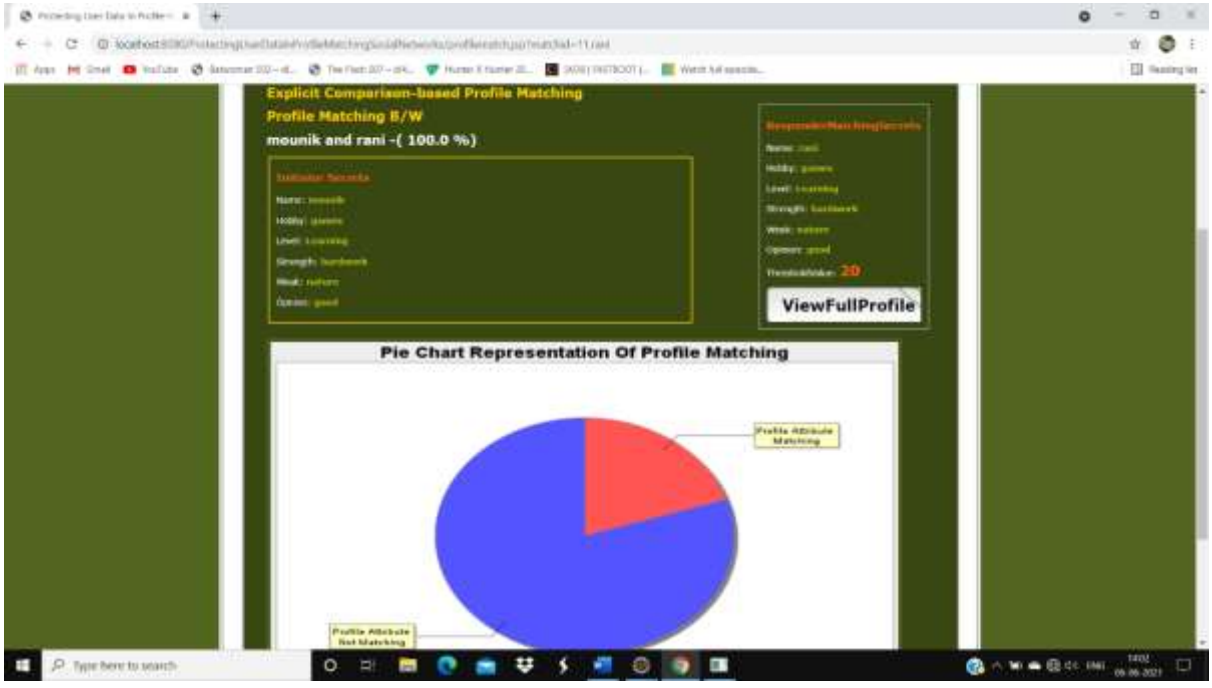


Fig 4: Explict comparrision-based on profile matching

This is the Explict comparrision-based on profile matching page. Here the user can see the percentage of there matching attributes.

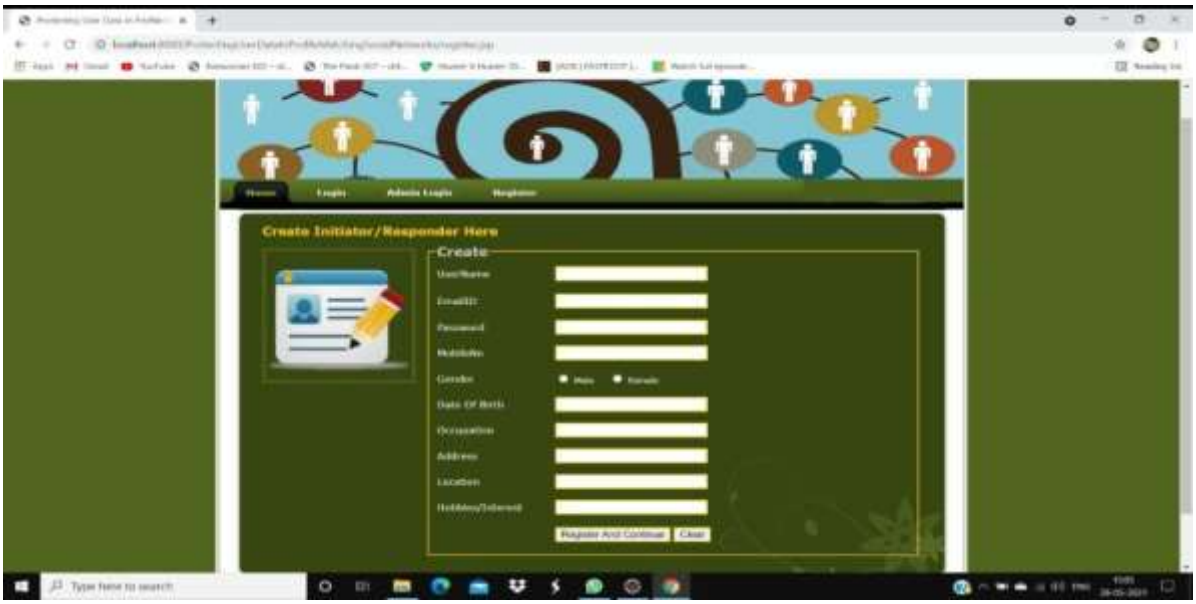


Fig 5: Register page

This is the register page. Here few details like username, email-id, password, mobile number, gender, date of birth, occupation, address, location, hobbies have to be filled by the user before registering.

Result Analysis:

Here in our project first we should register our profile by basic information like name, sex, age, gender, mobile number, hobbies etc. Through this information the database matches our profile with other users and show us the matching users. If intrested we can send them the request and we can also check the requests which the other users sent us.

The user can also see the matching attributes between different uses and can see the percentage of matching and the user data is encrypted and it is shown in the above figure and the data is not shared in any shared to any particular third party applications.

The user data is protected with multiple servers so the hackers has to break in multiple servers which is protected and the data is encrypted so user data is protected.

V . Conclusion

In this paper, we present three solutions for privacy preserving user profile matching with homomorphic encryption technique and multiple servers. Our solutions allow a user to find out some matching users with the help of the multiple servers without revealing to anyone privacy of the query and the queried user profiles. Security analyses have showed that our privacy-enhanced protocol and optimized two-party protocol achieve user profile privacy and user query privacy. We have implemented our solutions and conduct an extensive performance analysis. The experiment results have showed that our protocols, in particular the optimized two-party protocol, are practical and feasible.

VI . References

1. F. Rezaeibagha, Y. Mu, Ke Huang, Lanxiang Chenless Secure and Efficient Data Aggregation for IoT Monitoring Systems Published 2021.

2. Xun Yi Russell Paulet Elisa Bertino Fang-Yu Rao Practical Anonymous Subscription with Revocation Based on Broadcast Encryption Published in**:** 2020 IEEE 36th International Conference on Data Engineering (ICDE).

3. Xun Yi Elisa Bertino Fang-Yu Rao Kwok-Yan Lam Surya Nepal Athman Bouguettaya Privacy-Preserving User Profile Matching in Social Networks Published in**:** IEEE Transactions on Knowledge and Data Engineering (Volume: 32, Issue: 8, Aug. 1 2019).

4. Xun Yi Fang-Yu Rao Gabriel Ghinita Elisa Bertino Privacy-Preserving Spatial Crowdsourcing Based on Anonymous Credentials 19th IEEE International Conference on Mobile Data Management (MDM), 2018.

5. L. Sun, L. Zhang, X. Ye, Randomized bit vector: Privacy-preserving encoding mechanism, in CIKM 2018, pp. 1263-1272.

6. Fang-Yu Rao Gabriel Ghinita Elisa Bertino Hybrid Differentially-Private String Matching, IEEE 38th International Conference on Distributed Computing Systems (ICDCS), accepted in 2018.

7. F. Y. Rao, On the security of a variant of ElGamal encryption scheme, IEEE Trans. Dependable and Secure Computing, 2017.

8. X. Yi, E. Bertino, F. Y. Rao, A. Bouguettaya, Practical privacy-preserving user profile matching in social networks, in ICDE 2016, pp. 373-384.

9. X. Yi, A. Bouguettaya, D. Georgakopoulos, A. Song. Privacy protection for wireless medical sensor data, IEEE Transactions on Dependable and Secure Computing, accepted in 2015.

10. Chenyun Dai Fang-Yu RaoTraian Marius Truta Elisa Bertino Privacy-Preserving Assessment of Social Network Data Trustworthiness ,2014.

11. J Cao, FY Rao, M Kuzu, E Bertino, M Kantarcioglu Efficient tree pattern queries on encrypted XML documents Proceedings of the Joint EDBT/ICDT 2013 workshops, 2013.

12. C Dai, FY Rao, G Ghinita, E Bertino Privacy-preserving assessment of location data trustworthiness Proceedings of the 19th ACM SIGSPATIAL, 2011.

13. D. Cristofaro and G. Tsudik, Practical private set intersection protocols with linear complexity, in Financial Cryptography and Data Security'10, 2010.

14. Z. Yang, B. Zhang, J. Dai, A. Champion, D. Xuan, and D. Li, Esmalltalker: A distributed mobile system for social networking in physical proximity, in IEEE ICDCS, 2010.

15. D. Dachman-Soled, T. Malkin, M. Raykova, and M. Yung, Efficient robust private set intersection, in ACNS 2009, pp. 125-142

ONE PAGE PROFILE

1. DG. SAITEJA

Sai Teja is pursuing his Bachelor of Technology in the Stream of Computer science and engineering at St.Martin's Engineering College.He completed his intermediate from Sri Chaitanya Junior Kalasala and completeed his schooling from St Peter's Model School. His technical skills include C, C++, Java, HTML, Python. His area of interest is Data Science. He has Completed few certificate courses from online platforms like Coursera on Python Programming, HTML, CSS, Data Analysis.

2. G.M. VAIBHAVI

G.M. Vaibhavi is currently pursuing her Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. She completed her intermediate from Sri Gayatri Junior College and 10th class from St Francis Girls High School. Her participations include: National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020, National Workshop on "Android App Development" of Technex'20, IIT Varnasi in association with Innovation Technologies on 27th and 28th February 2020, Women online workshop on "Women in Cyber Security and Privacy in 2020" which was conducted from 6th to 10th July 2020, "One Day Webinar on Internet of Things and Its Applications" conducted by Anand Institute of Higher Technology on 21st May 2020, she has also taken part in 'Anti - Drug' Camping conducted by Lush Life Bistro on 19th August 2017 and Online Session conducted by Institution's Innovation Council of MHRD's Innovation Cell on Leadership talk on 16th May 2020. Her areas of interest are Python, Artificial Intelligence and Machine Learning. She completed few certification courses from online platforms like Coursera and CursaApp.

3. K.MOUNIK

K.MOUNIK is currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. He completed her intermediate from Narayana Junior College and 10th class from Bhashyam high school. His participations include: National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020, National Workshop on "Android App Development" of Technex'20, IIT Varnasi in association with Innovation Technologies on 27th and 28th February 2020, "One Day Webinar on Internet of Things and Its Applications" conducted by Anand Institute of Higher Technology on 21st May 2020, she has also taken part in 'Anti - Drug' Camping conducted by Lush Life Bistro on 19th August 2017 and Online Session conducted by Institution's Innovation Council of MHRD's Innovation Cell on Leadership talk on 16th May 2020. His areas of interest are Python, Artificial Intelligence and Machine Learning. She completed few certification courses from online platforms like Coursera and CursaApp.

4. K. KIRANMAYI

K.kiranmayi is currently pursuing her Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. She completed her intermediate from Gouthami Junior College and 10th class from Z P High School. Her participations include: National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020, National Workshop on "Android App Development" of Technex'20, "One Day Webinar on Internet of Things and Its Applications" conducted by Anand Institute of Higher Technology on 21st May 2020, she completed Online Session conducted by Institution's Innovation Council of MHRD's Innovation Cell on Leadership talk on 16th May 2020. Her areas of interest are Python, Artificial Intelligence and Machine Learning.She took a part in Employability skill development program conducted by Zensar.She also completed few certification courses from online platforms like Coursera and CursaApp.

APPENDICES

```
<%@ page import="java.sql.*,databaseconnection.*"%>


<%
//Profile Matching
Connection con=null;
Statement st = null;
ResultSet rs = null;
String id=null;
String name = request.getParameter("name");
String password = request.getParameter("password");



try{
        con=databasecon.getconnection();
        st = con.createStatement();
        String qry ="select * from admin where (name='"+name+"' AND
password='"+password+"') ";
        rs = st.executeQuery(qry);
        if(!rs.next()){


        response.sendRedirect("Admin.jsp?msg=Enter correct username, password ");
        }
        else{


                session.setAttribute("username",rs.getString("name"));


                response.sendRedirect("AdminHome.jsp");
        }
        con.close();
        st.close();
```

```
}
catch(Exception ex){
	out.println(ex);
}
%>
<%@ page import="java.sql.*,databaseconnection.*"%>

<%
//Profile Matching
Connection con=null;
Statement st = null;
ResultSet rs = null;
String id=null;
String name = request.getParameter("name");
String password = request.getParameter("password");



try{
	 con=databasecon.getconnection();
	st = con.createStatement();
	String qry ="select * from profile where (name='"+name+"' AND
password='"+password+"') OR (mobile='"+name+"' AND password='"+password+"') ";
	rs = st.executeQuery(qry);
	if(!rs.next()){
	out.println("Enter correct username, password ");

	}
	else{
		id=rs.getString("id");
```

```
                session.setAttribute("userid",id);
                session.setAttribute("username",rs.getString("name"));
                session.setAttribute("useremail",rs.getString("email"));
                session.setAttribute("mymobile",rs.getString("mobile"));
                response.sendRedirect("user.jsp");
        }
        con.close();
        st.close();




}
catch(Exception ex){
        out.println(ex);
}
%>


<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN"
"http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">
<html xmlns="http://www.w3.org/1999/xhtml">
<head>
<meta http-equiv="Content-Type" content="text/html; charset=utf-8" />
<title>Protecting User Data In Profile-Matching Social Networks</title>
<meta name="keywords" content="tea and meal, free css templates, green color, white
background, CSS, XHTML" />
<meta name="description" content="Tea and Meal - Green Color, White Background, Free CSS
Template provided by templatemo.com" />
<link href="templatemo_style.css" rel="stylesheet" type="text/css" />
<script language="JavaScript">
function validation()
{
var a = document.form.name.value;
```

```
var b = document.form.email.value;
var c = document.form.password.value;
var d = document.form.mobile.value;

var e = document.form.gender.value;

var f = document.form.birth.value;
var g = document.form.occupation.value;
var h = document.form.address.value;
var i = document.form.location.value;
var j = document.form.hobby.value;

if(a=="")
{
alert("Enter your UserName");
document.form.name.focus();
return false;
}
if(b=="")
{
alert("Enter the emailid");
document.form.email.focus();
return false;
}
if (b.indexOf("@", 0) < 0)
{
alert("Please enter a valid e-mail address.");
document.form.email.focus();
return false;
}
if (b.indexOf(".", 0) < 0)
```

```
{
alert("Please enter a valid e-mail address.");
document.form.email.focus();
return false;
}
if(c=="")
{
alert("Enter your Password");
document.form.password.focus();
return false;
}
if(d=="")
{
alert("Enter the MobileNo");
document.form.mobile.focus();
return false;
}
if(isNaN(d))
   {
      alert("Please enter the Correct Mobile number");
                        document.form.mobile.focus();
      return false;
   }
if(document.form.gender[0].checked==false&&document.form.gender[1].checked==false)
{
alert("Please select any one ");
return false;
}
if(f=="")
{
alert("Enter the Date Of Birth");
```

```
document.form.birth.focus();
return false;
}

        if(g=="")
{
alert("Enter your Occupation");
document.form.occupation.focus();
return false;
}
if(h=="")
{
alert("Enter your address");
document.form.address.focus();
return false;
}
if(i=="")
{
alert("Enter your Location");
document.form.location.focus();
return false;
}
if(j=="")
{
alert("Enter your Hobby");
document.form.hobby.focus();
return false;
}
}
</script>
</head>
```

```
<body>
<h1 align="center"><font color="#FFFFFF" size="5">Protecting User Data In Profile-Matching Social Networks</font></h1>
<div id="templatemo_container">
        <div id="templatemo_header">
        <div id="site_title"></div>
   </div> <!-- end of header -->

   <div id="templatemo_menu">
      <ul>
        <li class="current"><a href="index.html"><b>Home</b></a></li>
        <li><a href="login.jsp"><b>Login</b></a></li>
        <li><a href="Admin.jsp"><b>Admin Login</b></a></li>
        <li><a href="register.jsp"><b>Register</b></a></li>

      </ul>
   </div> <!-- end of menu -->

 <!-- end of top dishes -->

  <div id="templatemo_content"> <span class="top"></span>
   <div id="templatemo_innter_content">
     <div id="templatemo_content_left">
      <h2><font color="#FFCC00"> Create Initiator/Responder Here</font></h2>
      <img src="images/sign-up.png" alt="image" height="188"/>
      <fieldset style="border: 2px solid #FFCC00">
      <legend><font color="#CCCCCC"><strong><font size="4">Create</font></strong></font></legend>
      <table width="398" height="160">
        <form action="insertregister.jsp" method="post" name="form" onsubmit="return validation();">
```

```
<tr>
  <td width="148" height="37"><font size="2">UserName</font></td>
  <td width="238"><input type="text" name="name" id="s" size="25" style="border: 2px solid #FFCC00" /></td>
</tr>
<tr>
  <td width="148" height="37"><font size="2">EmailID</font></td>
  <td width="238"><input type="text" name="email" id="s" size="25" style="border: 2px solid #FFCC00" /></td>
</tr>
<tr>
  <td height="35"><font size="2">Password</font></td>
  <td><input type="password" name="password" id="s" size="25" style="border: 2px solid #FFCC00"/></td>
</tr>
<tr>
  <td height="35"><font size="2">MobileNo</font></td>
  <td><input type="text" name="mobile" id="s" size="25" style="border: 2px solid #FFCC00"/></td>
</tr>
<tr>
  <td height="35"><font size="2">Gender</font></td>
  <td><input type="radio" name="gender" style="border: 2px solid #FFCC00" value="Male" size="25"  id="s" />
    Male     <input type="radio" name="gender" value="Female"  id="s" size="25" style="border: 2px solid #FFCC00"/>
    Female</td>
</tr>
<tr>
  <td height="35"><font size="2">Date Of Birth</font></td>
  <td><input type="text" name="birth" id="s" size="25" style="border: 2px solid #FFCC00"/></td>
</tr>
<tr>
```

```
          <td height="35"><font size="2">Occupation</font></td>
          <td><input type="text" name="occupation" id="s" size="25" style="border: 2px solid #FFCC00"/></td>
        </tr>
        <tr>
          <td height="35"><font size="2">Address</font></td>
          <td><input type="text" name="address" id="s" size="25" style="border: 2px solid #FFCC00"/></td>
        </tr>
        <tr>
          <td height="35"><font size="2">Location</font></td>
          <td><input type="text" name="location" id="s" size="25" style="border: 2px solid #FFCC00"/></td>
        </tr>
        <tr>
          <td height="35"><font size="2">Hobbies/Interest</font></td>
          <td><input type="text" name="hobby" id="s" size="25" style="border: 2px solid #FFCC00"/></td>
        </tr>
        <tr>
          <td height="39"></td>
          <td><input type="submit" name="submit" class="button2" value="Register And Continue" style="border: 2px solid #FFCC00">
            <input type="reset" name="reset" class="button2" value="Clear" style="border: 2px solid #FFCC00">
          </td>
        </tr>
      </form>
    </table>
    </fieldset>
    <br>
        <br>
  </div>
  <div id="templatemo_content_center">
```

```
      <div class="right_column_section"> </div>
    </div>
    <!-- end of content right -->
    <div class="cleaner"> </div>
  </div>
  <div class="cleaner" style="background: #fff;"> </div>
 </div>


<div id="templatemo_footer">


 </div>
<!--  Free CSS Templates by TemplateMo.com  -->
</div><!-- end of container -->
</body>
</html>
<%@ page import="java.sql.*,databaseconnection.*"%>
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN"
"http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">
<html xmlns="http://www.w3.org/1999/xhtml">
<head>
<meta http-equiv="Content-Type" content="text/html; charset=utf-8" />
<title>Protecting User Data In Profile-Matching Social Networks</title>
<meta name="keywords" content="tea and meal, free css templates, green color, white
background, CSS, XHTML" />
<meta name="description" content="Tea and Meal - Green Color, White Background, Free CSS
Template provided by templatemo.com" />
<link href="templatemo_style.css" rel="stylesheet" type="text/css" />
<script language="JavaScript">
function validation()
{
var a = document.form.name.value;
```

```
var b = document.form.email.value;
var c = document.form.password.value;
var d = document.form.mobile.value;

var e = document.form.gender.value;

var f = document.form.birth.value;
var g = document.form.occupation.value;
var h = document.form.address.value;
var i = document.form.location.value;
var j = document.form.hobby.value;

if(a=="")
{
alert("Enter your UserName");
document.form.name.focus();
return false;
}
if(b=="")
{
alert("Enter the emailid");
document.form.email.focus();
return false;
}
if (b.indexOf("@", 0) < 0)
{
alert("Please enter a valid e-mail address.");
document.form.email.focus();
return false;
}
if (b.indexOf(".", 0) < 0)
```

```
{
alert("Please enter a valid e-mail address.");
document.form.email.focus();
return false;
}
if(c=="")
{
alert("Enter your Password");
document.form.password.focus();
return false;
}
if(d=="")
{
alert("Enter the MobileNo");
document.form.mobile.focus();
return false;
}
if(isNaN(d))
   {
      alert("Please enter the Correct Mobile number");
                        document.form.mobile.focus();
      return false;
   }
if(document.form.gender[0].checked==false&&document.form.gender[1].checked==false)
{
alert("Please select any one ");
return false;
}
if(f=="")
{
alert("Enter the Date Of Birth");
```

```
document.form.birth.focus();
return false;
}

        if(g=="")
{
alert("Enter your Occupation");
document.form.occupation.focus();
return false;
}
if(h=="")
{
alert("Enter your address");
document.form.address.focus();
return false;
}
if(i=="")
{
alert("Enter your Location");
document.form.location.focus();
return false;
}
if(j=="")
{
alert("Enter your Hobby");
document.form.hobby.focus();
return false;
}
}
</script>
</head>
```

```
<body>
<h1 align="center"><font color="#FFFFFF" size="5">Protecting User Data In Profile-Matching Social Networks</font></h1>
<div id="templatemo_container">
        <div id="templatemo_header">
        <div id="site_title"></div>
   </div> <!-- end of header -->

   <div id="templatemo_menu">
      <ul>
        <li class="current"><a href="index.html"><b>Home</b></a></li>
        <li><a href="login.jsp"><b>Login</b></a></li>
        <li><a href="Admin.jsp"><b>Admin Login</b></a></li>
        <li><a href="register.jsp"><b>Register</b></a></li>

      </ul>
   </div> <!-- end of menu -->

 <!-- end of top dishes -->

  <div id="templatemo_content"> <span class="top"></span>
   <div id="templatemo_innter_content">
     <div id="templatemo_content_left">
      <h2><font color="#FFCC00">Create Attribute Values</font></h2>
      <img src="images/attribute ladder.gif" alt="image" height="358" width="250"/>
      <form action="insertregister1.jsp" method="post" name="form" onsubmit="return validation();">
        <fieldset style="border: 2px solid #FFCC00">
        <legend><font color="#CCCCCC"><strong><font size="4">Create</font></strong></font></legend>
        <table width="450" height="160">
```

```
<tr bgcolor="#FF0000">
  <td><font size="2" color="#FFCC00"><strong>Attributes</strong></font></td>
  <td><font size="2" color="#FFCC00"><strong>Values</strong></font></td>
</tr>
<tr>
  <td width="180" height="37"><font size="2">UserName</font></td>
  <td width="206"><input type="text" name="name"
value="<%=session.getAttribute("name")%>" id="s" size="25" style="border: 2px solid
#FFCC00" /></td>
</tr>
<tr>
  <td height="35"><font size="2">Hobby/Interests</font></td>
  <td><input type="text" name="hobby" value="<%=session.getAttribute("hobby")%>"
id="s" size="25" style="border: 2px solid #FFCC00"/></td>
</tr>
<tr>
  <td height="35"><font size="2">Level Of Ur Interests</font></td>
  <td><select name="level" style="border: 2px solid #FFCC00">
     <option value="Select"></option>
     <option value="Learning">Learning</option>
     <option value="Middle">Middle</option>
     <option value="WellKnown">WellKnown</option>
   </select></td>
</tr>
<tr>
  <td height="35"><font size="2">Strength</font></td>
  <td><input type="text" name="strength" id="s" size="25" style="border: 2px solid
#FFCC00"/>
  </td>
</tr>
<tr>
  <td height="35"><font size="2">Weak</font></td>
```

```
<td><input type="text" name="weak" id="s" size="25" style="border: 2px solid #FFCC00"/></td>
</tr>
<tr>
<td height="35"><font size="2">Duration Of Ur Interest</font></td>
<td><input type="text" name="duration" id="s" size="25" style="border: 2px solid #FFCC00"/></td>
</tr>
<tr>
<td height="35"><font size="2">Opinion</font></td>
<td><input type="text" name="opinion" id="s" size="25" style="border: 2px solid #FFCC00"/></td>
</tr>
</table>
<fieldset style="border: 2px solid #FFCC00">
<table width="450" height="97">
<tr>
<td width="190" height="46"><font size="2">Staff Name</font></td>
<td width="248"><input type="text" name="staff" id="s" size="25" style="border: 2px solid #FFCC00" /></td>
</tr>
<tr>
<td height="35"><font size="2">Institute Name</font></td>
<td><input type="text" name="institute" id="s" size="25" style="border: 2px solid #FFCC00"/></td>
</tr>
<tr>
<td height="39"></td>
<td><input type="submit" name="submit" class="button2" value="Insert Attribute Values" style="border: 2px solid #FFCC00">
<input type="reset" name="reset" class="button2" value="Clear" style="border: 2px solid #FFCC00">
</td>
</tr>
```

```
        </table>
        </fieldset>
        </fieldset>
        <br>
      </form>
    </div>
    <div id="templatemo_content_center">
      <div class="right_column_section"> </div>
    </div>
    <!-- end of content right -->
    <div class="cleaner"> </div>
  </div>
  <div class="cleaner" style="background: #fff;"> </div>
 </div>


<div id="templatemo_footer">


 </div>
<!--  Free CSS Templates by TemplateMo.com  -->
</div><!-- end of container -->
</body>
</html>


<%@ page import="java.sql.*,databaseconnection.*"%>
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN"
"http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">
<html xmlns="http://www.w3.org/1999/xhtml">
<head>
<meta http-equiv="Content-Type" content="text/html; charset=utf-8" />
<title>Protecting User Data In Profile-Matching Social Networks</title>
```

```
<meta name="keywords" content="tea and meal, free css templates, green color, white background, CSS, XHTML" />
<meta name="description" content="Tea and Meal - Green Color, White Background, Free CSS Template provided by templatemo.com" />
<link href="templatemo_style.css" rel="stylesheet" type="text/css" />
<script language="JavaScript">
function validation()
{
var a = document.form.name.value;
var b = document.form.email.value;
var c = document.form.password.value;
var d = document.form.mobile.value;

var e = document.form.gender.value;

var f = document.form.birth.value;
var g = document.form.occupation.value;
var h = document.form.address.value;
var i = document.form.location.value;
var j = document.form.hobby.value;

if(a=="")
{
alert("Enter your UserName");
document.form.name.focus();
return false;
}
if(b=="")
{
alert("Enter the emailid");
document.form.email.focus();
return false;
```

```
}
if (b.indexOf("@", 0) < 0)
{
alert("Please enter a valid e-mail address.");
document.form.email.focus();
return false;
}
if (b.indexOf(".", 0) < 0)
{
alert("Please enter a valid e-mail address.");
document.form.email.focus();
return false;
}
if(c=="")
{
alert("Enter your Password");
document.form.password.focus();
return false;
}
if(d=="")
{
alert("Enter the MobileNo");
document.form.mobile.focus();
return false;
}
if(isNaN(d))
   {
      alert("Please enter the Correct Mobile number");
                        document.form.mobile.focus();
      return false;
   }
```

```
if(document.form.gender[0].checked==false&&document.form.gender[1].checked==false)
{
alert("Please select any one ");
return false;
}
if(f=="")
{
alert("Enter the Date Of Birth");
document.form.birth.focus();
return false;
}

        if(g=="")
{
alert("Enter your Occupation");
document.form.occupation.focus();
return false;
}
if(h=="")
{
alert("Enter your address");
document.form.address.focus();
return false;
}
if(i=="")
{
alert("Enter your Location");
document.form.location.focus();
return false;
}
if(j=="")
```

```
{
alert("Enter your Hobby");
document.form.hobby.focus();
return false;
}
}
</script>
<script type="text/javascript">
	function fun(){
		var a=document.f1.threshold.value;
		var b=document.f1.share.value;
		if((b==1) || (b>5)){
			alert("shares value should be less than 6 and  greater than 1");
			document.f1.share.focus();
			return false;
		}
		if((a>b) || (a==b)){
			alert("Threshold value should be less than shares value");
			document.f1.threshold.focus();
			return false;
		}


	}
</script>
</head>
<body>
<h1 align="center"><font color="#FFFFFF" size="5">Protecting User Data In Profile-Matching Social Networks</font></h1>
<div id="templatemo_container">
	<div id="templatemo_header">
	<div id="site_title"></div>
```

```
    </div> <!-- end of header -->

    <div id="templatemo_menu">
      <ul>
        <li class="current"><a href="index.html"><b>Home</b></a></li>
        <li><a href="login.jsp"><b>Login</b></a></li>
        <li><a href="Admin.jsp"><b>Admin Login</b></a></li>
        <li><a href="register.jsp"><b>Register</b></a></li>

      </ul>
    </div> <!-- end of menu -->

 <!-- end of top dishes -->


  <div id="templatemo_content"> <span class="top"></span>
    <div id="templatemo_innter_content">
     <div id="templatemo_content_left">
      <h2><font color="#FFCC00">Initiator/Responder Secret Sharing</font></h2>
      <img src="images/cc.png" alt="image" height="188"/>
      <form action="insertregister2.jsp" method="post" name="f1" onsubmit="return fun();">
        <fieldset style="border: 2px solid #FFCC00">
        <legend><font color="#CCCCCC"><strong><font
size="4">Create</font></strong></font></legend>
        <table width="551" height="160">
          <tr>
            <td width="216" height="37"><font size="2">UserName</font></td>
            <td width="275"><input type="text" name="name"
value="<%=session.getAttribute("name")%>" id="s" size="25" style="border: 2px solid
#FFCC00" /></td>
          </tr>
          <tr>
```

```
          <td height="35"><font size="2">Threshold Value (Threshold lessthan
            Share)</font></td>
          <td><input type="text" name="threshold"  id="s" size="25" style="border: 2px solid
#FFCC00"/></td>
        </tr>
        <tr>
          <td height="35"><font size="2">Number of shares</font></td>
          <td><input type="text" name="share" id="s" size="25" style="border: 2px solid
#FFCC00"/>
          </td>
        </tr>
        <tr>
          <td height="35"><font size="2">Secrets </font></td>
          <td><textarea  name="secret"  id="s" rows="3" cols="40" style="border: 2px solid
#FFCC00"/>
            <%=session.getAttribute("mysecret")%></textarea>
          </td>
        </tr>
        <tr>
          <td height="39"></td>
          <td><input type="submit" name="submit" class="button2" value="Share"
style="border: 2px solid #FFCC00">
            <input type="reset" name="reset" class="button2" value="Clear" style="border: 2px
solid #FFCC00">
          </td>
        </tr>
      </table>
      </fieldset>
      <br>
    </form>
  </div>
  <div id="templatemo_content_center">
    <div class="right_column_section"> </div>
```

```
    </div>
    <!-- end of content right -->
    <div class="cleaner"> </div>
   </div>
   <div class="cleaner" style="background: #fff;"> </div>
  </div>


<div id="templatemo_footer">

  </div>
<!--  Free CSS Templates by TemplateMo.com  -->
</div><!-- end of container -->
</body>
</html>

<%@ page import="java.sql.*,java.lang.*,databaseconnection.*"%>
<%@ page import="java.sql.*" %>
<%@ page import="java.io.*" %>
<%@ page import="java.lang.*" %>
<%@ page import="java.awt.*" %>
<%@ page import="org.jfree.chart.ChartFactory" %>
<%@ page import="org.jfree.chart.ChartUtilities" %>
<%@ page import="org.jfree.chart.JFreeChart" %>
<%@ page import="org.jfree.chart.plot.PlotOrientation"%>
<%@ page import="org.jfree.data.*" %>
<%@ page import="org.jfree.data.jdbc.JDBCCategoryDataset"%>

<%@ page  import="java.awt.*" %>
<%@ page  import="java.io.*" %>
<%@ page  import="org.jfree.chart.*" %>
```

```
<%@ page  import="org.jfree.chart.entity.*" %>
<%@ page  import ="org.jfree.data.general.*"%>
<%
                                        String matchid = request.getParameter("matchid");
                                        String temp[]=null;
                                        temp=matchid.split(",");
                                        session.setAttribute("matchname",temp[1]);
                                        String userid = (String)session.getAttribute("userid");
                                        int matchid1=100+Integer.parseInt(temp[0]);
                                        int userid1=100+Integer.parseInt(userid);
                                        int myshare=0,matchshare=0,nomatchshare=0;
                                        StringBuffer mysecret=new StringBuffer();
                                        StringBuffer matchsecret=new StringBuffer();
                                        StringBuffer nomatchsecret=new StringBuffer();
                                        String threshold =null;
                                        ResultSet rs = null;
                                        try {
                                        Connection con=databasecon.getconnection();

                                        Statement st=con.createStatement();

                        for(int j=0;j<2;j++)
                        {
                                                if(j==0){
                                                        String qry ="select * from privacy where
id = '"+userid1+"' ";

                                                        rs = st.executeQuery(qry);
                                                        rs.next();
                                                        myshare=rs.getInt("share");

                                                        //my secret
                                                                for(int i=0;i<myshare;i++,userid1--)
```

```
{
                                                    if(i==0){
                                                        String sql="select
AES_DECRYPT(hobby,'key') from privacy where id = '"+userid1+"'";
                                                        rs =
st.executeQuery(sql);
                                                        rs.next();

        mysecret.append(rs.getString(1));

                                                    }
                                                    if(i==1){
                                                        String sql="select
AES_DECRYPT(level,'key') from privacy where id = '"+userid1+"'";
                                                        rs =
st.executeQuery(sql);
                                                        rs.next();
                                                        mysecret.append("--
");

        mysecret.append(rs.getString(1));

                                                    }
                                                    if(i==2){
                                                        String sql="select
AES_DECRYPT(strength,'key') from privacy where id = '"+userid1+"'";
                                                        rs =
st.executeQuery(sql);
                                                        rs.next();
                                                        mysecret.append("--
");

        mysecret.append(rs.getString(1));

                                                    }
```

```
                                                    if(i==3){
                                                        String sql="select
AES_DECRYPT(weak,'key') from privacy where id = '"+userid1+"'";
                                                        rs =
st.executeQuery(sql);
                                                        rs.next();
                                                        mysecret.append("--
");

        mysecret.append(rs.getString(1));

                                                    }
                                                    if(i==4){
                                                        String sql="select
AES_DECRYPT(opinion,'key') from privacy where id = '"+userid1+"'";
                                                        rs =
st.executeQuery(sql);
                                                        rs.next();
                                                        mysecret.append("--
");

        mysecret.append(rs.getString(1));

                                                    }

                                            }
                                            //my
                                            }
                                            if(j==1){
                                                String qry ="select * from privacy where
id = '"+matchid1+"' ";
                                                rs = st.executeQuery(qry);
                                                rs.next();
                                                matchshare=rs.getInt("share");
                                                threshold=rs.getString("threshold");
```

```
                                                    //match secret
                                                        for(int
i=0;i<matchshare;i++,matchid1--)
                                                    {
                                                                if(i==0){
                                                                    String sql="select
AES_DECRYPT(hobby,'key'),hobby from privacy where id = '"+matchid1+"'";
                                                                    rs =
st.executeQuery(sql);
                                                                    rs.next();


        matchsecret.append(rs.getString(1));

        nomatchsecret.append(rs.getString(2));


                                                                }
                                                                if(i==1){
                                                                    String sql="select
AES_DECRYPT(level,'key'),level from privacy where id = '"+matchid1+"'";
                                                                    rs =
st.executeQuery(sql);
                                                                    rs.next();

        matchsecret.append("--");

        matchsecret.append(rs.getString(1));

        nomatchsecret.append("--");

        nomatchsecret.append(rs.getString(2));


                                                                }
                                                                if(i==2){
                                                                    String sql="select
AES_DECRYPT(strength,'key'),strength from privacy where id = '"+matchid1+"'";
```

```
                                                    rs =
st.executeQuery(sql);
                                                    rs.next();

        matchsecret.append("--");

        matchsecret.append(rs.getString(1));

        nomatchsecret.append("--");

        nomatchsecret.append(rs.getString(2));

                                                }
                                                if(i==3){
                                                    String sql="select
AES_DECRYPT(weak,'key'),weak from privacy where id = '"+matchid1+"'";
                                                    rs =
st.executeQuery(sql);
                                                    rs.next();

        matchsecret.append("--");

        matchsecret.append(rs.getString(1));

        nomatchsecret.append("--");

        nomatchsecret.append(rs.getString(2));
                                                }
                                                if(i==4){
                                                    String sql="select
AES_DECRYPT(opinion,'key'),opinion from privacy where id = '"+matchid1+"'";
                                                    rs =
st.executeQuery(sql);
                                                    rs.next();

        matchsecret.append("--");
```

```
matchsecret.append(rs.getString(1));

nomatchsecret.append("--");

nomatchsecret.append(rs.getString(2));
                                                                    }

                                                }

                                                }
                                        }


                                }
                                catch (Exception e)
                                {
                                        out.println(e.getMessage());
                                }
%>
<%
String mytemp[]=null;
String matchtemp[]=null;
String nomatchtemp[]=null;
int count=0;

mytemp=(mysecret.toString()).split("--");
matchtemp=(matchsecret.toString()).split("--");
nomatchtemp=(nomatchsecret.toString()).split("--");
System.out.print(mytemp[0]);
```

```
if(mytemp[0].equals(matchtemp[0])){
        count=count+1;
}
if(mytemp[1].equals(matchtemp[1])){
        count=count+1;
}
if(mytemp[2].equals(matchtemp[2])){
        count=count+1;
}
if(mytemp[3].equals(matchtemp[3])){
        count=count+1;
}
if(mytemp[4].equals(matchtemp[4])){
        count=count+1;
}

//match chart
                double ans1=0.0d,ans2=0.0d;
                ans1=(double)((double)count/(double)5)*100;
                ans2=(double)((double)(5-count)/(double)5)*100;
            final DefaultPieDataset data = new DefaultPieDataset();
                data.setValue("Profile Attribute Matching", new Double(ans1));
                data.setValue("Profile Attribute Not Matching", new Double(ans2));

                JFreeChart chart = ChartFactory.createPieChart
                ("Pie Chart Representation Of Profile Matching", data, true, true, false);

                try {
                        final ChartRenderingInfo info = new
                        ChartRenderingInfo(new StandardEntityCollection());
```

```
                    final File file1 = new File("");
                    ChartUtilities.saveChartAsPNG(file1, chart, 600, 475, info);

                    } catch (Exception e) {
                            out.println(e);
                    }
//chart
%>
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN"
"http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">
<html xmlns="http://www.w3.org/1999/xhtml">
<head>
<meta http-equiv="Content-Type" content="text/html; charset=utf-8" />
<title>Protecting User Data In Profile-Matching Social Networks</title>
<meta name="keywords" content="tea and meal, free css templates, green color, white
background, CSS, XHTML" />
<meta name="description" content="Tea and Meal - Green Color, White Background, Free CSS
Template provided by templatemo.com" />
<link href="templatemo_style1.css" rel="stylesheet" type="text/css" />

</head>
<body>
<h1 align="center"><font color="#FFFFFF" size="5">Protecting User Data In Profile-Matching
Social Networks</font></h1>
<div id="templatemo_container">
        <div id="templatemo_header">
        <div id="site_title"></div>
   </div> <!-- end of header -->

   <div id="templatemo_menu">
      <ul>
        <li class="current"><a href="user.jsp"><b>Home</b></a></li>
        <li><a href="Profile.jsp"><b>Profile</b></a></li>
```

```
                    <li><a href="viewres.jsp"><b>View Message From Responder</b></a></li>
   <li><a href="index.html"><b>Logout</b></a></li>

     </ul>
  </div> <!-- end of menu -->

<!-- end of top dishes -->


 <div id="templatemo_content"> <span class="top"></span>
  <div id="templatemo_innter_content">
   <div id="templatemo_content_left">
    <h2><font color="#FFCC00">Explicit Comparison-based Profile Matching</font></h2>
    <h2><font color="#FFCC00">Profile Matching B/W </font></h2>
    <h2>
     <%=session.getAttribute("username")%>
     and
     <%=temp[1]%>
     -(
     <%=ans1%>
     %)</h2>
    <fieldset style="border: 2px solid #FFCC00;">
    <h3><font color="#FF6600"><strong>Initiator Secrets</strong></font></h3>
    <p>Name:<a href=""><strong>
     <%=session.getAttribute("username")%>
     </strong></a></p>
    <p>Hobby:<a href=""><strong>
     <%=mytemp[0]%>
     </strong></a></p>
    <p>Level:<a href=""><strong>
```

```
  <%=mytemp[1]%>
  </strong></a></p>
 <p>Strength:<a href=""><strong>
  <%=mytemp[2]%>
  </strong></a></p>
 <p>Weak:<a href=""><strong>
  <%=mytemp[3]%>
  </strong></a></p>
 <p>Opnion:<a href=""><strong>
  <%=mytemp[4]%>
  </strong></a></p>
 </fieldset>
 <br>
 <p><img src="images/match.png" width="800" height="500"></p>
 <div class="cleaner_with_height"> </div>
</div>
<!-- end of content left -->
<div id="templatemo_content_right">
 <div class="right_column_section"> <br>
  <fieldset>
  <h3><font color="#FF6600"><strong>ResponderMatchingSecrets</strong></font></h3>
  <p>Name:<a href=""><strong>
   <%=temp[1]%>
   </strong></a></p>
  <%
                                for(int q=0;q<3;q++){

                                }
                                if(mytemp[0].equals(matchtemp[0])){
                                        %>
  <p>Hobby:<a href=""><strong>
```

```
 <%=matchtemp[0]%>
 </strong></a></p>
<%
                                    }
                                    else{
                                          %>
<p>Hobby:<a href=""><strong>
 <%=nomatchtemp[0]%>
 </strong></a></p>
<%}



                                    if(mytemp[1].equals(matchtemp[1])){
                                          %>
<p>Level:<a href=""><strong>
 <%=matchtemp[1]%>
 </strong></a></p>
<%
                                    }
                                    else{
                                          %>
<p>Level:<a href=""><strong>
 <%=nomatchtemp[1]%>
 </strong></a></p>
<%}



                                    if(mytemp[2].equals(matchtemp[2])){
                                          %>
<p>Strength:<a href=""><strong>
```

```
 <%=matchtemp[2]%>
 </strong></a></p>
<%
                                    }
                                    else{
                                          %>
<p>Strength:<a href=""><strong>
 <%=nomatchtemp[2]%>
 </strong></a></p>
<%}



                                    if(mytemp[3].equals(matchtemp[3])){
                                          %>
<p>Weak:<a href=""><strong>
 <%=matchtemp[3]%>
 </strong></a></p>
<%
                                    }
                                    else{
                                          %>
<p>Weak:<a href=""><strong>
 <%=nomatchtemp[3]%>
 </strong></a></p>
<%}

                                    if(mytemp[4].equals(matchtemp[4])){
                                    %>
<p>Opinion:<a href=""><strong>
 <%=matchtemp[4]%>
```

```
          </strong></a></p>
        <%
                                        }
                                        else{
                                                %>
        <p>Opinion:<a href=""><strong>
          <%=nomatchtemp[4]%>
          </strong></a></p>
        <%}




                %>
        <p>ThresholdValue:<a href=""><strong> <font color="#FF6600" size="4">
          <%=threshold%>
          </font> </strong></a></p>
        <h1><a href="viewprofile.jsp?id=<%=temp[0]%>,<%=temp[1]%>"><span><strong><font
color="#000000">ViewFullProfile</font></strong>
          </span></a></h1>
        </fieldset>
       </div>
     </div>
     <!-- end of content right -->
     <div class="cleaner"> </div>
   </div>
   <div class="cleaner" style="background: #fff;"> </div>
  </div>

<div id="templatemo_footer">


  </div>
<!--  Free CSS Templates by TemplateMo.com  -->
```

```
</div><!-- end of container -->
</body>
</html>

/*
CSS Credit: http://www.templatemo.com/
*/

body {
	margin: 0;
	padding: 10px;
	line-height: 1.5em;
	font-family: Verdana, Arial, san-serif;
	font-size: 11px;
	color: #333333;
	background: #526621;
}

a:link, a:visited { color: #b7bd19; text-decoration: none; font-weight: bold; }
a:active, a:hover { color: #d8df44; text-decoration: underline; }

img {
	padding: 0px;
	margin: 0px;
}

p {
	margin: 0px;
	padding: 0px;
	text-align: justify;
}
```

```
h1 {
	margin: 0 0 15px 0;
}

.cleaner {
	clear: both;
	width: 100%;
	height: 1px;
	font-size: 1px;
}

.cleaner_with_height {
	clear: both;
	width: 100%;
	height: 30px;
	font-size: 1px;
}

.cleaner_with_divider {
	clear: both;
	width: 100%;
	height: 15px;
	border-bottom: 1px solid #333;
	margin-bottom: 25px;
	font-size: 1px;
}

#templatemo_container {
	width: 960px;
	margin: 0 auto;
```

```
        padding: 0 5px;
        background: url(images/templatemo_main_bg.jpg) repeat-y;
}


/* header */
#templatemo_header {
        width: 920px;
        height: 155px;
        padding: 0 20px 0 20px;
        background: url(images/ccc.jpg) no-repeat;
}

#templatemo_header #site_title {
        float: left;
        font-size: 30px;
        font-weight: bold;
        color: #fff;
        padding: 60px 0 10px 0;
        width: 325px;
        height: 55px;

}
/* end of header */
/* menu */
#templatemo_menu {
        clear: both;
        width: 960px;
        margin: 0;
        height: 45px;
        background: url(images/templatemo_menu_bg.jpg) right no-repeat;
```

```
}

#templatemo_menu ul {
	padding: 0 0 0 10px;
	margin: 0 auto;
	height: 45px;
	list-style: none;
}

#templatemo_menu ul li {
	float:left;
	padding-right: 5px;
}

#templatemo_menu li a {
	float: left;
	display: block;
	color: #fff;
	font-size: 12px;
	height: 45px;
	line-height: 40px;
	text-align: center;
	padding: 0px 0 0 0px;
}
#templatemo_menu  li a b {
	float: left;
	display: block;
	padding: 0px 24px 0 24px;
}
#templatemo_menu  li.current a, #templatemo_menu  li a:hover {
	color: #b4c927;
```

```
        text-decoration: none;
        background: url(images/templatemo_menu_hover_right.jpg) right top no-repeat;
}
#templatemo_menu  li.current a b, #templatemo_menu li a:hover b {
        color: #b4c927;
        text-decoration: none;
        background: url(images/templatemo_menu_hover_left.jpg) left top no-repeat;
}
/* end of menu */

/* top dishes */
#templatemo_top_dishes {
        clear: both;
        width: 960px;
        padding: 50px 0px;

}

#templatemo_top_dishes h1 {
        color: #1b2308;
        font-size: 24px;
        margin: 0 20px 15px 20px;
        padding: 0 0 15px 0;
        border-bottom: 1px dotted #1b2308;
}

#templatemo_top_dishes h2 {
        font-size: 14px;
        color: #1f1f1f;
        margin: 0;
        padding: 0 0 5px 0;
```

```
}

#templatemo_top_dishes p {
	margin: 0px;
	padding: 0px;
}

#templatemo_top_dishes .top_dishes_box {
	float: left;
	width: 215px;
	margin-left: 20px;
}

#templatemo_top_dishes .top_dishes_box img {
	margin-bottom: 15px;
	border: 5px solid #e1e0e0;
}
/* end of banner */

/* content */
#templatemo_content {
	position: relative;
	color: #fff;
	width: 920px;
	padding: 0;
	margin-left: 20px;
	background: url(images/templatemo_content_bg_middle.jpg) repeat-y;
}

#templatemo_innter_content {
	background: url(images/templatemo_content_bg_bottom.jpg) bottom center no-repeat;
```

```
}

#templatemo_content .top {
	position: absolute;
	display: block;
	top: 0;
	left: 0;
	width: 920px;
	height: 15px;
	background:url(images/templatemo_content_bg_top.jpg) bottom center no-repeat;
}

#templatemo_content .bottom {
	position: absolute;
	float: left;
	bottom: 0;
	left: 0;
	width: 920px;
	height: 175px;
	background: url(images/templatemo_content_bg_bottom.jpg) bottom center no-repeat;
}

#templatemo_content #templatemo_content_left {
	float: left;
	padding: 10px 0 0 35px;
	width: 850px;
}

#templatemo_content #templatemo_content_right {
	float: right;
	padding: 40px 35px 0 0;
```

```
        width: 245px;

}

 #templatemo_content_left h1 {
         font-size: 24px;
         padding: 3px 0 15px 0;
         margin: 0 0 15px 0;
 }


#templatemo_content_left p {
        padding-bottom: 10px;
        margin: 0px;
}

#templatemo_content_left img {
        float: left;
        margin: 3px 15px 0 0;
        border: 5px solid #4b5e1e;
}

#templatemo_content_right h1 {
        color: #374712;
        font-size: 20px;
        height: 30px;
        margin: 0px;
        padding: 15px 0 0 20px;
        background: url(images/templatemo_header_bg.jpg) no-repeat;
}
```

```
#templatemo_content_right h2 {
	color: #b7bd19;
	font-size: 16px;
	margin: 0 0 5px 0;
	padding: 0 0 5px 0;
}

#templatemo_content_right img {
	border: 5px solid #4b5e1e;
	margin: 0 0 5px 0;
}


#templatemo_content_right p {
	margin: 0 0 5px 0;
	padding: 0 0 5px 0;
}

#templatemo_content_right .right_column_section {
	clear: both;
	margin: 20px;
}
/* left column */

/* footer */
#templatemo_footer {
	clear: both;
	color: #333;
	width: 960px;
	margin-top: 30px;
	padding: 20px 0px 20px 0;
```

```
        text-align: center;
        background: #ced1c8;
}

#templatemo_footer a {
        color: #333;
        font-weight: normal;
}
/* end of footer */
```

A

PROJECT REPORT

On

HUMAN ACTIVITY RECOGNITION USING MACHINE LEARNING WITH DATA ANALYSIS

Submitted by

Ms. K G. Neha Reddy (17K81A0583) **Mr. Lohan Vaddepally (17K81A0594)**

Mr. P. Arvind Chary (17K81A05A3) **Mr. B. Vinod Kumar (18K85A0504)**

in partial fulfillment for the award of the degreeof

BACHELOR OF TECHNOLOGY

IN

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Under the Guidance of

Mr. G. Mallikarjun

Assistant Professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ST.MARTIN'S ENGINEERING COLLEGE
An Autonomous Institute

Dhulapally, Secunderabad – 500 100

JUNE 2021

BONAFIDE CERTIFICATE

This is to certify that the project entitled **Human Activity Recognition using Machine Learning with Data Analysis** , is being submitted by **K G Neha Reddy 17K81A0583, Lohan Vaddepally 17K81A0594, Padala Arvind Chary 17K81A05A3**, **B Vinod Kumar 18K85A0504,** in partial fulfillment of the requirement for the award of the degree of **BACHELOR OF TECHNOLOGY IN Computer Science and Engineering** is recorded of bonafide work carried out by them. The result embodied in this report have been verified and found satisfactory.

.

Guide

Mr. G. MALLIKARJUN

Department of CSE

Head of the Department

Dr. M. NARAYANAN

Department of CSE

Internal Examiner

External Examiner

Place:

Date:

DECLARATION

We, the student of **Bachelor of Technology** in Department of Computer Science and Engineering', session: 2017 – 2021, St. Martin's Engineering College, Dhulapally, Kompally, Secunderabad, hereby declare that work presented in this Project Work entitled Human Activity Recognition using Machine Learning is the outcome of our own bonafide work and is correct to the best of our knowledge and this work has been undertaken taking care of Engineering Ethics. This result embodied in this project report has not been submitted in any university for award of any degree.

K G. Neha Reddy	17K81A0583
Lohan Vaddepally	17K81A0594
P. Arvind Chary	17K81A05A3
B. Vinod Kumar	18K85A0504

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible and whose encouragements and guidance have crowded effects with success.

We extended our deep sense of gratitude to Principal**, Dr. P. SANTOSH KUMAR PATRA**, St. Martin's Engineering College, Dhulapally, for permitting us to undertake this project.

We are also thankful to **Dr. M.NARAYANAN**, Head of the Department, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

We would like to thank **Dr. T. POONGOTHAI,** Department of Computer Science and Engineering (AI & ML) for her encouragement and insightful comments and as well as our project coordinators **Dr. B.RAJALINGAM**, Associate Professor and **Dr. GOVINDA RAJULU.G** Associate Professor, in Department of Computer Science and Engineering for their valuable support.

We would like to express our sincere gratitude and indebtedness to our project supervisor **Mr. MALLIKARJUN.G,** Assistant Professor, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

Finally, we express thanks to all those who have helped us successfully to completing this project. Furthermore, we would like to thank our family and friends for their moral support and encouragement.

We express thanks to all those who have helped us in successfully completing the project.

K G. Neha Reddy	17K81A0583
Lohan Vaddepally	17K81A0594
P. Arvind Chary	17K81A05A3
B. Vinod Kumar	18K85A0504

ABSTRACT

The significant intent is to generate the model for anticipating the activities of a human that ensures the aversion of human life . Activity Recognition is monitoring the liveliness of a person by using smart phone. Smart phones are used in a wider manner and it becomes one of the ways to identify the human's environmental changes by using the sensors in smart mobiles . Smart phones are equipped in detecting sensors like compass sensor, gyroscope, GPS sensor and accelerometer. The contraption is demonstrated to examine the state of an individual. Human Activity Recognition framework collects the raw data from sensors and observes the human movement using different deep learning approach. Deep learning models are proposed to identify motions of humans with plausible high accuracy by using sensed data. Human Activity Recognition Dataset from UCI dataset storehouse is utilized. The performance of a framework is analyzed using Convolutional Neural Network with Long-Short Term Memory and Recurrent Neural Network with Long-Short Term Memory using only the raw data.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES

LIST OF ACRONYMS

<RNN>	Recurrent Neural Network
<LSTM>	Long Short-Term Memory
<HAR>	Human Activity Recognition
<SVM>	Support Vector Machine
<CNN>	Convolutional Neural Network
<KNN>	K Nearest Neighbor

1. INTRODUCTION

Smart phones have become a most useful tool in our daily life for communication with advanced technology provided intelligent assistance to the user in their everyday activities The portable working framework with computing ability and interconnectivity, application programming interfaces (API) for executing outsiders' tools and applications, mobile phones have highlighted such as cameras, GPS, web browsers so on., and implanted sensors such as accelerometers, gyroscope and Magnetometer which permits the improvement of applications in view of client's specific area, movement and context.

To develop resourceful smart phone application, it is imperative to utilize context recognition and situational attention of the gadget's client. Activity Recognition is one such platform for these devices which can dealt by the implicit sensors and it is being used in various areas like business, medicinal services, security, transportation and so forth Different kinds of sensors incorporate wearable sensors which can identify the movement and Bluetooth sensor which empowers the exchange of information from one gadget to another utilizing the information correspondence channels.

Detecting and recording become conceivable with these portable sensors which help to recognize the subjects continually. Activities can also be put away when monitored in their favored environment. Human Movement Recognition is a significant yet challenging examination area with numerous applications in healthcare, smart environment and country security. PC vision-based strategies have generally been utilized for human movement monitoring. An increasingly proficient methodology is to process the information from inertial estimation unit sensors worn on client's body.

A model is built up which fits for perceiving various activities under real world conditions utilizing information gathered by a solitary triaxial accelerometer build into a phone. A triaxial accelerometer that returns a gauge of acceleration along the x, y and z axes from which speed and displacement can be evaluated . Activity Recognition interests to perceive the moves accomplished by an individual given a fixed of perceptions of itself and encompassing environments Recognition might be executed as an instance through exploiting the data recovered from inertial sensors. In some smart devices, the sensors are inserted with the aid of default and to classify a set of physical activities like standing, laying, walking, sitting, scrolling upstairs and scrolling downstairs by means of handling inertial frame markers for hardware with confined resources.

The Overall execution of classifiers with controlled training realities considering about the limited memory accessible on the smart gadgets. Accumulating the training records and it tends to be directly utilized for category steps, which diminishes the weight at the clients

1.1 PROJECT OVERVIEW

Human activity recognition, or HAR for short, is a broad field of study concerned with identifying the specific movement or action of a person based on sensor data. The sensor data may be remotely recorded, such as video, radar, or other wireless methods. It contains data generated from accelerometer , gyroscope and other sensors of Smart phone to train supervised predictive models using machine learning techniques like SVM , Random Forest and decision tree to generate a model. Which can be used to predict the kind of movement being carried out by the person which is divided into six categories walking, walking upstairs, walking down-stairs, sitting, standing and laying.

1.2 PROJECT OBJECTIVE

We can determine what is normal and what is abnormal activity for them therefore indicating whether or not they require attention from facility staff.

Innovative approaches to recognize activities of daily living (ADL) is essential input part for development of more interactive human-computer applications. Methods for understanding Human Activity Recognition (HAR) are developed by interpreting attributes derived from motion location, physiological signals and environmental information.

Effectiveness of machine learning methods are compared with published Multi Class Hardware-Friendly Support Vector Machine (MC-HF-SVM) recognition accuracy

1.3 ORGANIZATION OF CHAPTERS

This documentation consists of 10 different chapter and they are:

1. Introduction – This chapter covers the overview of our project and its objectives.

2. Literature Survey – This includes the details of our survey.

3. Software and Hardware Requirements – We specify our software and hardware requirements here.

4. Software Development Analysis – This section includes the problem definition and details of the modules we used in our project

5. Project System Design – This chapter includes the design part of our project which includes uml diagrams.

6. Project Coding – This section contains the details of our project code.

7. Project Testing – The details of test cases and testing are included in this chapter.

8. Output Screens – This contains the screenshots of how our project looks like when executed.

9. Experimental Results – This chapter contains the screenshots of our results.

10. Conclusion and Future Enhancements – This covers the conclusion of our project and the possible future Developments

2.LITERATURE SURVEY

2.1 SURVEY ON BACKGROUND

1. A New Approach to HAR Using Machine Learning.

AUTHORS: ***Leandro Bezerra, A.H de Souza, Pedro Pedrosa***.

In this paper we attempt to classify and compare the different published models to Recognize the Human Activities. In this paper we also evaluated a new approach of feature selection along with the PCA technique and compare the performance of several machine learning techniques for activity recognition.

The learning algorithms used in the comparison are k-Nearest Neighbors (kNN), Artificial Neural Network Multilayer Perceptron (MLP), Support Vector Machines (SVM), Bayes classifier, Minimal Learning Machine (MLM) and Minimal Learning Machine with Nearest Neighbors (MLM-NN).

2. Robust indoor human activity recognition using wireless signals.

AUTHORS: ***Wang, Y., Jiang, X., Cao, R., Wang, X***.

In this paper we attempt to classify published data sets into different classes based on the activities performed by a person. Here the whole data set is collected from two different categories: the first one is collected from laboratory under predetermined stable circumstances and the second one is collected from different persons of different heights and weights from real time areas.

The data that is collected contains readings from different phone sensors like Accelerometer, Gyroscope, Magnetometer and linear Accelerometer. These readings are again classified based on the spatial axis of the smart phone.

3. Recognition of Daily Human Activity Using an Artificial Neural Network and Smart watch

AUTHORS: ***Min-Cheol Kwon, Sunwoong Choi***

In this paper we attempt to design the models to predict the activities performed by Humans. Here Human Activity Recognition model (HAR) is designed as two models. One model used CNN with LSTM and other model is deployed as RNN with LSTM. Initially, the dataset called Human Activity Recognition [HAR] is collected from UCI machine learning Repository. The dataset is pre-processed using noise filters. After Pre-processing, the data is splitted as fixed windows. Feature engineering

technique is applied to window data. The window data is splitted as 80% training set and 20% testing set. Feature Engineering has following sections. Primarily, the dataset is loaded which has three main types of signals such as total acceleration, body acceleration and gyroscope.

4. Tracking the evolution of smartphone sensing for monitoring human movement.

AUTHORS: ***del Rosario, M.B., Redmond, S.J., Lovell, N.H***

The hyper-parameters for the MLM, MLM-NN, MLP and SVM were chosen using 10-fold cross-validation. SVM was trained using the SVM OpenCV toolbox with default settings for the hyper-parameters. The grid search is between 2 −2 and 211 for each hyper-parameter. Radial basis function and linear function are the kernels used. The strategy one-versus-all is adopted in SVM.

MLP is learned using the back-propagation algorithm and a range of hidden units from 1 to 100. Bayes classifier is validated using gaussian (normal) probability distribution. kNN method was configured with 1, 3 and 5 nearest neighbors. MLM and MLM-NN are trained with three distance metrics (Euclidean, Manhattan and Mahalanobis) and ranging the K number of reference points from 5% to 100% (with a step size of 5%) of the training data.

5. Human activities recognition with rgb-depth camera using hmm. In: Engineering in Medicine and Biology Society (EMBC)

AUTHORS: ***Dubois, A., Charpillet, F.***

Gradient boosting is an AI method for relapse and order issues, which creates an expectation model as a group of powerless forecast models, normally choice trees. The goal of any directed learning algorithm is to characterize a misfortune work and limit it. Gradient boosting machines are in light of a ensemble of choice trees where numerous weak learner trees are utilized in mix as a group to give preferred forecasts over singular trees. Boost has unrivalled regularization and better treatment of missing qualities and also much improved proficiency.

In machine learning, Support vector machines are models that are related with learning algorithms for investigating and grouping information. A SVM algorithm makes a model that assigns guides to one or other divisions of class.]. This model maps the precedents which isolate into the different classes. SVM's can likewise play out a non-direct order. The activity of SVM depends on the hyper plane. The task of SVM depends on the hyper plane that gives biggest least separation to the preparation models. The attributes are taken after implementing dimensionality reduction technique.

2.2. CONCLUSION ON SURVEY

Human Activity Recognition is a platform to combine sensors of smartphones and smartwatches to classify various human activities was proposed. It recognizes activities in real-time. Moreover, this approach is light-weight, computationally inexpensive, and able to run on handheld devices. The results showed that there is no clear winner, but naive Bayes performs best in our experiment in both the classification accuracy and efficiency. The overall accuracy lies between 84.6% and 89.4%, at which the differences are negligible.

Thus, this platform is able to recognize various human activities. However, all of the tested classifiers confused walking and using the stairs activities The second conclusion is that adding the smartwatch's sensor data to the recognition system improves its accuracy with at least six percentage point. Finally, it is computations that the best sampling frequency is in the field of 10 Hz. This work could be further extended by incorporating more sensors (e.g., heart rate sensor), recognizing high-level activities (e.g., shopping or eating dinner) or extrapolating these trained classifiers to other people.

3. SOFTWARE AND HARDWARE REQUIREMENT

The project involved analyzing the design of few applications so as to make the application more users friendly. To do so, it was really important to keep the navigations from one screen to the other well-ordered and at the same time reducing the amount of typing the user needs to do.

3.1. SOFTWARE REQUIREMENTS

For developing the application, the following are the Software Requirements:

Operating Systems supported: Windows 10

Technologies and Languages used to Develop: Python

3.2. HARDWARE REQUIREMENTS

For developing the application, the following are the Hardware Requirements:

- Processor : I3 Processor
- RAM : 4 GB
- Space on Hard Disk : 500 GB

4.SOFTWARE DEVELOPMENT ANALYSIS

4.1. OVERVIEW OF PROBLEM

We can determine what is normal and what is abnormal activity for them therefore indicating whether or not they require attention from facility staff.

Innovative approaches to recognize activities of daily living (ADL) is essential input part for development of more interactive human-computer applications. Methods for understanding Human Activity Recognition (HAR) are developed by interpreting attributes derived from motion location, physiological signals and environmental information.

4.2. DEFINE THE PROBLEM

The purpose of being able to classify what activity a person is undergoing at a given time is to allow computers to provide assistance and guidance to a person prior to or while undertaking a task.

The difficulty lies in how diverse our movements are as we perform our day-to-day tasks. There have been many attempts to use the various machine learning algorithms to accurately classify a person's activity, so much so that Google have created an Activity Recognition API for developers to embed into their creation of mobile applications.

4.3. MODULES OVERVIEW

The System Design Document describes the system requirements, operating environment, system and subsystem architecture, files and database design, input formats, output layouts, human-machine interfaces, detailed design, processing logic, and external interfaces.

MODULES:

This application has four modules which are listed in the following.

1. Data Collection
2. Preprocessing
3. Feature Extraction
4. Standardization

4.4. DEFINE THE MODULES

In this project there are four modules to achieve our expected result. These are the major functionalities of the project. The Data Preprocessing and Feature Extraction process are important in the project for users. There is only a single user, who owns a smartphone.

1. Data Collection

The first step is to collect multivariate time series data from the phones and the watch's sensors. The sensors are sampled with a constant frequency of 30 Hz. After that, the sliding window approach is utilized for segmentation, where the time series is divided into subsequent windows of fixed duration without inter window gaps (Banos et al., 2014). The sliding window approach does not require preprocessing of the time series, and is therefore ideally suited to real-time applications.

2. Data Preprocessing

Filtering is performed afterwards to remove noisy values and outliers from the accelerometer time series data, so that it will be appropriate for the feature extraction stage. There are two basic types of filters that are usually used in this step: average filter (Sharma et al., 2008) or median filter (Thiemjarus, 2010). Since the type of noise dealt with here is similar to the salt and pepper noise found in images, that is, extreme acceleration values that occur in single snapshots scattered throughout the time series. Therefore, a median filter of order 3 (window size) is applied to remove this kind of noise.

3. Feature Extraction

Here, each resulting segment will be summarized by a fixed number of features, i.e., one feature vector per segment. The used features are extracted from both time and frequency domains. Since, many activities have a repetitive nature, i.e., they consist of a set of movements that are done periodically like walking and running. This frequency of repetition, also known as dominant frequency, is a descriptive feature and thus, it has been taken into consideration.

4. Standardization

Since, the time domain features are measured in (m/s 2), while the frequency ones in (Hz), therefore, all features should have the same scale for a fair comparison between them, as some classification algorithms use distance metrics. In this step, Z-Score standardization is used, which will transform the attributes to have zero mean and unit variance, and is defined as

$$x_{new} = (x - \mu)/\sigma$$

where μ and σ are the attribute's mean and standard deviation respectively.

5. PROJECT SYSTEM DESIGN

5.1. SYSTEM ARCHITECTURE

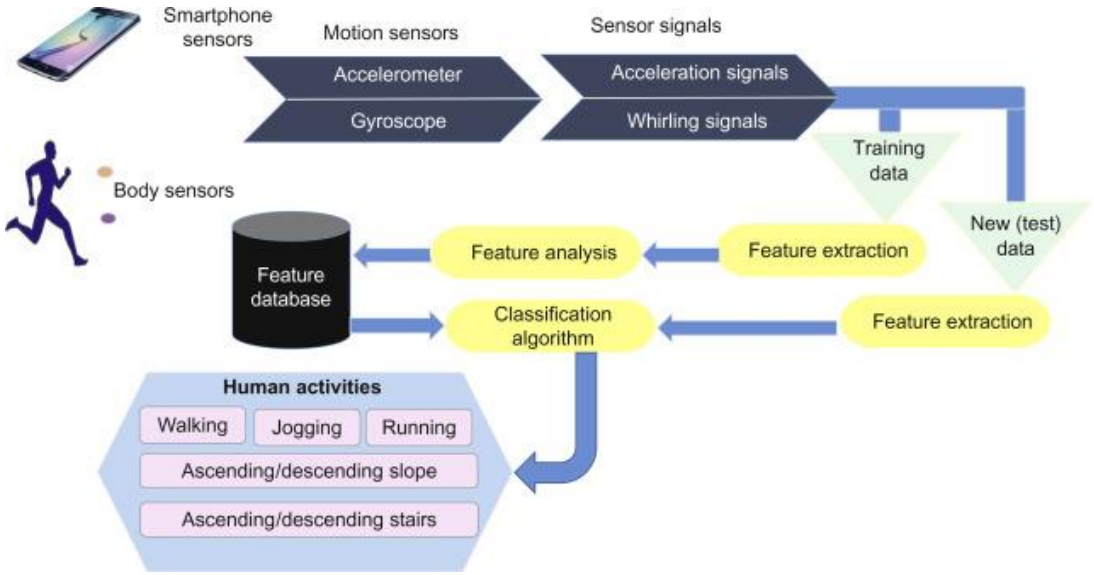


Fig. 5.1: User Functional Model Image

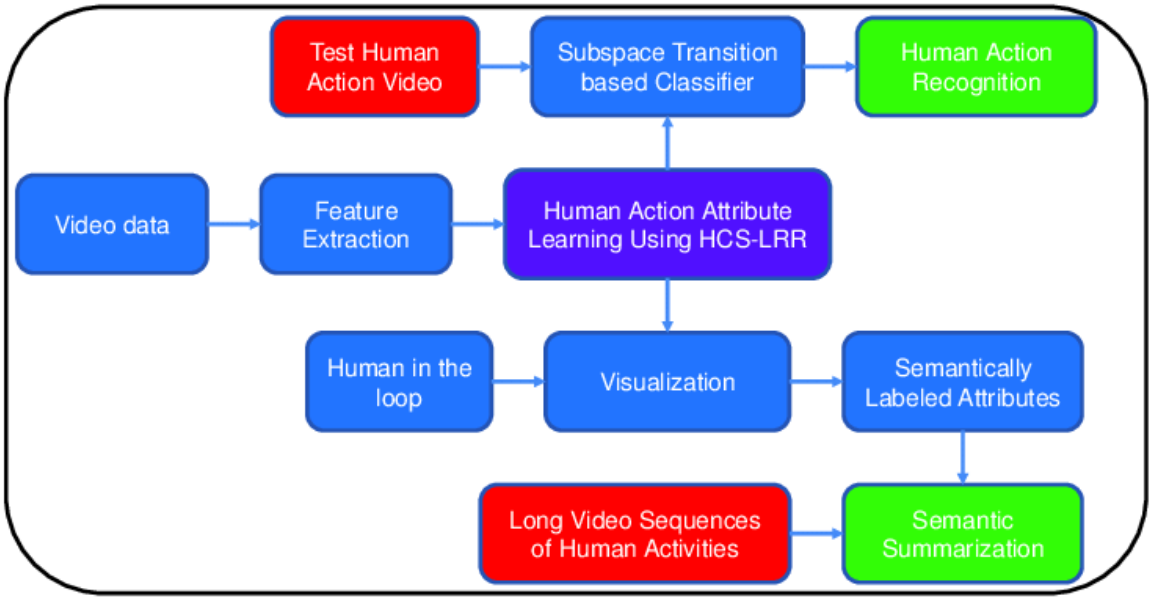


Fig. 5.2: Block Diagram of Model

5.2. UML DIAGRAMS

UML stands for Unified Modeling Language. UML is a standardized general-purpose modeling language in the field of object-oriented software engineering. The standard is managed, and was created by, the Object Management Group.

The goal is for UML to become a common language for creating models of object oriented computer software. In its current form UML is comprised of two major components: a Meta-model and a notation. In the future, some form of method or process may also be added to; or associated with, UML.

The Unified Modeling Language is a standard language for specifying, Visualization, Constructing and documenting the artifacts of software system, as well as for business modeling and other non-software systems.

The UML represents a collection of best engineering practices that have proven successful in the modeling of large and complex systems.

The UML is a very important part of developing objects oriented software and the software development process. The UML uses mostly graphical notations to express the design of software projects.

USE CASE DIAGRAM:

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.

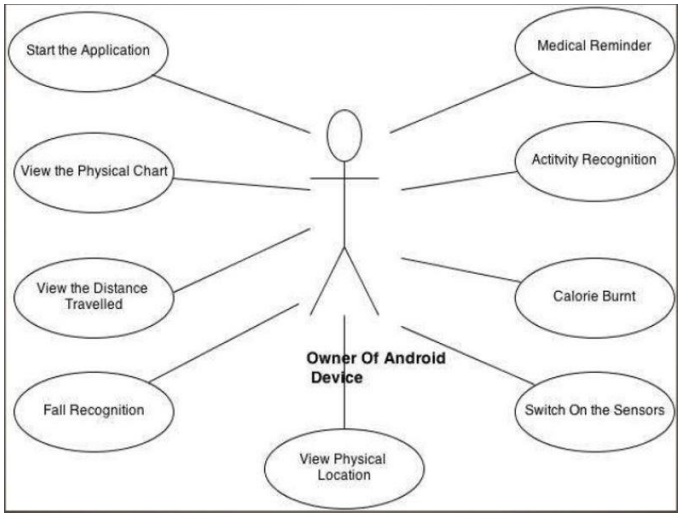


Fig. 5.3: Use Case Diagram

CLASS DIAGRAM:

In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.

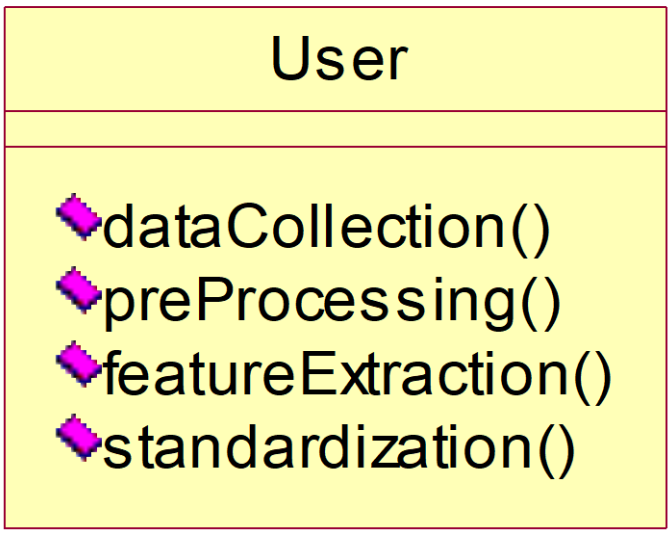


Fig. 5.4: Class Diagram

SEQUENCE DIAGRAM:

A sequence diagram in Unified Modeling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.

Fig.5.5: Sequence Diagram

ACTIVITY DIAGRAM:

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modeling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.

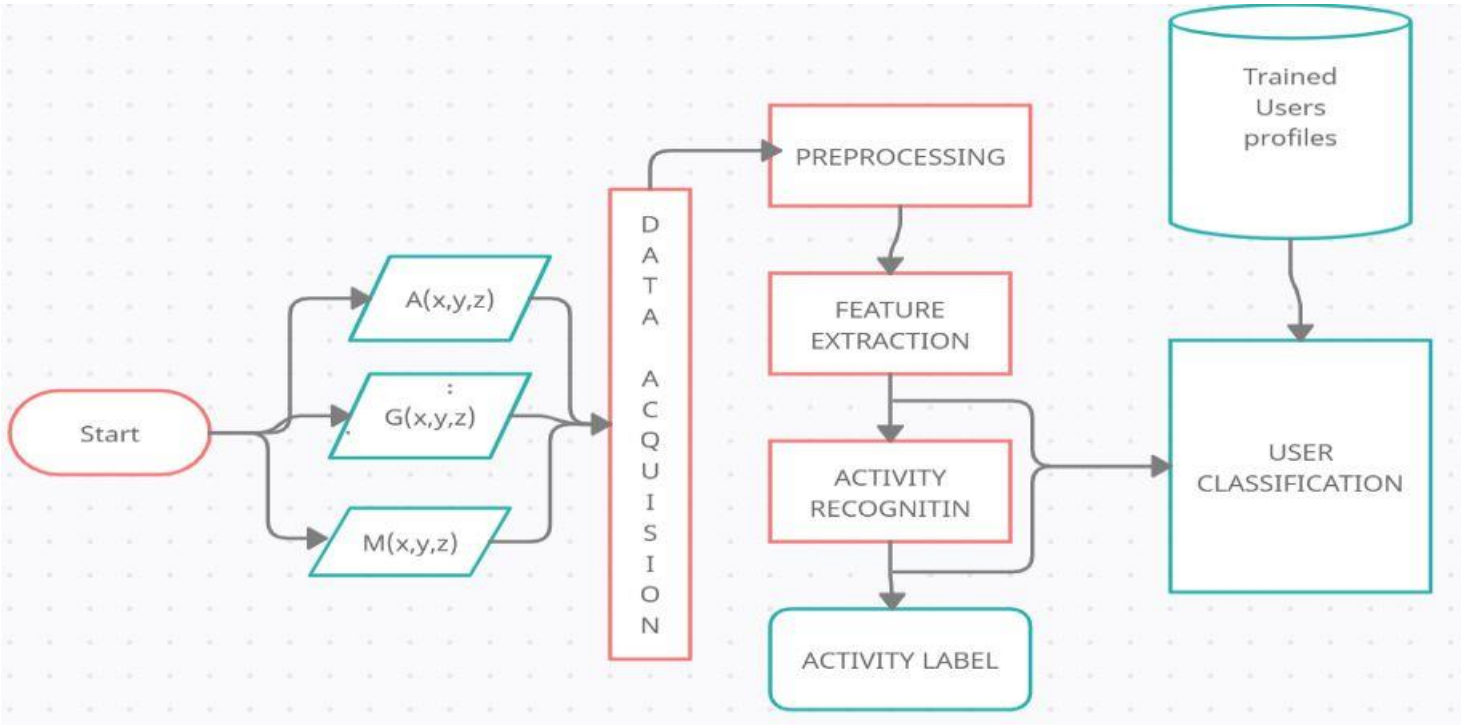


Fig.5.6: Activity Diagram

OBJECT DIAGRAM:

An object diagram is a UML structural diagram that shows the instances of the classifiers in models. Object diagrams use notation that is similar to that used in class diagrams. Class diagrams show the actual classifiers and their relationships in a system.

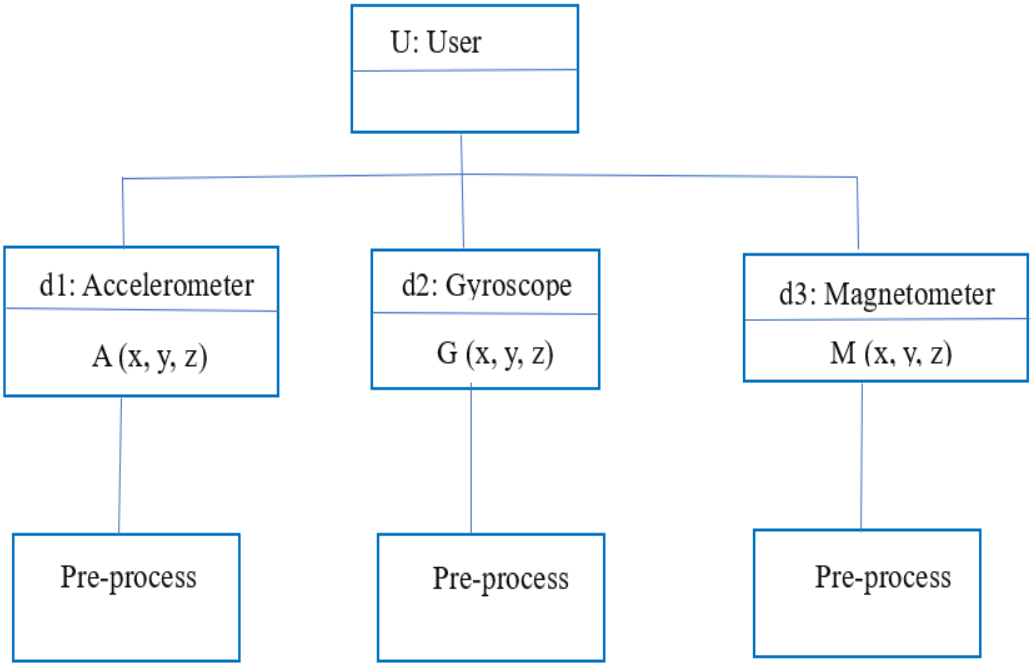


Fig.5.7: Object Diagram

6. PROJECT CODING

6.1. CODE TEMPLATE:

Load and concatenate all files into Single Data Frame

```
/*
        df=pd.DataFrame()

        for file in files:
            df_temp=pd.read_csv(path+file,header=1)
            df=pd.concat([df,df_temp], sort=False)
         df
   */
```

PREPROCESSING

Split the Data and Labels

```
/*

X_train = train_df[train_df.columns[:9]]
Y_train = train_df[train_df.columns[9:10]]


*/
```

Encode Labels into Numbers

```
/*

from sklearn.preprocessing import LabelEncoder
encoder=LabelEncoder()
y_train=encoder.fit_transform(y_train)
y_train[:10]

*/
```

Split training data into training and validation data

```
/*

def train_test_split(X, y, split_size=0.8):
    split= int(len(X) * split_size)
    train_x = X[:split]
    train_y = y[:split]
    test_x = X[split:]
    test_y = y[split: ]
    return train_x, test_x, train_y, test_y

X_train,X_test,y_train, y_test =train_test_split(X_train, y_train)

print("X_train shape ", X_train.shape)
print("Y train shape ", y_train.shape)
print("X_test shape ", X_test.shape)
print("y_test shape ", y_test.shape)


*/
```

Create and Compile LSTM Model

```
/*
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense,Flatten, LSTM
from tensorflow.keras.regularizers import l2
from tensorflow.keras.optimizers import Adam


model=Sequential()
model.add(LSTM(32, return_sequences=True, input_shape = (n_time_steps, n_features),
            kernel_regularizer = l2(0.000001), bias_regularizer = l2(0.000001), name='lstm_1'))
model.add(Flatten(name='flatten'))
model.add(Dense(64, activation='relu',kernel_regularizer = l2(0.000001), bias_regularizer = l2(0.000001),
name='dense_1' ))
model.add(Dense(len(np.unique(y_train)), activation='softmax',
              kernel_regularizer = l2(0.000001), bias_regularizer = l2(0.000001), name='output'))
model.summary()


*/
```

6.2. OUTLINE FOR VARIOUS FILES

urllib.request : defines functions and classes which help in opening URLs (mostly HTTP) in a complex world

sklearn.preprocessing : package provides several common utility functions and transformer classes to change raw feature vectors into a representation.

TimeseriesGenerator : They are used to easily handle time series.

keras.models : This is the class from which all layers inherit

keras.regularizers : Layer that reshapes inputs into the given shape

class.LSTMCell : Cell class for the LSTM layer

keras.optimizers : An optimizer is one of the two arguments required for compiling a Keras model

6.3. CLASS WITH FUNCTIONALITY

```
#check for unique labels

train_df.Activity.unique{
   /*
        Identifies uniquely defined activities
         */
}
array(['walking', 'standing', 'jogging', 'sitting', 'biking', 'upstairs',

                                'downstairs', 'upsatirs'], dtype=object)


#Time Series generation class

TimeseriesGenerator{

/*

X_train.to_numpy(), y_train, length=n_time_steps, batch_size=1024

Converts data and labels into Time Series Sequence

*/
```

```
}
#Creating Model Checkpoints
/*
ModelCheckpoint{
'model.h5', save_weights_only=False, save_best_only=True, verbose=1
Creates checkpoints to save the best results of the model
*/
}
```

6.4. METHODS INPUTS AND OUTPUT PARAMETERS

- public Input()
- public Add()
- public Sequential()
- public LSTM()
- public Flatten()
- public Dense()
- public Summary()
- public Compile()
- public ModelCheckpoint()

7. PROJECT TESTING

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub-assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of tests. Each test type addresses a specific testing requirement.

7.1. VARIOUS TEST CASES

Unit testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

Integration testing

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

Functional test

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items:

Valid Input : identified classes of valid input must be accepted.

Invalid Input : identified classes of invalid input must be rejected.

Functions : identified functions must be exercised.

Output : identified classes of application outputs must be exercised.

Systems/Procedures : interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

System Test

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration-oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

7.2. BLACK BOX

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box .you cannot “see” into it. The test provides inputs and responds to outputs without considering how the software works.

7.3. WHITE BOX TESTING

White Box Testing is a testing in which in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is purpose. It is used to test areas that cannot be reached from a black box level.

8. OUTPUT SCREENS

8.1. USER INTERFACES

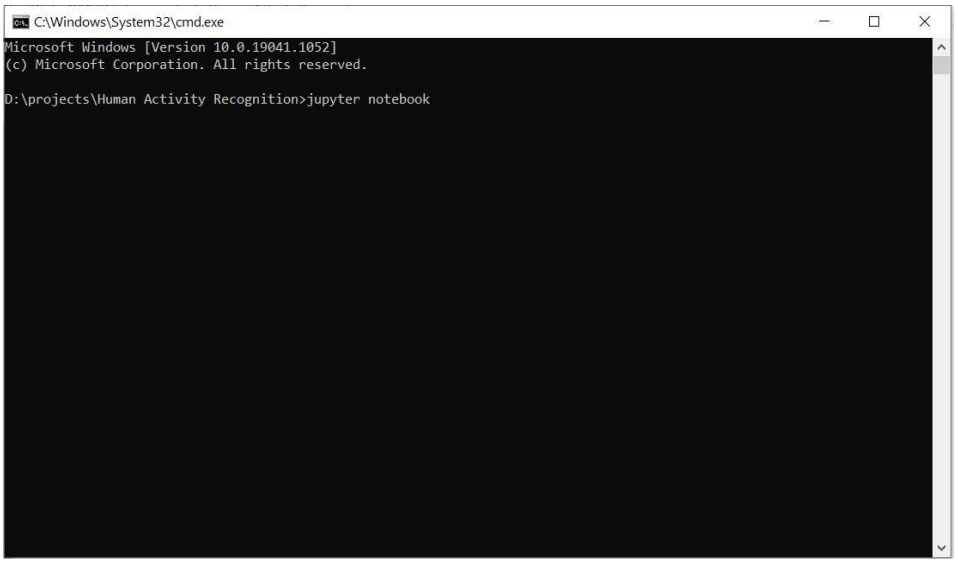


Fig 8.1.1. Open Jupyter Notebook

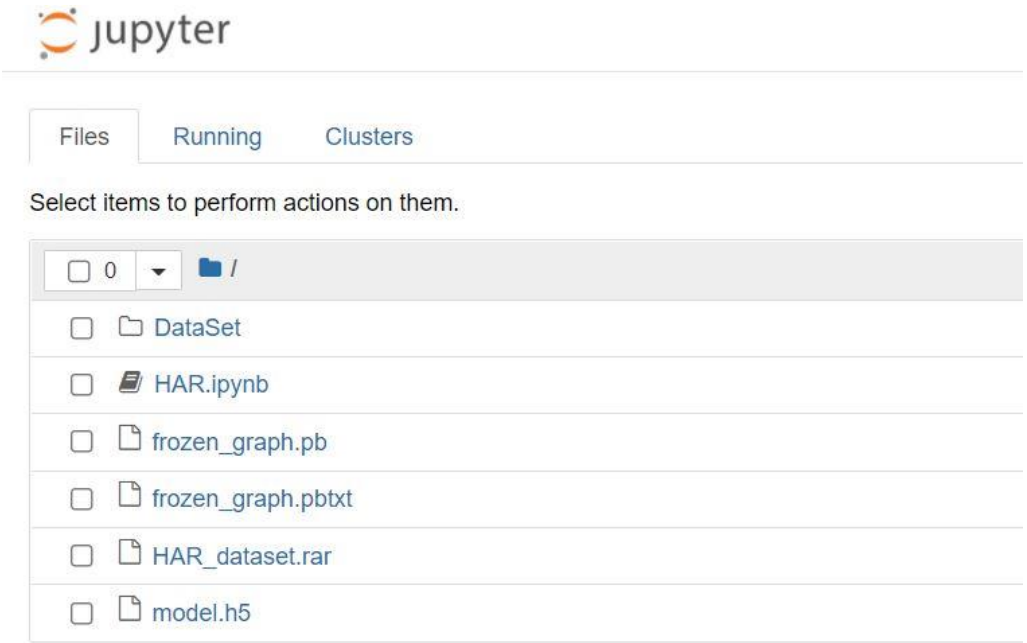


Fig 8.1.2. Dataset folder in Jupyter Notebook

8.2. OUTPUT SCREENS

Human Activity Recognition

step1 Install required Modules

```
In [1]: !pip install tensorflow
        !pip install keras==2.4.3
        !pip install numpy==1.19.5
        !pip install matplotlib==3.3.3
```

```
Requirement already satisfied: tensorflow in c:\users\hp\appdata\local\programs\python\python39\lib\site-packages (2.5.0)
Requirement already satisfied: protobuf>=3.9.2 in c:\users\hp\appdata\local\programs\python\python39\lib\site-packages (from
tensorflow) (3.17.1)
Requirement already satisfied: wrapt~=1.12.1 in c:\users\hp\appdata\local\programs\python\python39\lib\site-packages (from te
nsorflow) (1.12.1)
Requirement already satisfied: wheel~=0.35 in c:\users\hp\appdata\local\programs\python\python39\lib\site-packages (from tens
orflow) (0.36.2)
Requirement already satisfied: typing-extensions~=3.7.4 in c:\users\hp\appdata\local\programs\python\python39\lib\site-packag
es (from tensorflow) (3.7.4.3)
Requirement already satisfied: flatbuffers~=1.12.0 in c:\users\hp\appdata\local\programs\python\python39\lib\site-packages (f
rom tensorflow) (1.12)
Requirement already satisfied: opt-einsum~=3.3.0 in c:\users\hp\appdata\local\programs\python\python39\lib\site-packages (fro
m tensorflow) (3.3.0)
Requirement already satisfied: numpy~=1.19.2 in c:\users\hp\appdata\local\programs\python\python39\lib\site-packages (from te
nsorflow) (1.19.5)
```

Fig 8.2.1. Output Screen

step 2 : Download the DataSet

```
2]: import urllib.request

    print('Downloading dataset')

    url = 'https://www.utwente.nl/en/eemcs/ps/dataset-folder/sensors-activity-recognition-dataset-shoaib.rar'
    urllib.request.urlretrieve(url, 'HAR_dataset.rar')

    print('Download completed')
```

```
Downloading dataset
Download completed
```

Fig 8.2.2. Dataset Download

9. EXPERIMENTAL RESULTS

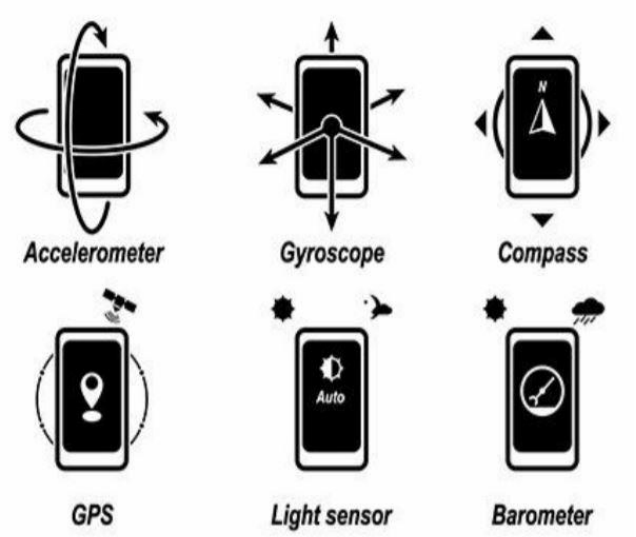


Fig 9.1. Smart Phone Sensors

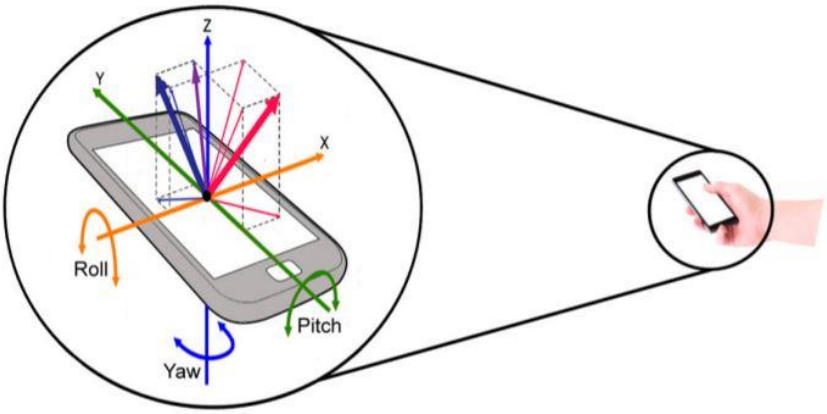


Fig 9.2. Accelerometer Spatial Axis

Visualize the difference in change in values w.r.t activity ¶

In [6]:
```
# change in values while walking

plt.plot(np.arange(0,100),df.Ax[df['Unnamed: 69']=="walking"][:100], label='X-axis')
plt.plot(np.arange(0,100),df.Ay[df['Unnamed: 69']=="walking"][:100], label='Y-axis')
plt.plot(np.arange(0,100),df.Az[df['Unnamed: 69']=="walking"][:100], label='Z-axis')
plt.legend()
```

Out[6]: <matplotlib.legend.Legend at 0x21e4a22dd60>

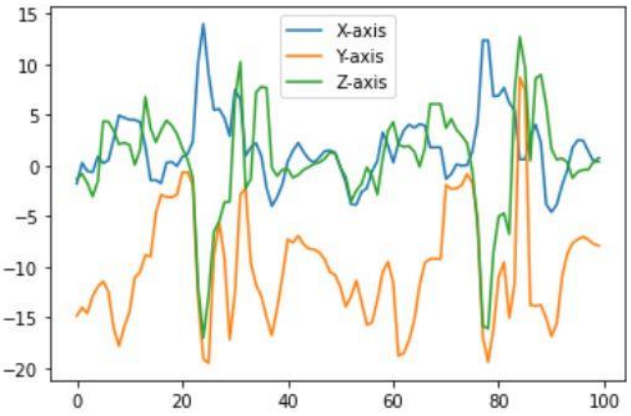


Fig 9.3. Walking Activity Visualization

In [7]:
```
# change in values while standing

plt.plot(np.arange(0,100),df.Ax[df['Unnamed: 69']=="standing"][:100], label='X-axis')
plt.plot(np.arange(0,100),df.Ay[df['Unnamed: 69']=="standing"][:100], label='Y-axis')
plt.plot(np.arange(0,100),df.Az[df['Unnamed: 69']=="standing"][:100], label='Z-axis')
plt.legend()
```

Out[7]: <matplotlib.legend.Legend at 0x21e22c40130>

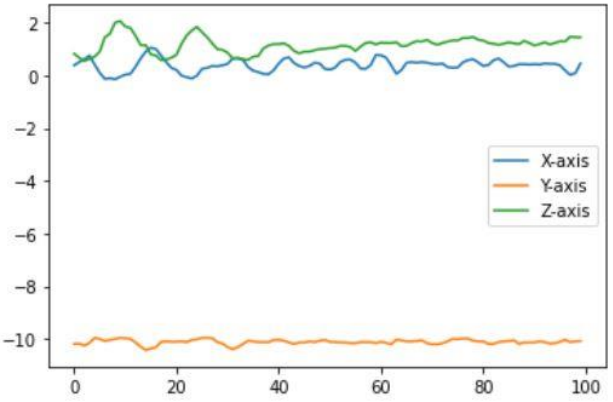


Fig 9.4. Standing Activity Visualization

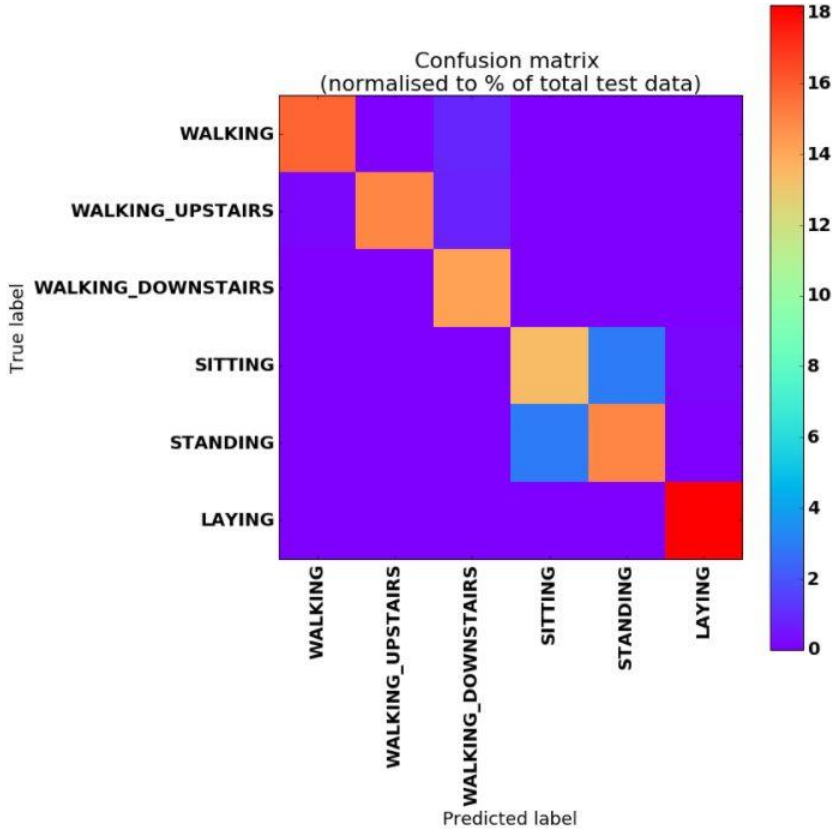


Fig 9.5. Confusion Matrix

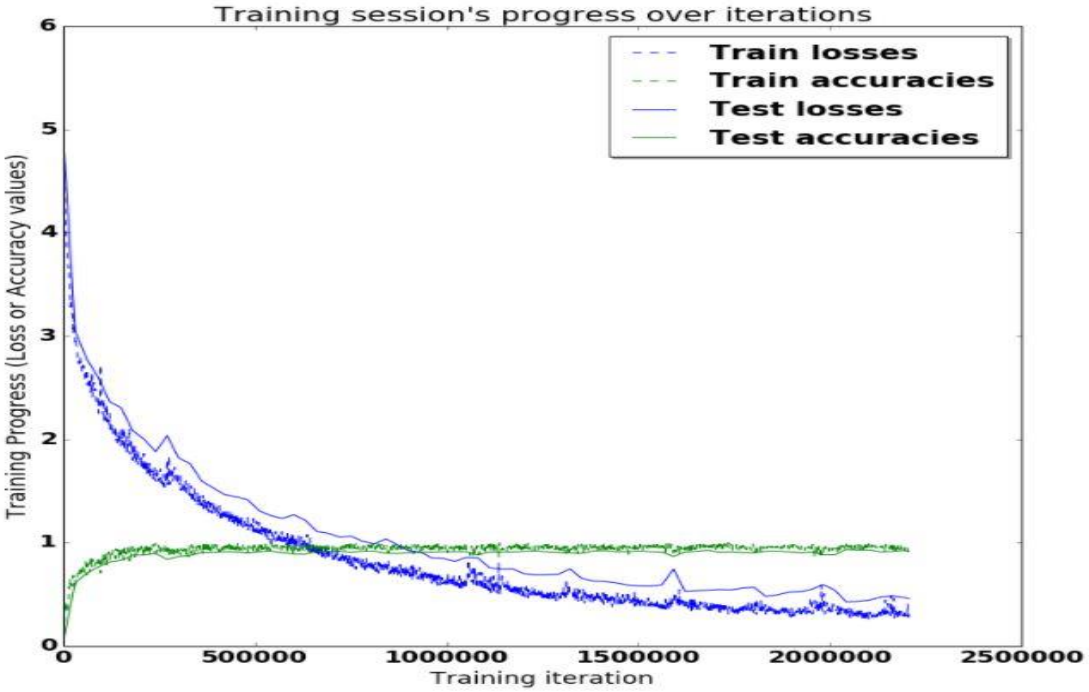


Fig 9.6. Accuracy Graph

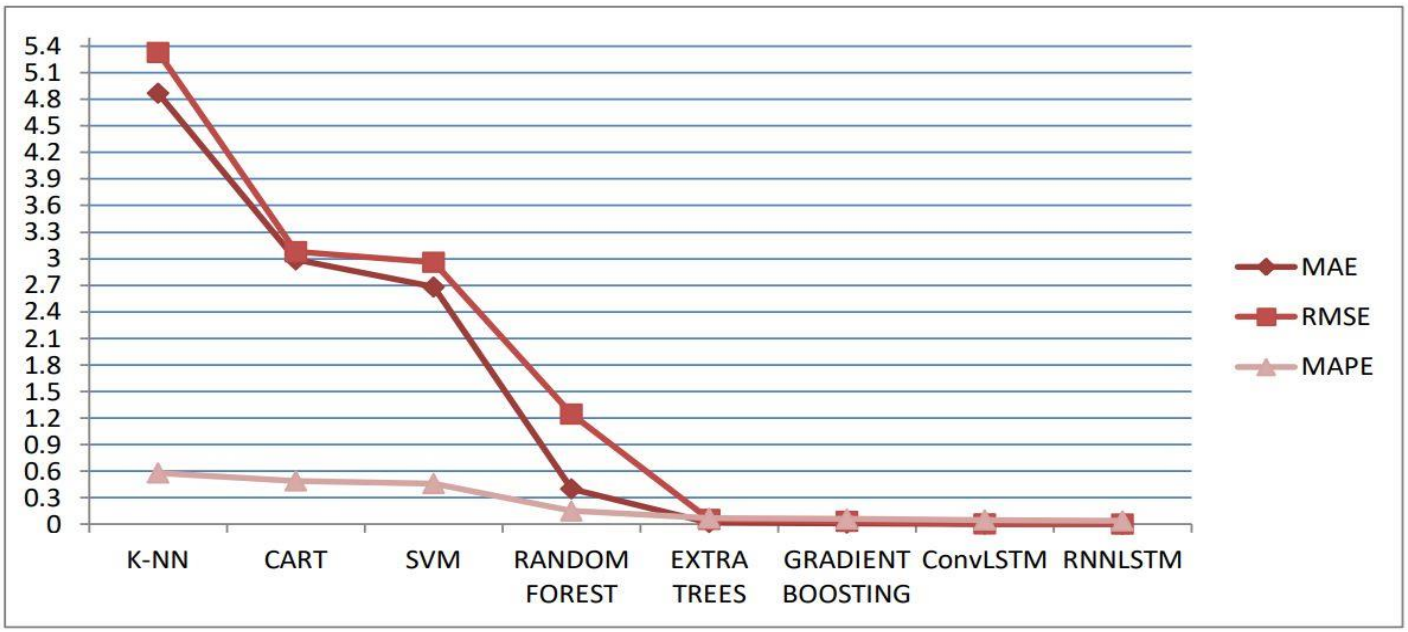


Fig 9.7. Comparison of MAE, RMSE, MAPE for different Models

Start Training

```
In [29]: history = model.fit(train_gen, epochs=5, validation_data=test_gen, callbacks=callbacks)

         Epoch 1/5
         985/985 [==============================] - 365s 367ms/step - loss: 0.4113 - accuracy: 0.8648 - val_loss: 0.0806 - val_accurac
         y: 0.9809

         Epoch 00001: val_loss improved from inf to 0.08060, saving model to model.h5
         Epoch 2/5
         985/985 [==============================] - 334s 339ms/step - loss: 0.0958 - accuracy: 0.9790 - val_loss: 0.2233 - val_accurac
         y: 0.9492

         Epoch 00002: val_loss did not improve from 0.08060
         Epoch 3/5
         985/985 [==============================] - 336s 342ms/step - loss: 0.0737 - accuracy: 0.9824 - val_loss: 0.1040 - val_accurac
         y: 0.9581

         Epoch 00003: val_loss did not improve from 0.08060
         Epoch 4/5
         985/985 [==============================] - 301s 305ms/step - loss: 0.0834 - accuracy: 0.9822 - val_loss: 0.0868 - val_accurac
         y: 0.9627
```

Fig 9.8. Testing Of Model For 5 Epochs

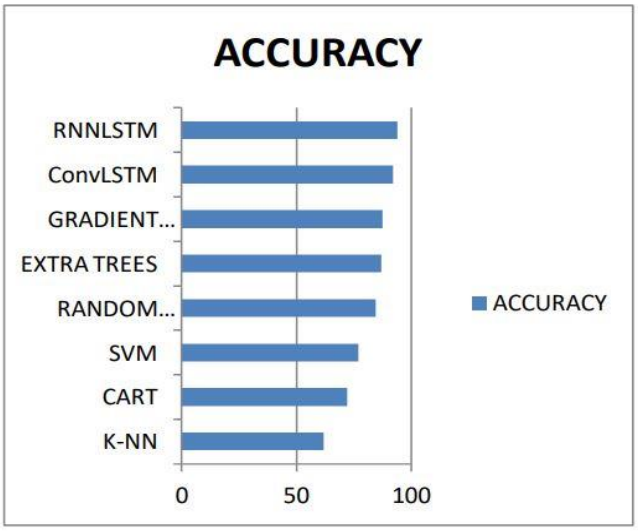


Fig 9.9. Accuracy of Classifiers

CLASSIFIER	ACCURACY (%)	Mean Absolute Error [MAE]	Root Mean Square Error [RMSE]	Mean Absolute Percentage Error [MAPE] (%)
K- Nearest Neighbor [K-NN]	61.893	4.87	5.33	58
Classification and Regression Trees [CART]	72.14	2.987	3.08	49.4
Support Vector Machine [SVM]	76.96	2.68	2.96	46.7
Random Forest	84.66	0.4	1.245	15.8
Extra Trees	86.90	0.02	0.0554	6.90
Gradient Boosting	87.61	0.0128	0.0336	6.78
Convolutional Neural Network with Long Short Term Memory [ConvLSTM]	**92.24**	**0.0007**	**0.0044**	**5.91**
Recurrent Neural Network with Long Short Term Memory [RNNLSTM]	**93.89**	**0.0004**	**0.0024**	**4.78**

Table . 9.1: Performance Metrics of Classifiers

10. CONCLUSION AND FUTURE ENHANCEMENT

Smart phones are pervasive and winding up increasingly modern. This has been changing the scene of individuals' day by day life and has opened the entryways fascinating information mining applications. Human action acknowledgment is a center structure hinder behind these applications. It takes the crude sensor's perusing as sources of info and predicts a client's movement action.

This project exhibits an extensive overview of the ongoing advances up to 93.89% on different movement acknowledgment with PDA's sensors. The information of various classifiers was utilized for assessing acknowledgment execution.

Consolidating several classifiers, best classifiers utilizing the normal of probabilities strategy ended up being the best classifier for movement acknowledgment, outperforming all other classifiers. Further demonstrated the acknowledgment strategy can identify exercises autonomous of cell phone’s position.

For future work, movement acknowledgment task in a few different ways. To start with plan to perceive extra exercises. Second, we want to gather information from more clients of different ages. Third, intend to remove more highlights that could more likely segregate various exercises.

REFERENCES

[1] Wang, A_new_approach_to_Human_Activity_Recognition_using_Machine_Learning_techniques Jiang, X., Cao, R, 17195 (2019).

[2] Aha, D.W., Kibler, D., Albert, M.K.: Instance-based learning algorithms. Machine Learning 6(1), 37–66 (2020).

[3] N. D. Lane, E. Miluzzo, H. Lu, D. Peebles, T. Choudhury, and A. T.Campbell, "A survey of mobile phone sensing", IEEE Commun. Mag, Vol. 48, PP.140–150, Sep, 2010.

[4] Kwapisz, Jennifer R Weiss, Gary M and Moore Samuel A, "Cell phone-based biometric identification, Biometrics: Theory Applications and Systems (BTAS)", Fourth IEEE International Conference, 2010.

[5] T. Brezmes, J. L. Gorricho, and J. Cotrina, "Activity recognition from accelerometer data on mobile phones", IWANN '09: Proc. the 10th International Work Conference on Artificial Neural Networks, 796– 799, 2009.

[6] Casale Pierluigi, Pujol Oriol and RadevaPetia, "Human activity recognition from accelerometer data using a wearable device", Pattern Recognition and Image Analysis, Springer, 289-296, 2011.

[7] Poppe Ronald, "A survey on vision-based human action recognition", Image and vision computing, Elsevier, 2010.

[8] Krishnan Narayanan C, Colbry Dirk, Juillard Colin and Panchanathan Sethuraman, "Real time human activity recognition using tri-axial accelerometers", Sensors signals and information processing workshop, 2008.

[9] Khan, Adil Mehmood, Lee, Young-Koo Lee, Sungyoung Y and Kim, Tae-Seong, A triaxial accelerometer-based physical-activity recognition via augmented-signal features and a hierarchical recognizer, Information Technology in Biomedicine, IEEE Transactions, 2010.

[10] Badawi, H., Saddik, A.E.: Towards a context-aware biofeedback activity recommendation mobile application for healthy lifestyle. Procedia Computer Science 21, 382 – 389 (2013).

[11] Kwapisz, Jennifer R Weiss, Gary M and Moore Samuel A, "Activity recognition using cell phone accelerometers", ACM SigKDD Explorations Newsletter, 2011.

[12] T. S. Saponas, J. Lester, J. E. Froehlich et al., "Ilearn on the Iphone: Real-time human activity classification on commodity mobile phones," CSE Technical Report, 2008.

[13] J. Goldman et al, "Participatory sensing: A citizen-powered approach to illuminating the patterns that shape our world", 2009.

[14] Westerterp, Klaas R, "Assessment of physical activity: a critical appraisal, European journal of applied physiology", The National Centre for Biotechnology Information, 2009.

[15] de Souza J´unior, A.H., Corona, F., Barreto, G.A., Miche, Y., Lendasse, A.: Minimal learning machine: A novel supervised distance-based approach for regression and classification. Neurocomputing 164, 34 – 44 (2015).

[16] Wang, Y., Jiang, X., Cao, R., Wang, X.: Robust indoor human activity recognition using wireless signals. Sensors 15(7), 17195 (2015).

[17] Mesquita, D.P.P., Gomes, J.P.P., Junior, A.H.S.: Ensemble of minimal learning machines for pattern

classification. In: Advances in Computational Intelligence - 13th International Work-Conference on Artificial Neural Networks, IWANN 2015, Palma de Mallorca, Spain, June 10-12, 2015. Proceedings, Part II. pp. 142–152 (2015).

[18] Attal, F., Mohammed, S., Dedabrishvili, M., Chamroukhi, F., Oukhellou, L., Amirat, Y.: Physical human activity recognition using wearable sensors. Sensors 15(12), 29858 (2015).

[19] Bayat, A., Pomplun, M., Tran, D.A.: A study on human activity recognition using accelerometer data from smartphones. Procedia Computer Science 34, 450 – 457 (2014).

[20] Chen, K.H., Chen, P.C., Liu, K.C., Chan, C.T.: Wearable sensor-based rehabilitation exercise assessment for knee osteoarthritis. Sensors 15(2), 4193

PUBLICATION

JOURNAL:

HUMAN ACTIVITY RECOGNITION USING MACHINE LEARNING WITH DATA ANALYSIS

CONFERENCE:

- INTERNATIONAL CONFERENCE ON INNOVATIONS IN COMPUTER NETWORKS,COMPUTATIONAL INTELLIGENCE AND IOT [ICICCI-2021].
- Paper ID: ICICCI – 21 – 0146

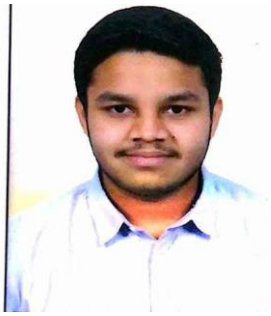
K G Neha Reddy is currently pursuing her Bachelor of Technology in the Computer Science and Engineering stream at St Martin's Engineering college. She has completed her Grade 10th and 12th from Army Public School, bollorum.

Her Participations include: National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020, "Webinar on Leadership Talks" conducted by MHRD's Innovation Cell on 16th May 2020 and India's First Leadership Talk conducted by Institution's Innovation Council (IIC) of MHRD's Innovation Cell on 6th June 2020 and participated in "Anti Drug's Campaign" conducted by St. Martin's Engineering College in 2018. Her areas of interest are animation, Cyber Security, Python, Artificial Intelligence, Machine Learning and Deep Learning. She completed a few certification courses from online platforms like Coursera and CurseApp.

Padala Arvind Chary is currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. . He completed his intermediate from Sri Chaitanya Junior College and 10th class from Ravindra Bharathi School. He is one of the members of Employability Skill Development Program initiated by Zensar.

His technical skills include C, C++ and Java. He is a member student of Smart Interviews. His participations include: National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020, "Know More - Teach More ", the Global Webinar on Cloud & Big Data – Changing the way we work which was conducted Global Education & Careers Forum (GECF) on 12th August 2020, a two day hackaton program Esummit 2017 conducted at MLRIT and IIC Online Sessions conducted by Institution's Innovation Council (IIC) of MHRD's Innovation Cell, New Delhi to promote Innovation, IPR, Entrepreneurship, and Start-ups among HEIs from 28th April to 22nd May 2020.

He was also a member of Street Cause for two years 2018-20 and member in TAM (Technical Awareness Month) program. His areas of interest are Python, Full Stack Development, Cryptocurrency, Artificial Intelligence, Machine Learning and Blockchain He also completed few certification courses from online platforms like Coursera and CursaApp.



Lohan Vaddepally is currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. He completed his intermediate education from Sri Chaitanya Junior College and 10th standard from Ken Crest International School. He took part in Employability Skill development Program conducted by Zensar. He is also a student of Smart Interviews.

His participations include: National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020, "Webinar on Leadership Talks" conducted by MHRD's Innovation Cell on 16th May 2020 and India's First Leadership Talk conducted by Institution's Innovation Council (IIC) of MHRD's Innovation Cell on 6th June 2020 and participated in "Anti Drug's Campaign" conducted by St. Martin's Engineering College in 2018. His areas of interest are Data Science, Cyber Security, Python, Artificial Intelligence, Machine Learning and Deep Learning. He completed few certification courses from online platforms like Coursera and CursaApp.

Bhukya Vinod Kumar is currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College He completed his Diploma in civil Engineering from Vivekananda group of institutions and 10th standard from ZPH School Domakonda. He took part in Employability Skill development Program conducted by Zensar. He is also a student of Smart Interviews.

His participations include: National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020, "Webinar on Leadership Talks" conducted by MHRD's Innovation Cell on 16th May 2020. His areas of interest are Data Science, Cyber Security, Python, Artificial Intelligence, Machine Learning and Deep Learning. He Achieved few certificates in college athletic games (400m and 800m).

APPENDICES

```
!pip install tensorflow
!pip install keras==2.4.3
!pip install numpy==1.19.5
!pip install matplotlib==3.3.3


import urllib.request

print('Downloading dataset')

url = 'https://www.utwente.nl/en/eemcs/ps/dataset-folder/sensors-activity-recognition-dataset-shoaib.rar'
urllib.request.urlretrieve(url, 'HAR_dataset.rar')

print('Download completed')



!pip3 install pyunpack
!pip3 install patool
from pyunpack import Archive
Archive('HAR_dataset.rar').extractall('.')


import tensorflow as tf
tf.__version__


import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import os


path='DataSet/'
files=[file for file in os.listdir(path) if file.endswith('.csv')]
files


# load a file into memory
df=pd.read_csv(path+files[0],header=1)
df.head()


df['Unnamed: 69'].unique()


# change in values while walking
```

```
plt.plot(np.arange(0,100),df.Ax[df['Unnamed: 69']=="walking"][:100], label='X-axis')
plt.plot(np.arange(0,100),df.Ay[df['Unnamed: 69']=="walking"][:100], label='Y-axis')
plt.plot(np.arange(0,100),df.Az[df['Unnamed: 69']=="walking"][:100], label='Z-axis')
plt.legend()



# change in values while standing

plt.plot(np.arange(0,100),df.Ax[df['Unnamed: 69']=="standing"][:100], label='X-axis')
plt.plot(np.arange(0,100),df.Ay[df['Unnamed: 69']=="standing"][:100], label='Y-axis')
plt.plot(np.arange(0,100),df.Az[df['Unnamed: 69']=="standing"][:100], label='Z-axis')
plt.legend()


# change in values while biking

plt.plot(np.arange(0,100),df.Ax[df['Unnamed: 69']=="biking"][:100], label='X-axis')
plt.plot(np.arange(0,100),df.Ay[df['Unnamed: 69']=="biking"][:100], label='Y-axis')
plt.plot(np.arange(0,100),df.Az[df['Unnamed: 69']=="biking"][:100], label='Z-axis')
plt.legend()


# change in values while jogging

plt.plot(np.arange(0,100),df.Ax[df['Unnamed: 69']=="jogging"][:100], label='X-axis')
plt.plot(np.arange(0,100),df.Ay[df['Unnamed: 69']=="jogging"][:100], label='Y-axis')
plt.plot(np.arange(0,100),df.Az[df['Unnamed: 69']=="jogging"][:100], label='Z-axis')
plt.legend()


df=pd.DataFrame()

for file in files:
    df_temp=pd.read_csv(path+file,header=1)
    df=pd.concat([df,df_temp], sort=False)
df


for i in df.columns:
    print(i)


# split out left and right pocket data

left_pocket = df[df.columns[1:10]]
left_pocket


right_pocket = df[df.columns[15:24]]
right_pocket.columns=left_pocket.columns
```

```
right_pocket


# concatenate left and right split in one data frame

train_df=pd.concat([left_pocket,right_pocket],sort=False)
train_df


# restore labels in dataframe
labels=pd.concat([df['Unnamed: 69'],df['Unnamed: 69']],axis=0, sort=False)
labels.columns=['Activity']
train_df['Activity']=labels
train_df


#check for unique labels
train_df.Activity.unique()


train_df.Activity.loc[(train_df.Activity == 'upsatirs')] = 'upstairs'
train_df.Activity.unique()



X_train = train_df[train_df.columns[:9]]
y_train = train_df[train_df.columns[9:10]]



X_train

y_train


y_train.Activity.unique()


from sklearn.preprocessing import LabelEncoder

encoder=LabelEncoder()
y_train=encoder.fit_transform(y_train)
y_train[:10]


def train_test_split(X, y, split_size=0.8):
    split= int(len(X) * split_size)
    train_x = X[:split]
    train_y = y[:split]
    test_x = X[split:]
    test_y = y[split: ]
    return train_x, test_x, train_y, test_y
```

```
X_train,X_test,y_train, y_test =train_test_split(X_train, y_train)

print("X_train shape ", X_train.shape)
print("Y train shape ", y_train.shape)
print("X_test shape ", X_test.shape)
print("y_test shape ", y_test.shape)


from tensorflow.keras.preprocessing.sequence import TimeseriesGenerator

n_time_steps = 100
n_features = 9

train_gen = TimeseriesGenerator(X_train.to_numpy(), y_train, length=n_time_steps, batch_size=1024)
test_gen = TimeseriesGenerator(X_test.to_numpy(), y_test, length=n_time_steps, batch_size=1024)


from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense,Flatten, LSTM
from tensorflow.keras.regularizers import l2
from tensorflow.keras.optimizers import Adam

model=Sequential()
model.add(LSTM(32, return_sequences=True, input_shape = (n_time_steps, n_features),
          kernel_regularizer = l2(0.000001), bias_regularizer = l2(0.000001), name='lstm_1'))
model.add(Flatten(name='flatten'))
model.add(Dense(64, activation='relu',kernel_regularizer = l2(0.000001), bias_regularizer = l2(0.000001),
name='dense_1' ))
model.add(Dense(len(np.unique(y_train)), activation='softmax',
           kernel_regularizer = l2(0.000001), bias_regularizer = l2(0.000001), name='output'))
model.summary()


#compile
model.compile(loss='sparse_categorical_crossentropy', optimizer=Adam(), metrics=['accuracy'])


# prepare callbacks
from tensorflow.keras.callbacks import ModelCheckpoint

callbacks= [ModelCheckpoint('model.h5', save_weights_only=False, save_best_only=True, verbose=1)]


history = model.fit(train_gen, epochs=5, validation_data=test_gen, callbacks=callbacks)


# stopping training here

# loading the best saved model

from tensorflow.keras.models import load_model
```

```
model=load_model('model.h5')
model.summary()


from tensorflow.python.framework.convert_to_constants import convert_variables_to_constants_v2
import numpy as np
#path of the directory where you want to save your model
frozen_out_path = ''
# name of the .pb file
frozen_graph_filename = "frozen_graph"
# Convert Keras model to ConcreteFunction
full_model = tf.function(lambda x: model(x))
full_model = full_model.get_concrete_function(
    tf.TensorSpec(model.inputs[0].shape, model.inputs[0].dtype))
# Get frozen ConcreteFunction
frozen_func = convert_variables_to_constants_v2(full_model)
frozen_func.graph.as_graph_def()
layers = [op.name for op in frozen_func.graph.get_operations()]
print("-" * 60)
print("Frozen model layers: ")
for layer in layers:
    print(layer)
print("-" * 60)
print("Frozen model inputs: ")
print(frozen_func.inputs)
print("Frozen model outputs: ")
print(frozen_func.outputs)
# Save frozen graph to disk
tf.io.write_graph(graph_or_graph_def=frozen_func.graph,
                  logdir=frozen_out_path,
                  name=f"{frozen_graph_filename}.pb",
                  as_text=False)
# Save its text representation
tf.io.write_graph(graph_or_graph_def=frozen_func.graph,
                  logdir=frozen_out_path,
                  name=f"{frozen_graph_filename}.pbtxt",
                  as_text=True)
```

A

PROJECT REPORT

On

DETECTION AND RECOGNITION OF CROP DISEASES AND INSECT PESTS BASED ON DEEP LEARNING IN HARSH ENVIRONMENTS

Submitted by

1)Ms.CH.Ravali(17K81A0567) 2)Ms.Ch.Aishwarya(17K81A0569)

3)Mr.K.Prem Sai(17K81A0589) 4)Ms.M.Shirisha (17K81A0595)

in partial fulfillment for the award of the degreeof

BACHELOR OF TECHNOLOGY

IN

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Under the Guidance of

Dr. M. Narayanan

Professor & HOD(CSE)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ST.MARTIN'S ENGINEERING COLLEGE
An Autonomous Institute

Dhulapally, Secunderabad – 500 100

JUNE 2021

BONAFIDE CERTIFICATE

This is to certify that the project entitled Detection and Recognition of Crop Diseases and Insect Pests Based On Deep Learning In Harsh Environments, is being submitted by **1.CH.Ravali (17K81A0567), 2.Ch.Aishwarya (17K81A0569), 3.K.Prem Sai (17K81A0589) 4. M.Shirisha (17K81A0595)** in partial fulfillment of the requirement for the award of the degree of **BACHELOR OF TECHNOLOGY IN COMPUTER SCIENCE AND ENGINEERING** is recorded of bonafide work carried out by them. The result embodied in this report have been verified and found satisfactory.

Dr.M.NARAYANAN
Professor & HOD(CSE)
Department of CSE

Head of the Department
Dr.M.NARAYANAN
Department of CSE

Internal Examiner

External Examiner

Place: Dhulapally
Date: 21-06-2021

DECLARATION

We, the student of **Bachelor of Technology** in Department of Computer Science and Engineering', session: <2017 – 2021>, St. Martin's Engineering College, Dhulapally, Kompally, Secunderabad, hereby declare that work presented in this Project Work entitled Detection and Recognition of Crop Diseases and Insect Pests Based On Deep Learning In Harsh Environments is the outcome of our own bonafide work and is correct to the best of our knowledge and this work has been undertaken taking care of Engineering Ethics. This result embodied in this project report has not been submitted in any university for award of any degree.

CH.Ravali 17K81A0567
Ch.Aishwarya 17K81A0569
K.Prem Sai 17K81A0589
M.Shirisha 17K81A0595

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible and whose encouragements and guidance have crowded effects with success.

We extended our deep sense of gratitude to Principal**, Dr. P. SANTOSH KUMAR PATRA**, St. Martin's Engineering College, Dhulapally, for permitting us to undertake this project.

We are also thankful to **Dr. M.NARAYANAN**, Head of the Department, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

We would like to thank **Dr. T. POONGOTHAI,** Department of Computer Science and Engineering (AI & ML) for her encouragement and insightful comments and as well as our project coordinators **Dr. B.RAJALINGAM**, Associate Professor and **Dr. G. GOVINDARAJULU,** Associate Professor, in Department of Computer Science and Engineering for their valuable support.

We would like to express our sincere gratitude and indebtedness to our project supervisor Dr.M.Narayanan, Professor & HOD (CSE), Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

Finally, we express thanks to all those who have helped us successfully to completing this project. Furthermore, we would like to thank our family and friends for their moral support and encouragement.

We express thanks to all those who have helped us in successfully completing the project.

CH.Ravali 17K81A0567

Ch.Aishwarya 17K81A0569

K.Prem Sai 17K81A0589

M.Shirisha 17K81A0595

ABSTRACT

Agricultural diseases and insect pests are one of the most important factors that seriously threaten agricultural production. Early detection and identification of pests can effectively reduce the economic losses caused by pests. In this project, convolutional neural network is used to automatically identify crop diseases. The data set comes from the public data set of the AI Challenger Competition in 2018, with 15 disease images of 3 crops. In this project, the Inception-ResNet-v2 model is used for training. The cross-layer direct edge and multi-layer convolution in the residual network unit to the model. After the combined convolution operation is completed, it is activated by the connection into the ReLu function. The experimental results show that the overall recognition accuracy is 86.1% in this model, which verifies the effectiveness. After the training of this model, we designed and implemented the Wechat applet of crop diseases and insect pests recognition. Then we carried out the actual test. The results show that the system can accurately identify crop diseases, and give the corresponding guidance.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF OUTPUT SCREENS

LIST OF ABBREVIATIONS

<ANN>	Artificial Neural Network
<AI>	Artificial Intelligence
<CGI>	Computer Generated Imagery
<CNN>	Convolutional Neural Network
<DCNN>	Deep Convolutional Neural Network
<DA>	Data Augmentation
<GB>	Giga Bytes
<IP>	Internet Protocol
<IPM>	Integrated Pest Management
<KNN>	K-Nearest Neighbors
<MAP>	Mean Average Precision
<NB>	Naïve Bayer
<NBAIR>	National Bureau Of Agricultural Insect Resources
<OO>	Object Oriented
<RAM>	Random Access Memory
<ReLu>	Rectified Linear Unit
<ResNet>	Residual Neural Network
<SSD>	Solid State Drive
<SVM>	Support Vector Machine
<UML>	Unified Modeling Language

<V2>	Version 2
<V3>	Version 3
<V4>	Version 4
<VGG>	Visual Geometry Group

1. INTRODUCTION

Image classification has been around for a long time and is a common research subject. In reality, it is used in the majority of major applications. In this paper, we use Convolutional Neural Networks to solve the problem of plant disease identification by analyzing the leaf of a plant (CNN).Plants, deep down in the food chain, are a big factor that ensures the nature of life on Earth. Since they are exposed to various natural environments, many of these plants are susceptible to various diseases. As shown in Figure 1, these diseases have resulted in significant agricultural losses. Detecting and treating these diseases at an early stage saves a lot of money and time.

We devised a system that uses deep learning to evaluate, identify, and classify any disease that might have affected a plant by taking an image of the leaf as a solution to this issue. The following is the processing flow:

1. In the given picture, a leaf is detected and cropped out.
2. After that, the extracted leaf is run through a classifier to determine which plant it belongs to.
3. The leaf is then screened for disease classes, if any, based on the results of the previous step.

Artificial intelligence's rapid growth in recent years has made life easier, and AI has become a well-known technology. AlphaGo, for example, beat the world Go champion. Siri and Alexa, Apple's and Amazon's voice assistants' are all examples of artificial intelligence technology portrayed by deep learning in a variety of fields. Image recognition has advanced significantly in recent years as a primary research topic in computer vision and artificial intelligence. The aim of image recognition in agricultural applications is to recognize and classify various types of images, as well as to analyze crop types, disease types, and severity, among other things. Then, in a timely and efficient manner, we can devise appropriate countermeasures to address various problems in agricultural production so as to ensure and increase crop yields and contribute to the betterment of agriculture. At the moment, crop disease research is primarily divided into two directions. The conventional physical approach, which is primarily focused on spectral detection to identify various diseases, is the first. Different diseases and insect pests cause different types of leaf damage, resulting in different spectral absorption and reflection of diseased and healthy crops' leaves. The other choice is to classify images using computer vision technology. To put it another way, the characteristics of disease images are collected using computer technology, and the identification is done using different characteristics of diseased and healthy plants. This paper develops a framework based on the Wechat applet to help farmers recognize and diagnose pests and diseases more easily and rapidly. The software will detect disease on the leaves of diseased crops, making it easier for farmers to consider disease and insect pest situations and seek expert advice. The machine uploads the image first, and then sends the data to the backend for processing through the network frontend. The aim of image preprocessing is to

improve the quality of the incoming image. First and foremost, the image is zoomed to meet the model input requirements; a large image would have a negative impact on recognition quality. Second, the image is cut randomly and the pixels are optimized in order to improve recognition quality. After the recognition is complete, the name and status of the crop with the highest matching degree will be granted. If the crop is unhealthy, the appropriate instructions will be given and sent to the mobile phone.

1.1 PROJECT OVERVIEW

In order to enable farmers to identify and detect pests and diseases conveniently and quickly, this project establishes a system based on Wechat applet. The program can identify the disease on the leaves of crops with diseases, which is convenient for farmers to understand the situation of diseases and insect pests and to obtain expert guidance. The system first uploads the image, and then transmits the image data to the back-end for processing through the network frontend. Image preprocessing is mainly to optimize the incoming image. First of all, the image is zoomed to meet the requirements of the model input, too large image will seriously affect the efficiency of recognition. Secondly, in order to achieve higher recognition efficiency, the image is cut randomly and the pixels are optimized. Finally, the name and status of the crop with the highest matching degree will be given after the recognition is completed.

1.2 PROJECT OBJECTIVES

Our project objectives include:

- Image classification has been around for a long time and is a common research subject.
- In reality, it is used in the majority of major applications.
- In this project, we use Convolutional Neural Networks to solve the problem of plant disease identification by analysing the leaf of a plant (CNN).
- Plants, deep down in the food chain, are a big factor that ensures the nature of life on Earth.
- Since they are exposed to various natural environments, many of these plants are susceptible to various diseases.
- These diseases have resulted in significant agricultural losses.

- Detecting and treating these diseases at an early stage saves a lot of money and time.

- We devised a system that uses deep learning to evaluate, identify, and classify any disease that might have affected a plant by taking an image of the leaf as a solution to this issue.

1.3 ORGANIZATION OF CHAPTERS

This documentation consists of 10 different chapters. Chapter 1 is Introduction. This chapter covers the overview of our project and its objectives. Chapter 2 is Literature Survey. This includes the details of our survey. Chapter 3 is Software and Hardware Requirements. We specify our software and hardware requirements here. Chapter 4 is Software Development Analysis. This section includes the problem definition and details of the modules we used in our project. Chapter 5 is Project System Design. This chapter includes the design part of our project which includes UML diagrams. Chapter 6 is Project Coding. This section contains the details of our project code. Chapter 7 is Project Testing. The details of test cases and testing are included in this chapter. Chapter 8 is Output Screens. This contains the screenshots of how our project looks like when executed. Chapter 9 is Experimental Results. This chapter contains the screenshots of our results. Chapter 10 is Conclusion and Future Enhancements. This covers the conclusion of our project and the possible future developments.

2. LITERATURE SURVEY

2.1 SURVEY ON BACKGROUND

1. Image Recognition of Crop Diseases and Insect Pests Based on Deep Learning

AUTHORS: Mingyuan Xin and Yong Wang

Deep learning algorithms have the advantages of clear structure and high accuracy in image recognition. Accurate identification of pests and diseases in crops can improve the pertinence of pest control in farmland, which is beneficial to agricultural production. It proposes a DCNN-G model based on deep learning and fusion of Google data analysis, using this model to train 640 data samples, and then using 5000 test samples for testing, selecting 80% as the training set and 20% as the test set, and compare the accuracy of the model with the conventional recognition model.

2. Tomato Diseases and Pests Detection Based on Improved Yolo V3 Convolutional Neural Network

AUTHORS: Jun Liu and Xuewei Wang

Tomato is affected by various diseases and pests during its growth process. If the control is not timely, it will lead to yield reduction or even crop failure. Based on the application of deep learning object detection, not only can save time and effort, but also can achieve real-time judgment, greatly reduce the huge loss caused by diseases and pests, which has important research value and significance. Based on the latest research results of detection theory based on deep learning object detection and the characteristics of tomato diseases and pests images, this study will build the dataset of tomato diseases and pests under the real natural environment, optimize the feature layer of Yolo V3 model by using image pyramid to achieve multi-scale feature detection, improve the detection accuracy and speed of Yolo V3 model, and detect the location and category of diseases and pests of tomato accurately and quickly.

3. Automatic greenhouse insect pest detection and recognition based on a cascaded deep learning classification method

AUTHORS: Dan Jeric Arcega Rustia, Jun-Jee Chao, Lin-Ya Chiu, Ya-Fang Wu, Jui-Yung Chung, Ju-Chun Hsu, Ta-Te Lin

Inspection of insect sticky paper traps is an essential task for an effective integrated pest management (IPM) programme. However, identification and counting of the insect pests stuck on the traps is a very cumbersome task. Therefore, an efficient approach is needed to alleviate the problem and to provide timely information on insect pests. In this research, an automatic method for the multi-class recognition of small-size greenhouse insect pests on sticky paper trap images acquired by wireless imaging devices is proposed. The developed algorithm features a cascaded approach that uses a convolutional neural network (CNN) object detector and CNN image classifiers, separately. The object detector was trained for detecting objects in an image, and a CNN classifier was applied to further filter out non-insect objects from the detected objects.

4. Protecting the Farming Land from Insects Damage to Growing Crops using Deep Convolutional Neural Network

AUTHORS: N.Abirami, P.Kavinilavan, M.Pooja, R.Vigneshand T.Kavitha

Rice cultivation is one of the most important economic sectors for Indian economy. With the increase in world population, the demand for the rice cultivation is also increasing. In order to increase the growth of rice crop, it is necessary to detect the pests in an earlier stage to minimize the pest growth. But our farmers are still struggling to protect the crops from external threats particularly from insects in agricultural lands. To overcome this problem, we are providing a solution to protect the crops in the farming lands using deep networks. Hence, the lives of farmers are saved from their struggle. In this paper, we proposed a system that will help the farmers in detecting rice crop pest using deep convolutional neural network with VGG16 architecture. Then, the proposed model is compared with the existing models GoogleNet and AlexNet.

5. Insect classification and detection in field crops using modern machine learning techniques

AUTHORS: Thenmozhi Kasinathan, Dakshayani Singaraju, Srinivasulu Reddy Uyyala

The agriculture sector has an immense potential to improve the requirement of food and supplies healthy and nutritious food. Crop insect detection is a challenging task for farmers as a significant portion of the crops are damaged, and the quality is degraded due to the pest attack. Traditional insect identification has the drawback of requiring well-trained taxonomists to identify insects based on morphological features accurately. Experiments were conducted for classification on nine and 24 insect classes of Wang and Xie dataset using the shape features and applying machine learning techniques such as artificial neural networks (ANN), support vector machine (SVM), k-nearest neighbours (KNN), naive bayes (NB) and convolutional neural network (CNN) model. This paper presents the insect pest detection algorithm that consists of foreground extraction and contour identification to detect the insects for Wang, Xie, Deng, and IP102 datasets in a highly complex background. The 9-fold cross-validation was applied to improve the performance of the classification models.

6. A comparative study of fine-tuning deep learning models for plant disease identification

AUTHORS: E. C. Too, L. Yujian, S. Njuki, and L. Yingchun

Deep learning has recently attracted a lot of attention with the aim to develop a quick, automatic and accurate system for image identification and classification. In this work, the focus was on fine-tuning and evaluation of state-of-the-art deep convolutional neural network for image-based plant disease classification. An empirical comparison of the deep learning architecture is done. The architectures evaluated include VGG 16, Inception V4, ResNet with 50, 101 and 152 layers and DenseNets with 121 layers. The data used for the experiment is 38 different classes including diseased and healthy images of leaves of 14 plants from plantVillage. Fast and accurate models for plant disease identification are desired so that accurate measures can be applied early. Thus, alleviating the problem of food security. In our experiment, DenseNets has tendency's to consistently improve in accuracy with growing number of epochs, with no signs of overfitting and performance deterioration. Moreover, DenseNets requires a considerably less number of parameters and reasonable computing time to achieve state-of-the-art performances.

7. Crop pest classification based on deep convolutional neural network and transfer learning

AUTHORS: K. Thenmozhi, U. Srinivasulu Reddy

The growth of most important field crops such as rice, wheat, maize, soybean, and sugarcane are affected due to attack of various pests and the crop production is reduced due to different types of insects. The classification and identification of all types of crop insects correctly is a difficult task for the farmers due to the similar appearance in the earlier stage of crop growth. To address this issue, Convolutional neural network (CNN) with deep architectures is being applied as it performs automatic feature extraction and learns complex high-level features in image classification applications. This study proposed an efficient deep CNN model to classify insect species on three publicly available insect datasets. The National Bureau of Agricultural Insect Resources (NBAIR) dataset used as first insect dataset that consists of 40 classes of field crop insect images, while the second and third dataset (Xie1, Xie2) contains 24 and 40 classes of insects respectively. The proposed model was evaluated and compared with pre-trained deep learning architectures such as AlexNet, ResNet, GoogleNet and VGGNet for insect classification. Transfer learning was applied to fine-tune the pre-trained models. The data augmentation techniques such as reflection, scaling, rotation, and translation are also applied to prevent the network from overfitting. The effectiveness of hyper parameters was analysed in the proposed model to improve accuracy. The results demonstrated that the proposed CNN model is effective in classifying various types of insects in field crops than pre-trained models and can be implemented in the agriculture sector for crop protection.

8. A Deep Learning Model for Recognition of Pest Insect in Maize Plantations

AUTHORS: Witenberg S. R. Souza, Adao Nunes Alves, Dibio Borges

This work approaches recognition of insect pests in maize plantations. It presents a novel dataset of field-based images for primary and secondary insect pests, with original and augmented images to be used for supervised classification. It also proposes a modification on a residual deep learning model (Inception-V3), called Inception-V3* here, which provides faster learning and better accuracy than the original model. Tests are run for two experiments, primary pests and all (primary and secondary). Pre-trained weights from

ImageNet are used via transfer learning and AlexNet and residual models (Inception-V3 and modified Inception-V3*) are evaluated.

9. Application of Deep Learning in Integrated Pest Management: A Real-Time System for Detection and Diagnosis of Oilseed Rape Pests

AUTHORS: Yong He, Hong Zeng, Yangyang Fan, Shuaisheng Ji, Jianjian Wu

We proposed an approach to detect oilseed rape pests based on deep learning, which improves the mean average precision (mAP) to 77.14%; the result increased by 9.7% with the original model. We adopt this model to mobile platform to let every farmer able to use this program, which will diagnose pests in real time and provide suggestions on pest controlling. We designed an oilseed rape pest imaging database with 12 typical oilseed rape pests and compared the performance of five models, SSD w/Inception is chosen as the optimal model. Moreover, for the purpose of the high mAP, we have used data augmentation (DA) and added a dropout layer. The experiments are performed on the Android application we developed, and the result shows that our approach surpasses the original model obviously and is helpful for integrated pest management.

10. Rice Pest and Disease Detection Using Convolutional Neural Network

AUTHORS: Eusebio L. Mique, Thelma D. Palaoag

Detection of rice pest and diseases, and proper management and control of pest infested rice fields may result to a higher rice crop production. This study proposed an application that will help farmers in detecting rice insect pests and diseases using Convolutional Neural Network (CNN) and image processing. It looked into the different pests that attack rice fields; information on how they can be controlled and managed was considered; farmers' knowledge in different rice pests and diseases, and how they control these pests was regarded in this study; the study also looked into the reporting mechanism of farmers to government agencies. Using CNN and image processing, the application that detects rice pests and diseases was

developed. The searching and comparison of captured images to a stack of rice pest images was implemented using a model based on CNN.

2.2 CONCLUSIONS ON SURVEY

These readings provide basic background information about various techniques and algorithms in deep learning in various stages such as DCNN-G model and fusion of Google data analysis and in another stage feature layer of Yolo V3 model used by using image pyramid to achieve multi-scale feature detection in every stage these various models used to test the diseases in plants and to detect ,every model was used to increase the accuracy in predicting diseases .So here also deep learning is used as it have the advantages of clear structure and high accuracy in image recognition. So deep learning used for detecting diseases. Every reference gave an example to how to use deep learning with different algorithms giving us a basic idea about which the best to use.

3. SOFTWARE AND HARDWARE REQUIREMENTS

3.1 SOFTWARE REQUIREMENTS

- **Operating System :** Windows 7
- **Coding Language :** Python 3.7

3.2 HARDWARE REQUIREMENTS

- **Processor :** i3, RAM-4 GB
- **Storage Type :** 128 GB SSD

4. SOFTWARE DEVELOPMENT ANALYASIS

4.1 OVERVIEW OF PROBLEM

Image classification has been around for a long time and is a common research subject. In reality, it is used in the majority of major applications. In this project, we use Convolutional Neural Networks to solve the problem of plant disease identification by analysing the leaf of a plant (CNN).Plants, deep down in the food chain, are a big factor that ensures the nature of life on Earth. Since they are exposed to various natural environments, many of these plants are susceptible to various diseases. These diseases have resulted in significant agricultural losses. Detecting and treating these diseases at an early stage saves a lot of money and time. We devised a system that uses deep learning to evaluate, identify, and classify any disease that might have affected a plant by taking an image of the leaf as a solution to this issue.

4.2 DEFINE THE PROBLEM

In recent years, this problem is on the rise and seriously threatens the development of planting industry. Timely diagnosis and prevention of crop diseases has become particularly important. At present, agricultural workers often use books and network, contact local experts and use other methods to protect and manage crop diseases. But for various reasons, misjudgements and other problems often occur, resulting in agricultural production is deeply affected. At present, the research on crop diseases is mainly divided into two directions. The first one is the traditional physical method, which is mainly based on spectral detection to identify different diseases. Different types of diseases and insect pests cause different leaf damage, which leads to different spectral absorption and reflection of leaves eroded by diseases and healthy crops. The other one is to use computer vision technology to identify images. That is to say, the characteristics of disease images are extracted by using computer related technology, and the recognition is carried out through the different characteristics of diseased plants and healthy plants.

4.3 MODULES OVERVIEW

This project develops a framework based on the Wechat applet to help farmers recognize and diagnose pests and diseases more easily and rapidly. The software will detect disease on the leaves of diseased crops, making it easier for farmers to consider disease and insect pest situations and seek expert advice. We implemented the project by dividing it into different modules. The system uploads the image first, and then sends the data to the backend for processing via the network frontend. The purpose of image preprocessing

is to improve the quality of the incoming image. First and foremost, the image is zoomed to suit the model input requirements; a huge image will have a negative impact on recognition efficiency. Second, the image is cut randomly and the pixels are adjusted in order to improve recognition efficiency. After the recognition is complete, the name and status of the crop with the highest matching degree will be presented.

4.4 DEFINE THE MODULES

This project mainly consists of six modules. They are:

1. Upload crop disease Dataset – This is the first module of our project where we need to upload a folder that contains different crops belonging to different diseases.

2. Image Processing & Normalization – The uploaded dataset will undergo the process of image processing and normalization where each every image is processed and normalized.

3. Build Crop Diseases Recognitions Model – This module builds the Convolutional Neural Network algorithm which is used to identify crop diseases.

4. Upload Test Image & Predict Disease – This module takes a leaf as an input and identifies if the leaf is in a healthy state or not.

5. Accuracy & Loss Graph – This module represents the accuracy and loss graph against the number of iterations that the algorithm has undergone.

6. Exit – This will let us come out of the project implementation and closes the project.

4.5 MODULE FUNCTIONALITY

1. Upload Crop Disease Dataset – This is the first module of our project. First, we need to click on Upload Crop Disease Dataset button. Then, a window will open and we need to upload the crop diseases dataset. Go to the folder which contains the crop diseases and select that folder and then click on “Select Folder”. Now, the crop diseases dataset will be uploaded. Upon successfully uploading the dataset folder, we can see the message “dataset uploaded”.

2. Image Processing & Normalization – This is the second module of our project. After the message is displayed as mentioned in above module, click on Image Processing & Normalization button. The uploaded dataset will undergo the process of image processing and normalization where each every image is

processed and normalized. It will read all images and then process images to normalize by converting each image pixel value between 0 and 1 and for that normalization we will divide image pixels with 255 and then get value as 0 or 1 as all images pixel value will be between 0 to 255. Then a random image of a leaf will be opened to represent the successful execution of this process. After closing the image, the message "image processing completed" will be displayed.

3. Build Crop Diseases Recognition Model – This is the third module of our project. After the message is displayed as mentioned above, now click on Build Crop Diseases Recognition Model button. This module builds the Convolutional Neural Network algorithm. Convolutional neural network is used to automatically identify crop diseases. The cross-layer direct edge and multi-layer convolution in the residual network unit to the model. After the combined convolution operation is completed, it is activated by the connection into the ReLu function. After building the model, a message representing the accuracy will be displayed.

4. Upload Test Image & Predict Disease – This is the fourth module of our project. After the message is displayed as mentioned above, now click on Upload Test Image & Predict Disease. Now a window will open for us to upload a leaf image to predict the disease. Go to folder location where the leaf image whose disease must be predicted, is present. Select that image and click on open. Now, the image will be processed and then compare it with the uploaded dataset. After successful completion of this process, a window will be opened which contains the image of the leaf we uploaded and also shows the name of the disease along with the crop name.

5. Accuracy & Loss Graph – This is the fifth module of our project. After closing the above leaf window, now click on Accuracy & Loss Graph button. A window will be opened now which shows the accuracy and loss graph against the iterations. X-axis represents epoch/iterations and Y-axis represents accuracy/loss. Green line represents accuracy and blue line represents loss. This graph represents that with each increasing iteration, accuracy is getting better and better and loss is getting decreased.

6. Exit – This is the sixth module of our project. After closing the graph window, we can click on Exit button to close the program. This will let us come out of the project implementation and closes the project.

5. PROJECT SYSTEM DESIGN

5.1 SYSTEM ARCHITECTURE

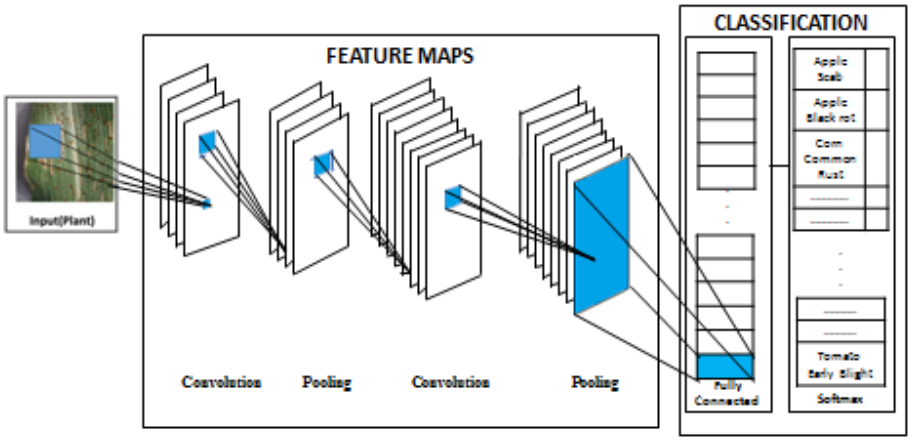


Fig 5.1: Proposed System

Fig 5.1 represents the proposed system of our project. Convolutional neural networks are made up of three components in most cases. Convolution layer, for feature extraction. The pooling layer, also known as the convergence layer, is primarily used for feature selection. By reducing the number of features, the number of parameters is decreased. The summary and output of the characteristics are carried out by the full connection layer. A convolution layer is made up of a convolution mechanism and the ReLU nonlinear activation function. Figure 1 depicts a standard CNN model architecture for pattern recognition.

The input layer is the image on the left, which the machine interprets as the input of several matrices. The convolution layer follows, with ReLU as its activation feature. The pooling layer has no activation function. Many different combinations of convolution and pooling layers can be established. When building the model, the combination of convolution layer and convolution layer, or convolution layer and pool layer, can be quite flexible. However, the most popular CNN is made up of a number of convolutional and pooling layers. Finally, a full connection layer serves as a classifier, mapping the learned feature representation to the sample label space.

5.2 UML DIAGRAMS

UML stands for Unified Modeling Language. UML is a standardized general-purpose modeling language in the field of object-oriented software engineering. The standard is managed, and was created by, the Object Management Group.

The goal is for UML to become a common language for creating models of object oriented computer software. In its current form UML is comprised of two major components: a Meta-model and a notation. In the future, some form of method or process may also be added to; or associated with, UML.

The Unified Modeling Language is a standard language for specifying, Visualization, Constructing and documenting the artifacts of software system, as well as for business modeling and other non-software systems.

The UML represents a collection of best engineering practices that have proven successful in the modeling of large and complex systems.

The UML is a very important part of developing objects oriented software and the software development process. The UML uses mostly graphical notations to express the design of software projects.

GOALS:

The Primary goals in the design of the UML are as follows:

1. Provide users a ready-to-use, expressive visual modeling Language so that they can develop and exchange meaningful models.
2. Provide extendibility and specialization mechanisms to extend the core concepts.
3. Be independent of particular programming languages and development process.
4. Provide a formal basis for understanding the modeling language.
5. Encourage the growth of OO tools market.
6. Support higher level development concepts such as collaborations, frameworks, patterns and components.
7. Integrate best practices.

USE CASE DIAGRAM

Fig 5.2 represented below depicts Use Case Diagram of our project. A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.

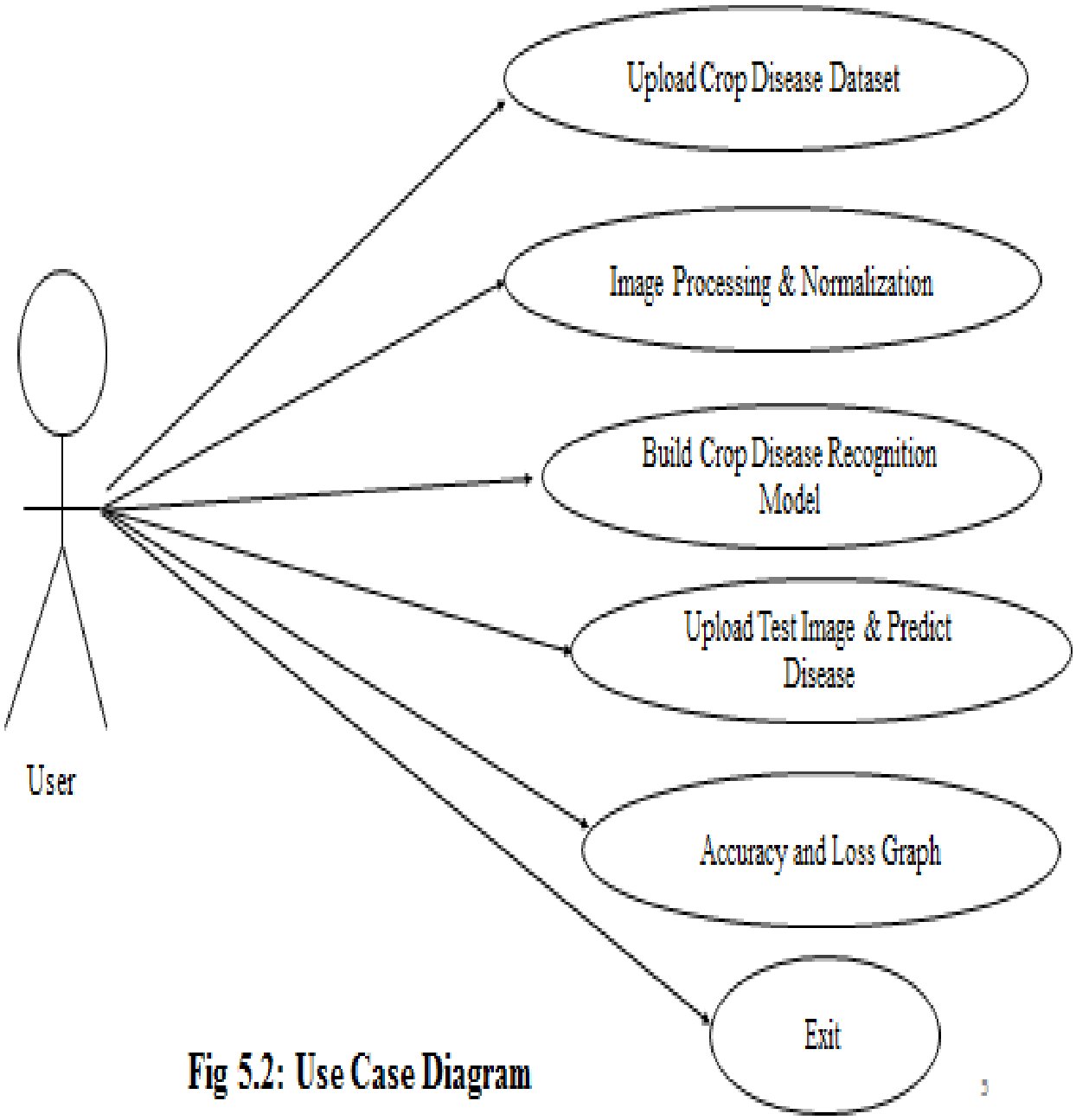


Fig 5.2: Use Case Diagram

CLASS DIAGRAM

Fig 5.3 represented below depicts Class Diagram of our project. In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.

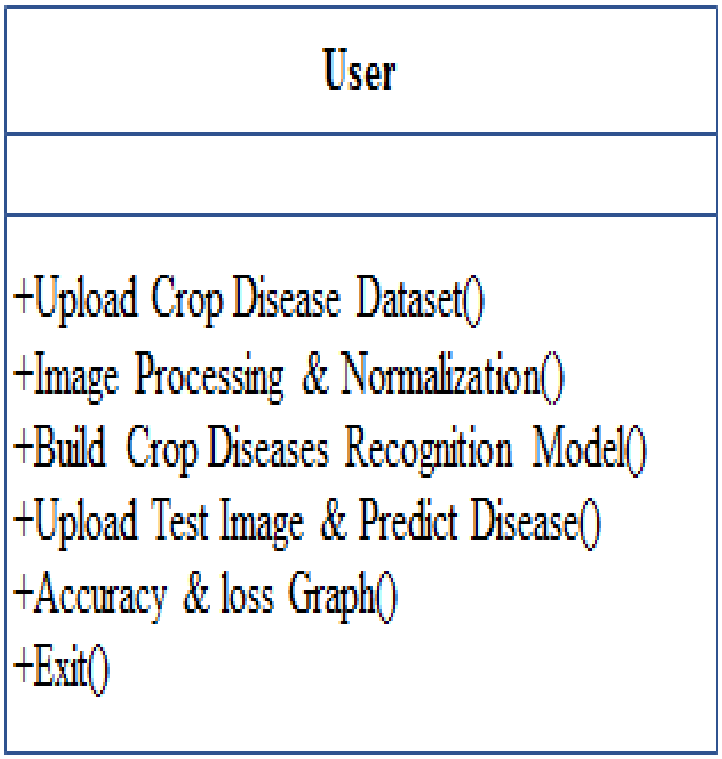


Fig 5.3: Class Diagram

SEQUENCE DIAGRAM

Fig 5.4 represented below depicts Sequence Diagram of our project. A sequence diagram in Unified Modeling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.

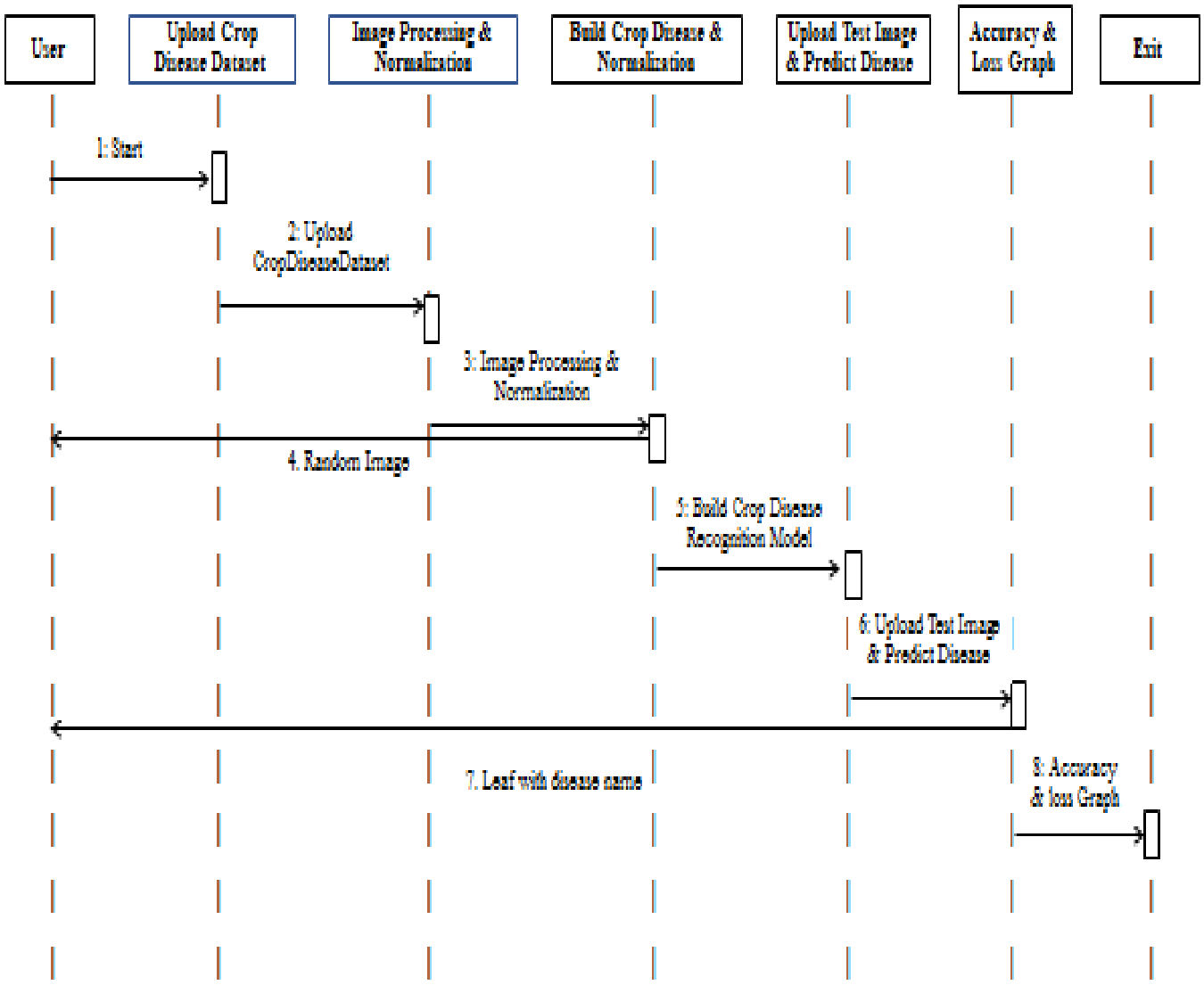


Fig 5.4: Sequence Diagram

ACTIVITY DIAGRAM

Fig 5.5 represents Activity diagram of our project. Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modeling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.

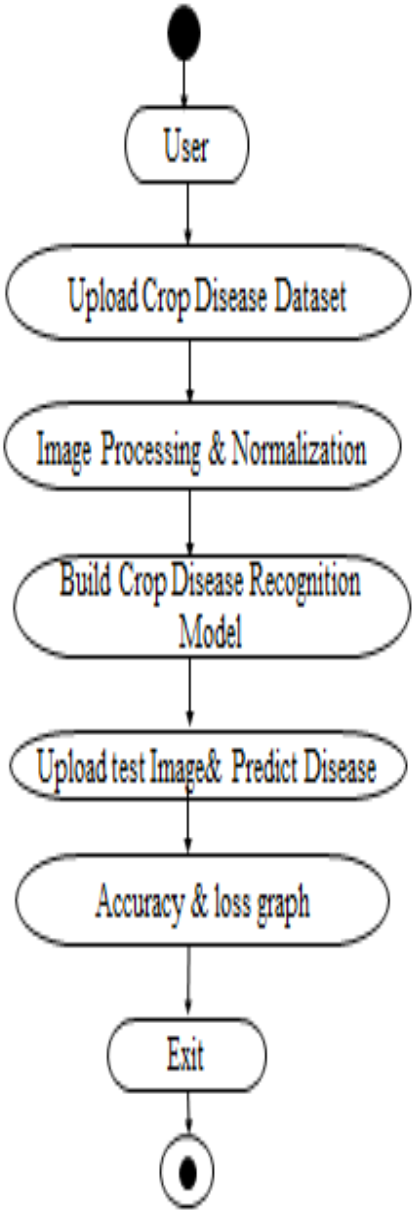


Fig 5.5: Activity Diagram

COMMUNICATION DIAGRAM

Fig 5.6 represents Communication Diagram of our project. UML Communication Diagrams, previously known as collaboration diagrams are a type of behavioural diagram that shows the interactions that take place between objects in a piece of software or system. This type of diagram emphasizes the messages exchanged between objects.

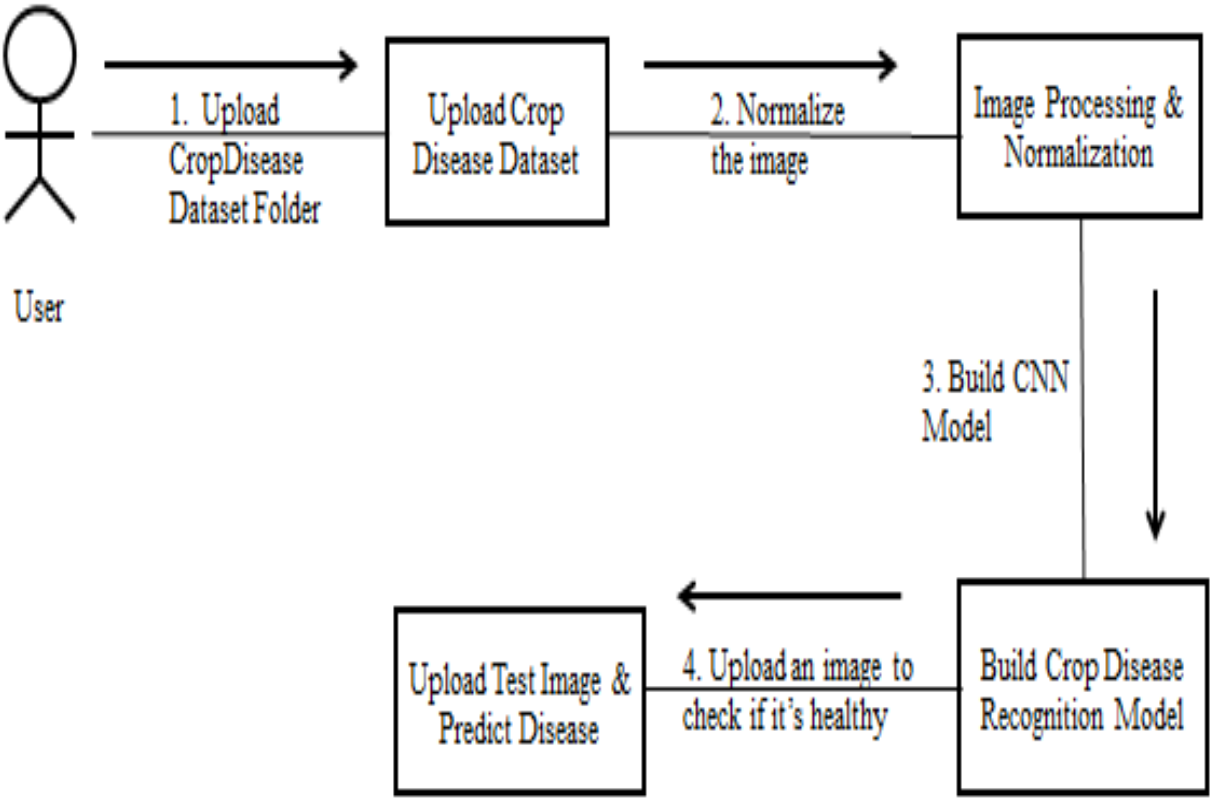


Fig 5.6: Communication Diagram

5.3 SOFTWARE ENVIRONMENT

What is Python?

Below are some facts about Python.

Python is currently the most widely used multi-purpose, high-level programming language.

Python allows programming in Object-Oriented and Procedural paradigms. Python programs generally are smaller than other programming languages like Java.

Programmers have to type relatively less and indentation requirement of the language, makes them readable all the time.

Python language is being used by almost all tech-giant companies like – Google, Amazon, Facebook, Instagram, Dropbox, Uber… etc.

The biggest strength of Python is huge collection of standard library which can be used for the following –

Machine Learning

- GUI Applications (like Kivy, Tkinter, PyQt etc.)
- Web frameworks like Django (used by YouTube, Instagram, Dropbox)
- Image processing (like Opencv, Pillow)
- Web scraping (like Scrapy, BeautifulSoup, Selenium)
- Test frameworks
- Multimedia

Advantages of Python:

Let's see how Python dominates over other languages.

1. Extensive Libraries

Python downloads with an extensive library and it contain code for various purposes like regular expressions, documentation-generation, unit-testing, web browsers, threading, databases, CGI, email, image manipulation, and more. So, we don't have to write the complete code for that manually.

2. Extensible

As we have seen earlier, Python can be **extended to other languages**. You can write some of your code in languages like C++ or C. This comes in handy, especially in projects.

3. Embeddable

Complimentary to extensibility, Python is embeddable as well. You can put your Python code in your source code of a different language, like C++. This lets us add **scripting capabilities** to our code in the other language.

4. Improved Productivity

The language's simplicity and extensive libraries render programmers **more productive** than languages like Java and C++ do. Also, the fact that you need to write less and get more things done.

5. IOT Opportunities

Since Python forms the basis of new platforms like Raspberry Pi, it finds the future bright for the Internet of Things. This is a way to connect the language with the real world.

6. Simple and Easy

When working with Java, you may have to create a class to print **'Hello World'**. But in Python, just a print statement will do. It is also quite **easy to learn, understand,** and **code.** This is why when people pick up Python, they have a hard time adjusting to other more verbose languages like Java.

7. Readable

Because it is not such a verbose language, reading Python is much like reading English. This is the reason why it is so easy to learn, understand, and code. It also does not need curly braces to define blocks, and **indentation is mandatory.** This further aids the readability of the code.

8. Object-Oriented

This language supports both the **procedural and object-oriented** programming paradigms. While functions help us with code reusability, classes and objects let us model the real world. A class allows the **encapsulation of data** and functions into one.

9. Free and Open-Source

Like we said earlier, Python is **freely available.** But not only can you **download Python** for free, but you can also download its source code, make changes to it, and even distribute it. It downloads with an extensive collection of libraries to help you with your tasks.

10. Portable

When you code your project in a language like C++, you may need to make some changes to it if you want to run it on another platform. But it isn't the same with Python. Here, you need to **code only**

once, and you can run it anywhere. This is called **Write Once Run Anywhere (WORA)**. However, you need to be careful enough not to include any system-dependent features.

11. Interpreted

Lastly, we will say that it is an interpreted language. Since statements are executed one by one, **debugging is easier** than in compiled languages.

Any doubts till now in the advantages of Python? Mention in the comment section.

Purpose:

We demonstrated that our approach enables successful segmentation of intra-retinal layers—even with low-quality images containing speckle noise, low contrast, and different intensity ranges throughout—with the assistance of the ANIS feature.

Python

Python is an interpreted high-level programming language for general-purpose programming. Created by Guido van Rossum and first released in 1991, Python has a design philosophy that emphasizes code readability, notably using significant whitespace.

Python features a dynamic type system and automatic memory management. It supports multiple programming paradigms, including object-oriented, imperative, functional and procedural, and has a large and comprehensive standard library.

- Python is Interpreted − Python is processed at runtime by the interpreter. You do not need to compile your program before executing it. This is similar to PERL and PHP.
- Python is Interactive − you can actually sit at a Python prompt and interact with the interpreter directly to write your programs.

Python also acknowledges that speed of development is important. Readable and terse code is part of this, and so is access to powerful constructs that avoid tedious repetition of code. Maintainability also ties into this may be an all but useless metric, but it does say something about how much code you have to scan, read and/or understand to troubleshoot problems or tweak behaviors. This speed of development, the ease with which a programmer of other languages can pick up basic Python skills and the huge standard library is key to another area where Python excels. All its tools have been quick to implement, saved a lot of time, and several of them have later been patched and updated by people with no Python background - without breaking.

Modules Used in Project :

Tensorflow

TensorFlow is a free and open-source software library for dataflow and differentiable programming across a range of tasks. It is a symbolic math library, and is also used for machine learning applications such as neural networks. It is used for both research and production at Google.

TensorFlow was developed by the Google Brain team for internal Google use. It was released under the Apache 2.0 open-source license on November 9, 2015.

Numpy

Numpy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays.

It is the fundamental package for scientific computing with Python. It contains various features including these important ones:

- A powerful N-dimensional array object
- Sophisticated (broadcasting) functions
- Tools for integrating C/C++ and Fortran code
- Useful linear algebra, Fourier transform, and random number capabilities

Besides its obvious scientific uses, Numpy can also be used as an efficient multi-dimensional container of generic data. Arbitrary data-types can be defined using Numpy which allows Numpy to seamlessly and speedily integrate with a wide variety of databases.

Pandas

Pandas is an open-source Python Library providing high-performance data manipulation and analysis tool using its powerful data structures. Python was majorly used for data munging and preparation. It had very little contribution towards data analysis. Pandas solved this problem. Using Pandas, we can accomplish five typical steps in the processing and analysis of data, regardless of the origin of data load, prepare, manipulate, model, and analyze. Python with Pandas is used in a wide range of fields including academic and commercial domains including finance, economics, Statistics, analytics, etc.

Matplotlib

Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms. Matplotlib can be used in Python scripts, the Python and IPython shells, the Jupyter Notebook, web application servers, and four

graphical user interface toolkits. Matplotlib tries to make easy things easy and hard things possible. You can generate plots, histograms, power spectra, bar charts, error charts, scatter plots, etc., with just a few lines of code. For examples, see the sample plots and thumbnail gallery.

For simple plotting the pyplot module provides a MATLAB-like interface, particularly when combined with IPython. For the power user, you have full control of line styles, font properties, axes properties, etc, via an object oriented interface or via a set of functions familiar to MATLAB users.

Install Python Step-by-Step in Windows and Mac:

Python a versatile programming language doesn't come pre-installed on your computer devices. Python was first released in the year 1991 and until today it is a very popular high-level programming language. Its style philosophy emphasizes code readability with its notable use of great whitespace.
The object-oriented approach and language construct provided by Python enables programmers to write both clear and logical code for projects. This software does not come pre-packaged with Windows.

How to Install Python on Windows and Mac:

There have been several updates in the Python version over the years. The question is how to install Python? It might be confusing for the beginner who is willing to start learning Python but this tutorial will solve your query. The latest or the newest version of Python is version 3.7.4 or in other words, it is Python 3.
Note: The python version 3.7.4 cannot be used on Windows XP or earlier devices.

Before you start with the installation process of Python. First, you need to know about your **System Requirements**. Based on your system type i.e. operating system and based processor, you must download the python version. My system type is a **Windows 64-bit operating system**. So the steps below are to install python version 3.7.4 on Windows 7 device or to install Python 3. Download the Python Cheatsheet here. The steps on how to install Python on Windows 10, 8 and 7 are **divided into 4 parts** to help understand better.

6. PROJECT CODING

6.1 CODING ALGORITHM

```
# Import required modules

# Declare global variables

def uploadDataset():

        # CropDiseaseDataset folder containing different crops belonging to different diseases will be
        uploaded.


def imageProcessing():

        # Uploaded images will be processed.


def cnnModel():

        # Convolutional Neural Networks algorithm is built.

        global model

           global accuracy

           text.delete('1.0', END)

           if os.path.exists('model/model.json'):

              with open('model/model.json', "r") as json_file:

                 loaded_model_json = json_file.read()

                 model = model_from_json(loaded_model_json)

              json_file.close()

              model.load_weights("model/model_weights.h5")

              model.make_predict_function()
```

```
        print(model.summary())
        f = open('model/history.pckl', 'rb')
        accuracy = pickle.load(f)
        f.close()
        text.insert('accuracy')
        acc = accuracy['accuracy']
        acc = acc[9] * 100
        text.insert(END,"CNN Crop Disease Recognition Model Prediction Accuracy = "+str(acc))

else:
        model = Sequential() #resnet transfer learning code here
        model.add(Convolution2D(32, 3, 3, input_shape = (64, 64, 3), activation = 'relu'))
        model.add(MaxPooling2D(pool_size = (2, 2)))
        model.add(Convolution2D(32, 3, 3, activation = 'relu'))
        model.add(MaxPooling2D(pool_size = (2, 2)))
        model.add(Flatten())
        model.add(Dense(output_dim = 256, activation = 'relu'))
        model.add(Dense(output_dim = 15, activation = 'softmax'))
        model.compile(optimizer = 'adam', loss = 'categorical_crossentropy', metrics = ['accuracy'])
        print(model.summary())
        hist = model.fit(X, Y, batch_size=16, epochs=10, validation_split=0.2, shuffle=True,
verbose=2)
        model.save_weights('model/model_weights.h5')
        model_json = model.to_json()
        with open("model/model.json", "w") as json_file:
```

```
            json_file.write(model_json)
        json_file.close()
        f = open('model/history.pckl', 'wb')
        pickle.dump(hist.history, f)
        f.close()
        f = open('model/history.pckl', 'rb')
        accuracy = pickle.load(f)
        f.close()
        acc = accuracy['accuracy']
        acc = acc[9] * 100
        text.insert(END,"CNN Crop Disease Recognition Model Prediction Accuracy = "+str(acc))

def predict():
    # A leaf image will be uploaded to predict the state of leaf.

def graph():
    # Graph representing accuracy and loss versus number of iterations.

def close():
    # Exit from the project.
# Define title window
# Create buttons for all the methods used
main.config()
main.mainloop()
```

6.2 OUTLINE FOR VARIOUS FILES

We used Python programming to implement our project. A single python file is used to implement our code. This file consists of various modules that we have used. Our project modules are - Upload crop disease Dataset, Image Processing & Normalization, Build Crop Diseases Recognitions Model, Upload Test Image & Predict Disease, Accuracy & Loss Graph and Exit. We also used various python modules like tkinter, matplotlib, numpy, tensorflow, keras, sklearn, os, cv2 and pickle.

6.3 METHODS INPUT AND OUTPUT PARAMETERS

In our project code, we implemented six different methods. They are:

1. uploadDataset()

2. imageProcessing()

3. cnnModel()

4. predict()

5. graph()

6. close()

Our first method uploadDataset() doesn't take any input parameters but after successful execution, it displays a message "dataset loaded". Second method imageProcessing() doesn't have any input parameters and after successful completion, it displays a message " image processing completed". cnnModel() doesn't have any input parameters. After building the CNN algorithm, the accuracy of our project is displayed. predict() doesn't have any input parameters but upon successful completion, it displays the disease name along with the crop name. graph() also don't have any input parameters but it displays a graph showing accuracy and loss versus iterations. close() don't have any parameters but upon clicking this button, it will close the project window.

7. PROJECT TESTING

7.1 VARIOUS TEST CASES

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub-assemblies, assemblies and/or a finished product. It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

TYPES OF TESTS

Unit testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

Integration testing

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

Functional test

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items:

Valid Input : identified classes of valid input must be accepted.

Invalid Input : identified classes of invalid input must be rejected.

Functions : identified functions must be exercised.

Output : identified classes of application outputs must be exercised.

Systems/Procedures : interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

System Test

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

Unit Testing

Unit testing is usually conducted as part of a combined code and unit test phase of the software lifecycle, although it is not uncommon for coding and unit testing to be conducted as two distinct phases.

Test strategy and approach

Field testing will be performed manually and functional tests will be written in detail.

Test objectives

- All field entries must work properly.
- Pages must be activated from the identified link.
- The entry screen, messages and responses must not be delayed.

Features to be tested

- Verify that the entries are of the correct format
- No duplicate entries should be allowed
- All links should take the user to the correct page.

Integration Testing

Software integration testing is the incremental integration testing of two or more integrated software components on a single platform to produce failures caused by interface defects.

The task of the integration test is to check that components or software applications, e.g. components in a software system or – one step up – software applications at the company level – interact without error.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

Acceptance Testing

User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

7.2 BLACK BOX TESTING

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box .you cannot "see" into it. The test provides inputs and responds to outputs without considering how the software works.

7.3 WHITE BOX TESTING

White Box Testing is a testing in which in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is purpose. It is used to test areas that cannot be reached from a black box level.

8. OUTPUT SCREENS

8.1 USER INTERFACES

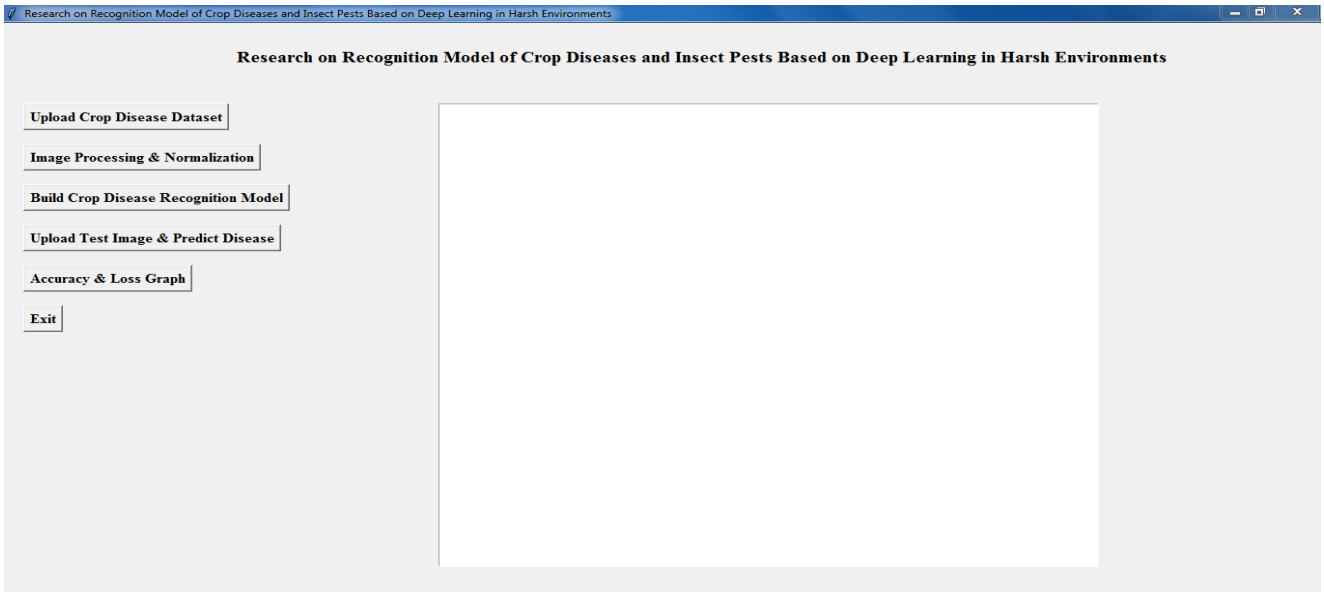


Fig 8.1: User Interface

In fig 8.1, click on 'Upload Crop Disease Dataset' button to upload dataset images.

8.2 OUTPUT SCREENS

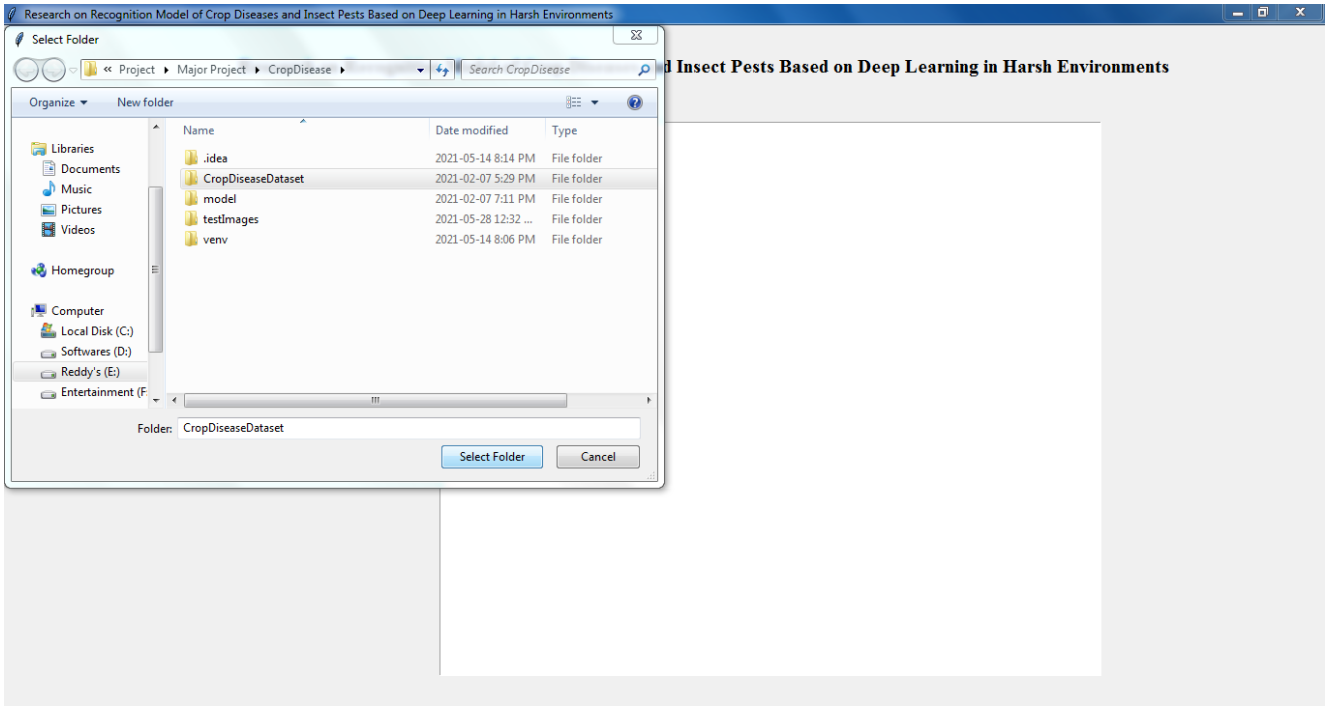


Fig 8.2: Uploading CropDiseaseDataset

In fig 8.2, selecting and uploading ‘CropDiseaseDataset’ folder and then click on ‘SelectFolder’ button to load dataset and to get below screen.

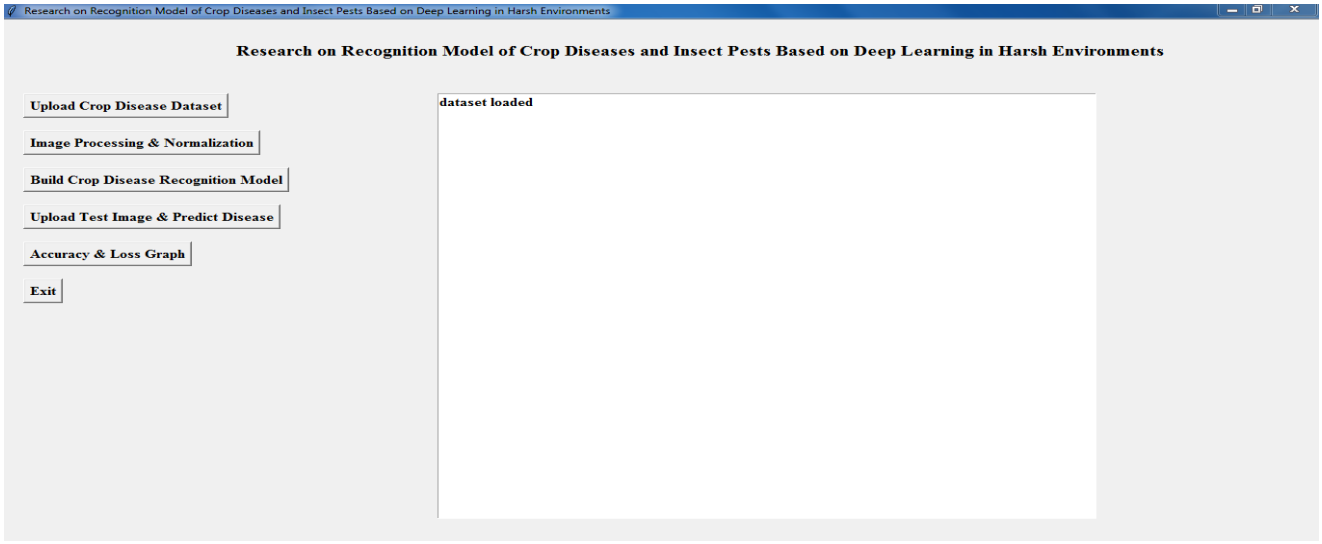


Fig 8.3: Dataset uploaded

In fig 8.3, dataset loaded and now click on ‘Image Processing & Normalization’ button to read all images and then process images to normalize by converting each image pixel value between 0 and 1 and for that normalization we will divide image pixels with 255 and then get value as 0 or 1 as all images pixel value will be between 0 to 255.

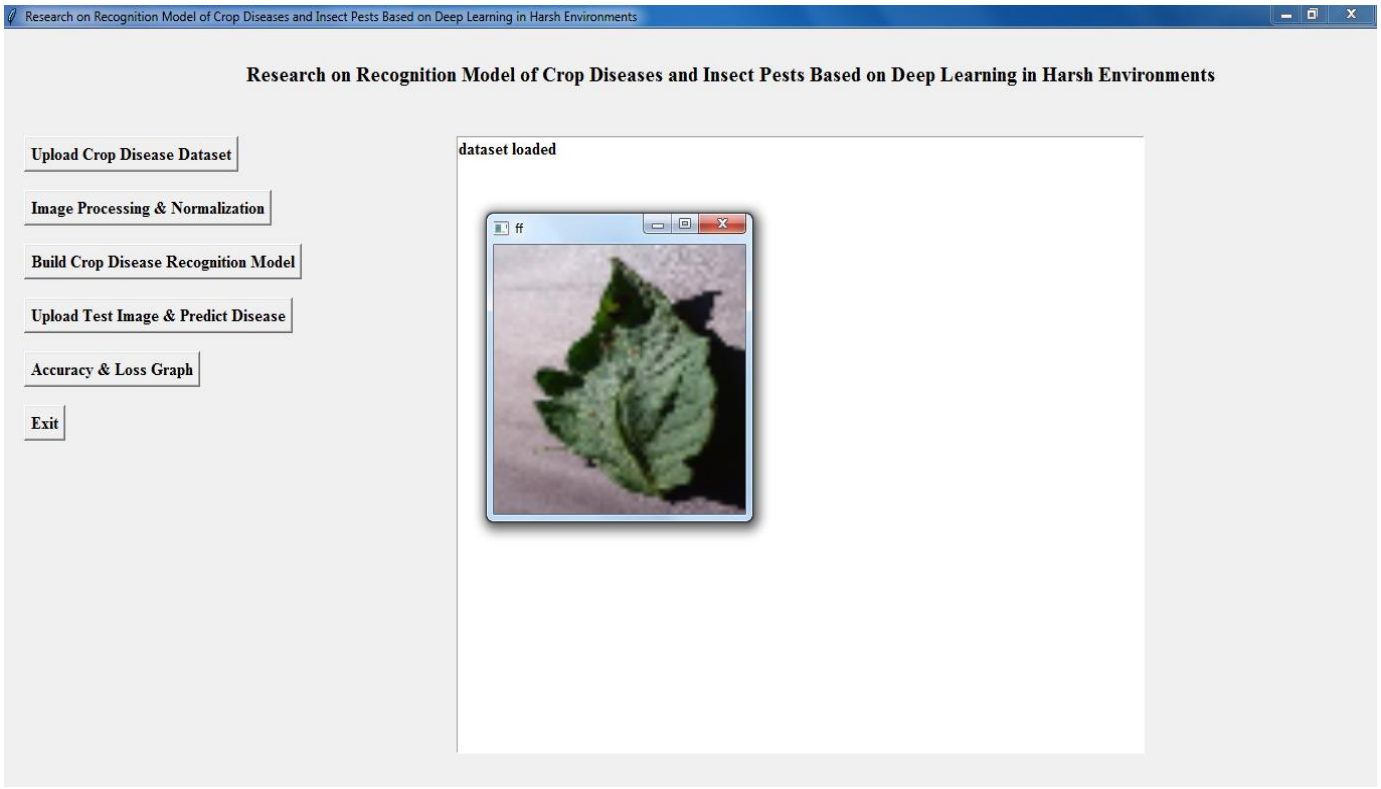


Fig 8.4: Image Processing and Normalization

In fig 8.4, after applying normalization we are just displaying one random image from dataset to check whether images loaded and process properly or not and now you close above image to get below screen.

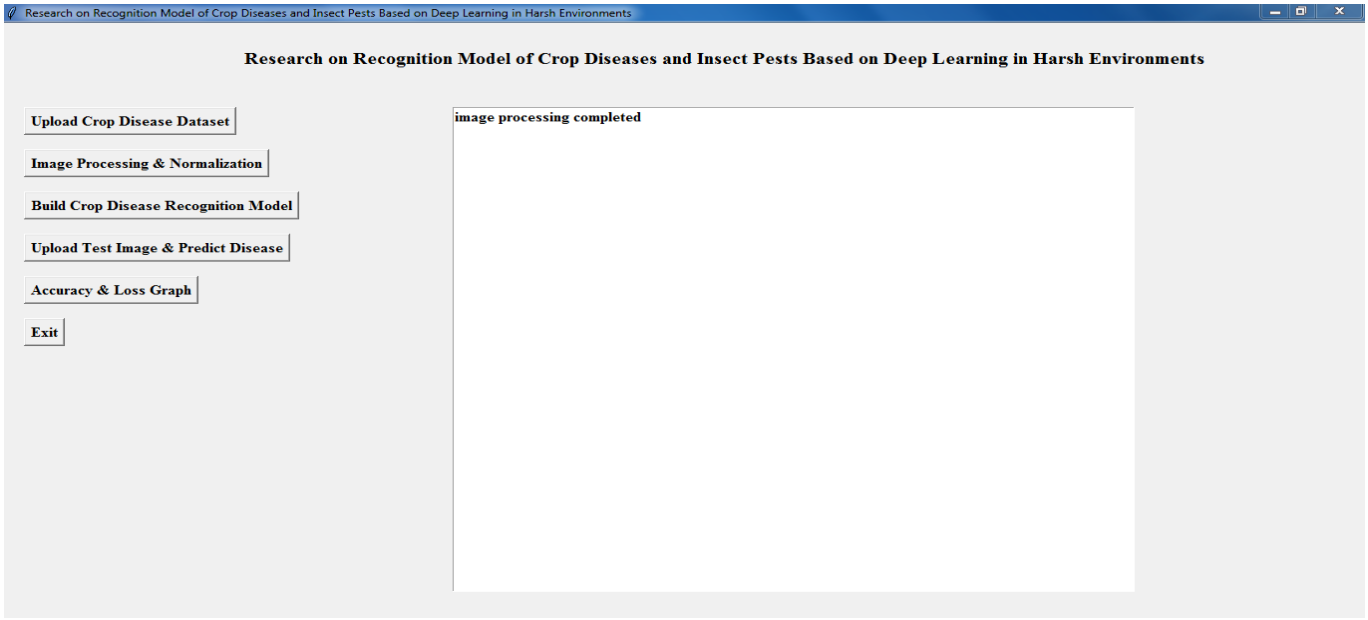


Fig 8.5: Images Processed

In fig 8.5, all images process successfully and now dataset images are ready and now click on 'Build Crop Disease Recognition Model' button to build CNN model

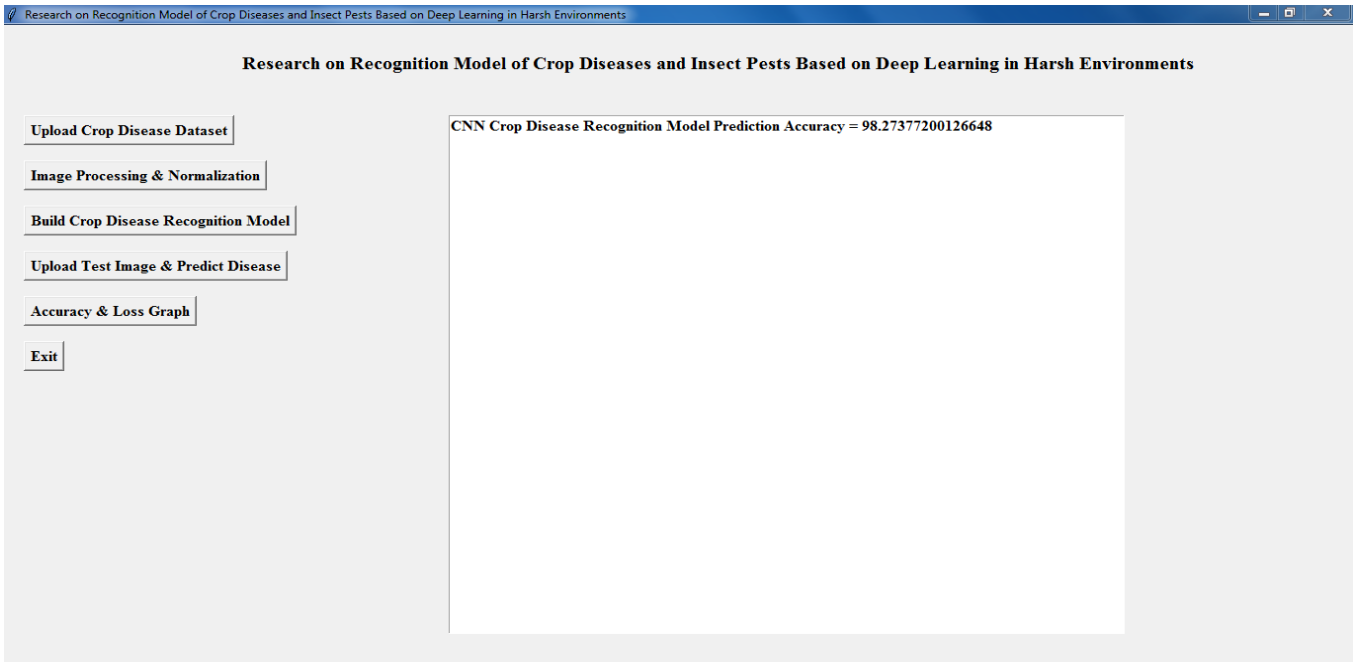


Fig 8.6: Built CNN Model

In fig 8.6, CNN model is generated and its prediction accuracy is 98%.

9. EXPERIMENTAL RESULTS

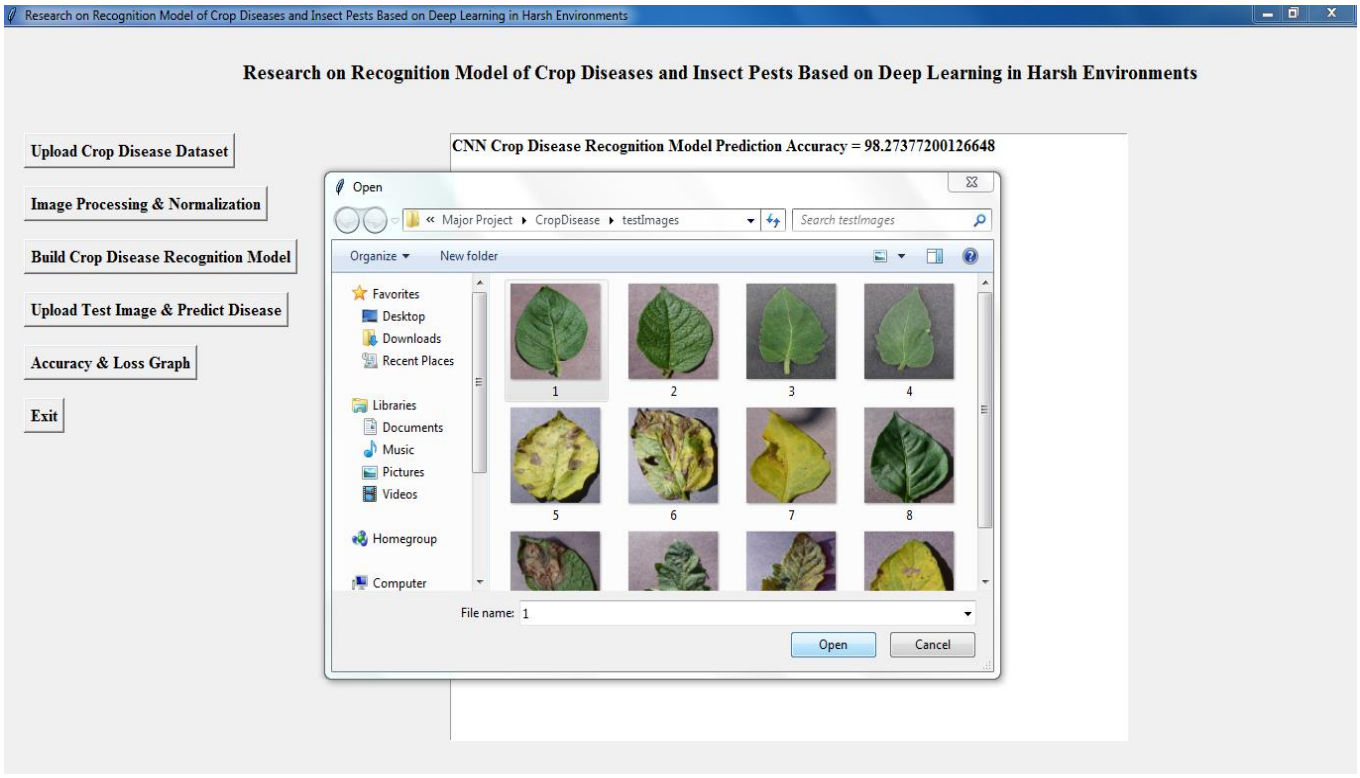


Fig 9.1: Predict Disease with Sample Imag e-1

In fig 9.1, selecting and uploading '1.JPG' image file and then click on 'Open' button to get below prediction result.

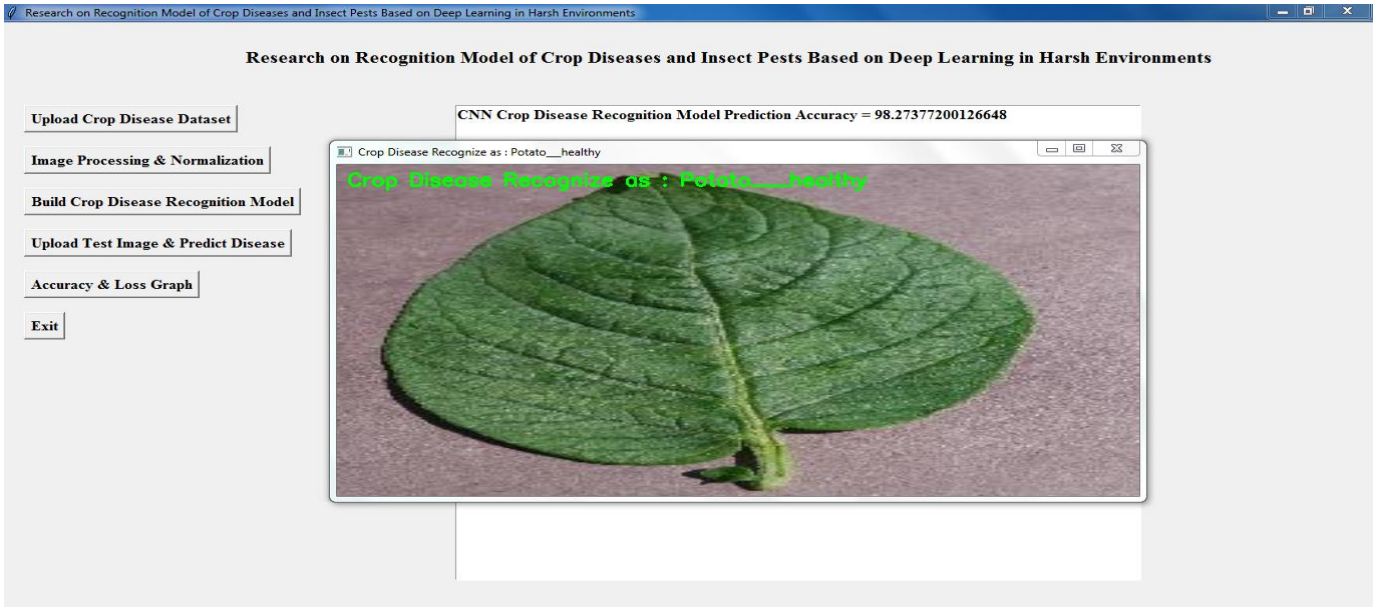


Fig 9.2: Sample Image-1 Disease Recognised

In fig 9.2, potato leaf is predicted as healthy and now we try with other image

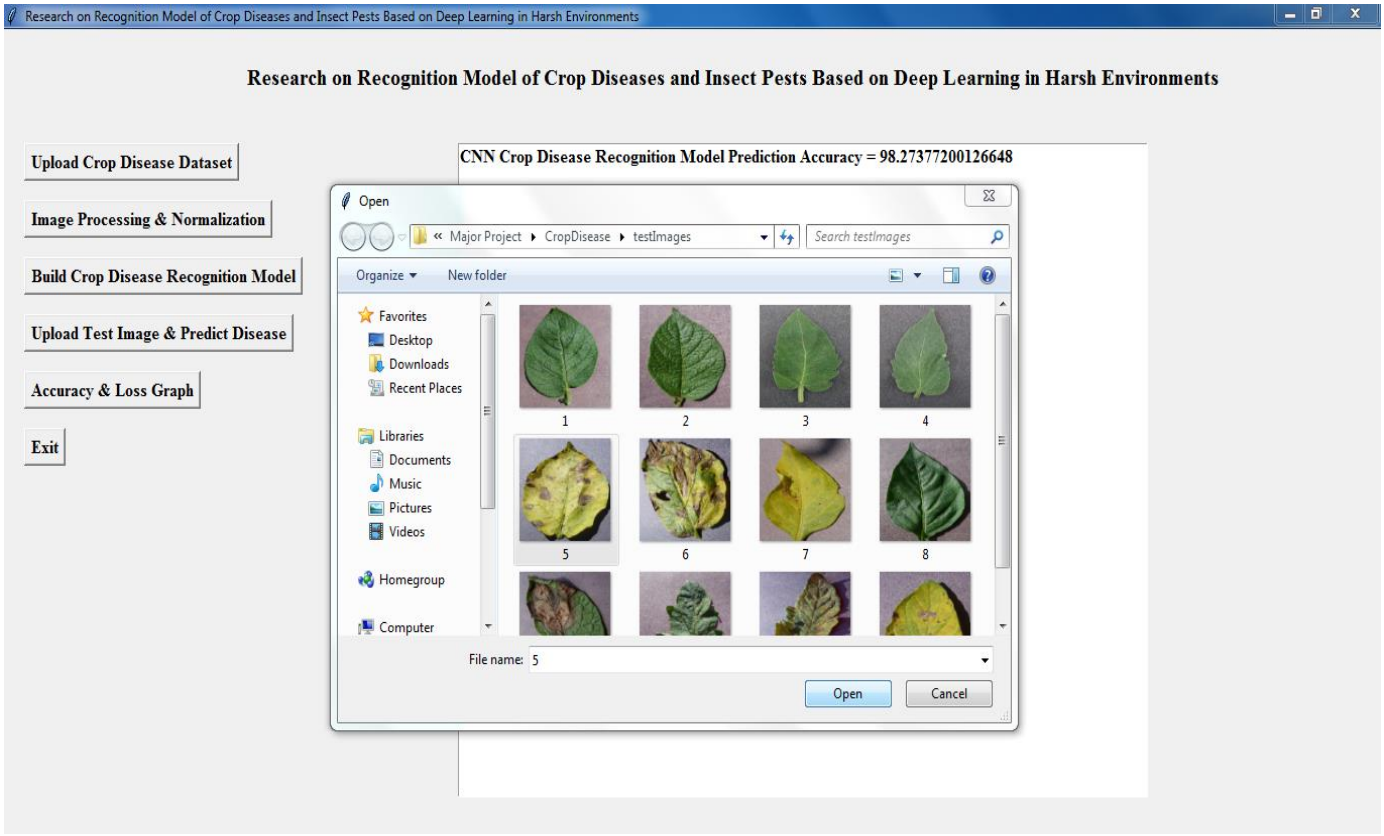


Fig 9.3: Predict Disease with Sample Image-2

In fig 9.3, selecting and uploading '5.JPG' file and click 'Open' button to get below result.

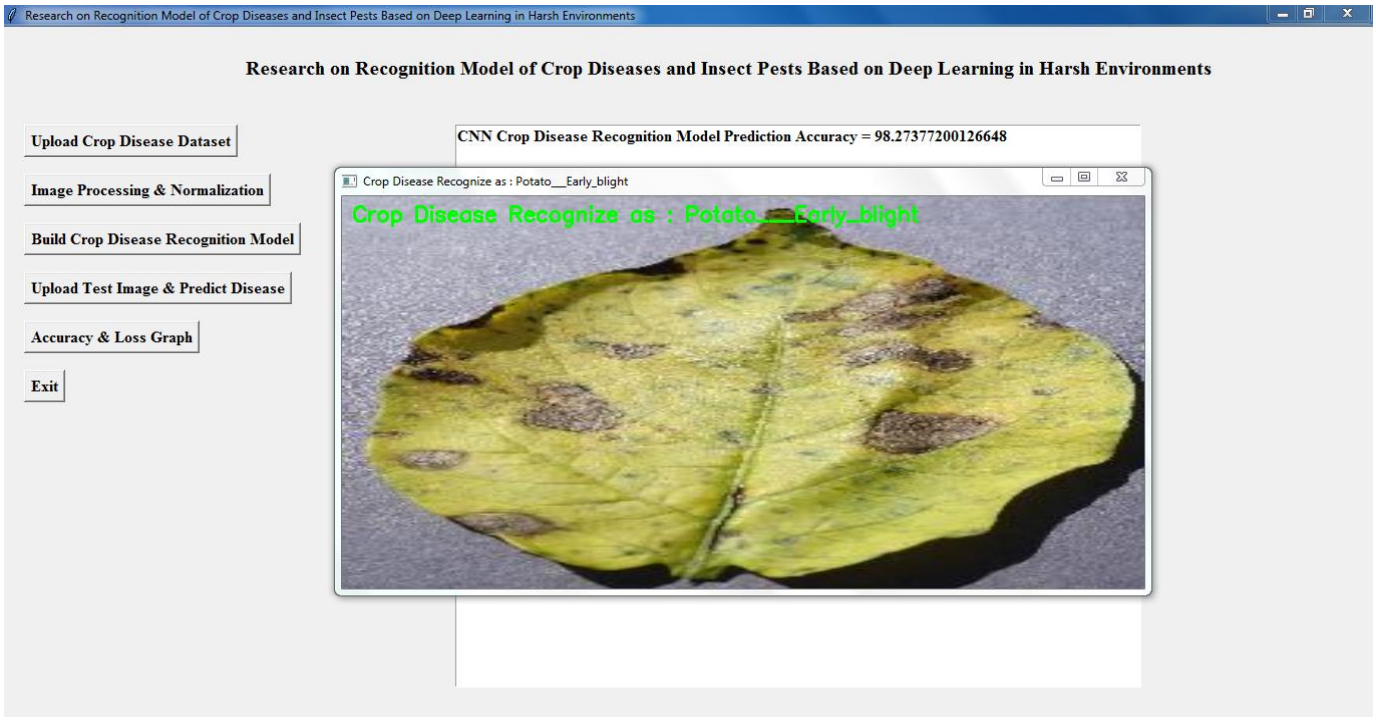


Fig 9.4: Sample Image-2 Disease RecognisedI

n fig 9.4, potato EARLY BLIGHT disease is detected or recognize and similarly you can upload any other image and get result and now click on 'Accuracy & Loss Graph' button to get below graph.

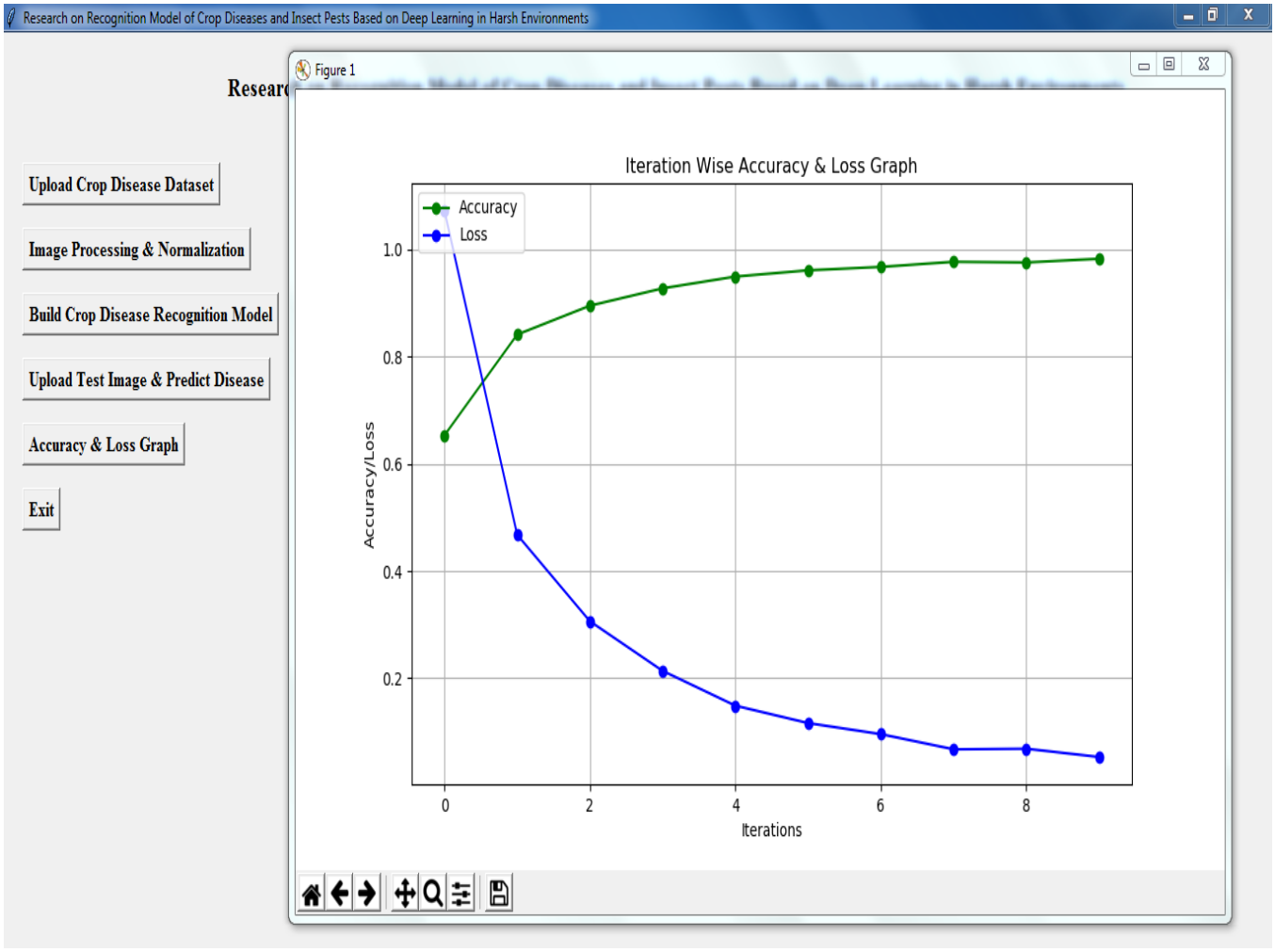


Fig 9.5: Accuracy/Loss Graph

In fig 9.5, x-axis represents epoch/iterations and y-axis represents accuracy/loss and green line represents accuracy and blue line represents loss and from above graph we can see with each increasing iteration accuracy is getting better and better and loss getting decrease.

10. CONCLUSION AND FUTURE ENHANCEMENT

In this project, 15 kinds of disease recognition of 3 kinds of crops were studied. The Inception-ResNet-v2 model is constructed by using deep learning theory and convolution neural network technology. Experiments show that the model can effectively identify the data set, and the overall recognition accuracy is as high as 86.1%. The results show that the recognition accuracy of this hybrid network model is relatively higher than the traditional model, and it can be effectively applied to the identification and detection of plant diseases and insect pests. In the future work, there are two directions should be improved:

1) Extended data set. In this project, only 15 diseases of 3 crop species were studied, and other species and diseases were not involved, such as rice and wheat, and their related diseases. Therefore, the next step is to obtain more crop species and disease images for research.

2) Optimize the model. Through the experiment of this project, we can see that Inception-resnet-v2 this kind of mixed network has absorbed the corresponding advantage. This model has achieved good recognition accuracy, and is worthy of further study and optimization. At the same time, we should design a network model which can classify crop images with higher accuracy.

REFERENCES

1. Mingyuan Xin and Yong Wang, "Image Recognition of Crop Diseases and Insect Pests Based on Deep Learning" Volume 2021 Article ID 5511676.
2. Jun Liu and, Xuewei Wang, "Tomato Diseases and Pests Detection Based on Improved Yolo V3 Convolutional Neural Network," Front. Plant Sci., 16 June 2020.
3. Dan Jeric Arcega Rustia, Jun-Jee Chao, Lin-Ya Chiu, Ya-Fang Wu, Jui-Yung Chung, Ju-Chun Hsu, Ta-Te Lin," Automatic greenhouse insect pest detection and recognition based on a cascaded deep learning classification method, "First published: 20 November 2020
4. N.Abirami , P.Kavinilavan , M.Pooja , R.Vigneshand T.Kavitha, "Protecting the Farming Land from Insects Damage to Growing Crops using Deep Convolutional Neural Network, "Published online: 28 April 2021.
5. Thenmozhi Kasinathan, Dakshayani Singaraju, Srinivasulu Reddy Uyyala, "Insect classification and detection in field crops using modern machine learning techniques," Information Processing in Agriculture, Available online 8 October 2020.
6. E. C. Too, L. Yujian, S. Njuki, and L. Yingchun, ''A comparative study of fine-tuning deep learning models for plant disease identific00at12ion,'' Comput. Electron. Agricult., vol. 161, pp. 272–279, Jun. 2019.
7. K. Thenmozhi, U. Srinivasulu Reddy, "Crop pest classification based on deep convolutional neural network and transfer learning," Computers and Electronics in Agriculture 164(4):104906, September 2019.
8. Witenberg S. R. Souza, Adao Nunes Alves, Dibio Borges "A Deep Learning Model for Recognition of Pest Insect in Maize Plantations, "October 2019 DOI:10.1109/SMC.2019.8914428 Conference: 2019
9. Yong He, Hong Zeng, Yangyang Fan, Shuaisheng Ji, Jianjian Wu, "Application of Deep Learning in Integrated Pest Management: A Real-Time System for Detection and Diagnosis of Oilseed Rape Pests" Volume 2019 |Article ID 4570808
10. Eusebio L. Mique, Thelma D. Palaoag, "Rice Pest and Disease Detection Using Convolutional Neural Network" ICISS '18: 2018 International Conference on Information Science and System Jeju Republic of Korea 27 April, 2018- 29 April, 2018

PUBLICATIONS

Research on Recognition Model of Crop Diseases and Insect Pests Based on Deep Learning in Fields

Chaganti Ravali[1], Chennam Aishwarya Reddy[2], Mada Shirisha[3],

Katukam Prem Kumar[4], Dr. M. Narayanan[5], Dr. G. JawaherlalNehru[6]

[1234]UG Scholar, [5]Professor, [6]Assistant Professor

Department of Computer Science and Engineering

St. Martin's Engineering College, Secunderabad – 500 100, India

E-Mail: [1]chinni.chaganti666@gmail.com, [2]aishwaryachennam@gmail.com, [3]madashirisha78@gmail.com, [4]katukamprem28@gmail.com, narayanan_baba@yahoo.com

Abstract

Agricultural diseases and insect pests are two of the most significant factors threatening agricultural development. Early detection and identification of pests will significantly minimize pest-related economic losses. In this paper, a convolutional neural network is used to classify crop diseases automatically. The data comes from the AI Challenger Competition's public data collection from 2018, which includes 27 disease images from ten crops. In this paper, the CNN-Inception-ResNet-v2 model is used for training. The residual network unit to the model has a cross-layer direct edge and multi-layer convolution. The connection into the ReLu feature activates it after the combined convolution process is completed. The overall identification accuracy is 98 %, according to the experimental findings. The proposed model's effectiveness is verified by the experimental findings.

Keywords: ResNet-v2, Keras, Recognition of pests and diseases, Deep Learning, Convolutional Neural Network (CNN).

I INTRODUCTION

Image classification has been around for a long time and is a common research subject. In reality, it is used in the majority of major applications. In this paper, we use Convolutional Neural Networks to solve the problem of plant disease identification by analysing the leaf of a plant (CNN).Plants, deep down

in the food chain, are a big factor that ensures the nature of life on Earth. Since they are exposed to various natural environments, many of these plants are susceptible to various diseases. As shown in Figure 1, these diseases have resulted in significant agricultural losses. Detecting and treating these diseases at an early stage saves a lot of money and time.

We devised a system that uses deep learning to evaluate, identify, and classify any disease that might have affected a plant by taking an image of the leaf as a solution to this issue. The following is the processing flow:

1. In the given picture, a leaf is detected and cropped out.

2. After that, the extracted leaf is run through a classifier to determine which plant it belongs to.

3. The leaf is then screened for disease classes, if any, based on the results of the previous step.

Artificial intelligence's rapid growth in recent years has made life easier, and AI has become a well-known technology. AlphaGo, for example, beat the world Go champion. Siri and Alexa, Apple's and Amazon's voice assistants' are all examples of artificial intelligence technology portrayed by deep learning in a variety of fields. Image recognition has advanced significantly in recent years as a primary research topic in computer vision and artificial intelligence. The aim of image recognition in agricultural applications is to recognize and classify various types of images, as well as to analyse crop types, disease types, and severity, among other things. Then, in a timely and efficient manner, we can devise appropriate countermeasures to address various problems in agricultural production so as to ensure and increase crop yields and contribute to the betterment of agriculture.

At the moment, crop disease research is primarily divided into two directions. The conventional physical approach, which is primarily focused on spectral detection to identify various diseases, is the first. Different diseases and insect pests cause different types of leaf damage, resulting in different spectral absorption and reflection of diseased and healthy crops' leaves. The other choice is to classify images using computer vision technology. To put it another way, the characteristics of disease images are collected using computer technology, and the identification is done using different characteristics of diseased and healthy plants.

This paper develops a framework based on the Wechat applet to help farmers recognize and diagnose pests and diseases more easily and rapidly. The software will detect disease on the leaves of diseased crops, making it easier for farmers to consider disease and insect pest situations and seek expert advice. The machine uploads the image first, and then sends the data to the backend for processing through the network frontend. The aim of image preprocessing is to improve the quality of the incoming image. First and foremost, the image is zoomed to meet the model input requirements; a large image would have

a negative impact on recognition quality. Second, the image is cut randomly and the pixels are optimized in order to improve recognition quality. After the recognition is complete, the name and status of the crop with the highest matching degree will be granted. If the crop is unhealthy, the appropriate instructions will be given and sent to the mobile phone.

II LITERATURE SURVEY

Crop disease and insect pest detection and control are ongoing research topics. Many sensor networks and automated control systems have been proposed as technology advances.

Deep artificial neural networks (including recurrent ones) have won various pattern recognition and machine learning competitions in recent years. This historical overview condenses pertinent research, much of it from the previous millennium. The depth of their credit assignment paths, which are chains of potentially learnable, causal connections between behaviour and consequences, distinguishes shallow and deep learners. Deep supervised learning (including a recap of backpropagation history), unsupervised learning, reinforcement learning & evolutionary computation, and indirect quest for short programs encoding deep and wide networks are all discussed [1].

A paper by Sue Han Lee makes use of Caffe framework and Back-Propagation mechanism in training the model. One of the approaches used in understanding the internal working of the CNN model and visualizing the selected filters was done using Deconvolutional Network. The Deconvolutional Network provides a function that enables us to see the feature map at each layer by deconvoluting and un pooling down to input image pixel. The models were trained on two datasets, MalayaKew Dataset (original) and addition of local leaf data to original dataset. Experimental results show that the model trained on the modified dataset yielded only a small increase in the accuracy of the model when compared to the model trained on the original dataset [2].

In another paper that proposes a plant identification system the standard VGGNet-16 architecture was taken and modified by adding more convolution layers for learning the combined species and organ features resulting the high-level fusion architecture model. The high-level fusion architecture model and Fine-tuned VGGNet-16 were trained on PlantClef2015 dataset. According to the experimental results, the High- Level fusion architecture didn't perform better compared to the fine-tuned VGGNet-16 [3].

In paper [4], an autonomous monitoring device based on a low-cost image sensor is proposed, which is capable of capturing and sending images of the trap contents to a remote control station at the frequency required by the trapping application. Our self-contained monitoring device will be able to cover vast areas

while using very little resources. This will be the most important aspect of our research, since the overall monitoring system's operating life could be extended to months of continuous service without any maintenance (i.e., battery replacement). The number of individuals detected at each trap would be calculated using the images delivered by image sensors, which would be time-stamped and processed in the control station.All of the data will be conveniently processed at the control station and accessible through the Internet via the control station's available network services (WiFi, WiMAX, 3G/4G, etc.).

Acoustic detection technology is a much more efficient alternative to these time-consuming tasks. Sugarcane takes 10-18 months to mature, making it vulnerable to a variety of insects, pests, and diseases. Its production is reduced by 19 to 20% as a result of their attacks. Control of insect pests and diseases is critical for increasing crop production. Sugarcane is infested by approximately 288 insects, of which nearly two dozen cause significant losses in both quality and quantity. We will focus on a pest control and monitoring system for efficient sugarcane crop production in this project, as sugarcane is a staple crop grown in Pune. Sugarcane is primarily affected by top shoot borer, stalk borer, rood borer, and sugarcane woolly aphid. Red rot, Smut, Grassy Shoot, and Wilt are the most common diseases that affect sugarcane crops. The system employs an acoustic device sensor that tracks pest noise levels and alerts the farmer via an alarm when the noise level exceeds a set threshold. The information is disseminated through a network of wireless sensors linked to a computer in the control room. Field data is transmitted and received using the ZigBee 802.15.4 optical communication interface standard. The machine covers a wide area while using very little energy [5].

One of the research papers has introduced an Apple Leaf Diseases Classifier that classifies leaf diseases using a combination of Alex Precursor and Cascade Inception architecture. On the basis of efficienncy, accuracy, convergence rate (with respect to epochs), and computational resources needed, it also provides a good comparison between their model and other models such as SVM, Back Propagation (BP) Neural Network, AlexNet, GoogleNet, Resnet-20, and VGGNet-16.These parameters were put to the test on a dataset of 1053 images of apple disease. The classifier that was created had a 97.62 percent accuracy rate [6].

III CONVOLUTIONAL NEURAL NETWORKS

Crop diseases and insect pests are increasingly being identified using deep neural networks. The structure of a deep neural network is modelled after that of a biological neural network, which is an artificial neural network that mimics the brain and uses learnable parameters to replace the connections between neurons. A branch of the feed forward neural network, the convolutional neural network is one of the most commonly utilized deep neural network structures. The deeper AlexNet network, which first appeared in 2012, marked the beginning of the modern convolutional neural network. The performance of the AlexNet

network model demonstrates the value of convolutional neural networks. Convolutional neural networks have grown in popularity since then, with applications in financial supervision, text and speech recognition, smart homes, medical diagnosis, and a variety of other fields.

Convolutional neural networks are made up of three components in most cases. Convolution layer, for feature extraction. The pooling layer, also known as the convergence layer, is primarily used for feature selection. By reducing the number of features, the number of parameters is decreased. The summary and output of the characteristics are carried out by the full connection layer. A convolution layer is made up of a convolution mechanism and the ReLU nonlinear activation function. Figure 1 depicts a standard CNN model architecture for pattern recognition.

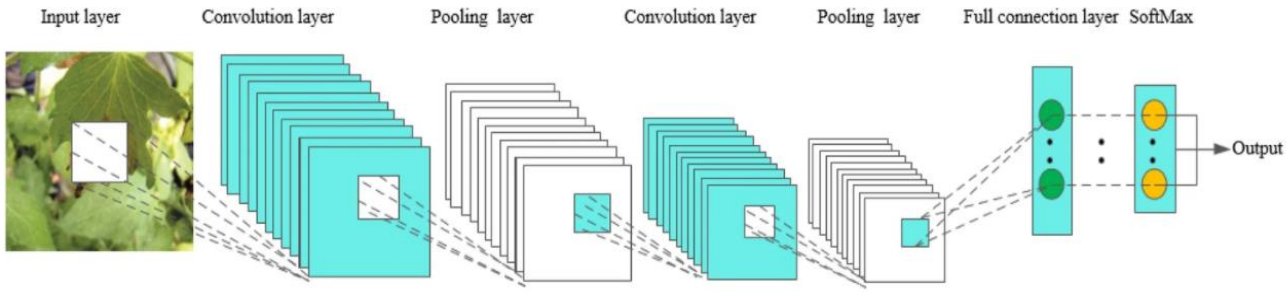


Figure 1. A typical Convolution Neural Network architecture

The input layer is the image on the left, which the machine interprets as the input of several matrices. The convolution layer follows, with ReLU as its activation feature. The pooling layer has no activation function. Many different combinations of convolution and pooling layers can be established. When building the model, the combination of convolution layer and convolution layer, or convolution layer and pool layer, can be quite flexible. However, the most popular CNN is made up of a number of convolutional and pooling layers. Finally, a full connection layer serves as a classifier, mapping the learned feature representation to the sample label space.

Crop Disease Recognition Model

In this paper, a model for identifying crop diseases and insect pests in a complex Internet of Things environment is created. The environmental information and image information of the scene are collected through the deployment of sensors and cameras in a complex mountainous area, and a basic database of crop pest identification is created. The image knowledge is learned and recognized using the deep learning network model, which is then used to classify and collect leaf images, as well as identify pests and diseases

The Structure of Crop Disease Recognition Model

The basic model of crop disease identification used in this paper is the Inception-ResNet-v2 network. This hy brid network not only has the depth advantage of residual network, but also retains the unique

characteristics of multi-convolution core of inception network. Although there is no substantial increase in accuracy after adding the residual unit to the inception network, it effectively solves the problems of gradient disappearance and gradient explosion. In addition, the model's convergence speed is accelerated. Also, the training efficiency and the small-range promotion performance are improved. Figure 2 depicts the model's structure.

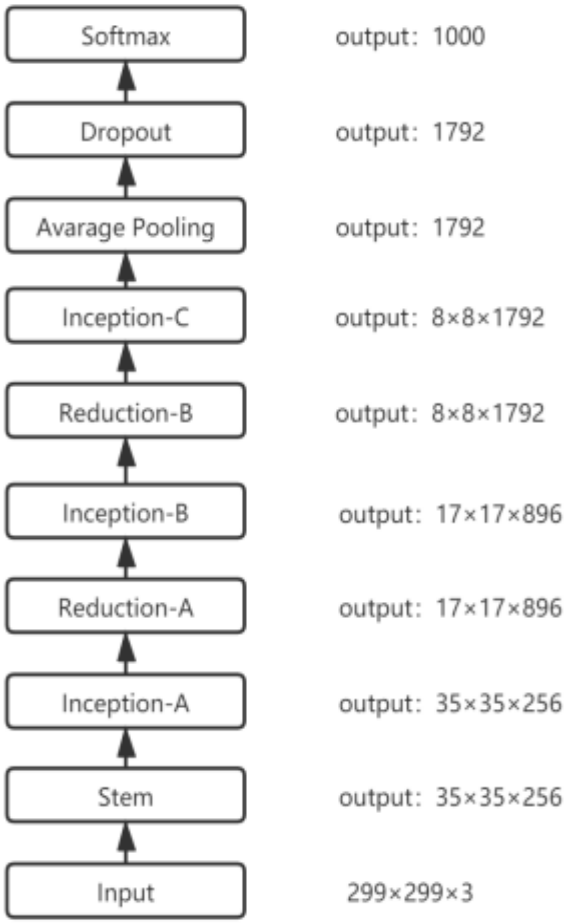


Figure 2: The structure of Inception-ResNet-v2.

Dataset

The data set utilized in this paper comes from the 2018 Artificial Intelligence Challenger Competition's Crop Disease Recognition Competition. The dataset contains 47363 photos of 27 illnesses affecting ten different crops (mainly tomatoes, potatoes, corn, etc.). The data set is separated into three sections: 70% for the training set, 10% for the validation set, and 20% for the test set. Only the leaves of a single crop are shown in each image. Figure 3 shows some examples of photographs.

Figure 3: Sample pictures in data set

Image Preprocessing

The goal of picture preprocessing is to reduce the amount of meaningless data in a data set that interferes with model recognition and to expand the data set to a certain extent. A neural network can provide a more effective training impact. In this approach, the image's recognisability may be successfully increased, and the model's recognition accuracy can be increased. Geometric space transformation and pixel colour transformation are currently the most often utilized preprocessing approaches. The former offers functions such as flip, crop, rotate, and zoom. Changing contrast, adding Gaussian noise, colour dithering, and other effects are among the latter. Because of the uneven distribution of data sets, we mostly apply the light transformation and random clipping methods in this paper. Enhance the picture's feature information as well as the data set's scale. The model's influence is lessened by the background factor and the data quantity problem. It has the potential to improve the model's learning effect as well as its stability.

Normalized Processing

After the preceding stages are completed, the data set's picture will be normalized. Normalization is a vital aspect of the convolutional neural network that cannot be overlooked. It scales each dimension's properties to the same range. On the one hand, it is simple to calculate data and increase operational efficiency. The link between diverse traits, on the other hand, is removed.

IV EXPERIMENTAL AND RESULTS

This experiment uses Windows as its operating system. Python is the programming language, and TensorFlow is the deep learning framework. Tab. 1 shows the specific equipment arrangement.

TABLE 1.Experimental environment

Configure	Param
CPU	Intel(R)Core (TM) i7-6200u
Anaconda	Anaconda 3.6
TensorFlow	1.2.1
Operating System	Windows 10
Hard disk	512GSSD
RAM	8G

Training Strategy

For migration, we employ the Inception-ResNet-v2 model in this paper. The network weight parameters learned from a large number of data sets are then transferred to their own network and fine-tuned. The steps in the procedure are as follows.

1) First, the pre-training model is loaded. The convolution and pooling layer parameters in the original model are kept as initial parameters, and the last fully connected layer is frozen. Create a new full connection layer to solve the target task's categorization challenge.

2) Configure the parameters. Set the learning rate and batch size to 0.001 and 32, respectively. The Dropout is set to 0.5 and the workout count is set at 5 epoch.

3) The loss layer's loss function is a cross-entropy loss function. The Adam optimization algorithm is used by the optimizer to update the weights and biases.

4) The image from the preprocessed train set and the image from the preprocessed verification set are then sent in a batch size image for training.

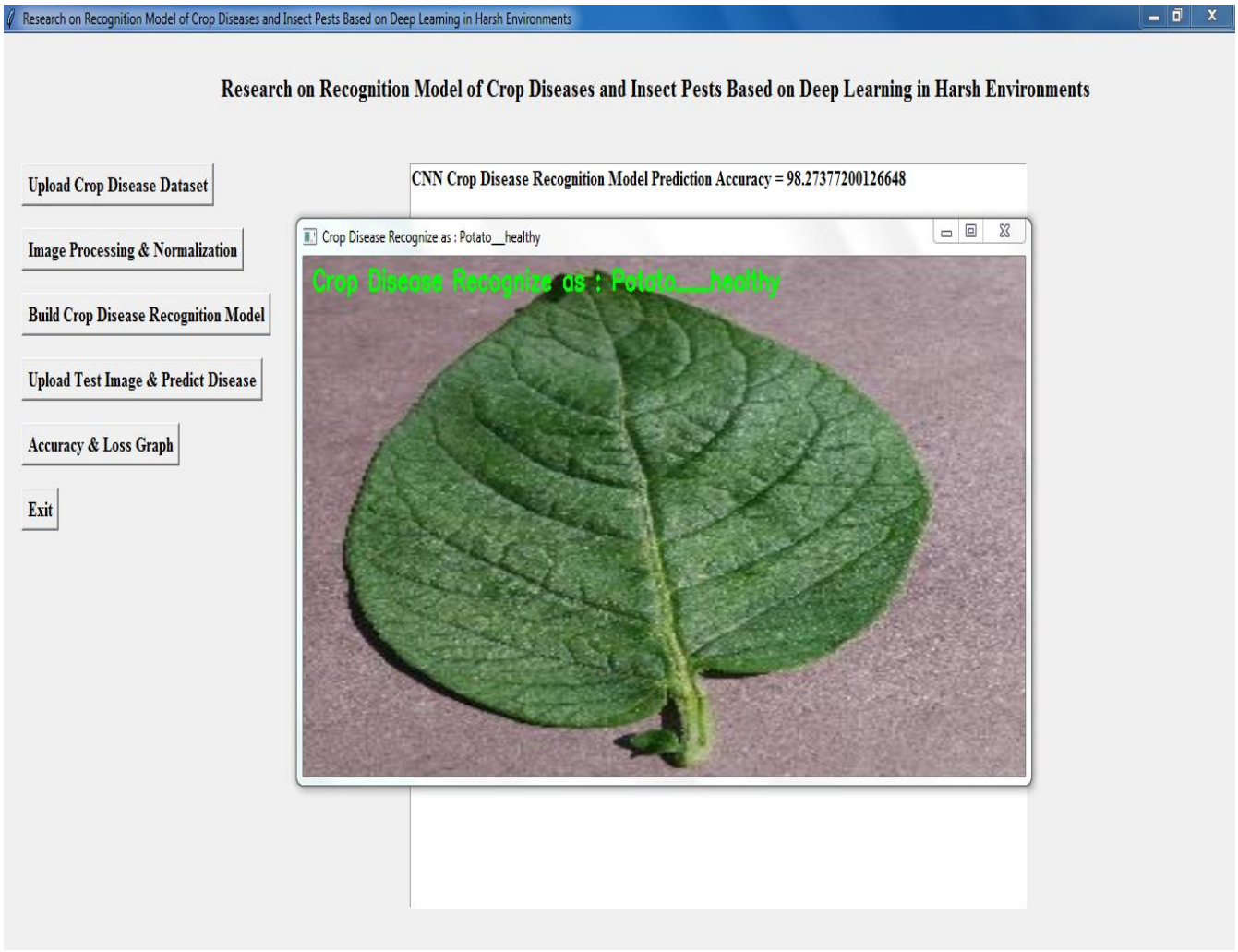
5) On the test set, the recognition and classification are finished after the model training. A list of the performance measures used to examine the dataset.

Implementation

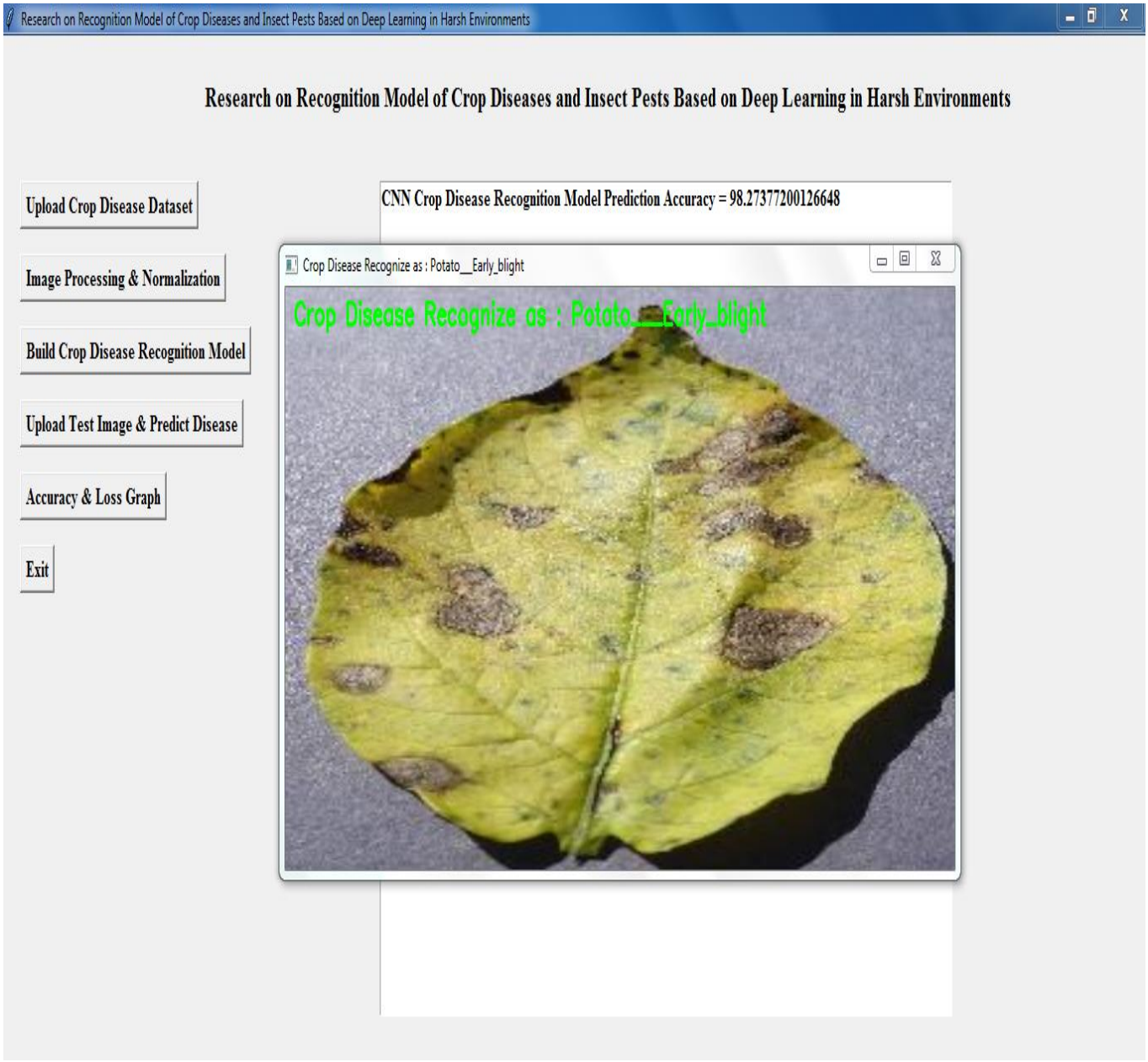
This paper develops a framework based on the Wechat applet to help farmers recognize and diagnose pests and diseases more easily and rapidly. The software will detect disease on the leaves of diseased crops, making it easier for farmers to consider disease and insect pest situations and seek expert advice. The system uploads the image first, then sends the data to the backend for processing via the network frontend. The purpose of image preprocessing is to improve the quality of the incoming image. First and foremost, the image is zoomed to suit the model input requirements; a huge image will have a negative impact on recognition efficiency. Second, the image is cut randomly and the pixels are adjusted in

order to improve recognition efficiency. After the recognition is complete, the name and status of the crop with the highest matching degree will be presented. If the crop is unhealthy, the appropriate instructions will be given and sent to the cell phone.

The system's crop disease recognition result is displayed in Screenshot 1 and the identification result is potato_healthy and Screenshot 2 shows crop disease as potato_early_blight a common disease of potato plant. After verification, the recognition is accurate.



Screenshot 1. Crop Disease Recognition as Potato healthy



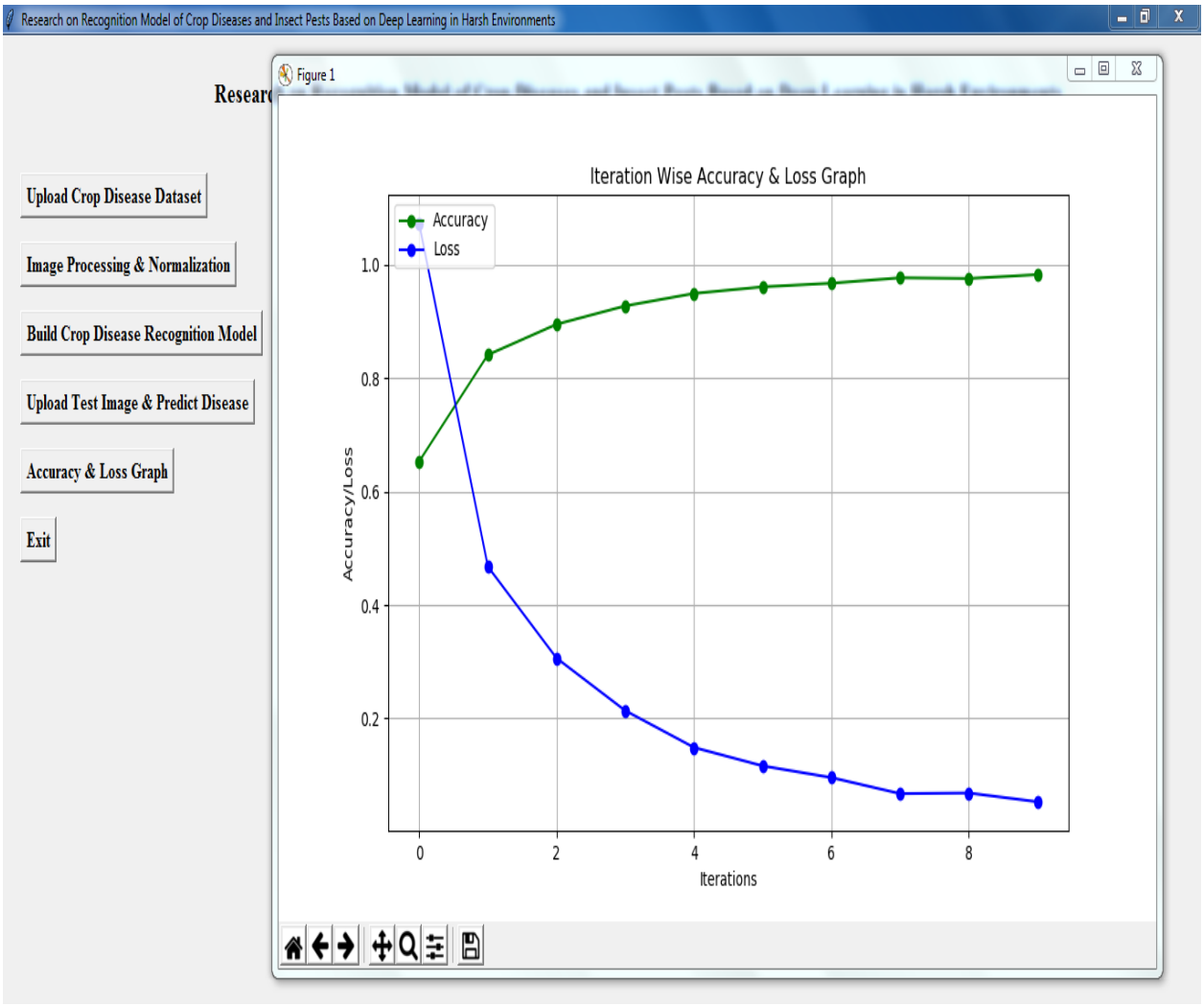
Screenshot 2. Crop Disease Recognition as Potato_Early_Blight

V RESULTS AND ANALYSIS

The Top1 accuracy in classification problems is the evaluation index utilized in this paper. It refers to the accuracy rate ACC of the model and actual class with the highest recognition probability. Formula 1, where N is the number of samples and R is the number of right predictions, is shown.

$$Acc = \frac{R}{N} \qquad (1)$$

The dataset's photos are preprocessed before being trained. Verification is done after each epoch iteration. Screenshot 3 depicts the image convergence process. The curve of the convolution neural network training model utilized in this paper, as seen in the graph, remains steady after three epochs of training, and its accuracy and loss keep a relatively stable state. The ultimate accuracy is 98 %, and the recognition impact is satisfactory.



Screenshot 3. Iteration wise accuracy and loss

VI CONCLUSION AND FUTURE WORK

This paper looked at 27 different types of disease recognition in 10 different crops. Deep learning theory and convolution neural network technologies are used to build the CNN-Inception-ResNet-v2 model. Experiments reveal that the model is capable of recognizing the data set, with an overall recognition

accuracy of 98 percent. The results reveal that this hybrid network model has greater recognition accuracy than the standard model, and it can be used to efficiently identify and detect plant diseases and insect pests.

There are two areas that should be improved in future work:

1) A larger data set. Only 27 diseases of ten crop species were evaluated in this research; additional species and diseases, such as rice and wheat, and their linked diseases, were not included. As a result, the next step in the study process is to gather more crop species and disease photos for research.

2) Improve the model. We can observe from the results of this paper's experiment that the Inception-resnet-v2 mixed network has absorbed the equivalent benefit. This model has high recognition accuracy and deserves more research and optimization. At the same time, we should create a network model that can accurately classify crop photos with higher accuracy.

REFERNCES

[1]. Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks*, *61*, 85-117.

[2]. Lee, S. H., Chan, C. S., Mayo, S. J., & Remagnino, P. (2017). How deep learning extracts and learns leaf features for plant classification. *Pattern Recognition*, *71*, 1-13.

[3]. Lee, Sue Han, et al. "Plant Identification System based on a Convolutional Neural Network for the LifeClef 2016 Plant Classification Task." CLEF (Working Notes). 2016.

[4]. López, O.; Rach, M.M.; Migallon, H.; Malumbres, M.P.; Bonastre, A.; Serrano, J.J. Monitoring Pest Insect Traps by Means of Low-Power Image Sensor Technologies. Sensors 2012, 12, 15801-15819

[5]. Khalid Zamir Rasib*, Sundaisa Abru and Arif Malik, “Agri Res & Tech: Open Access J 24(3): ARTOAJ.MS.ID.556276 (2020)”, 00104-00114.

[6]. Liu, Bin, et al. "Identification of apple leaf diseases based on deep convolutional neural networks." Symmetry 10.1 (2017).

ONE PAGE PROFILES

1. CH. Ravali

CHAGANTI RAVALI is currently pursuing her Bachelor of Technology in the stream of Computer Science and Engineering at St.Martin's Engineering College. She completed her intermediate from Narayana Junior College and schooling from Gayatri Vidyaalaya. Her technical skills include C,C++and Python. She took part in Employability Skill development Program conducted by Zensar. Her participations include: National Level Three Day Online Workshop on "AI & MI in Speech and Audio Processing" which was conducted from 10th to 12th December 2020, Women online workshop on "Women in Cyber Security and Privacy in 2020" which was conducted from 6th to 10th July 2020, "One Day Webinar on Internet of Things and Its Applications" conducted by Anand Institute of Higher Technology on 21st May 2020,IIC Online Sessions conducted by Institution's Innovation Council(IIC) of MHRD's Innovation Cell on 16th May 2020 Leadership talk, Two days National Workshop on "Android App Development" an outreach workshop on Technex'20,IIT Varanasi in association with Innovians Technologies(India) held at Megha Institute Of Engineering & Technology For Women, on 27th & 28th Feb,2020 and "Machine Learning Using Python" a three day workshop organized by the Department of Electronics and Communication Engineering at Vignana Bharathi Institute of Technology under Equity Action Plan,TEQIP-III,JNTUH conducted on 5th to 7th March 2020. Her areas of interest is python. She has completed few certification courses from online platforms like Coursera, CursaApp and SoloLearn.

2. Ch. Aishwarya

Chennam Aishwarya Reddy is currently pursuing her Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. She completed her intermediate from Sri Medha V Junior College and 10th class from Chanakya School. She is one of the members of Coders Club in our college. Her responsibilities in that group include mentoring and motivating students to take coding as a serious hobby. Her technical skills include C, Python and Java. She also has a basic understanding of C++. She took part in Employability Skill development Program conducted by Zensar. She is also a student of Smart Interviews. Her participations include: National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020, "Know More - Teach More ", the Global Webinar on Cloud & Big Data – Changing the way we work which was conducted Global Education & Careers Forum (GECF) on 12th August 2020, Women online workshop on "Women in Cyber Security and Privacy in 2020" which was conducted from 6th to 10th July 2020, "Know More - Teach More ", the Global Webinar on Cyber Threats and Defense Techniques conducted by GECF on 22nd July 2020, "One Day Webinar on Internet of Things and Its Applications" conducted by Anand Institute of Higher Technology on 21st May 2020 and IIC Online Sessions conducted by Institution's Innovation Council (IIC) of MHRD's Innovation Cell, New Delhi to promote Innovation, IPR, Entrepreneurship, and Start-ups among HEIs from 28th

April to 22nd May 2020. Her areas of interest are Python, Artificial Intelligence, Machine Learning and Deep Learning. She completed few certification courses from online platforms like Coursera, CursaApp and SoloLearn.

3. K. Prem Sai

Katukam Prem Sai is pursuing his Bachelor of Technology in the stream of Computer science and engineering at St.Martin's Engineering College. He completed his intermediate from Sri Chaitanya Junior Kalasala and schooling from Gowtham Model School. His technical skills include C, C++, Java, HTML and Python. His area of interest is Data Science. He has completed few certificate courses from online platforms like Coursera on Python Programming, AI For Everyone, HTML5, CSS, Data Analysis, Managing Project Risks and Changes. His participations include National Level Three Day Online Workshop on "AI & ML in speech and audio processing" which was conducted from 10th and 12th December, 2020, Leadership Talk with Mr.Mahesh Babu CEO Mahindra Electric Mobility Ltd. He even worked as a Transaction Risk Investigator role at Amazon for 6 months.

4. M. Shirisha

MADA SHIRISHA is currently pursuing her Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. She completed her intermediate from Sri Chaitanya Junior College and completed her schooling from Sri Sai Chaitanya Talent School. Her technical skills include C, C++ and Python. Her participations include: National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020, "Women online workshop on "Women in Cyber Security and Privacy in 2020" which was conducted from 6th to 10th July 2020, "One Day Webinar on Internet of Things and Its Applications" conducted by Anand Institute of Higher Technology on 21st May 2020,Leadership Talk conducted by Institution's Innovation cell by MHRD'S INNOVATION CELL on 6th Jun 2020 ,Two days National Workshop on "Android App Development" an outreach workshop on Technex'20,IIT Varanasi in association with Innovians Technologies(India) held at Megha Institute Of Engineering & Technology For Women, on 27th & 28th Feb,2020 and online Quiz program in compiler design on 18th May 2020 conducted by St.Peter's Engineering College. Her areas of interest is python.She has completed few certification courses from online platforms like Coursera, CursaApp and SoloLearn.

APPENDICES

```
from tkinter import messagebox

from tkinter import *

from tkinter import simpledialog

import tkinter

import matplotlib.pyplot as plt

import numpy as np

from tkinter import ttk

from tkinter import filedialog

from keras.utils.np_utils import to_categorical

from keras.models import Sequential

from keras.layers.core import Dense,Activation,Dropout, Flatten

from sklearn.metrics import accuracy_score

import os

import cv2

from keras.layers import Convolution2D

from keras.layers import MaxPooling2D

import pickle

from keras.models import model_from_json


main = Tk()

main.title("Research on Recognition Model of Crop Diseases and Insect Pests Based on Deep Learning in Harsh Environments")


main.geometry("1300x1200")
```

```
global filename

global X, Y

global model

global accuracy


plants = ['Pepper__bell___Bacterial_spot', 'Pepper__bell___healthy', 'Potato___Early_blight',
'Potato___healthy', 'Potato___Late_blight', 'Tomato_Bacterial_spot', 'Tomato_Early_blight',
'Tomato_healthy', 'Tomato_Late_blight', 'Tomato_Leaf_Mold', 'Tomato_Septoria_leaf_spot',
'Tomato_Spider_mites_Two_spotted_spider_mite', 'Tomato__Target_Spot',
'Tomato__Tomato_mosaic_virus', 'Tomato__Tomato_YellowLeaf__Curl_Virus']


def uploadDataset():

    global X, Y

    global filename

    text.delete('1.0', END)

    filename = filedialog.askdirectory(initialdir=".")

    text.insert(END,'dataset loaded\n')


def imageProcessing():

    text.delete('1.0', END)

    global X, Y

    X = np.load("model/myimg_data.txt.npy")

    Y = np.load("model/myimg_label.txt.npy")

    Y = to_categorical(Y)

    X = np.asarray(X)

    Y = np.asarray(Y)
```

```
    X = X.astype('float32')
    X = X/255
    indices = np.arange(X.shape[0])
    np.random.shuffle(indices)
    X = X[indices]
    Y = Y[indices]
    text.insert(END,'image processing completed\n')
    img = X[20].reshape(64,64,3)
    cv2.imshow('ff',cv2.resize(img,(250,250)))
    cv2.waitKey(0)

def cnnModel():
    global model
    global accuracy
    text.delete('1.0', END)
    if os.path.exists('model/model.json'):
        with open('model/model.json', "r") as json_file:
            loaded_model_json = json_file.read()
            model = model_from_json(loaded_model_json)
        json_file.close()
        model.load_weights("model/model_weights.h5")
        model.make_predict_function()
        print(model.summary())
        f = open('model/history.pckl', 'rb')
        accuracy = pickle.load(f)
```

```
    f.close()
    text.insert('accuracy')
    acc = accuracy['accuracy']
    acc = acc[9] * 100
    text.insert(END,"CNN Crop Disease Recognition Model Prediction Accuracy = "+str(acc))

else:
    model = Sequential() #resnet transfer learning code here
    model.add(Convolution2D(32, 3, 3, input_shape = (64, 64, 3), activation = 'relu'))
    model.add(MaxPooling2D(pool_size = (2, 2)))
    model.add(Convolution2D(32, 3, 3, activation = 'relu'))
    model.add(MaxPooling2D(pool_size = (2, 2)))
    model.add(Flatten())
    model.add(Dense(output_dim = 256, activation = 'relu'))
    model.add(Dense(output_dim = 15, activation = 'softmax'))
    model.compile(optimizer = 'adam', loss = 'categorical_crossentropy', metrics = ['accuracy'])
    print(model.summary())
    hist = model.fit(X, Y, batch_size=16, epochs=10, validation_split=0.2, shuffle=True, verbose=2)
    model.save_weights('model/model_weights.h5')
    model_json = model.to_json()
    with open("model/model.json", "w") as json_file:
        json_file.write(model_json)
    json_file.close()
    f = open('model/history.pckl', 'wb')
    pickle.dump(hist.history, f)
```

```
        f.close()
        f = open('model/history.pckl', 'rb')
        accuracy = pickle.load(f)
        f.close()
        acc = accuracy['accuracy']
        acc = acc[9] * 100
        text.insert(END,"CNN Crop Disease Recognition Model Prediction Accuracy = "+str(acc))


def predict():
    global model
    filename = filedialog.askopenfilename(initialdir="testImages")
    img = cv2.imread(filename)
    img = cv2.resize(img, (64,64))
    im2arr = np.array(img)
    im2arr = im2arr.reshape(1,64,64,3)
    test = np.asarray(im2arr)
    test = test.astype('float32')
    test = test/255
    preds = model.predict(test)
    predict = np.argmax(preds)
    img = cv2.imread(filename)
    img = cv2.resize(img, (800,400))
    cv2.putText(img, 'Crop Disease Recognize as : '+plants[predict], (10, 25),
cv2.FONT_HERSHEY_SIMPLEX,0.7, (0, 255, 0), 2)
    cv2.imshow('Crop Disease Recognize as : '+plants[predict], img)
```

```
    cv2.waitKey(0)


def graph():
    acc = accuracy['accuracy']
    loss = accuracy['loss']
    plt.figure(figsize=(10,6))
    plt.grid(True)
    plt.xlabel('Iterations')
    plt.ylabel('Accuracy/Loss')
    plt.plot(acc, 'ro-', color = 'green')
    plt.plot(loss, 'ro-', color = 'blue')
    plt.legend(['Accuracy', 'Loss'], loc='upper left')
    #plt.xticks(wordloss.index)
    plt.title('Iteration Wise Accuracy & Loss Graph')
    plt.show()


def close():
    main.destroy()
    text.delete('1.0', END)
    font = ('times', 15, 'bold')
title = Label(main, text='Research on Recognition Model of Crop Diseases and Insect Pests Based on Deep Learning in Harsh Environments')
#title.config(bg='powder blue', fg='olive drab')
title.config(font=font)
title.config(height=3, width=120)
```

```
title.place(x=0,y=5)

font1 = ('times', 13, 'bold')

ff = ('times', 12, 'bold')


uploadButton = Button(main, text="Upload Crop Disease Dataset", command=uploadDataset)

uploadButton.place(x=20,y=100)

uploadButton.config(font=ff)


processButton = Button(main, text="Image Processing & Normalization", command=imageProcessing)

processButton.place(x=20,y=150)

processButton.config(font=ff)


modelButton = Button(main, text="Build Crop Disease Recognition Model", command=cnnModel)

modelButton.place(x=20,y=200)

modelButton.config(font=ff)


predictButton = Button(main, text="Upload Test Image & Predict Disease", command=predict)

predictButton.place(x=20,y=250)

predictButton.config(font=ff)


graphButton = Button(main, text="Accuracy & Loss Graph", command=graph)

graphButton.place(x=20,y=300)

graphButton.config(font=ff)


exitButton = Button(main, text="Exit", command=close)
```

```
exitButton.place(x=20,y=350)

exitButton.config(font=ff)

font1 = ('times', 12, 'bold')

text=Text(main,height=30,width=85)

scroll=Scrollbar(text)

text.configure(yscrollcommand=scroll.set)

text.place(x=450,y=100)

text.config(font=font1)

main.config()

main.mainloop()
```

A

PROJECTREPORT

On

STOCK MARKET TREND PREDICTION USING KNN ALGORITHM

Submitted by

1) Mr. B. Varun Aditya (17K81A0564

2) Mr. K. Harsha Vardhan (17K81A0585)

3) Mr. D. Raghavender Reddy (17K81A0571)

In partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Under the Guidance of

Mr. J. Manikandan

Assistant professor (CSE)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ST.MARTIN'S ENGINEERING COLLEGE
An Autonomous Institute

Dhulapally, Secunderabad – 500 100

JUNE 2021

BONAFIDE CERTIFICATE

This is to certify that the project entitled STOCK MARKET ANALYSIS USING KNN ALGORITHM, is being submitted by **B. VarunAditya(17K81A0564)**, **K.HarshaVardhan(17K81A0585), D. Raghavender Reddy(17K81A0571)** in partial fulfillment of the requirement for the award of the degree of **BACHELOR OF TECHNOLOGY IN COMPUTER SCIENCE AND ENGINEERING** is recorded of bonafide work carried out by them. The result embodied in this report have been verified and found satisfactory.

J. MANIKANDAN

Department of CSE

Head of the Department

Dr.M.NARAYANAN

Department of CSE

Internal Examiner

External Examiner

Place:

Date:

DECLARATION

We, the student of **Bachelor of Technology** in Department of Computer Science and Engineering', session: <2017 – 2021>, St. Martin's Engineering College, Dhulapally, Kompally, Secunderabad, hereby declare that work presented in this Project Work entitled **STOCK MARKET TREND PREDICTION USING KNN** is the outcome of our own bonafide work and is correct to the best of our knowledge and this work has been undertaken taking care of Engineering Ethics. This result embodied in this project report has not been submitted in any university for award of any degree.

B.VarunAditya17K81A0564
K.HrashaVardhan17K81A0585
D.Raghavender Reddy 17K81A0571

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible and whose encouragements and guidance have crowded effects with success.

We extended our deep sense of gratitude to Principal**, Dr. P. SANTOSH KUMARPATRA**, St. Martin's Engineering College, Dhulapally, for permitting us to undertake this project. We are also thankful to **Dr. M.NARAYANAN**, Head of the Department, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project. We would like to thank **Dr. T. POONGOTHAI,** Department of Computer Science and Engineering (AI & ML) for her encouragement and insightful comments and as well as our project coordinators **Dr. B.RAJALINGAM**, Associate Professor and **Mr. J.SUDHAKAR**, Assistant Professor, in Department of Computer Science and Engineering for their valuable support. We would like to express our sincere gratitude and indebtedness to our project supervisor <Guide Name, Designation>, Computer Science and Engineering,St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.Finally, we express thanks to all those who have helped us successfully to completing this project. Furthermore, we would like to thank our family and friends for their moral support and encouragement. We express thanks to all those who have helped us in successfully completing the project.

B.VarunAditya 17K81A0564

K.HarshaVardhan 17K81A0585

D.Raghavender Reddy 17K81A0571

ABSTRACT

This paper examines a hybrid model which combines a K-Nearest Neighbors (KNN) approach with a probabilistic method for the prediction of stock price trends. One of the main problems of KNN classification is the assumptions implied by distance functions. The assumptions focus on the nearest neighbors which are at the centroid of data points for test instances. This approach excludes the non-centric data points which can be statistically significant in the problem of predicting the stock price trends. For this it is necessary to construct an enhanced model that integrates KNN with a probabilistic method which utilizes both centric and non-centric data points in the computations of probabilities for the target instances. The embedded probabilistic method is derived from Bayes' theorem. The prediction outcome is based on a joint probability where the likelihood of the event of the nearest neighbors and the event of prior probability occurring together and at the same point in time where they are calculated. The proposed hybrid KNN Probabilistic model was compared with the standard classifiers that include KNN, Naive Bayes, One Rule (OneR) and Zero Rule (ZeroR). The test results showed that the proposed model outperformed the standard classifiers which were used for the comparisons.

Predicting the Stock Market has been the bane and goal of investors since its existence. Everyday billions of dollars are traded on the exchange, and behind each dollar is an investor hoping to profit in one way or another. Entire companies rise and fall daily based on the behaviour of the market. Should an investor be able to accurately predict market movements, it offers a tantalizing promises of wealth and influence. It is no wonder then that the Stock Market and its associated challenges find their way into the public imagination every time it misbehaves. The 2008 financial crisis was no different, as evidenced by the flood of films and documentaries based on the crash. If there was a common theme among those productions, it was that few people knew how the market worked or reacted. Perhaps a better understanding of stock market prediction might help in the case of similar events in the future.Despite its prevalence, Stock Market prediction remains a secretive and empirical art. Few people, if any, are willing to share what successful strategies they have. A chief goal of this project is to add to the academic understanding of stock market prediction. The hope is that with a greater understanding of how the market moves, investors will be better equipped to prevent another financial crisis. The project will evaluate some existing strategies from a rigorous scientific perspective and provide a quantitative evaluation of new strategies.

Keywords: Stock Price Prediction, K-Nearest Neighbours, Baye's Theorem, Naive Baye's, Probabilistic Method.

TABLEOFCONTENTS

LISTOFTABLES

LIST OF FIGURES

LISTOFACRONYMS

<KNN>	K Nearest Neighbor
<RSME>	Root Mean Square Error
<MAE>	Mean Average Error
<OHLC>	Open-High-Low-Close
<GUI>	Graphical user interface

INTRODUCTION

1.1 PROJECT OVERVIEW

Analyzing financial data in securities has been an important and challenging issue in the investment community. Stock price efficiency for public listed firms is difficult to achieve due to the opposing effects of information competition among major investors and the adverse selection costs imposed by their information advantage.

There are two main schools of thought in analyzing the financial markets. The first approach is known as fundamental analysis. The methodology used in fundamental analysis evaluates a stock by measuring its intrinsic value through qualitative and quantitative analysis. This approach examines a company's financial reports, management, industry, micro and macro-economic factors.

The second approach is known as technical analysis. The methodology used in technical analysis for forecasting the direction of prices is through the study of historical market data. Technical analysis uses a variety of charts to anticipate what are likely to happen. The stock charts include candlestick charts, line charts, bar charts, point and figure charts, OHLC (open-high-low-close) charts and mountain charts. The charts are viewable in different time frames with price and volume. There are many types of indicators used in the charts, including resistance, support, breakout, trending and momentum.

Several alternatives to approach this type of problem have been proposed, which range from traditional statistical modelling to methods based on computational intelligence and machine learning. They categorized the papers reviewed in the following areas: time series, optimization, hybrid methods, pattern recognition and classification. Within the context of financial trading discipline, the survey showed that most of the research was being conducted in the field of technical analysis. An integrated fundamental and technical analysis model was examined to evaluate the stock price trends by focusing on macro-economic analysis. It also analyzed the company behaviour and the associated industry in relation to the economy which in turn provide more information for investors in their investment decisions.

A nearest neighbor search (NNS) method produced an intended result by the use of KNN technique with technical analysis. This model applied technical analysis on stock market data which include historical price and trading volume. It applied technical indicators made up of stop loss, stop gain and RSI filters. The KNN algorithm part applied the distance function on the collected data. This model was compared with the buy-and-hold strategy by using the fundamental analysis approach.

The KNN algorithm method is used on the stock data. Also, mathematical calculations and visualization models are provided and discussed below.

The KNN algorithm is used to measure the distance between the given test instance and all the instances in the data set, this is done by choosing the 'k' closest instances and then predict the class value based on these nearest neighbors. The 'k' is assigned as number of neighbors voting on the test instance. As such

KNN is often referred to as case based learning or an instance-based learning where each training instance is a case from the problem domain. KNN is also referred to as a lazy learning algorithm due to the fact that there is no learning of the model required and all of the computation works happen at the time a prediction is requested. KNN is a non-parametric machine learning algorithm as it makes no assumptions about the functional form of the problem being solved. Each prediction is made for a new instance (x) by searching through the entire training set for the 'k' most nearest instances and applying majority voting rule to determine the prediction outcome.

A variety of distance functions are available in KNN which include Euclidean, Manhattan, Minkowski and Hamming. The Euclidean distance function is probably the most commonly used in any distance-based algorithm. It is defined as $d(x,y) = \sqrt{\sum} (\boldsymbol{xi} - \boldsymbol{yi}) \boldsymbol{k2i=1}$

(1) Where, x and y are two data vectors and k is the number of attributes. The Manhattan distance function is defined as $d(x,y) = (\sum |\boldsymbol{xi} - \boldsymbol{yi}|) \boldsymbol{ki=1}$

(2) The Minkowski distance function is defined as $d(x,y) = (\sum (|\boldsymbol{xi} - \boldsymbol{yi}|) \boldsymbol{p}) \boldsymbol{1/pki=1}$) (3) The distance functions for Euclidean, Manhattan and Minkowski are used for numerical attributes. The Hamming distance function is defined as $d(x,y) = \sum |\boldsymbol{xiki=1} - \boldsymbol{yi}|$ x = y => D = 0 x ≠ y => D = 1

(4) Hamming distance is usually used for categorical attributes. The Hamming distance between two data vectors is the number of attributes in which they differ.

Advantages of KNN Algorithm:

• It is simple to implement.

• It is robust to the noisy training data

• It can be more effective if the training data is large.

Disadvantages of KNN Algorithm:

• Always needs to determine the value of K which may be complex some time.

• The computation cost is high because of calculating the distance between the data points for all the training samples.

The probabilistic method calculates the prior probabilities of Profit class and Loss class based on the number of instances in the data set. The outcome from the earlier KNN approach is used as an input as the probabilities of Profit class and Loss class by the nearest neighbors method. The joint probabilities of Profit class and Loss class can then be calculated using the outcomes of the prior probabilities and the calculated KNN's probabilities. Finally, the predictive decision is made by comparing the joint probabilities of Profit class and Loss class

1.2. PROJECT OBJECTIVES

- The aim of our project is to help the stock brokers and investors for investing money or stocks. The prediction plays a very important role in stock business which is complicated and challenging process due to dynamic nature of stock market.

- As per the discussed works above, predictions of the stock prices based on KNN algorithm The aim of this research is to improve the statistical fitness of the proposed model to overcome a KNN problem due to its computation approach.

- We have compared a hybrid KNN-Probabilistic model with four standard algorithms, our results showed that the proposed KNN-Probabilistic model leads to significantly better results compared to the other classification algorithms.

- To provide 100% accuracy when compared to other approaches and the predicted results are nearly 90% true.

- The rate of prediction should also increase by using this algorithm.

- To make sure that the predicted price results are genuine and true.

1.3 ORGANIZATION OF CHAPTERS

This documentation consists of 10 differrent chapter and they are;

1. Introduction – This chapter covers the overview of our project and its objectives.

2. Literature Survey – This includes the details of our survey.

3. Software and Hardware Requirements – We specify our software and hardware requirements here.

4. Software Development Analysis – This section includes the problem definition and details of the modules we used in our project.

5. Project System Design – This chapter includes the design part of our project which includes uml diagrams.

6. Project Coding – This section contains the details of our project code.

7. Project Testing – The details of test cases and testing are included in this chapter.

8. Output Screens – This contains the screenshots of how our project looks like when executed.

9. Experimental Results – This chapter contains the screenshots of our results.

10. Conclusion and Future Enhancements – This covers the conclusion of our project and the possible future developments.

2. LITERATURE SURVEY

2.1 SURVEY ON BACKGROUND

Financial services companies are developing their products to serve future prediction. There are a large amount of financial information sources in the world that can be valuable research areas, one of these areas is stock prediction and also called stock market mining. Stock prediction becomes increasingly important especially if number of rules could be created to help making better investment decisions in different stock markets

1.Sneh Kalra et al. in 2019, in this paper authors, did research on the **fluctuation of stock market prices** with respect to the relevant new articles of a company. They used classifier Naïve Bayes to separate negative or positive statements for prediction purposes based on daily news variance the social media data, blogs data may be considered for future work .

AUTHOR: Sneh Kalra

2. Aditya Menon et al. in 2019, this paper is focused on a review of **neural model for forecast the stock trend** after reviewing on a neural model they think that The long short term memory algorithm for predicting the economic information in confluence into the trendy era, this would be prioritized algorithm for forecasting

AUTHOR: Aditya Menon

3. Ashish Sharma et al. in 2017, they found that **regression analysis** is mostly used for stock market trend prediction they survey of regression technique for stock prediction using stock market data. In the future result could improve by using more numbers variables

AUTHOR: **Ashish Sharma**

4. Mu yen chen et al. in 2019, authors did research to calculate the impact of news articles on the **stock prices using deep learning approach LSTM (long short-term memory)** and they think this study can predict the stock market trend.

AUTHOR: Mu yen chen

5.Andrea Picasso et al. in 2019, in this research, authors worked which will alliance the **economic and elemental analysis for market trend prediction** through the various kind of application and automation methods neural network is machine learning technique the problem of trend stock and those are charts with forecasting data. As an input data sentiment of a news article is exploited. According to their research the problem in the most problematic accomplishment among the use of information about news astral one-off. To overcome this problem in the future the proper feature fusion technique will be suitable for the future

AUTHOR: Andrea Picasso

6.Gangadhar Shobha et al. in 2018, this paper provided a full **overview of machine learning techniques** which will help to reader for use of equations and concept the author discussed about three type of all machine learning technique and also various kind of metrics like accuracy, confusion matrix, recall, RMSE, precision and quintile of errors. The author thinks that this review can help those people who are new to machine learning because most of the people confuse to use most of the machine learning techniques for prediction or others .

AUTHOR: Gangadhar Shobha

7.Suryoday Basak et al. in 2018, the author developed an **experimental framework for predicting stock prices** whether the price goes up or down in this experiment author uses the two algorithms name as a random forest classifier and Gradient boosted decision' n trees, and they got more accuracy in comparison to others research papers where others experiments got 50% to 67% results on the other hand according to the author of this paper they got 78% result accuracy for long term window. In the future, they could use the build boosted tree model for short term data window .

AUTHOR: Suryoday Basak

8.Arash Negahdari kia et al. in 2018, as the stock prediction so many experiments and models, have been developed for prediction purpose on historical data like as in this paper the author present **HyS3 graph-based semi-supervised model and through a network views Kruskal based graph algorithm called ConKruG.** In the future they think social media data, Twitter data could be used for the prediction of stock for better results using these algorithms.

AUTHOR: Arash Negahdari

9.Bruno Miranda et al. in 2019, in this paper the survey of bibliographic techniques that focus on text area for research the author works on the **prediction of financial market values by using the machine learning models support vector machine (SVM) and neural network** with data set from North American market new models may have opportunities for north American market data for prediction purpose in future .

AUTHOR: Bruno Miranda

10. K. Hiba Sadia et al. in 2019, author aim for this paper us to preprocess the raw data firstly then they are doing a **comparison between random forest and SVM algorithm** the main aim of the author is to find out the best algorithm for stock trend prediction in the last they have given the best-fit algorithm for future stock forecasting which is random forest algorithm for future work they think that for getting more accuracy in result the adding of more parameters can be good .

AUTHOR: . K. Hiba Sadia

11.A. Akash et al. in 2019, the author introduced two more algorithm name as **"LS SVM"** which is least square support vector machine and another one is "PSO" (particle swarm optimization) the work "PSO" basically select the bestunbounded parameter with the "LS SVM" to reduce the overfitting and some technical indicators which will basically enhance the result accuracy. On the other hand at the same time, the proposed algorithm is being compared with artificial neural network model .

AUTHOR: A. Akash

12.Aparna Nayak et al.in 2016, in this paper authors, woked to **predict the stock market trend by using the supervised learning methods**, here authors predicted the data based on daily live data which is directly calling by the program using yahoo financial website and also predicting the monthly based prediction where in this paper they got a better result for daily live prediction instead of monthly prediction further future work they think if we consider more sentiments to the monthly [prediction that would also generate the best result.

AUTHOR: Aparna Nayak

13.Nuno Oliveira et al. in 2016, in this paper the author purposed a methodology by which they can access the value of **stock prediction and microblogging** data they used, for stock prices and return indices and some more like a portfolio. For this experiment, they have used huge data of Twitter, for all this experimental work they use Kalman filter to merge the microblogging data and some external sources and as a result, they found twitter data and blogging data were relevant for the purpose of forecasting these datasets were very useful. This result can be improved by using some more and different data such as social media datasets and others.

AUTHOR: **Nuno Oliveira**

14. Han lock Siew et al. in 2017, the author in this paper used regression technique for finding the **accuracy in the forecasting of a stock trend** for all this experiment they used WEKA software which used for data mining and machine learning algorithms to execute them, the dataset they used which contains heterogeneous values and which is used for handling of currency values and functional ratios. The dataset for calculating the stock movement is collected from Bursa Malaysia for forecasting purposes. For the future extension, the authors thought that the forecasting using the regression method can be improved by using the more standardized ordinal format of data .

AUTHOR: Han lock Siew

15. Smruti rekha das et al. in 2019, in this paper authors, used **firefly method for forecasting the stock prices** as an input dataset author collect from four different websites name as NSE-India, BSE, S&P 500 and FTSE, and all collected dataset is well transformed by using proper mathematical formulas by using the backpropagation, neural network and more two methods used for prediction, forecasting according to the time horizon of alternate days 1 day, 3 days, 5 days and so on. For future work, there may be some chance to get more accurate results by giving more parameters to the implemented algorithms .

AUTHOR: Smruti rekha das

16.Dattatray P.Gandhmal et al. in 2019, authors had written the paper on the review of stock prediction techniques. In this paper, the authors reviewed about 50 published research papers according to the publication years, and the authors suggested the best technique for prediction. **KNN and fuzzybased techniques** as the authors suggested are the best techniques according to the review such as KNN, SVM, SVR, and much more but these two techniques can be more effective for the purpose of using historical data. In the future, they will review more papers to get the best-fit algorithm for prediction.

AUTHOR: Dattatray P.Gandhmal

2.2 CONCLUSION OF SURVEY

These readings provide basic background information about various techniques and algorithms in machine learning in various stages such as KNN model and Linear regression model and in another stage Baye's theorem used by using in stock market trend prediction in every stage these various models used to test the stock market dataset provided by finance companies and other websites . So here also machine learning is used as it have the advantages of easily identifies trends and patterns, no human intervention needed (automation), continuous Improvement, handling multi-dimensional and multi-variety data,andWide Applications.. So machine learning used for predicting values and trends. Every reference gave an example to how to use machine learning with different algorithms giving us a basic idea about which the best to use.

3. SOFTWARE AND HARDWARE REQUIREMENTS

3.1 SOFTWARE REQUIREMENTS

- Operating System : Windows XP.
- Platform : PYTHON TECHNOLOGY
- Tool : Spyder, Python 3.5
- Front End : Anaconda
- Back End : python anaconda script

3.2HARDWARE REQUIREMENTS

- System: Pentium IV 2.4 GHz.
- Hard Disk : 40 GB.
- Monitor : 15 inch VGA Color.
- Mouse : Logitech Mouse.
- Ram : 512 MB
- Keyboard : Standard Keyboard.

4. SOFTWARE DEVLOPMENT ANALYSIS

4.1 PROBLEM OVERVIEW

The problem statement we have chosen to work with in this project follows from the hypothesis that the KNN algorithm is a more precise way of predicting closing prices than the MA formula is.

The problem statement is formulated below: Is using the KNN algorithm a more precise way of predicting the future closing prices of equities than using the more common method of MA?

This report will explore the ability of using the KNN algorithm and the MA formula as methods of predicting stock market movements. Several steps will be followed as we attempt to answer and explore the research question. –

As an initial step, the algorithm and the formula will be implemented in a programming language that fits the purpose of the research question and then tested to ensure that it performs as it should. - Data will then be gathered from the Stockholm stock exchange where the focused data will be on four Swedish companies' equities and their closing prices. This data will be collected from the yahoo finance stock dataset. The range of data will cover two years in order to get a sufficient amount of training data needed to get as accurate results as expected.

The results of the gathered data from the algorithm and the MA method will then be accumulated and displayed graphically through graphs, charts and explanatory text, in order to clearly display the accuracy.

The error calculations from running the algorithm and the MA method is the numerical quantification of the project. It is necessary in order to test the correctness of the data collected and in order to allow an evaluation of whether or not the method of using the KNN algorithm is a well working prediction model when attempting to predict stock market movements compared to the MA formula.

4.2 DEFINE THE PROBLEM

Predicting the Stock Market has been the bane and goal of investors since its existence. Everyday billions of dollars are traded on the exchange, and behind each dollar is an investor hoping to profit in one way or another. Entire companies rise and fall daily based on the behaviour of the market. Should an investor be able to accurately predict market movements, it offers a tantalizing promises of wealth and influence. It is no wonder then that the Stock Market and its associated challenges find their way into the public imagination every time it misbehaves. The 2008 financial crisis was no different, as evidenced by the flood of films and documentaries based on the crash. If there was a common theme among those productions, it was that few people knew how the market worked or reacted. Perhaps a better understanding of stock market prediction might help in the case of similar events in the future.

Despite its prevalence, Stock Market prediction remains a secretive and empirical art. Few people, if any, are willing to share what successful strategies they have. A chief goal of this project is to add to the academic understanding of stock market prediction. The hope is that with a greater understanding of how the market moves, investors will be better equipped to prevent another financial crisis. The project will evaluate some existing strategies from a rigorous scientific perspective and provide a quantitative evaluation of new strategies.

4.3 MODULES OVERVIEW

This project develops a framework based on the stock market prediction to make it easy and to understand the trade done by the traders. The software will predict the upcoming trends and prices of the stock of an organization. We implemented this project by dividing it into different modules. Firstly, the data is loaded. Secondly, the relations between the stocks. Next the data is pre-processed. Next the KNN with uniform weights runs and followed by KNN with distance weights. Then the data which we have collected should be uploaded and we get the predicted results. At the last it also shows the accuracy of the algorithm. So here we have used total seven different types of modules which performs different operations to get the results.

4.4 DEFINE THE MODULE

The project mainly consists of seven modules . They are,

1. Download the dataset.
2. Co-relation of data.
3. Data pre-processing.
4. Run KNN with uniform weights.
5. Run KNN with distance weights.
6. Upload test data.
7. KNN accuracy

4.5 MODULE FUNCTIONALITY

1. Download dataset : This module shows that we have to download the stock dataset into the loader.

2. Co-relation of data: This modules shows the correlation between Apple and Competitor Stock market Dataset.show the trend in the technology industry rather than show how competing stocks affect each other.

3. Data pre-processing: It pre-process the raw data, that means it removes all the noisy data and helps in removing the duplicate data. It fills the missing values, spilt the labels, trained data and test data.

4. Run KNN with uniform weights: It generates the KNN model with uniform weights and calculate the model accuracy.

5. Run KNN with distance weights: It generates the KNN model with distance weights and calculate the model accuracy.

6. Test data upload: here we have to upload the historical stock data of an organisation which we want to predict.

7.KNN accuracy: It shows the accuracy of the algorithm and also develops a graph of both the KNN models to show which is performing better.

5. PROJECT SYSTEM DESIGN

5.1 SYSTEM ARCHITECTURE

In this research paper, we implemented trend forecasting of stock market "the stock market trend prediction" using supervised methods like "linear regression", "random forest algorithm", "support vector machine algorithm", and Knearestneighbor algorithm. Here is the proposed work system architecture, which we have depicted through the stepwise structure, In the first step we are starting by giving raw data to our trained algorithm which will pre-process the data by using python libraries which is also a feature extraction part, Where it will the cleanup data by using data pre-processing method after that we have divided our data into two parts where 70% of our data is trained and remaining 30% of data which is for testing by using the trained algorithm after all this process we will only get our predicted data, as shown

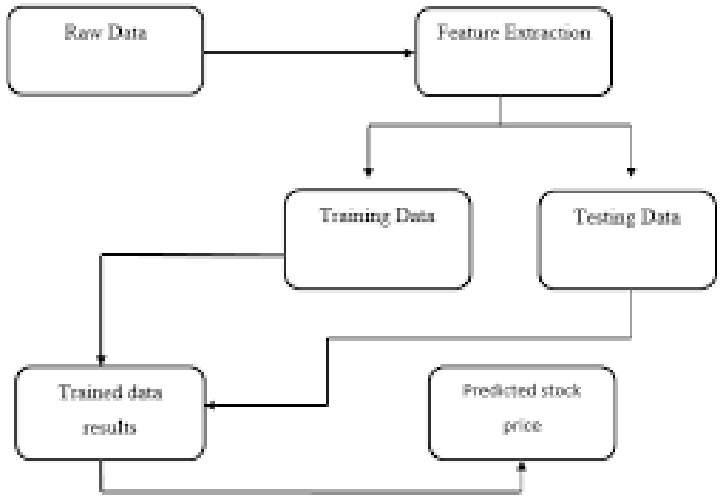


FIG 1 SYSTEM ARCHITECTURE

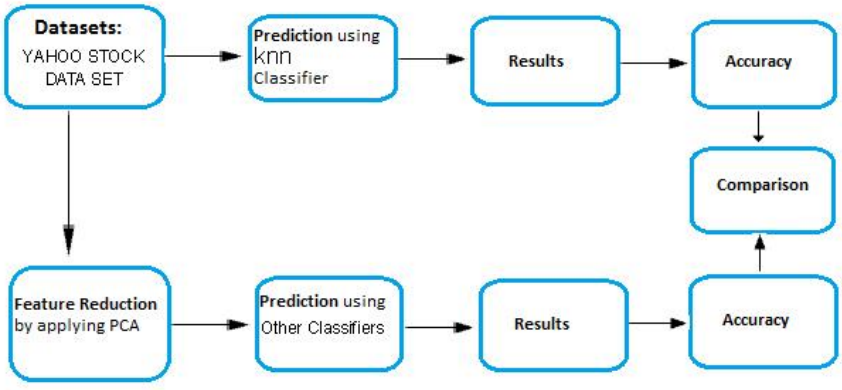


Fig. 1. Block diagram of the proposed research approach

KNN ARCHITECTURE AND E-R DIAGRAMS

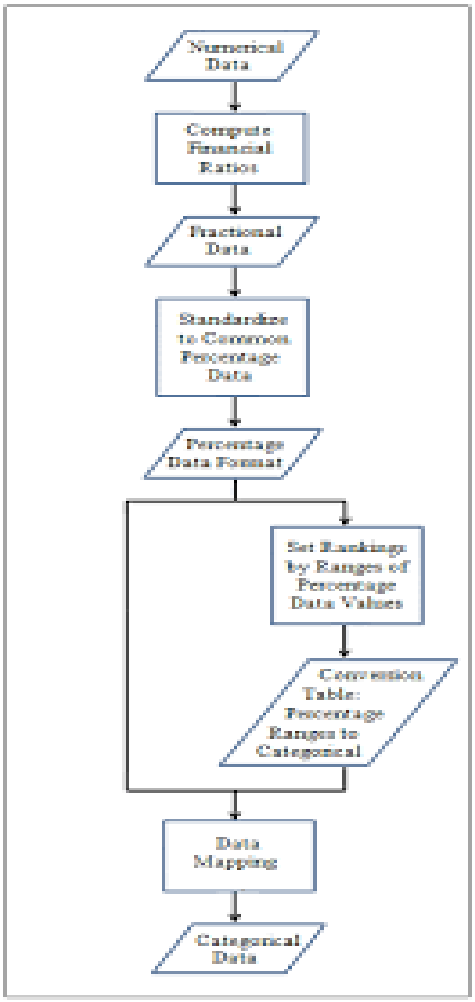


FIG 1.Classification by KNN

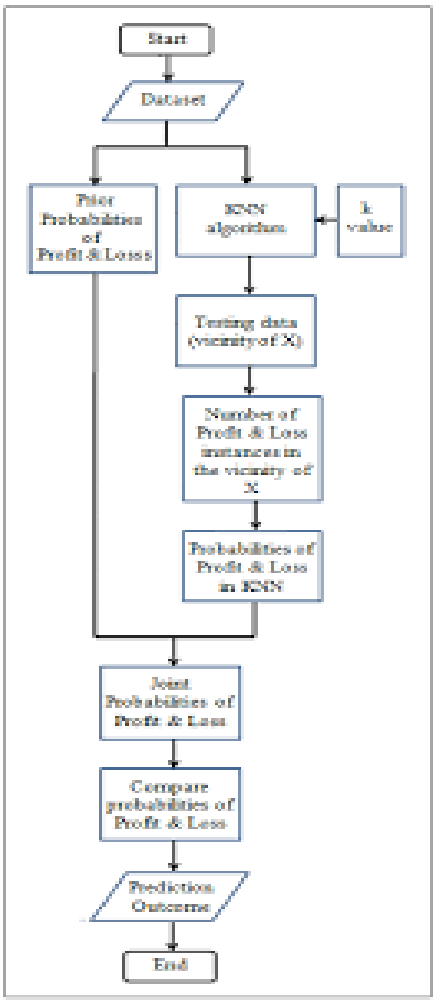


FIG 2.Classification by the probabilistic method

The steps adopted for classification by KNN are illustrated as follows:

Steps:

• Classification: KNN

• Initialization of k value on nearest neighbors

• Compute the distance between the X query instance and all the training samples.

• Sort the distance values

• Determine the nearest neighbors to the query instance based on the k value

• Calculate the number of Profit instances of the nearest neighbors in the vicinity of X query instance

• Calculate the number of Loss instances of the nearest neighbors in the vicinity of X query instance

The steps adopted for classification by the probabilistic method is illustrated as follows:

Steps:

• Classification: Probabilistic method

• Calculate the prior probabilities of Profit class and Loss class from the data set

• Calculate the KNN's probabilities of Profit class and Loss class based on the number of Profit nearest neighbors and the number of Loss nearest neighbors.

• Calculate the joint probabilities from the prior probabilities and KNN's probabilities on Profit class and Loss class

• Compare the joint probabilities of Profit class and Loss class

• Select the predictive value from the class values with the highest joint probability

5.2 UML DIAGRAMS

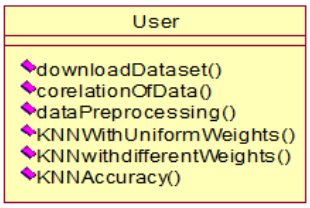
UML Diagrams: A use case diagram in the Unified Modelling Language (UML) is a type of behavioural diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted. Use Case diagrams are formally included in two modelling languages defined by the OMG: the Unified Modelling Language (UML) and the Systems Modelling Language (Sys ML). Types of UML Diagrams There are several types of UML diagrams and each one of them serves a different purpose. The two most broad categories that encompass all other types are Behavioral UML diagram and Structural UML diagram. As the name suggests, some UML diagrams try to analyze and depict the structure of a system or process, whereas other describe the behavior of the system, its actors, and its building components. The different types are broken down as follows:32

A. **Class Diagram**: Class UML diagram is the most common diagram type for software documentation. Since most software being created nowadays is still based on the Object-Oriented Programming paradigm, using class diagrams to document the software turns out to be a common-sense solution. This happens because OOP is based on classes and the relations between them. In a nutshell, class diagrams contain classes, alongside with their attributes (also referred to as data fields) and their behaviors (also referred to as member functions). More specifically, each class has 3 fields: the class name at the top, the class attributes right below 41 the name, the class operations/behaviors at the bottom. The relation between different classes (represented by a connecting line), makes up a class diagram

UML Diagrams

Class diagram

Fig 1



B. Use Case Diagram: A cornerstone part of the system is the functional requirements that the system fulfills. Use Case diagrams are used to analyze the system's high-level requirements. These requirements are expressed through different use cases. We notice three main components of this UML diagram: Functional requirements – represented as use cases; a verb describing an action Actors – they interact with the system; an actor can be a human being, an organization or an internal or external application Relationships between actors and use cases – represented using straight arrows.

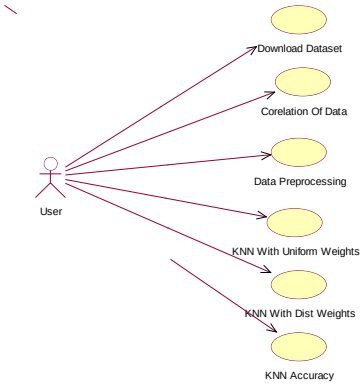


Fig 2 Use case diagram

C. Sequence UML Diagram: Sequence diagrams are probably the most important UML diagrams among not only the computer science community but also as design-level models for business application development. The sequence diagram number the actions starting from the data input to the optimal prediction. There is an arrow direction to show the sequence of flow for the action taking to arrive at the optimal prediction.

Here in the proposed diagram the sequential operations to be performed are given.

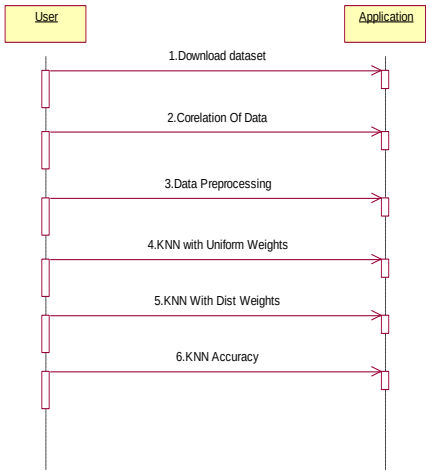


Fig 3 **Sequence diagram**

D. Collaboration diagram is a diagram that shows the interactions between elements at run-time in much the same manner as a Sequence diagram. However, Communicationdiagrams are used to visualize inter-object relationships, while Sequence diagrams are more effective at visualizing processing over time.

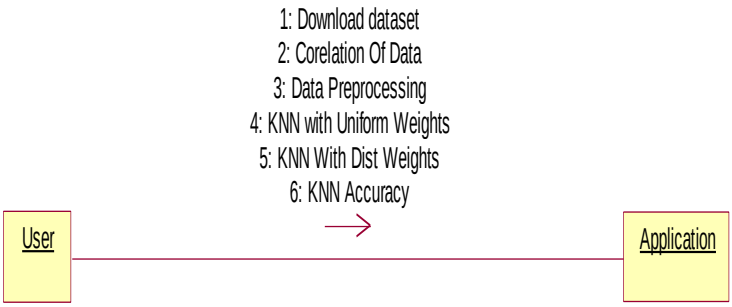
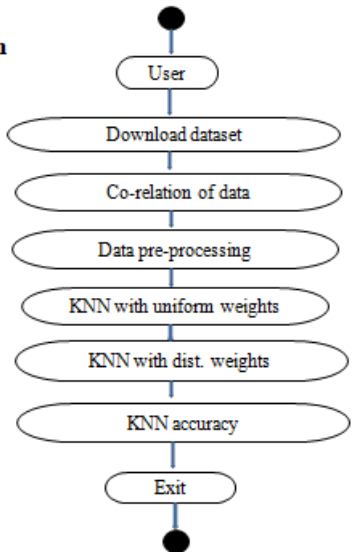


Fig 4 **Collaboration diagram**

E. Activity Diagram: An activity diagram is essentially a flowchart, showing flow of controls from one activity to another. Unlike a traditional flowchart, it can model the dynamic functional view of a system. An activity diagram represents an operation on some classes in the system that results to changes in the state of the system

Activity diagram

Fig 5



6. PROJECT CODING

6.1 CODE TEMPLATES

#import required modules
#Declare global variables

def loaddataset():

#yahoostockdataset folder containing stock data of apple and other organisatiowill be uploaded

def dfcorr():
#co-relation data is observed here.

def datapreprocess():
#datapreprocessing is done here.

def uniformKNN():
#knnwithuniform weights is runned.

def distKNN():
#knnwith distance weights is runned.

def predModel():
#testdata is uploaded for which wehave to predict the prices and the results are shown.

def graph():
#graph representing the accuracy level of the knn algorithm and comparing both uniform and distance knn

#exit from project

#Define the window

#create buttons for all methods used

Main.config()

Main.mainloop()

6.2 OUTLINE FOR VARIOUS FILES

We used python programming to implement our project. A simple python file is used to implement our code. This file consists of our modules that we have used. Our project modules are download the dataset, co-relation of data, data pre-processing, Run KNN with uniform weights, Run KNN with distance weights, upload test data, KNN accuracy. We also used various python modules like tkinter, pandas, matpolib, numpy, sklearn & imutils.

6.3 METHODS INPUT AND OUTPUT PARAMETERS

In our project code, we implemented six different methods. They are:

1. loadDataset()
2. datacorrelation()
3. datapreprocess()
4. knnuniformweights()
5. knndistweights()
6. testdataupload()
7. graph()

Our first method loadDataset() doesn't take any input parameters but after successful execution, it displays a message "dataset loaded". Second method datacorelation() doesn't have any input parameters and after successful completion, it displays a message " successful ". Third method datapreprocess() doesn't have any input parameters and after successful completion, it displays a message " data is preprocessed ". Fourth knnuniformweights() doesn't have any input parameters. After building the KNN algorithm, the accuracy of our project is displayed. knndistweights() doesn't have any input parameters. After building the

KNN algorithm, the accuracy of our project is displayed. testdataupload() have input parameters, we have to upload the data for which we are going to predict the future prices. Then on successful completion, it displays the predicted prices of the stock. graph() also don't have any input parameters but it displays a graph showing accuracy and loss versus iterations. close() don't have any parameters but upon clicking this button, it will close the project window.

7. PROJECT TESTING

7.1 VARIOUS TEST CASES

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub-assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

A. Test strategy and approach

Field testing will be performed manually and functional tests will be written in detail. Test objectives

• All field entries must work properly.

• Pages must be activated from the identified link.

• The entry screen, messages and responses must not be delayed.

Features to be tested50

• Verify that the entries are of the correct format

• No duplicate entries should be allowed

• All links should take the user to the correct page.

Integration Testing

Software integration testing is the incremental integration testing of two or more integrated software components on a single platform to produce failures caused by interface defects. The task of the integration test is to check that components or software applications, e.g. components in a software system or – one step up – software applications at the company level – interact without error.

Test Results: All the test cases mentioned above passed successfully. No defects encountered

Acceptance Testing

User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

B. Verification and Validation

The testing process is a part of broader subject referring to verification and validation. We have to acknowledge the system specifications and try to meet the customer's requirements and for this sole purpose, we have to verify and validate the product to make sure everything is in place. Verification and validation are two different things. One is performed to ensure that the software correctly implements a specific functionality and other is done to ensure if the customer requirements are properly met or not by the end product. Verification of the project was carried out to ensure that the project met all the requirement and specification of our project. We made sure that our project is up to the standard as we planned at the beginning of our project development.

7.2 WHITE BOX TESTING

White Box Testing is a testing in which in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is purpose. It is used to test areas that cannot be reached from a black box level.

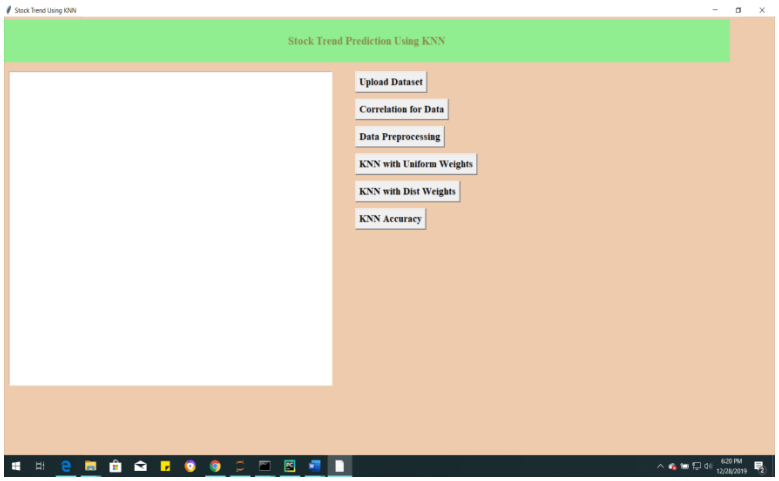
7.3 BLACK BOX TESTING

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box .you cannot "see" into it. The test provides inputs and responds to outputs without considering how the software works.

8. OUTPUT SCREENS

8.1 USER INTERFACE

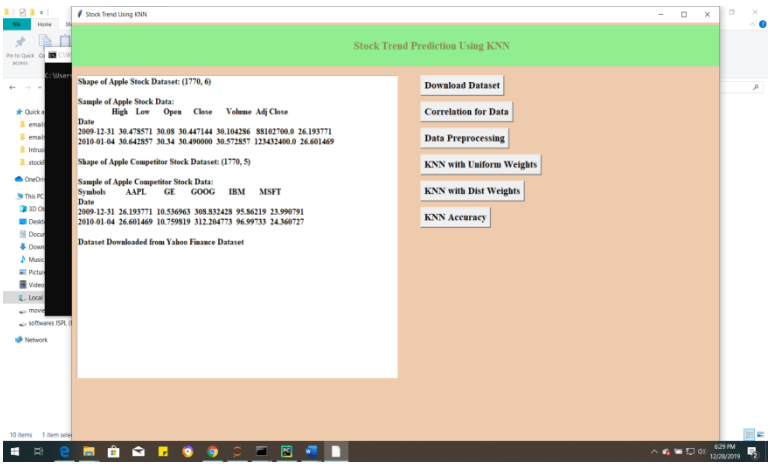
double click on 'run.bat' file to get below screen



(fig1) user interface

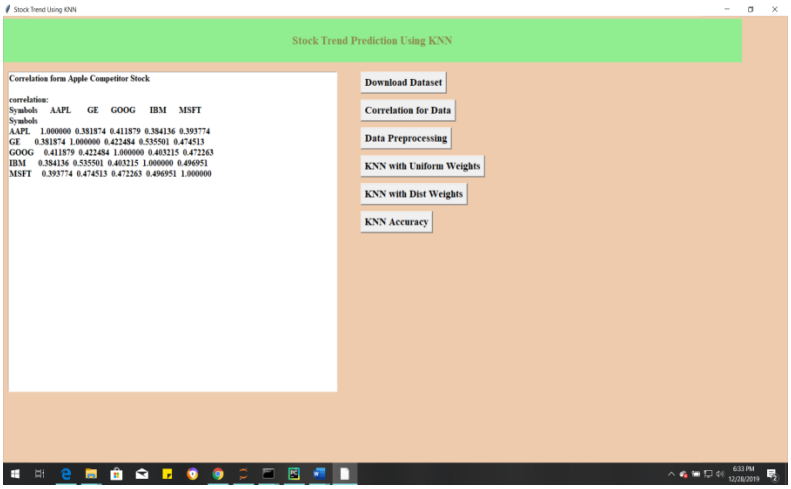
In above screen click on 'Download Button' download the Apple Stock and competitors data from Yahoo Finance Dataset (fig1)

8.2 OUTPUT SCREENS



(fig2) loading dataset

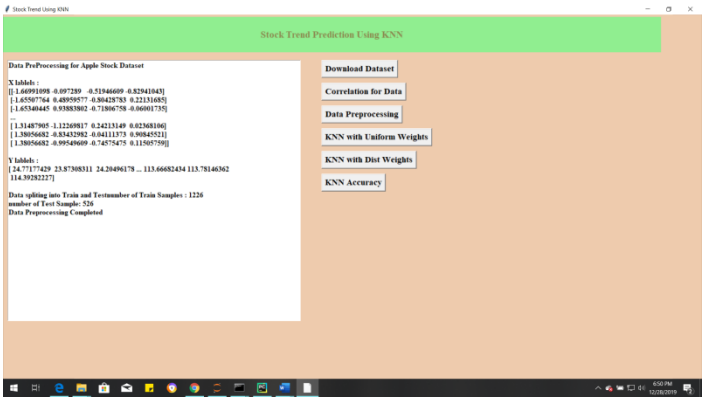
In above screen I am Downloading of Apple Stock and Apple competitor Stock Data from Yahoo Finance Dataset.(fig2)



(fig3) corealtion of dataset

Now click on 'Correlation Data' Button to find the correlation between Apple and Competitor Stock market Dataset.show the trend in the technology industry rather than show how competing stocks affect each other.

Now click on 'Data PreProcessing button to drop missing values, split labels split train and test

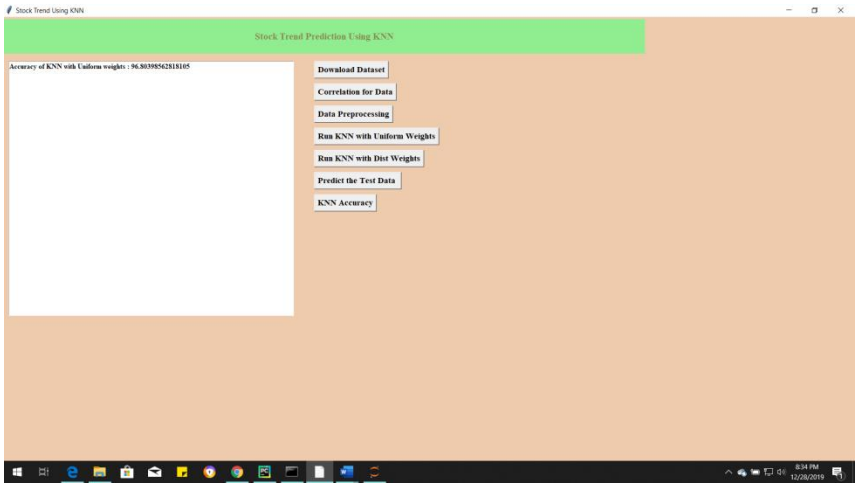


(fig4) data preprocessing

After pre-processing all missing values are dropped, Separating the label here, Scalling of X, find Data Series of late X and early X (train) for model generation and evaluation, Separate label and identify it as y and Separation of training and testing of model.

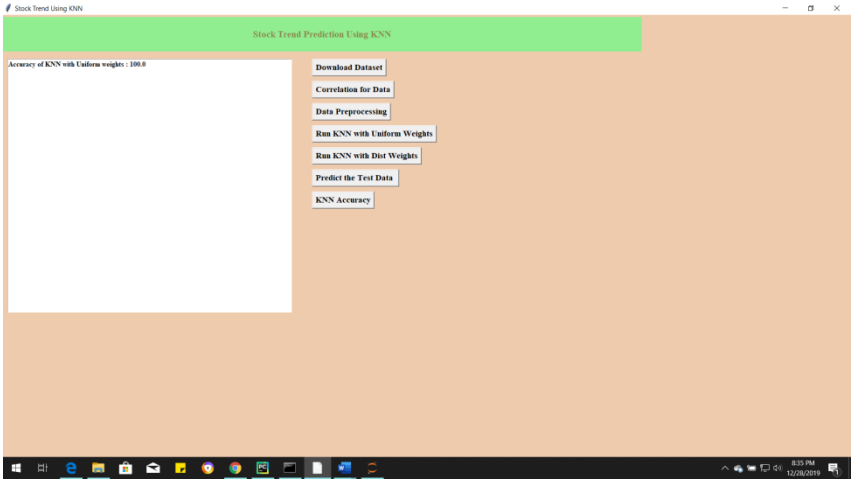
In above screen we can see dataset contains total 1752 records and 1226 used for training and 526 used for testing.

Now click on 'Run KNN with Uniform Weights' to generate KNN model with uniform weights and calculate its model accuracy



(fig5) run KNN with uniform weights

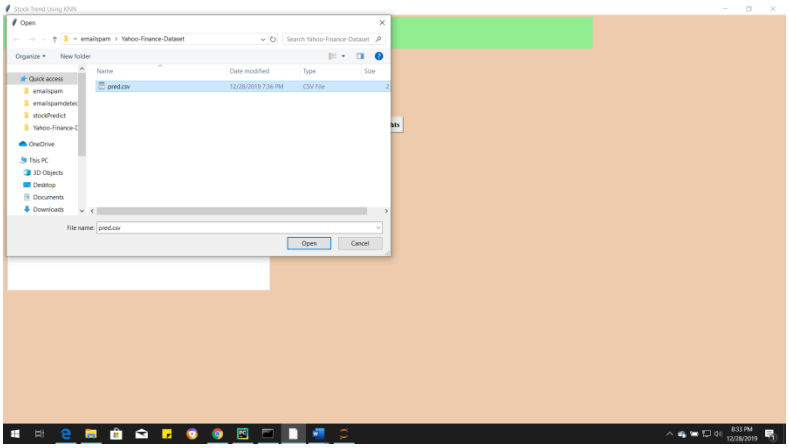
In above screen we can see with KNN with uniform weights got 96.8% accuracy, now click on 'Run KNN with distance weights' to calculate accuracy



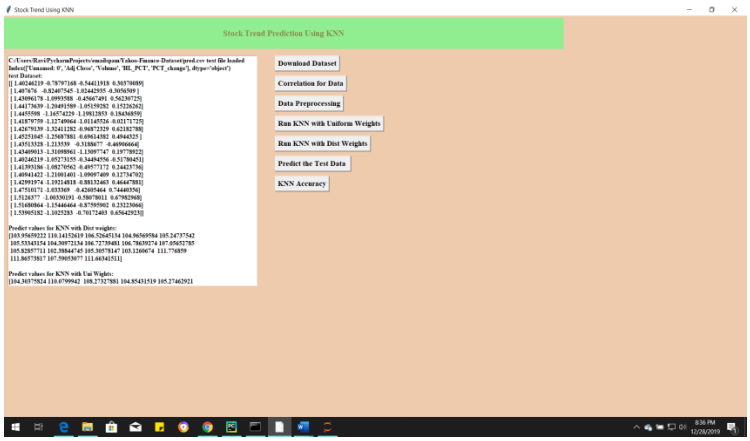
(fig6) run KNN with distance weights

In above screen we got 100% accuracy, now we will click on 'Predict Test Data'button to upload test data and to predict whether test data stock market for both models.

accuracy score (>0.95) for most of the models.

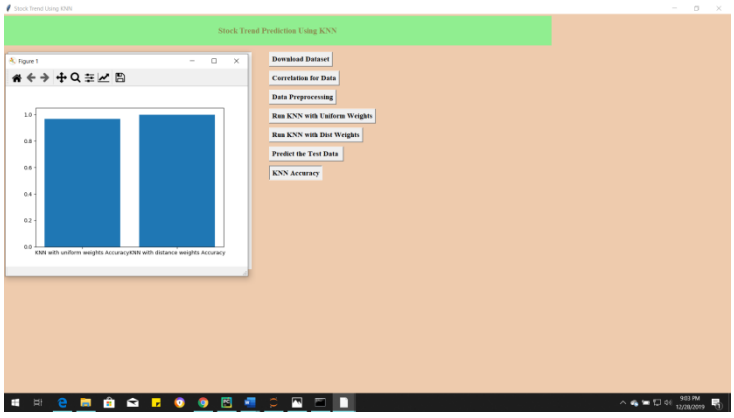


(fig7) test data upload



(fig8) predicted prices

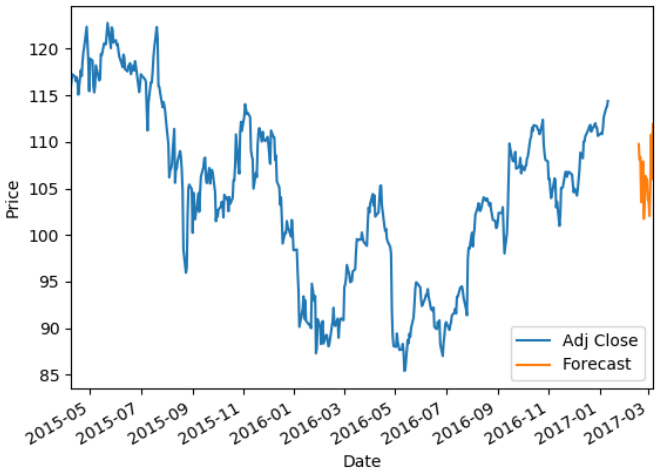
In above screen for each test data we got forecast values for Apple Stock for each test record. Now click on 'KNN Accuracy' button to save the predicted values for each model save in the local directory and Accuracy comparison to both the models



(fig9) KNN accuracy

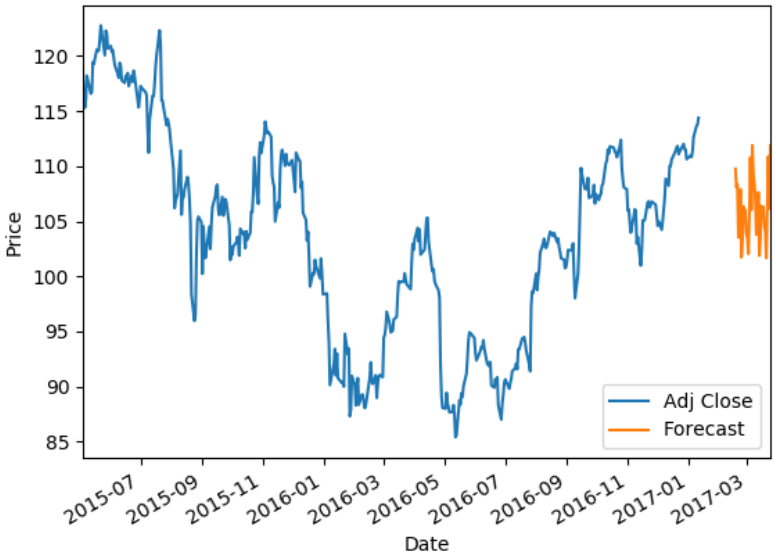
From above graph we can see that distance weights has little bit better better accuracy compare to Uniform weights, in above graph x-axis contains algorithm name and y-axis represents accuracy of that algorithms

Plotting the Prediction for KNN with Uniform Weights:



(fig10)

Plotting the Prediction for KNN with Distance Weights:



(fig11)

9. EXPERIMENTAL RESULTS

The proposed model was tested and compared with four other standard algorithms, including KNN, Naïve Bayes, OneR and ZeroR. The test examined how accurate the tested algorithms predict the stock price trends, and evaluated the MAE and RMSE. Table 5 presents the test results. The hybrid KNN-Probabilistic model has allowed us to achieve an estimated accuracy of 89.1725%, exceeding the stand alone KNN reported accuracy of 86.6667% and the Naive Bayes accuracy of 76.1194%. The accuracy rates for OneR and ZeroR classifiers were 71.6418% and 64.1791% respectively. KNN-Probabilistic model has MAE rate of 0.0667% and RMSE rate of 0.2582% which are much lower than the other classifiers.

Comparison of results with each algorithm

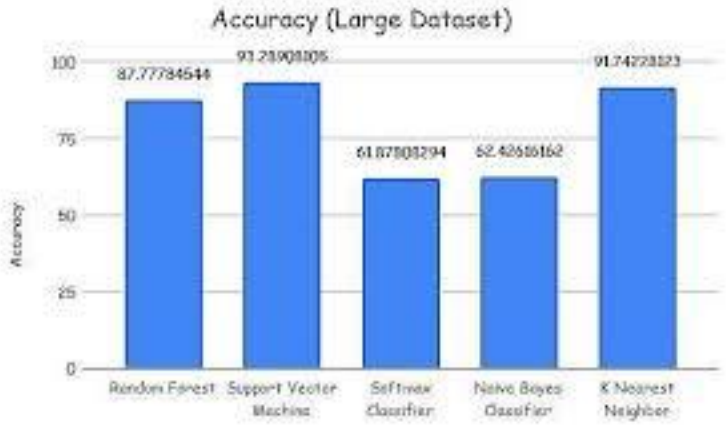


Fig 12 Accuracy of algorithms graph

Overall, KNN-Probabilistic model has better accuracy rate and error rates than the other classifiers used for comparisons. The test demonstrated that the hybrid mechanism of KNN and probabilistic method produced significantly improved results, compared with each of the KNN and Naïve Bayes classifiers

Here are some screenshots of the results obtained:

Classifier	Accuracy (%)	MAE	RMSE
KNNProbabilistic	93.3333	0.0667	0.2582
KNN	86.6667	0.1333	0.3651
Naive Bayes	76.1194	0.1726	0.2824
One R	71.6418	0.5325	0.6139
Zero R	64.1791	0.4619	0.4805

Table 1.Comparision of the algorithms

Overall, KNN-Probabilistic model has better accuracy rate and error rates than the other classifiers used for comparisons. The test demonstrated that the hybrid mechanism of KNN and probabilistic method produced significantly improved results, compared with each of the KNN and Naïve Bayes classifier.

10. CONCLUSION AND FUTURE ENHANCEMENT

The aim of this research is to improve the statistical fitness of the proposed model to overcome a KNN problem due to its computation approach. The KNN classifier can compute the empirical distribution over the Profit and Loss class values in the k number of nearest neighbors. However, the outcome is less than adequate due to sparse data. The KNN classifier has under fitting issue as it does not cater to generalization of sparse data outside the range of nearest neighborhood. We have compared a hybrid KNN-Probabilistic model with four standard algorithms on the problem of predicting the stock price trends. Our results showed that the proposed KNN-Probabilistic model leads to significantly better results compared to the standard KNN algorithm and the other classification algorithms. The limitation of the proposed model is that it applies a binary classification technique. The actual output of this binary classification model is a prediction score in twoclass. The score indicates the model's certainty that the given observation belongs to either the Profit class or Loss class.

For future work, the knowledge component is to transform the binary classification into multiclass classification. The multiclass classification involves observation and analysis of more than the existing two statistical class values. Additional research will include the application of the probabilistic model to multiclass data in order to provide more specific information of each class value. The newly formed multiclass classification will contain five class labels named “Sell”, “Underperform”, “Hold”, “Outperform”, and “Buy”. In numerical values for mapping purpose, we will convert “Sell” to -2 which implies strongly unfavorable; “Underperform” to -1 which implies moderately unfavorable; “Hold” to 0 which implies neutral; “Outperform” to 1 which implies moderately favorable; and “Buy” to 2 which implies strongly favourable.

REFERENCES

[1] Kalra S, Prasad JS. Efficacy of News Sentiment for Stock Market Prediction. Proc Int Conf Mach Learn Big Data, Cloud Parallel Comput Trends, Prespectives Prospect Com 2019. Published online 2019:491-496. doi:10.1109/COMITCon.2019.8862265

[2] Menon A, Singh S, Parekh H. A review of stock market prediction using neural networks. 2019 IEEE Int Conf Syst Comput Autom Networking, ICSCAN 2019. Published online 2019:1-6. doi:10.1109/ICSCAN.2019.8878682

[3] Sharma A, Bhuriya D, Singh U. Survey of stock market prediction using machine learning approach. Proc Int Conf Electron Commun Aerosp Technol ICECA 2017. 2017;2017-Janua:506-509. doi:10.1109/ICECA.2017.8212715

[4] Chen MY, Liao CH, Hsieh RP. Modeling public mood and emotion: Stock market trend prediction with anticipatory computing approach. Comput Human Behav. 2019;101(September 2018):402-408. doi:10.1016/j.chb.2019.03.021

[5] Picasso A, Merello S, Ma Y, Oneto L, Cambria E. Technical analysis and sentiment embeddings for market trend prediction. Expert Syst Appl. 2019;135:60-70. doi:10.1016/j.eswa.2019.06.014

[6] Shobha G, Rangaswamy S. Machine Learning. Vol 38. 1st ed. Elsevier B.V.; 2018. doi:10.1016/bs.host.2018.07.004

[7] Basak S, Kar S, Saha S, Khaidem L, Dey SR. Predicting the direction of stock market prices using tree-based classifiers. North Am J Econ Financ. 2019;47(December 2017):552-567. doi:10.1016/j.najef.2018.06.013

[8] Kia AN, Haratizadeh S, Shouraki SB. A hybrid supervised semi-supervised graph-based model to predict one-day ahead movement of global stock markets and commodity prices. Expert Syst Appl. 2018;105:159-173. doi:10.1016/j.eswa.2018.03.037

[9] Henrique BM, Sobreiro VA, Kimura H. Literature review: Machine learning techniques applied to financial market prediction. Expert Syst Appl. 2019;124:226-251. doi:10.1016/j.eswa.2019.01.012

[10] Zhou F, Zhou H min, Yang Z, Yang L. EMD2FNN: A strategy combining empirical mode decomposition and factorization machine based neural network for stock market trend prediction. Expert Syst Appl. 2019;115:136-151. doi:10.1016/j.eswa.2018.07.065

[11] Sirimevan N, Mamalgaha IGUH, Jayasekara C, Mayuran YS, Jayawardena C. Stock Market Prediction Using Machine Learning Techniques. 2019 Int Conf Adv Comput ICAC 2019. 2019;(4):192-197. doi:10.1109/ICAC49085.2019.9103381

[12] Jadhav AA, Biradar N, Bhaldar H, Mathpati MS, Wadekar R, Scholar R. International Journal of Innovative Research in Computer and Communication Engineering Design and Analysis of Triple Band Miniaturized Antenna for Wearable Application. 2019;(March). doi:10.15680/IJIRCCE.2019

[13] Nayak A, Pai MMM, Pai RM. Prediction Models for Indian Stock Market. Procedia Comput Sci. 2016;89:441- 449. doi:10.1016/j.procs.2016.06.096

[14] Oliveira N, Cortez P, Areal N. The impact of microblogging data for stock market prediction: Using Twitter to predict returns, volatility, trading volume and survey sentiment indices. Expert System appl.

PUBLICATIONS

Stock Market Analysis Using KNN Algorithm

Varun Aditya[1], Raghavender Reddy[2], Harsha Vardhan[3],

Mr J. Manikandan[5],

[1234]UG Scholar, [5]Professor, [6]Assistant Professor

Department of Computer Science and Engineering

St. Martin's Engineering College, Secunderabad – 500 100, India

E-Mail: [1]varun28baindla@gmail.com , [2]raghavender2348@gmail.com , hharsha8581@mail.com , manikandancse@smec.ac.in .

Abstract:

This paper examines a hybrid model which combines a K-Nearest Neighbours (KNN) approach with a probabilistic method for the prediction of stock price trends. One of the main problems of KNN classification is the assumptions implied by distance functions. Theassumptions focus on the nearest neighbours which are at the centroid of data points for test instances. This approach excludes the non-centric data pointswhich can be statistically significant in the problem of predicting the stock price trends. For this it is necessary to construct an enhanced model that integrates KNN with a probabilistic method which utilizes both centricand non-centric data points in the computations of probabilities for the target instances. The embedded probabilistic method is derived from Bayes' theorem. The prediction outcome is based on a joint probability where the likelihood of the event of the nearest neighbours and the event of prior probability occurring together and at the same point in time where they are calculated. The proposed hybrid KNN Probabilistic model was compared with the standard classifiers that include KNN, Naive Bayes, One Rule (One-R) and Zero Rule(ZeroR). The test results showed that the proposed model outperformed the standard classifiers which were used for the comparisons.

Keywords:

Stock Price Prediction, K-Nearest Neighbours, Bayes' Theorem, Naive Bayes, Probabilistic method.

I INTRODUCTION

Analyzing financial data in securities has been an important and challenging issue in the investment community. Stock price efficiency for public listed firms is difficult to achieve due to the opposing effects of information competition among major investors and the adverse selection costs imposed by their information advantage.

There are two main schools of thought in analyzing the financial markets. The first approach is known as fundamental analysis. The methodology used in fundamental analysis evaluates a stock by measuring its intrinsic value through qualitative and quantitative2 analysis.

This approach examines a company's financial reports, management, industry, micro and macro-economic factors. The second approach is known as technical analysis. The methodology used in technical analysis for forecasting the direction of prices is through the study of historical market data. Technical analysis uses a variety of charts to anticipate what are likely to happen. The stock charts include candlestick charts, line charts, bar charts, point and figure charts, OHLC (open-high-low-close) charts and mountain charts. The charts are viewable in different time frames with price and volume. There are many types of indicators used in the charts, including resistance, support, breakout, trending and momentum.

Several alternatives to approach this type of problem have been proposed, which range from traditional statistical modelling to methods based on computational intelligence and machine learning. Vanstone and Tan surveyed the works in the domain of applying soft computing to financial trading and investment. They categorized the papers reviewed in the following areas: time series, optimization, hybrid methods, pattern recognition and classification. Within the context of financial trading discipline, the survey showed that most of the research was being conducted in the field of technical analysis. An integrated fundamental and technical analysis model was examined to evaluate the stock price trends by focusing on macro-economic analysis. It also analyzed the company behaviour and the associated industry in relation to the economy which in turn provide more information for investors in their investment decisions.

A nearest neighbor search (NNS) method produced an intended result by the use of KNN technique with technical analysis. This model applied technical analysis on stock market data which include historicalprice and trading volume. It applied technical indicators made up of stop loss, stop gain and RSI filters. The KNN algorithm part applied the distance function on the collected data. This model was compared with the buy-and-hold strategy by using the fundamental analysis approach.

Fast Library for Approximate Nearest Neighbors (FLANN) is used to perform the searches for choosing the best algorithm found to work best among a collection of algorithms in its library. Majhi etal. examined the FLANN model to predict the S&P 500 indices, and the FLANN model was established by performing fast approximate nearest neighbor searches in high dimensional spaces.

II LITERATURE SURVEY

Financial services companies are developing their products to serve future prediction. There are a large amount of financial information sources in the world that can be valuable research areas, one of these areas is stock prediction and also called stock market mining. Stock prediction becomes increasingly important especially if number of rules could be created to help making better investment decisions in different stock markets

Sneh Kalra et al. in 2019, in this paper authors, did research on the fluctuation of stock market prices with respect to the relevant new articles of a company. They used classifier Naïve Bayes to separate negative or positive statements for prediction purposes based on daily news variance the social media data, blogs data may be considered for future work [1].

Aditya Menon et al. in 2019, this paper is focused on a review of neural model for forecast the stock tread after reviewing on a neural model they think that The long short term memory algorithm for predicting the economic information in confluence into the trendy era, this would be prioritized algorithm for forecasting [2].

Andrea Picasso et al. in 2019, in this research, authors worked which will alliance the economic and elemental analysis for market trend prediction through the various kind of application and automation methods neural network is machine learning technique the problem of trend stock and those are charts with forecasting data. As an input data sentiment of a news article is exploited. According to their research the problem in the most problematic accomplishment among the use of information about news astral one-off. To overcome this problem in the future the proper feature fusion technique will be suitable for the future [3].

Gangadhar Shobha et al. in 2018, this paper provided a full overview of machine learning techniques which will help to reader for use of equations and concept the author discussed about three type of all machine learning technique and also various kind of metrics like accuracy, confusion matrix, recall, RMSE, precision and quintile of errors. The author thinks that this review can help those people who are new to machine learning because most of the people confuse to use most of the machine learning techniques for prediction or others [4].

Arash Negahdari kia et al. in 2018, as the stock prediction so many experiments and models, have been developed for prediction purpose on historical data like as in this paper the author present HyS3 graph-based semi-supervised model and through a network views Kruskal based graph algorithm called ConKruG. In the future they think social media data, Twitter data could be used for the prediction of stock for better results using these algorithms [5]

III SYSTEM STUDY

Predicting the Stock Market has been the bane and goal of investors since its existence. Everyday billions of dollars are traded on the exchange, and behind each dollar is an investor hoping to profit in one way or another. Entire companies rise and fall daily based on the behaviour of the market. Should an investor be able to accurately predict market movements, it offers a tantalizing promises of wealth and influence. It is no wonder then that the Stock Market and its associated challenges find their way into the public imagination every time it misbehaves. The 2008 financial crisis was no different, as evidenced by the flood of films and documentaries based on the crash. If there was a common theme among those productions, it was that few people knew how the market worked or reacted. Perhaps a better understanding of stock market prediction might help in the case of similar events in the future.

Despite its prevalence, Stock Market prediction remains a secretive and empirical art. Few people, if any, are willing to share what successful strategies they have. A chief goal of this project is to add to the academic understanding of stock market prediction. The hope is that with a greater understanding of how the market moves, investors will be better equipped to prevent another financial crisis. The project will evaluate some existing strategies from a rigorous scientific perspective and provide a quantitative evaluation of new strategies.

SYSTEM ARCHITECTURE

Here is the proposed work system architecture, which we have depicted through the stepwise structure, In the first step we are starting by giving raw data to our trained algorithm which will pre-process the data by using python libraries which is also a feature extraction part, Where it will the cleanup data by using data pre-processing method after that we have divided our data into two parts where 70% of our data is trained and remaining 30% of data which is for testing by using the trained algorithm after all this process we will only get our predicted data, as shown in figure 1.

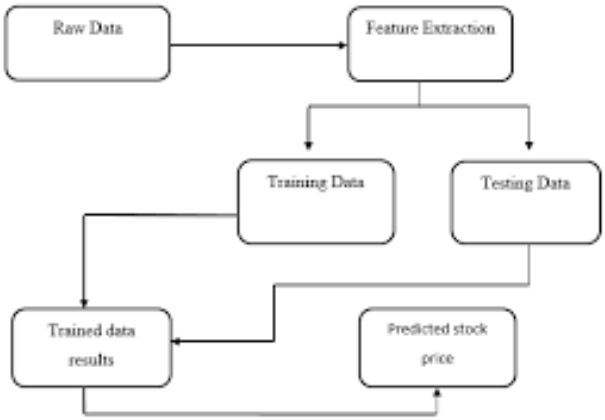


Fig 13 System architecture

Data Preparation Raw data is at high risk for noise, lost values, and inconsistencies. Data quality affects the results of data mining. To help enhance the affection of information or, as a result, which is the output

of extraction the unstructured information returns in advance to enhance the effectiveness or simplification of extraction operations. This is the only step that is very tough in the data extraction that negotiates and tries to rebuild and alteration of the original information.

Data integration The chance for your information or data anatomizing function would include in data integration, that means the information will alliance from many cradles into a compatible stash, such information in the stash. Which can combine a huge amount of information, data files, and more in some files. So we can get many hurdles at the time of information compilation. The combination of tricks could deceptive. Whereas the other word companies which could be matched from many sources of data. Which is basically called business index the causes. Such as, a computer engineer or analyst cannot be sure about the person's information details in one database, as well as the customer number for the same business as a refers. Huge information and data repositories usually contain data at a very huge level which is called metadata. Which we say that data is about data, which normally can help to obtain or abate the mistakes in integration process retrenchment which is also trouble the adjective which should not any requirement, therefore, that is not based on the \\ " other table, such as annual income"

Data transformation In data conversion, data is converted or compiled into appropriate mining forms. Data modification may include some points such as: 1. Formalize, the affection of measured details that can be depicted in the range of 0 to 1. 2. Self-composed, unwanted information or can say noise that casually contains the data. These processes having bent and turning of information. 3. Agregassion, where summarizing or collating data is used. The information which deals on the basis of daily could be compiled or calculate the total income of annually and monthly subtotal. Which is commonly maintained to build an information set of to the anatomizing whole collective grained. 4. The data which is generalized, where zero level and 'old' (green) information which cut out with the high place of information by the exploit of conceptual ordering. Such as, a separate feature which is arterial might be tailored to other ordering abstraction, such as a rural region same as, number values, or age, may be included in the map of high-level concepts, such as young, middle-aged, and older.

Removing extra values or cleaning of data The dataset which is basically anatomized by the extraction method which is not completed by the process (deficit of adjective which is specific interest affection which contains aggregated into), bodacious/ noisy (means that which have error values and external data which is exactly different what we expect). Capricious not compatible with another fact (having inconsistency into code area which basically used to clarify dataset). The noisy changeable which are at the same area of huge existing in the real-world source of data or its repositories, the occurring of incomplete information have many reasons. Attributes in concern might not eternally operable. As the data of transaction which depends on customer information, some information cannot be encaged where it may be advised arrogant during login. Appropriate information cannot be recorded due to misunderstandings, or due to mechanical malfunction. Data that did not match any other recorded data may have been deleted. In addition, historical recording or modification of data may not be considered. Missing data, especially duplicates with missing values some of the symbols, might be entered. Given information might be arresting, with not corrective liability set of data, due to a given array of dataset tools given could improper. That could error of people and machine automation arising from the presence of information. Inaccuracy in information transfer might be raise. That is technical boundaries, likewise bounded buffer assize in linking and confide or syncing. The information which is not correct might have appeared from the changeable name or meetings and use of information canon. Coequal requires information is to be fine. The information purification methods beaver "clean up" information through filling with lost values, easily out sensitive information, recognizable and deleting sales, Disassociate capriciously. Noisy information might have some disarrangement about the extraction process. Albeit maximum extraction methods which are having some way of negotiation with imperfections and

worthless information which is not every time solid. On the other hand, they can focus to abate adding enough information there is a task at hand. Accordingly, a practical preprocessing way to use own information in other ways to clean up data.

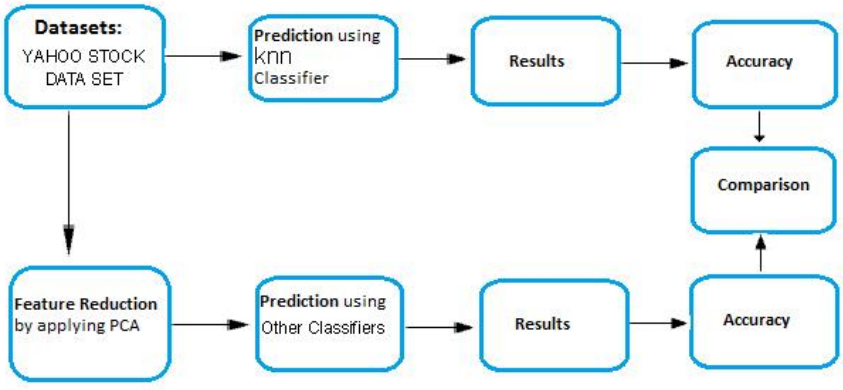


Fig. 1. Block diagram of the proposed research approach

Fig14 KNN architecture

IV EXPERIMENTAL RESULTS

In this experiment we implement the four supervised machine learning algorithms, name as linear regression, support vector machine, random forest, and k-nearest neighbor using these algorithms we predict the stock market trends.

To conduct experiment author has used Yahoo Finance stock Dataset and below is some example records of that dataset which contains request signatures. I have also used same dataset and this dataset is available inside ‘dataset’ folder.

Dataset example

01-05-2017	116.860001	115.809998	116.610001	22193600.0	111.393003	111.393303
01-092017	119.430000	117.940002	117.949997	118.989998	118.989998	113.666824

['High', 'Low', 'Open', 'Close', 'Volume', 'Adj Close']

Above list are the columns of Yahoo finance

Above two records are the APPLE stock form the Yahoo Finance Dataset. For the rest of analysis, we will use the Closing Price which remarks the final price in which the stocks are traded by the end of the day.we analyse stocks using two key measurements: Rolling Mean and Return Rate.

Double click on ‘run.bat’ file to get below screen

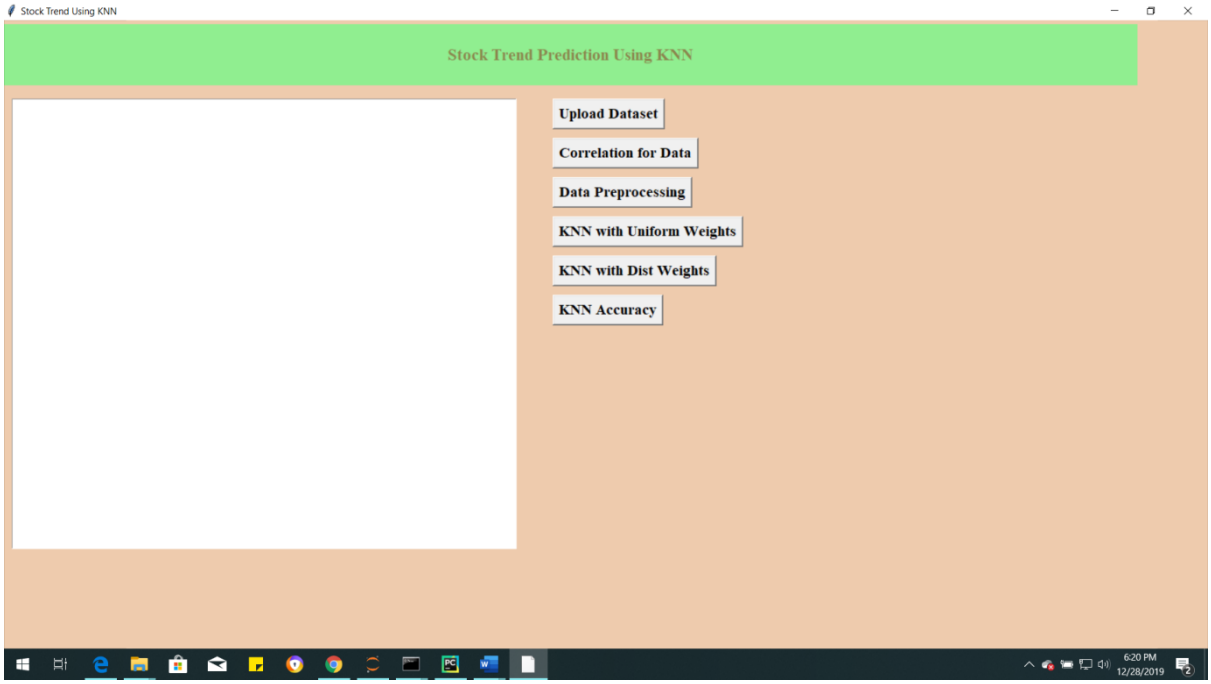


Fig 15 load dataset

In above screen click on ‘Download Button’ download the Apple Stock and competitors data from Yahoo Finance Dataset

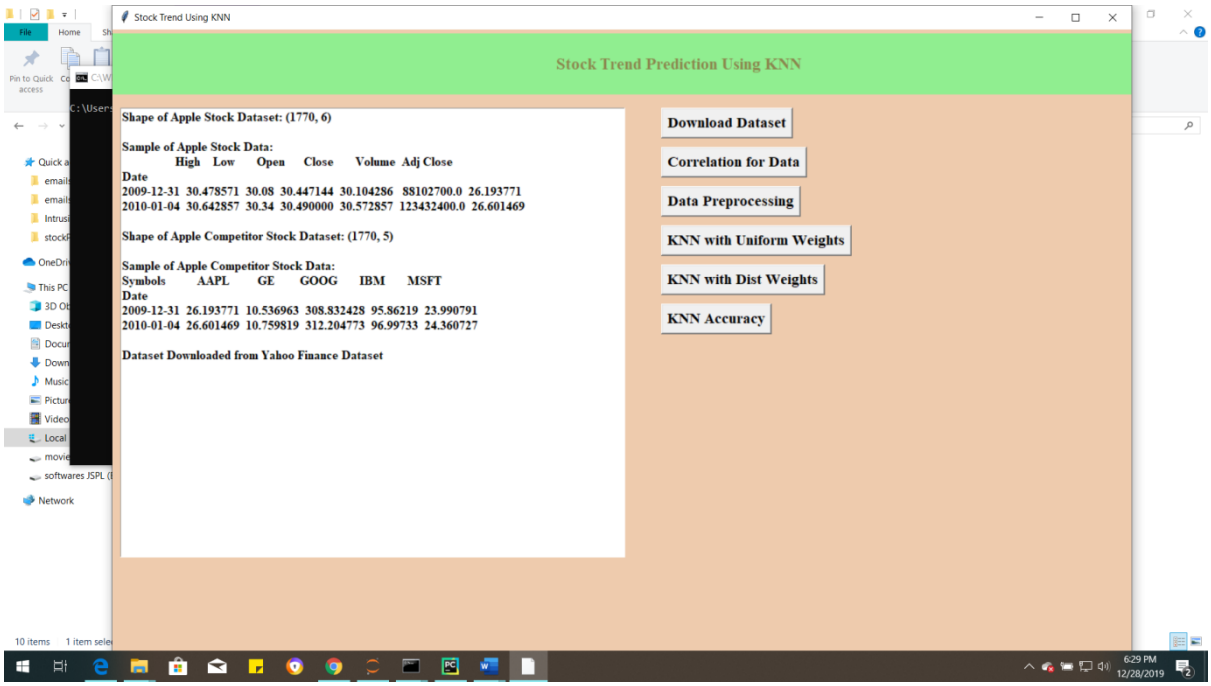


Fig 16 data set downloaded

In above screen I am Downloading of Apple Stock and Apple competitor Stock Data from Yahoo Finance Dataset.

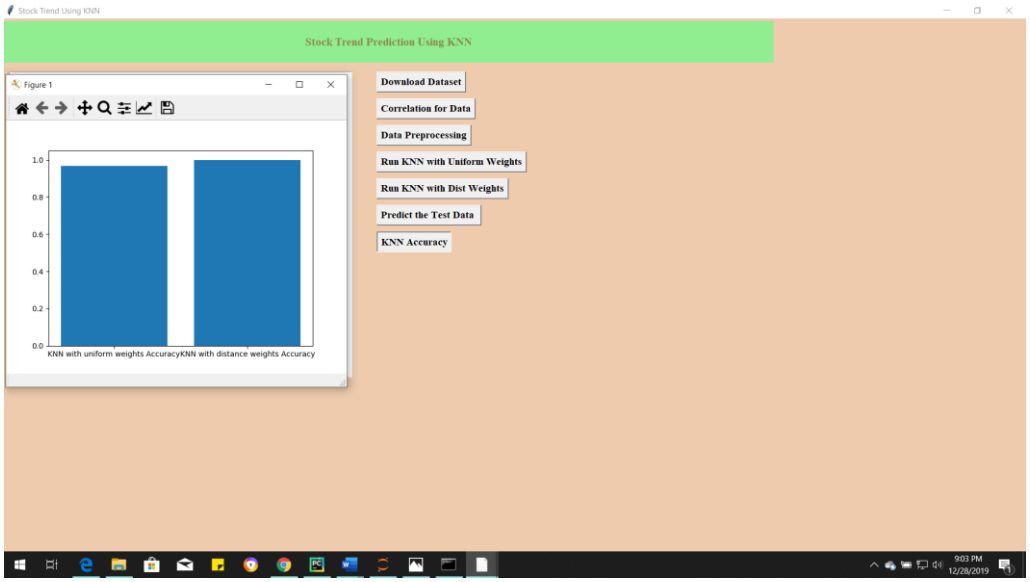


Fig 17 accuracy of KNN

From above graph we can see that distance weights has little bit better better accuracy compare to Uniform weights, in above graph x-axis contains algorithm name and y-axis represents accuracy of that algorithms.

V. RESULTS

The proposed model was tested and compared with four other standard algorithms, including KNN, Naïve Bayes, OneR and ZeroR. The test examined how accurate the tested algorithms predict the stock price trends, and evaluated the MAE and RMSE. Table 5 presents the test results. The hybrid KNN-Probabilistic model has allowed us to achieve an estimated accuracy of 89.1725%, exceeding the stand alone KNN reported accuracy of 86.6667% and the Naive Bayes accuracy of 76.1194%. The accuracy rates for OneR and ZeroR classifiers were 71.6418% and 64.1791% respectively. KNN-Probabilistic model has MAE rate of 0.0667% and RMSE rate of 0.2582% which are much lower than the other classifiers.

6.1 Comparison of results with each algorithm

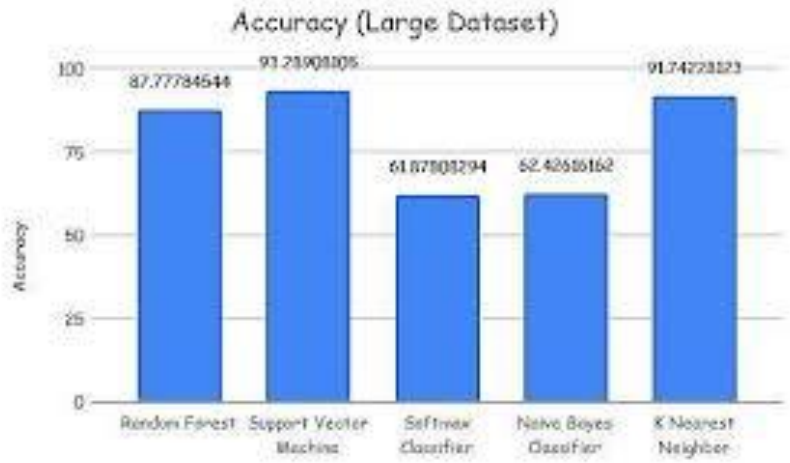


Fig 18 accuracy of algorithms graph

Overall, KNN-Probabilistic model has better accuracy rate and error rates than the other classifiers used for comparisons. The test demonstrated that the hybrid mechanism of KNN and probabilistic method produced significantly improved results, compared with each of the KNN and Naïve Bayes classifiers

Here are some screenshots of the results obtained:

Classifier	**Accuracy (%)**	**MAE**	**RMSE**
KNNProbabilistic	93.3333	0.0667	0.2582
KNN	86.6667	0.1333	0.3651
Naive Bayes	76.1194	0.1726	0.2824
One R	71.6418	0.5325	0.6139
Zero R	64.1791	0.4619	0.4805

Table 2 comparision of algorithms

Overall, KNN-Probabilistic model has better accuracy rate and error rates than the other classifiers used for comparisons. The test demonstrated that the hybrid mechanism of KNN and probabilistic method produced significantly improved results, compared with each of the KNN and Naïve Bayes classifiers.

VI. CONCLUSION AND FUTURE WORK

The aim of this research is to improve the statistical fitness of the proposed model to overcome a KNN problem due to its computation approach. The KNN classifier can compute the empirical distribution over the Profit and Loss class values in the k number of nearest neighbors. However, the outcome is less than adequate due to sparse data. The KNN classifier has under fitting issue as it does not cater to generalization of sparse data outside the range of nearest neighborhood. We have compared a hybrid KNN-Probabilistic model with four standard algorithms on the problem of predicting the stock price trends. Our results showed that the proposed KNN-Probabilistic model leads to significantly better results compared to the standard KNN algorithm and the other classification algorithms. The limitation of the proposed model is that it applies a binary classification technique. The actual output of this binary classification model is a prediction score in twoclass. The score indicates the model's certainty that the given observation belongs to either the Profit class or Loss class.

For future work, the knowledge component is to transform the binary classification into multiclass classification. The multiclass classification involves observation and analysis of more than the existing two statistical class values. Additional research will include the application of the probabilistic model to multiclass data in order to provide more specific information of each class value. The newly formed multiclass classification will contain five class labels named “Sell”, “Underperform”, “Hold”, “Outperform”, and “Buy”. In numerical values for mapping purpose, we will convert “Sell” to -2 which implies strongly unfavorable; “Underperform” to -1 which implies moderately unfavorable; “Hold” to 0 which implies neutral; “Outperform” to 1 which implies moderately favorable; and “Buy” to 2 which implies strongly favourable

VII REFERNCES

[1] Kalra S, Prasad JS. Efficacy of News Sentiment for Stock Market Prediction. Proc Int Conf Mach Learn Big Data, Cloud Parallel Comput Trends, Prespectives Prospect Com 2019. Published online 2019:491-496. doi:10.1109/COMITCon.2019.8862265

[2] Menon A, Singh S, Parekh H. A review of stock market prediction using neural networks. 2019 IEEE Int Conf Syst Comput Autom Networking, ICSCAN 2019. Published online 2019:1-6. doi:10.1109/ICSCAN.2019.8878682

[3] Sharma A, Bhuriya D, Singh U. Survey of stock market prediction using machine learning approach. Proc Int Conf Electron Commun Aerosp Technol ICECA 2017. 2017;2017-Janua:506-509. doi:10.1109/ICECA.2017.8212715

[4] Chen MY, Liao CH, Hsieh RP. Modeling public mood and emotion: Stock market trend prediction with anticipatory computing approach. Comput Human Behav. 2019;101(September 2018):402-408. doi:10.10 16/j.chb.2019.03.021

[5] Picasso A, Merello S, Ma Y, Oneto L, Cambria E. Technical analysis and sentiment embeddings for market trend prediction. Expert Syst Appl. 2019;135:60-70. doi:10.1016/j.eswa.2019.06.014

[6] Shobha G, Rangaswamy S. Machine Learning. Vol 38. 1st ed. Elsevier B.V.; 2018. doi:10.1016/bs.host.2018.07.004

ONE PAGE PROFILES

Varun Aditya Baindla is pursuing his Bachelor of Technology in the stream of Computer science and engineering at St. Martin's Engineering College. He completed his intermediate from Sri Gayatri Educational Institutions and completed his schooling from C.M.R Model High School. His responsibilities in that group include mentoring and motivating students to take coding as a serious hobby. His participations include National Level Three Day Online Workshop on “AI & ML in speech and audio processing” which was conducted from 10th and 12th December, 2020, Leadership Talk with Mr.Mahesh Babu CEO Mahindra Electric Mobility Ltd. His technical skills include C, C++, Java, Python, HTML and CSS. His areas of interest are Python, Artificial Intelligence, Machine Learning and Deep Learning. He completed 2 one month internships at Lasya Infotech Pvt.Ltd. During the period, he has successfully completed Mini Project and Major Project, at our Development Center. He has completed few certificate courses from online platforms like Coursera on Managing project risk and changes, Data Science math skills, Matrix algebra for engineers, Leadership and emotional intelligence and AWS fundamentals.

Dadi Raghavender Reddy is currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. She completed her intermediate from Little Flower Junior College and 10^{th} class from Sri Chaitanya School. His participations include National Level Three Day Online Workshop on "AI & ML in speech and audio processing" which was conducted from 10^{th} and 12^{th} December, 2020, Leadership Talk with Mr Mahesh Babu CEO Mahindra Electric Mobility Ltd. His technical skills include C, Python and Java. She also has a basic understanding of C++. His areas of interest are Python, Data science and Machine Learning. He completed 2 one month internships at Lasya Infotech Pvt.Ltd. During the period, he has successfully completed Mini Project and Major Project, at our Development Center.

Harsha Vardhan Kajeepuram is currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. She completed his intermediate from Sri Chaitanya Junior College and 10th class from Bashyam brooks School. His participations include National Level Three Day Online Workshop on "AI & ML in speech and audio processing" which was conducted from 10th and 12th December, 2020, Leadership Talk with Mr Mahesh Babu CEO Mahindra Electric Mobility Ltd. His technical skills include C, Python and Java. She also has a basic understanding of C++. His areas of interest are Python, Data science and Machine Learning. He completed 2 one month internships at Lasya Infotech Pvt.Ltd. During the period, he has successfully completed Mini Project and Major Project, at our Development Center.

APPENDICES

\

```
from tkinter import messagebox
from tkinter import *
from tkinter import simpledialog
import tkinter
from tkinter import filedialog
from imutils import paths
from tkinter.filedialog import askopenfilename

import pandas as pd
import datetime
import pandas_datareader.data as web
from pandas import Series, DataFrame
import matplotlib.pyplot as plt
from matplotlib import style
import matplotlib as mpl
from matplotlib import cm as cm
import math
import numpy as np
from sklearn import preprocessing
from sklearn.model_selection import train_test_split
from sklearn.neighbors import KNeighborsRegressor
import seaborn as sns



main = tkinter.Tk()
main.title("Stock Trend Using KNN")
main.geometry("1300x1200")
```

```
global dataFrame, dfreg
global moving_avg
global dfcomp
global clfknn
global clfknndist
global X, y, X_train, y_train, X_test, y_test,X_pred
global distknn, uniknn, knnunipred, knndistpred

def loadDataset():
    text.delete('1.0', END)
    global dataFrame
    global dfcomp
    start = datetime.datetime(2010, 1, 1)
    end = datetime.datetime(2017, 1, 11)

    dataFrame = web.DataReader("AAPL", 'yahoo', start, end)

    text.insert(END, "Shape of Apple Stock Dataset: "+str(dataFrame.shape)+"\n\n")
    text.insert(END, "Sample of Apple Stock Data: \n"+str(dataFrame.head(2))+"\n\n")

    dfcomp = web.DataReader(['AAPL', 'GE', 'GOOG', 'IBM', 'MSFT'], 'yahoo', start=start, end=end)['Adj
Close']
    text.insert(END, "Shape of Apple Competitor Stock Dataset: " + str(dfcomp.shape) + "\n\n")
    text.insert(END, "Sample of Apple Competitor Stock Data: \n" + str(dfcomp.head(2)) + "\n\n")

    text.insert(END, "Dataset Downloaded from Yahoo Finance Dataset\n\n")

def dfcorr():
    text.delete('1.0', END)
    global dfcomp
```

```
    text.insert(END, "Correlation form Apple Competitor Stock\n\n")
    retscomp = dfcomp.pct_change()
    corr = retscomp.corr()
    text.insert(END, "correlation: \n"+str(corr)+"\n\n")


def dataPreProcess():
    text.delete('1.0', END)
    global dataFrame,dfreg
    global X, y, X_train, X_test, y_train, y_test,X_pred

    text.insert(END,"Data PreProcessing for Apple Stock Dataset\n\n")
    dfreg = dataFrame.loc[:,["Adj Close","Volume"]]
    dfreg["HL_PCT"] = (dataFrame["High"] - dataFrame["Low"]) / dataFrame["Close"] * 100.0
    dfreg["PCT_change"] = (dataFrame["Close"] - dataFrame["Open"]) / dataFrame["Open"] * 100.0

    # Drop missing value
    dfreg.fillna(value=-99999, inplace=True)
    # We want to separate 1 percent of the data to forecast
    forecast_out = int(math.ceil(0.01 * len(dfreg)))

    # Separating the label here, we want to predict the AdjClose
    forecast_col = 'Adj Close'
    dfreg['label'] = dfreg[forecast_col].shift(-forecast_out)
    X = np.array(dfreg.drop(['label'], 1))

    # Scale the X so that everyone can have the same distribution for linear regression
    X = preprocessing.scale(X)
    # Finally We want to find Data Series of late X and early X (train) for model generation and evaluation
    X_pred = X[-forecast_out:]
    X = X[:-forecast_out]
```

```
    # Separate label and identify it as y
    y = np.array(dfreg['label'])
    y = y[:-forecast_out]

    text.insert(END, "X lablels : \n"+str(X)+"\n\n")
    text.insert(END, "Y lablels : \n"+str(y)+"\n\n")
    text.insert(END, "Data spliting into Train and Test")
    X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.3)

    text.insert(END, "number of Train Samples : " + str(len(X_train)) + "\n")
    text.insert(END, "number of Test Sample: " + str(len(X_test)) + "\n")

    text.insert(END, "Data Preprocessing Completed\n\n")

def uniformKNN():
    text.delete('1.0',END)
    global clfknn
    global uniknn

    # KNN Regression
    clfknn = KNeighborsRegressor(n_neighbors=5)
    clfknn.fit(X_train, y_train)

    uniknn = clfknn.score(X_train, y_train)

    text.insert(END, "Accuracy of KNN with Uniform weights : "+str(uniknn*100)+"\n\n")

def distKNN():
    text.delete('1.0', END)
    global clfknndist,knndistpred
    global distknn,knnunipred
```

```
    # KNN Regression
    clfkndist = KNeighborsRegressor(n_neighbors=5,weights='distance')
    clfkndist.fit(X_train, y_train)
    distknn = clfkndist.score(X_train, y_train)

    text.insert(END, "Accuracy of KNN with Uniform weights : "+str(distknn*100)+"\n\n")
def predModel():
    text.delete('1.0', END)
    global clfkndist,clfknn,knnunipred,kndistpred
    global X, y, X_train, X_test, y_train, y_test


    filename = filedialog.askopenfilename(initialdir="Yahoo-Finance-Dataset")
    test = pd.read_csv(filename)
    text.insert(END, filename + " test file loaded\n"+str(test.columns)+"\n");
    x_pred = np.array(test.drop(['Unnamed: 0'],1))

    text.insert(END, "test Dataset: \n"+str(x_pred)+"\n\n");

    kndistpred = clfkndist.predict(x_pred)

    text.insert(END, "Predict values for KNN with Dist weights: \n" + str(kndistpred) + "\n\n");

    knnunipred = clfknn.predict(x_pred)

    text.insert(END, "Predict values for KNN with Uni Wights: \n" + str(knnunipred) + "\n\n");

def graph():
    text.delete('1.0', END)

    global uniknn,distknn
```

```
global knnunipred,knndistpred
global dfreg

dfreg['Forecast'] = np.nan
last_date = dfreg.iloc[-1].name
last_unix = last_date
next_unix = last_unix + datetime.timedelta(days=1)

for i in knnunipred:
    next_date = next_unix
    next_unix += datetime.timedelta(days=1)
    dfreg.loc[next_date] = [np.nan for _ in range(len(dfreg.columns) - 1)] + [i]
dfreg['Adj Close'].tail(500).plot()
dfreg['Forecast'].tail(500).plot()
plt.legend(loc=4)
plt.xlabel('Date')
plt.ylabel('Price')
plt.savefig('knnUniformPredGraph.png')
plt.close()

for i in knndistpred:
    next_date = next_unix
    next_unix += datetime.timedelta(days=1)
    dfreg.loc[next_date] = [np.nan for _ in range(len(dfreg.columns) - 1)] + [i]
dfreg['Adj Close'].tail(500).plot()
dfreg['Forecast'].tail(500).plot()
plt.legend(loc=4)
plt.xlabel('Date')
plt.ylabel('Price')
plt.savefig('knnDistPredGraph.png')
plt.close()
```

```
    height = [uniknn,distknn]
    bars = ('KNN with uniform weights Accuracy', 'KNN with distance weights Accuracy')
    y_pos = np.arange(len(bars))
    plt.bar(y_pos, height)
    plt.xticks(y_pos, bars)
    plt.show()


font = ('arial', 12, 'bold')
title = Label(main, text='Stock Trend Prediction Using KNN')
title.config(bg='palegreen', fg='blue2')
title.config(font=font)
title.config(height=3, width=120)
title.place(x=0,y=5)

font1 = ('arial', 10, 'bold')
uploadButton = Button(main, text="Download Dataset", command=loadDataset)
uploadButton.place(x=700,y=100)
uploadButton.config(font=font1)

corrButton = Button(main, text="Correlation for Data", command=dfcorr)
corrButton.place(x=700,y=150)
corrButton.config(font=font1)

ppButton = Button(main, text="Data Preprocessing", command=dataPreProcess)
ppButton.place(x=700,y=200)
ppButton.config(font=font1)

uniformButton = Button(main, text="Run KNN with Uniform Weights", command=uniformKNN)
uniformButton.place(x=700,y=250)
```

```
uniformButton.config(font=font1)

distButton = Button(main, text="Run KNN with Dist Weights", command=distKNN)
distButton.place(x=700,y=300)
distButton.config(font=font1)

predButton = Button(main, text="Predict the Test Data ", command=predModel)
predButton.place(x=700,y=350)
predButton.config(font=font1)

graphButton = Button(main, text="KNN Accuracy", command=graph)
graphButton.place(x=700,y=400)
graphButton.config(font=font1)

font1 = ('arial', 10, 'bold')
text=Text(main,height=30,width=80)
scroll=Scrollbar(text)
text.configure(yscrollcommand=scroll.set)
text.place(x=10,y=100)
text.config(font=font1)


main.config(bg='peachbuff2')
main.mainloop()
```

A

PROJECT REPORT

On

FAKE IMAGE DETECTION

Submitted by

1)Mr.V.Vineeth Chandra (16K81A05B7)

2)Mr.G.Vijay(16K81A0523)

3)Ms.Ch.Indira Priya Darshini (16K81A0575)

in partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Under the Guidance of

Y.Chandra Mouli

Assistant professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ST.MARTIN'S ENGINEERING COLLEGE
An Autonomous Institute

Dhulapally, Secunderabad – 500 100

JUNE 2021

BONAFIDE CERTIFICATE

This is to certify that the project entitled Fake Image Detection, is being submitted by 1.**Mr.V.VineethChandra(16K81A05B7),2.MR.G.Vijay(16K81A0523),3.Ms.Ch.Indira Priya Darshini (16K81A0575)** in partial fulfillment of the requirement for the award of the degree of **BACHELOR OF TECHNOLOGY IN Computer Science And Engineering** is recorded of bonafide work carried out by them. The result embodied in this report have been verified and found satisfactory.

Y.Chandra Mouli
Department of CSE

Head of the Department
Dr.M.NARAYANAN
Department of CSE

Internal Examiner

External Examiner

Place:

Date:

DECLARATION

We, the student of **Bachelor of Technology** in Department of Computer Science and Engineering', session: 2016 – 2021, St. Martin's Engineering College, Dhulapally, Kompally, Secunderabad, hereby declare that work presented in this Project Work entitled Fake Image Detection is the outcome of our own bonafide work and is correct to the best of our knowledge and this work has been undertaken taking care of Engineering Ethics. This result embodied in this project report has not been submitted in any university for award of any degree.

V.Vineeth Chandra 16J81A05B7

G.Vijay 16K81A0523

Ch.Indira Priya Darshini 16K81A0575

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible and whose encouragements and guidance have crowded effects with success.

We extended our deep sense of gratitude to Principal**, Dr. P. SANTOSH KUMAR PATRA**, St. Martin's Engineering College, Dhulapally, for permitting us to undertake this project.

We are also thankful to **Dr. M.NARAYANAN**, Head of the Department, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

We would like to thank **Dr. T. POONGOTHAI,** Department of Computer Science and Engineering (AI & ML) for her encouragement and insightful comments and as well as our project coordinators **Dr. B.RAJALINGAM**, Associate Professor and **Dr.GOVINDA RAJULU**, Associate Professor, in Department of Computer Science and Engineering for their valuable support.

We would like to express our sincere gratitude and indebtedness to our project supervisor **Y.CHANDRA MOULI**, Assistant Professor, Computer Science and Engineering,St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

Finally, we express thanks to all those who have helped us successfully to completing this project. Furthermore, we would like to thank our family and friends for their moral support and encouragement.

We express thanks to all those who have helped us in successfully completing the project.

V.Vineeth Chandra	16J81A05B7
G.Vijay	16K81A0523
Ch.Indira Priya Darshini	16K81A0575

ABSTRACT

Now-a-days biometric systems are useful in recognizing person's identity but criminals change their appearance in behaviour and psychological to deceive recognition system. To overcome from this problemwe are using new technique called Deep Texture Features extraction from images and then building train machine learning model using CNN (Convolution Neural Networks) algorithm. This technique refer as LBPNet or NLBPNet as this technique heavily dependent on features extraction using LBP (Local Binary Pattern) algorithm. In this project we are designing LBP Based machine learning Convolution Neural Network called LBPNET to detect fake face images. Here first we will extract LBP from images and then train LBP descriptor images with Convolution Neural Network to generate training model. Whenever we upload new test image then that test image will be applied on training model to detect whether test image contains fake image or non-fake image. Below we can see some details on LBP.

TABLEOFCONTENTS

LIST OF FIGURES

LIST OF OUTPUT SCREENS

LIST OF ABBREVIATIONS

<CNN>	Convolutional Nural Network
<UML>	Unified Modeling Language
<RAM>	Random Access Memory
<SSD>	Solid State Drive
<DCNN>	Deep Convolutional Neural Network
<GAN>	Generative Adversarial Net

1.INTRODUCTION

Recently, the generative model based on deep learning such as the generative adversarial net (GAN) is widely used to synthesize the photo-realistic partial or whole content of the image and video. Furthermore, recent research of GANs such as progressive growth of GANs (PGGAN) and Big GAN could be used to synthesize a highly photo-realistic image or video so that the human cannot 20 recognize whether the image is fake or not in the limited time. In general, the generative applications can be used to perform the image translation tasks . However, it may lead to a serious problem once the fake or synthesized image is improperly used on social network or platform. For instance, cycle GAN is used to synthesize the fake face image in a pornography video .Furthermore, GANs may be used to create a speech video with the synthesized facial content of any famous politician, causing severe problems on the society, political, and co mmercial activities. Therefore, an effective fake face image detection technique is desired. In this paper, we have extended our previous study associated with paper ID #1062 to effectively and efficiently address these issues.

In traditional image forgery detection approach, two types of forensics scheme are widely used: active schemes and passive schemes. With the active schemes, the externally additive signal (i.e., watermark) will be embedded in the source image without visual artifacts. In order to identify whether the image has tampered or not, the watermark extraction process will be performed on the target image to restore the watermark[6]. The extracted watermark image can be used to localize or detect the tampered regions in the target image. However, there is no "source image" for the generated images by GANs such that the active image forgery detector cannot be used to extract the watermark image. The second one-passive image forgery detector–uses the statistical information in the source image that will be highly consistency between different images. With this property, the intrinsic statistical informationcan be used to detect the fake region in the image[7][8]. However, the passive image forgery detector cannot be used to identify the fake image

generated by GANs since they are synthesized from the low-dimensional random vector. Nothing change in the generated image by GANs because the fake image is not modified from its original image

1.1PROJECT OVERVIEW

Intuitively, we can adopt the deep neural network to detect the fake image generated by GAN. Recently, there are some studies that investigate a deep learning-based approach for fake image detection in a supervised way. In other words, fake image detection can be treated as a binaryclassification problem (i.e., fake or real image). For example, the convolution neural network (CNN) network is used to learn the fake image detector . In [10], the performance of the fake face image detection can be further improved by adopting the most advanced CNN–Xception network . However, there are many GANs proposed year by year. For example, recently proposed GANs such as [1][12][13][14][15][16][3][2] can be used to produce the photo-realistic images. It is hard and very time-consuming to collect all training samples of all GANs. In addition, such a supervised learning strategy will tend to learn the discriminative features for a fake image generated by each GANs. In this situation, the learned detector may not be effective for the fake image generated by another new GAN excluded in the training phase. Local binary patterns (LBP) is a type of visual descriptor used for classification in computer vision and is a simple yet very efficient texture operator which labels the pixels of an image by thresholding the neighborhood of each pixel and considers the result as a binary number. Due to its discriminative power and computational simplicity, LBP texture operator has become a popular approach in various applications. It can be seen as a unifying approach to the traditionally divergent statistical and structural models of texture analysis. Perhaps the most important property of the LBP operator in real-world applications is its robustness to monotonic gray-scale changes caused, for example, by illumination variations. Another important property is its computational simplicity, which makes it possible to analyze images in challenging real-time settings.

1.2 PROJECT OBJECTIVES

we have proposed a fake feature network based the pairwise learning, to detect the fake face/general images generated by state-of-the-art GANs successfully.

The proposed CFFN can be used to learn the middle- and high-level and discriminative fake feature by aggregating the cross-layer feature representations into the last fully connected layers.

The proposed pairwise learning can be used to improve the performance of fake image detection further.

With the proposed pairwise learning, the proposed fake image detector should be able to have the ability to identify the fake image generated by a new GAN.

Our experimental results demonstrated that the proposed method outperforms other state-of-the-art schemes in terms of precision and recall rate.

1.3 ORGANIZATION OF CHAPTERS

This documentation consists of 10 different chapter and they are:

1. Introduction – This chapter covers the overview of our project and its objectives.
2. Literature Survey – This includes the details of our survey.
3. Software and Hardware Requirements – We specify our software and hardware requirements here.
4. Software Development Analysis – This section includes the problem definition and details of the modules we used in our project.
5. Project System Design – This chapter includes the design part of our project which includes uml diagrams.
6. Project Coding – This section contains the details of our project code.
7. Project Testing – The details of test cases and testing are included in this chapter.

8. Output Screens – This contains the screenshots of how our project looks like when executed.

9. Experimental Results – This chapter contains the screenshots of our results.

10. Conclusion and Future Enhancements – This covers the conclusion of our project and the possible future developments.

2. LITERATURE SURVAY

2.1 SURVEY ON BACKGROUND

1. A Classified Adversarial Network for Multi-Spectral Remote Sensing Image Change Detection

AUTHORS: **Wu, Y.; Bai, Z.; Miao, Q.; Ma, W.; Yang, Y.; Gong**

Adversarial training has demonstrated advanced capabilities for generating image models. In this paper, we propose a deep neural network, named a classified adversarial network (CAN), for multi-spectral image change detection. This network is based on generative adversarial networks (GANs). The generator captures the distribution of the bitemporal multi-spectral image data and transforms it into change detection results, and these change detection results (as the fake data) are input into the discriminator to train the discriminator.

2. A Classified Adversarial Network for Multi-Spectral Remote Sensing Image Change Detection.

AUTHORS: **Yue; Bai, Zhuangfei; Miao, Qiguang; Ma, Wenping; Yang, Yuelei; Gong, Maoguo.**

Person re-identification (re-ID) is a fundamental problem in the field of computer vision. The performance of deep learning-based person re-ID models suffers from a lack of training data. In this work, we introduce a novel image-specific data augmentation method on the feature map level to enforce feature diversity in the network. Furthermore, an attention assignment mechanism is proposed to enforce that the person re-ID classifier focuses on nearly all important regions of the input person image.

3. **Deep Fake Image Detection Based on Pairwise Learning**.

AUTHORS: **Hsu, C.-C.; Zhuang, Y.-X.; Lee, C.-Y.**

Generative adversarial networks (GANs) can be used to generate a photo-realistic image from a low-dimension random noise. Such a synthesized (fake) image with inappropriate

content can be used on social media networks, which can cause severe problems. With the aim to successfully detect fake images, an effective and efficient image forgery detector is necessary. However, conventional image forgery detectors fail to recognize fake images generated by the GAN-based generator since these images are generated and manipulated from the source image. Therefore, in this paper, we propose a deep learning-based approach for detecting the fake images by using the contrastive loss. First, several state-of-the-art GANs are employed to generate the fake–real image pairs. Next, the reduced DenseNet is developed to a two-streamed network structure to allow pairwise information as the input.

4.**Evading deepfakeimage detectors with white-and black-box attacks**.

AUTHORS**: Nicholas Carlini and Hany Farid.**

It is now possible to synthesize highly realistic images of people who don't exist. Such content has, for example, been implicated in the creation of fraudulent social-media profiles responsible for dis-information campaigns. Significant efforts are, therefore, being deployed to detect synthetically-generated content. One popular forensic approach trains a neural network to distinguish real from synthetic content.

5.**Adversarial perturbations fool deepfake detectors**.

AUTHORS**: Apurva Gandhi and Shomik Jain.**

This work uses adversarial perturbations to enhance deepfake images and fool common deepfake detectors. We created adversarial perturbations using the Fast Gradient Sign Method and the Carlini and Wagner L2 norm attack in both blackbox and whitebox settings. Detectors achieved over 95% accuracy on unperturbed deepfakes, but less than 27% accuracy on perturbed deepfakes. We also explore two improvements to deepfake detectors: (i) Lipschitz regularization, and (ii) Deep Image Prior (DIP). Lipschitz

regularization constrains the gradient of the detector with respect to the input in order to increase robustness to input perturbations.

6. **Fake Faces Identification via Convolutional Neural Network.**

AUTHORS**: Huaxiao Mo, B.C.; Luo, W.**

Generative Adversarial Network (GAN) is a prominent generative model that are widely used in various applications. Recent studies have indicated that it is possible to obtain fake face images with a high visual quality based on this novel model. If those fake faces are abused in image tampering, it would cause some potential moral, ethical and legal problems. In this paper, therefore, we first propose a Convolutional Neural Network (CNN) based method to identify fake face images generated by the current best method [20], and provide experimental evidences to show that the proposed method can achieve satisfactory results with an average accuracy over 99.4%.

7**. Detection of GAN-Generated face identification using convolutional neural network.**

AUTHORS: **Marra, F.; Gragnaniello, D.; Cozzolino, D.; Verdoliva.**

Generative adversarial networks (GANs) can be used to generate a photo-realistic image from a low-dimension random noise. Such a synthesized (fake) image with inappropriate content can be used on social media networks, which can cause severe problems. With the aim to successfully detect fake images, an effective and efficient image forgery detector is necessary. However, conventional image forgery detectors fail to recognize fake images generated by the GAN-based generator since these images are generated and manipulated from the source image.

8. **Understanding Digital Image Processing.**

AUTHORS: Tyagi, V.

Digital images processing deals with processing of images which are stored in digital form. A digital image is a 2-dimensional representation in the form of a matrix. This chapter describes matrix representation of intensity (grayscale) image and color images. Introduction to digital image processing and its history is also given. There are a number of operations that can be performed on digital images. A brief introduction of basic image processing operations is provided. The anatomy of the human eye and various phenomena associated with human visual perception are described.

9**. A robust content based digital signature for image authentication.**

AUTHORS: Schneider, M., Chang, S.

A methodology for designing content based digital signatures which can be used to authenticate images is presented. A continuous measure of authenticity is presented which forms the basis of this methodology. Using this methodology signature systems can be designed which allow certain types of image modification (e.g. lossy compression) but which prevent other types of manipulation. Some experience with content based signatures is also presented.Key Method The idea of signature based authentication is extended to video, and a system to generate signatures for video sequences is presented.

10. Digital Watermarking and Steganography Second Edition.

AUTHORS: Cox, I.J., Miller, M.L., Bloom, J.A., Kalker, T

Digital watermarking and steganography technology greatly reduces the instances of this by limiting or eliminating the ability of third parties to decipher the content that he has taken. The many techiniques of digital watermarking (embedding a code) and steganography (hiding information) continue to evolve as applications that necessitate them do the same.

2.2 CONCLUSIONS ON SURVEY

These readings provide basic background information about various techniques and algorithms in deep learning in various stages such as CNN model and fusion of Google data analysis and in another stage feature layer of Yolo V3 model used by using image pyramid to achieve multi-scale feature detection in every stage these various models. So here also deep learning is used as it have the advantages of clear structure and high accuracy in image recognition. So deep learning used for detecting fake images. Every reference gave an example to how to use deep learning with different algorithms giving us a basic idea about which the best to use.

3. SOFTWARE AND HARDWARE REQUIRMENTS

3.1 SOFTWARE REQUIRMENTS

For developing the application the following are the Software Requirements:

Operating Systems supported

Windows 7

Technologies and Languages used to Develop

Python

3.2 HARDWARE REQUIRMENTS

For developing the application the following are the Hardware Requirements:

Processor: DUAL CORE

RAM: 2 GB

Space on Hard Disk: 5MB

4.SOFTWARE DEVELOPMENT ANALYSIS

4.1 OVERVIEW OF PROBLEM

With the advent of social networking services such as Facebook and Instagram, there has been a huge increase in the volume of image data generated in the last decade.

Use of image (and video) processing software like GNU GIMP, ADOBE PHOTOSHOP to create doctored images and videos is a major concern for internet companies like Facebook.

These images are prime sources of fake news and are often used in malevolent ways such as for mob incitement.

Before action can be taken on basis of a questionable image, we must verify its authenticity.

The IEEE Information Forensics and Security Technical Committee (IFS-TC) launched a detection and localization forensics challenge, the FIRST IMAGE in Forensics Challenge 2013 to solve this problem.

4.2 DEFINE THE PROBLEM

we have proposed a fake feature network based the pairwise learning, to detect the fake face/general images generated by state-of-the-art GANs successfully.

The proposed CFFN can be used to learn the middle- and high-level and discriminative fake feature by aggregating the cross-layer feature representations into the last fully connected layers.

The proposed pairwise learning can be used to improve the performance of fake image detection further.

With the proposed pairwise learning, the proposed fake image detector should be able to have the ability to identify the fake image generated by a new GAN.

Our experimental results demonstrated that the proposed method outperforms other state-of-the-art schemes in terms of precision and recall rate.

4.3 MODULES OVERVIEW

In this project we are designing LBP Based machine learning Convolution Neural Network called LBPNET to detect fake face images. Here first we will extract LBP from images and then train LBP descriptor images with Convolution Neural Network to generate training model. Whenever we upload new test image then that test image will be applied on training model to detect whether test image contains fake image or non-fake image. Below we can see some details on LBP.

4.4 DEFINE THE MODULES

This Project Mainly consists of for modules. They are:

1. Generate Image Train And Test Model.

2. Upload Test Image.

3. Classify Picture In Image.

4. Exit.

4.5 MODULE FUNCTONALITY

1. Generate Image Train And Test Model- This is the first module of our project where we need to generate image train and test model.
2. Upload Test Image- In this module we basically upload a test image from the data base.
3. Classify picture In Image- This module takes a image as an input and identifies weather the image is fake or not.
4. Exit- This will let us come out of the project implementation and closes the project.

5. PROJECT SYSTEM DESIGM

5.1 SYSTEM ARCHITETURE

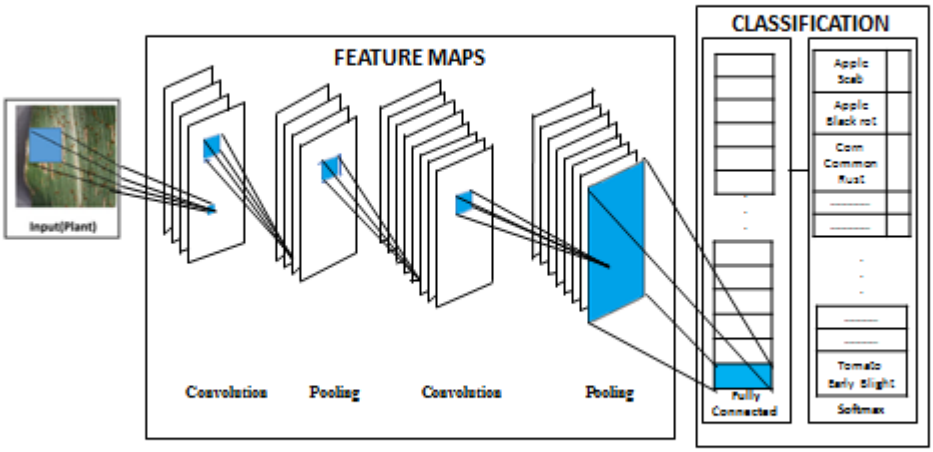


Fig 5.1: CNN Architecture

Convolutional neural networks are made up of three components in most cases. Convolution layer, for feature extraction. The pooling layer, also known as the convergence layer, is primarily used for feature selection. By reducing the number of features, the number of parameters is decreased. The summary and output of the characteristics are carried out by the full connection layer. A convolution layer is made up of a convolution mechanism and the ReLU nonlinear activation function. Figure 1 depicts a standard CNN model architecture for pattern recognition.

The input layer is the image on the left, which the machine interprets as the input of several matrices. The convolution layer follows, with ReLU as its activation feature. The pooling layer has no activation function. Many different combinations of convolution and pooling layers can be established. When building the model, the combination of convolution layer and convolution layer, or convolution layer and pool layer, can be quite flexible. However, the most popular CNN is made up of a number of convolutional and pooling layers. Finally, a full connection layer serves as a classifier, mapping the learned

feature representation to the sample label space.

FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are,

- **ECONOMICAL FEASIBILITY**
- **TECHNICAL FEASIBILITY**
- **SOCIAL FEASIBILITY**

ECONOMICAL FEASIBILITY

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

TECHNICAL FEASIBILITY

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

SOCIAL FEASIBILITY

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

5.2 UML DIAGRAMS

Usecase Diagram

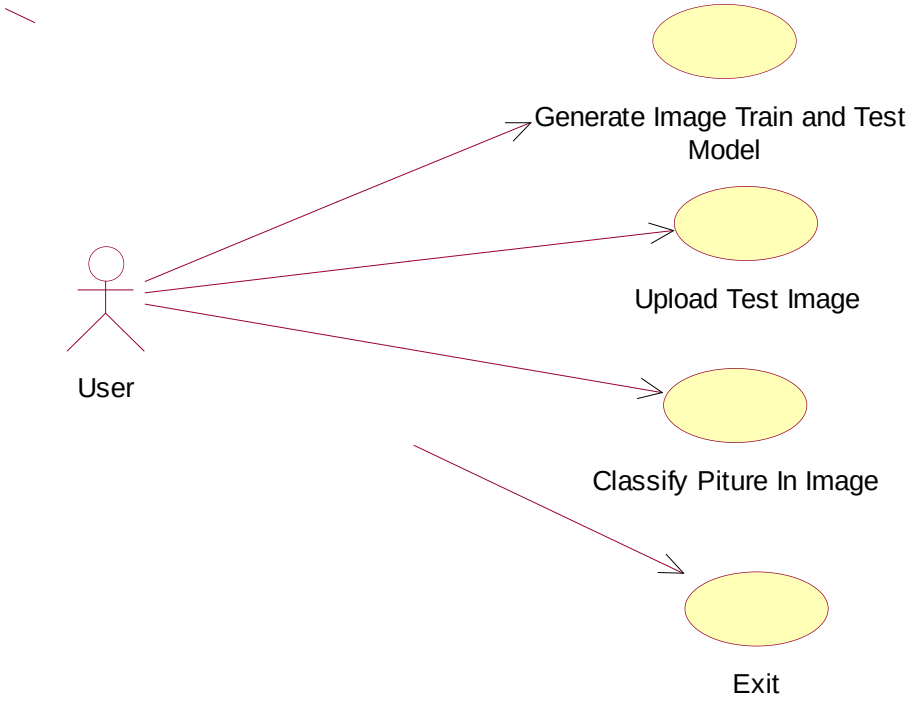


Figure5.2: Usecase Diagram

Upload a image in train and test model. Then upload test image then after classify picture in image then the result will be displayed as if the image is fake the image contains fake faces.

Class Diagram

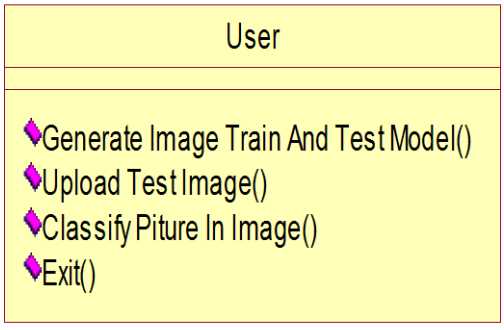


Figure5.3:Class Diagram

Generate image train and test model and then upload test image after that an pop up will be displayed that test file uploaded successfully after that classify picture in image should be selected thereafter if the image contains any fake content then this image contains fake faces will be displayed if the image does not contains fake faces then the image does not contains fake faces will be displayed

Sequence Diagram

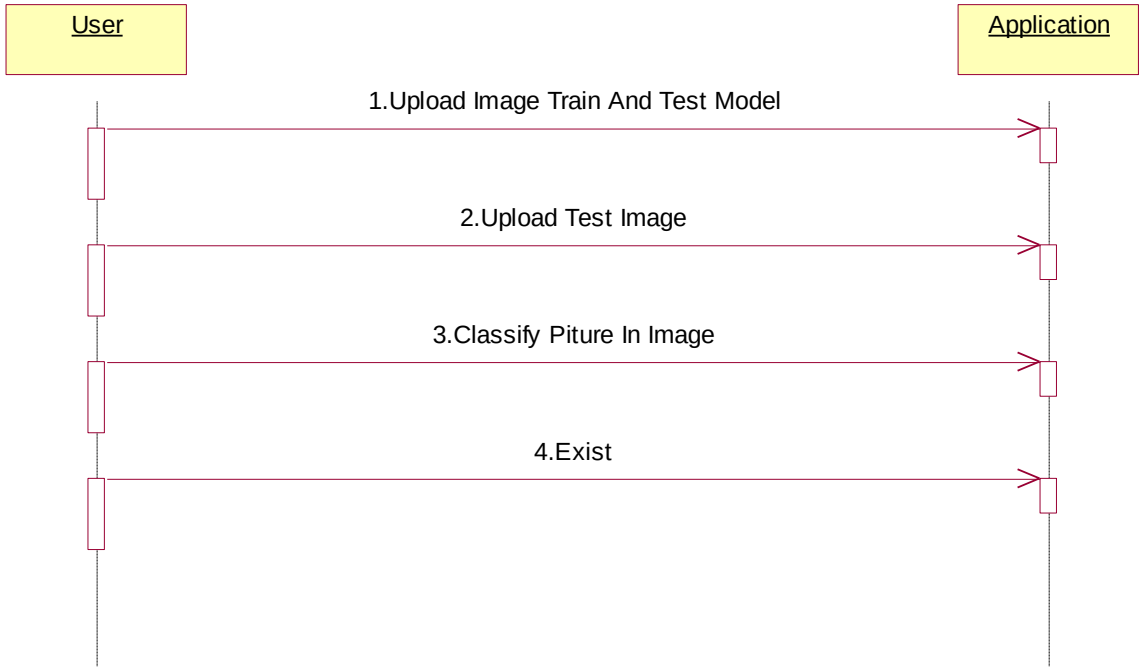


FIGURE5.4:SEQUENCE DIAGRAM

Upload a image in train and test model. Then upload test image then after classify picture in image then the result will be displayed as if the image is fake the image contains fake faces.

Collaboration Diagram

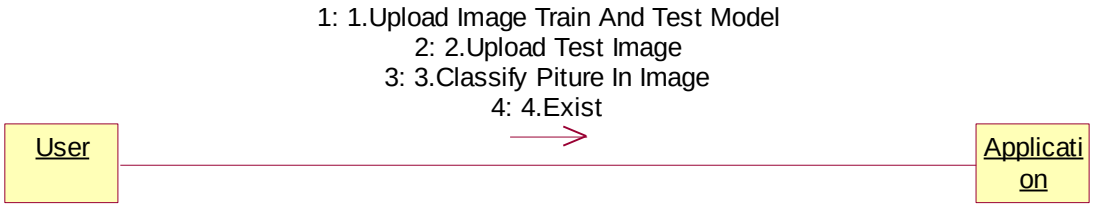


Figure 5.5:Collaboration Diagram

Generate image train and test model and then upload test image after that an pop up will be displayed that test file uploaded successfully after that classify picture in image should be selected thereafter if the image contains any fake content then this image contains fake

6. PROJECT CODING

6.1 CODE TEMPLATES

```
global main
main.destroy()
font = ('times', 16, 'bold')
title = Label(main, text='Fake Image Identification', justify=LEFT)
title.config(bg='lavender blush', fg='DarkOrchid1')
title.config(font=font)
title.config(height=3, width=120)
title.place(x=100,y=5)
title.pack()
font1 = ('times', 14, 'bold')
model = Button(main, text="Generate image Train & Test Model",
command=generateModel)
model.place(x=200,y=100)
model.config(font=font1)
uploadimage = Button(main, text="Upload Test Image", command=upload)
uploadimage.place(x=200,y=150)
uploadimage.config(font=font1)

classifyimage = Button(main, text="Classify Picture In Image",
command=classify)
classifyimage.place(x=200,y=200)
classifyimage.config(font=font1)
exitapp = Button(main, text="Exit", command=exit)
exitapp.place(x=200,y=250)
exitapp.config(font=font1)
```

```
main.config(bg='light coral')
main.mainloop()
```

6.2 OUTLINE FOR VARIOUS FILES

We used Python programming to implement our project. A single python file is used to implement our code. This file consists of various modules that we have used. Our project modules are- Generate Image Train and Test Model, Upload Test Image, Classify Picture Image and Exit. The LBP feature vector, in its simplest form, is created in the following manner:

Divide the examined window into cells (e.g. 16x16 pixels for each cell).

For each pixel in a cell, compare the pixel to each of its 8 neighbors (on its left-top, left-middle, left-bottom, right-top, etc.). Follow the pixels along a circle, i.e. clockwise or counter-clockwise.

Where the center pixel's value is greater than the neighbor's value, write "0". Otherwise, write "1". This gives an 8-digit binary number (which is usually converted to decimal for convenience).

Compute the histogram, over the cell, of the frequency of each "number" occurring (i.e., each combination of which pixels are smaller and which are greater than the center). This histogram can be seen as a 256-dimensional feature vector.

Optionally normalize the histogram.

Concatenate (normalized) histograms of all cells. This gives a feature vector for the entire window.

The feature vector can now be processed using the Support vector machine, extreme learning machines, or some other machine learning algorithm to classify images. Such classifiers can be used for face recognition or texture analysis.

6.3 METHODS INPUT AND OUTPUT PARAMETERS

In our project code we implemented four different methods. They are:

1. Generate Image Train and Test Model.
2. Upload Test Image.
3. Classify Picture Image.
4. Exit.

7. PROJECT TESTING

7.1 VARIOUS TST CASES

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

7.2 BLACK BOX TESTING

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box .you cannot “see” into it. The test provides inputs and responds to outputs without considering how the software works.

7.3 WHITE BOX TESTING

White Box Testing is a testing in which in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is purpose. It is used to test areas that cannot be reached from a black box level.

Functional test

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items:

Valid Input : identified classes of valid input must be accepted.

Invalid Input : identified classes of invalid input must be rejected.

Functions : identified functions must be exercised.

Output : identified classes of application outputs must be exercised.

Systems/Procedures : interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

Test strategy and approach

Field testing will be performed manually and functional tests will be written in detail.

Test objectives

- All field entries must work properly.
- Pages must be activated from the identified link.
- The entry screen, messages and responses must not be delayed.

Features to be tested

- Verify that the entries are of the correct format
- No duplicate entries should be allowed
- All links should take the user to the correct page.

Integration Testing

Software integration testing is the incremental integration testing of two or more integrated software components on a single platform to produce failures caused by interface defects.

The task of the integration test is to check that components or software applications, e.g. components in a software system or – one step up – software applications at the company level – interact without error.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

Acceptance Testing

User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements.

8. OUTPUT SCREENS

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output.

It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making.

1. Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is designed so that people will find the system can use easily and effectively. When analysis design computer output, they should Identify the specific output that is needed to meet the requirements.

2.Select methods for presenting information.

3.Create document, report, or other formats that contain information produced by the system.

The output form of an information system should accomplish one or more of the following objectives.

- Convey information about past activities, current status or projections of the
- Future.
- Signal important events, opportunities, problems, or warnings.
- Trigger an action.
- Confirm an action.

8.1 USER INTERFACE

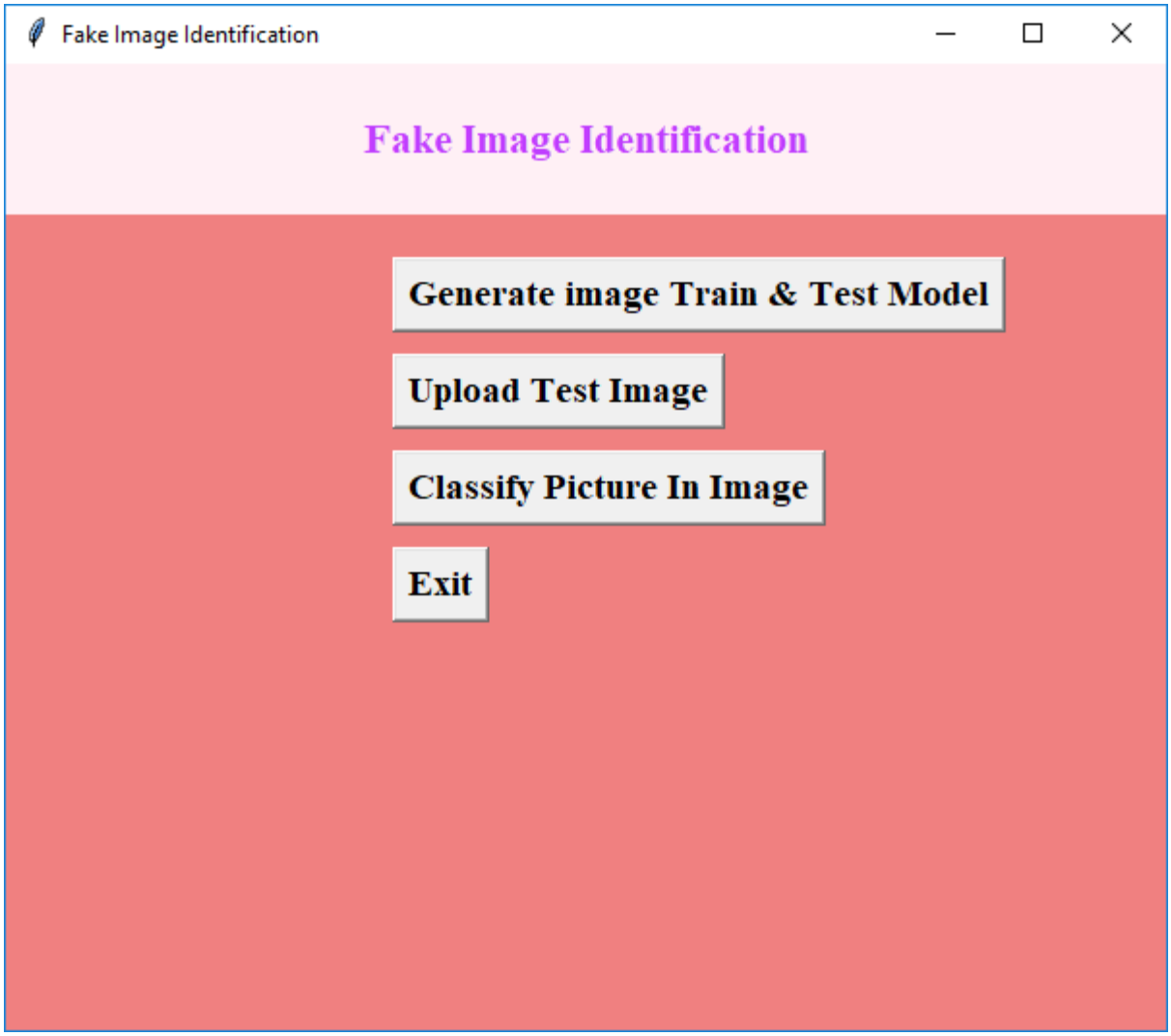


FIGURE:8.1 USER INTERFACE

In above screen click on ‘Generate Image Train & Test Model’ button to generate CNN model using LBP images contains inside LBP folder.

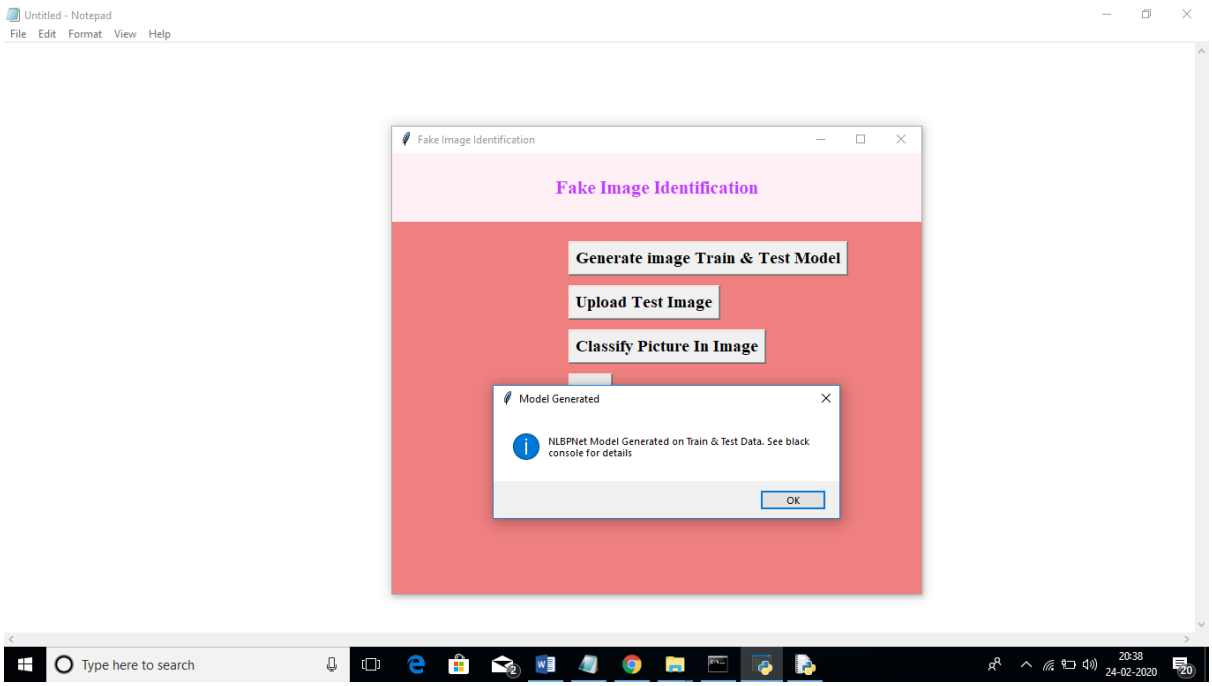


FIGURE:8.2 OUTPUT SCREEN

In above screen we can see CNN LBPNET model generated. Now click on 'Upload Test Image' button to upload test image.

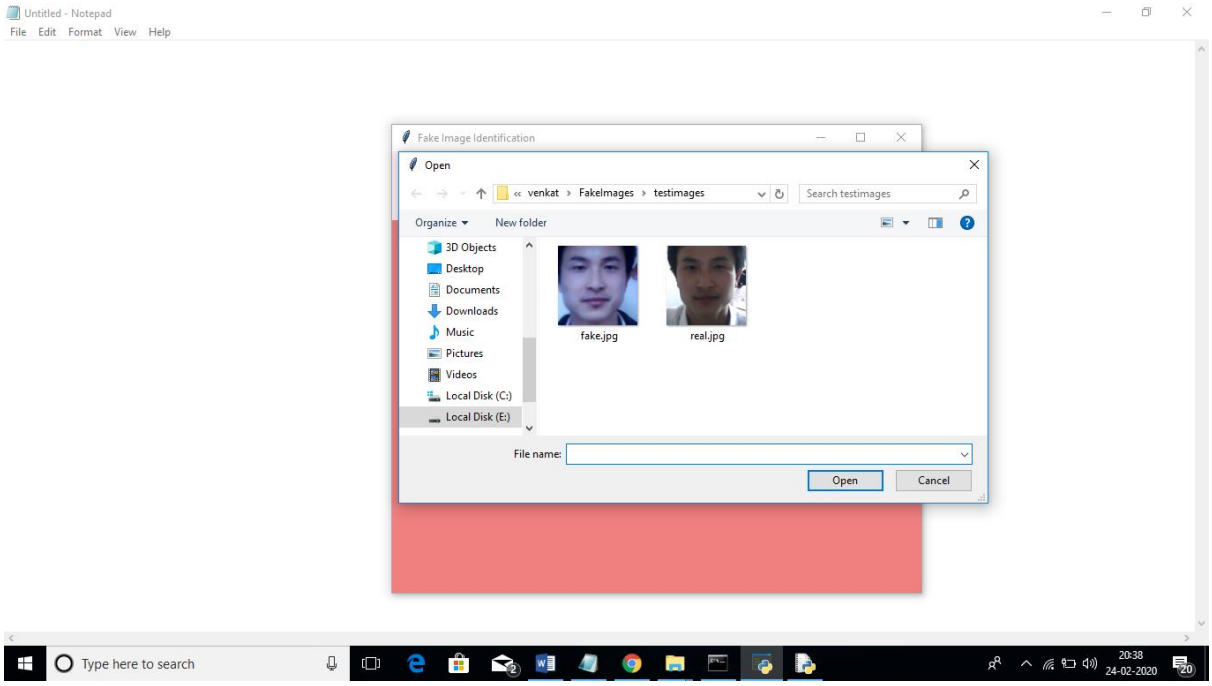


FIGURE:8.3 IMAGE FOULDER

In above screen we can see two faces are there from same person but in different appearances. For simplicity I gave image name as fake and real to test whether application can detect it or not. In above screen I am uploading fake image and then click on 'Classify Picture In Image' button to get below result.

8.2 OUTPUT SCREENS

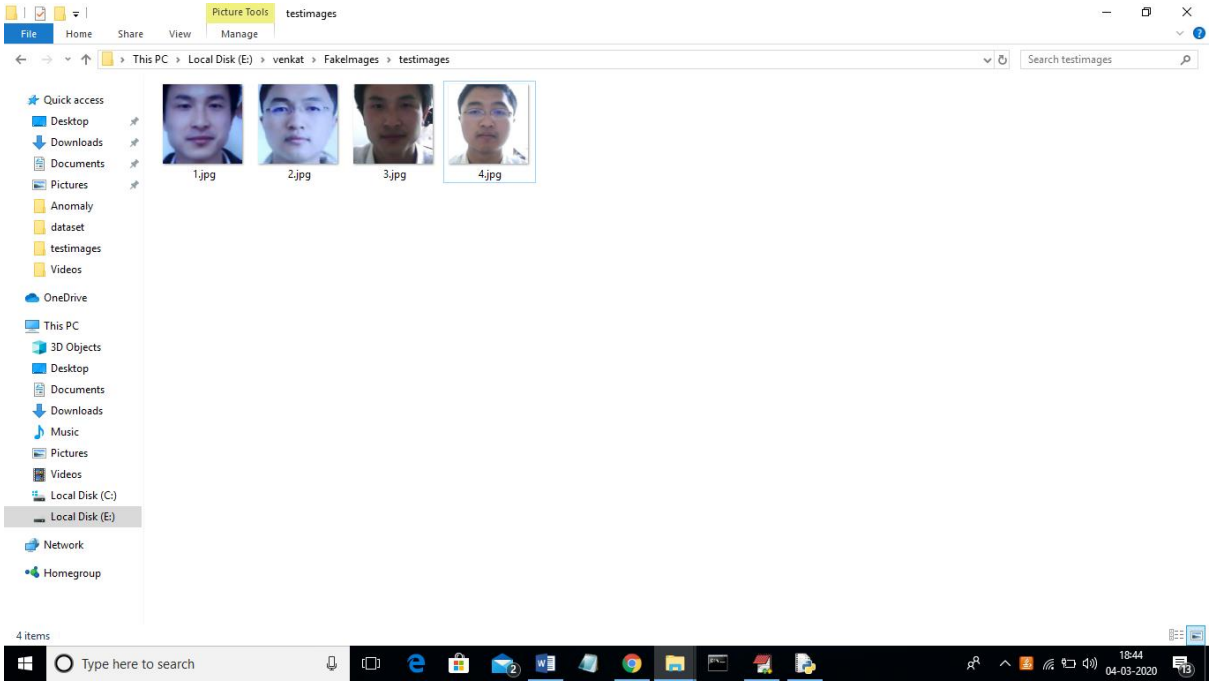


FIGURE:8.4 IMAGE DATABASE

In above screen we can see all real face will have normal light and in fake faces peoples will try some editing to avoid detection but this application will detect whether face is real or fake.

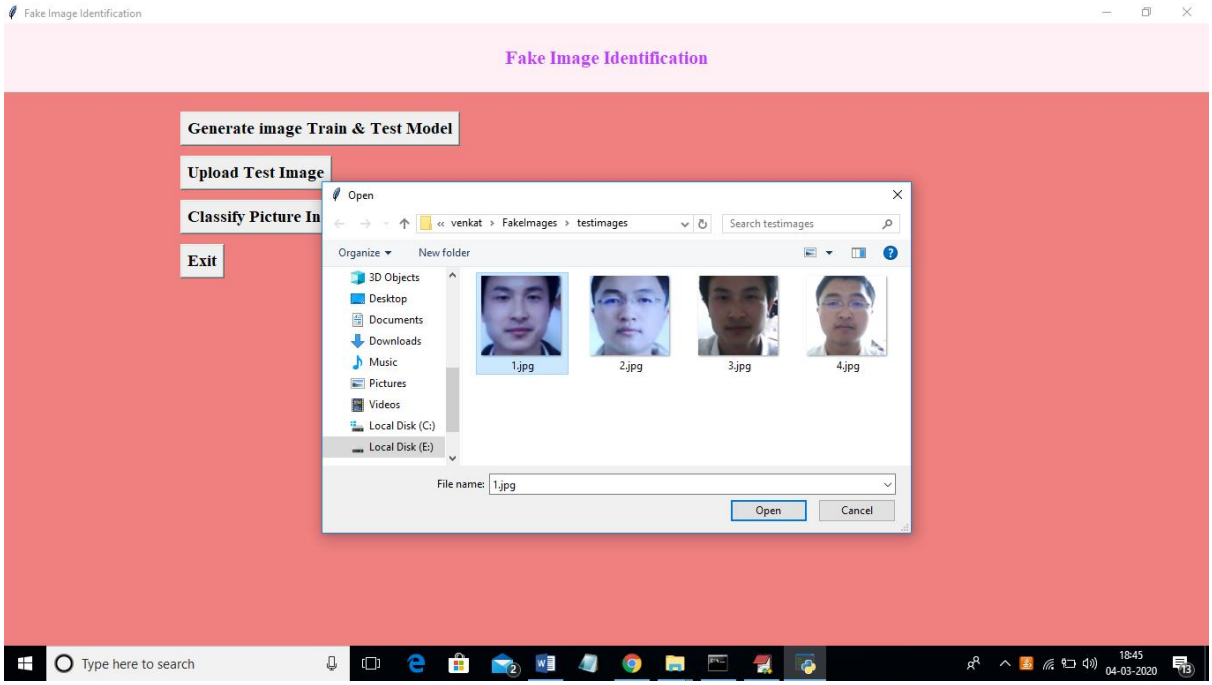


FIGURE:8.5 IMAGES

In above screen I am uploading 1.jpg and after upload click on open button to get below screen.

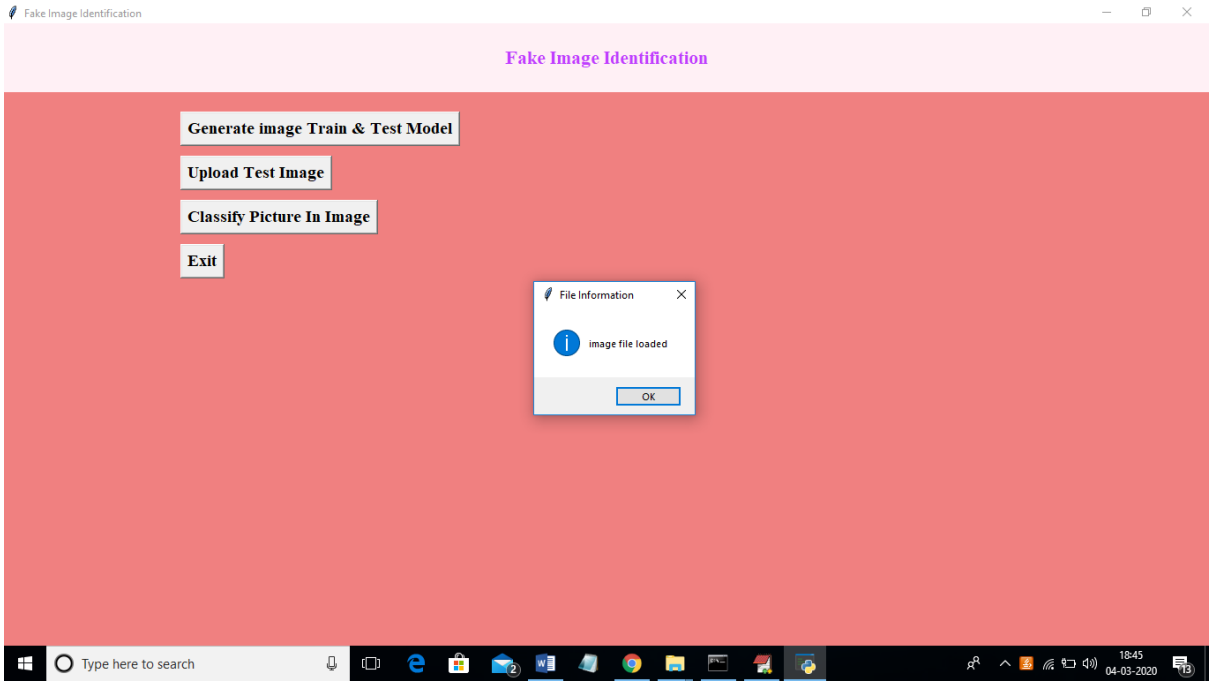


FIGURE:8.6 POPUP DISPLAY

And now click on ‘classify Picture in Image’ to get below details.

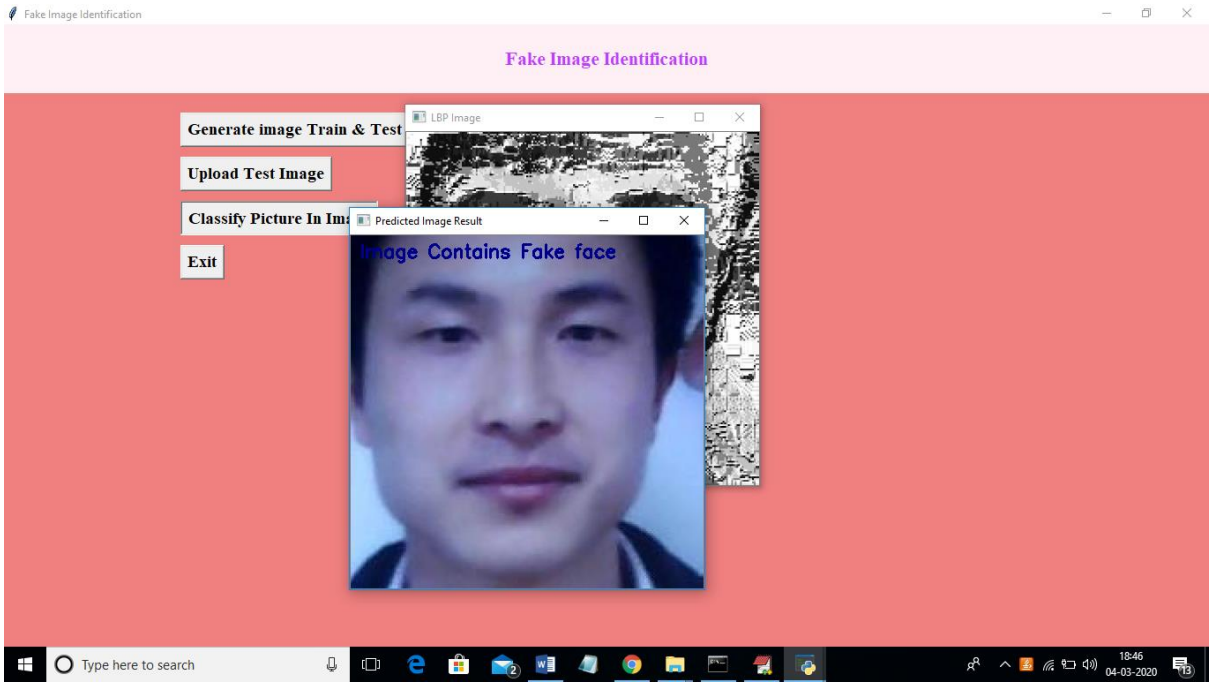


FIGURE:8.7 OUTPUT DISPLAY

In above screen we are getting result as image contains Fake face. Similarly u can try other images also. If u want to try new images then u need to send those new images to us so we will make CNN model to familiar with new images so it can detect those images also.

9. EXPERIMANTAL RESULTS

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making.

1. Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is designed so that people will find the system can use easily and effectively. When analysis design computer output, they should Identify the specific output that is needed to meet the requirements.

2.Select methods for presenting information.

3.Create document, report, or other formats that contain information produced by the system.

The output form of an information system should accomplish one or more of the following objectives. Convey information about past activities, current status or projections of the Future.

10. CONCLUSION AND FUTURE ENHANCEMENT

In this project, we have proposed a novel common fake feature network based the pairwise learning, to detect the fake face/general images generated by state-of-the-art GANs successfully. The proposed CFFN can be used to learn the middle- and high-level and discriminative fake feature by aggregating the cross-layer feature representations into the last fully connected layers. The proposed pairwise learning can be used to improve the performance of fake image detection further. With the proposed pairwise learning, the proposed fake image detector should be able to have the ability to identify the fake image generated by a new GAN. Our experimental results demonstrated that the proposed method outperforms other state-of-the-art schemes in terms of precision and recall rate. Future work is the accuracy can be improved if we go for techniques for image detection.

REFERENCE

1.Wu, Y.; Bai, Z.; Miao, Q.; Ma, W.; Yang, Y.; Gong, M. A Classified Adversarial Network for Multi-Spectral Remote Sensing Image Change Detection. *Remote Sens.* **2020**, *12*, 2098. https://doi.org/10.3390/rs12132098

2.Wu Y, Bai Z, Miao Q, Ma W, Yang Y, Gong M. A Classified Adversarial Network for Multi-Spectral Remote Sensing Image Change Detection. 2020; 12(13):2098. https://doi.org/10.3390/rs12132098

3.Wu, Yue; Bai, Zhuangfei; Miao, Qiguang; Ma, Wenping; Yang, Yuelei; Gong, Maoguo. 2020. "A Classified Adversarial Network for Multi-Spectral Remote Sensing Image Change Detection. 12, no. 13: 2098. https://doi.org/10.3390/rs12132098

4. Hsu, C.-C.; Zhuang, Y.-X.; Lee, C.-Y. Deep Fake Image Detection Based on Pairwise Learning. *Appl. Sci.* **2020**, *10*, 370. https://doi.org/10.3390/app10010370

5.Hsu C-C, Zhuang Y-X, Lee C-Y. Deep Fake Image Detection Based on Pairwise Learning. *Applied Sciences*. 2020; 10(1):370. https://doi.org/10.3390/app10010370

6.Hsu, Chih-Chung; Zhuang, Yi-Xiu; Lee, Chia-Yen. 2020. "Deep Fake Image Detection Based on Pairwise Learning" *Appl. Sci.* 10, no. 1: 370. https://doi.org/10.3390/app10010370. 7.Nicholas Carlini and Hany Farid, “Evading deepfakeimage detectors with white-and black-box attacks,” in Proceedings of the CVPRW, 2020.

8.Apurva Gandhi and Shomik Jain, “Adversarial perturbations fool deepfake detectors,” in IJCNN, 2020.

9. Huaxiao Mo, B.C.; Luo, W. Fake Faces Identification via Convolutional Neural Network

Proc. of the ACM Workshop on Information Hiding and Multimedia Security. ACM, 2018, pp. 43–47.

10. Marra, F.; Gragnaniello, D.; Cozzolino, D.; Verdoliva, L. Detection of GAN-Generated face identification using convolutional neural network. Appl. Sci. 2018,8, 2610.

11.Fake Images over Social Networks. Proc. of the IEEE Conference on Multimedia Information Processing and Retrieval, 2019, 274 pp. 384–389. doi:10.1109/MIPR.2018.00084github, “deepfake” Accessed Apr 2, 2020.

12.Apostolos Modas, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard, “Sparsefool: a few pixels make a big difference,” in CVPR, 2019.

13.Dang, L.; Hassan, S.; Im, S.; Lee, J.; Lee, S.; Moon, H. Deep learning based computer generated faceidentification using convolutional neural network. Appl. Sci. 2018,8, 2610.

14.Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of theIEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA,21–26 July 2017;pp. 1610–02357

APPENDICES

```
from tkinter import *
import tkinter
from tkinter import filedialog
import numpy as np
from tkinter.filedialog import askopenfilename
import pandas as pd
from keras.optimizers import Adam
from keras.models import model_from_json
from tkinter import simpledialog

from keras.models import Sequential
from keras.layers import Convolution2D
from keras.layers import MaxPooling2D
from keras.layers import Flatten
from keras.layers import Dense,Activation,BatchNormalization
import os
from keras.preprocessing import image
from keras.preprocessing.image import ImageDataGenerator
from tkinter import messagebox
```

```
import cv2
from imutils import paths
import imutils
import cv2
import numpy as np



main = tkinter.Tk()
main.title("Fake Image Identification") #designing main screen
main.geometry("600x500")

global filename
global loaded_model

def get_pixel(img, center, x, y):
    new_value = 0
    try:
        if img[x][y] >= center:
            new_value = 1
    except:
        pass
    return new_value
```

```
def lbp_calculated_pixel(img, x, y):
    center = img[x][y]
    val_ar = []
    val_ar.append(get_pixel(img, center, x-1, y+1))     # top_right
    val_ar.append(get_pixel(img, center, x, y+1))       # right
    val_ar.append(get_pixel(img, center, x+1, y+1))     # bottom_right
    val_ar.append(get_pixel(img, center, x+1, y))       # bottom
    val_ar.append(get_pixel(img, center, x+1, y-1))     # bottom_left
    val_ar.append(get_pixel(img, center, x, y-1))       # left
    val_ar.append(get_pixel(img, center, x-1, y-1))     # top_left
    val_ar.append(get_pixel(img, center, x-1, y))       # top

    power_val = [1, 2, 4, 8, 16, 32, 64, 128]
    val = 0
    for i in range(len(val_ar)):
        val += val_ar[i] * power_val[i]
    return val


def upload(): #function to upload tweeter profile
    global filename
```

```
    filename = filedialog.askopenfilename(initialdir="testimages")
    messagebox.showinfo("File Information", "image file loaded")


def generateModel():
    global loaded_model
    if os.path.exists('model.json'):
        with open('model.json', "r") as json_file:
            loaded_model_json = json_file.read()
            loaded_model = model_from_json(loaded_model_json)


        loaded_model.load_weights("model_weights.h5")
        loaded_model._make_predict_function()
        print(loaded_model.summary())
        messagebox.showinfo("Model Generated", "CNN Model Generated on Train & Test Data. See black console for details")
    else:
        classifier = Sequential()
        classifier.add(Convolution2D(32, (3, 3), border_mode='valid', input_shape=(48, 48, 1)))
        classifier.add(BatchNormalization())
        classifier.add(Activation("relu"))
        classifier.add(Convolution2D(32, (3, 3),  border_mode='valid'))
```

```
classifier.add(BatchNormalization())
classifier.add(Activation("relu"))
classifier.add(MaxPooling2D(pool_size=(2, 2)))
classifier.add(Flatten())
classifier.add(Dense(128))
classifier.add(BatchNormalization())
classifier.add(Activation("relu"))
classifier.add(Dense(2))
classifier.add(BatchNormalization())
classifier.add(Activation("softmax"))
# model5 the model
classifier.compile(optimizer='adam', loss='categorical_crossentropy',
metrics=['accuracy'])
files = []
filename = 'LBP/train/Fake'
label = []
for root, dirs, directory in os.walk(filename):
    for i in range(len(directory)):
        files.append(filename+"/"+directory[i]);
        label.append([1,0])

filename = 'LBP/train/Real'
for root, dirs, directory in os.walk(filename):
```

```
    for i in range(len(directory)):
        files.append(filename+"/"+directory[i]);
        label.append([0,1])


print(len(files))
X = np.ndarray(shape=(len(files), 48,48,1), dtype=np.float32)
Y = np.ndarray(shape=(len(files),2),dtype=np.float32)
print(X.shape)
print(Y.shape)
for i in range(len(files)):
    img = cv2.imread(files[i])
    img = cv2.cvtColor(img, cv2.COLOR_BGR2GRAY)
    img = np.resize(img, (48,48,1))
    im2arr = np.array(img)
    im2arr = im2arr.reshape(1,48,48,1)
    X[i] = im2arr
    Y[i] = label[i]
print("shape == "+str(X.shape))
#X = X.reshape(X.shape[0],48, 48,3)
classifier.fit(X, Y,epochs = 10)
classifier.save_weights('model_weights.h5')
model_json = classifier.to_json()
```

```
        with open("model.json", "w") as json_file:
            json_file.write(model_json)
        print(X.class_indices)
        print(classifier.summary)
        messagebox.showinfo("Model Generated", "NLBPNet Model Generated on Train & Test Data. See black console for details")


def classify():
    name = os.path.basename(filename)
    image_file = filename;
    img_bgr = cv2.imread(image_file)
    height, width, channel = img_bgr.shape
    img_gray = cv2.cvtColor(img_bgr, cv2.COLOR_BGR2GRAY)
    img_lbp = np.zeros((height, width,3), np.uint8)
    for i in range(0, height):
        for j in range(0, width):
            img_lbp[i, j] = lbp_calculated_pixel(img_gray, i, j)
    cv2.imwrite('testimages/lbp_'+name, img_lbp)
    img = cv2.imread('testimages/lbp_'+name)
    img = cv2.cvtColor(img, cv2.COLOR_BGR2GRAY)
    img = np.resize(img, (48,48,1))
    im2arr = np.array(img)
```

```
    im2arr = im2arr.reshape(1,48,48,1)
    preds = loaded_model.predict(im2arr)
    print(str(preds)+" "+str(np.argmax(preds)))
    predict = np.argmax(preds)
    msg = ""
    if predict == 0:
        msg = "Image Contains Fake face"
    if predict == 1:
        msg = "Image Contains Real face"
    imagedisplay = cv2.imread(filename)
    orig = imagedisplay.copy()
    output = imutils.resize(orig, width=400)
    cv2.putText(output, msg, (10, 25),  cv2.FONT_HERSHEY_SIMPLEX,0.7, (139, 0, 0),
2)
    cv2.imshow("Predicted Image Result ", output)
    imagedisplay = cv2.imread('testimages/lbp_'+name)
    orig = imagedisplay.copy()
    output = imutils.resize(orig, width=400)
    os.remove('testimages/lbp_'+name)
    cv2.imshow("LBP Image", output)
    cv2.waitKey(0)
```

```
def exit():
    global main
    main.destroy()

font = ('times', 16, 'bold')
title = Label(main, text='Fake Image Identification', justify=LEFT)
title.config(bg='lavender blush', fg='DarkOrchid1')
title.config(font=font)
title.config(height=3, width=120)
title.place(x=100,y=5)
title.pack()

font1 = ('times', 14, 'bold')
model = Button(main, text="Generate image Train & Test Model",
command=generateModel)
model.place(x=200,y=100)
model.config(font=font1)

uploadimage = Button(main, text="Upload Test Image", command=upload)
uploadimage.place(x=200,y=150)
uploadimage.config(font=font1)
```

```
classifyimage = Button(main, text="Classify Picture In Image", command=classify)
classifyimage.place(x=200,y=200)
classifyimage.config(font=font1)

exitapp = Button(main, text="Exit", command=exit)
exitapp.place(x=200,y=250)
exitapp.config(font=font1)

main.config(bg='light coral')
main.mainloop()
```

A

PROJECT REPORT

On

Blockchain E-Voting Done Right Privacy and Transparency with Public Blockchain

Submitted by

1)Mr. K.Saisriram (17K81A0591) **2)Mr. R.S.V.Suvesith(17K81A05A6)**

3)Ms. Swathi Tomar (17K81A05B2) **4) Mr. T.Srikanth Reddy(17K81A05B3)**

in partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Under the Guidance of

P. Sabitha

Assistant Professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ST.MARTIN'S ENGINEERING COLLEGE

An Autonomous Institute

Dhulapally, Secunderabad – 500 100

JUNE 2021

BONAFIDE CERTIFICATE

This is to certify that the project entitled **Blockchain E-Voting Done Right Privacy and Transparency with Public Blockchain**, is being submitted by **1. Mr. Kothapalli Saisriram 17K81A0591, 2. Mr. Ramayanam Sai Veer Suvesith 17K81A05A6, 3. Ms. Swathi Tomar 17K81A05B2, 4. Mr. Tummaluru Srikanth Reddy 17K81A05B3** in partial fulfillment of the requirement for the award of the degree of **BACHELOR OF TECHNOLOGY IN DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING** is recorded of bonafide work carried out by them. The results embodied in this report have been verified and found satisfactory.

P.Sabitha
Department of CSE

Head of the Department
Dr.M.NARAYANAN
Department of CSE

Internal Examiner

External Examiner

Place:

Date:

DECLARATION

We, the student of **Bachelor of Technology** in Department of Computer Science and Engineering, session: 2017 – 2021, St. Martin's Engineering College, Dhulapally, Kompally, Secunderabad, hereby declare that work presented in this Project Work entitled Blockchain E-Voting Done Right Privacy and Transparency with Public Blockchain is the outcome of our own bonafide work and is correct to the best of our knowledge and this work has been undertaken taking care of Engineering Ethics. This result embodied in this project report has not been submitted in any university for award of any degree.

Kothapalli Saisriram	17K81A0591
Ramayanam Sai Veer Suvesith	17K81A05A6
Swathi Tomar	17K81A05B2
Tummaluru Srikanth Reddy	17K81A05B3

ABSTRACT

With the advanced technology and developments since the 20th century, new procedure of casting votes in an election is developed every now and then. This project uses advanced technology like block chain and homomorphic encryption in order to make the election more safe and secure. By implementing the idea of block chain e-voting the elections can be made more fair, as it double checks the votes casted by the voters before and after the elections. Moreover, it eliminates the chances of malpractices as images of voters are taken into consideration. Hence, a voter can only vote once and can recheck their vote.

At present the voting is done using paper ballots and electronic voting but it has problems mainly regarding security, credibility, transparency, reliability, and functionality. So, block chain e-voting can deliver an answer to all these problems and further can add advantages like immutability and decentralization.

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible and whose encouragement and guidance have crowded effects with success.

We extended our deep sense of gratitude to Principal**, Dr. P. SANTOSH KUMAR PATRA**, St. Martin's Engineering College, Dhulapally, for permitting us to undertake this project.

We are also thankful to **Dr. M.NARAYANAN**, Head of the Department, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

We would like to thank **Dr. T. POONGOTHAI,** Department of Computer Science and Engineering (AI & ML) for her encouragement and insightful comments and as well as our project coordinators **Dr. B.RAJALINGAM**, Associate Professor and **Dr.G.GOVINDARAJULU,** Associate Professor, in the Department of Computer Science and Engineering for their valuable support.

We would like to express our sincere gratitude and indebtedness to our project supervisor **Ms. P. Sabitha,** Assistant Professor, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

Finally, we express thanks to all those who have helped us successfully to complete this project. Furthermore, we would like to thank our family and friends for their moral support and encouragement.

We express thanks to all those who have helped us in successfully completing the project.

Kothapalli Sai Sriram	17K81A0591
Ramayanam Sai Veer Suvesith	17K81A05A6
Swathi Tomar	17K81A05B2
Tummaluru Srikanth Reddy	17K81A05B3

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF OUTPUT SCREENS

LIST OF ACRONYMS

<API>	Application Programming Interface
<FHE>	Fully Homomorphic Encryption
<HTML>	Hyper Text Markup Language
<GB>	Giga Bytes
<GUI>	Graphical User Interface
<CSS>	Cascading Style Sheets

1.Introduction

1.1 PROJECT OVERVIEW:

There are a number of people and parties complaining about the unfair election process and the violence at the booth, blockchain e-voting is a solution to this. In this project we are using the public python Blockchain API to store and manage voting data as Blockchain provides secure and tamper proof of data storage. The admin is responsible to add new party and candidate details and can view party details and vote count. The user has to sign up with the application by using username as his ID and then upload his face photo which is captured from the webcam. After registration, users can go for a login which validates user id and after successful login user can go for cast vote module.

One of the reasons that electoral officials have been slow to migrate voting online is fear that election integrity could be compromised by hackers.But that's where blockchain comes in, which promises to combine much-needed ballot security with voting convenience.Blockchain integrates cryptography into software in a unique way.

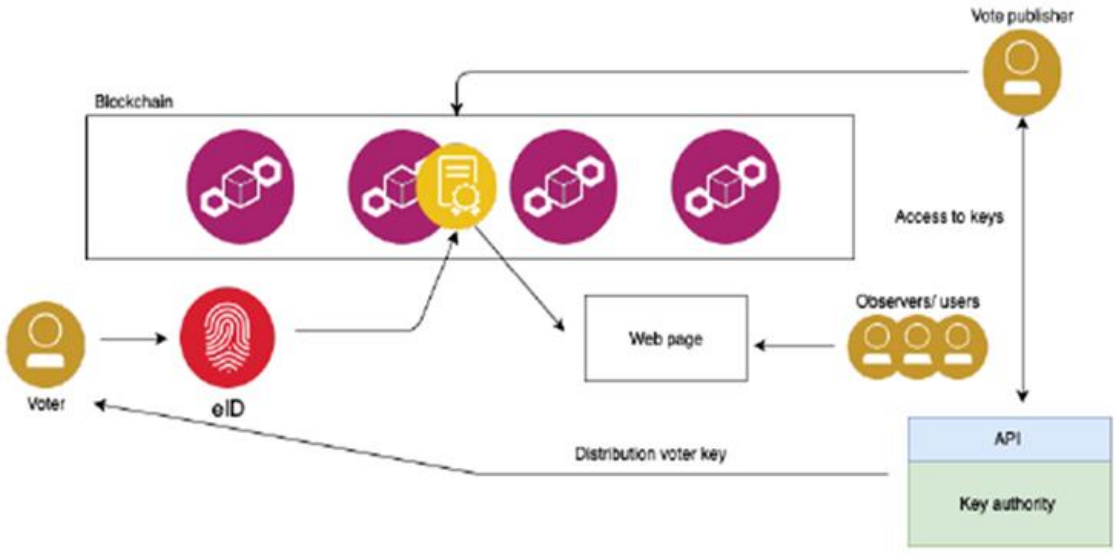


Fig 1.1: Architecture Diagram

1.2 Project Objectives

To create a E-voting platform with Blockchain using homomorphic encryption. To create a tamper-free record that can easily be checked to ensure votes are accurately recorded. Due to the secure and immutable nature of blockchain, votes may be cast by computer or mobile device instead of having voters show up at a local polling place or cast a mail-in-ballot to be processed manually by election officials. The main advantage is that there is no need for confidence in the centralized authority that created the elections. This authority cannot affect the election results in our system. After the start of the voting, the platform behaves as fully independent and decentralized without possibilities to affect the voting process. The data are fully transparent, but the identity of voters is secured by homomorphic encryption.

1.3 Organization Of Chapters

This document contains the following 10 Chapters.

Chapter -1 (Introduction): The material in this chapter of the Introduction is about the prologue of our project, which provides specifics about the project's overview and aims.

Chapter-2 (Literature Survey): The Literature Survey chapter explores into the details of a few existing papers that are relevant to our project research, and we round up the chapter with a list of linked references.

Chapter-3 (S/W and H/W Requirements): This chapter outlines the project's requirements. Both software and hardware requirements have been specified.

Chapter-4 (Software Development Analysis): We examine the overview and definition of the problem, as well as the overview, definition, and functionality of modules, in the software development analysis chapter.

Chapter-5 (Project System Design): This chapter contains a visual representation of our project. Our project's system design is represented through UML diagrams.

Chapter-6 (Project Coding): This chapter collects information regarding the project's coding. The code templates, files utilised in the code, class functionality, methods, inputs, and output parameters are all gathered.

Chapter-7 (Project Testing): This chapter details our project's testing process, including several types of testing such as black box and white box, as well as project test cases.

Chapter-8 (Output Screens): Our project's output screens are presented in this chapter. In which the project's user interface and outcomes are well understood.

Chapter-9 (Experimental Results): In this chapter, the project's experimental outcomes are presented.

Chapter-10 (Conclusion And Future Enhancement): This chapter concludes the project documentation, as well as future enhancements, references, paper publishing, student biography, and appendices.

2.Literature Survey

2.1 Survey On Background

[1] S Muralidhara, B A Usha, "Review of Blockchain Security and Privacy", Computing Methodologies and Communication (ICCMC) 2021 5th International Conference on, pp. 526-533, 2021.

Blockchain is a cutting-edge innovative technology that allows people to record message exchanges or transactions on a sophisticated, decentralised, distributed ledger without the need for a central controlling body, as in traditional banking systems. All users in the blockchain network can view the recorded transactions, and no user or node may change them.Many additional applications, including as IoT, healthcare, industrial, supply chain management, and so on, have embraced blockchain. Blockchain technology is poised to transform nearly every facet of our high-tech life. Blockchains ensure that our frameworks are more effective by eliminating third parties. They ensure that our frameworks are more unbiased by avoiding supervision.Furthermore, if properly implemented, they have the potential to make our frameworks more trustworthy and secure. The goal of this analysis is to have a better understanding of blockchain technology and the problems that come with it in terms of security and privacy.

[3]B Vivekanadam, "Analysis of Recent Trend and Applications in BlockChain Technology ", Journal of ISMAC, vol. 2, no. 04, pp. 200-206, 2020.

Blockchain is a digital ledger in which each item, referred to as a block, is joined into a single list, referred to as a chain. Bitcoin's backbone technology is known as this. It's also known as a unified collection of digital wallets. To record these transactions, blockchains are largely employed by cryptocurrencies like Bitcoin and other applications.A blockchain is a collection of distributed databases that contain all public transactions, records, and digital events. This information is then shared among participants. Every transaction is double-checked and cannot be reversed. The key characteristics of this technology are its dependability, efficiency, fault tolerance, and scalability. When the three traits are combined, manufacturing, government, and finance are some of the applications (i.e., Efficiency, Scalability and Security). Each transaction that is applied to a blockchain is validated by employing several computers. These tools that are used to validate these types of blockchain transactions create a peer-to-peer network. They collaborate to verify that any transaction is legal until it is put to the blockchain, and these systems are unable to add faulty blocks to the chain. When a new block is added, it can be linked to a previous block using a cryptographic hash, ensuring that the chain is unbreakable and that each block is permanently stored.

[7] N. Kshetri and J. Voas, “Blockchain-Enabled E-Voting,” IEEE Software, vol. 35, pp. 95–99, jul 2018.

BEV (blockchain-enabled e-voting) has the potential to eliminate election fraud while also increasing voter access. Eligible voters used a computer or smartphone to vote anonymously. BEV employs a tamper-proof personal ID and an encrypted key. This article examines some BEV deployments, as well as the approach's possible benefits and drawbacks. E-voting is one of the important public sectors that blockchain technology has the potential to disrupt. The concept behind blockchain-enabled e-voting (BEV) is straightforward. BEV gives each voter a "wallet" containing a user credential, to use a digital currency analogy. Each voter is given a single "coin" that represents one vote. When a voter casts a ballot, the voter's coin is transferred to the candidate's wallet.

[12] P. McCorry, S. F. Shahandashti, and F. Hao, “A Smart Contract for Boardroom Voting with Maximum Voter Privacy,” in Lecture Notes in Computer Science, ch. FCDS, pp. 357–375, Springer, Cham, 2017.

We demonstrate the use of the Blockchain to construct a decentralised and self-tallying internet voting mechanism with maximal voter anonymity. The Open Vote Network is a smart contract for Ethereum that is appropriate for boardroom elections. This is the first implementation of Blockchain e-voting that does not rely on any trusted authority to calculate the tally or safeguard the voter's privacy, unlike prior suggested Blockchain e-voting protocols. The Open Vote Network, on the other hand, is a self-tallying protocol in which each voter has complete control over the privacy of their own vote, which can only be violated by a full collusion involving all other voters. The protocol's execution is governed by the same consensus process that protects the Ethereum blockchain. To illustrate its viability, we put the implementation through its paces on Ethereum's official test network. We also give a financial and computational analysis of the cost of execution.

[15] Z. Brakerski and V. Vaikuntanathan, “Efficient Fully Homomorphic Encryption from (Standard) LWE,” SIAM Journal on Computing, vol. 43, pp. 831–871, jan 2014.

Without knowing the secret key, anybody may convert an encryption of a message into an encryption of any (efficient) function of that message using a fully homomorphic encryption (FHE) approach. We describe a multilevel FHE method that is entirely based on the (normal) premise of learning with mistakes. The greatest assessment depth of levelled FHE schemes is constrained from the start. However, by assuming "poor circular security," this requirement can be lifted.) The security of our technique is based on the worst-case difficulty of "short vector problems" on arbitrary lattices, using established findings. In two ways, our

construction improves on past efforts: 1. Using a novel relinearization methodology, we show that "somewhat homomorphic" encryption may be built on.

[14] Croman, K., et al.: On scaling decentralized blockchains. In: Clark, J., Meiklejohn, S., Ryan, P.Y.A., Wallach, D., Brenner, M., Rohloff, K. (eds.) FC 2016. LNCS, vol. 9604, pp. 106–125. Springer, Heidelberg (2016).

With the growing popularity of blockchain-based cryptocurrencies, scalability has become a major worry. We look at how inherent and situational limitations in Bitcoin's present peer-to-peer overlay network hinder its capacity to handle significantly larger throughputs and reduced latencies. Our findings show that adjusting block size and intervals is simply the first step in developing next-generation, high-load blockchain protocols, and that considerable advancements would need a fundamental rethinking of technological techniques. For such techniques, we provide a structured viewpoint on the design space. We list and quickly explore a number of previously presented protocol proposals, as well as suggest some new ideas and open challenges, from this perspective.

[11] Buterin, V. A Next-Generation Smart Contract and Decentralized Application Platform. 1 May 2018.

The implementation and usage of smart contracts in businesses need a holistic approach. This review explains how this technology is now being used, as well as the issues that are preventing it from being used in modern organisations. This paper presents a comprehensive evaluation of prior research that illustrates the use of smart contracts in businesses, including frameworks, methodology, functioning prototypes, and simulations. This article focuses on determining the existing state and use of smart-contract technology in a company. While much work is being made in building technology that supports smart contracts, little is known about how they are used in businesses. We identify features of smart-contract applications in several fields of modern businesses in this study. We go on to break down and categorise the hurdles and issues that are preventing smart-contract implementation.

[13]A. Azaria, A. Ekblaw, T. Vieira and A. Lippman, "Medrec: Using blockchain for medical data access and permission management", *Proceedings of 2nd International Conference on Open and Big Data*, pp. 25-30, 2016.

Years of heavy regulation and bureaucratic inefficiency have slowed innovation for electronic medical records (EMRs). We now face a critical need for such innovation, as personalization and data science prompt patients to engage in the details of their healthcare and restore agency over their medical data. In this paper, we propose MedRec: a novel, decentralized record management system to handle EMRs, using blockchain technology. Our system gives patients a comprehensive, immutable log and easy access to their medical information across providers and treatment sites. Leveraging unique blockchain properties, MedRec manages authentication, confidentiality, accountability and data sharing- crucial considerations when handling sensitive information. A modular design integrates with providers' existing, local data storage solutions, facilitating interoperability and making our system convenient and adaptable. We incentivize medical stakeholders (researchers, public health authorities, etc.) to participate in the network as blockchain "miners". This provides them with access to aggregate, anonymized data as mining rewards, in return for sustaining and securing the network via Proof of Work. MedRec thus enables the emergence of data economics, supplying big data to empower researchers while engaging patients and providers in the choice to release metadata. The purpose of this short paper is to expose, prior to field tests, a working prototype through which we analyze and discuss our approach..

[2] S. Velliangiri and P. Karthikeyan, "Blockchain Technology: Challenges and Security issues in Consensus algorithm", 2020 International Conference on Computer Communication and Informatics (ICCCI -2020), Jan. 22 – 24, 2020.

Blockchain is picking up footing and one of the omnipresent themes these days can be named. Despite the fact that faultfinders are challenging its durability, safety and stability, it has officially changed the way of life of numerous people in a few territories due to its overwhelming impact on projects, moreover, organizations. Given that the highlights of blockchain innovation ensure more solid and convenient administrations, it is imperative to think about the security furthermore, protection concerns and complications behind the inventive innovation. The range of blockchain applications extends to open and public services including health, financial, automotive, Internet of Things (IoT) and risk management. Many talk about using the blockchain data system for various applications in the spotlight. Nonetheless, a comprehensive review on the point of view of specialized applications and applications has not yet been an expert. In this paper, we seek to perform a thorough study on the technology of blockchain by analyzing its architecture of varying consensus algorithms over and above challenges and opportunities for security and data protection within blockchains.

[10] B. Singhal, G. Dhameja, and P. S. Panda, "How Blockchain Works," in Beginning Blockchain, pp. 31–148, Berkeley, CA: Apress, 2018.

Blockchain is the new wave of disruption that has already started to redesign business, social and political interactions, and any other way of value exchange. Again, it is not just a change, but a rapid phenomenon that is already in motion. As of this writing, more than 40 top financial institutions and many different firms across industries have started to explore Blockchain to lower transaction cost, speed up transaction time, reduce the risk of fraud, and eliminate the middleman or intermediary services. Some are trying to reimagine the legacy systems and services to take them to a new level and also come up with new kinds of service offerings.

2.2 Conclusions On Survey

The idea of adapting digital voting systems to make the public electoral process cheaper, faster and easier, is a compelling one in modern society. Making the electoral process cheap and quick, normalizes it in the eyes of the voters, removes a certain power barrier between the voter and the elected official and puts a certain amount of pressure on the elected official. It also opens the door for a more direct form of democracy, allowing voters to express their will on individual bills and propositions. In this system, we introduced a unique, blockchain-based electronic voting system that utilizes smart contracts to enable secure and cost efficient elections while guaranteeing voters privacy. We have outlined the systems architecture, the design, and a security analysis of the system. By comparison to previous work, we have shown that the blockchain technology offers a new possibility for democratic countries to advance from the pen and paper election scheme, to a more cost and time-efficient election scheme, while increasing the security measures of today's scheme and offering new possibilities of transparency. Using a private blockchain, it is possible to send hundreds of transactions per second onto the blockchain, utilizing every aspect of the smart contract to ease the load on the blockchain. For countries of greater size, some measures must be taken to withhold greater throughput of transactions per second, for example the parent & child architecture which reduces the number of transactions stored on the blockchain at a 1:100 ratio without compromising the network's security. Our election scheme allows individual voters to vote at a voting district of their choosing while guaranteeing that each individual voter's vote is counted from the correct district, which could potentially increase voter turnout.

3.Software and Hardware Requirements

3.1 Software Requirements

- Operating System : Windows XP.
- Platform : PYTHON TECHNOLOGY
- Tool : Python 3.5
- Front End : HTML,CSS
- Back End : python script

3.2 Hardware Requirements

- System : INTEL i3.
- Hard Disk : 512 GB.
- Ram : 4 GB.

4.Software Development Analysis

4.1 Overview Of Problem

In every democracy, the security of an election is a matter of national security. The computer security field has for a decade studied the possibilities of electronic voting systems, with the goal of minimizing the cost of having a national election, while fulfilling and increasing the security conditions of an election. From the dawn of democratically elected candidates, the voting system has been based on pen and paper. Replacing the traditional pen and paper scheme with a new election system is critical to limit fraud and having the voting process traceable and verifiable. Electronic voting machines have been viewed as flawed, by the security community, primarily based on physical security concerns. Anyone with physical access to such a machine can sabotage the machine, thereby affecting all votes cast on the aforementioned machine.Enter blockchain technology. A blockchain is a distributed, immutable, incontrovertible, public ledger.

4.2 Define The Problem

Anonymous vote-casting: Each vote may or may not contain any choice per candidate, should be anonymous to everyone including the system administrators, after the vote is submitted through the system. Individualized ballot processes: How a vote is depicted within the involving net applications or databases continues to be AN open discussion. whereas a transparent text message is the worst plan, a hashed token is wont to offer obscurity and integrity. Meanwhile, the vote ought to be non-reputable, that can't be bonded by the token resolution. Ballot casting verifiability by (and only by) the voter. The elector ought to be ready to see and verify his/her own vote, when he/she submitted the vote. This is often vital to realize so as to forestall, or a minimum of to note, any potential malicious activity. This counter live, except for providing suggestions that of non-repudiation, can surely boost the sensation of trust of the voters. These issues are partly self-addressed in some recent applications.High initial setup costs: Though sustaining and maintaining on-line selection systems is way cheaper than ancient elections, initial deployments could be pricey, particularly for businesses. Increasing security problems: Cyber attacks cause an excellent threat to the general public polls. Nobody would settle for the responsibility if an associate degree hacking try succeeds throughout an election. Lack of transparency and trust: How can people surely trust the results, when everything is done online? Perceptual problems cannot be ignored. Voting delays or inefficiencies related to remote voting. Timing is very important in voting schemes; technical capabilities and the infrastructures should be reliable and run at the highest possible performance to let remote voting be synchronous.

4.3 Modules Overview

In this project we are using the public python Blockchain API to store and manage voting data as Blockchain provides secure and tamper proof of data storage and to implement this project we have designed following modules.

1)Admin Module

2)User Module

4.4 Define the Modules

Admin module: This user is responsible to add new party and candidate details and can view party details and vote count.

User Module: This user has to sign up with the application by using username as his ID and then upload his face photo which capture from the webcam. After registering a user can go for a login which validates user id and after successful login user can go for cast vote module.

4.5 Module Functionality

The cast vote module has the following functionality:

1) First the user will be connected to his PC webcam and then the image will be captured.

2) Using Python face recognition API the picture captured by the webcam is compared against the user's profile image to validate the user.

3) If a user does not casted vote then the user can give a vote to the desired candidate by clicking a link beside party name or candidate name.

4) Upon giving a vote application will capture voter and candidate details and then encrypt the data and then store it in Blockchain.

5.Project System Design

5.1 UML DIAGRAMS

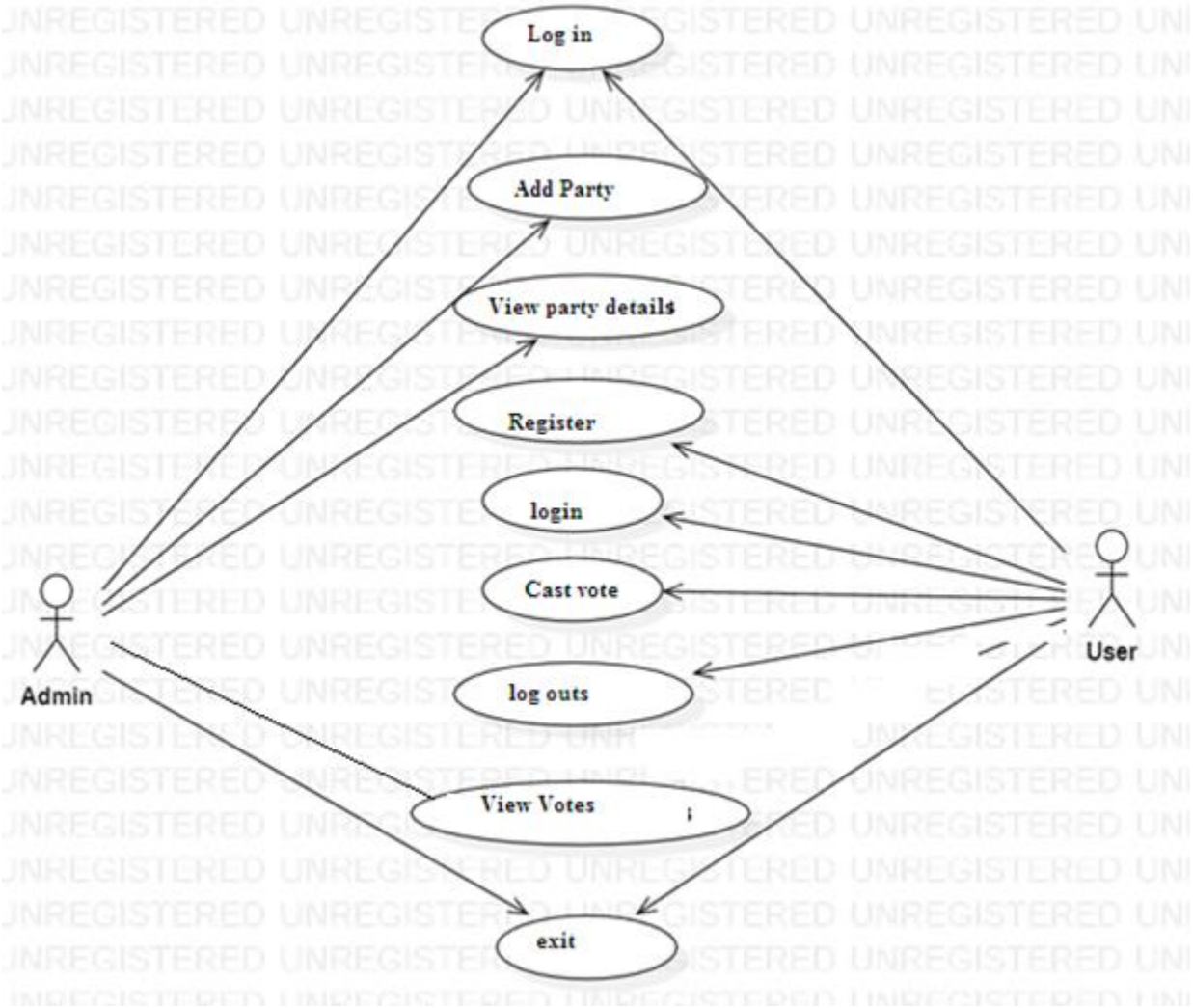


Fig-5.1: Use Case Diagram

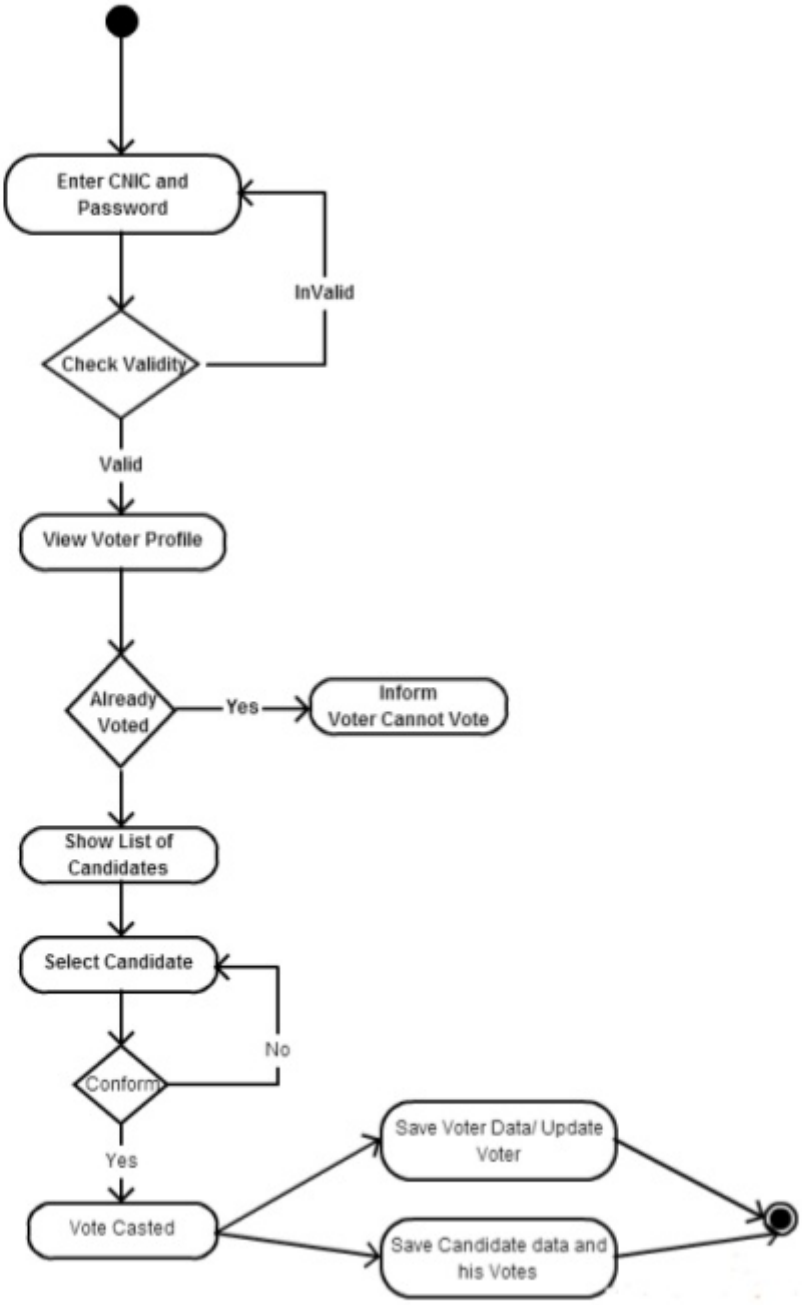


Fig-5.2:User Activity Diagram

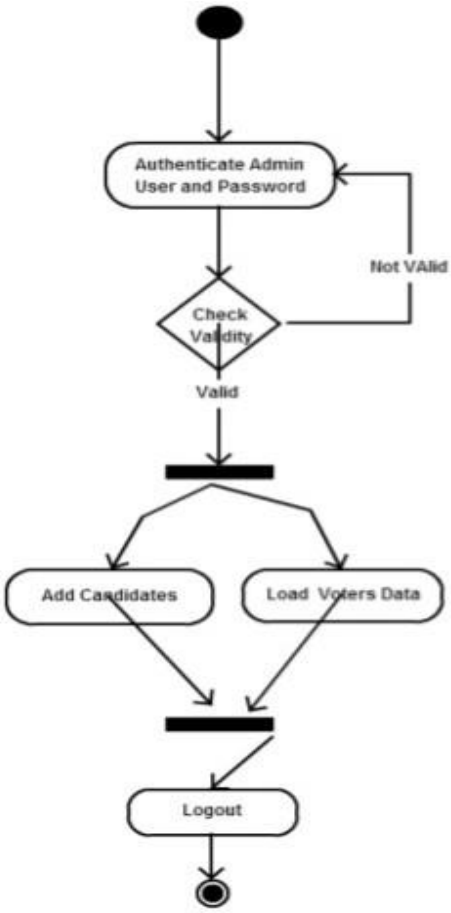


Fig-5.3: Admin Activity Diagram

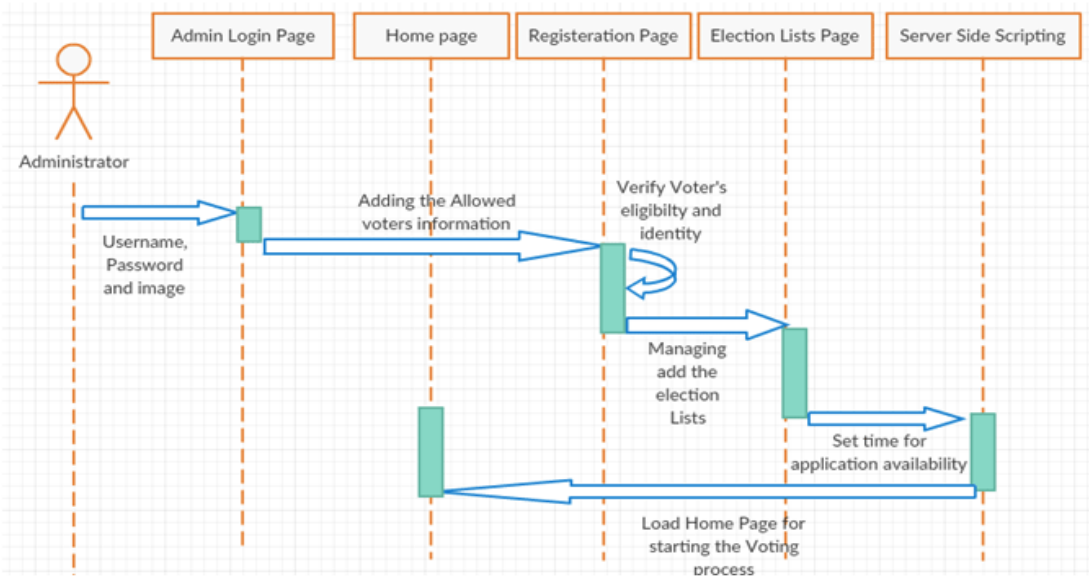


Fig-5.4: Sequence Diagram

Fig-5.5: Class Diagram

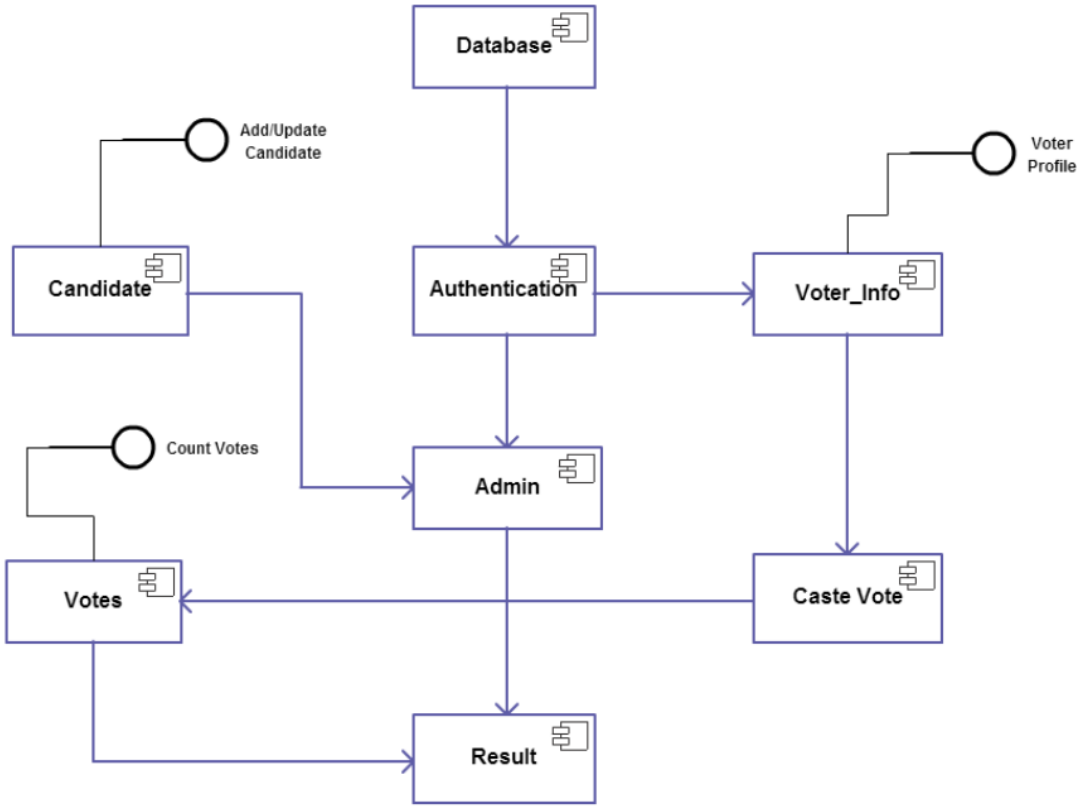


Fig 5.6 Component Diagram

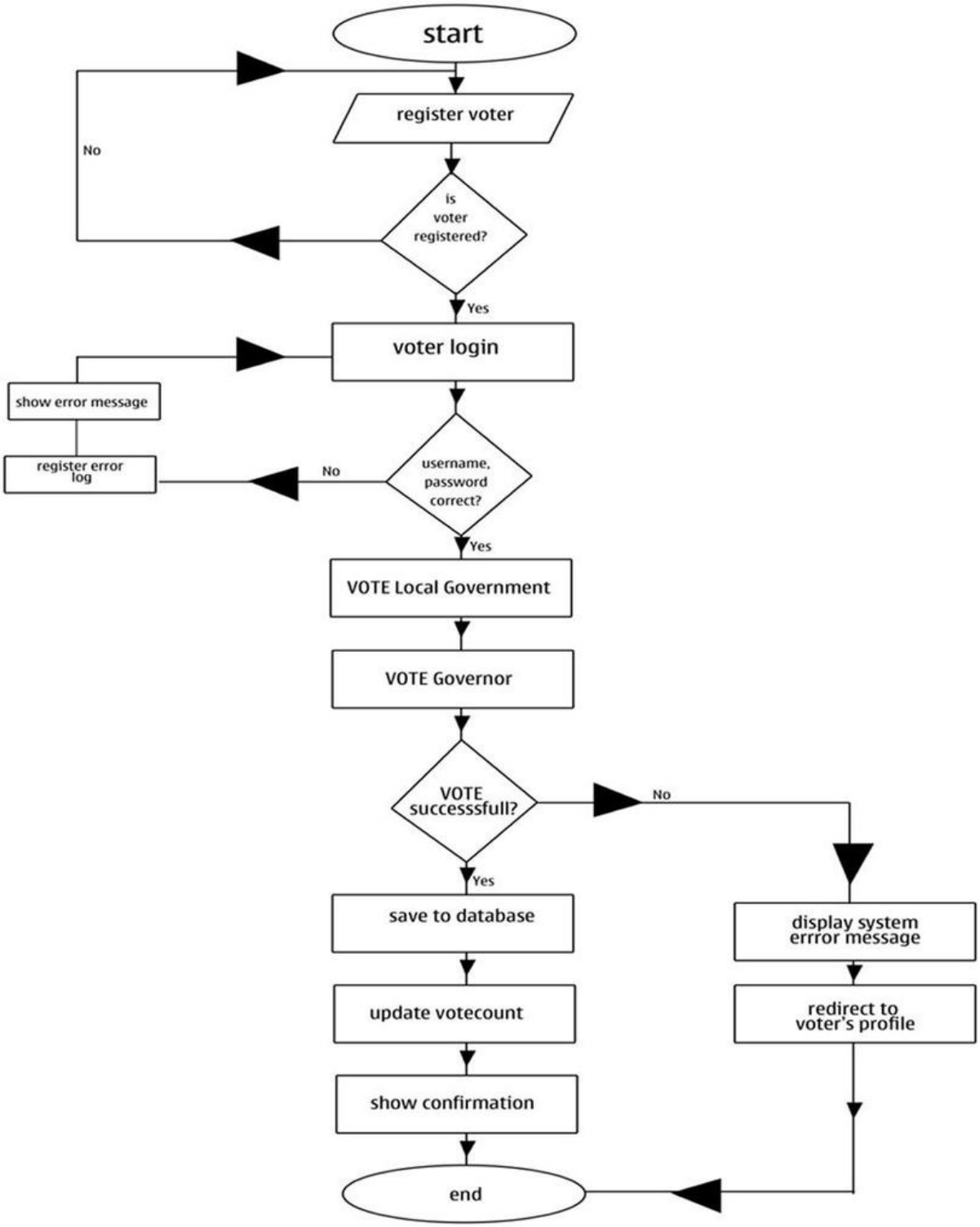


Fig-5.7: State Chart Diagram

6.Project Coding

6.1 CODE TEMPLATES

```
def ValidateUser(request):
    if request.method == 'POST':
        global load_model
        global classifier
        user = ''
        with open("session.txt", "r") as file:
            for line in file:
                user = line.strip('\n')
        file.close()
        option = 0
        status = "unable to predict user"
        img = cv2.imread('C:/Python/EVoting/test.png')
        faces =
face_detection.detectMultiScale(img,scaleFactor=1.1,minNeighbors=5,minSize=(30,30),flags=cv2.C
ASCADE_SCALE_IMAGE)
        print("==================="+str(len(faces)))
        if len(faces) > 0:
            flag = checkUser(user)
            picture_of_me = face_recognition.load_image_file("test.png")
            my_face_encoding = face_recognition.face_encodings(picture_of_me)[0]
            # my_face_encoding now contains a universal 'encoding' of my facial features that can be
compared to any other picture of a face!
            userimg=user+".png"
            unknown_picture = face_recognition.load_image_file(userimg)
            unknown_face_encoding = face_recognition.face_encodings(unknown_picture)[0]
```

```
        # Now we can see the two face encodings are of the same person with `compare_faces`!
        results = face_recognition.compare_faces([my_face_encoding], unknown_face_encoding)
        flag2=results[0]
        if flag == 0:
            if flag2:
                status = "User Validated as "+user
                option = 1
            else:
    status = "<h3><font size='3' color='black'>Unable to predict. Please retry"
        else:
            status = "<font size='3' color='black'>You already casted your vote"
    else:
        status = "<font size='3' color='black'>unable to detect face"
    if option == 1:
        output = getOutput(status)
        context= {'data':output}
        return render(request, 'VotePage.html', context)
    else:
        context= {'data':status}
        return render(request, 'UserScreen.html', context)
```

6.2 OUTLINE FOR VARIOUS FILES

HTML FILES:

- AddParty.html
- AdminScreen.html
- Login.html
- UserScreen.html
- ViewVotes.html

- index.html
- Admin.html
- CastVote.html
- Register.html
- ViewParty.html
- VotePage.html

PYTHON FILES:

- admin.py
- apps.py
- models.py
- tests.py
- urls.py
- views.py

6.3 CLASS WITH FUNCTIONALITY

Creating Block:

```
class Block:
    def __init__(self, index, transactions, timestamp, previous_hash):
        self.index = index
        self.transactions = transactions
        self.timestamp = timestamp
        self.previous_hash = previous_hash
        self.nonce = 0

    def compute_hash(self):
  block_string = json.dumps(self.__dict__, sort_keys=True)
        return sha256(block_string.encode()).hexdigest()
```

Creating Blockchain

```
class Blockchain:
    # difficulty of our PoW algorithm
    difficulty = 2 #using difficulty 2 computation

    def __init__(self):
        self.unconfirmed_transactions = []
        self.chain = []
        self.create_genesis_block()
        self.peer = []
        self.translist = []

    def create_genesis_block(self): #create genesis block
        genesis_block = Block(0, [], time.time(), "0")
        genesis_block.hash = genesis_block.compute_hash()
        self.chain.append(genesis_block)

    @property
    def last_block(self):
        return self.chain[-1]

    def add_block(self, block, proof): #adding data to block by computing new and previous hashes
        previous_hash = self.last_block.hash

        if previous_hash != block.previous_hash:
            return False
```

```
        if not self.is_valid_proof(block, proof):
            return False

        block.hash = proof
        #print("main "+str(block.hash))
        self.chain.append(block)
        return True

    def is_valid_proof(self, block, block_hash): #proof of work
        return (block_hash.startswith('0' * Blockchain.difficulty) and block_hash ==
block.compute_hash())

    def proof_of_work(self, block): #proof of work
        block.nonce = 0

        computed_hash = block.compute_hash()
        while not computed_hash.startswith('0' * Blockchain.difficulty):
            block.nonce += 1
            computed_hash = block.compute_hash()

        return computed_hash

    def add_new_transaction(self, transaction):
        self.unconfirmed_transactions.append(transaction)

    def addPeer(self, peer_details):
        self.peer.append(peer_details)
```

```
def addTransaction(self,trans_details): #add transaction
    self.translist.append(trans_details)

def mine(self):#mine transaction
    if not self.unconfirmed_transactions:
        return False

    last_block = self.last_block

    new_block = Block(index=last_block.index + 1,
                      transactions=self.unconfirmed_transactions,
                      timestamp=time.time(),
                      previous_hash=last_block.hash)

    proof = self.proof_of_work(new_block)
    self.add_block(new_block, proof)

    self.unconfirmed_transactions = []
    return new_block.index

def save_object(self,obj, filename):
    with open(filename, 'wb') as output:
        pickle.dump(obj, output, pickle.HIGHEST_PROTOCOL)
```

6.4 METHODS INPUT AND OUTPUT PARAMETERS

User Registration Page:

```
<table align="center" width="80" >

                    <tr><td><font size="3" color="black">Username</b></td><td><input type="text" name="username" style="font-family: Comic Sans MS" size="30"/></td></tr>

            <tr><td><font size="3" color="black">Password</b></td><td><input type="password" name="password" style="font-family: Comic Sans MS" size="30"/></td></tr>

            <tr><td><font size="3" color="black">Contact No</b></td><td><input type="text" name="contact" style="font-family: Comic Sans MS" size="20"/></td></tr>

            <tr><td><font size="3" color="black">Email ID</b></td><td><input type="text" name="email" style="font-family: Comic Sans MS" size="40"/></td></tr>

            <tr><td><font size="3" color="black">Address</b></td><td><input type="text" name="address" style="font-family: Comic Sans MS" size="60"/></td></tr>

            <tr><td><font size="3" color="black">Profile Image</b></td><td><input type="file" name="image" style="font-family: Comic Sans MS" size="60"/></td></tr>

                    <tr><td></td><td><input type="submit" value="Register">

                    </td>

                    </table>
```

User Login Page:

```
<table align="center" width="80" >

                    <tr><td><font size="3" color="black">Username</b></td><td><input type="text" name="username" style="font-family: Comic Sans MS" size="30"/></td></tr>

            <tr><td><font size="3" color="black">Password</b></td><td><input type="password" name="password" style="font-family: Comic Sans MS" size="30"/></td></tr>

            <tr><td></td><td><input type="submit" value="Login">

            </td>

</table>
```

Add Candidate:

```
<form name="f1" method="post" action={% url 'AddPartyAction' %} enctype="multipart/form-data" OnSubmit="return validate()">

        {% csrf_token %}<br/>

  <h3><b>Add Party Candidate Screen</b></h3>

        <font size="" color="black"><center>{{ data }}</center></font>

                                                <table align="center" width="80" >

                        <tr><td><font size="3" color="black">Candidate Name</b></td><td><input type="text" name="t1" style="font-family: Comic Sans MS" size="30"/></td></tr>



                <tr><td><font size="3" color="black">Party Name</b></td><td><select name="t2">

                <option value="Congress">Congress</option>

                <option value="BJP">BJP</option>

                </select>

                </td></tr>

                <tr><td><font size="3" color="black">Area Name</b></td><td><input type="text" name="t3" style="font-family: Comic Sans MS" size="20"/></td></tr>

                <tr><td><font size="3" color="black">Profile Image</b></td><td><input type="file" name="t4" style="font-family: Comic Sans MS" size="60"/></td></tr>

                        <tr><td></td><td><input type="submit" value="Add Party">

                        </td>

                </table>
```

View Candidates:

```
def ViewParty(request):

    if request.method == 'GET':

        output = '<table border=1 align=center>'
```

```
        output+='<tr><th><font size=3 color=black>Candidate Name</font></th>'
        output+='<th><font size=3 color=black>Party Name</font></th>'
        output+='<th><font size=3 color=black>Area Name</font></th>'
        output+='<th><font size=3 color=black>Image</font></th>'
        con = pymysql.connect(host='127.0.0.1',port = 3306,user = 'root', password = '', database = 'evoting',charset='utf8')
        with con:
            cur = con.cursor()
            cur.execute("select * FROM addparty")
            rows = cur.fetchall()
            for row in rows:
                cname = row[0]
                pname = str(row[1])
                area = row[2]
                image = row[3]
                output+='<tr><td><font size=3 color=black>'+cname+'</font></td>'
                output+='<td><font size=3 color=black>'+pname+'</font></td>'
                output+='<td><font size=3 color=black>'+area+'</font></td>'
                output+='<td><img src=/static/profiles/'+cname+'.png width=200 height=200></img></td></tr>'
        output+="</table><br/><br/><br/><br/><br/><br/>"
        context= {'data':output}
        return render(request, 'ViewParty.html', context)
```

View Votes:

```
def ViewVotes(request):
    if request.method == 'GET':
        output = '<table border=1 align=center>'
```

```
    output+='<tr><th><font size=3 color=black>Candidate Name</font></th>'
    output+='<th><font size=3 color=black>Party Name</font></th>'
    output+='<th><font size=3 color=black>Area Name</font></th>'
    output+='<th><font size=3 color=black>Image</font></th>'
    output+='<th><font size=3 color=black>Vote Count</font></th>'
    con = pymysql.connect(host='127.0.0.1',port = 3306,user = 'root', password = '', database = 'evoting',charset='utf8')
    with con:
        cur = con.cursor()
        cur.execute("select * FROM addparty")
        rows = cur.fetchall()
        for row in rows:
            cname = row[0]
            count = getCount(cname)
            pname = str(row[1])
            area = row[2]
            image = row[3]
            output+='<tr><td><font size=3 color=black>'+cname+'</font></td>'
            output+='<td><font size=3 color=black>'+pname+'</font></td>'
            output+='<td><font size=3 color=black>'+area+'</font></td>'
            output+='<td><img src=/static/profiles/'+cname+'.png width=200 height=200></img></td>'
            output+='<td><font size=3 color=black>'+str(count)+'</font></td></tr>'
    output+="</table><br/><br/><br/><br/><br/>"
    context= {'data':output}
    return render(request, 'ViewVotes.html', context)
```

7. PROJECT TESTING

7.1 VARIOUS TEST CASES

1)Unit testing: Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .It is done after the completion of an individual unit before integration. This is a structural testing that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

2)Integration testing: Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfactory, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

3)Functional testing: Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items:

- Valid Input : identified classes of valid input must be accepted.
- Invalid Input : identified classes of invalid input must be rejected.
- Functions : identified functions must be exercised.
- Output : identified classes of application outputs must be exercised.
- Systems/Procedures : interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identifying Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

4)System Testing: System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration-oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

7.2 BLACK BOX TESTING

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a test in which the software under test is treated as a black box. You cannot "see" into it. The test provides inputs and responds to outputs without considering how the software works.

7.3 WHITE BOX TESTING

White Box Testing is a testing in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It has a purpose. It is used to test areas that cannot be reached from a black box level.

8.OUTPUT SCREENS

8.1 USER INTERFACES

Fig-8.1: Home Page

Fig-8.2: Admin Home Page

Fig-8.3: User Home Page

Fig-8.4: Adding Candidate Details

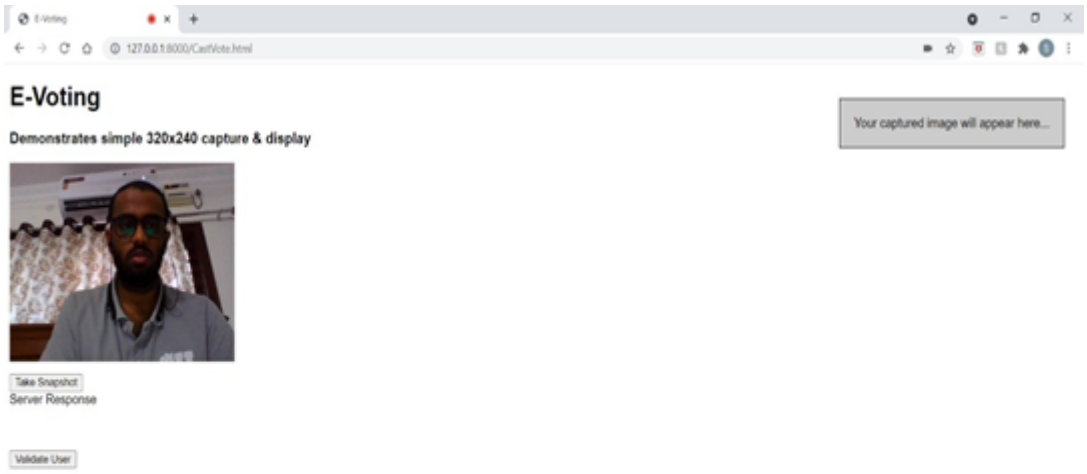


Fig-8.5: Validation of User

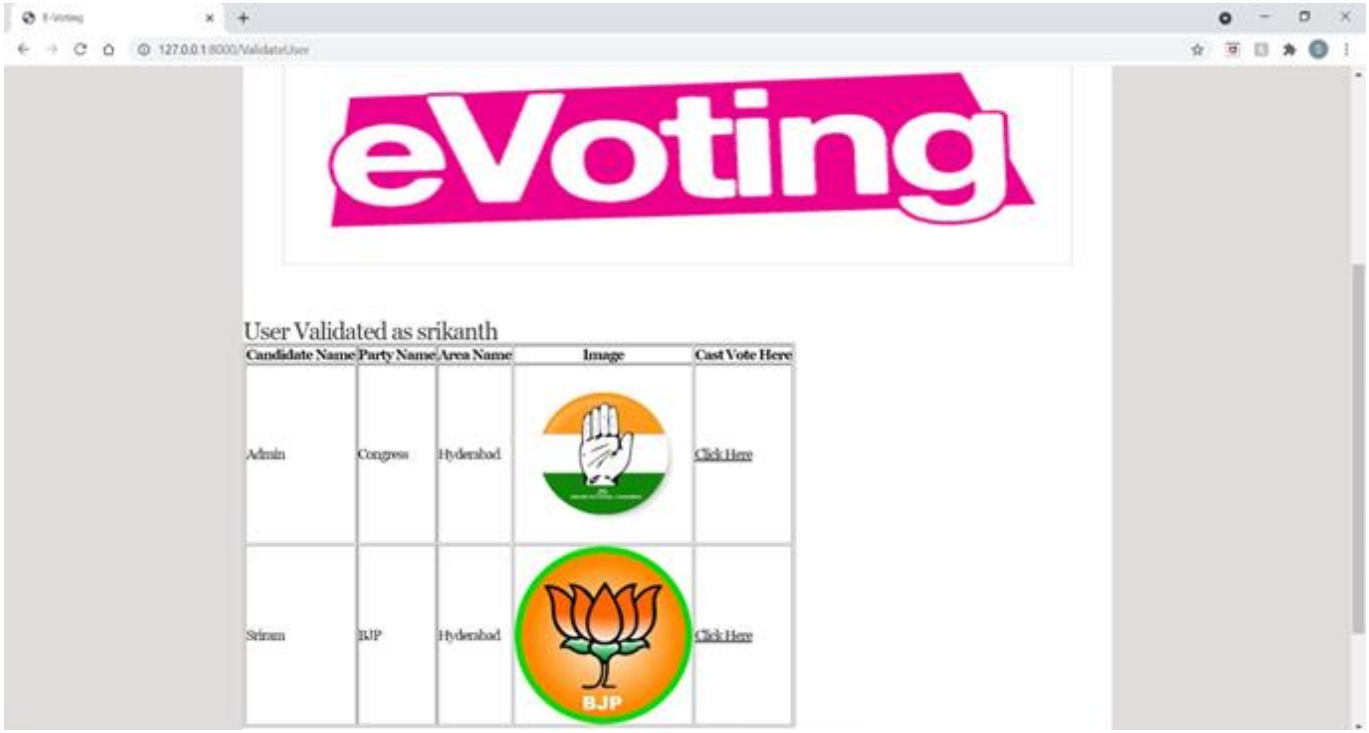


Fig-8.6: Casting Vote

8.2 OUTPUT SCREENS:

Fig-8.7: View Candidate Details

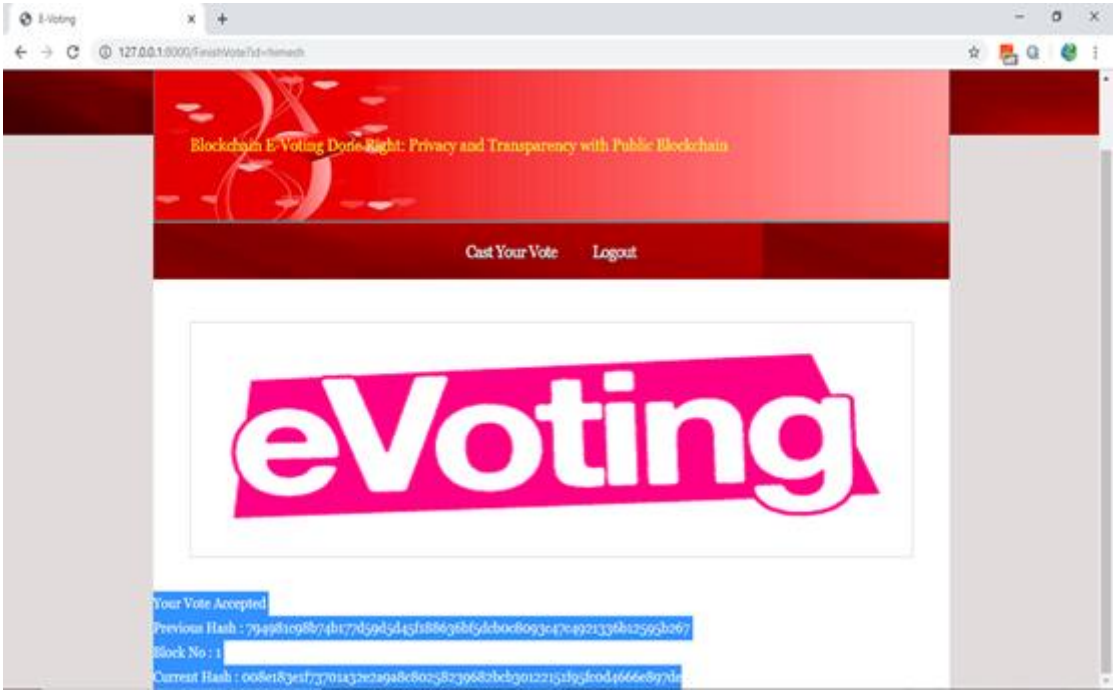


Fig-8.8: Vote Accepted

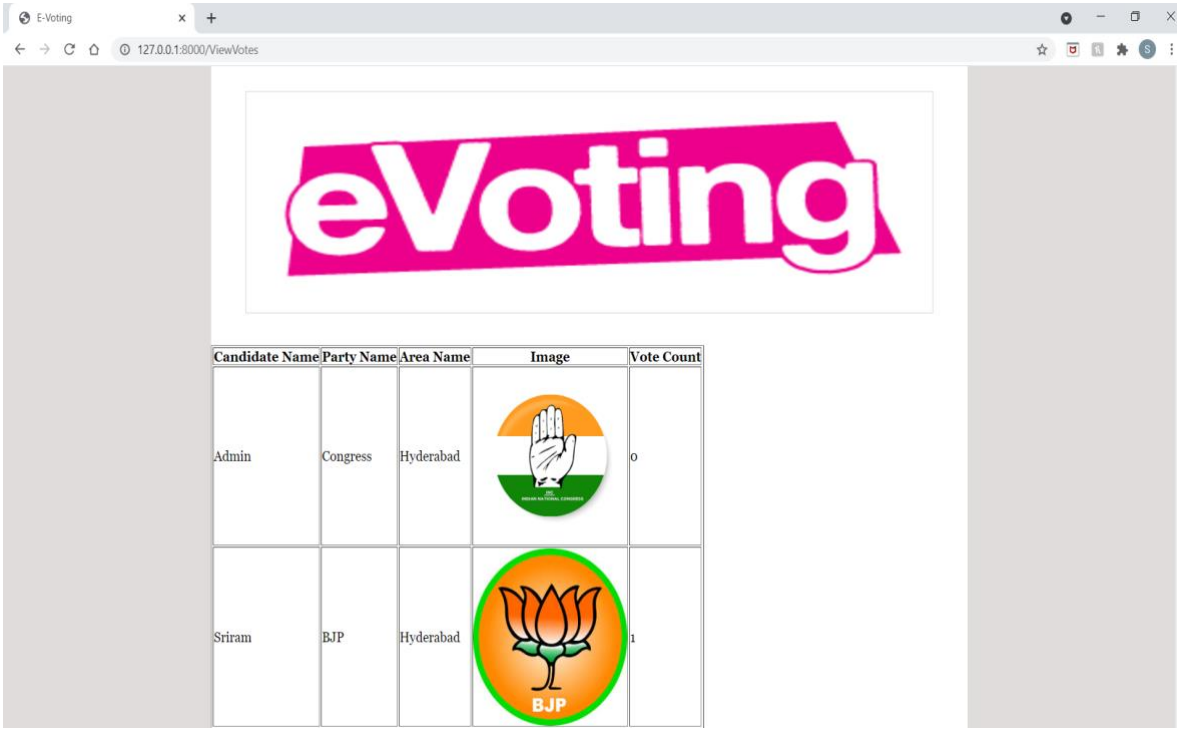


Fig-8.9: View Vote Count

9. EXPERIMENTAL RESULTS

In this project, we are storing and managing voting data using public python Blockchain APIs since Blockchain enables safe and tamper-proof data storage. There is no reason to have faith in the centralised body that oversaw the elections. In our system, this authority has no influence on election results. Although the data is completely visible, the identity of voters is protected through homomorphic encryption.

We introduced a one-of-a-kind, blockchain-based electronic voting system that uses smart contracts to ensure secure and cost-effective elections while protecting voters' privacy. The system's architecture, design, and security analysis have all been detailed. By comparing our findings to previous research, we have demonstrated that blockchain technology provides a new opportunity for democratic countries to move away from the pen and paper election system and toward a more cost- and time-effective election system, while also improving security and transparency.Our election system allows individual voters to vote in the voting district of their choice while ensuring that their votes are counted from the right district, potentially increasing voter turnout.

10. CONCLUSION AND FUTURE ENHANCEMENT

In this project we introduced a blockchain-based electronic voting system that utilizes smart contracts to enable secure and cost-efficient election while guaranteeing voters privacy. By comparison to previous work, we have shown that the blockchain technology offers a new possibility for democratic countries to advance from the pen and paper election scheme, to a more cost- and time- efficient election scheme. Using the above technologies our project has achieved transparency, immutability and decentralization.

Although there are minor variations in network delays, they are so minor that a public block chain has greater benefits in such an election system owing to its openness of data and the fact that anybody can see them in real time.OTP or Advanced Biometrics can be used to ensure more secure user authentication. Technologies such as Blockchain and Decentralization are growing at an exponential rate, and the time between the actual usage of Blockchain in general elections by nations and now gives a good chance to enhance the suggested model and create a more resilient system.

REFERENCES

[1] S Muralidhara, B A Usha, "Review of Blockchain Security and Privacy", Computing Methodologies and Communication (ICCMC) 2021 5th International Conference on, pp. 526-533, 2021.

[2] S. Velliangiri and P. Karthikeyan, "Blockchain Technology: Challenges and Security issues in Consensus algorithm", 2020 International Conference on Computer Communication and Informatics (ICCCI -2020), Jan. 22 – 24, 2020.

[3] B Vivekanadam, "Analysis of Recent Trend and Applications in BlockChain Technology ", Journal of ISMAC, vol. 2, no. 04, pp. 200-206, 2020.

[4] Pan, X., Pan, X., Song, M., Ai, B., & Ming, Y. (2020). Blockchain technology and enterprise operational capabilities: An empirical test. International Journal of Information Management, 52, 101946.

[5] M. Pawlak, J. Guziur, and I. Boniszewska-Miranda, "Voting Process with Blockchain Technology: Auditable Blockchain Voting System," in Lecture Notes on Data Engineering and Communications Technologies, pp. 233–244, Springer, Cham, 2019.

[6] Agora, "Agora Whitepaper," 2018.

[7] N. Kshetri and J. Voas, "Blockchain-Enabled E-Voting," IEEE Software, vol. 35, pp. 95–99, jul 2018.

[8] S. Landers, "Netvote: A Decentralized Voting Platform - Netvote Project - Medium," 2018.

[9] R. Perper, "Sierra Leone is the first country to use blockchain during an election - Business Insider," 2018.

[10] B. Singhal, G. Dhameja, and P. S. Panda, "How Blockchain Works," in Beginning Blockchain, pp. 31–148, Berkeley, CA: Apress, 2018.

[11]Buterin, V. A Next-Generation Smart Contract and Decentralized Application Platform. 1 May 2018.

[12] P. McCorry, S. F. Shahandashti, and F. Hao, "A Smart Contract for Boardroom Voting with Maximum Voter Privacy," in Lecture Notes in Computer Science, ch. FCDS, pp. 357–375, Springer, Cham, 2017.

[13]A. Azaria, A. Ekblaw, T. Vieira and A. Lippman, "Medrec: Using blockchain for medical data access and permission management", *Proceedings of 2nd International Conference on Open and Big Data*, pp. 25-30, 2016.

[14] Croman, K., et al.: On scaling decentralized blockchains. In: Clark, J., Meiklejohn, S., Ryan, P.Y.A., Wallach, D., Brenner, M., Rohloff, K. (eds.) FC 2016. LNCS, vol. 9604, pp. 106–125. Springer, Heidelberg (2016).

[15] Z. Brakerski and V. Vaikuntanathan, “Efficient Fully Homomorphic Encryption from (Standard) LWE,” SIAM Journal on Computing, vol. 43, pp. 831–871, jan 2014.

PUBLICATIONS

Electronic Voting System using Public Blockchain with Privacy, Transparency and Security

Kothapalli Saisriram1, Ramayanam Sai Veer Suvesith2, Swathi Tomar3, Tummaluru Srikanth Reddy4

P.Sabitha5

1,2,3,4 UG Scholar, 5Assistant Professor

Department of Computer Science and Engineering,

St. Martin's Engineering College,

Near Forest Academy, Dulapally, Kompally, Secunderabad, Telangana 500 014, India

E-Mail:saisriram8921@gmail.com1, suvesith.ramayanam@gmail.com2, swathitomar04@gmail.com3
tsrikanth2511@gmail.com4, psabithacse@smec.ac.in5

Abstract: With the advanced technology and developments since the 20th century, a new procedure of casting votes in an election is developed every now and then. This project uses advanced technology like block chain and homomorphic encryption in order to make the election more safe and secure. By implementing the idea of block chain e-voting the elections can be made more fair, as it double checks the votes casted by the voters before and after the elections. Moreover, it eliminates the chances of malpractices as images of voters are taken into consideration. Hence, a voter can only vote once and can recheck their vote.

At present the voting is done using paper ballots and electronic voting but it has problems mainly regarding security, credibility, transparency, reliability, and functionality. So, block chain e-voting can deliver an answer to all these problems and further can add advantages like immutability and decentralization.

KEYWORDS: blockchain: immutability: decentralization: transparency: homomorphic encryption

1.INTRODUCTION

There are a number of people and parties complaining about the unfair election process and the violence at the booth, blockchain e-voting is a solution to this. In this project we are using the public python Blockchain API to store and manage voting data as Blockchain provides secure and tamper proof of data storage. The admin is responsible to add new party and candidate details and can view party details and vote count. The user has to sign up with the application by using username as his ID and then upload his face photo which

is captured from the webcam. After registration, the user can go for a login which validates user id and after successful login user can go for cast vote module.

The main advantage is that there is no need for confidence in the centralized authority that created the elections. This authority cannot affect the election results in our system. After the start of the voting, the platform behaves as fully independent and decentralized without possibilities to affect the voting process. The data are fully transparent, but the identity of voters is secured by homomorphic encryption.

2. Literature Survey

We demonstrate the use of the Blockchain to construct a decentralised and self-tallying internet voting mechanism with maximal voter anonymity. The Open Vote Network is a smart contract for Ethereum that is appropriate for boardroom elections. This is the first implementation of Blockchain e-voting that does not rely on any trusted authority to calculate the tally or safeguard the voter's privacy, unlike prior suggested Blockchain e-voting protocols. The Open Vote Network, on the other hand, is a self-tallying protocol in which each voter has complete control over the privacy of their own vote, which can only be violated by a full collusion involving all other voters. The protocol's execution is governed by the same consensus process that protects the Ethereum blockchain. To illustrate its viability, we put the implementation through its paces on Ethereum's official test network. We also give a financial and computational analysis of the cost of execution.

[7] P. McCorry, S. F. Shahandashti, and F. Hao, “A Smart Contract for Boardroom Voting with Maximum Voter Privacy,” in Lecture Notes in Computer Science, ch. FCDS, pp. 357–375, Springer, Cham, 2017.

Without knowing the secret key, anybody may convert an encryption of a message into an encryption of any (efficient) function of that message using a fully homomorphic encryption (FHE) approach. We describe a multilevel FHE method that is entirely based on the (normal) premise of learning with mistakes. The greatest assessment depth of levelled FHE schemes is constrained from the start. However, by assuming "poor circular security," this requirement can be lifted.) The security of our technique is based on the worst-case difficulty of "short vector problems" on arbitrary lattices, using established findings. In two ways, our construction improves on past efforts: 1. Using a novel relinearization methodology, we show that “somewhat homomorphic” encryption may be built on.

[12] Z. Brakerski and V. Vaikuntanathan, “Efficient Fully Homomorphic Encryption from (Standard) LWE,” SIAM Journal on Computing, vol. 43, pp. 831–871, jan 2014.

With the growing popularity of blockchain-based cryptocurrencies, scalability has become a major worry. We look at how inherent and situational limitations in Bitcoin's present peer-to-peer overlay network hinder its capacity to handle significantly larger throughputs and reduced latencies. Our findings show that adjusting block size and intervals is simply the first step in developing next-generation, high-load blockchain protocols, and that considerable advancements would need a fundamental rethinking of technological techniques. For such techniques, we provide a structured viewpoint on the design space. We list and quickly explore a number of previously presented protocol proposals, as well as suggest some new ideas and open challenges, from this perspective.

[10] Croman, K., et al.: On scaling decentralized blockchains. In: Clark, J., Meiklejohn, S., Ryan, P.Y.A., Wallach, D., Brenner, M., Rohloff, K. (eds.) FC 2016. LNCS, vol. 9604, pp. 106–125. Springer, Heidelberg (2016).

The implementation and usage of smart contracts in businesses need a holistic approach. This review explains how this technology is now being used, as well as the issues that are preventing it from being used in modern organisations. This paper presents a comprehensive evaluation of prior research that illustrate the use of smart contracts in businesses, including frameworks, methodology, functioning prototypes, and simulations. This article focuses on determining the existing state and use of smart-contract technology in a company. While much work is being made in building technology that supports smart contracts, little is known about how they are used in businesses. We identify features of smart-contract applications in several fields of modern businesses in this study. We go on to break down and categorise the hurdles and issues that are preventing smart-contract implementation.

[4] Buterin, V. A Next-Generation Smart Contract and Decentralized Application Platform. 1 May 2018.

3. Proposed Methodology

A.Block chain

The blockchain component represents the whole infrastructure for data storage and voting. A public blockchain, such as Ethereum, or a private blockchain, such as Hyperledger, can both be used to build a blockchain. The public blockchain's benefits include the fact that it makes all transaction and block information available to all users.

This assurance is expressed in the perspective of a typical user who isn't tech-savvy and wants to see everything. The private blockchain can provide the same level of trust, but an organisation must demonstrate it with data. The suggested design does not impose any limitations on the types of blockchains

that can be employed. Both kinds of blockchain have the ability to deliver the same level of trust. The blockchain organisation chooses the platform to be used.

B.Smart Contract

The voting's security is based on blockchain, and the processing is handled by a smart contract that is also a part of the blockchain. After a setup, the smart contract is deployed to the blockchain network. There are times, candidates, and other attributes in the configuration. The candidate does not have to be a person; it can be anything that embodies the election's goal. The published smart contract can't be updated or updated, which adds to the voting's transparency. A list of users who are qualified to vote is included in the smart contract. The access list must adhere to a key distribution procedure, which is carried out by the key authority. The data is encrypted using homomorphic encryption and stored in a blockchain. To make the encrypted election results, the smart contract must contain the public key of homomorphic encryption. The private key is kept separate and is used to view the results after the polls have closed. This key can be provided to numerous parties that are in charge of evaluating votes. Nobody understands what's going on since the outcomes aren't revealed until the end. We employed Zero-knowledge proof to ensure that the votes had valid values, such that a voter may only cast one vote and cannot cast two or more.

C.Homomorphic Encryption

Homomorphic encryptions allow complex mathematical operations to be performed on encrypted data without compromising the encryption. This represents the translation of one data collection into another while maintaining relationships between components in both sets in mathematics. The term "similar structure" is taken from Greek terms. Because the data in a homomorphic encryption method has the same structure, equivalent mathematical operations on encrypted or decrypted data will produce the same results.

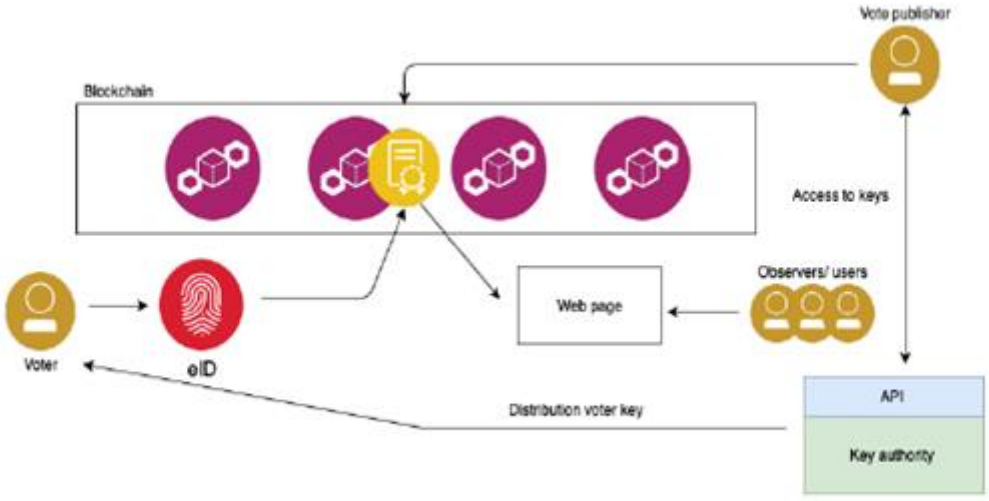


Fig-1: Architecture Diagram

4. Result and Discussion

The suggested blockchain voting system takes into account all voting needs and is suitable for any election, such as president, student parliament, and so on. The technology allows for multiple rounds of voting and, ideally, makes use of a public blockchain. Other types of blockchain can be used in place of the public blockchain, but the stored data (votes) must be easily verifiable by anybody. We identify three important roles in our proposed system: vote publisher, key authority, and voter. These three characters can indicate a firm, an organisation, or a user. Because the functions of vote publisher and key authority can be combined into one, they can belong to the same organisation or person. Depending on the vote arrangement, the voter attends the elections. The vote publisher is in charge of voting configuration, which is included in the smart contract. Before publishing the smart contract, the vote publisher must have all of the cipher keys. It is necessary for the vote publisher and the key authority to work closely together. All cipher keys are created and distributed by the key authority to a voter and vote publisher. The distribution method must be secure and not subject to unauthorised access.

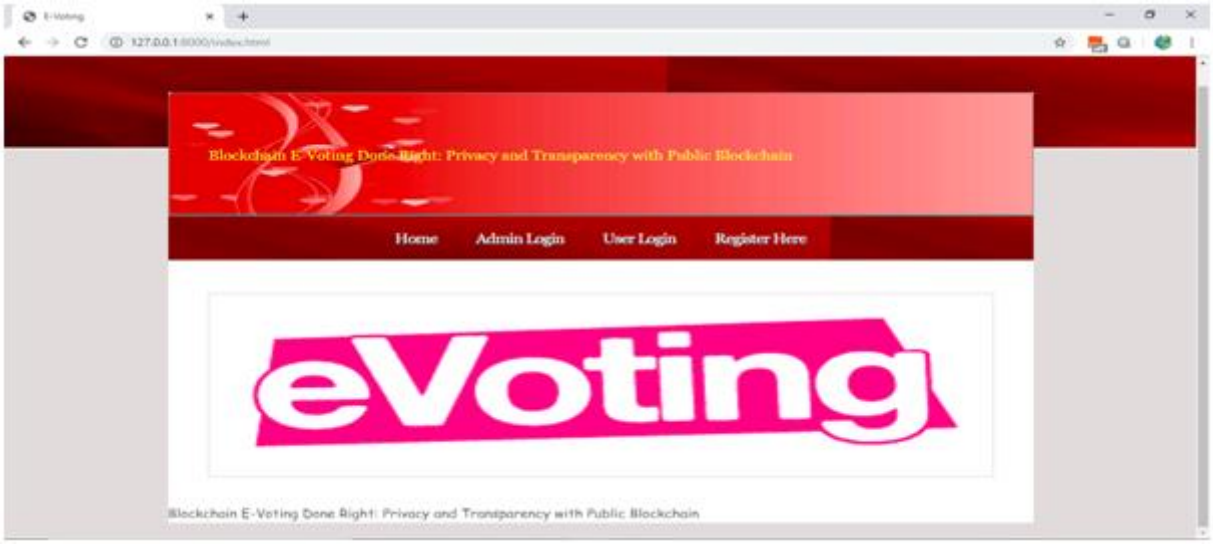


Fig-2: Home Page

Fig-3: Admin Module

The Home page consists of Admin Login, User Login and New Registration modules. After logging into the admin account successfully we can find admin home page modules such as Add Party Details, View Party Details, View Votes and Logout.

Fig-4: New User registration

Fig-5: User Homepage

The user needs to provide his basic details for registering his account. After filling the details the user account will be created successfully after submitting the details required. After registering the user can go for login with valid credentials and once after successful login user can go for cast your vote module to cast a vote and the user will be validated before casting his vote.

Fig-6: Vote accepted

In the above screen as this is the first vote so a block will be added to Blockchain with block No as 1 and we can see Blockchain created a chain of blocks with previous and current hash code validation.

5. Conclusion

In this paper, we proposed an electronic voting mechanism based on blockchain that uses intelligent contracts to allow safe and cost effective voting while ensuring privacy for voters. It has been seen that blockchain technology provides a new opportunity to resolve restrictions and obstacles in adopting electronic voting systems that guarantee the protection and fairness of elections and set the foundation for transparency. Though we see minor changes in network time, they are so insignificant that the distributed blockchain in such an election environment has more benefits because the results are transparent and can be viewed by anyone in real time. A private blockchain is a little quick, but decreases the integrity of the whole scheme by a limited centralization, since it just operates when and where the administrators want it. This eliminates all the malpractices and will develop a secure voting system which is fast and is equally efficient. It increases the security and credibility of the election process.

6. References

[1] M. Pawlak, J. Guziur, and I. Boniszewska-Miranda, “Voting Process with Blockchain Technology: Auditable Blockchain Voting System,” in Lecture Notes on Data Engineering and Communications Technologies, pp. 233–244, Springer, Cham, 2019.

[2] B. Singhal, G. Dhameja, and P. S. Panda, “How Blockchain Works,” in Beginning Blockchain, pp. 31–148, Berkeley, CA: Apress, 2018.

[3] N. Kshetri and J. Voas, “Blockchain-Enabled E-Voting,” IEEE Software, vol. 35, pp. 95–99, jul 2018.

[4] Buterin, V. A Next-Generation Smart Contract and Decentralized Application Platform. 1 May 2018.

[5] B. Singhal, G. Dhameja, and P. S. Panda, “How Blockchain Works,” in Beginning Blockchain, pp. 31–148, Berkeley, CA: Apress, 2018.

[6] R. Perper, “Sierra Leone is the first country to use blockchain during an election - Business Insider,” 2018.

[7] P. McCorry, S. F. Shahandashti, and F. Hao, “A Smart Contract for Boardroom Voting with Maximum Voter Privacy,” in Lecture Notes in Computer Science, ch. FCDS, pp. 357–375, Springer, Cham, 2017.

[8] A. Azaria, A. Ekblaw, T. Vieira and A. Lippman, "Medrec: Using blockchain for medical data access and permission management", Proceedings of 2nd International Conference on Open and Big Data, pp. 25-30, 2016.

[9] K. Bhargavan, A. Delignat-Lavaud, C. Fournet, A. Gollamudi, N. Kobeissi and S. Zanella-Béguelin, "Formal verification of smart contracts: Short paper", Proceedings of the 2016 ACM Workshop on Programming Languages and Analysis for Security, pp. 91-96, 2016

[10] Croman, K., et al.: On scaling decentralized blockchains. In: Clark, J., Meiklejohn, S., Ryan, P.Y.A., Wallach, D., Brenner, M., Rohloff, K. (eds.) FC 2016. LNCS, vol. 9604, pp. 106–125. Springer, Heidelberg (2016).

[11] G. Wood et al., "Ethereum: A secure decentralised generalised transaction ledger," Ethereum project yellow paper, vol. 151, pp. 1–32, 2014.

[12] Z. Brakerski and V. Vaikuntanathan, "Efficient Fully Homomorphic Encryption from (Standard) LWE," SIAM Journal on Computing, vol. 43, pp. 831–871, jan 2014.

[13] O. Goldreich and Y. Oren, "Definitions and properties of zero-knowledge proof systems," Journal of Cryptology, vol. 7, no. 1, pp. 1–32, 1994.

[14] Aiello, W., and J. Hastad, Perfect Zero-Knowledge Languages Can Be Recognized in Two Rounds, Proc., pp. 439–448, 1987.

[15]Goldreich, O., S. Micali, and A. Wigderson, Proof that Yield Nothing but their Validity and a Methodology of Cryptographic Protocol Design," Proc. of 27th Symposium on Foundations of Computer Science, pp 174{187, Toronto, 1986.}

ALL FOUR STUDENTS' ONE PAGE PROFILE

Kothapalli Saisriram is currently pursuing her Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. He completed his intermediate from Sri Chaitanya Junior College and 10th class from Geetanjali Talent School. His technical skills include C, C++, Python. He also has a basic understanding of Java. He took part in Employability Skill development Program conducted by Zensar. He is a student of Smart Interviews. He did an Internship as Programming Trainer in C++ at GCS Technologies. His responsibilities include delivering content and helping students clearing their doubts in programming. He is an active member of Rotaract Club of new age Engineers. He is also a member of Cyber Security Club in our College. His participations include: National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020, one day Seminar on "Career Conclave on Future Technologies" which was conducted by EmergeX on 5th July 2020, two-day workshop on "Machine Learning" which was conducted at Indian Institute of Technology Hyderabad (IITH) on 14th and 15th February 2020, Leadership Talk conducted by MHRD's Innovation Cell on 16th May 2020. His areas of interest are Python, Data Science, Machine Learning and BlockChain. He completed few certification courses from online platforms like Coursera, CursaApp and SoloLearn.

Ramayanam Sai Veer Suvesith is currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. He completed his intermediate from Sri Chaitanya Junior College and 10th class from St.Peter's High School. He is one of the members of Cyber Security Club in our college. His technical skills include C, Python and Java. He also has a basic understanding of C++. He took part in Employability Skill development Program conducted by Zensar. He is also a student of Smart Interviews . His participations include: National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020, HTML and CSS Workshop of TAM event held from 5th January 2018 to 3rd February 2018. He is also a member of Rotaract Club of New Age Engineers. He is also an active participant in coding contests held by different platforms such as HackerRank, Codechef, and Codeforces. His areas of interest are Python and Full Stack Projects. He completed few certification courses from online platforms like Coursera, CursaApp and Solo Learn.

Swathi Tomar is currently pursuing her Bachelor of Technology specialized in the stream of Computer Science and Engineering at St. Martin's Engineering College. She completed her intermediate from Kendriya Vidyalaya (AFS), Hakimpet and 10th class from Kendriya Vidyalaya-2, Jodhpur. Her technical skills include Python, HTML, SQL and Java. She also has a basic understanding of C, C++. She has received a NPTEL certificate in Discrete Mathematics from IIT Madras (2019). She took part in Employability Skill development Program conducted by Zensar. She is also a student of Smart Interviews. She is a member of Rotaract club of new age engineers. Her participations include: National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020, Technovation (2018) Participation and show casting the project model in national level project expo and competition, Inter college basketball tournament (2018). Her areas of interest are Python, Artificial Intelligence, Machine Learning and Deep Learning. She completed few certification courses from online platforms like Coursera, Cursa, Udemy and Guvi.

Tummaluru Srikanth Reddy is currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. He completed his intermediate from Sri Chaitanya Junior College and 10th class from Rao's My Techno School. His technical skills include Python, Java and Html. He also has a basic understanding of C++. He took part in Employability Skill development Program conducted by Zensar. He is also a student of Smart Interviews. His participations include: National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020, HTML and CSS Workshop of TAM event held from 5th January 2018 to 3rd February 2018, National Level Project Expo and Competition "TECHNOVATION - 2018" Organized by Department of Mechanical Engineering & Computer Science and Engineering on 28th March, 2018, He Showcased His project "Sports Events Management Website" in this Competition. He is also an active participant in coding contests held by different platforms such as HackerRank, Codechef, and Codeforces. His areas of interest are Python, Full Stack Projects and Competitive Programming. He completed few certification courses from online platforms like Coursera, CursaApp and Internshala.

APPENDICES

```
from django.shortcuts import render

from django.template import RequestContext

from django.contrib import messages

import pymysql

from django.http import HttpResponse

from django.core.files.storage import FileSystemStorage

import os

from keras.utils.np_utils import to_categorical

from keras.layers import  MaxPooling2D

from keras.layers import Dense, Dropout, Activation, Flatten

from keras.layers import Convolution2D

from keras.models import Sequential

from keras.models import model_from_json

from Blockchain import *

from Block import *

import face_recognition

from datetime import date

import cv2

import numpy as np

import pyaes, pbkdf2, binascii, os, secrets
```

```
import base64

global load_model

load_model = 0

global classifier

blockchain = Blockchain()

if os.path.exists('blockchain_contract.txt'):

    with open('blockchain_contract.txt', 'rb') as fileinput:

        blockchain = pickle.load(fileinput)

    fileinput.close()

face_detection = cv2.CascadeClassifier('haarcascade_frontalface_default.xml')

def getKey(): #generating key with PBKDF2 for AES

    password = "s3cr3t*c0d3"

    passwordSalt = '76895'

    key = pbkdf2.PBKDF2(password, passwordSalt).read(32)

    return key

def encrypt(plaintext): #AES data encryption

    aes = pyaes.AESModeOfOperationCTR(getKey(),
pyaes.Counter(311295470350000473029524339676541953981242398445663228841721636378460
56248223))

    ciphertext = aes.encrypt(plaintext)

    return ciphertext

def decrypt(enc): #AES data decryption
```

```
    aes = pyaes.AESModeOfOperationCTR(getKey(),
pyaes.Counter(311295470350000473029524339676541953981242398445663228841721636378460562
48223))
    decrypted = aes.decrypt(enc)
    return decrypted
def AddParty(request):
    if request.method == 'GET':
        return render(request, 'AddParty.html', {})
def index(request):
    if request.method == 'GET':
        return render(request, 'index.html', {})
def Login(request):
    if request.method == 'GET':
        return render(request, 'Login.html', {})
def CastVote(request):
    if request.method == 'GET':
        return render(request, 'CastVote.html', {})
def Register(request):
    if request.method == 'GET':
        return render(request, 'Register.html', {})
def Admin(request):
    if request.method == 'GET':
```

```
        return render(request, 'Admin.html', {})
def WebCam(request):
    if request.method == 'GET':
        data = str(request)
        formats, imgstr = data.split(';base64,')
        imgstr = imgstr[0:(len(imgstr)-2)]
        data = base64.b64decode(imgstr)
        with open('C:/Python/EVoting/test.png', 'wb') as f:
            f.write(data)
        f.close()
        context= {'data':"done"}
        return HttpResponse("Image saved")
def checkUser(name):
    flag = 0
    for i in range(len(blockchain.chain)):
        if i > 0:
            b = blockchain.chain[i]
            data = b.transactions[0]
            data = base64.b64decode(data)
            data = str(decrypt(data))
            data = data[2:len(data)-1]
```

```
            print(data)
            arr = data.split("#")
            if arr[0] == name:
                flag = 1
                break
    return flag
def getOutput(status):
    output = '<h3><br/>'+status+'<br/><table border=1 align=center>'
    output+='<tr><th><font size=3 color=black>Candidate Name</font></th>'
    output+='<th><font size=3 color=black>Party Name</font></th>'
    output+='<th><font size=3 color=black>Area Name</font></th>'
    output+='<th><font size=3 color=black>Image</font></th>'
    output+='<th><font size=3 color=black>Cast Vote Here</font></th></tr>'
    con = pymysql.connect(host='127.0.0.1',port = 3306,user = 'root', password = '', database =
'evoting',charset='utf8')
    with con:
        cur = con.cursor()
        cur.execute("select * FROM addparty")
        rows = cur.fetchall()
        for row in rows:
            cname = row[0]
            pname = str(row[1])
```

```
            area = row[2]
            image = row[3]
            output+='<tr><td><font size=3 color=black>'+cname+'</font></td>'
            output+='<td><font size=3 color=black>'+pname+'</font></td>'
            output+='<td><font size=3 color=black>'+area+'</font></td>'
            output+='<td><img src=/static/profiles/'+cname+'.png width=200 height=200></img></td>'
            output+='<td><a href=\'FinishVote?id='+cname+'\'><font size=3 color=black>Click Here</font></a></td></tr>'
    output+="</table><br/><br/><br/><br/><br/><br/>"
    return output
def FinishVote(request):
    if request.method == 'GET':
        cname = request.GET.get('id', False)
        voter = ''
        with open("session.txt", "r") as file:
            for line in file:
                user = line.strip('\n')
        file.close()
        today = date.today()
        data = str(user)+"#"+str(cname)+"#"+str(today)
        enc = encrypt(str(data))
        enc = str(base64.b64encode(enc),'utf-8')
```

```
        blockchain.add_new_transaction(enc)
        hash = blockchain.mine()
        b = blockchain.chain[len(blockchain.chain)-1]
        print("Previous Hash : "+str(b.previous_hash)+" Block No : "+str(b.index)+" Current Hash : "+str(b.hash))
        bc = "Previous Hash : "+str(b.previous_hash)+"<br/>Block No : "+str(b.index)+"<br/>Current Hash : "+str(b.hash)
        blockchain.save_object(blockchain,'blockchain_contract.txt')
        context= {'data':'<font size=3 color=black>Your Vote Accepted<br/>'+bc}
        return render(request, 'UserScreen.html', context)
def getCount(name):
    count = 0
    for i in range(len(blockchain.chain)):
        if i > 0:
            b = blockchain.chain[i]
            data = b.transactions[0]
            data = base64.b64decode(data)
            data = str(decrypt(data))
            data = data[2:len(data)-1]
            arr = data.split("#")
            if arr[1] == name:
                count = count + 1
```

```
            break
    return count
def ViewVotes(request):
    if request.method == 'GET':
        output = '<table border=1 align=center>'
        output+='<tr><th><font size=3 color=black>Candidate Name</font></th>'
        output+='<th><font size=3 color=black>Party Name</font></th>'
        output+='<th><font size=3 color=black>Area Name</font></th>'
        output+='<th><font size=3 color=black>Image</font></th>'
        output+='<th><font size=3 color=black>Vote Count</font></th>'
        con = pymysql.connect(host='127.0.0.1',port = 3306,user = 'root', password = '', database =
'evoting',charset='utf8')
        with con:
            cur = con.cursor()
            cur.execute("select * FROM addparty")
            rows = cur.fetchall()
            for row in rows:
                cname = row[0]
                count = getCount(cname)
                pname = str(row[1])
                area = row[2]
                image = row[3]
```

```
            output+='<tr><td><font size=3 color=black>'+cname+'</font></td>'
            output+='<td><font size=3 color=black>'+pname+'</font></td>'
            output+='<td><font size=3 color=black>'+area+'</font></td>'
            output+='<td><img src=/static/profiles/'+cname+'.png width=200 height=200></img></td>'
            output+='<td><font size=3 color=black>'+str(count)+'</font></td></tr>'
        output+="</table><br/><br/><br/><br/><br/>"
        context= {'data':output}
        return render(request, 'ViewVotes.html', context)
def ViewParty(request):
    if request.method == 'GET':
        output = '<table border=1 align=center>'
        output+='<tr><th><font size=3 color=black>Candidate Name</font></th>'
        output+='<th><font size=3 color=black>Party Name</font></th>'
        output+='<th><font size=3 color=black>Area Name</font></th>'
        output+='<th><font size=3 color=black>Image</font></th>'
        con = pymysql.connect(host='127.0.0.1',port = 3306,user = 'root', password = '', database =
'evoting',charset='utf8')
        with con:
            cur = con.cursor()
            cur.execute("select * FROM addparty")
            rows = cur.fetchall()
            for row in rows:
```

```
            cname = row[0]
            pname = str(row[1])
            area = row[2]
            image = row[3]
            output+='<tr><td><font size=3 color=black>'+cname+'</font></td>'
            output+='<td><font size=3 color=black>'+pname+'</font></td>'
            output+='<td><font size=3 color=black>'+area+'</font></td>'
            output+='<td><img src=/static/profiles/'+cname+'.png width=200
height=200></img></td></tr>'
        output+="</table><br/><br/><br/><br/><br/>"
        context= {'data':output}
        return render(request, 'ViewParty.html', context)
def AddPartyAction(request):
    if request.method == 'POST':
      cname = request.POST.get('t1', False)
      pname = request.POST.get('t2', False)
      area = request.POST.get('t3', False)
      myfile = request.FILES['t4']
      fs = FileSystemStorage()
      filename = fs.save('C:/Python/EVoting/EVotingApp/static/profiles/'+cname+'.png', myfile)
      db_connection = pymysql.connect(host='127.0.0.1',port = 3306,user = 'root', password = '', database
= 'evoting',charset='utf8')
```

```
    db_cursor = db_connection.cursor()
    student_sql_query = "INSERT INTO addparty(candidatename,partyname,areaname,image)
VALUES('"+cname+"','"+pname+"','"+area+"','"+cname+"')"
    db_cursor.execute(student_sql_query)
    db_connection.commit()
    print(db_cursor.rowcount, "Record Inserted")
    if db_cursor.rowcount == 1:
     context= {'data':'Party Details Added'}
     return render(request, 'AddParty.html', context)
    else:
     context= {'data':'Error in adding party details'}
     return render(request, 'AddParty.html', context)
def Signup(request):
   if request.method == 'POST':
    username = request.POST.get('username', False)
    password = request.POST.get('password', False)
    contact = request.POST.get('contact', False)
    email = request.POST.get('email', False)
    address = request.POST.get('address', False)
    myfile = request.FILES['image']
    fs = FileSystemStorage()
    filename = fs.save('C:/Python/EVoting/'+username+'.png', myfile)
```

```
    db_connection = pymysql.connect(host='127.0.0.1',port = 3306,user = 'root', password = '', database = 'evoting',charset='utf8')

    db_cursor = db_connection.cursor()

    student_sql_query = "INSERT INTO register(username,password,contact,email,address) VALUES('"+username+"','"+password+"','"+contact+"','"+email+"','"+address+"')"

    db_cursor.execute(student_sql_query)

    db_connection.commit()

    print(db_cursor.rowcount, "Record Inserted")

    if db_cursor.rowcount == 1:

     context= {'data':'Signup Process Completed'}

     return render(request, 'Register.html', context)

    else:

     context= {'data':'Error in signup process'}

     return render(request, 'Register.html', context)
def AdminLogin(request):
   if request.method == 'POST':

      username = request.POST.get('username', False)

      password = request.POST.get('password', False)

      if username == 'admin' and password == 'admin':

         file = open('session.txt','w')

         file.write(username)

         file.close()
```

```
            context= {'data':'Welcome '+username}
            return render(request, 'AdminScreen.html', context)
        if status == 'none':
            context= {'data':'Invalid login details'}
            return render(request, 'Admin.html', context)
def UserLogin(request):
    if request.method == 'POST':
        username = request.POST.get('username', False)
        password = request.POST.get('password', False)
        status = 'none'
        con = pymysql.connect(host='127.0.0.1',port = 3306,user = 'root', password = '', database =
'evoting',charset='utf8')
        with con:
            cur = con.cursor()
            cur.execute("select * FROM register")
            rows = cur.fetchall()
            for row in rows:
                if row[0] == username and row[1] == password:
                    status = 'success'
                    break
        if status == 'success':
            file = open('session.txt','w')
```

```
        file.write(username)
        file.close()
        context= {'data':'Welcome '+username}
        return render(request, 'UserScreen.html', context)
    if status == 'none':
        context= {'data':'Invalid login details'}
        return render(request, 'Login.html', context)
```

A

PROJECT REPORT

On

Vehicle Pattern Recognition using Machine & Deep Learning

Submitted by

1)Ms.Anupama (17K81A05C7) 2)Mr.Sairam (17K81A05D0)

3)Mr.Rithik Reddy (17K81A05E2)4)Mr.Manisyam(17K81A05F7)

in partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Under the Guidance of

Mrs.Swetha.P

Assistant Professor, Department of CSE

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ST.MARTIN'S ENGINEERING COLLEGE
An Autonomous Institute

Dhulapally, Secunderabad – 500 100

JUNE 2021

BONAFIDE CERTIFICATE

This is to certify that the project entitled Vehicle Pattern Recognition using Machine & Deep Learning, is being submitted by 1.Ms.Anupama 17K81A05C7, 2.Mr.Sairam 17K81A05D0, 3.Mr.Rithik Reddy 17K81A05E2, 4. Mr.Manisyam 17K81A05F3 in partial fulfillment of the requirement for the award of the degree of **BACHELOR OF TECHNOLOGY IN COMPUTER SCIENCE AND ENGINEERING** is recorded of bonafide work carried out by them. The result embodied in this report have been verified and found satisfactory.

Mrs.Swetha.P

Department of CSE

Head of the Department

Dr.M.NARAYANAN

Department of CSE

Internal Examiner

External Examiner

Place:

Date:

DECLARATION

We, the student of **Bachelor of Technology** in Department of Computer Science and Engineering', session: 2017 – 2021, St. Martin's Engineering College, Dhulapally, Kompally, Secunderabad, hereby declare that work presented in this Project Work entitled Vehicle Pattern Recognition using Machine & Deep Learning is the outcome of our own bonafide work and is correct to the best of our knowledge and this work has been undertaken taking care of Engineering Ethics. This result embodied in this project report has not been submitted in any university for award of any degree.

Anupama	17K81A05C7
Sairam	17K81A05D0
Rithik Reddy	17K81A05E2
Manisyam	17K81A05F3

ABSTRACT

Intelligent transportation systems have acknowledged a ration of attention in the last decades. In this area vehicle classification and localization is the key task. In this task the biggest challenge is to discriminate the features of different vehicles. Further, vehicle classification and detection is a hard problem to identify and locate because wide variety of vehicles don't follow the lane discipline. In this article, to identify and locate, we have created a convolution neural network from scratch to classify and detect objects using a modern convolution neural network based on fast regions. Vehicle type testing is an important part of intelligent transportation systems. Its function is to detect the type of vehicle and provide information for road monitoring and traffic planning. Vehicle type detection, as a key technology to construct video surveillance of traffic conditions, has long been widely concerned by researchers at home and abroad. In this project we have considered three types of vehicles like bus, car and bike for classification and detection. Our approach will use the entire image as input and create a bounding box with probability estimates of the feature classes as output. The results of the experiment have shown that the projected system can considerably improve the accuracy of the detection.

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible and whose encouragements and guidance have crowded effects with success.

We extended our deep sense of gratitude to Principal**,** **Dr. P. SANTOSH KUMAR PATRA**, St. Martin's Engineering College, Dhulapally, for permitting us to undertake this project.

We are also thankful to **Dr. M.NARAYANAN**, Head of the Department, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

We would like to thank **Dr. T. POONGOTHAI,** Department of Computer Science and Engineering (AI & ML) for her encouragement and insightful comments and as well as our project coordinators **Dr. B.RAJALINGAM**, Associate Professor and **Dr. N. SATHEESH**, Professor, in Department of Computer Science and Engineering for their valuable support.

We would like to express our sincere gratitude and indebtedness to our project supervisor **P.SWETHA**, Assistant Professor, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

Finally, we express thanks to all those who have helped us successfully to completing this project. Furthermore, we would like to thank our family and friends for their moral support and encouragement.

We express thanks to all those who have helped us in successfully completing the project.

Anupama	17K81A05C7
Sairam	17K81A05D0
Rithik Reddy	17K81A05E2
Manisyam	17K81A05F3

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF OUTPUT SCREENS

LIST OF ABBREVIATION

AVI	Audio Video Interlace
BMP	Bitmap
CPU	Central Processing Unit
GB	Giga Bytes
GUI	Graphical User Interface
SVM	Support Vector Machine
CNN	Convolutional Neural Networks
KNN	K-Nearest Neighbors
CMMR	Car Make And Model Recognition
R-CNN	Region Based Convolutional Neural Networks

1.INTRODUCTION

Machine learning is a subset of computer vision which is one of the computer world's fastest rising new developments. There are several applications that might be utilized for this concept such as self-driving cars, learning robots or a medical system that diagnose medical images. One application of machine learning and computer vision is CMMR. Modern cars represent a new age for mobility and means of primary transport for us on day-to-day basis. With other surveillance aspects of urban life taking leaps forward such as face recognition systems at airports and other security related places, similarly car identification is also considered as a step forward in advanced surveillance systems. From perspective of computer vision, car identification is considered as hierarchical identification based on make, model and the production year range for specific models in assembly.

With rapid consumer demand of new and unique cars models, for each production year has leads to cars manufacturing having very large quantities of varying shapes and sizes. This altogether yields appearance differences in unlimited poses that demand today’s car recognition algorithms to be very robust in terms of outside conditions, deformities and occlusions. In modern design language, cars tend to have distinctive properties such light styles, seats, configuration options and whether its sports model or economy model, all of which are recognizable from appearance. In comparison to a human face recognition, the car classification and recognition points at inferring that if the two cars belong to same make and model rather than each person having a unique face to identify, thus making this an interesting as well as less researched challenge with the primary focus on identifying model from a single image.

Till now, many methods have been proposed to read vehicle features for CMMR. The question of multiple theoretical stances on several methods has been discussed in this article and examined that how the recent and prominent researches have progressed. The three major approaches were either to perform CMMR with state of the art image classification algorithms by considering human feature engineering with localisation, or direct image recognition process with automatic feature engineering on the whole image without localization based on deep learning for detecting a known car in image individually based on its classification and lastly the fine-grained classification approach that combined feature engineering, localization automatically of sub-parts and yet their challenges faced so far in the state of the art.

1.1 PROJECT OVERVIEW

This project focuses attention towards the review of various applications and approaches in the field on image processing up to and including recent advancement of deep learning using convolutional neural networks that can be used as tools for tackling the obstacles of Car Make and Model Recognition (CMMR) in real-world environment images. Such algorithms for CMMR system are typically designed to detect specific features in images that used to be formed by feature engineering processes and are now being replaced with deep learning. The review consists of three types of algorithms. The first set explores the traditional methods that use feature extraction to localise cars in various applications and attempt to provide solution for recognizing car characteristics with feature matching over whole images in database. The next set under consideration was deep learning since it demonstrated promising results due to automatic feature engineering although still being an area under consistent research and improvement over the past few years. This paper refers to how the deep learning systems have contributed towards successful CMMR and not a comparison of deep learning architectures. The last section of this review is focused in fine-grained classification with deep learning. This is conducted especially considering the cars that are generally built up of many different parts and identifying them based on fine-grained parts from very recent researches and whether it is a viable method for attaining better overall classification accuracy score.

1.2 PROJECT OBJECTIVES

This project mainly focuses on the detection algorithm based on vehicle target apparent feature information, that is, detects and classifies the vehicle target in the actual traffic picture.Its main difficulty lies in the picture of the vehicle target will change due to lighting, angle of view and the interior of the vehicle

. This project mainly focuses on the detection algorithm based on vehicle target apparent feature information, that is, detects and classifies the vehicle target in the actual traffic video or picture. Its main difficulty lies in the picture or video frame of the vehicle target will change due to lighting, angle of view and the interior of the vehicle and other.

The use of deep convolutional networks (CNNs) has achieved amazing success in the field of vehicle object detection. CNNs have a strong ability to learn image features and can perform multiple related tasks, such as classification and bounding box regression.

The detection method can be generally divided into two categories. The image input to the convolutional network must be fixed-size, and the deeper structure of the network requires a long training time and consumes a large amount of storage memory

.

1.3 ORGANIZATION OF CHAPTERS

Besides the introduction, the thesis is organized in other six chapters as follows:

The review is made in the context of hand gesture recognition systems with a particular attention on those implementations that assess the scalability and performances or their implementations. Most of the related work is on convolution neural network, whereas a small part is on cloud solutions. It will be possible to notice that only a small subset of the literature actually focuses on the analysis of the systems in mass crises scenarios. Chapter 3, SOFTWARE AND HARDWARE REQUIREMENTS: this chapter discuss about the software and hardware required for the execution of the project. Chapter 4, SOFTWARE DEVELOPMENT ANALYASIS: this chapter explains the assumptions and technical specifications of the project. Chapter 5, PROJECT SYSTEM DESIGN: this chapter explains all the software development process with DFD and UML diagrams clearly. Chapter 6, PROJECT CODING: this chapter explains the design of the system, roles and responsibilities, as well as the requirements of a HGR management solution based on CNN. Chapter 7, PROJECT TESTING: this chapter explains various test cases to test the project working. Chapter 8, OUTPUT SCREENS: explains a step-by-step process of the project execution. Chapter 9, EXPERIMENTAL RESULTS: tests and results are shown and explained in this chapter. The results are analyzed in the context of the thesis project and followed by discussion on systems throughput and resiliency, as well as the approaches to testing and analysis. Chapter 10, CONCLUSION AND FUTURE ENHANCEMENT: the chapter ends the project with a short summary of the main concepts mentioned in the thesis as well as the relevant results.

2 LITERATURE SURVEY

Starting from the most famous classification algorithms that are still used today as building blocks of modern feature engineering tools, CMMR had been a challenge for many to achieve and the review begins with the most basic CMMR approaches to realise the progress that have been made today.

2.1 SURVEY ON BACKGROUND

Initially[1] Cheung & Chu (2008) proposed the methodology for CMMR is by defining interest points on cars for matching features in two images using ScaleInvariant Feature Transformation (SIFT) algorithm. The images being feature matched that represented by red lines where the points of interest were plotted on two images between test and training image and then based on geometry, the points that were in same location on the image, represented interest points on cars. The points were called inliers and the models in dataset matching to the most number of inliers that were same in test image inliers represented the classified car match. RANSAC model was used to determine symmetrical points in images that ensured that the points of interest belong to car. The disadvantage however, was that the recognition worked at same angle as dataset only.

According to [2]Chen et al., (2015), although SIFT approach is a common feature extraction algorithm but it can be slow for real-time and some vehicles also have similar shapes even if manufactured by different manufacturers thus leading to inaccuracies. Therefore, an enhancement to Speeded Up Robust Features (SURF) algorithm initially proposed by Bay, Tuytelaars & Gool (2008) was introduced as Symmetrical-SURF. This helped to form region of interest on the axis of highest symmetry in image and detect vehicles comfortably in noisy environments such as roads. A grid was formed within the bounding box, this extracted the features based on boxes inside grid and Support Vector Machine (SVM) classifier was used for feature classification. Each of the grid within the bounding box was used independently to represent certain features of the car which increased accuracy. The grid helped in a manner that of one part of car was hidden behind a person or another car, even then the remaining grids contributed to feature extraction and helped classifier in making identification of car.

Emami, Fathi & Raahemifar (2014) also proposed [3]CMMR from the back of the vehicle like Cheung & Chu (2008), but recently the trend had evolved to determine the car location in image symmetry from number plate rather than the entire car itself which was not accurate always due to mechanical deformities in some cases. The scale of car was predicted from number plate once detected. Using Hue Saturation Value (HSV) colour detection, the red colour was detected to represent taillight. The Region of Interest (ROI) was defined by taillights, badge and bumper. Location of number plate with reference to taillights determined the class of vehicle such as truck, car or van. The classification reduced choices of dataset resulting in faster classification.

Features of taillights such as orientation, height/width and equidistance were used with Sobel edges algorithm. The k-Nearest Neighbour (k-NN) classifier was used and performed 96.3% overall classification accuracy on 280 test images on multiple camera angles but suffered with 53.1% accuracy in night time.

Moreover, Yang (2013) proposed[4] an adaptive Harris corner detection and recognition method to identify car makes and models based on the front of the car. In this method, the number plate of car was considered as reference point to detect symmetrical features in image using Harris corner detection algorithm. This algorithm detected differences in pixels that translated to corners or edges of objects in given image. With an assumption that car logo is placed directly above the car number plate so the position of logo was estimated from refence of number plate and the Adaptive-Harris corner algorithm was developed to extract logo features which were to be later classified with SVM.

With adaptive method, the threshold of[5] Harris corner detection was increased or decreased based on light conditions in image, since too dark images would create low gradient pixels and cause unwanted corner detections. System utilised Graphics Processing Unit (GPU) for faster speeds and was tested on 1096 images consisting of 12 models moving at speeds of 20km/h and the overall classification accuracy determined was 99.5%. The limitation for this system was that it can only determine make of the vehicle from logo. Along with the different approaches commenced by various researchers on CMMR, Deep Learning had immensely improved accuracies due to GPU enabled training with NVidia CUDA capabilities since 2012. Until then the overall classification accuracies were ranging 70% on the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) and with deep learning the percentage increased. The following researches used deep learning architectures.

Henceforth, when AlexNet was proposed [6] by Krizhevshy, Sutskever & Hinton (2012), it broke all previous ILSVRC records which is basically an international image classification competition held annually. Till now all researches mentioned required manual feature engineering and definition which was difficult on large datasets such as ImageNet dataset which consists of 1.2 million images and 1000 classes which has today become a default dataset to benchmark and compare classification accuracies by many authors.

With [7] this advancement, Yang et al., (2015) realised the potential of deep learning and its application to CMMR. It was stated that deep learning needs huge dataset of images and the lack of huge training sets for deep learning hindered success. So, Yang et al., (2015) proposed a Large-Scale Car Dataset which was made to be sufficient for fine-grained classification also. The dataset when trained of deep learning architectures, was found to provide CMMR capabilities to AlexNet and similar networks with interesting findings. It was found through conducted tests that models trained with specific car parts and especially tail lights gave higher accuracies rather than training cars overall for classifications and during fine grained it was found that front and back-side poses were the most reliable angles for training networks. Verifications were made by

visualizing neuron activation in last fully connected layers of networks. Total of 44,481 images having 281 car models were used in testing models like AlexNet GoogLeNet and Overfeat for surveillance camera view. Front view was found to be best suited for such camera since 98% accuracy in various weather conditions was acheived.

Szegedy[8] et al., (2015) proposed a new model with a very deep architecture as target. The model was constructed with a network-in-network approach. It was designed to use maximum system resources and consisted of 22 layers as compared to 8 layers proposed by Krizhevsky, Sutskever & Hinton (2012) but used ReLU for activations. The norm was to increase number of layers in deep architectures to increase accuracy by in GoogLeNet, 1x1 convolution layers were added in each network and reduce the dimensionality of the network while increasing the width of network. The width of network is the number of units at a level in architecture. GoogLeNet saved computation by looking for low-level cues first like colour or texture of object and then form bounding box internally on the expected region with high probability of containing trained object and then only performing full convolution network to form ensemble of predictions. This saved from unnecessary computations. Model performed 6.7% top-5 error in ILSVRC and showed that it could be trained to perform CMMR.

2.2 CONCLUSIONS ON SURVEY

From the literature review conducted above it was observed that with each new research, more advancement has been made towards image recognition tasks. It was observed that initially the trend for CMMR was to localize the car in the image by using number plates or car taillights as reference points and estimating the car positions. Then image feature extraction was engineered by hand defined filters which was not always perfect and did not perform robust to multiple cars in an image.

Then Deep Learning was rediscovered due to GPU capabilities enhancing the training times that once took weeks to train now took days only. With fast deep learning capabilities, many new algorithms were discovered such as AlexNet and GoogLeNet. Although deep learning provided good image classification scores but from reviews it was seen that when used standalone, it suffered in multiple instances such as two or three cars in an image but such architectures became building blocks of modern systems and are still used today even in the latest of researches so they hold vital and critical contribution to the subject.

The latest work with fine-grained classification, pointed critical parts of cars body that helped to identify the car by its special characteristics. For instance, comparing the highest scoring localization method, R-CNN with fine grained method introduced by Krause et al., (2016), that even if detection is slightly inaccurate but fine grain still classifies correctly with localization because R-CNN had to treat the entire object to be detected in image as one bounding box hindering it differentiate in fine detail of objects like car bumper lights and shapes.

The only limitation in today's technology is that it requires pose normalisation that all the images of cars should have similar pose and must be aligned because the patch discrimination is performed based on geometric constraints on image.

In future researches, the limitation of alignment and pose has to be considered and methods of finding the most discriminative patches amongst all patches detected in different car models have to be researched without limiting the pose. For example, if different poses are used and there are no geometric constraints then the wheels of car in an individual image might be considered as a distinctive feature of the car but as a whole when applied to the entire dataset in unsupervised learning, all the images would have wheels and detected in all the cars and the patch would no longer be discriminative as a whole since all the cars would have it and that's why geometric constraints and similar poses are used in all the state of the arts.

3 SOFTWARE AND HARDWARE REQUIREMENTS

3.1 SOFTWARE REQUIREMENTS:

- **Operating system** : Windows 10.
- **Coding Language** : Python.
- **Front-End** : Python.
- **Designing** : Html,css,javascript.
- **Data Base** : MySQL.

3.2 HARDWARE REQUIREMENTS:

- **System** : Intel i3 or more.
- **Hard Disk** : 1 TB.
- **Mouse** : Optical Mouse.
- **Ram** : 8 GB.

PYTHON

Python is a **high-level, interpreted**, **interactive** and **object-oriented scripting language**. Python is designed to be highly readable. It uses English keywords frequently where as other languages use punctuation, and it has fewer syntactical constructions than other languages.

- **Python is Interpreted:** Python is processed at runtime by the interpreter. You do not need to compile your program before executing it. This is similar to PERL and PHP.
- **Python is Interactive:** You can actually sit at a Python prompt and interact with the interpreter directly to write your programs.
- **Python is Object-Oriented:** Python supports Object-Oriented style or technique of programming that encapsulates code within objects.
- **Python is a Beginner's Language:** Python is a great language for the beginner-level programmers and supports the development of a wide range of applications from simple text processing to WWW browsers to games.

History of Python

Python was developed by Guido van Rossum in the late eighties and early nineties at the National Research Institute for Mathematics and Computer Science in the Netherlands.

Python is derived from many other languages, including ABC, Modula-3, C, C++, Algol-68, SmallTalk, and Unix shell and other scripting languages.

Python is copyrighted. Like Perl, Python source code is now available under the GNU General Public License (GPL).

Python is now maintained by a core development team at the institute, although Guido van Rossum still holds a vital role in directing its progress.

Python Features

Python's features include:

- **Easy-to-learn:** Python has few keywords, simple structure, and a clearly defined syntax. This allows the student to pick up the language quickly.
- **Easy-to-read:** Python code is more clearly defined and visible to the eyes.
- **Easy-to-maintain:** Python's source code is fairly easy-to-maintain.
- **A broad standard library:** Python's bulk of the library is very portable and cross-platform compatible on UNIX, Windows, and Macintosh.

- **Interactive Mode:** Python has support for an interactive mode which allows interactive testing and debugging of snippets of code.
- **Portable:** Python can run on a wide variety of hardware platforms and has the same interface on all platforms.
- **Extendable:** You can add low-level modules to the Python interpreter. These modules enable programmers to add to or customize their tools to be more efficient.
- **Databases:** Python provides interfaces to all major commercial databases.
- **GUI Programming:** Python supports GUI applications that can be created and ported to many system calls, libraries and windows systems, such as Windows MFC, Macintosh, and the X Window system of Unix.
- **Scalable:** Python provides a better structure and support for large programs than shell scripting.

Python has a big list of good features:

- It supports functional and structured programming methods as well as OOP.
- It can be used as a scripting language or can be compiled to byte-code for building large applications.
- It provides very high-level dynamic data types and supports dynamic type checking.
- IT supports automatic garbage collection.

MySQL

MySQL is an open-source relational database management system (RDBMS). Its name is a combination of "My", the name of co-founder Michael Widenius's daughter, and "SQL", the abbreviation for Structured Query Language. A relational database organizes data into one or more data tables in which data types may be related to each other; these relations help structure the data. SQL is a language programmers use to create, modify and extract data from the relational database, as well as control user access to the database. In addition to relational databases and SQL, an RDBMS like MySQL works with an operating system to implement a relational database in a computer's storage system, manages users, allows for network access and facilitates testing database integrity and creation of backups.

MySQL was created by a Swedish company, MySQL AB, founded by Swedes David Axmark, Allan Larsson and Finland Swede Michael "Monty" Widenius. Original development of MySQL by Widenius and Axmark began in 1994. The first version of MySQL appeared on 23 May 1995. It was initially created for personal usage from mSQL based on the low-level language ISAM, which the creators considered too slow and inflexible. They created a new SQL interface, while keeping the same API as mSQL. By keeping the API consistent with the mSQL system, many developers were able to use MySQL instead of the (proprietarily licensed) mSQL antecedent.

Features

- A broad subset of ANSI SQL 99, as well as extensions
- Cross-platform support
- Stored procedures, using a procedural language that closely adheres to SQL/PSM[79]
- Triggers
- Cursors
- Updatable views
- Online Data Definition Language (DDL) when using the InnoDB Storage Engine.
- Information schema
- Performance Schema that collects and aggregates statistics about server execution and query performance for monitoring purposes.
- A set of SQL Mode options to control runtime behavior, including a strict mode to better adhere to SQL standards.
- X/Open XA distributed transaction processing (DTP) support; two phase commit as part of this, using the default InnoDB storage engine

- Transactions with savepoints when using the default InnoDB Storage Engine. The NDB Cluster Storage Engine also supports transactions.
- ACID compliance when using InnoDB and NDB Cluster Storage Engines[81]
- SSL support
- Query caching
- Sub-SELECTs (i.e. nested SELECTs)
- Built-in replication support
  - Asynchronous replication: master-slave from one master to many slaves[82][83] or many masters to one slave[84]
  - Semi synchronous replication: Master to slave replication where the master waits on replication[85][86]
  - Synchronous replication: Multi-master replication is provided in MySQL Cluster.[87]
  - Virtual Synchronous: Self managed groups of MySQL servers with multi master support can be done using: Galera Cluster[88] or the built in Group Replication plugin[89]

MySQL as a service

Some cloud platforms offer MySQL "as a service". In this configuration, application owners do not have to install and maintain the MySQL database on their own. Instead, the database service provider takes responsibility for installing and maintaining the database, and application owners pay according to their usage.[101] Notable cloud-based MySQL services are the Amazon Relational Database Service; Oracle MySQL Cloud Service, Azure Database for MySQL, Rackspace; HP Converged Cloud; Heroku and Jelastic. In this model the database service provider takes responsibility for maintaining the host and database.

Advantages of MySQL

1.Data Security

MySQL is globally renowned for being the most secure and reliable database management system used in popular web applications including WordPress, Drupal, Joomla, Facebook and Twitter. The data security and support for transactional processing that accompany the recent version of MySQL can greatly benefit any business, especially if it is an eCommerce business that involves frequent money transfers.

2. **On-Demand Scalability**

MySQL offers unmatched scalability to facilitate the management of deeply embedded apps using a smaller footprint, even in massive warehouses that stack terabytes of data. On-demand flexibility is the star feature of MySQL. This open-source solution allows complete customization to eCommerce businesses with unique database server requirements.

3. **High Performance**

MySQL features a distinct storage-engine framework that facilitates system administrators to configure the MySQL database server for a flawless performance. Whether it is an eCommerce website that receives a million queries every single day or a high-speed transactional processing system, MySQL is designed to meet even the most demanding applications while ensuring optimum speed, full-text indexes and unique memory caches for enhanced performance.

4. **Round-the-Clock Uptime**

MySQL comes with the assurance of 24×7 uptime and offers a wide range of high-availability solutions, including specialized cluster servers and master/slave replication configurations.

5. **Comprehensive Transactional Support**

MySQL tops the list of robust transactional database engines available on the market. With features such as complete atomic, consistent, isolated, durable transaction support; multi-version transaction support; and unrestricted row-level locking, it is the go-to solution for full data integrity. It guarantees instant deadlock identification through server-enforced referential integrity.

6. **Complete Workflow Control**

With an average download and installation time of less than 30 minutes, MySQL means usability from day one. Whether your platform is Linux, Microsoft, Macintosh or UNIX, MySQL is a comprehensive solution with self-management features that automate everything from space expansion and configuration to data design and database administration.

7. **Reduced Total Cost of Ownership**

By migrating current database apps to MySQL, enterprises enjoy significant cost savings on new projects. The dependability and ease of management can save troubleshooting time that is otherwise wasted in fixing downtime issues and performance problems.

8. **The Flexibility of Open Source**

All the fears and worries that arise in an open-source solution can be brought to an end with MySQL's round-the-clock support and enterprise indemnification. The secure processing and trusted software of MySQL combine to provide effective transactions for large-volume projects. It makes maintenance, debugging and upgrades fast and easy while enhancing the end-user experience.

4 SOFTWARE DEVELOPMENT ANALYASIS

4.1 Overview of Problem

In recent times, there has been a drastic change in people's lifestyles and with an increase in incomes and lower cost of automobiles there is a huge increment in the number of cars on the roads which has led to traffic and commotion. The manual efforts to keep people from breaking traffic rules such as the speed limit are not enough. There is not enough police and man force available to track the traffic and vehicles on roads and check them for speed control. Hence, we require technologically advanced speed calculators installed that effectively detect cars on the road and calculate their speeds.

To implement the above idea two basic requirements, need to be met which are the effective detection of the cars on roads and their velocity measurement. For this purpose, we can use OpenCV software which uses the Haar cascade to train our machine to detect the object, in this case the car.

4.2 Defining the problem

we have developed a Haar cascade to detect cars on the roads, whose velocities are then measured using a python script. The real-time application of this project proves to be much useful as it is easy to implement, fast to process and efficient with low cost development. Also, the tool might be useful to apply in simulation tools to measure velocities of cars. This can be further developed to identify all kinds of vehicles as well as to check anyone who breaks a traffic light.

The improvements in the project can be done by creating a bigger haar cascade since bigger the haar cascade developed, more the number of vehicles that can be detected on the roads. Better search algorithms can allow a faster search and better detection of these vehicles for better efficiency.

In order to support traffic management system in our country we need to build economical traffic monitoring systems. In recent times image and video processing has been applied to the field of traffic management system. This paper explicitly concentrates on the speed of the vehicles, which is one of the important parameters to make roads safe

4.3 Modules An Overview

Modules are a way to organize your course by weeks, units, chapters, topics or whatever organizational structure works for your course. With modules, you create a one-directional linear flow of what you would like your students to do. Modules can be accessed by clicking the Modules button in the Course Tools Menu along the left side of any course. You may also choose to have the Modules page display as your course Home page.You are already familiar with modules, as this orientation course is set up using Modules. Each module can contain files, discussions, assignments, quizzes, and any other learning materials that you would like to use. You can easily add items to a module that you have already created in the course. You can also create new items on the fly within the module. This allows you to create the structure of the course while developing new learning materials.Modules can easily be reordered to fit the flow of the course by simply dragging and dropping. Elements within the modules can also be reorganized by dragging and dropping.

Modules can be released on specific dates. You can also create release conditions (e.g. a module cannot be accessed until a previous module has been completed). All modules appear on the Module page. At this time, you cannot hide modules. You can collapse modules so that only module headings display; however, any user can expand a module to display contents in his/her account. If a module is set to release on a specific date, students will be able to see a list of module contents, but the list will be grayed out and items will remain inaccessible until the release date.

1. TensorFlow
2. Pandas
3. Scikit – learn
4. Matplotlib
5. NumPy

4.4 DEFINE THE MODULES

TensorFlow:

TensorFlow is a free and open-source software library for dataflow and differentiable programming across a range of tasks.

It is a symbolic math library, and is also used for machine learning applications such as neural networks.

It is used for both research and production at Google.

TensorFlow was developed by the Google Brain team for internal Google use.

It was released under the Apache 2.0 open-source license on November 9, 2015.

Its particular focus is on training & interference of deep neural network.

It is written in Python, C++.

It works on Linux, macOS, Windows, Android, JavaScript.

Pandas is an open-source Python Library providing high-performance data manipulation and analysis tool using its powerful data structures.

Python was majorly used for data munging and preparation.

It had very little contribution towards data analysis. Pandas solved this problem.

Using Pandas, we can accomplish five typical steps in the processing and analysis of data, regardless of the origin of data load, prepare, manipulate, model, and analyze.

Python with Pandas is used in a wide range of fields including academic and commercial domains including

finance, economics, Statistics, analytics, etc

In particular, it offers data structures and operations for manipulating numeric tables and time series

The name is derived from the term “panel data”, an econometrics term for data sets. It’s name is play on the phase “python data analysis” itself.

It is written in Python, C.

It is initially released on 11th January 2008.

Scikit – learn:

Scikit-learn provides a range of supervised and unsupervised learning algorithms via a consistent interface in Python.

It is licensed under a permissive simplified BSD license and is distributed under many Linux distributions, encouraging academic and commercial use.

It is free software machine learning library for the python programming language.

It is initially released in June 2007.

Matplotlib:-

It is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms.

Matplotlib can be used in Python scripts, the Python and I Python shells, the Jupyter Notebook, web application servers, and four graphical user interface toolkits.

Matplotlib tries to make easy things easy and hard things possible.

You can generate plots, histograms, power spectra, bar charts, error charts, scatter plots, etc., with just a few lines of code. For examples, see the sample plots and thumbnail gallery.

For simple plotting the pyplot module provides a MATLAB-like interface, particularly when combined with I Python.

For the power user, you have full control of line styles, font properties, axes properties, etc, via an object oriented interface or via a set of functions familiar to MATLAB users

It is written in Python

It is released in 2003

It was originally Written by John D.Hunder

NumPy:

NumPy is a general-purpose array-processing package.

It provides a high performance multidimensional array object, and tools for working with these arrays.

It is the fundamental package for scientific computing with Python. It contains various features including these important ones:

- A powerful N-dimensional array object
- Sophisticated (broadcasting) functions
- Tools for integrating C/C++ and Fortran code
- Useful linear algebra, Fourier transform, and random number capabilities

Besides its obvious scientific uses, NumPy can also be used as an efficient multi-dimensional container of generic data.

It is Written in Python & C.

4.5 Module Functionality

Modules present:

1.Upload Image

2.Train Dataset

3.Upload Test & Classify.

4.user

5.Exit

1.Upload Image:

we apply each component on all the preparation pictures. For each component, it finds the best limit which will characterize the countenances to positive and negative. Be that as it may, clearly, there will be blunders or misclassifications. We select the elements with least mistake rate, which implies they are the elements that best orders the auto and non-auto pictures.

• So now you take a picture. Take each 24x24 window. Apply 6000 elements to it. Check on the off chance that it is auto or not.

2.Train Dataset:

Now every single conceivable size and areas of every part is utilized to ascertain a lot of components. (Simply envision what amount of calculation it needs? Indeed, even a 24x24 window comes about more than 160000

3.Upload Test & Classify:

This velocity and the distance of the camera in feet from the car (i.e. the height of camera above the car) is printed on the output screen. For this use multiple object detection algorithms could have been used but the algorithm of developing the Haar cascade and its implementation proves to be the best since it is the least time consuming, most efficient and highly reliable.

4. User:

User will login through the login with the credentials and will do the whole process.

5. Exit:

Once after the whole process is done with the vehicle detection the user will logout.

5 PROJECT SYSTEM DESIGN

5.1 DFDS IN CASE OF DATABASE PROJECTS

A data flow diagram shows the way information flows through a process or system. It includes data inputs and outputs, data stores, and the various sub processes the data moves through. DFDs are built using standardized symbols and notation to describe various entities and their relationships.

For each data flow, at least one of the endpoints (source and / or destination) must exist in a process. The refined representation of a process can be done in another data-flow diagram, which subdivides this process into sub-processes.

The data-flow diagram is part of the structured-analysis modeling tools. When using UML, the activity diagram typically takes over the role of the data-flow diagram. A special form of data-flow plan is a site-oriented data-flow plan.

Data-flow diagrams can be regarded as inverted Petri nets, because places in such networks correspond to the semantics of data memories. Analogously, the semantics of transitions from Petri nets and data flows and functions from data-flow diagrams should be considered equivalent.

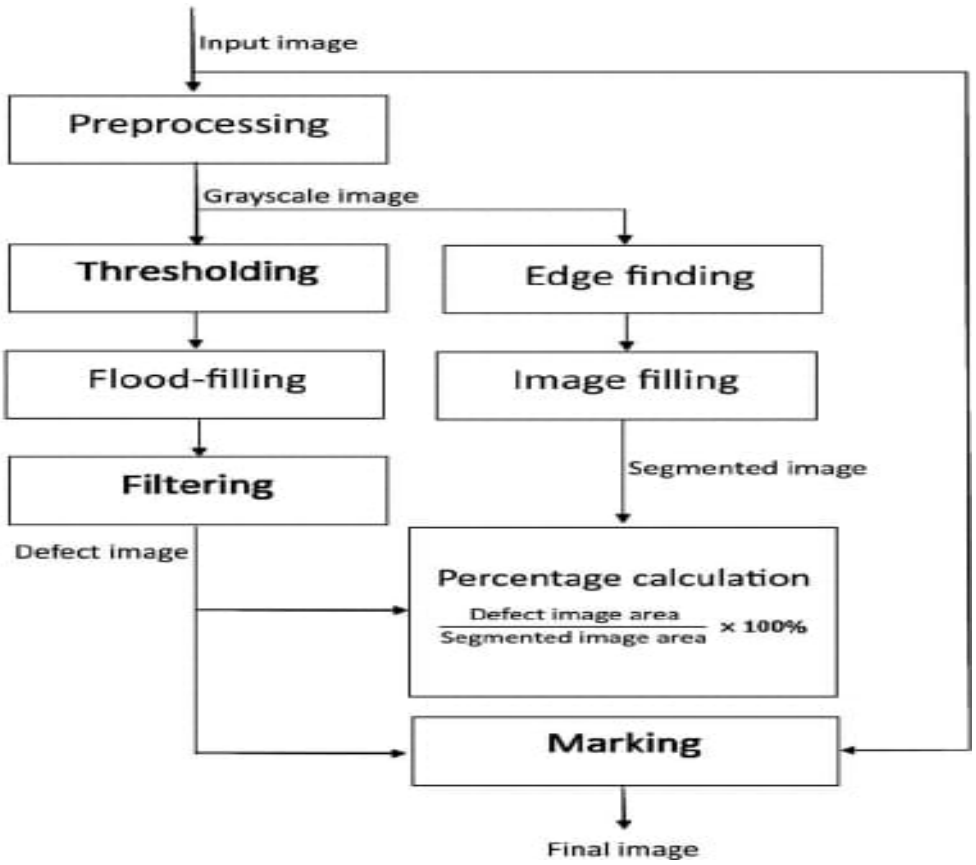


Fig.5.1.1 Data flow diagram

Data flow diagrams visually represent systems and processes that would be hard to describe in a chunk of text. You can use these diagrams to map out an existing system and make it better or to plan out a new system for implementation. Visualizing each element makes it easy to identify inefficiencies and produce the best possible system.

5.2 E-R DIADRAMS

ER Diagram stands for **Entity Relationship Diagram**, also known as **ERD** is a **diagram** that displays the relationship of entity sets stored in a database. ... **ER Diagrams** contain different symbols that use rectangles to represent entities, ovals to define attributes and diamond shapes to represent relationships.

In software engineering, an ER model is commonly formed to represent things a business needs to remember in order to perform business processes. Consequently, the ER model becomes an abstract data model, that defines a data or information structure which can be implemented in a database, typically a relational database.

An ER model is typically implemented as a database. In a simple relational database implementation, each row of a table represents one instance of an entity type, and each field in a table represents an attribute type. In a relational database a relationship between entities is implemented by storing the primary key of one entity as a pointer or "foreign key" in the table of another entity.

5.3 UML DIAGRAMS

UML stands for Unified Modeling Language. UML is a standardized general-purpose modeling language in the field of object-oriented software engineering. The standard is managed, and was created by, the Object Management Group.

The goal is for UML to become a common language for creating models of object oriented computer software. In its current form UML is comprised of two major components: a Meta-model and a notation. In the future, some form of method or process may also be added to; or associated with, UML.

The Unified Modeling Language is a standard language for specifying, Visualization, Constructing and documenting the artifacts of software system, as well as for business modeling and other non-software systems.

The UML represents a collection of best engineering practices that have proven successful in the modeling of large and complex systems.

The UML is a very important part of developing objects oriented software and the software development process. The UML uses mostly graphical notations to express the design of software projects.

GOALS:

The Primary goals in the design of the UML are as follows:

1. Provide users a ready-to-use, expressive visual modeling Language so that they can develop and exchange meaningful models.
2. Provide extendibility and specialization mechanisms to extend the core concepts.
3. Be independent of particular programming languages and development process.
4. Provide a formal basis for understanding the modeling language.
5. Encourage the growth of OO tools market.
6. Support higher level development concepts such as collaborations, frameworks, patterns and components.
7. Integrate best practices.

Characteristics of UML

The UML has the following features:

- It is a generalized modeling language.
- It is distinct from other programming languages like C++, Python, etc.
- It is interrelated to object-oriented analysis and design.
- It is used to visualize the workflow of the system.
- It is a pictorial language, used to generate powerful modeling artifacts.

USE CASE DIAGRAM:

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.

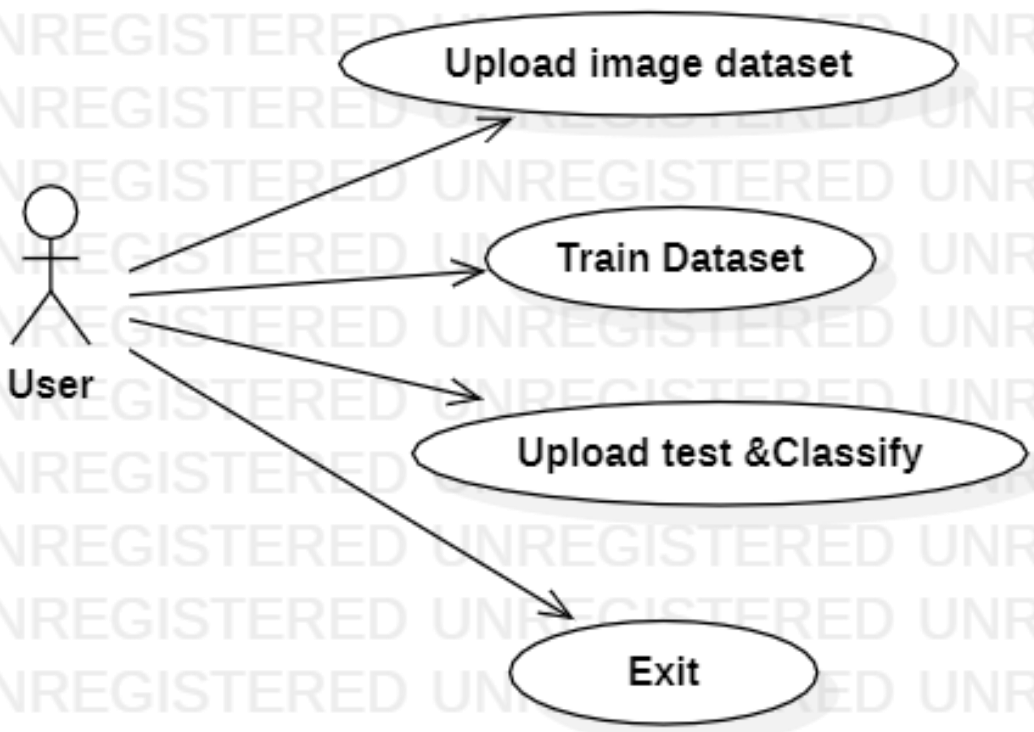


Fig. 5.2.1 Use-Case diagram

CLASS DIAGRAM:

In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.

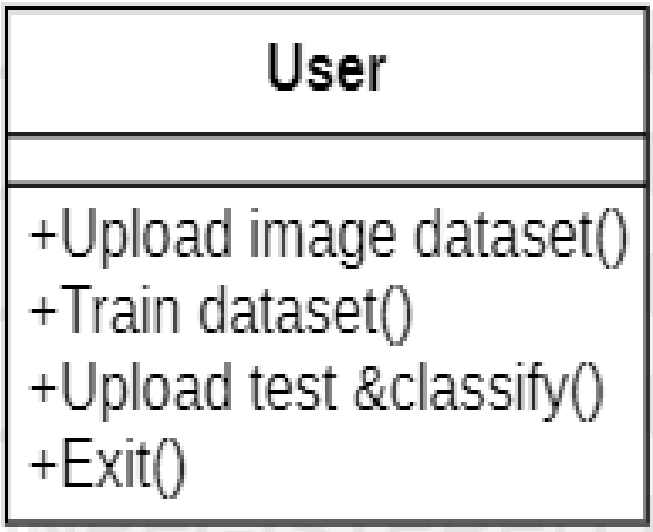


Fig.5.2.2 Class diagram

SEQUENCE DIAGRAM:

A sequence diagram in Unified Modeling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.

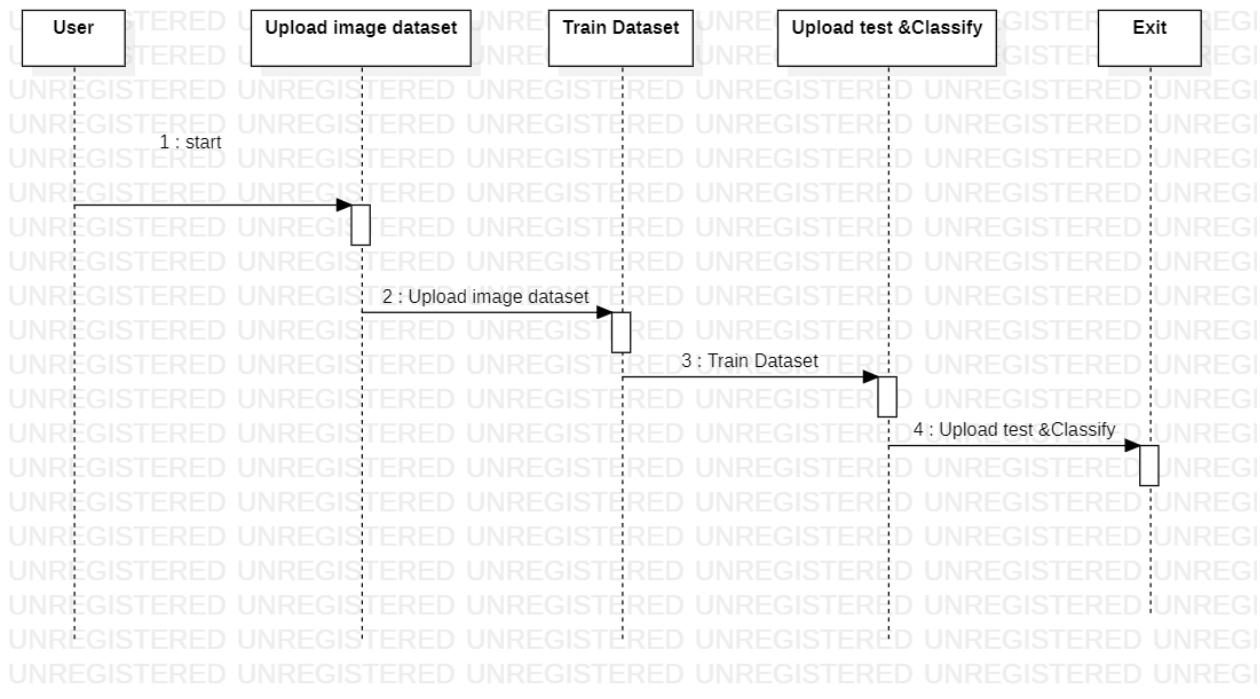


fig.5.2.3 Sequence diagram

ACTIVITY DIAGRAM:

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modeling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.

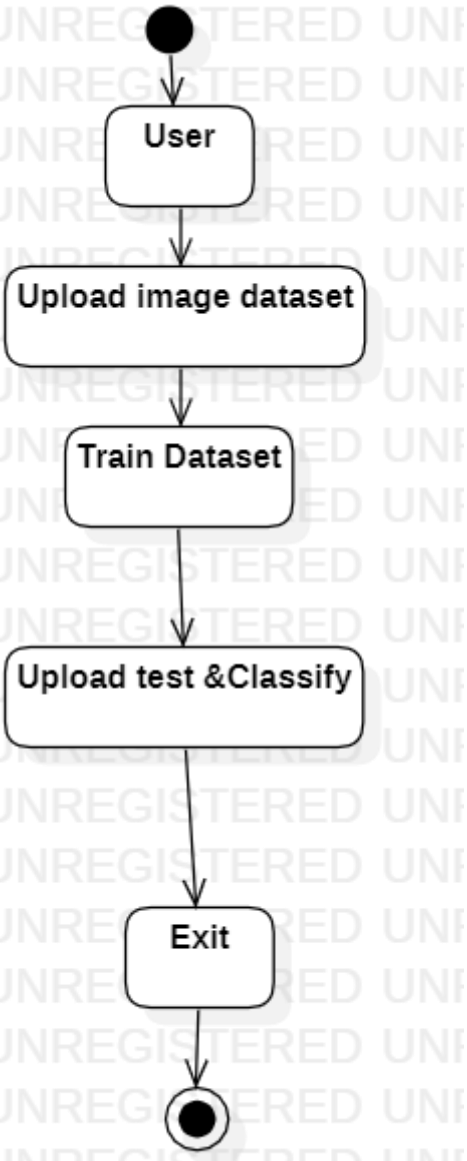


Fig.5.2.4 Activity Diagram

6 Project Coding

6.1 CODE TEMPLATES

```
def uploadDataset():
    global filename
    text.delete('1.0', END)
    filename = filedialog.askdirectory(initialdir=".")
    text.insert(END,'dataset loaded\n')
    X = np.load("model/X.txt.npy")
    Y = np.load("model/Y.txt.npy")
    X = np.asarray(X)
    Y = np.asarray(Y)
    img = X[20].reshape(64,64,3)
    cv2.imshow('ff',cv2.resize(img,(250,250)))
    cv2.waitKey(0)
def KNNCNN():
    X = np.load("model/X.txt.npy")
    Y = np.load("model/Y.txt.npy")
    print(X.shape)
    print(Y.shape)
    temp = X
    XX = np.reshape(temp, (temp.shape[0],(temp.shape[1]*temp.shape[2]*temp.shape[3])))
    pca = PCA(n_components = 180)
    XX = pca.fit_transform(XX)
    print(XX.shape)


    X_train, X_test, y_train, y_test = train_test_split(XX, Y, test_size=0.2)


    cls = KNeighborsClassifier()
    cls.fit(X_train, y_train)
    predict = cls.predict(X_test)
    acc = accuracy_score(y_test,predict)*100
    accuracy.append(acc)
    text.insert(END,'KNN Prediction Accuracy : '+str(acc)+"\n")
```

```
Y1 = to_categorical(Y)
cnn = Sequential()
cnn.add(Convolution2D(32, 3, 3, input_shape = (64, 64, 3), activation = 'relu'))
cnn.add(MaxPooling2D(pool_size = (2, 2)))
cnn.add(Convolution2D(32, 3, 3, activation = 'relu'))
cnn.add(MaxPooling2D(pool_size = (2, 2)))
cnn.add(Flatten())
cnn.add(Dense(128, activation = 'relu'))
cnn.add(Dense(5, activation = 'softmax'))
cnn.compile(optimizer = 'adam', loss = 'categorical_crossentropy', metrics = ['accuracy'])
print(cnn.summary())
cnn_history = cnn.fit(X, Y1, batch_size=16, epochs=10, validation_split=0.2, shuffle=True, verbose=2)
cnn_history = cnn_history.history
cnn_history = cnn_history['accuracy']
acc = cnn_history[9] * 100
accuracy.append(acc)
text.insert(END,'CNN Prediction Accuracy : '+str(acc)+"\n\n")

bars = ('KNN Inference','CNN Inference')
y_pos = np.arange(len(bars))
plt.bar(y_pos, [accuracy[6],accuracy[7]])
plt.xticks(y_pos, bars)
plt.show()
plt.title('KNN & CNN Inference Accuracy Performance Graph')
plt.show()
```

6.2 OUTLINE FOR VARIOUS FILES

We used Python programming to implement our project. A single python file is used to implement our code. This file consists of various modules that we have used. Our project modules are - Upload Image, Train Dataset, Upload Test & Classify, User and exit. We also used various python modules like pandas, matplotlib, numpy, tensorflow, sklearn.

Python too supports file handling and allows users to handle files i.e., to read and write files, along with many other file handling options, to operate on files. The concept of file handling has stretched over various other languages, but the implementation is either complicated or lengthy, but alike other concepts of Python,

this concept here is also easy and short. Python treats file differently as text or binary and this is important. Each line of code includes a sequence of characters and they form text file. Each line of a file is terminated with a special character, called the EOL or End of Line characters like comma {,} or newline character. It ends the current line and tells the interpreter a new one has begun. Let's start with Reading and Writing files.

Reading and writing data to files using Python is pretty straightforward. To do this, you must first open files in the appropriate mode. Here's an example of how to use Python's "with open(…) as …" pattern to open a text file and read its contents:

```
with open('data.txt', 'r') as f:
   data = f.read()
```

open() takes a filename and a mode as its arguments. r opens the file in read only mode.

NumPy introduces a simple file format for ndarray objects. This . npy file stores data, shape, dtype and other information required to reconstruct the ndarray in a disk file such that the array is correctly retrieved even if the file is on another machine with different architecture.

6.3 CLASS WITH FUNCTIONALITY

In our project code, we implemented six different methods. They are:

1.UploadImage()

2.TrainDataset()

3.UploadTest& Classify ()

4.user()

5.Exit()

Our first method upload Image() doesn't take any input parameters but after successful execution, it displays a message "Image loaded". Our second method TrainDataset () will take 1 parameters and it will train the whole dataset being considered.Our third .UploadTest& Classify () study all the images and will predict the car model based on the pattern.the last two are the important methods that are mainly used for the logging and logout purpose of the project. Once after the prediction is using the image user will exit with the help of the exit () method.

6.4 METHODS INPUT AND OUTPUT PARAMETERS

INPUT DESIGN

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the following things:

- What data should be given as input?
- How the data should be arranged or coded?
- The dialog to guide the operating personnel in providing input.
- Methods for preparing input validations and steps to follow when error occur.

OBJECTIVES

1.Input Design is the process of converting a user-oriented description of the input into a computer-based system. This design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerized system.

2. It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The goal of designing input is to make data entry easier and to be free from errors. The data entry screen is designed in such a way that all the data manipulates can be performed. It also provides record viewing facilities.

3.When the data is entered it will check for its validity. Data can be entered with the help of screens. Appropriate messages are provided as when needed so that the user will not be in maize of instant. Thus the objective of input design is to create an input layout that is easy to follow

OUTPUT DESIGN

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making.

1. Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is designed so that people will find the system can use easily and effectively. When analysis design computer output, they should Identify the specific output that is needed to meet the requirements.

2.Select methods for presenting information.

3.Create document, report, or other formats that contain information produced by the system.

The output form of an information system should accomplish one or more of the following objectives.

- Convey information about past activities, current status or projections of the
- Future.
- Signal important events, opportunities, problems, or warnings.
- Trigger an action.
- Confirm an action.

7.PROJECT TESTING

7.1 VARIOUS TEST CASES

In software engineering, a test case is a specification of the inputs, execution conditions, testing procedure, and expected results that define a single test to be executed to achieve a particular software testing objective, such as to exercise a particular program path or to verify compliance with a specific requirement.[1] Test cases underlie testing that is methodical rather than haphazard. A battery of test cases can be built to produce the desired coverage of the software being tested. Formally defined test cases allow the same tests to be run repeatedly against successive versions of the software, allowing for effective and consistent regression testing.

SYSTEM TEST

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

TYPES OF TESTS

Unit testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

Integration testing

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

Functional test

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items:

Valid Input : identified classes of valid input must be accepted.

Invalid Input : identified classes of invalid input must be rejected.

Functions : identified functions must be exercised.

Output : identified classes of application outputs must be exercised.

Systems/Procedures : interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

System Test

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

Unit Testing

Unit testing is usually conducted as part of a combined code and unit test phase of the software lifecycle, although it is not uncommon for coding and unit testing to be conducted as two distinct phases.

Test strategy and approach

Field testing will be performed manually and functional tests will be written in detail.

Test objectives

- All field entries must work properly.
- Pages must be activated from the identified link.
- The entry screen, messages and responses must not be delayed.

Features to be tested

- Verify that the entries are of the correct format
- No duplicate entries should be allowed
- All links should take the user to the correct page.

Integration Testing

Software integration testing is the incremental integration testing of two or more integrated software components on a single platform to produce failures caused by interface defects.

The task of the integration test is to check that components or software applications, e.g. components in a software system or – one step up – software applications at the company level – interact without error.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

Acceptance Testing

User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

7.2 BLACK BOX

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box .you cannot "see" into it. The test provides inputs and responds to outputs without considering how the software works.

FIG.7.2.1 blackbox testing

Unit Testing

Unit testing is usually conducted as part of a combined code and unit test phase of the software lifecycle, although it is not uncommon for coding and unit testing to be conducted as two distinct phases.

Test strategy and approach

Field testing will be performed manually and functional tests will be written in detail.

Test objectives

- All field entries must work properly.
- Pages must be activated from the identified link.
- The entry screen, messages and responses must not be delayed.

Features to be tested

- Verify that the entries are of the correct format
- No duplicate entries should be allowed
- All links should take the user to the correct page.

Integration Testing

Software integration testing is the incremental integration testing of two or more integrated software components on a single platform to produce failures caused by interface defects.

The task of the integration test is to check that components or software applications, e.g. components in a software system or – one step up – software applications at the company level – interact without error.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

Acceptance Testing

User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

7.3 WHITE BOX TESTING

White Box Testing is a testing in which in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is purpose. It is used to test areas that cannot be reached from a black box level.

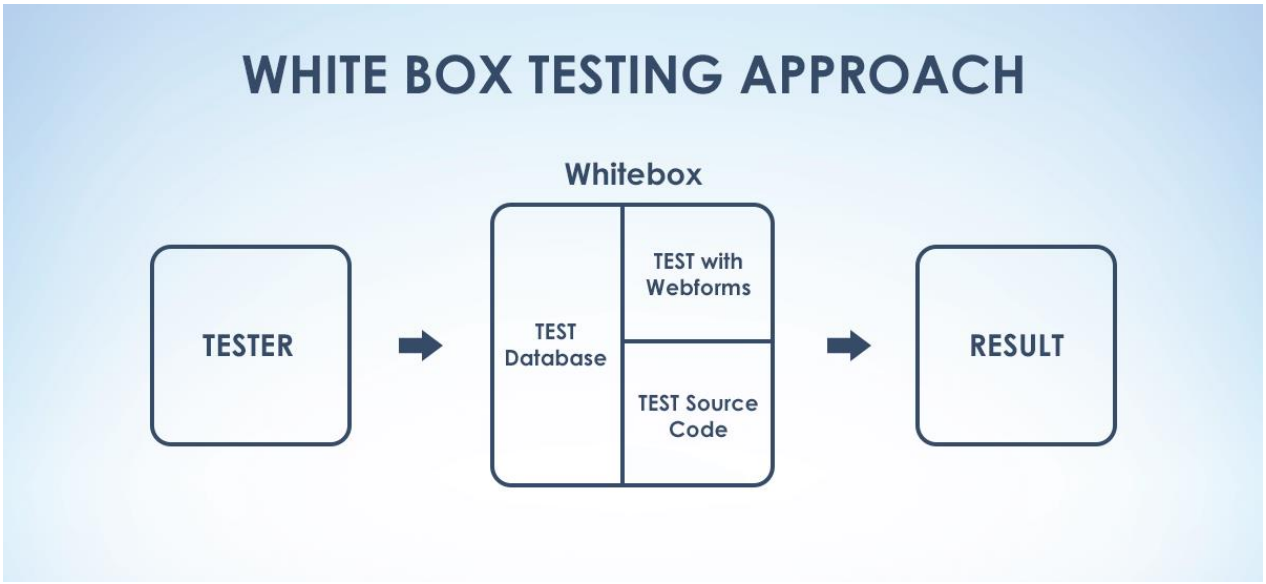


FIG.7.3.1.White box testing

The box testing approach of software testing consists of black box testing and white box testing. We are discussing here white box testing which also known as glass box is **testing, structural testing, clear box testing, open box testing and transparent box testing**.

It tests internal coding and infrastructure of a software focus on checking of predefined inputs against expected and desired outputs. It is based on inner workings of an application and revolves around internal structure testing. In this type of testing programming skills are required to design test cases. The primary goal of white box testing is to focus on the flow of inputs and outputs through the software and strengthening the security of the software.

The term 'white box' is used because of the internal perspective of the system. The clear box or white box or transparent box name denote the ability to see through the software's outer shell into its inner workings.

Developers do white box testing. In this, the developer will test every line of the code of the program. The developers perform the White-box testing and then send the application or the software to the testing team, where they will perform the black box testing and verify the application along with the requirements and identify the bugs and sends it to the developer.

The developer fixes the bugs and does one round of white box testing and sends it to the testing team. Here, fixing the bugs implies that the bug is deleted, and the particular feature is working fine on the application.

Here, the test engineers will not include in fixing the defects for the following reasons:

- Fixing the bug might interrupt the other features. Therefore, the test engineer should always find the bugs, and developers should still be doing the bug fixes.
- If the test engineers spend most of the time fixing the defects, then they may be unable to find the other bugs in the application.

8 OUTPUT SCREENS

8.1 USER INTERFACES

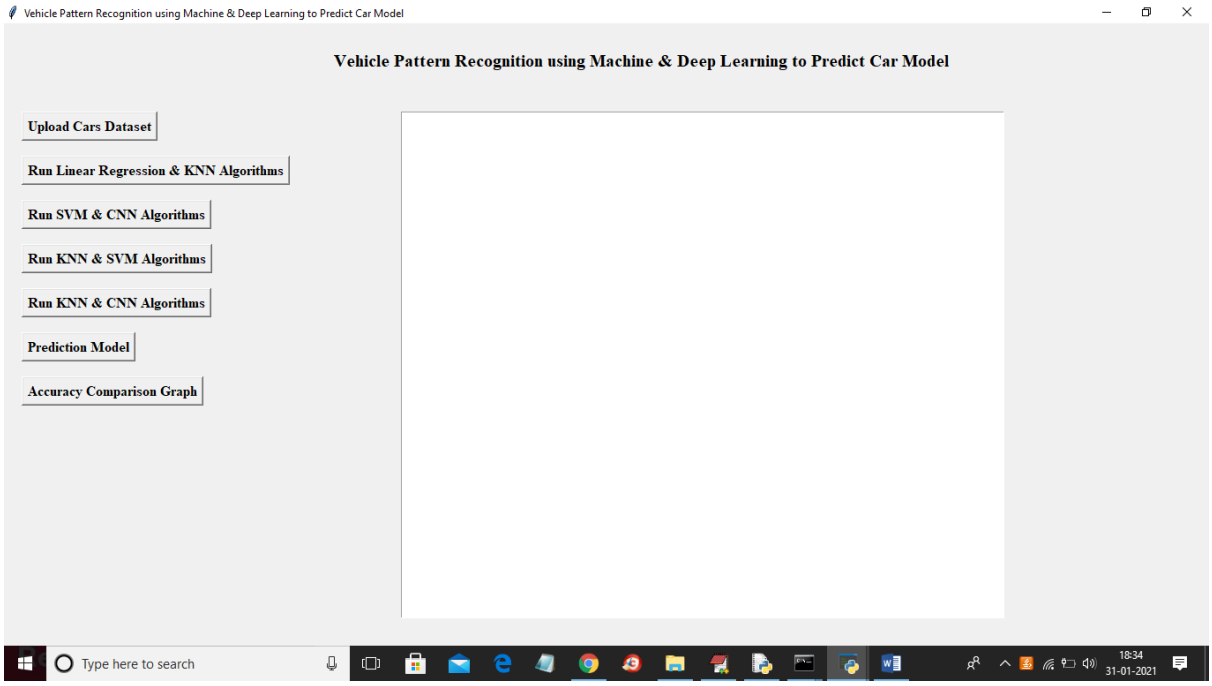


Figure 8.1.1 In above screen click on 'Upload Cars Dataset' button to load dataset

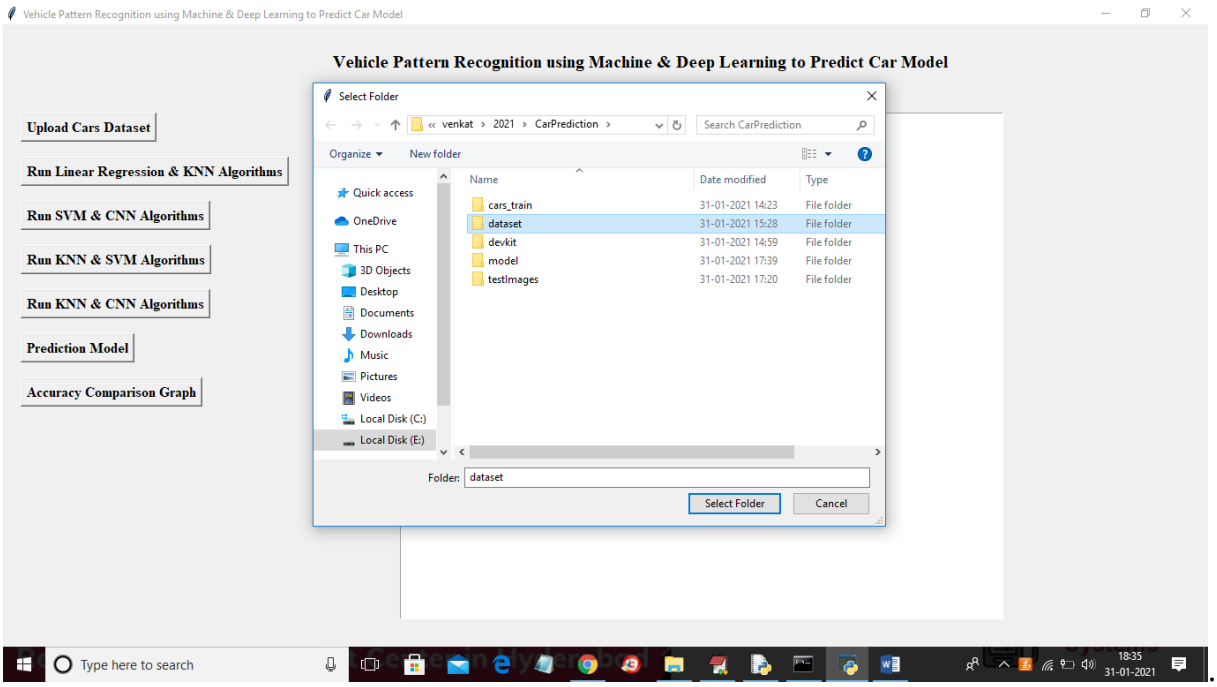
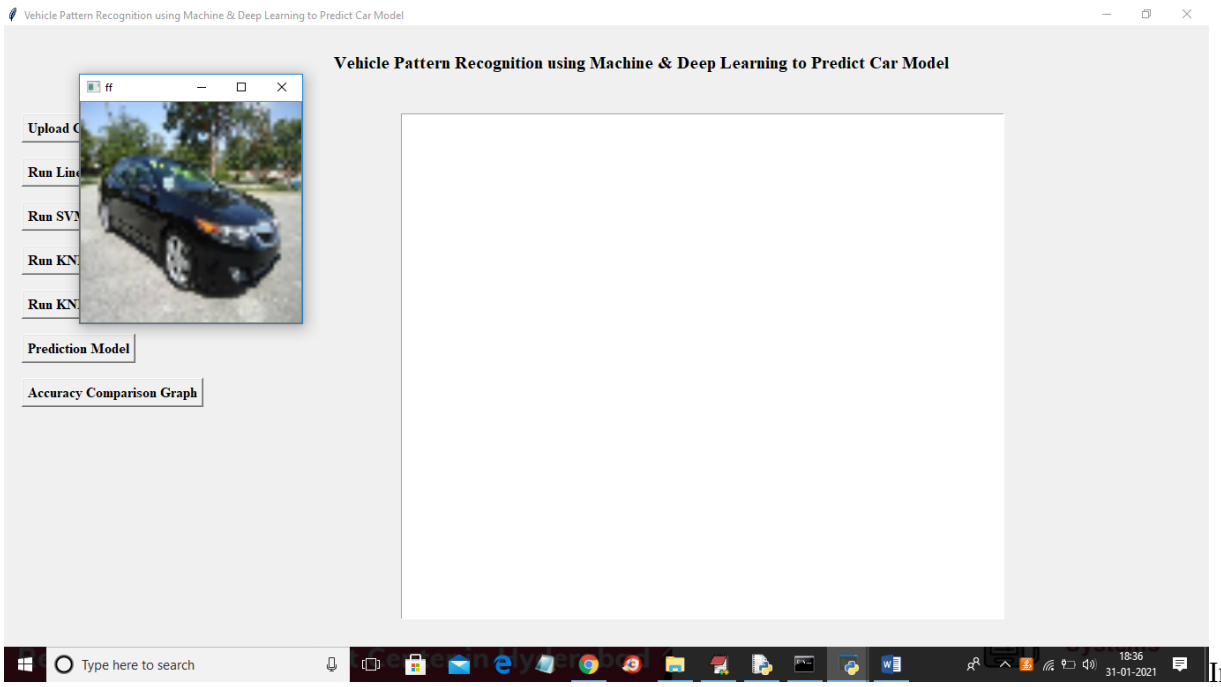


Figure 8.1.2 In above screen selecting and uploading 'dataset' folder and then click on 'Select Folder' button to load dataset and to get below screen



In above screen Figure 8.1.3 showing one sample image to see dataset loaded correctly and now close above image to get below screen

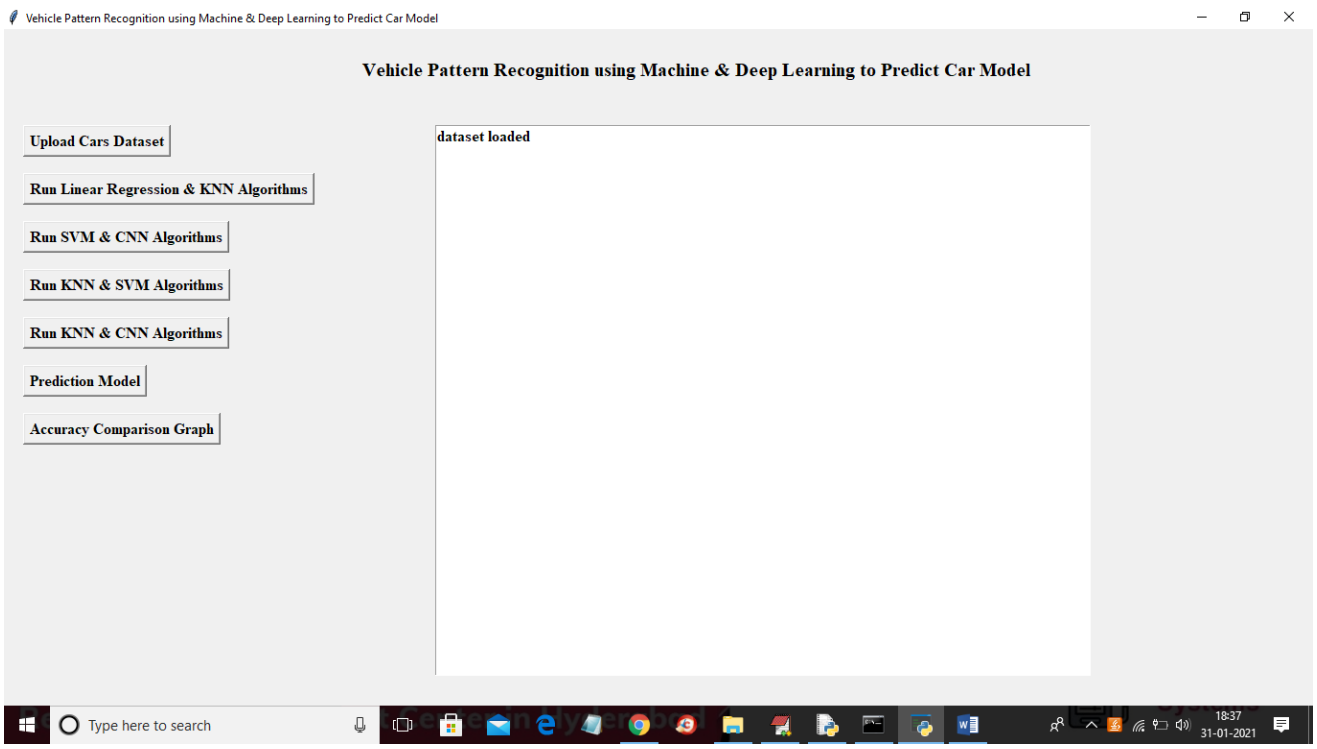


Figure 8.1.4 In above screen dataset loaded and now click on 'Run Linear Regression & KNN Algorithm' button to apply dataset on both algorithms and to calculate prediction accuracy

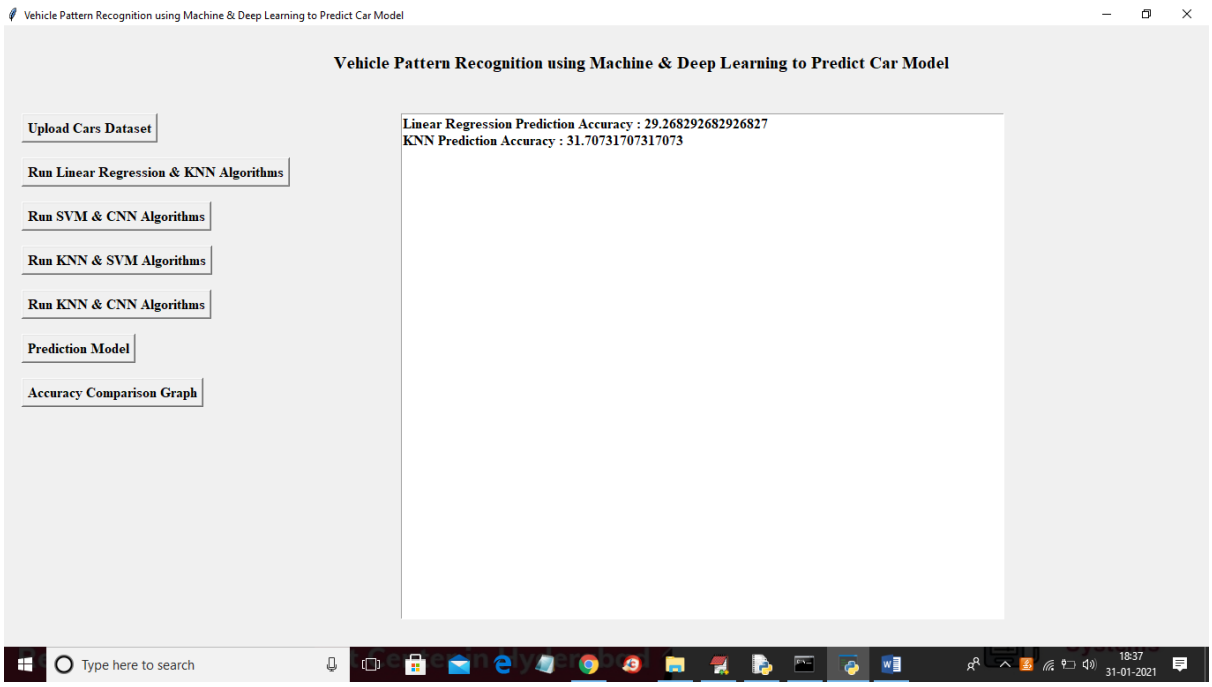


Figure 8.1.5 In above screen Linear regression accuracy is 29% and KNN accuracy is 31% and now click on 'Run SVM & CNN Algorithm' button to get pattern accuracy

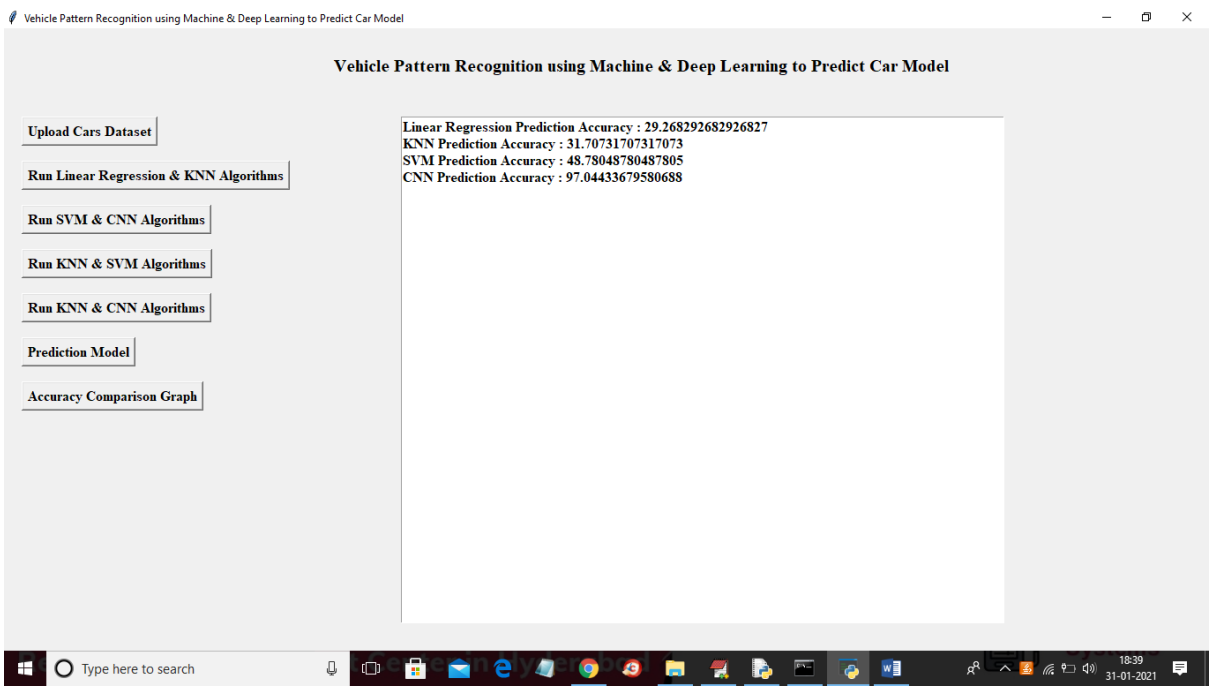


Figure 8.1.6 In above screen SVM accuracy is 48% and CNN is 97% for pattern features and now click on 'Run KNN & SVM Algorithm' button to run inference features from dataset

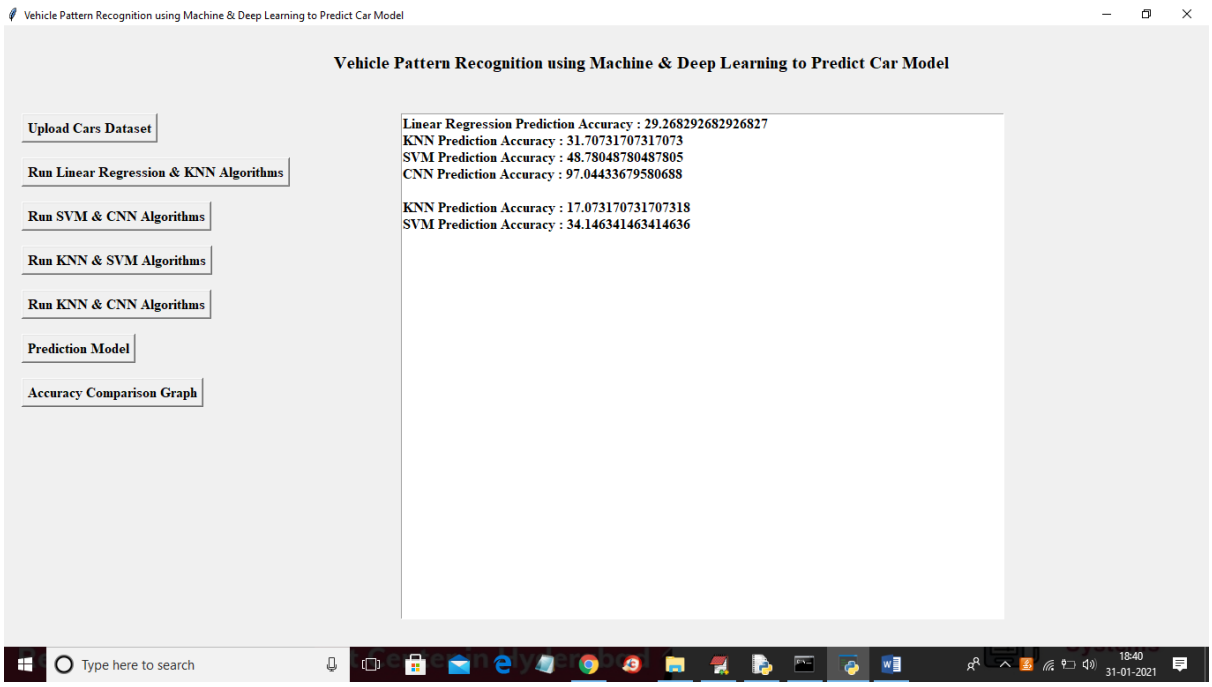


Figure 8.1.7 In above screen KNN accuracy is 17% and SVM accuracy is 34% and now click on 'Run KNN & CNN Algorithms' button to calculate its accuracy on inference features

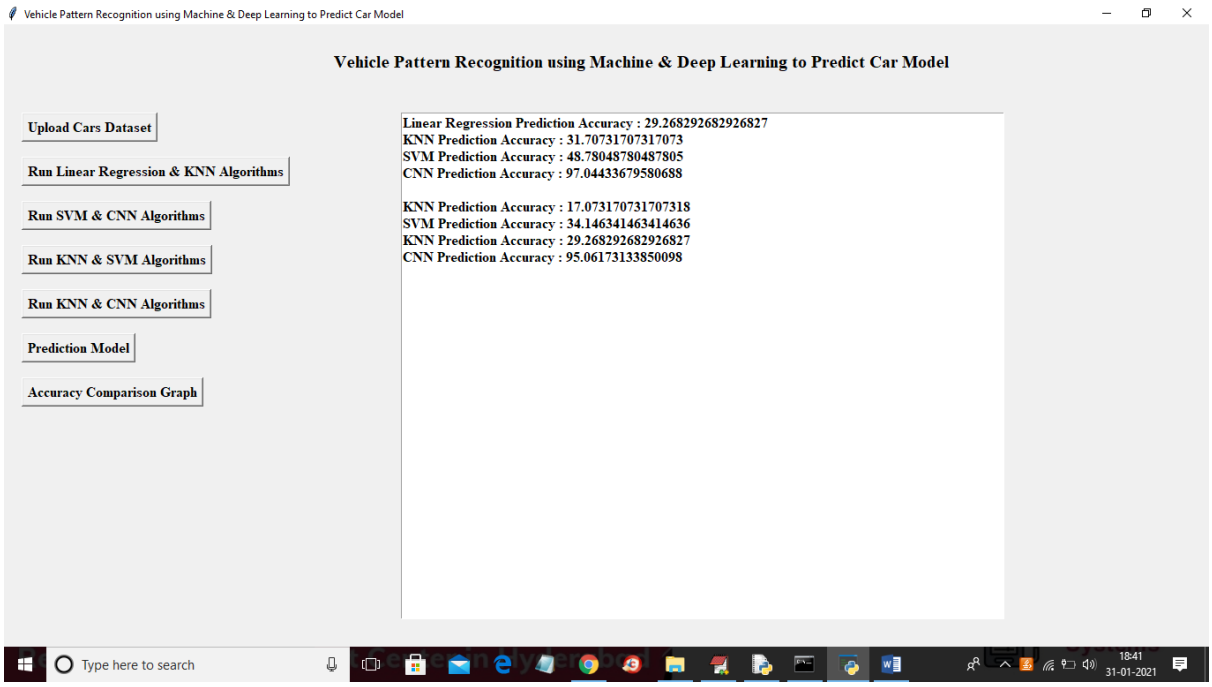


Figure 8.1.8 In above screen KNN inference accuracy is 29% and CNN accuracy is 95% and now click on 'Prediction Model' button to upload test image and get it predicted model

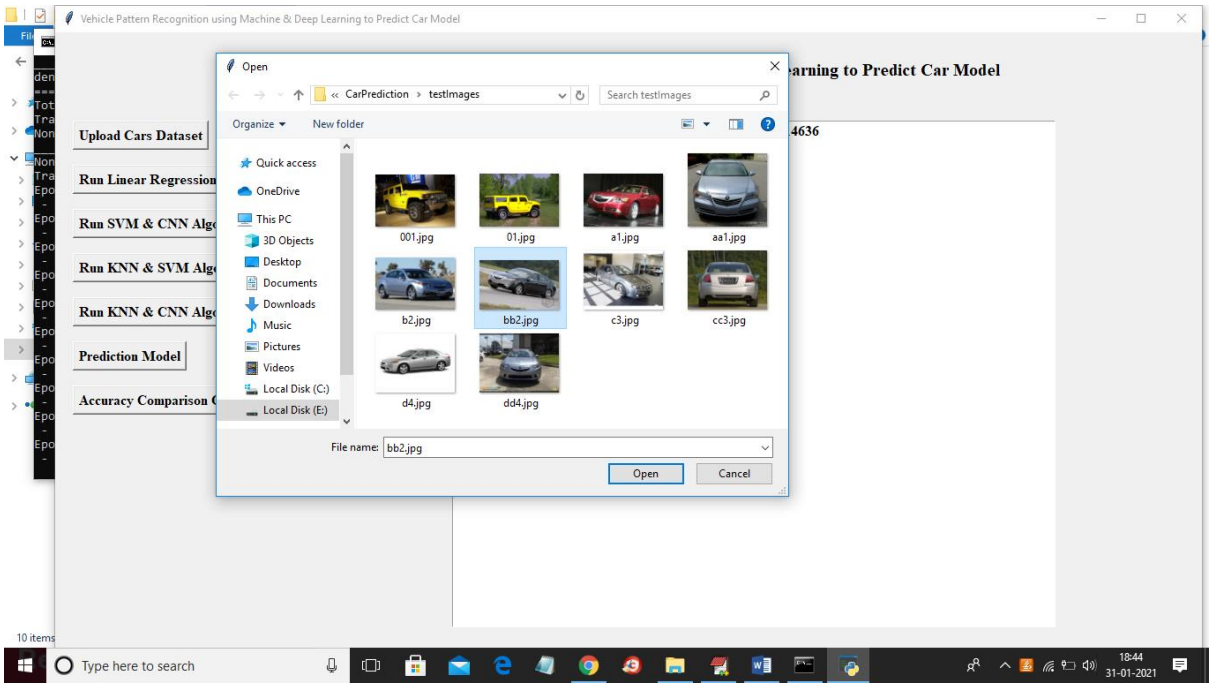


Figure 8.1.9 In above screen selecting and uploading ‘bb2.jpg’ and then click on ‘Open’ button to get below prediction result

8.2 OUTPUT SCREENS

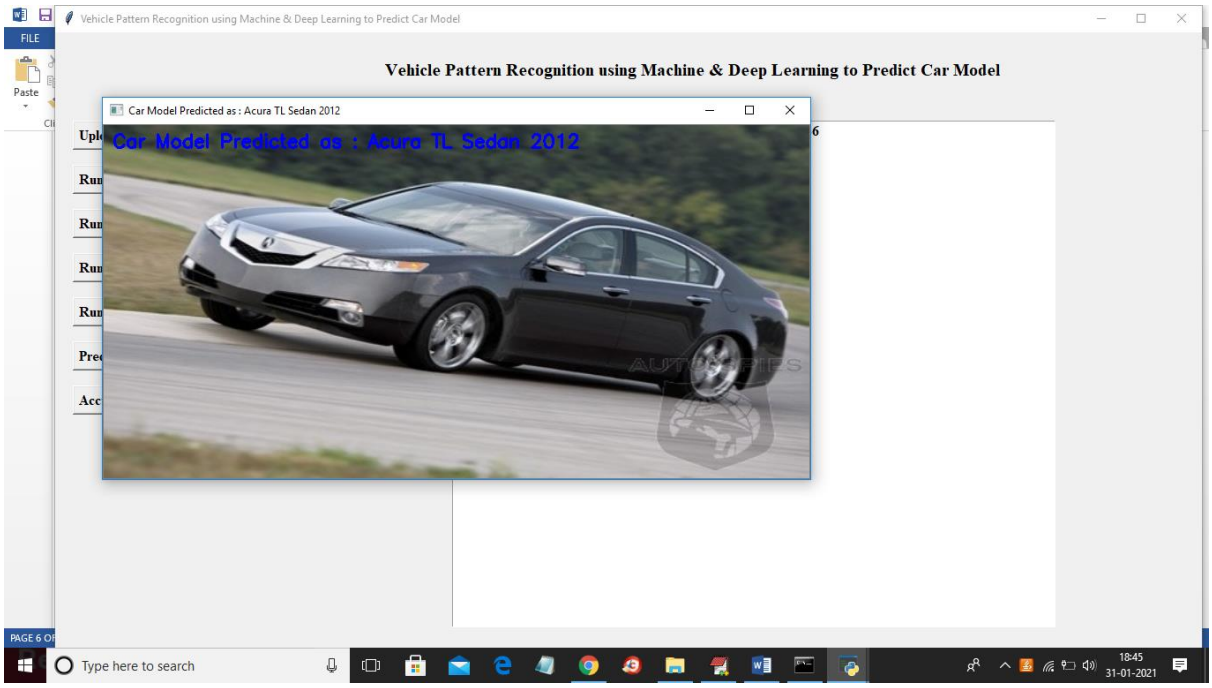


Figure 8.2.1 Similarly you can upload any image and get it model predicted. Now click on 'Accuracy Comparison Graph' button to get below screen

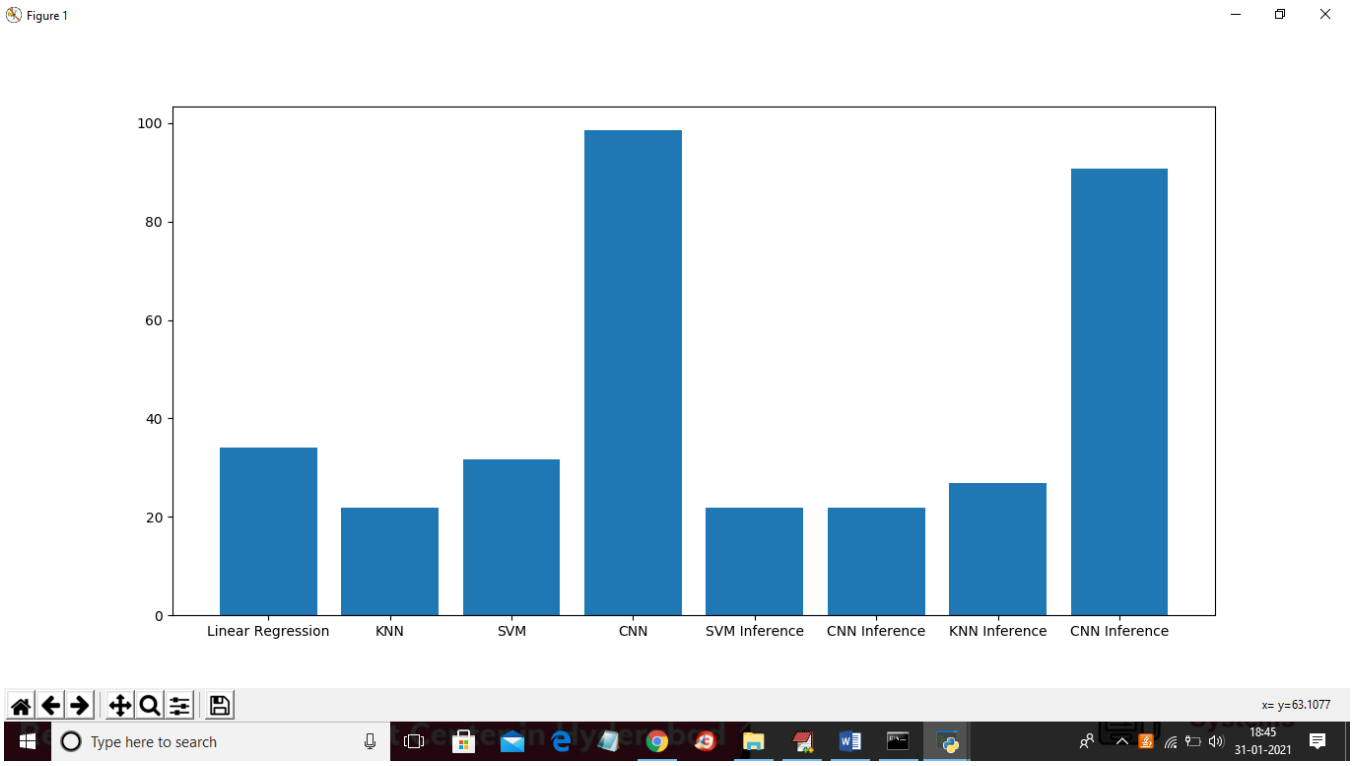


Figure 8.2.2 In above screen x-axis represents algorithm name and y-axis represents accuracy of that algorithm

9 EXPERIMENTAL RESULTS

HOMEPAGE

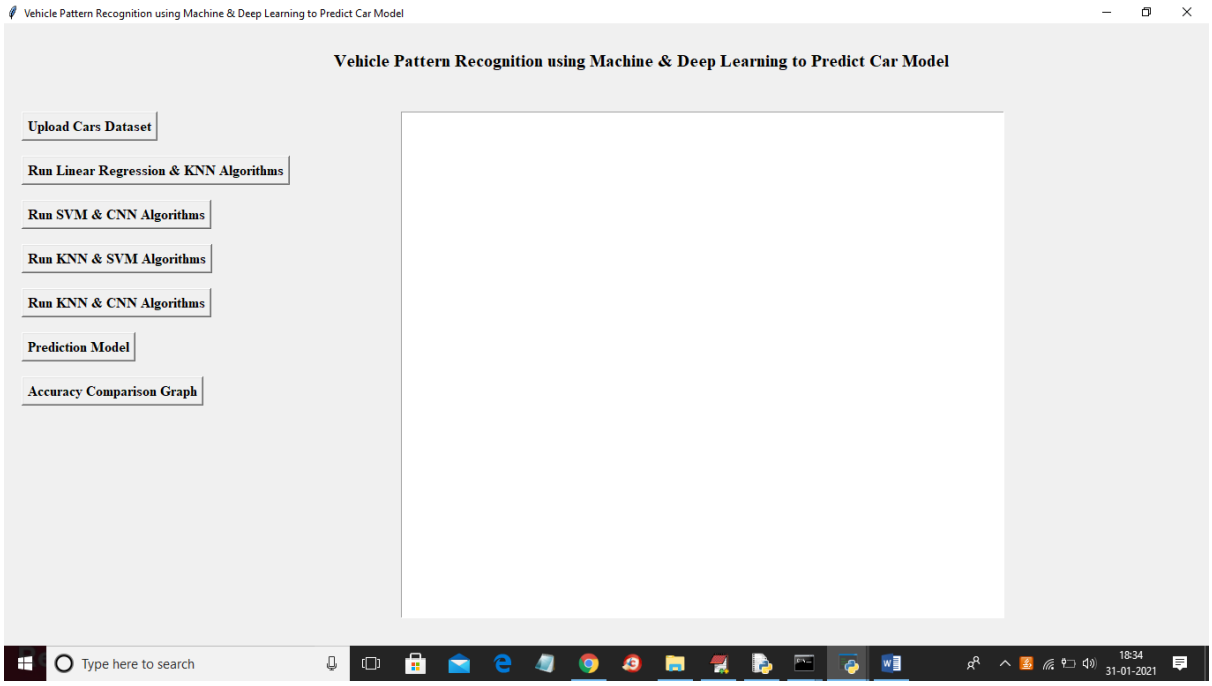


Figure 9.1 In above screen selecting and uploading 'dataset' folder and then click on 'Select Folder' button to load dataset and to get below screen

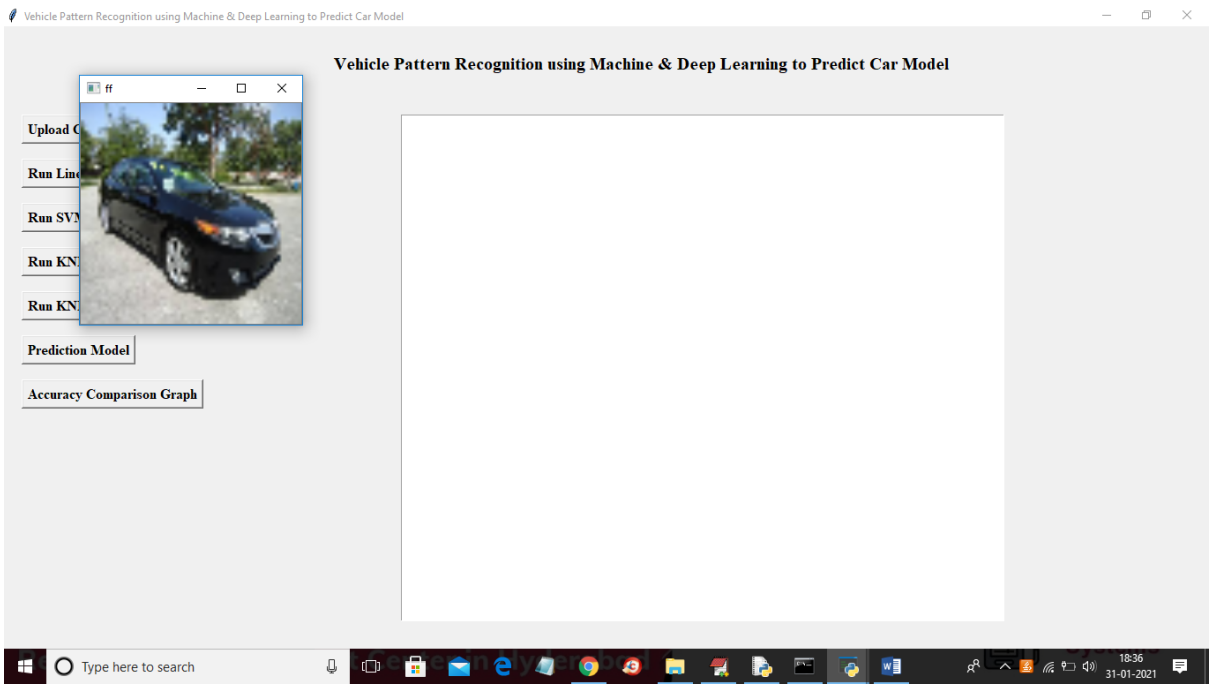


Figure 9.2 In above screen dataset loaded and now click on 'Run Linear Regression & KNN Algorithm' button to apply dataset on both algorithms and to calculate prediction accuracy

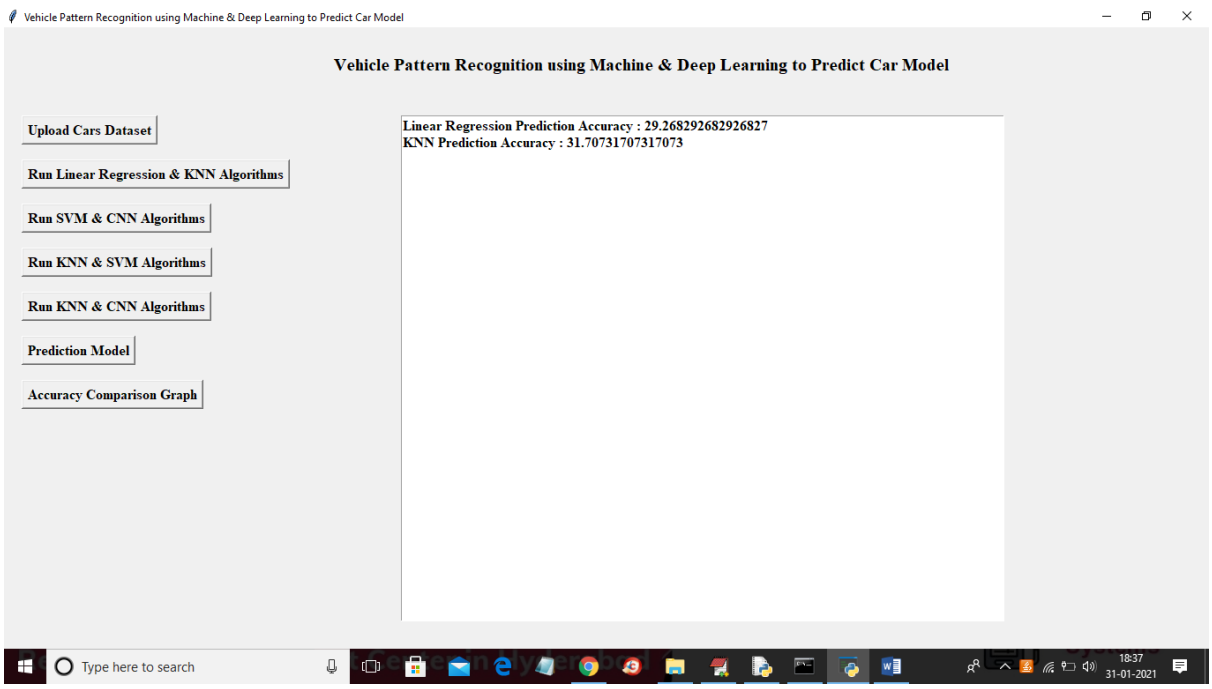


Figure 9.3 Running algorithms

OUTPUT

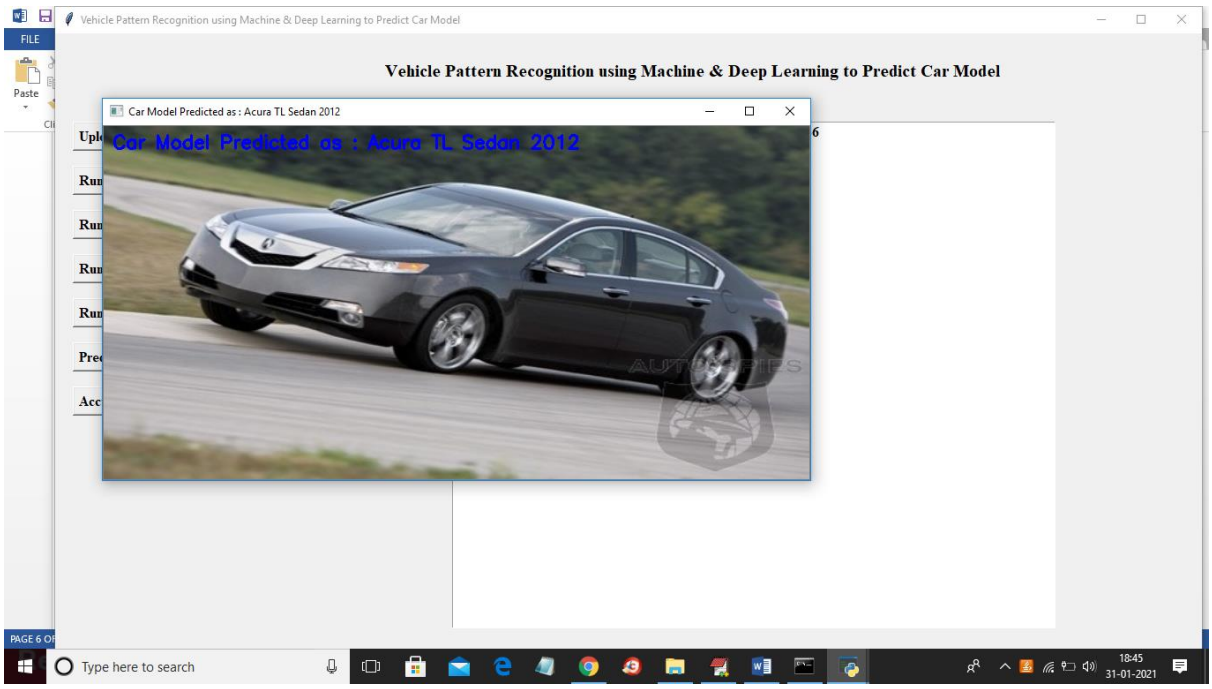


Figure 9.4 Final output

10.CONCLUSION AND FUTURE ENHANCEMENT

The latest work with fine-grained classification, pointed critical parts of cars body that helped to identify the car by its special characteristics.

The experimental results show that compared with the traditional machine learning methods, the model used in this paper has been improved both in average target detection accuracy and detection rate. the classification test result of this article is also suitable for vehicle type detection of three types of cars, minibus and suv in different scenarios and has achieved good results

Then deep learning was rediscovered due to GPU capabilities enhancing the training times that once took weeks to train now took days only. with fast deep learning capabilities, many new algorithms were discovered such as alexnet and googlenet.

This system proves very efficient in terms of recognition rate and processing speed, even in challenging situations like occluded or unclear images. hence, the system can be considered profoundly useful for vehicle analysis based on make and model and can be effectively used for vehicle monitoring.

We have developed a completely innovative convolutional neural network, that is simple but accurate and efficient. In object detection framework the convolutional features gathered from our system is better than state-of-art image classification network. Our method achieves accuracy by exchanging the flexibility characteristics with a faster RCNN, both during training and during testing. But our model hasn't considered the noise while the image is being captured .

In future the noise will be consider as a pre-processing step. The proposed model performed well without noise, providing accurate prediction of some test images. Although it is accurate, but that it is not 100% accurate. We hope that our system will benefit from progress in this area.

In future researches, the limitation of alignment and pose has to be considered and methods of finding the most discriminative patches amongst all patches detected in different car models have to be researched without limiting the pose.

REFERENCES

1. Bay, H., Tuytelaars, T. and Gool, L. (2008) SURF: Speeded Up Robust Features. Computer Vision and Image Understanding. 110(3). p. 346-359.

2 .Chen, L., Hsieh, J., Yan, Y. and Chen, D. (2015) Vehicle make and model recognition using sparse representation and symmetrical SURFs. Pattern Recognition. 48(6). p. 1979-1998.

3. Cheung, S. and Chu, A. (2008) Make and Model Recognition of Cars. Projects in Vision and Learning.

Emami, H., Fathi, M. and Raahemifar, K. (2014) Real Time Vehicle Make and Model Recognition Based on Hierarchical Classification. International Journal of Machine Learning and Computing.

4(2). p. 142-145. 4. Erhan, D., Szegedy, C., Toshev, A. and Anguelov, D. (2013) Scalable Object Detection Using Neural Networks.

5. Girshick, R., Donahue,J., Darrell, T. and Malik, J. (2014) Rich feature hierarchies for accurate object detection and semantic segmentation.

6. Krause, J., Gebru, T., Deng, J., Li, L. and Fei-Fei, L. (2014) Learning Features and Parts for Fine-Grained Recognition.

7. Krause, J., Jin, H., Yang, J. and Fei-Fei, L. (2016) Fine-Grained Recognition without Part Annotations.

8. Krause, J., Stark, M., Deng, J. and Fei-Fei, L. (2013) 3d object representations for fine-grained categorization. In 4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13).

9. Krizhevsky, A., Sutskever, I. and Hinton, G., E. (2012) ImageNet Classification with Deep Constitutional Neural Networks.

10. Redmon, J., Divvala, S., Girshick, R. and Farhadi, A. (2016) You Only Look Once: Unified, Real-Time Object Detection. 11. Ren, S., He, K., Girshick, R. and Sun, J. (2016) Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks.

12. Szegedy, C., Liu, W., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A. (2015) Going Deeper with Convolutions.

13. Wang, J., Yang, J., Yu, K., Huang, T. and Gong,Y. (2010) Locality-constrained Linear Coding for Image Classification. 14. Wang, Y., Choi, J., Morariu, V., I. and Davis, L., S. (2016) Mining Discriminative Triplets of Patches for FineGrained Classification.

15. Yang, H., Zhai, L., Li, L., Liu, Z., Luo, Y., Wang, Y., Lai, H. and Guan, H. (2013) An Efficient Vehicle Model Recognition Method. Journal of Software. 8(8).

16. Yang, L., Luo, P., Loy, C. and Tang, X. (2015) A LargeScale Car Dataset for Fine-Grained Categorization and Verification.

17. Zhou, Y., Nejati, H., Do, T., Cheung, N. and Cheah, L. (2016) Image-based Vehicle Analysis using Deep NeuralNetworkSystematicStudy.

PUBLICATIONS

JOURNAL (UGC APPROVED JOURNAL)

CONFERENCE (INTERNATIONAL CONFERENCE ON “INNOVATIONS IN COMPUTERS NETWORKS, COMPUTATIONAL INTELLIGENCE AND IOT” [ICICCI-21],C01 BATCH).

TOPIC:VEHICLE PATTERN RECOGNITION USING MACHEINE LEARNING AND DEEP LEARNING.

STUDENT PROFLE

Anupama Arkala is currently pursuing her Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. She completed her Intermediate from Narayana Junior College and 10th class from RBVRR High School. Her Technical Skills include C and Python. She also has a basic understanding of C++. She had also done Participation in Entrepreneurship ESUMMIT certified at MLRIT College and Machine Learning Workshop. Some External skills are: She is a college ambassador in youth marketing company called Grapevine and also she participated in district wise Athletics and Basketball competitions. She also completed some certifications in Artificial Intelligence, MYSQL Database and Java Script by the Net Ninja.

B.Sairam is currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin’s Engineering College. He completed his schooling till 10th grade in Sri Chaitanya techno school and he completed his Intermediate in Sri Chaitanya junior college. his technical skills include c, c++, java, python, MySQL. he has done an internship in java at electronics corporation of India limited (ecil). his participations include: a national level seminar on "recent trends in cloud computing, fog, and edge computing" on 18th and 19th of June 2021 and a national level three-day workshop on "ai & ml in speech & audio processing" from 10th to 12th of December 2020. he also took part in the employability skill development program conducted by zensar. he has also completed few certification courses from coursera, cursapp & sololearn**.**

G Rithik Reddy is currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. He completed his schooling till 10th grade in Sri Chaitanya techno school and he completed his Intermediate in Narayana junior college. his technical skills include c, c++, java, python, MySQL. he has done an internship in java at lasya infotech. his participations include: a national level seminar on "recent trends in cloud computing, fog, and edge computing" on 18th and 19th of June 2021 and a national level three-day workshop on "ai & ml in speech & audio processing" from 10th to 12th of December 2020. he also took part in the employability skill development program conducted by zensar. he has also completed few certification courses from coursera, cursapp & sololearn**.**

K Manisyam is currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin’s Engineering College He completed his schooling till 10th grade in Kakatheeya High school and he completed his Intermediate in Sri Chaitanya junior college. His technical skills include c, c++, java, python, MySQL. He has done an internship in java and python at lasya infotech. His participations include: a national level seminar on "recent trends in cloud computing, fog, and edge computing" on 18th and 19th of June 2021 and a national level three-day workshop on "AI & ML in speech & audio processing" from 10th to 12th of December 2020 and national level python programming conducted in 2020. He also took part in the employability skill development program conducted by zensar. he has also completed few certification courses from coursera, cursapp & sololear

APPENDICES

```
FROM TKINTER IMPORT MESSAGEBOX

FROM TKINTER IMPORT *

FROM TKINTER IMPORT SIMPLEDIALOG

IMPORT TKINTER

IMPORT MATPLOTLIB.PYPLOT AS PLT

IMPORT NUMPY AS NP

FROM TKINTER IMPORT TTK

FROM TKINTER IMPORT FILEDIALOG

IMPORT PANDAS AS PD

FROM SKLEARN.MODEL_SELECTION IMPORT TRAIN_TEST_SPLIT

FROM KERAS.UTILS.NP_UTILS IMPORT TO_CATEGORICAL

FROM KERAS.MODELS IMPORT SEQUENTIAL

FROM KERAS.LAYERS.CORE IMPORT DENSE,ACTIVATION,DROPOUT, FLATTEN

FROM SKLEARN.NEIGHBORS IMPORT KNEIGHBORSCLASSIFIER

FROM SKLEARN IMPORT SVM

FROM SKLEARN.METRICS IMPORT ACCURACY_SCORE

IMPORT CV2

FROM KERAS.LAYERS IMPORT CONVOLUTION2D

FROM KERAS.LAYERS IMPORT MAXPOOLING2D

FROM SKLEARN.LINEAR_MODEL IMPORT LOGISTICREGRESSION

FROM SKLEARN.DECOMPOSITION IMPORT PCA

MAIN = TK()
```

```
MAIN.TITLE("VEHICLE PATTERN RECOGNITION USING MACHINE & DEEP LEARNING TO PREDICT CAR MODEL")

MAIN.GEOMETRY("1300X1200")

GLOBAL FILENAME

GLOBAL X, Y

GLOBAL MODEL

GLOBAL X_TRAIN, X_TEST, Y_TRAIN, Y_TEST

ACCURACY = []

GLOBAL XX

GLOBAL CLASSIFIER

NAMES = ['AM GENERAL HUMMER SUV 2000', 'ACURA RL SEDAN 2012', 'ACURA TL SEDAN 2012', 'ACURA TL TYPE-S 2008', 'ACURA TSX SEDAN 2012']

FROM TKINTER IMPORT MESSAGEBOX

FROM TKINTER IMPORT *

FROM TKINTER IMPORT SIMPLEDIALOG

IMPORT TKINTER

IMPORT MATPLOTLIB.PYPLOT AS PLT

IMPORT NUMPY AS NP

FROM TKINTER IMPORT TTK

FROM TKINTER IMPORT FILEDIALOG

IMPORT PANDAS AS PD

FROM SKLEARN.MODEL_SELECTION IMPORT TRAIN_TEST_SPLIT

FROM KERAS.UTILS.NP_UTILS IMPORT TO_CATEGORICAL

FROM KERAS.MODELS IMPORT SEQUENTIAL

FROM KERAS.LAYERS.CORE IMPORT DENSE,ACTIVATION,DROPOUT, FLATTEN
```

```
FROM SKLEARN.NEIGHBORS IMPORT KNEIGHBORSCLASSIFIER

FROM SKLEARN IMPORT SVM

FROM SKLEARN.METRICS IMPORT ACCURACY_SCORE

IMPORT CV2

FROM KERAS.LAYERS IMPORT CONVOLUTION2D

FROM KERAS.LAYERS IMPORT MAXPOOLING2D

FROM SKLEARN.LINEAR_MODEL IMPORT LOGISTICREGRESSION

FROM SKLEARN.DECOMPOSITION IMPORT PCA

MAIN = TK()

MAIN.TITLE("VEHICLE PATTERN RECOGNITION USING MACHINE & DEEP LEARNING TO PREDICT CAR MODEL")

MAIN.GEOMETRY("1300X1200")

GLOBAL FILENAME

GLOBAL X, Y

GLOBAL MODEL

GLOBAL X_TRAIN, X_TEST, Y_TRAIN, Y_TEST

ACCURACY = []

GLOBAL XX

GLOBAL CLASSIFIER

NAMES = ['AM GENERAL HUMMER SUV 2000', 'ACURA RL SEDAN 2012', 'ACURA TL SEDAN 2012', 'ACURA TL TYPE-S 2008', 'ACURA TSX SEDAN 2012']
```

A

PROJECT REPORT

On

Blockchain E-Voting Done Right: Privacy and Transparency with Public Blockchain

Submitted by

1) M.Srujith (17K81A05G3) 2) M.Rohan (17K81A05F9)

3) K.Rohith (17K81A05E9) 4) M.Ranadeep (17K81A05F7)

in partial fulfillment for the award of the

degree of

BACHELOR OF TECHNOLOGY

IN

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Under the Guidance of

Mr.Ch.Sathyanarayana

Assistant Professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ST.MARTIN'S ENGINEERING COLLEGE
An Autonomous Institute

Dhulapally, Secunderabad – 500 100

JUNE 2021

BONAFIDE CERTIFICATE

This is to certify that the project entitled Blockchain E-Voting Done Right: Privacy and Transparency with Public Blockchain, is being submitted by **1.Mr.MUPPIDI SRUJITH 17K81A05G3 2.Mr.RANADEEP MANCHIKATLA 17K81A05F7 3.Mr.ROHAN MARUGAI 17K81A05F9 4.Mr.KARNATAKA ROHITH GOUD 17K81A5E9** in partial fulfillment of the requirement for the award of the degree of **BACHELOR OF TECHNOLOGY IN DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING** is recorded of bonafide work carried out by them. The result embodied in this report have been verified and found satisfactory.

Signature

Mr.Ch.Sathyanarayana

Department of CSE

Head of the Department

Dr.M.NARAYANAN

Department of CSE

Internal Examiner

External Examiner

Place:

Date:

DECLARATION

We, the student of **Bachelor of Technology** in Department of Computer Science and Engineering', session: 2017 – 2021, St. Martin's Engineering College, Dhulapally, Kompally, Secunderabad, hereby declare that work presented in this Project Work entitled Blockchain E-Voting Done Right: Privacy and Transparency with Public Blockchain is the outcome of our own bonafide work and is correct to the best of our knowledge and this work has been undertaken taking care of Engineering Ethics. This result embodied in this project report has not been submitted in any university for award of any degree.

M. Srujith	17K81A05G3
M. Ranadeep	17K81A05F7
K. Rohith Goud	17K81A05E9
M. Rohan	17K81A05F9

ABSTRACT

Some forms of voting have been here ever since. Mostly used form all over the world are paper ballots. Electronic voting schemes are being popular only in the last decade and they are still unsolved. E-voting schemes bring problems mainly regarding security, credibility, transparency, reliability, and functionality. Estonia is the pioneer in this field and maybe considered the state of the art. But there are only a few solutions using blockchain. Blockchain can deliver an answer to all of the mentioned problems and furthermore bring some advantages such as immutability and decentralization. The main problems of technologies utilizing blockchain for e-voting are their focus on only one field or lack of testing and comparison. The project presents a blockchain-based e-voting platform, which can be used for any kind of voting. It is fully utilizedby blockchain and all processes can be handled within it. After the start of the voting, the platform behaves as fully independent and decentralized without possibilities to affect the voting process. The data are fully transparent, but the identity of voters is secured by homomorphic encryption. The solution is tested and compared in three different blockchains. The results show, that both public and private blockchains can be used with only a little difference in the speed. The key novelty of our solution is a fully decentralized management of e-voting platform through blockchain, transparency of the whole process and at the same time security and privacy of the voters thanksto homomorphic encryption.

Keywords—blockchain; e-voting; smart contract; Ethereum; Hyperledger Composer; elections; homomorphic encryption

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible and whose encouragements and guidance have crowded effects with success.

We extended our deep sense of gratitude to Principal**,** **Dr. P. SANTOSH KUMAR PATRA**, St. Martin's Engineering College, Dhulapally, for permitting us to undertake this project.

We are also thankful to **Dr. M.NARAYANAN**, Head of the Department, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

We would like to thank **Dr. T. POONGOTHAI,** Department of Computer Science and Engineering (AI & ML) for her encouragement and insightful comments and as well as our project coordinators **Dr. B.RAJALINGAM**, Associate Professor and **Mr. J.SUDHAKAR**, Assistant Professor, in Department of Computer Science and Engineering for their valuable support.

We would like to express our sincere gratitude and indebtedness to our project supervisor Mr.Ch.Sathyanarayana, Assistant Professor, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

Finally, we express thanks to all those who have helped us successfully to completing this project. Furthermore, we would like to thank our family and friends for their moral support and encouragement.

We express thanks to all those who have helped us in successfully completing the project.

M. Srujith 17K81A05G3
M. Ranadeep 17K81A05F7
K. Rohith Goud 17K81A05E9
M. Rohan 17K81A05F9

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF OUTPUT SCREENS

LIST OF ABBREVIATIONS

AVI	Audio Video Interlace
BMP	Bitmap
CPU	Central Processing Unit
GB	Giga Bytes
GUI	Graphical User Interface
ZKP	Zero-knowledge proof

INTRODUCTION

The topic of e-voting systems is still at an early stage of development. This domain was not only for its recency but also because there are not many solutions that address problems of e-voting. Nowadays, popularity grows also in the development of e-Government. However, such a system is not feasible if basic services for citizens such as elections do not become electronic. "E-voting is one of the key public sectors that can be transformed by blockchain technology". Hand by hand with e-voting come also new challenges, which need to be addressed. One of them is e.g. securing the elections, which needs to be at least as safe as the classic voting systems with ballots. That is why it's decided to create safe elections in which voters do not have to worry about someone abusing the electoral system.

In recent years blockchain is often mentioned as an example of secure technology used in an online environment. Our e-voting system uses blockchain to manage all election processes. Its main advantage is that there is no need for confidence in the centralized authority that created the elections. This authority cannot affect the election results in our system. Another challenge in e-voting is the lack of transparency in the functioning of the system, leading to a lack of confidence in voters. This problem is solved by blockchain in a way of total transparency that allows everyone to see the stored data and processes such as how are these data handled. In the field of security, this technology is more suitable in every way than the classic e-voting platform without blockchain. The article is structured as follows. Section II shows a brief analysis of existing blockchain e-voting solutions. In Section III, It presents the design of our solution and describe all of its components. The evaluation is in Section IV and the results are discussed in Section V with further conclusions.

BLOCK CHAIN:

Blockchain is a distributed database of records, often referred to as a ledger. The basic principle of blockchain is in the immutability of the records already written in blocks. Advanced cryptography is used to ensure the chaining of blocks, providing data integrity. Another feature is the type of network communication. Network nodes communicate with a client client method. There is no third party to communicate between clients and therefore no trust in this person is required. The network participant's true identity is unknown. In this section, it will analyze some of the existing solutions for e-voting systems based on blockchain technology.

A. Agora

The only successful project that was partially demonstrated in the state elections was created by the Swiss company Agora, which has the best technology and research in this area. The elections took place in the western part of Sierra Leone. They were not completely decentralized and could not meet the eligibility requirements for elections via the Internet in the given area. The choices were made as regular options with paper ballots. Identity of citizens was verified by ID card and then they put their selection into the ballot box. These ballots were then manually inserted into a private blockchain Bulletin Board developed by Agora. A private blockchain differs from a public, such as Ethereum, in that, which nodes are allowed to validate transactions and blocks. In private, these operations can be only performed by trusted nodes that have been previously recognized as trusted. Other nodes, ordinary network users, are part of the network and can see all the data if allowed to. In the mentioned election method, there was no need to address the problems of anonymity of voters. Paper ballots were counted by officials of the State Election Commission. The number of paper votes was compared with the number of votes on the blockchain. Thus, the final counting of the votes did not proceed through blockchain. However, the solutions that Agora delivers have a perspective in this regard. Agora offers solutions for elections to governments and institutions. Agora uses its token to make choices. Governments must pay for each citizen a pre-determined amount for which a sufficient number of tokens for election is allocated to the voter. The electoral system is a multi-layer architecture using a blockchain, called Bulletin Board, which is based on Skip chain architecture. Bulletin Board data are cryptographically bounded to Bitcoin blockchain via the Cotena layer that provides immutability and decentralization of stored data. Bulletin Board uses nodes recognized as authorities (cothority). These confirm transactions where each such node in the network has a full copy of all transactions. The Cotena layer is a change-resistant mechanism built on the Bitcoin blockchain. This layer provides the integrity of data layer records to a decentralized system. Cotena has been designed to take advantage of Bitcoin's blockchain security and has introduced a design that has low memory requirements and low fees for using Bitcoin transactions. Skip chain allows software clients to navigate through large numbers of blocks both forward and backward. In doing so, it provides proof of transaction validity without the need for a full copy of the records contained in the blockchain.

B. Netvote

Netvote is a decentralized application based on blockchain technology for elections and works on the Ethereum network . Netvote provides an environment with decentralized applications (dApp) for network users. DApp designed for admin allows to create elections, set election rules, set voter registration rules, create ballot boxes and setup voting. DApp for voters allows them to sign up for elections and vote for the

selected candidate. After closing the elections, you can view the results using the appropriate application. The application allows the administrator to choose one of three types of voting. The first type is open elections where everyone who has an account in the Ethereum network can vote. The second type is private elections where only authenticated voters can vote. The last option allows only voters who have the required amount of correct tokens issued specifically for the elections. In Netvote, each election consists of multiple smart contracts deployed on the Ethereum network. These smart contracts are created by an administrator through his dApp. Every electoral district is a smart contract. The smart contract includes a list of ballots also represented by separate smart contracts. The voter registers in a given electoral district and all actions subsequently performed through his dApp are processed in that district. Netvote, as it's mentioned, includes the option of private elections where only authenticated voters can participate in. The solution that Netvote brings is called the Vote Gateway and serves to verify voters' identity. The voter sends his signed ballot through his dApp. If the voter is registered, the Vote Gateway selects the voter's private key from the vault where this unique private key is stored. Subsequently, the SHA3 cryptographic function is executed along with the ballot and the private key to create the anonymous identifier of the voter. A ballot mapped to an anonymous identity is sent to the electoral district via a transaction. Netvote is a good solution for both state elections and institutions. The drawback is the need for a lot of smart contracts and transactions which highers the costs of deployment and also brings scalability issues. Netvote uses Ethereum's public blockchain. The architecture of Netvote needs to be refactored. Maybe using private blockchain elections where the number of transactions processed per second increases significantly is also interesting.

C. OV-net

The OV-net (Open voting network) is a 2-rounddecentralized electoral protocol implemented on Ethereum blockchain. This protocol has several advantages. One of them is counting the votes that the protocol does itself without the necessary authority. Privacy is maximized. The only time a voter's choice might be revealed is if all the other nodes in the network are fraudulent. Each user can check others for compliance with the protocol. The protocol consists of five parts: 1. Setup - The election's administrator is responsible for uploading a valid voters' list to the smart contract when it is started. 2. Signing up - Voters will send their electoral key and use Zero-knowledge proof (ZKP) to confirm the electoral key. Ethereum confirms the correctness of the ZKP and stores the electoral key. 3. Voting - Voters will send an encrypted vote. This vote can be either 1 (yes) or 0 (no). Ethereum verifies that the vote is only one of the options 1 or 0, and then stores it. 4. Voting Count - Ethereum counts the votes when all of them have been sent. OV-net is a well-managed protocol that supports multiple necessary functions required for e-voting. However, it has several disadvantages. The only voting options are yes or no. The whole implementation is based on the

principle of two options for voting for the proper functioning of the ZKP. Our application needs to make choices for multiple candidates with their names. OV-net is also not eligible for a large number of voters because of its specific implementation. Another disadvantage is the need to vote for each voter. If one voter did not vote, then the elections could not be evaluated or counted. Several schemes for e-voting have been published in recent years, but most do not have documentation and they lack information about the internal operation of the service. One of the services is Follow my Vote. This system does not answer multiple questions about how it works in a blockchain network. Another service is Bit Congress, which was designed to work with multiple protocols like Bitcoin and Master coin. Finally, both applications were not put into operation.

1.1 PROJECT OVERVIEW

Despite the digitalization of several important aspects of modern life, elections are still largely conducted offline, on paper. Since the turn of the century, e-voting has been considered a promising and (eventually) inevitable development, which could speed up, simplify and reduce the cost of elections, and might even lead to higher voter turnouts and the development of stronger democracies. E-voting could take many forms: using the internet or a dedicated, isolated network; requiring voters to attend a polling station or allowing unsupervised voting; using existing devices, such as mobile phones and laptops, or requiring specialist equipment. It has a further choice; to continue trusting central authorities to manage elections or to use blockchain technology to distribute an open voting record among citizens.

1.2 PROJECT OBJECTIVES

The objective of Blockchain e-voting is a means of logging and verifying records that is transparent and distributed among users. Usually, votes are recorded, managed, counted and checked by a central authority. Blockchain-enabled e-voting (BEV) would empower voters to do these tasks themselves, by allowing them to hold a copy of the voting record. The historic record could then not be changed because other voters would see that the record differs from theirs. Illegitimate votes could not be added, because other voters would be able to scrutinize whether votes were compatible with the rules (perhaps because they have already been counted, or are not associated with a valid voter record). BEV would shift power and trust away from central actors. The project presents electronic voting system using blockchain, a secure and robust system that ensures anonymity of the voter, transparency in the process, and robust functioning

1.3 ORGANIZATION OF CHAPTERS

Besides the introduction, the thesis is organized in other six chapters as follows:

2. LITERATURE SURVEY: This review consist of the blockchain system using API to verify the voter ID and create a decentralized system where the data is fully transparent, the given papers provide solutions using blockchain.

3. SOFTWARE AND HARDWARE REQUIREMENTS: this discuss about the software and hardware required for the execution of the project.

4. SOFTWARE DEVELOPMENT ANALYASIS: this explains the assumptions and technical specifications of the project.

5. PROJECT SYSTEM DESIGN: this explains all the software development process with DFD and UML diagrams clearly.

6. PROJECT CODING: this explains the design of the system, roles and responsibilities.

7. PROJECT TESTING: this explains various test cases to test the project working.

8. OUTPUT SCREENS: explains a step by step process of the project execution.

9. EXPERIMENTAL RESULTS: tests and results are shown and explained in this chapter. The results are analysed in the context of the thesis project and followed by discussion on systems throughput and resiliency, as well as the approaches to testing and analysis.

10. CONCLUSION AND FUTURE ENHANCEMENT: this ends the project with a short summary of the main concepts mentioned in the thesis as well as the relevant results.

2 LITERATURE SURVEY

A literature survey or a literature review in a project report is that section which shows the various analyses and research made in the field of your interest and the results already published, taking into account the various parameters of the project and the extent of the project.

It is the most important part of your report as it gives you a direction in the area of your research. It helps you set a goal for your analysis - thus giving you your problem statement.

When you write a literature review in respect of your project, you have to write the researches made by various analysts - their methodology (which is basically their abstract) and the conclusions they have arrived at. You should also give an account of how this research has influenced your thesis.

2.1 SURVEY ON BACKGROUND

1. Adida, B., Helios (2008).“Web-based open-audit voting.”, in Proceedings of the 17th Conference on Security Symposium, ser. SS’08. Berkeley, CA, USA: USENIX Association, 2008. This paper proposes associated justify an adequate security model and criteria to judge comprehensibility. It additionally describe a web ballot theme, Pretty graspable Democracy, show that it satisfies the adequate security model which it’s a lot of graspable than Pretty smart Democracy, presently the sole theme that additionally satisfies the planned security model. 2. Chaum, D., Essex, A., Carback, R., Clark, J., Popoveniuc, S., Sherman, A. and Vora, P. (2008).“Scantegrity: End-to-end voter-veriable optical- scan voting.", IEEE Security Privacy, vol. 6, no. 3, pp. 40-46, May 2008. The paper describes Scantegrity that minimally impacts election procedures and is the first independent E2E verification mechanism that preserves optical scan as the underlying voting system and doesn’t interfere with a manual recount. 3. Dalia, K., Ben, R. , Peter Y. A, and Feng, H. (2012). “A fair and robust voting system by broadcast.”, 5th International Conference on E-voting, 2012. The paper proposes a recovery round to enable the election result to be announced if voters abort and also added a commitment round to ensure fairness. In addition, it also provided a computational security proof of ballot secrecy. 4. Bell, S., Benaloh, J., Byrne, M. D., Debeauvoir, D., Eakin, B., Kortum, P., McBurnett, N., Pereira, O., Stark, P. B., Wallach, D. S., Fisher, G., Montoya, J., Parker, M. and Winn, M. (2013).“Star-vote: A secure, transparent, auditable, and reliable voting system.”, in 2013 Electronic Voting Technology Workshop/Workshop on Trustworthy Elections (EVT/WOTE 13). Washington, D.C.: USENIX Association, 2013. The paper describes the STAR-Vote design, that may preferably be the next-generation electoral system for Travis County and maybe elsewhere.

2.2 CONCLUSIONS ON SURVEY

Recent major technical challenges relating to e-voting systems embrace, however not restricted to secure digital identity management. Any potential citizen ought to are registered to the electoral system before the elections. Their data ought to be in a very digitally processable format. Besides, their identity data ought to be unbroken personal in any involving information. Ancient E-voting system could face following problems:

- Anonymous vote-casting.
- Individualized ballot processes.
- Ballot casting verifiability by (and only by) the voter.
- High initial setup costs.
- Increasing security problems.
- Lack of transparency and trust.
- Voting delays or inefficiencies related to remote/absentee voting.

The paper proposes an electronic voting system using multi-chain. It shows how multichain can be configured to restrict transactions to only one vote between voter and contestant. A new entity – trusted third party – was introduced to keep the voting secret.

3 SOFTWARE AND HARDWARE REQUIREMENTS

3.1 SOFTWARE REQUIREMENTS

For developing the application the following are the Software Requirements:

- **Operating system:** Windows 10.
- **Coding Language :** Python.
- **Front-End :** Python.
- **Data Base :** MySQL.
- **Designing :** Html, css, javascript.

3.2 HARDWARE REQUIREMENTS

- **System :** Intel i5 Generation.
- **Hard Disk :** 1 TB HDD
- **Monitor :** 14’ Colour Monitor.
- **Mouse :** Optical Mouse.
- **Ram :** 8 GB Ram

PYTHON

Python is a general-purpose interpreted, interactive, object-oriented, and high-level programming language. An interpreted language, Python has a design philosophy that emphasizes code readability

(notably using whitespace indentation to delimit code blocks rather than curly brackets or keywords), and a syntax that allows programmers to express concepts in fewer lines of code than might be used in languages such as C++or Java. It provides constructs that enable clear programming on both small and large scales. Python interpreters are available for many operating systems. CPython, the reference implementation of Python, is open source software and has a community-based development model, as do nearly all of its variant implementations. CPython is managed by the non-profit Python Software Foundation. Python features a dynamic type system and automatic memory management. It supports multiple programming paradigms, including object-oriented , imperative, functional and procedural , and has a large and comprehensive standard library

DJANGO

Django is a high-level Python Web framework that encourages rapid development and clean, pragmatic design. Built by experienced developers, it takes care of much of the hassle of Web development, so you can focus on writing your app without needing to reinvent the wheel. It’s free and open source.

Django's primary goal is to ease the creation of complex, database-driven websites. Django emphasizes reusability and "pluggability" of components, rapid development, and the principle of don't repeat yourself. Python is used throughout, even for settings files and data models.

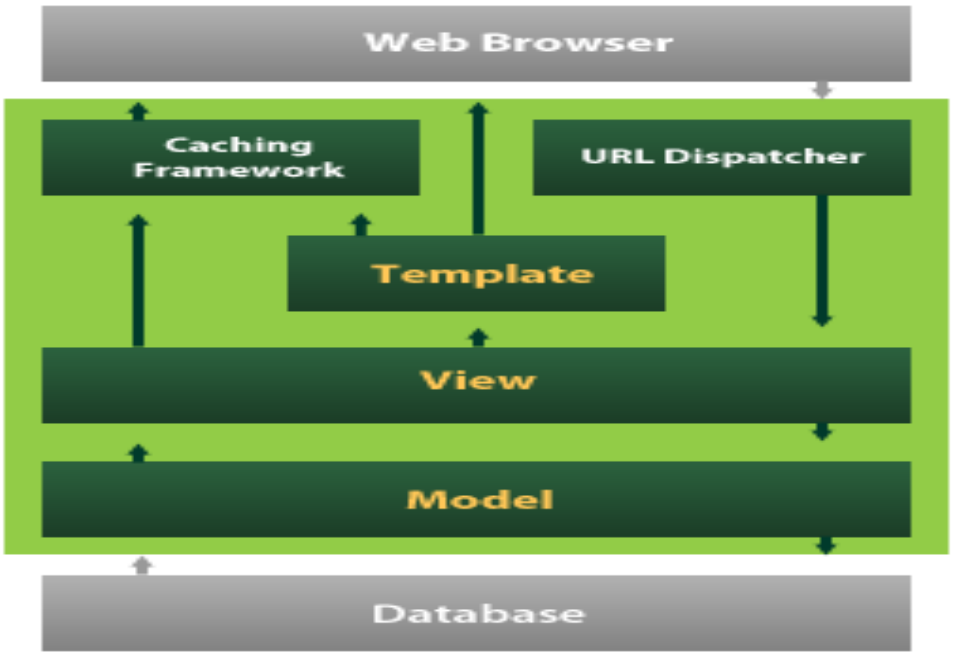


Figure 3(a)

Django also provides an optional administrative create, read, update and delete interface that is generated dynamically through introspection and configured via admin models

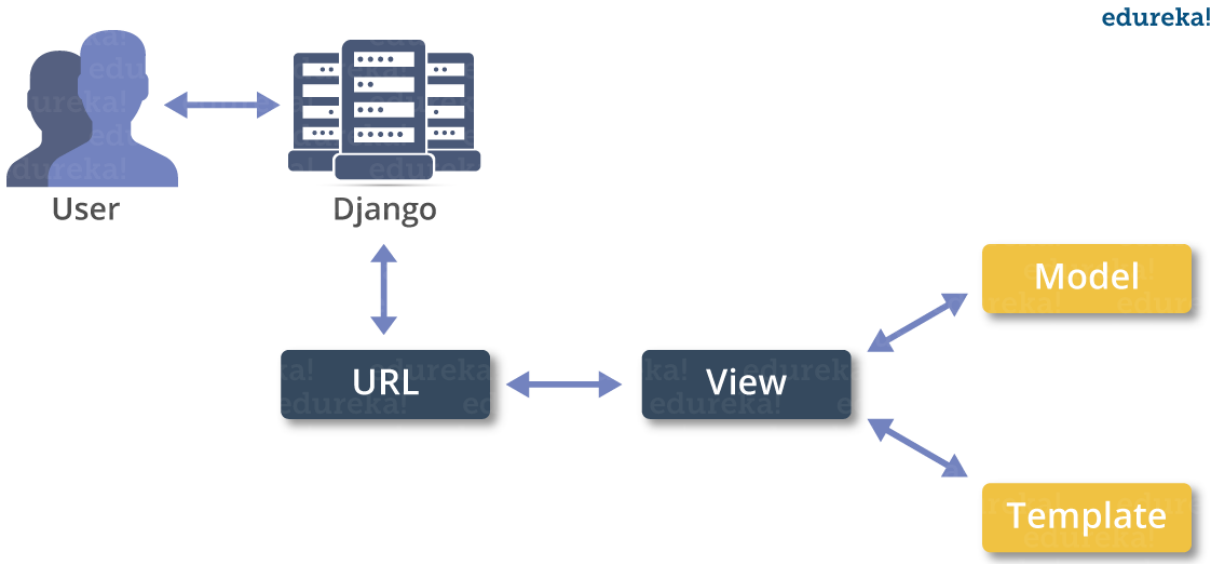


Figure 3(b)

4 SOFTWARE DEVELOPMENT ANALYASIS

4.1 OVERVIEW OF PROBLEM

E-voting schemes bring problems mainly regarding security, credibility, transparency, reliability, and functionality. Estonia is the pioneer in this field and may be considered the state of the art. But there are only a few solutions using blockchain. Blockchain can deliver an answer to all of the mentioned problems and furthermore bring some advantages such as immutability and decentralization. The main problems of technologies utilizing blockchain for e-voting are their focus on only one field or lack of testing and comparison The project presents a blockchain-based e-voting platform, which can be used for any kind of voting.

4.2 DEFINE THE PROBLEM

• **Anonymous vote-casting:** Each vote may or may not contain any choice per candidate, should be anonymous to everyone including the system administrators, after the vote is submitted through the system.

• **Individualized ballot processes:** How a vote are depicted within the involving net applications or databases continues to be AN open discussion. whereas a transparent text message is that the worst plan, a hashed token is wont to offer obscurity and integrity. Meanwhile, the vote ought to be non-reputable, that can't be bonded by the token resolution.

• **Ballot casting verifiability by (and only by) the voter:** The elector ought to be ready to see and verify his/her own vote, when he/she submitted the vote. this is often vital to realize so as to forestall, or a minimum of to note, any potential malicious activity. This counter live, except for providing suggests that of non-repudiation, can sure boost the sensation of trust of the voters. These issues area unit partly self-addressed in some recent applications. Yet, suggests that of e-voting is presently in use in many countries together with Brazil, uk, Japan, and Republic of Estonia.

• **High initial setup costs:** Though sustaining and maintaining on-line selection systems is way cheaper than ancient elections, initial deployments could be pricy, particularly for businesses.

• **Increasing security problems:** Cyber attacks cause an excellent threat to the general public polls. nobody would settle for the responsibility if associate degree hacking try succeeds throughout an election. The

DDoS attacks are documented and largely not the case within the elections. The citizen integrity commission of the u. s. gave an affidavit concerning the state of the elections within the North American country recently. Accordingly; Ronald Rivest explicit that “hackers have myriad ways in which of assaultive pick machines”. As associate degree example; barcodes on ballots and smartphones in pick locations may be utilized in the hacking method. Apple explicit that the people tend to mustn’t ignore the actual fact that computers are hackable, and also the evidences will simply be deleted. Double-voting or voters from the opposite regions also are some common issues.

4.3 MODULES OVERVIEW

The proposed system uses a web based application that is created to serve as a front end application which enables the users to interact with the system. The application is implemented in PHP language and uses MYSQL database as the backend for the application. In Blockchain, it is using public python Blockchain API’s to store and manage voting data as Blockchain provides secure and tamper proof of data storage and to implement this project which have designed following modules.

1. Admin Module

2. User Module

4.4 DEFINE THE MODULES

Admin module: this user responsible to add new party and candidate details and can view party details and vote count. Admin login to system by using username as ‘admin’ and password as ‘admin’. Performs functions such as creating elections , adding candidates ,displaying candidate’s and voter’s details ,verifying hashed blocks of data and viewing results.

User Module: This module performs two major functions of allowing the voters to cast votes and viewing graphical results instantly. The user has to sign up with the application by using username as his ID and then upload his face photo which capture from webcam. After registering user can go for login which validate user id and after successful login user can go for cast vote module which execute following functionality.

4.5 MODULE FUNCTIONALITY

The users are provided a user interface from where they can view the candidates, register votes, and check the reports. The website will be active only during the Election period.. The voter submits their personal credentials such as voter's name, address, identity card number, date of birth etc to the server. Once the voter credentials are verified by the application server, the voter can then proceed to view the candidates, cast votes and view results. The verification helps in the prevention of fraud. To confirm the identity of the voter ,a One Time Password(OTP) is send to the contact number of the voter. On verification the voter is allowed to cast the vote. On verification, the user can view the candidates electing the contest according to their legislature and can register their votes. These votes are flooded to the peer nodes in the form of transactions .Once all the nodes verify the transaction to be valid ,these nodes are added to blockchain. The Election Commission receives the votes in form of a report. The blockchain technology ensures that the votes are tamper proof and complete privacy is ensured.

First user will be connected to his PC webcam and then image will be capture

Using OpenCV application will detect face and then using CNN application will predict user identify and if user identity matched with CNN predicted face then application will display all voting candidates list.

If user not casted vote then user can give vote to desire candidate by clicking link beside party name or candidate name.

Upon giving vote application will capture voter and candidate details and then encrypt the data and then store in Blockchain

5 PROJECT SYSTEM DESIGN

5.1 DFDS IN CASE OF DATABASE PROJECTS

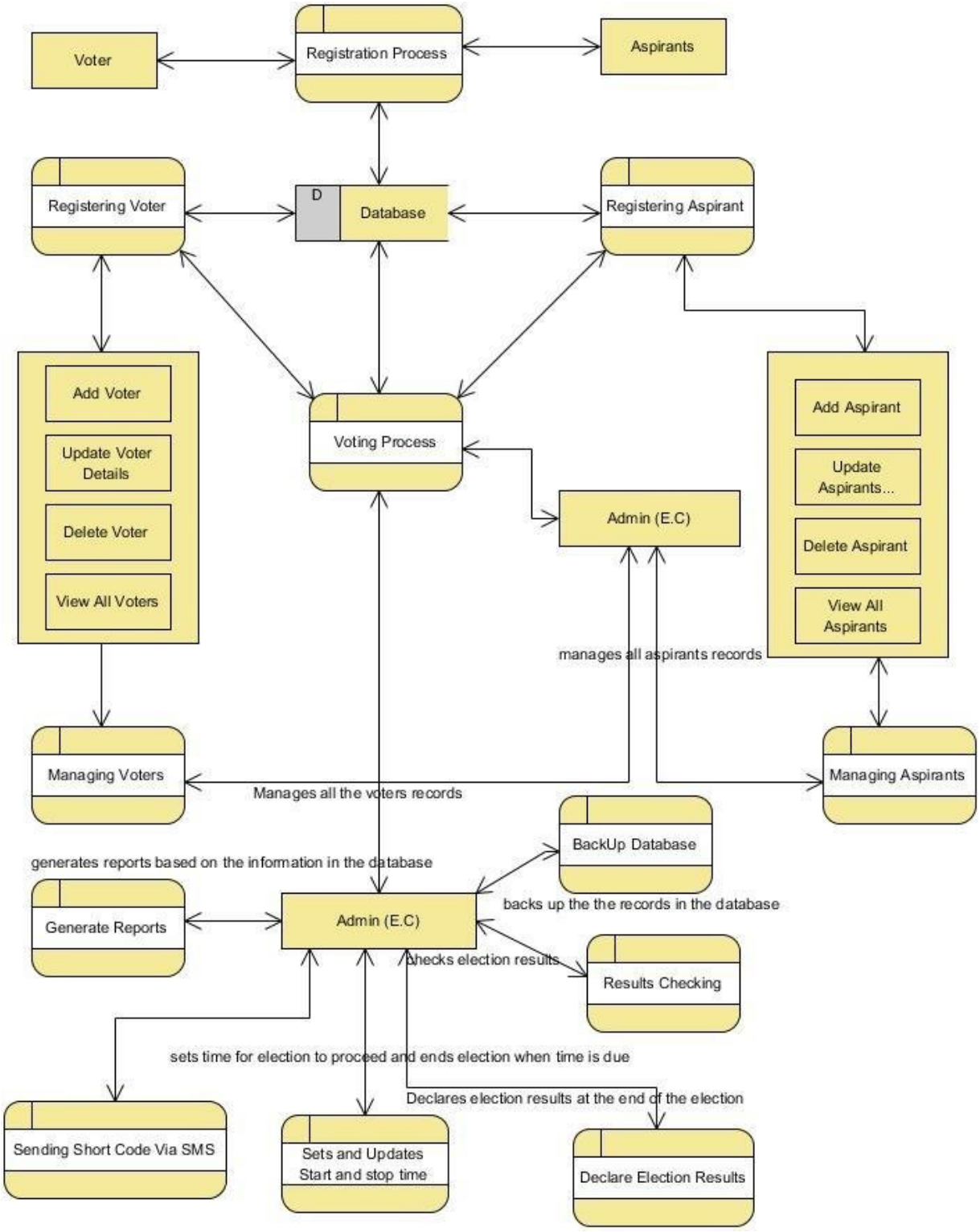


Figure 5.1

5.2 E-R DIAGRAMS

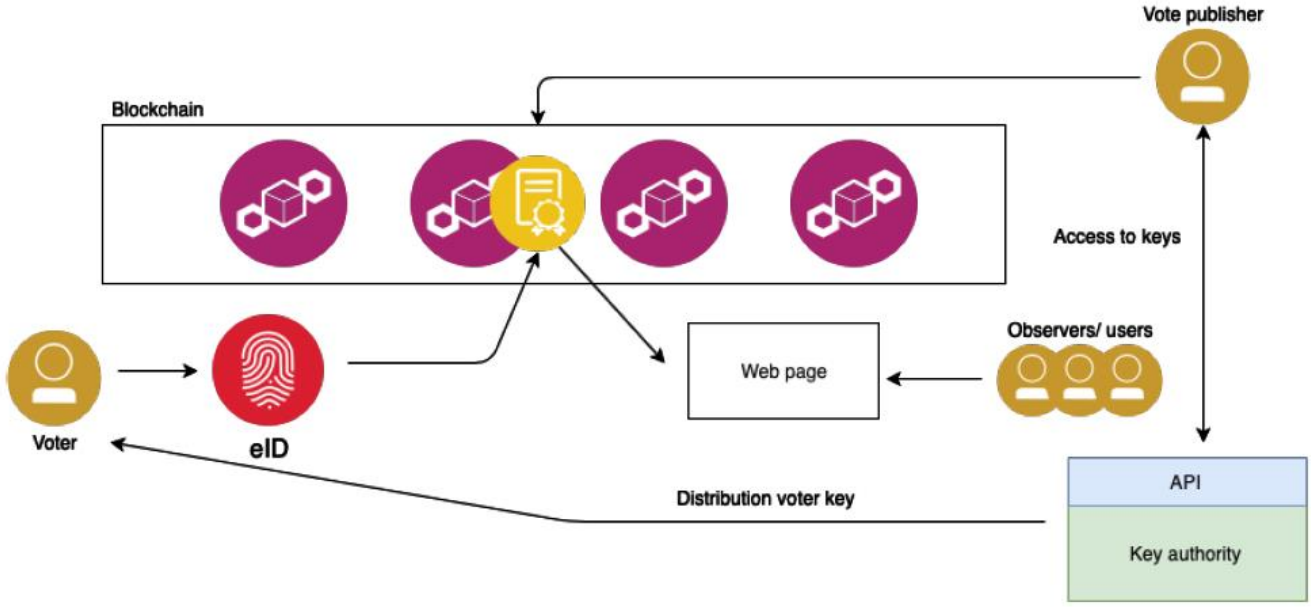


Figure 5.2(a)

5.2 UML DIAGRAMS

UML stands for Unified Modeling Language. UML is a standardized general-purpose modeling language in the field of object-oriented software engineering. The standard is managed, and was created by, the Object Management Group.

The goal is for UML to become a common language for creating models of object oriented computer software. In its current form UML is comprised of two major components: a Meta-model and a notation. In the future, some form of method or process may also be added to; or associated with, UML.

The Unified Modeling Language is a standard language for specifying, Visualization, Constructing and documenting the artifacts of software system, as well as for business modeling and other non-software systems.

The UML represents a collection of best engineering practices that have proven successful in the modeling of large and complex systems.

The UML is a very important part of developing objects oriented software and the software development process. The UML uses mostly graphical notations to express the design of software projects.

GOALS:

The Primary goals in the design of the UML are as follows:

1. Provide users a ready-to-use, expressive visual modeling Language so that they can develop and exchange meaningful models.
2. Provide extendibility and specialization mechanisms to extend the core concepts.
3. Be independent of particular programming languages and development process.
4. Provide a formal basis for understanding the modeling language.
5. Encourage the growth of OO tools market.
6. Support higher level development concepts such as collaborations, frameworks, patterns and components.
7. Integrate best practices.

USE CASE DIAGRAM:

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.

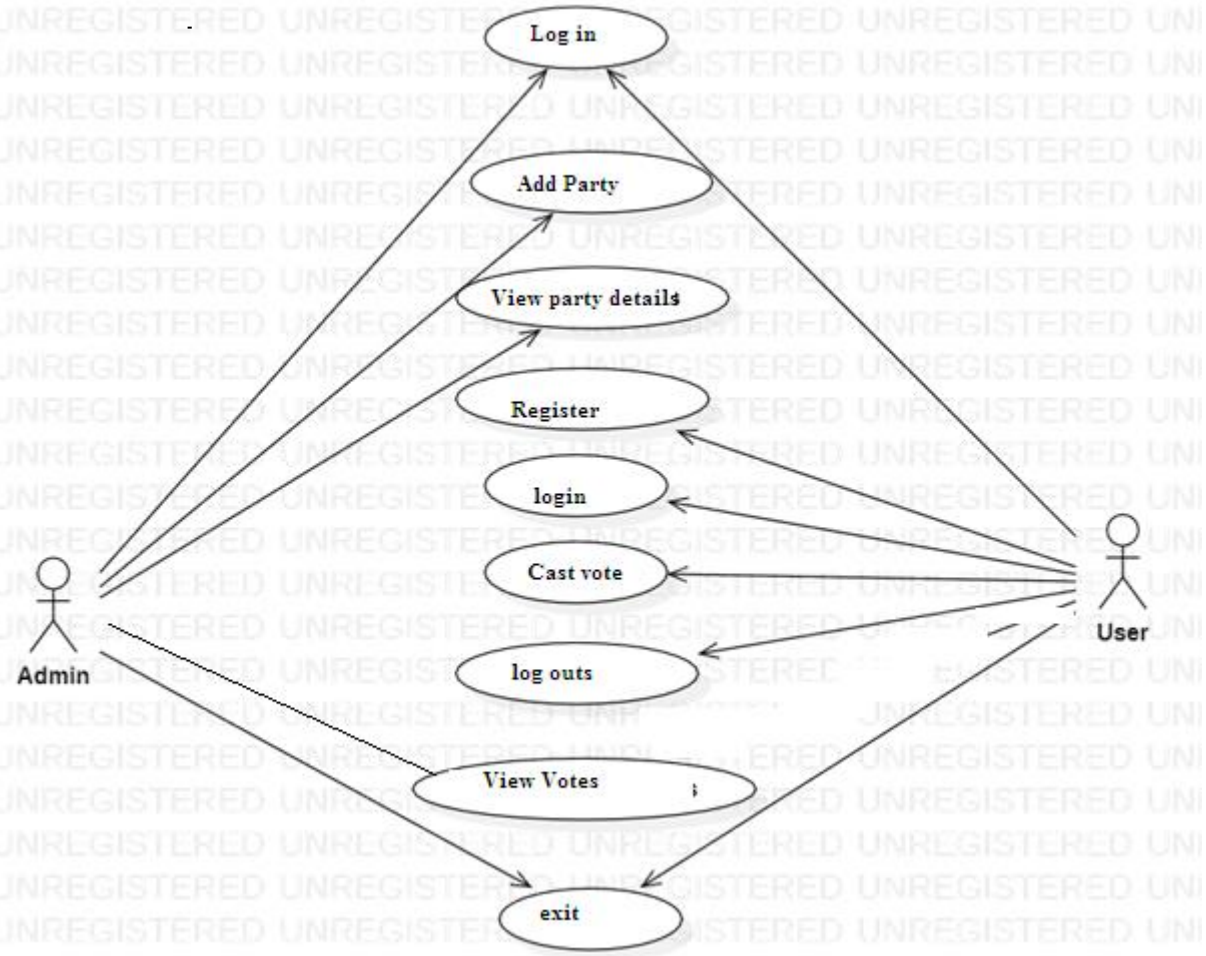


Figure 5.2(b)

CLASS DIAGRAM:

In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.

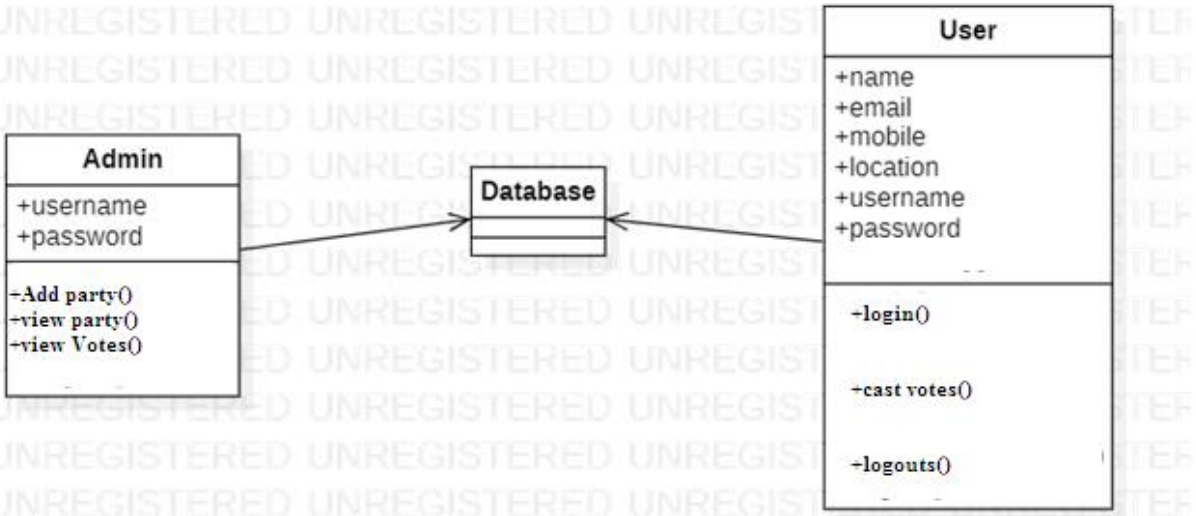


Figure 5.2(c)

ACTIVITY DIAGRAM:

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modeling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.

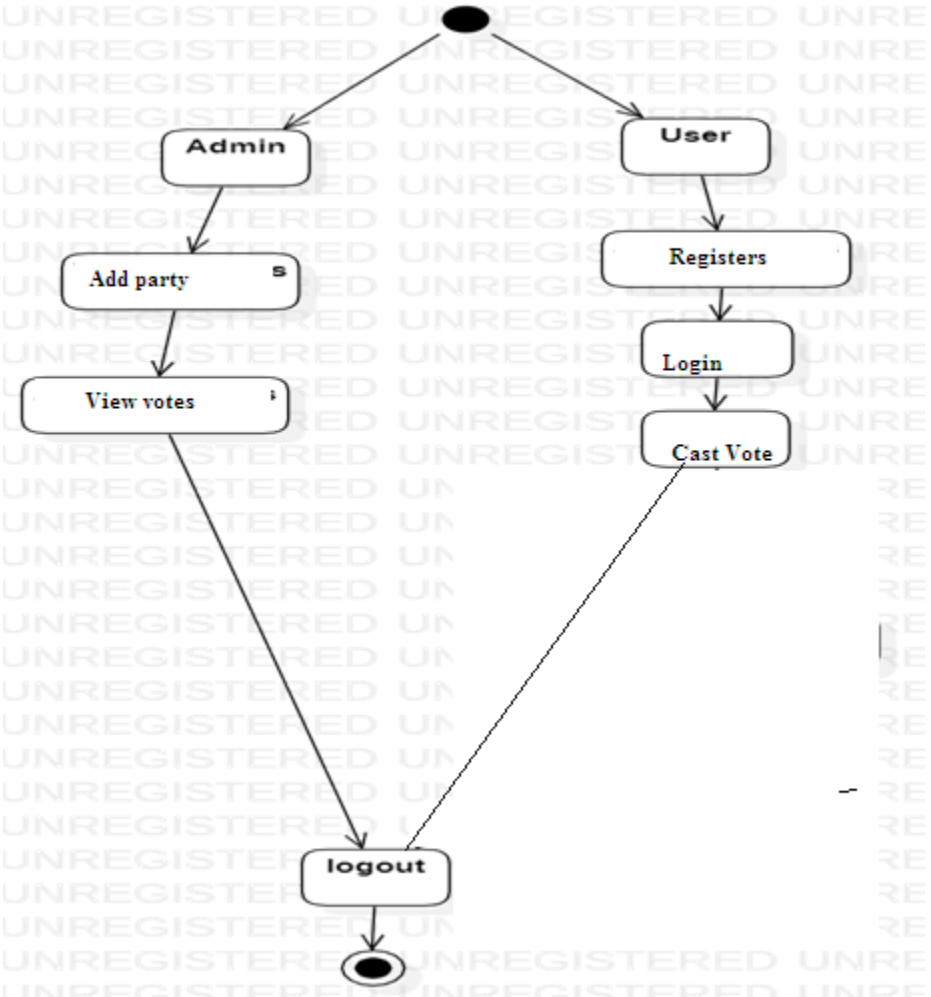


Figure 5.2(d)

6 PROJECT CODING

6.1 CODE TEMPLATE

views.py - Notepad

File Edit Format View Help

```
from django.shortcuts import render
from django.template import RequestContext
from django.contrib import messages
import pymysql
from django.http import HttpResponse
from django.core.files.storage import FileSystemStorage
import os
from keras.utils.np_utils import to_categorical
from keras.layers import  MaxPooling2D
from keras.layers import Dense, Dropout, Activation, Flatten
from keras.layers import Convolution2D
from keras.models import Sequential
from keras.models import model_from_json
from Blockchain import *
from Block import *
from datetime import date
import cv2
import numpy as np
import pyaes, pbkdf2, binascii, os, secrets
import base64

global load_model
load_model = 0
global classifier

blockchain = Blockchain()
if os.path.exists('blockchain_contract.txt'):
```

Ln 1, Col 1 100% Windows (CRLF) UTF-8

Figure 6.1(a)

views.py - Notepad

File Edit Format View Help

```
    with open('blockchain_contract.txt', 'rb') as fileinput:
        blockchain = pickle.load(fileinput)
    fileinput.close()

names = ['Azizullah Karimi', 'Hasibullah Atayi', 'khushhal Qasimyar', 'Nazif Mal',
         'pacha khan khogyani', 'Ramin paikar', 'SayedAhmad Seyar sawiz', 'Venket Rao',
'wahidullah Bdri', 'yahya Maqsidi']

face_detection = cv2.CascadeClassifier('haarcascade_frontalface_default.xml')


def getKey(): #generating key with PBKDF2 for AES
    password = "s3cr3t*c0d3"
    passwordSalt = '76895'
    key = pbkdf2.PBKDF2(password, passwordSalt).read(32)
    return key

def encrypt(plaintext): #AES data encryption
    aes = pyaes.AESModeOfOperationCTR(getKey(), pyaes.Counter
(311295470350000473029524339676541953981242398445663228841721636378460562482 23))
    ciphertext = aes.encrypt(plaintext)
    return ciphertext

def decrypt(enc): #AES data decryption
    aes = pyaes.AESModeOfOperationCTR(getKey(), pyaes.Counter
(311295470350000473029524339676541953981242398445663228841721636378460562482 23))
    decrypted = aes.decrypt(enc)
```

Ln 9, Col 39 100% Windows (CRLF) UTF-8

Figure 6.1(b)

views.py - Notepad

File Edit Format View Help

```
    return decrypted


def AddParty(request):
    if request.method == 'GET':
       return render(request, 'AddParty.html', {})

def index(request):
    if request.method == 'GET':
       return render(request, 'index.html', {})

def Login(request):
    if request.method == 'GET':
       return render(request, 'Login.html', {})

def CastVote(request):
    if request.method == 'GET':
       return render(request, 'CastVote.html', {})


def Register(request):
    if request.method == 'GET':
       return render(request, 'Register.html', {})

def Admin(request):
    if request.method == 'GET':
       return render(request, 'Admin.html', {})
```

Ln 9, Col 39 | 100% | Windows (CRLF) | UTF-8

Figure6.1(c)

views.py - Notepad

File Edit Format View Help

```
def WebCam(request):
    if request.method == 'GET':
        data = str(request)
        formats, imgstr = data.split(';base64,')
        imgstr = imgstr[0:(len(imgstr)-2)]
        data = base64.b64decode(imgstr)
        with open('C:/Python/EVoting/EVotingApp/static/photo/test.png', 'wb') as f:
            f.write(data)
        f.close()
        context= {'data':"done"}
        return HttpResponse("Image saved")

def checkUser(name):
    flag = 0
    for i in range(len(blockchain.chain)):
          if i > 0:
              b = blockchain.chain[i]
              data = b.transactions[0]
              data = base64.b64decode(data)
              data = str(decrypt(data))
              data = data[2:len(data)-1]
              print(data)
              arr = data.split("#")
              if arr[0] == name:
                  flag = 1
                  break
    return flag

def getOutput(status):
    output = '<h3><br/>'+status+'<br/><table border=1 align=center>'
    output+='<tr><th><font size=3 color=black>Candidate Name</font></th>'
    output+='<th><font size=3 color=black>Party Name</font></th>'
    output+='<th><font size=3 color=black>Area Name</font></th>'
    output+='<th><font size=3 color=black>Image</font></th>'
    output+='<th><font size=3 color=black>Cast Vote Here</font></th></tr>'
    con = pymysql.connect(host='127.0.0.1',port = 3308,user = 'root', password = 'root',
database = 'evoting',charset='utf8')
    with con:
        cur = con.cursor()
        cur.execute("select * FROM addparty")
```

Ln 9, Col 39 | 100% | Windows (CRLF) | UTF-8

Figure6.1(d)

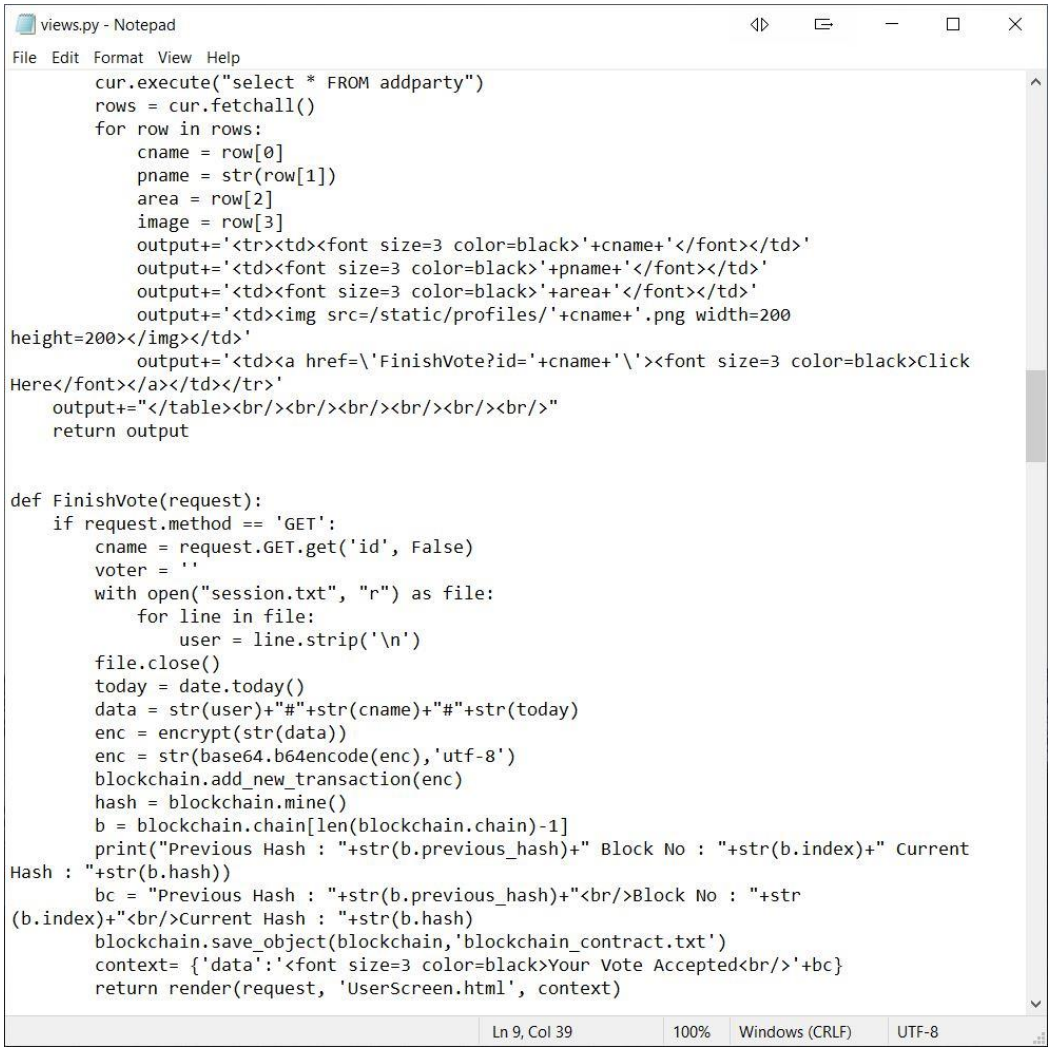


Figure6.1(e)

6.2 OUTLINE FOR VARIOUS FILES

The project uses Python programming to implement our project along with Django. A single python file is used to implement our code. This file consists of various modules that has been used. Our project modules are – Admin Module and User Modules. It also used various python modules like json, time, pickle, pyaes and base64.

6.3 METHODS INPUT AND OUTPUT PARAMETERS

INPUT DESIGN

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the following things:

- What data should be given as input?
- How the data should be arranged or coded?
- The dialog to guide the operating personnel in providing input.
- Methods for preparing input validations and steps to follow when error occur.

OBJECTIVES

1.Input Design is the process of converting a user-oriented description of the input into a computer-based system. This design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerized system.

2. It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The goal of designing input is to make data entry easier and to be free from errors. The data entry screen is designed in such a way that all the data manipulates can be performed. It also provides record viewing facilities.

3.When the data is entered it will check for its validity. Data can be entered with the help of screens. Appropriate messages are provided as when needed so that the user will not be in maize of instant. Thus the objective of input design is to create an input layout that is easy to follow

OUTPUT DESIGN

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making.

1. Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is designed so that people will find the system can use easily and effectively. When analysis design computer output, they should Identify the specific output that is needed to meet the requirements.

2.Select methods for presenting information.

3.Create document, report, or other formats that contain information produced by the system.

The output form of an information system should accomplish one or more of the following objectives.

- Convey information about past activities, current status or projections of the
- Future.
- Signal important events, opportunities, problems, or warnings.
- Trigger an action.
- Confirm an action.

7 PROJECT TESTING

7.1 VARIOUS TEST CASES

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub-assemblies, assemblies and/or a finished product. It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

TYPES OF TESTS

Unit testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

Integration testing

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

Functional test

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items:

Valid Input : identified classes of valid input must be accepted.

Invalid Input : identified classes of invalid input must be rejected.

Functions : identified functions must be exercised.

Output : identified classes of application outputs must be exercised.

Systems/Procedures : interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

System Test

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

Unit Testing

Unit testing is usually conducted as part of a combined code and unit test phase of the software lifecycle, although it is not uncommon for coding and unit testing to be conducted as two distinct phases.

Test strategy and approach

Field testing will be performed manually and functional tests will be written in detail.

Test objectives

- All field entries must work properly.
- Pages must be activated from the identified link.

- The entry screen, messages and responses must not be delayed.

Features to be tested

- Verify that the entries are of the correct format
- No duplicate entries should be allowed
- All links should take the user to the correct page.

Integration Testing

Software integration testing is the incremental integration testing of two or more integrated software components on a single platform to produce failures caused by interface defects.

The task of the integration test is to check that components or software applications, e.g. components in a software system or – one step up – software applications at the company level – interact without error.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

Acceptance Testing

User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

7.2 BLACK BOX TESTING

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box .you cannot "see" into it. The test provides inputs and responds to outputs without considering how the software works. The below Black-Box can be any software system you want to test. For Example, an operating system like Windows, a website like Google, a database like Oracle or even your own custom application. Under Black Box Testing, you can test these applications by just focusing on the inputs and outputs without knowing their internal code implementation

Figure7.2

Various approaches to black-box testing

There are a set of approaches for black-box testing.
Manual UI Testing: In this approach, a tester checks the system as a user. Check and verify the user data, error messages.

Automated UI Testing: In this approach, user interaction with the system is recorded to find errors and glitches. Testers can set record demand as per schedule.

Documentation Testing: In this approach, a tester purely checks the input and output of the software. Testers consider what system should perform rather than how. It is a manual approach to testing.

The tester doesn't need any technical knowledge to test the system. It is essential to understand the user's perspective.
Testing is performed after development, and both the activities are independent of each other.
It works for a more extensive coverage which is usually missed out by testers as they fail to see the bigger picture of the software.
Test cases can be generated before development and right after specification.
Black box testing methodology is close to agile.

7.3 WHITE BOX TESTING

The box testing approach of software testing consists of black box testing and white box testing. Here the white box testing which also known as glass box is **testing, structural testing, clear box testing, open box testing and transparent box testing**.

It tests internal coding and infrastructure of a software focus on checking of predefined inputs against expected and desired outputs. It is based on inner workings of an application and revolves around internal structure testing. In this type of testing programming skills are required to design test cases. The primary goal of white box testing is to focus on the flow of inputs and outputs through the software and strengthening the security of the software.

The term 'white box' is used because of the internal perspective of the system. The clear box or white box or transparent box name denote the ability to see through the software's outer shell into its inner workings.

Developers do white box testing. the developer will test every line of the code of the program. The developers perform the White-box testing and then send the application or the software to the testing team, where they will perform the black box testing and verify the application along with the requirements and identify the bugs and sends it to the developer.

The developer fixes the bugs and does one round of white box testing and sends it to the testing team. Here, fixing the bugs implies that the bug is deleted, and the particular feature is working fine on the application.

Here, the test engineers will not include in fixing the defects for the following reasons:

- Fixing the bug might interrupt the other features. Therefore, the test engineer should always find the bugs, and developers should still be doing the bug fixes.
- If the test engineers spend most of the time fixing the defects, then they may be unable to find the other bugs in the application.

8 OUTPUT SCREENS

8.1 USER INTERFACES

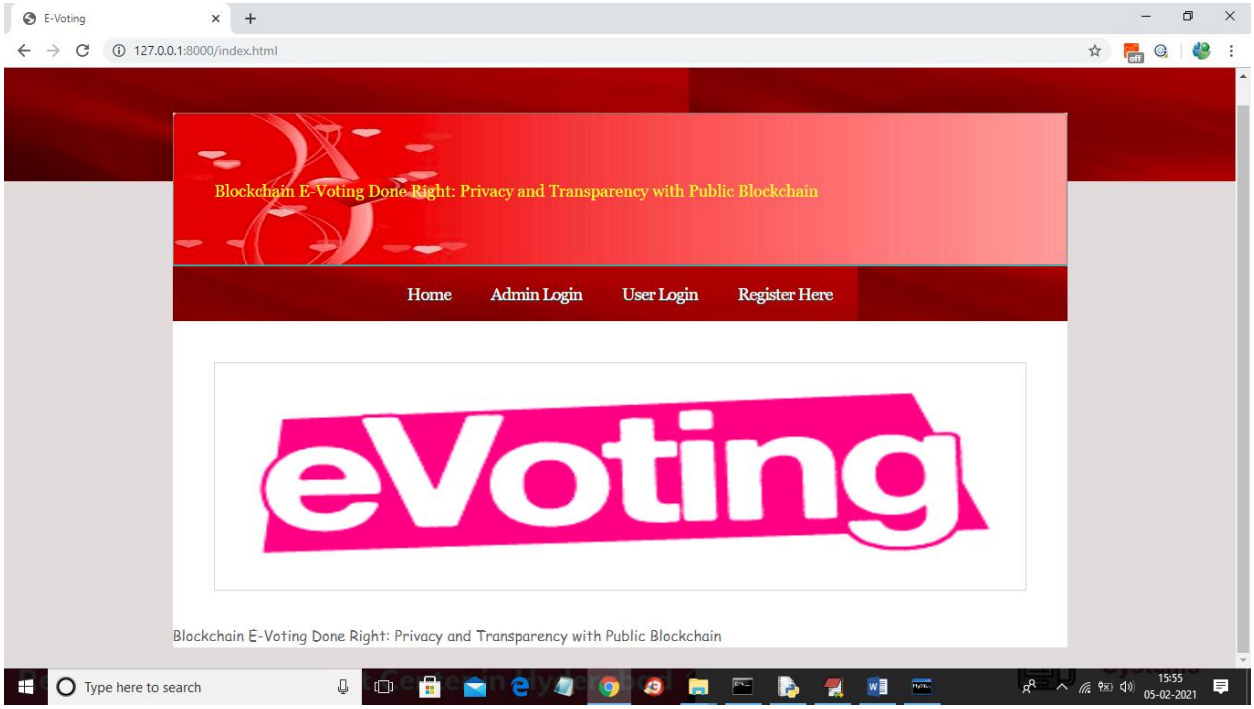


Figure8.1(a) In above screen click on 'Admin Login' link to get below screen

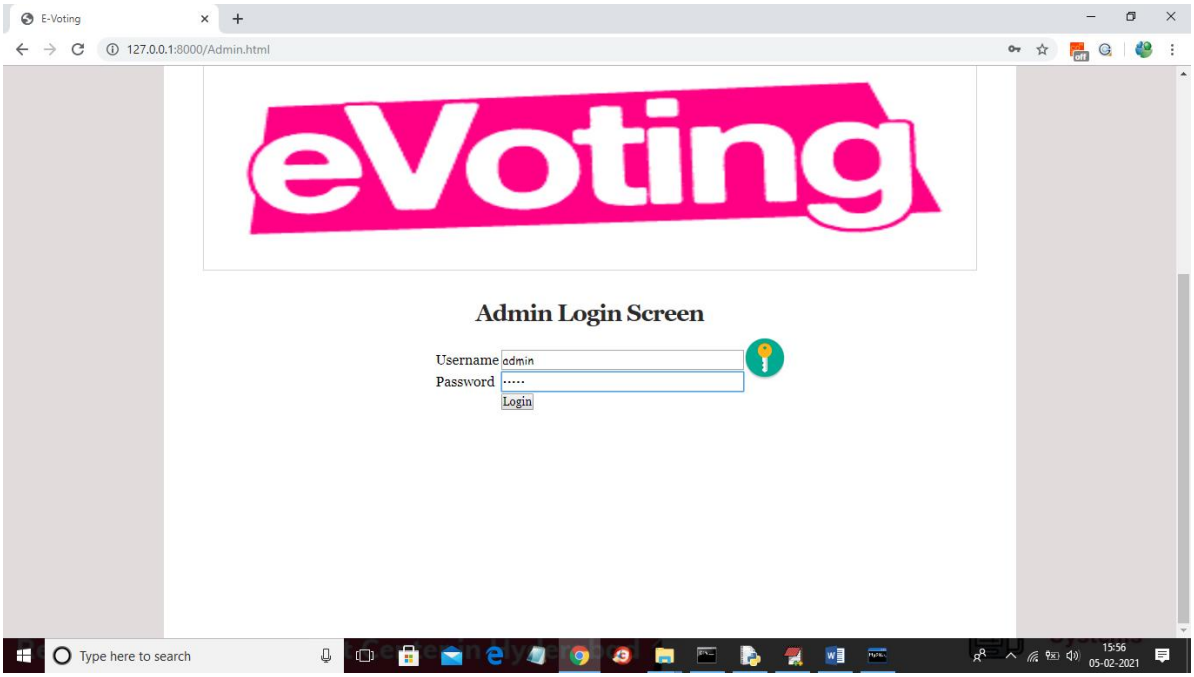


Figure8.1(b) In above screen login as admin by giving username as 'admin' and password as 'admin' and then click Login button to get below screen

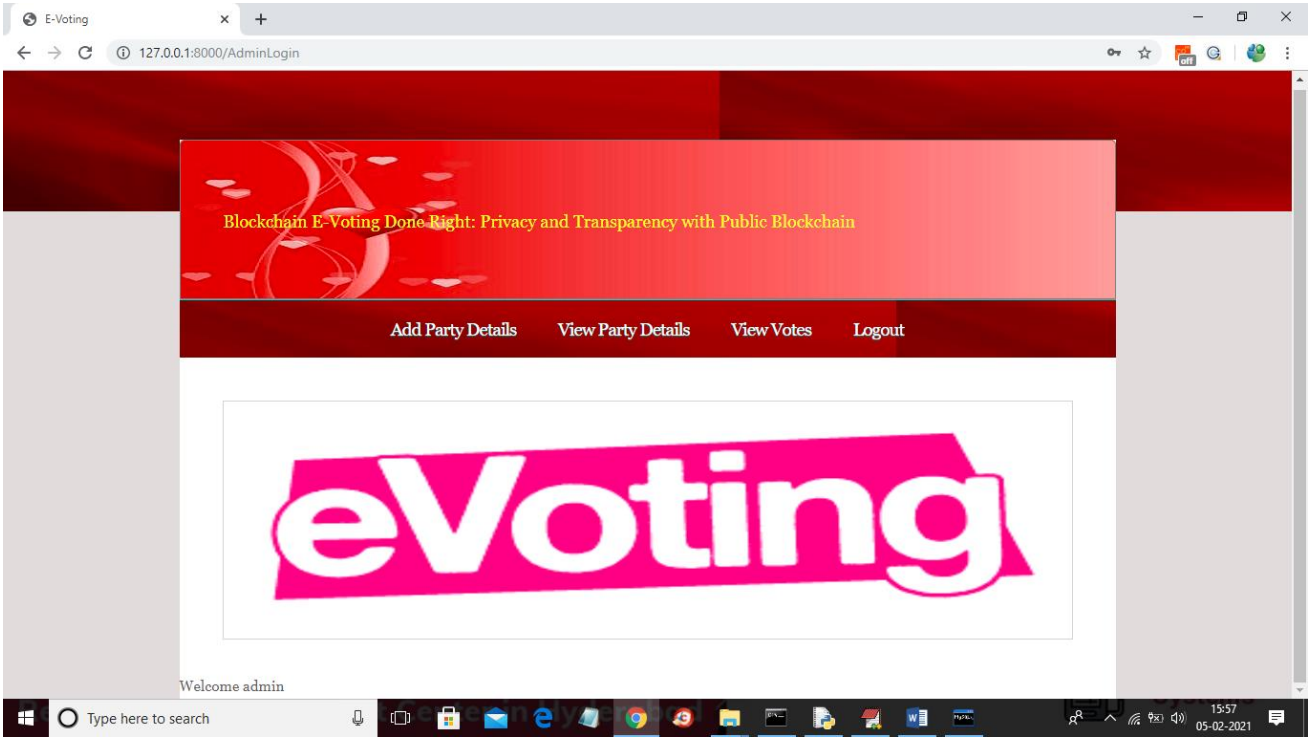


Figure8.1(c) In above screen admin can click on 'Add Party Details' link to add party details

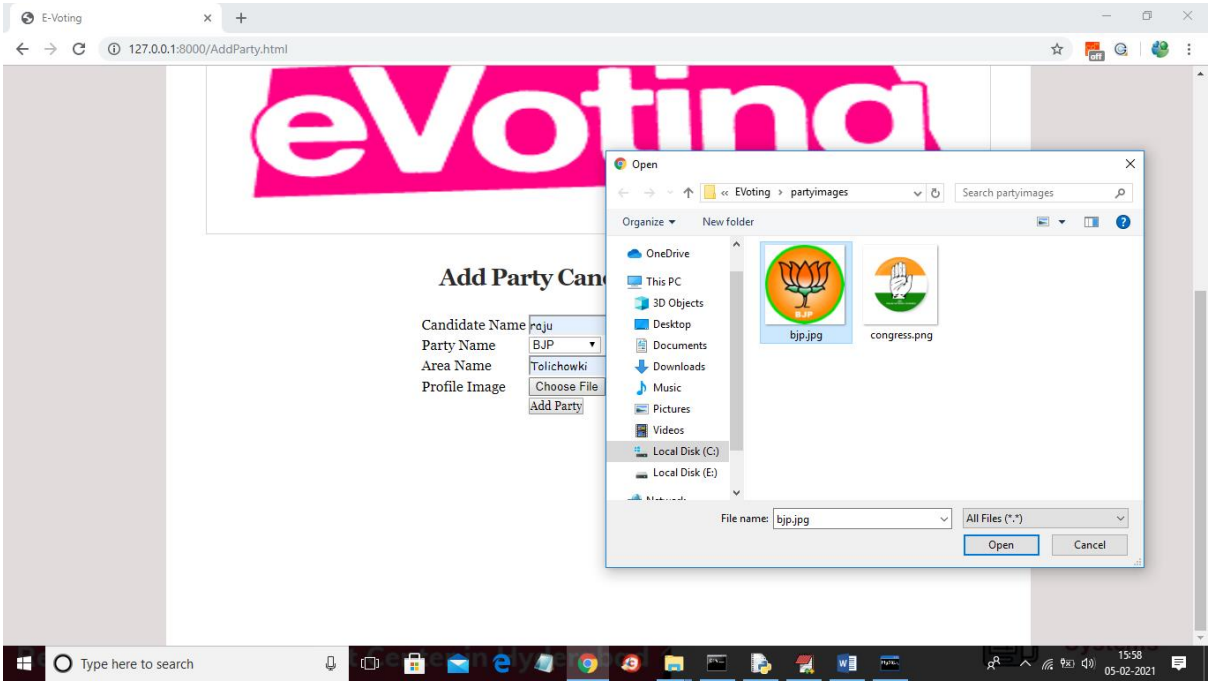


Figure8.1(d)In above screen adding party and candidate details and then upload image and click on 'Open' button then click on 'Add Party' button to add party details

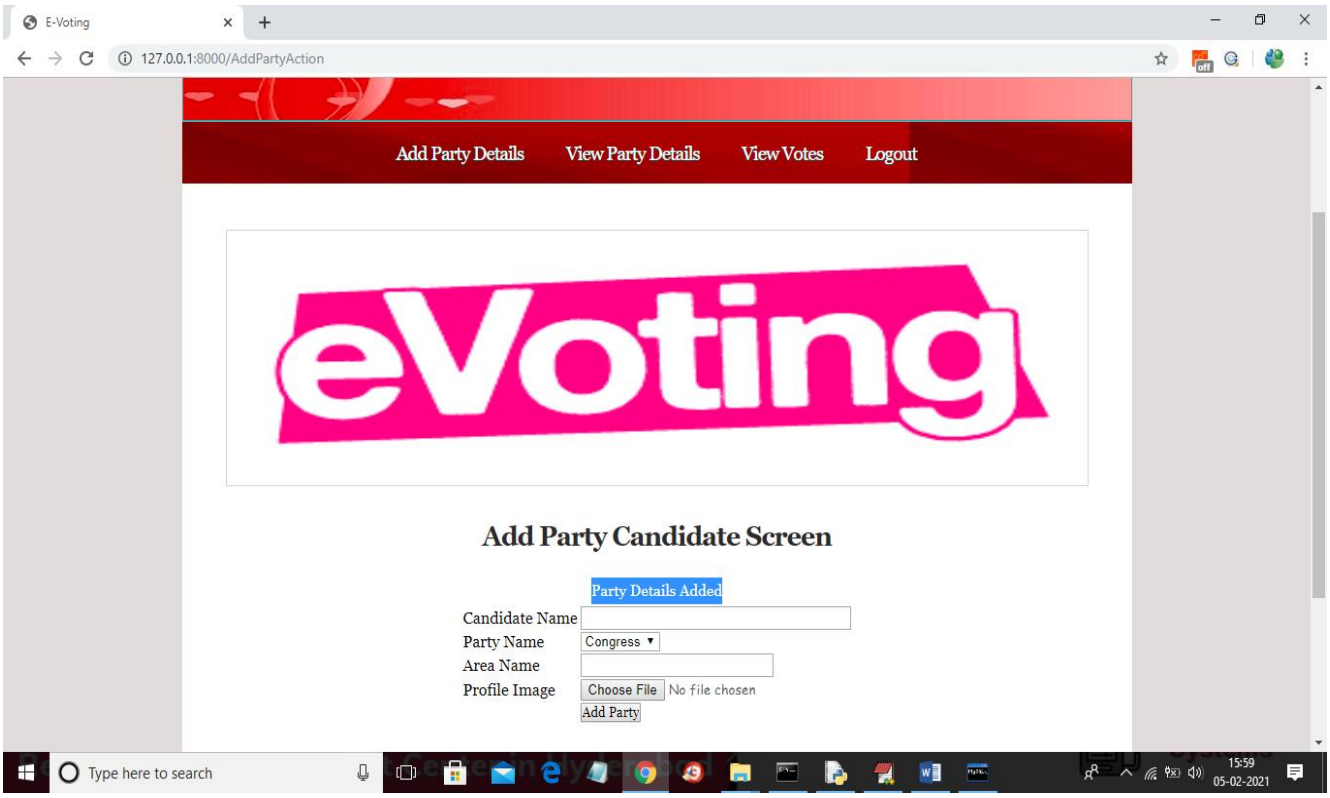


Figure8.1(e)In above screen part details added and similarly you can add any number of party members and now click on 'View Party Details' link to get below screen

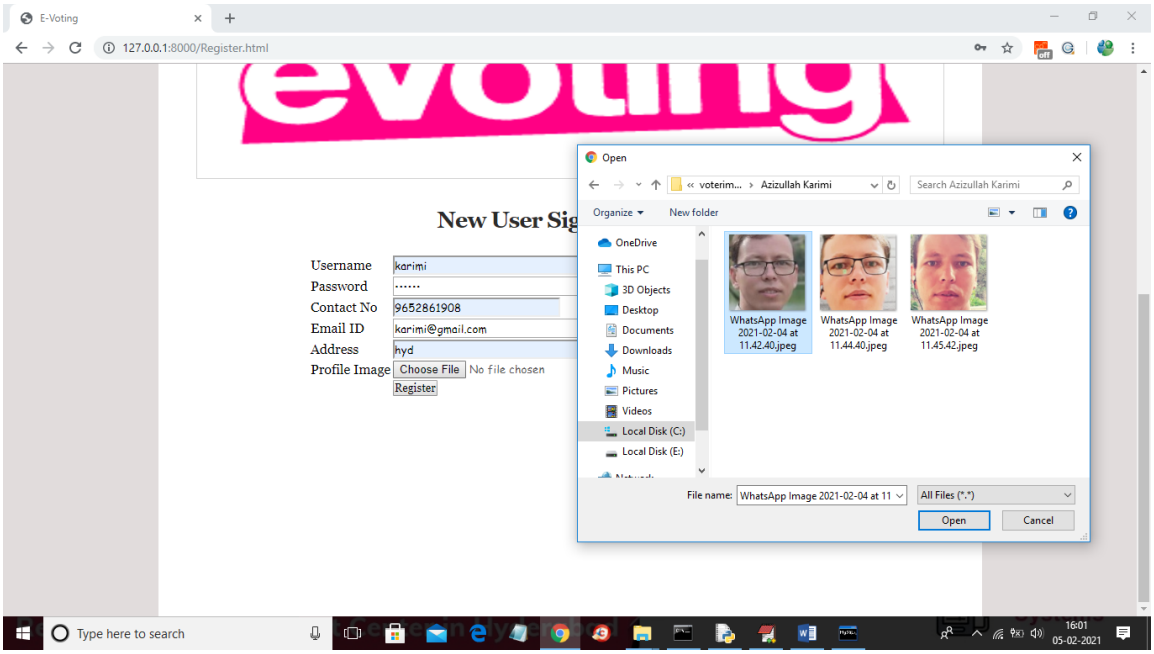


Figure8.1(f)In above screen adding new user and then selecting his face photo taken from webcam and then click on 'Register' button to complete signup process. Here you have given images taken from phone but it needs to capture from webcam for dataset as quality of webcam image and phone image vary and then problem comes in prediction.

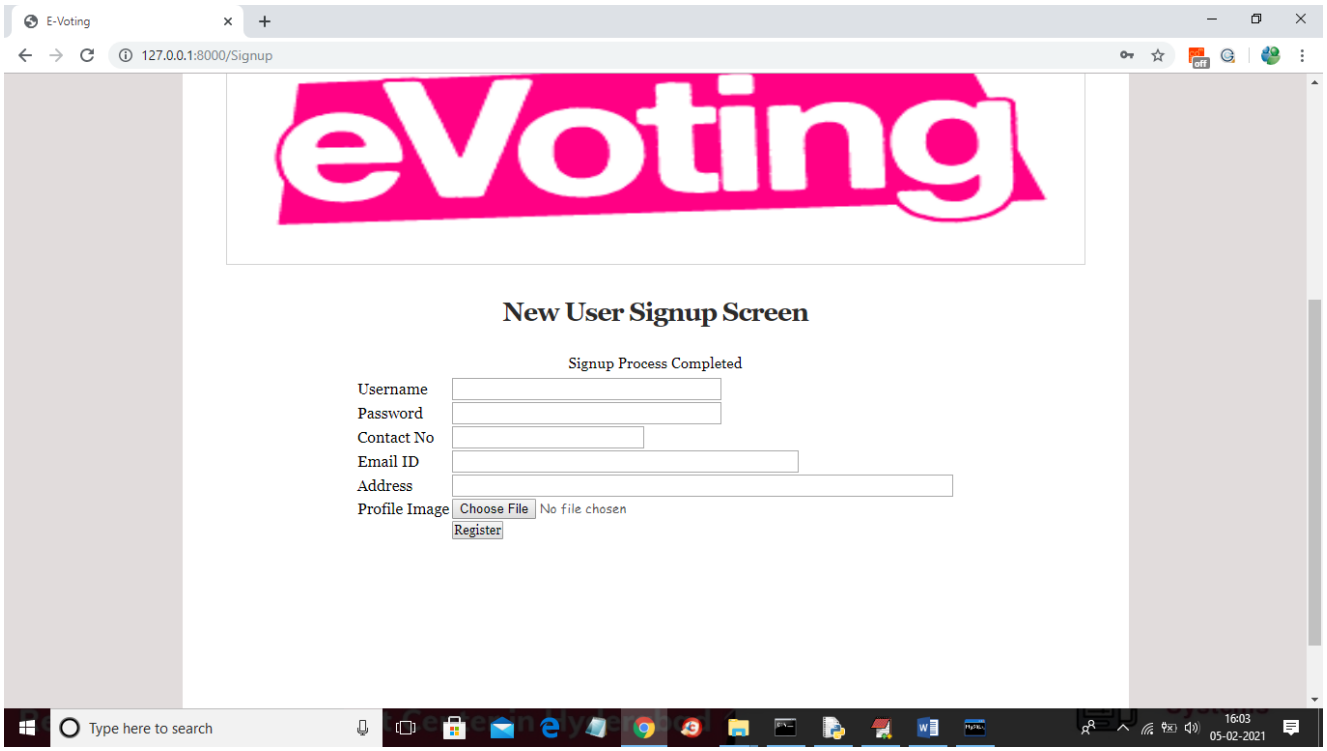


Figure8.1(g)In above screen signup process completed and now login as this user to cast vote

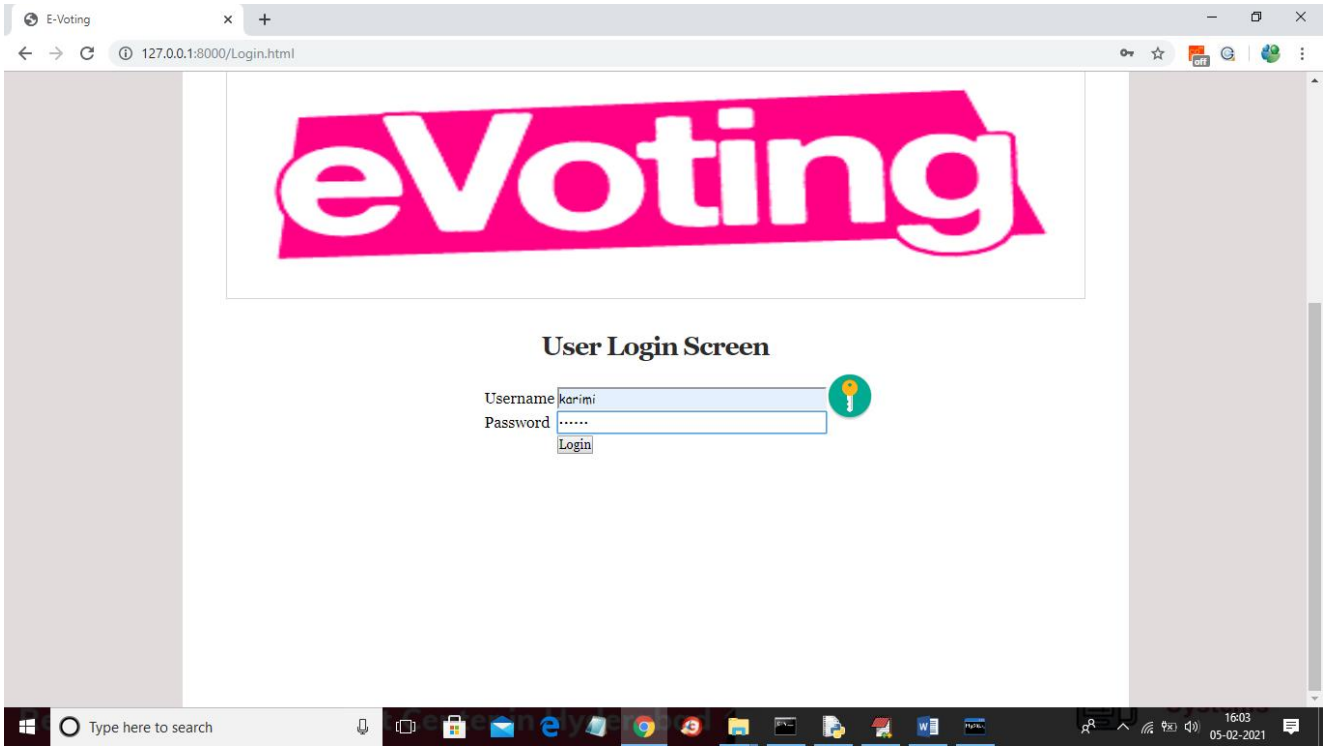


Figure8.1(h)In above screen application first authenticate user by using his login details and once after successful login then user will get below page

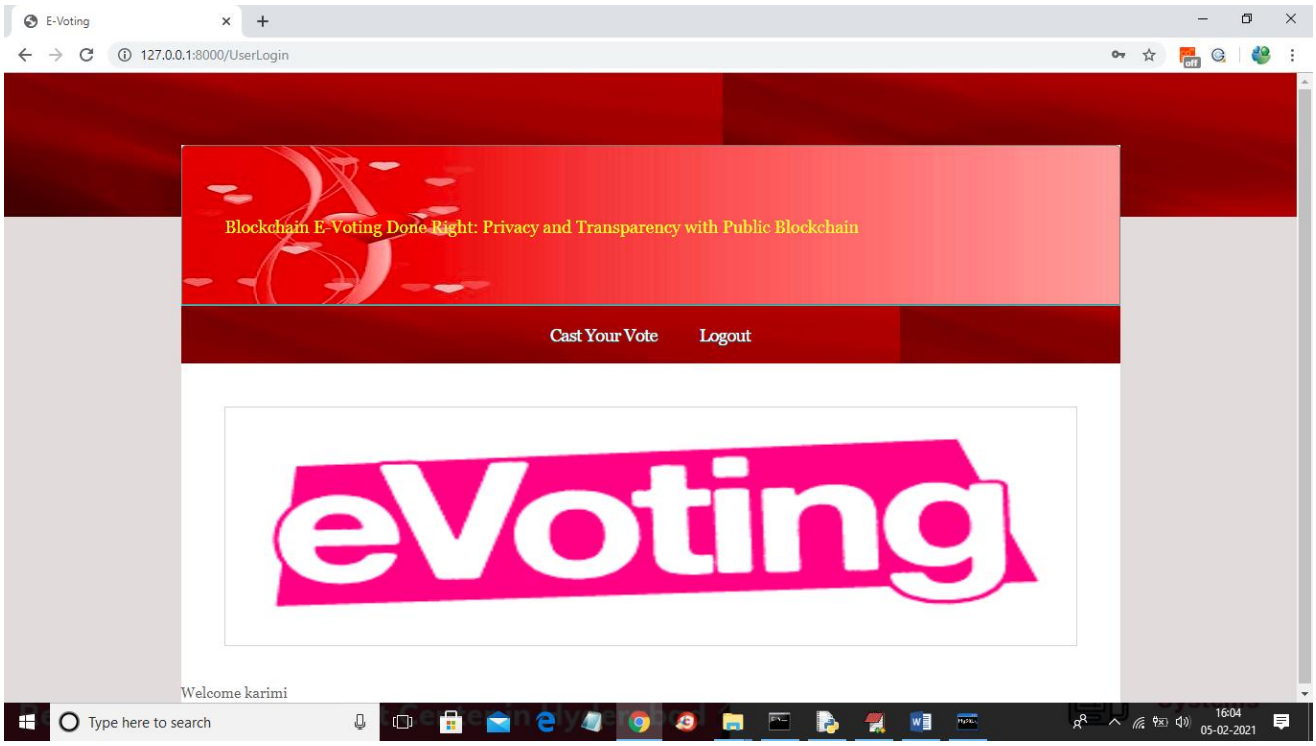


Figure8.1(i)In above screen user can click on 'Cast Your Vote' link to get below webcam screen

8.2 OUTPUT SCREENS

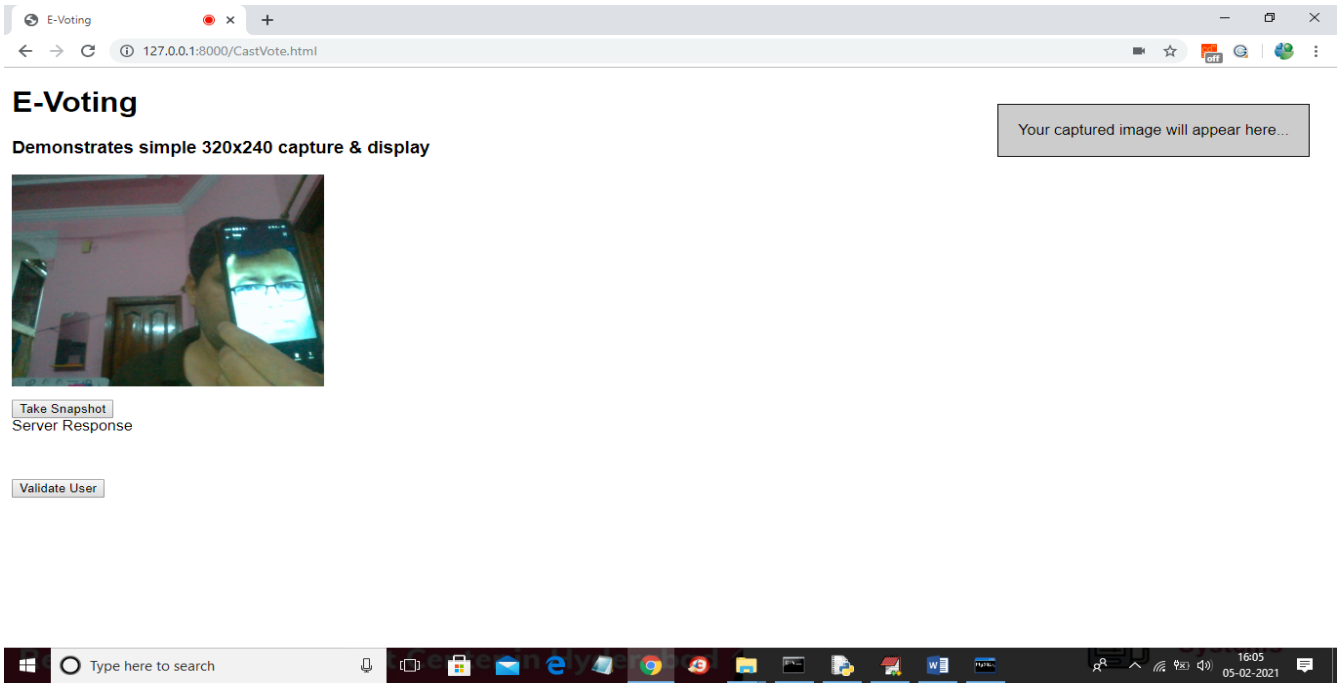


Figure8.2(a)In above screen webcam is running and then by showing person face, it needs to click on 'Take Snapshot' button to capture his face

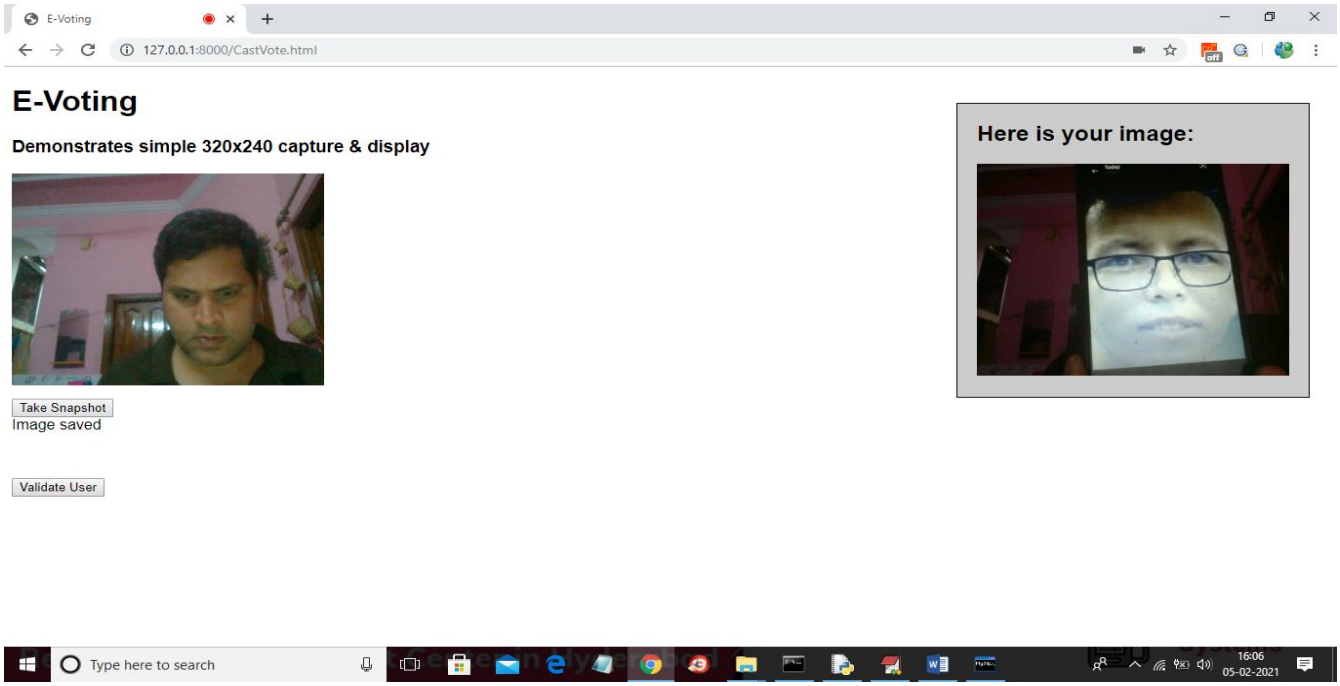


Figure8.2(b)In above screen person face is capture and now click on 'Validate User' button to validate user

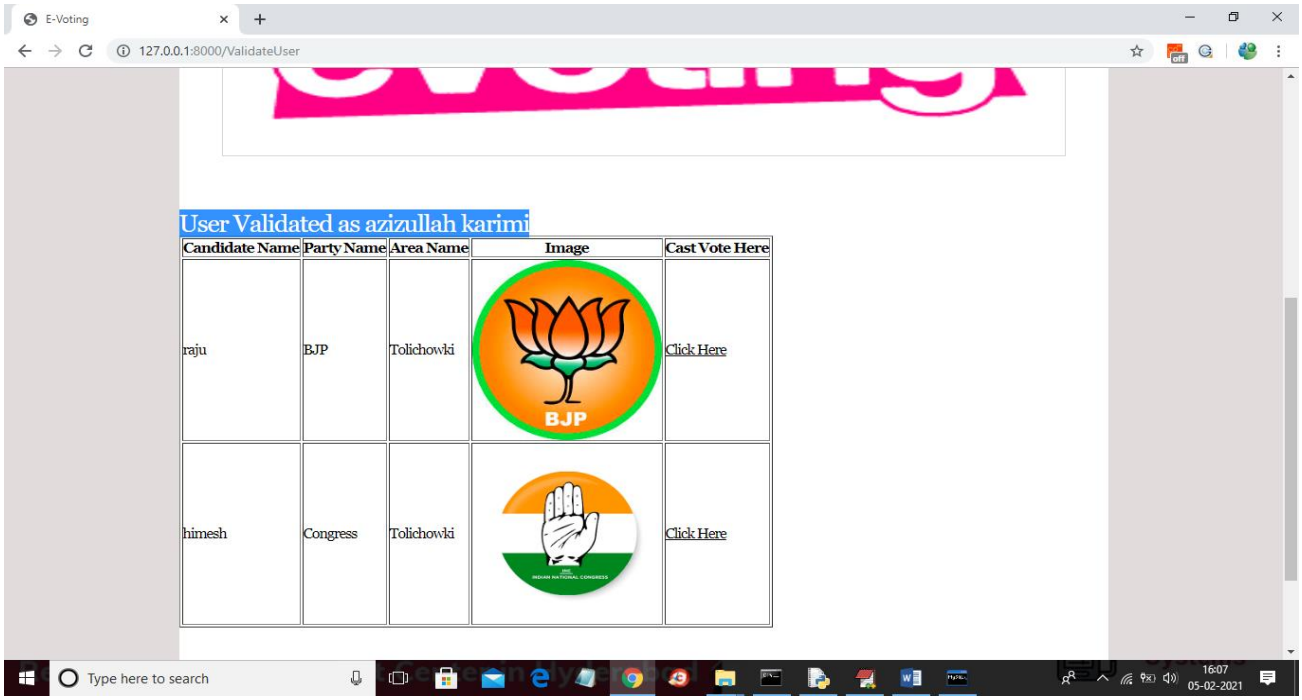


Figure8.2(c)In above screen in blue colour you can see user is identified as 'azizullah karimi' and then displaying list of candidates and now user can click on 'Click Here' option to cast his vote and to get below screen

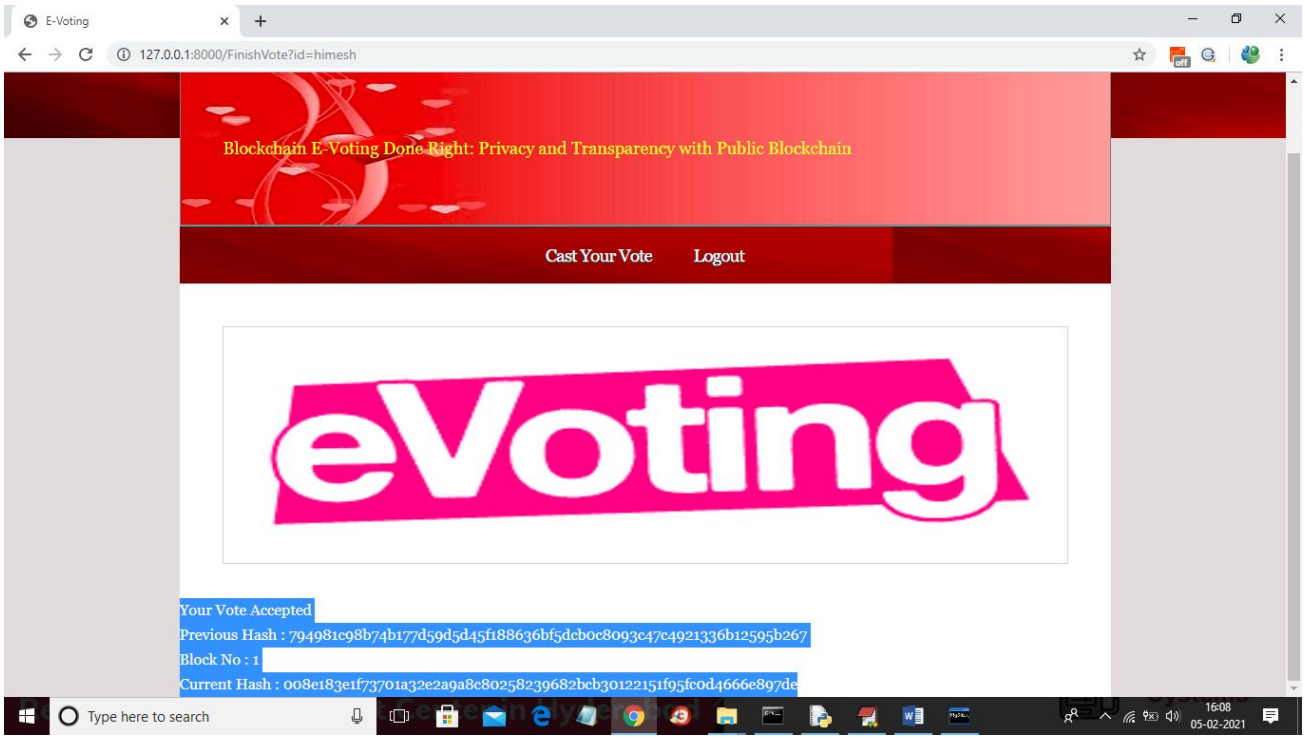


Figure8.2(d)In above screen as this is the first vote so block will be added to Blockchain with block No as 1 , you can see Blockchain created a chain of blocks with previous and current hash code validation. Now try again with same user to cast vote

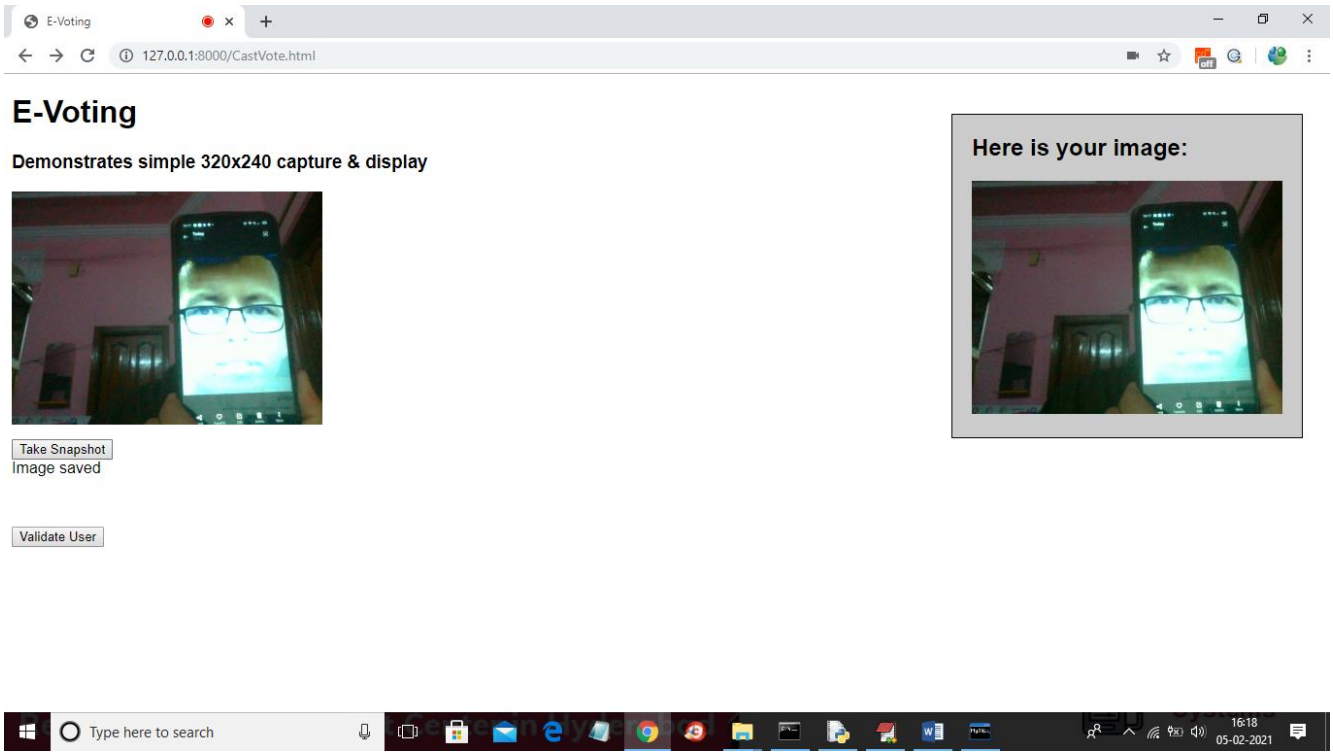


Figure8.2(e)In above screen same user trying again and below is the result

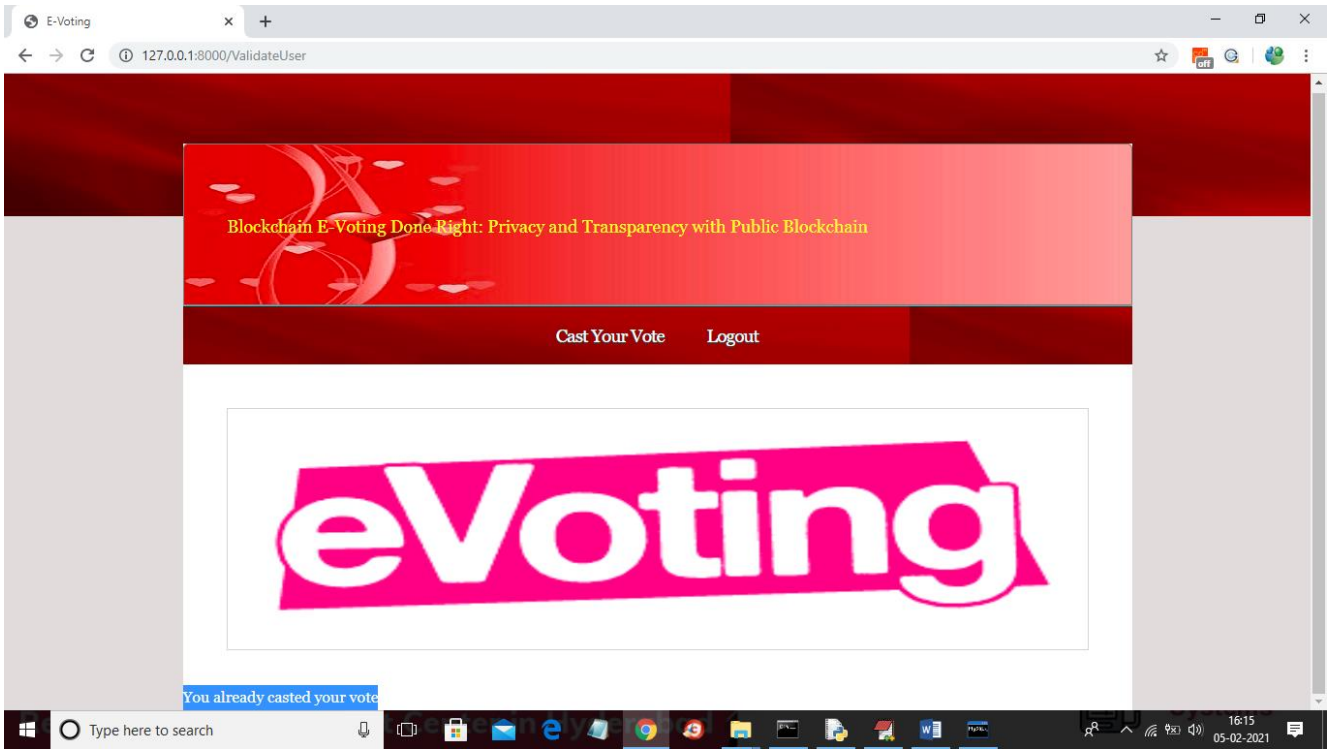


Figure8.2(f)In above screen if same user try again then will get message as 'You already casted you vote' and now logout and login as 'admin' to get vote count

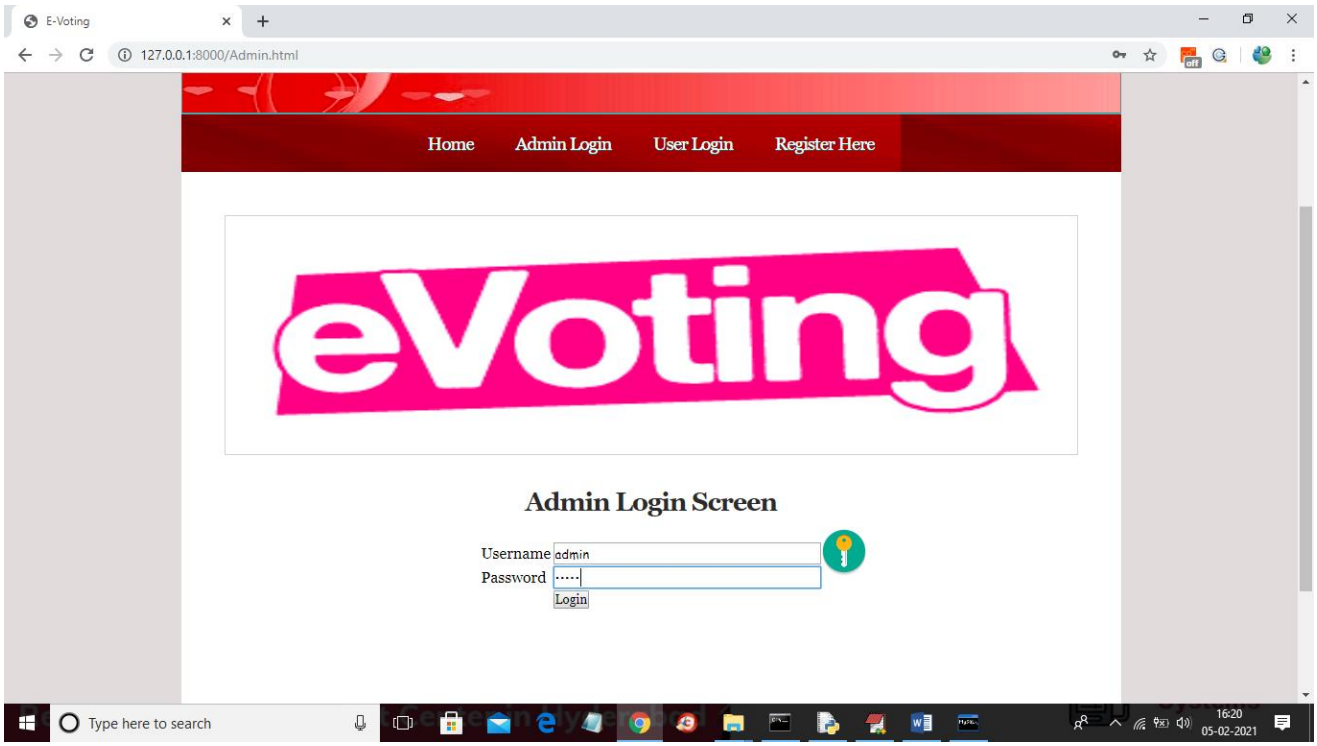


Figure8.2(g)In above screen login as admin and after login will get below screen

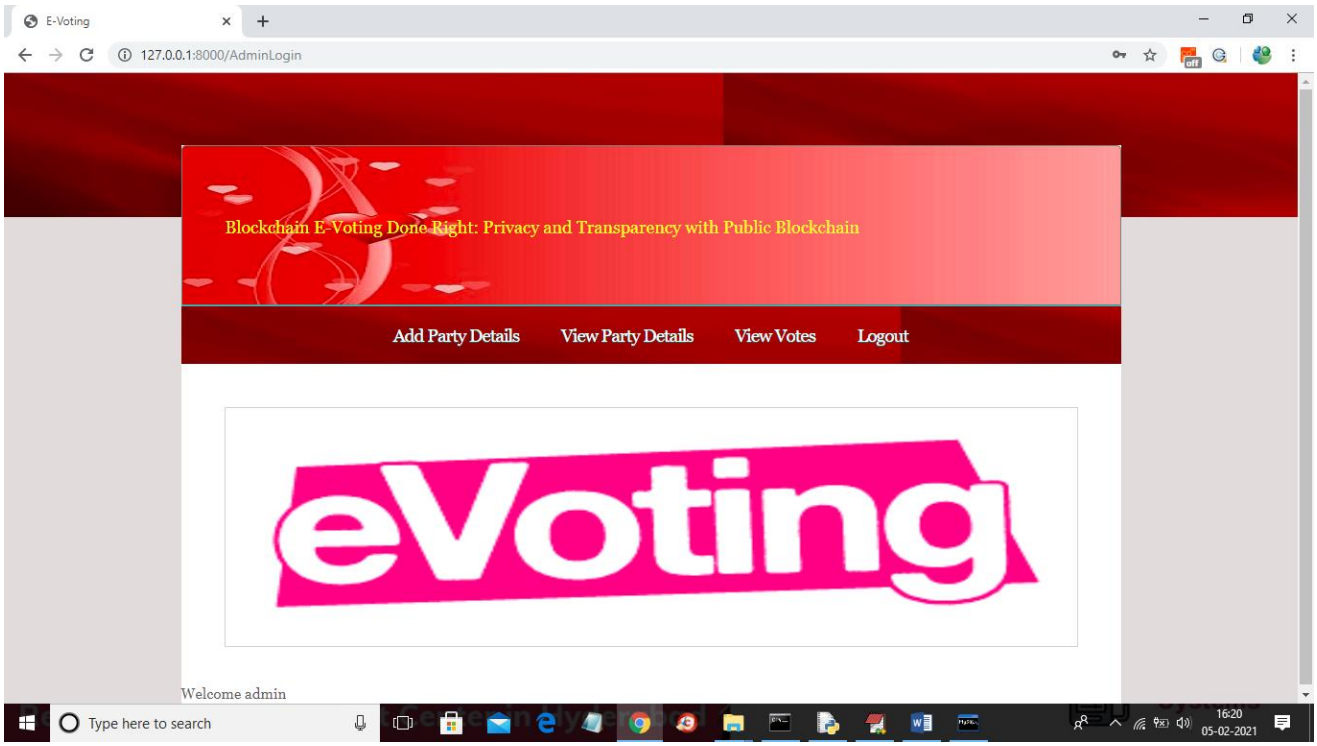


Figure8.2(h)In above screen admin can click on 'View Votes' link to get below screen

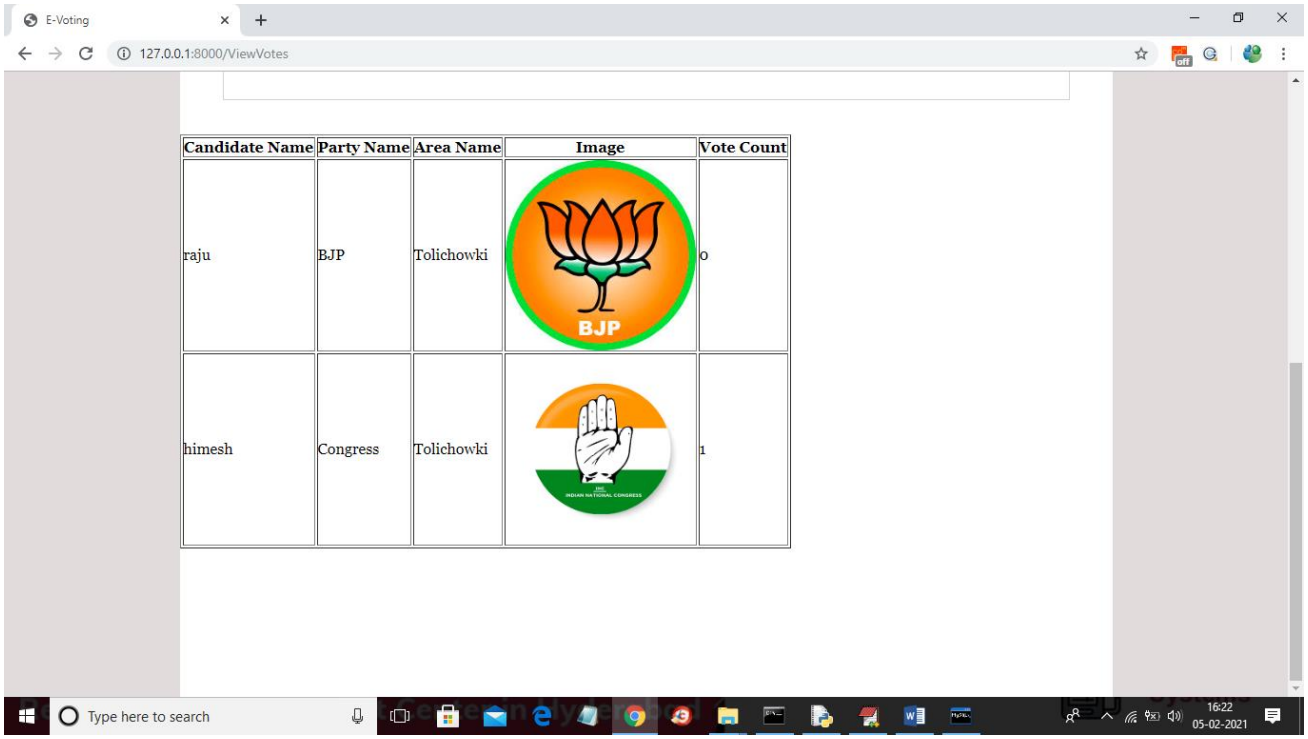


Figure8.2(i)In above screen admin can view all vote counts.

CONCLUSION AND FUTURE ENHANCEMENT

This project introduces a blockchain-based electronic voting system that utilizes smart contracts to enable secure and cost-efficient election while guaranteeing voters privacy. They have shown that the blockchain technology offers a new possibility to overcome the limitations and adoption barriers of electronic voting systems which ensures the election security and integrity and lays the ground for transparency. Using an Ethereum private blockchain, it is possible to send hundreds of transactions per second onto the blockchain, utilizing every aspect of the smart contract to ease the load on the blockchain. For countries of greater size, some additional measures would be needed to support greater throughput of transactions per second. The transparency of the block-chain enables more auditing and understanding of elections. These attributes are some of the requirements of a voting system. These characteristics come from decentralized networks, and can bring more democratic processes to elections, especially to direct election systems. For e-voting to become more open, transparent, and independently auditable, a potential solution would be to base it on blockchain technology. This project explores the potential of blockchain technology and its usefulness in the e-voting scheme. The blockchain will be publicly verifiable and distributed in a way that no one will be able to corrupt it.

REFERENCES

[1]Emre Yavuz ; Ali Kaan Koç ; Umut Can Çabuk ; Gökhan Dalkılıç (2018) Towards secure e-voting using ethereum blockchain. Electronic copy available at: https://ssrn.com/abstract=3648870

[2] KC Tam ,(2018) ,Transactions in Ethereum

[4] David Khoury,Elie F. Kfoury, Ali Kassem and Hamza Harb,(2018), Decentralized Voting Platform Based on Ethereum Blockchain

[5] Vaibhav Anasune, Pradeep Choudhari , Madhura Kelapure and Pranali Shirke Prasad Halgaonkar,"Online Voting: Voting System Using B-chain",(2019), Online Voting: Voting System Using Blockchain

[6] Linh Vo-Cao-Thuy, Khoi Cao-Minh, Chuong Dang-Le-Bao and Tuan A. Nguyen,(2019), Votereum: An Ethereum-based E-voting system :

[7] G Bhavan,i ,(2018) ,"Survey on Blockchain Based E-Voting Recording System Design"

[8]Friðrik Þ. Hjálmarsson , Gunnlaugur K . Hreiðarsson,(2018)"Blockchain-Based E-Voting System"

[9] Design:Rifa Hanifatunnisa and Budi Rahardjo,(2017) , Blockchain Based E-Voting Recording System

[10] N. Kshetri and J. Voas, "Blockchain-Enabled E-Voting," IEEE Software, vol. 35, pp. 95–99, jul 2018.

[11] M. Pawlak, J. Guziur, and A. Poniszewska-Maran´da, "Voting Process with Blockchain Technology: Auditable Blockchain Voting System," in Lecture Notes on Data Engineering and Communications Technologies, pp. 233–244, Springer, Cham, 2019.

[12] B. Singhal, G. Dhameja, and P. S. Panda, "How Blockchain Works," in

Beginning Blockchain, pp. 31–148, Berkeley, CA: Apress, 2018.

[13] Agora, "Agora Whitepaper," 2018.

[14] R. Perper, "Sierra Leone is the first country to use blockchain during an election - Business Insider," 2018.

[15] S. Nakamoto, “Bitcoin: A Peer-to-Peer Electronic Cash System,” tech. rep., 2008.

[16] G. Wood et al., “Ethereum: A secure decentralised generalised transac- tion ledger,” Ethereum project yellow paper, vol. 151, pp. 1–32, 2014.

[17] S. Landers, “Netvote: A Decentralized Voting Platform - Netvote Project

- Medium,” 2018.

[18] P. McCorry, S. F. Shahandashti, and F. Hao, “A Smart Contract for Boardroom Voting with Maximum Voter Privacy,” in Lecture Notes in Computer Science, ch. FCDS, pp. 357–375, Springer, Cham, 2017.

[19] Z. Brakerski and V. Vaikuntanathan, “Efficient Fully Homomorphic En- cryption from (Standard) LWE,” SIAM Journal on Computing, vol. 43, pp. 831–871, jan 2014.

[20] O. Goldreich and Y. Oren, “Definitions and properties of zero-knowledge proof systems,” Journal of Cryptology, vol. 7, no. 1, pp. 1–32, 2004

PUBLICATIONS

JOURNAL (UGC approved Journal)

CONFERENCE (2020 8th International Conference on Cyber and IT Service Management (CITSM))

DATE OF CONFERENCE: 23-24 OCT. 2020

INSPEC Accession Number: 20241450

TOPIC: Using Blockchain Data Security Management for E-Voting Systems

STUDENT PROFILE

Srujith Muppidi is a Bachelor of Technology Student at St Martin's Engineering College in Computer Science Engineering stream. He finished his schooling till 10th grade in Manair High School, Karimnagar and he completed his intermediate in Alphores Junior College. His technical skills include C, C++, Python, MySQL. He has completed one month internship program at Lasya IT solution pvt. Ltd, kompally. His participations include: a National Level Seminar on "Leadership talk" on 16th may 2020 Conducted by MHRD innovation cell and a National Level Three Day Workshop on "AI & ML in Speech & Audio Processing" From 10th to 12th of December 2020. He also placed in COGNIZANT and INFOSYS companies.

Karnataka Rohith Goud is a Bachelor of Technology Student at St Martin's Engineering College in Computer Science Engineering stream. He finished his schooling till 10th grade in Ekashila etechno school, Warangal .and he completed his intermediate in Sai chaitanya Junior College. His technical skills include C, C++, Java. He has completed one month internship program at Lasya IT solution pvt. Ltd, kompally. His participations include: a National Level Seminar on "Recent Trends in Cloud Computing, Fog, and Edge Computing" on 18th and 19th of June 2021 and a National Level Three Day Workshop on "AI & ML in Speech & Audio Processing" From 10th to 12th of December 2020.

Manchikatla Ranadeep is a Bachelor of Technology Student at St Martin's Engineering College in Computer Science Engineering stream. He finished his schooling till 10th grade in Narayana High School, Hyderabad and he completed his intermediate in Narayana Junior College. His technical skills include C, C++, Java. He has completed one month internship program at Lasya IT solution pvt. Ltd, kompally. His participations include: a National Level Seminar on "Recent Trends in Cloud Computing, Fog, and Edge Computing" on 18th and 19th of June 2021 and a National Level Three Day Workshop on "AI & ML in Speech & Audio Processing" From 10th to 12th of December 2020.

Marugai Rohan is a Bachelor of Technology Student at St Martin's Engineering College in Computer Science Engineering stream. He finished his schooling till 10th grade in St marks high School, hyderabad and he completed his intermediate in bhavans Sri aurobindo junior. His technical skills include C, C++, Java, Python, MySQL. He has completed one month internship program at Lasya IT solution pvt. Ltd, kompally. His participations include: a National Level Seminar on "Recent Trends in Cloud Computing, Fog, and Edge Computing" on 18th and 19th of June 2021 and a National Level Three Day Workshop on "AI & ML in Speech & Audio Processing" From 10th to 12th of December 2020. He also took part in the Employability Skill Development Program conducted by Zensar.

APPENDICES

```
from hashlib import sha256
import json
import time
import pickle
from datetime import datetime
import random
import pyaes, pbkdf2, binascii, os, secrets
import base64


class Block:
    def __init__(self, index, transactions, timestamp, previous_hash):
        self.index = index
        self.transactions = transactions
        self.timestamp = timestamp
        self.previous_hash = previous_hash
        self.nonce = 0


    def compute_hash(self):
        block_string = json.dumps(self.__dict__, sort_keys=True)
        return sha256(block_string.encode()).hexdigest()


class Blockchain:
```

```
# difficulty of our PoW algorithm
difficulty = 2 #using difficulty 2 computation

def __init__(self):
    self.unconfirmed_transactions = []
    self.chain = []
    self.create_genesis_block()
    self.peer = []
    self.translist = []

def create_genesis_block(self): #create genesis block
    genesis_block = Block(0, [], time.time(), "0")
    genesis_block.hash = genesis_block.compute_hash()
    self.chain.append(genesis_block)

@property
def last_block(self):
    return self.chain[-1]

def add_block(self, block, proof): #adding data to block by computing new and previous hashes
    previous_hash = self.last_block.hash

    if previous_hash != block.previous_hash:
        return False
```

```
    if not self.is_valid_proof(block, proof):
        return False

    block.hash = proof
    #print("main "+str(block.hash))
    self.chain.append(block)
    return True

def is_valid_proof(self, block, block_hash): #proof of work
    return (block_hash.startswith('0' * Blockchain.difficulty) and block_hash == block.compute_hash())

def proof_of_work(self, block): #proof of work
    block.nonce = 0

    computed_hash = block.compute_hash()
    while not computed_hash.startswith('0' * Blockchain.difficulty):
        block.nonce += 1
        computed_hash = block.compute_hash()

    return computed_hash

def add_new_transaction(self, transaction):
    self.unconfirmed_transactions.append(transaction)
```

```
def addPeer(self, peer_details):
    self.peer.append(peer_details)

def addTransaction(self,trans_details): #add transaction
    self.translist.append(trans_details)

def mine(self):#mine transaction
    if not self.unconfirmed_transactions:
        return False

    last_block = self.last_block

    new_block = Block(index=last_block.index + 1,
                      transactions=self.unconfirmed_transactions,
                      timestamp=time.time(),
                      previous_hash=last_block.hash)

    proof = self.proof_of_work(new_block)
    self.add_block(new_block, proof)

    self.unconfirmed_transactions = []
    return new_block.index
```

```
def save_object(self,obj, filename):
    with open(filename, 'wb') as output:
        pickle.dump(obj, output, pickle.HIGHEST_PROTOCOL)
```

A

PROJECT REPORT

On

ACCIDENT DETECTION USING IoT

Submitted by

1)Mr.K.Gaurav(17K81A05E1) **2)Mr.VedhaKalyan(17K81A05G4)**

3)Mr.K.Bhaskar(17K81A05E8) **4)Mr.K.Sumanth (17K81A05F1)**

in partial fulfillment for the award of the

degree of

BACHELOR OF TECHNOLOGY

IN

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Under the Guidance of

Mr.G.Mallikarjun

Assistant Professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ST.MARTIN'S ENGINEERING COLLEGE
An Autonomous Institute

Dhulapally, Secunderabad – 500 100

JUNE 2021

TABLE OF CONTENTS

BONAFIDE CERTIFICATE

This is to certify that the project entitled ACCIDENT DETECTION USING IOT is being submittedby**1.Mr.GauravKarwa(17K81A05E1),**2.**Mr.NandyalaVedhaKalyan(17K81A05G4)** 3.**Mr.Kanugu Bhaskar(17K81A05E8),**4.**Mr.Sai Sumanth Chowdary(17K81A05F1)** in partial fulfillment of the requirement for the award of the degree of **BACHELOR OF TECHNOLOGY IN** COMPUTER SCIENCE AND ENGINEERING is recorded of bonafide work carried out by them. The result embodied in this report have been verified and found satisfactory.

Assistant Professor
Mr.G.Mallikarjun
Department of CSE

Head of the Department
Dr.M.Narayanan
Department of CSE

Internal Examiner

External Examiner

Place:

Date:

DECLARATION

We, the student of **Bachelor of Technology** in Department of Computer Science and Engineering', session: 2017 – 2021, St. Martin's Engineering College, Dhulapally, Kompally, Secunderabad, hereby declare that work presented in this Project Work entitled ACCIDENT DETECTION USING IOT is the outcome of our own bonafide work and is correct to the best of our knowledge and this work has been undertaken taking care of Engineering Ethics. This result embodied in this project report has not been submitted in any university for award of any degree.

Mr. Gaurav Karwa(17K81A05E1)

Mr. Vedha Kalyan(17K81A05G4)

Mr. Kanugu Bhaskar(17K81A05E8)

Mr.Sumant Chowdary(17K81A05F1)

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible and whose encouragements and guidance have crowded effects with success.

We extended our deep sense of gratitude to Principal**,** **Dr. P. SANTOSH KUMAR PATRA**, St. Martin's Engineering College, Dhulapally, for permitting us to undertake this project.

We are also thankful to **Dr. M.NARAYANAN**, Head of the Department, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

We would like to thank **Dr. T. POONGOTHAI,** Department of Computer Science and Engineering (AI & ML) for her encouragement and insightful comments and as well as our project coordinators **Dr. B.RAJALINGAM**, Associate Professor and **Dr. N. SATHEESH**, Professor, in Department of Computer Science and Engineering for their valuable support.

We would like to express our sincere gratitude and indebtedness to our project supervisor G.Mallikarjun, Assistant Professor, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

Finally, we express thanks to all those who have helped us successfully to completing this project. Furthermore, we would like to thank our family and friends for their moral support and encouragement.

We express thanks to all those who have helped us in successfully completing the project.

Mr.Gaurav Karwa (17K81A05E1)

Mr.Vedha Kalyan (17K81A05G4)

Mr.Kanugu Bhaskar (17K81A05E8)

Mr.Sumanth Chowdary (17K81A05F1)

ABSTRACT

Most individuals involved in traffic accidents receive assistance from drivers, passengers, or other people who are travelling nearby. However, when an accident occurs in a thinly populated area or when the driver is the only person present in the vehicle and the crash results in loss of consciousness of the person, no one will be able to send a message to proper authorities within the golden period for medical treatment. This paper presents the application of wireless communication also called as vehicle to hospital communication using the LoRa technology. The main aim of the v2h communication is to prevent accidents by allowing vehicles in transit to send position data to nearby location. This proposed method is designed with the help of GPS and Lora. This system works based on IOT. Lora provides better performance while compared to bluetooth.

LIST OF FIGURES

LIST OF OUTPUT SCREENS

LIST OF ACRONYMS

MEMS	Micro-Electro-Mechanical System
LoRa	Low power High range
Lorawan	Low power High range wide area protocol
IoT	Internet of Things
GPS	Geo Positioning System

1.0 INTRODUCTION

1.1 PROJECT OVERVIEW

The entire project is mainly on the Lora based project,which is going to send the accident took place in a certain location to a particular vehicle using a special kit built .This project is to save a lot of lives that are lost in the scarely populated areas by sending their location to the near by hospitals.

1.2 OBJECTIVES OF THE STUDY

To pass the message from vehicle to hospital using LoRa wireless sensor networks. To send the exact location of the accident place as soon as the accident occurs. Mems sensor sends values to the Arduino board on the transmitter side. As soon as the threshold value is breached, we then find the GPS coordinates and send them to the Arduino, which in turn sends them to LORA module which transmits the data to the receiver.

On the receiver end we use the LORA to receive the GPS coordinates and send them to node mcu which is connected to computer which displays the output.

1.3 ORGANIZATION OF CHAPTERS

Besides the introduction, the thesis is organized in other six chapters as follows:

Chapter 2, LITERATURE SURVEY: the review is made in the context of the LORA where the entire flaws in the present system and the advantages of the project has been explained.

Chapter 3, SOFTWARE AND HARDWARE REQUIREMENTS: this chapter discuss

about the software and hardware required for the execution of the project.

Chapter 4, SOFTWARE DEVELOPMENT ANALYASIS: this chapter explains the

assumptions and technical specifications of the project.

Chapter 5, PROJECT SYSTEM DESIGN: this chapter explains all the software development process with dfd, E-R diagrams, and UML diagrams clearly.

Chapter 6, PROJECT CODING: this chapter explains the design of the system, roles and responsibilities, as well as the requirements of a EHRs management solution based on block chain.

Chapter 7, PROJECT TESTING: this chapter explains various test cases to test the project working.

Chapter 8, OUTPUT SCREENS: explains a step by step process of the project

execution.

Chapter 9, EXPERIMENTAL RESULTS: tests and results are shown and explained in this chapter. The results are analyzed in the context of the thesis project and followed by discussion on systems throughput and resiliency, as well as the approaches to testing and analysis.

Chapter 10, CONCLUSION AND FUTURE ENHANCEMENT: the chapter ends the project with a short summary of the main concepts mentioned in the thesis as well as the relevant results.

2. LITERATURE SURVEY

2.1 SURVEY ON BACKGROUND

A large number of road accidents take place all over the world. There are a lot of reasons for them to happen and some of them are due to collisions between the vehicles, vehicle to the divider, brake failures, fuel ignition, sudden health issues of the driver, etc. More than 1.2 million people lost their lives in the road accidents in 2019, according to world health organization (WHO). Worldwide, the road traffic accidents (RTAs) kill 3,000 people every day and injure over 3 million every year. It is also difficult to pair the data transmission from one vehicle to another vehicle. The advantage in wireless communication electronics has accelerated to develop many wireless network solutions to replace existing wired network.

In current method the system is designed with the help of Bluetooth module. A Bluetooth device can maximum have a range of 10metres. The Bluetooth devices can also be easily hacked hence it does not provide the security. Bluetooth, the technology comes with a few disadvantages, including slow data speeds, poor data security and shortened battery life. It was also difficult to pair the data transmission from one vehicle to another vehicle as the connectivity rage of the Bluetooth devices is very much less. There was no direct connectivity from any vehicle to any other devices that are far away from that particular vehicle.

2.2 CONCLUSIONS ON SURVEY

The proposed method is designed with the help of MEMS sensor, GPS and LoRa. The LoRa (Low Range) has a very large range of advantages. Few of them are listed below:

- LoRa converts symbols (binary data) to chirp signals that span the frequency range.
- Lora works on frequency communication model.
- This module will have long battery life.
- There will be no necessity of internet.
- Lora module can communicate to a range of 20km minimum.
- It is being used to communicate with many IoT devoices and nodes.
- Power consumption is very low when being compared to any other IoT protocols. Therefore, it is a best choice for the battery critical applications which indeed is a major requirement or all embedded system projects.
- The frequency band that is being used for LoRa is much easily available and also in use in most of the countries in the world.

These are the few advantages of this project.

With no wastage of time for calling the hospitals and the rescue team, the information is beingdirectly sent to the most nearest hospital in case of severe road accidents and there by trying to save people's lives.

3. SOFTWARE AND HARDWARE REQUIREMENTS

3.1 SOFTWARE REQUIREMENTS

The software specifications that are to be used are enlisted below:

- Arduino IDE
- Embedded C

3.2 HARDWARE REQUIREMENTS

The hardware equipments that are to be used are enlisted as follows:

- MEMS Sensor
- GPS Module
- Arduino Kit
- LORA Module
- Node MCU
- Windows System

4. SOFTWARE DEVELOPMENT ANALYASIS

4.1 OVERVIEW OF PROBLEM

The entire project is mainly on the Lora based project,which is going to send the accident took place in a certain location to a particular vehicle using a special kit built .This project is to save a lot of lives that are lost in the scarely populated areas by sending their location to the near by hospitals.

The main objective underlying during this project is to reduce the number of casualties at the time of any vehicle accidents. This will be achieved by using proper communication methodologies for the security. Usage of this type of advanced technologies in our daily using automobiles is additionally called because the “SMART CARS SYSTEM”. The new productive equipment that's getting used during this project is LORA which is answerable for the transmission and therefore the reception of the info or any quite information. [4] LoRa module is capable of supporting a good range of sensors and a protracted range wireless protocol. The transmission or sending the signal is finished from the vehicle that's being met with the accident and therefore the reception of that data is finished at the hospital that's being located nearest to the accident occurred site. As soon as any accident is being detected by the vehicle then the signal is being sent on to the closest hospital and with no wastage of your time the ambulance will be sent to the accident spot in order that no lives are going to be lost thanks to the delay. The project idea is proposed to scale back the implications of accidents in our daily lives.

4.2 DEFINE THE PROBLEM

Most individuals involved in traffic accidents receive assistance from drivers, passengers, or other people who are travelling nearby. However, when an accident occurs in a thinly populated area or when the driver is the only person present in the vehicle and the crash results in loss of consciousness of the person, no one will be able to send a message to proper authorities within the golden period for medical treatment. This paper presents the application of wireless communication also called as vehicle to hospital communication using the LoRa technology. The main aim of the v2h communication is to prevent accidents by allowing vehicles in transit to send position data to nearby location. This proposed method is designed with the help of GPS and Lora. This system works based on IOT. Lora provides better performance while compared to bluetooth.

A large number of road accidents take place all over the world. There are a lot of reasons for them to happen and some of them are due to collisions between the vehicles, vehicle to the divider, brake failures, fuel ignition, sudden health issues of the driver, etc. More than 1.2 million people lost their lives in the road accidents in 2019, according to world health organization (WHO). [5] Worldwide, the road traffic accidents (RTAs) kill 3,000 people every day and injure over 3 million every year. It is also difficult to pair the data transmission from one vehicle to another vehicle. The advantage in wireless communication electronics has accelerated to develop many wireless network solutions to replace existing wired network.

Existing System with Disadvantage

In current method the system is designed with the help of Bluetooth module. A Bluetooth device can maximum have a range of 10metres. The Bluetooth devices can also be easily hacked hence it does not provide the security. [6] Bluetooth, the technology comes with a few disadvantages, including slow data speeds, poor data security and shortened battery life. It was also difficult to pair the data transmission from one vehicle to another vehicle as the connectivity rage of the Bluetooth devices is very much less. There

was no direct connectivity from any vehicle to any other devices that are far away from that particular vehicle.

Proposed System with advantages

The proposed method is designed with the help of MEMS sensor, GPS and LoRa. The LoRa (Low Range) has a very large range of advantages. Few of them are listed below:

- [7] LoRa is the technology that modulates the data into electromagnetic waves.
- Lora works on frequency communication model.
- This module will have long battery life.
- There will be no necessity of internet.
- Lora module can communicate to a range of 20km minimum.
- It can be used to communicate with thousands of IoT nodes and devices.
- Power consumption is very low when being compared to any other IoT protocols. Therefore, it is a best choice for the battery critical applications which indeed is a major requirement or all embedded system projects.
- The frequency band that is being used for LoRa is much easily available and also in use in most of the countries in the world.

4.3 MODULES OVERVIEW

MEMS SENSOR:

MEMS inclinometers and accelerometers are low-cost, high precision inertial sensors that serve a wide variety of industrial applications. MEMS Sensor: It is a chip-based technology, known as a Micro Electro-Mechanical System sensor.that is composed of a suspended mass between a pair of capacitive plates.When tilt is applied to the sensor, the suspended mass creates a difference in electric potential which is measured as a change in capacitance.That signal is then amplified to produce a stable output signal in digital.

FIG 4.1 - ANGULAR DESCRIPTION OF MEMS SENSOR

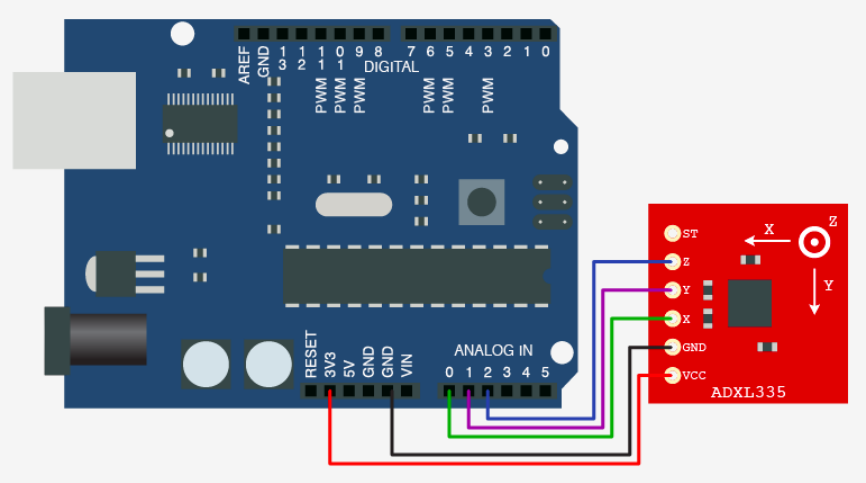


FIG 4.2 - MEMS SENSOR TO ARDUINO

GPS Sensor**:**

GPS stands for Global Positioning System.

The system contains satellites and ground based control installations. GPS sensor consists of surface mount chip which processes signals from GPS satellites using a small rectangular often mounted on the top of the GPS chip. GPS receiver can distinguish signals from atleast four satellites by comparing their received pseudo random bit sequences. These GPS module contain latitude, longitude, altitude and data recording time. These sentences are decoded by connecting microcontroller with GPS module and writing small program.

Board with GPS sensor

FIG 4.3 - GPS SENSOR

ARDUINO**:**

Arduino is an open-source platform used for building electronics projects. Arduino consists of both a physical programmable circuit board (often referred to as a microcontroller) and a piece of software, or IDE (Integrated Development Environment). The software runs on your computer, used to write and upload computer code to the physical board.

LORA:

LoRa devices and wireless radio frequency technology is a long range, low power wireless platform that has become the most widely used technology for Internet of Things (IoT) networks worldwide. Lora stands for long range and its range starts with 10km. They serve as the foundation for the specification are bi-directional communication, mobility and localization services. It maintains low-power characteristics and significantly increases communication range. It can be used as both transmitter and the receiver.

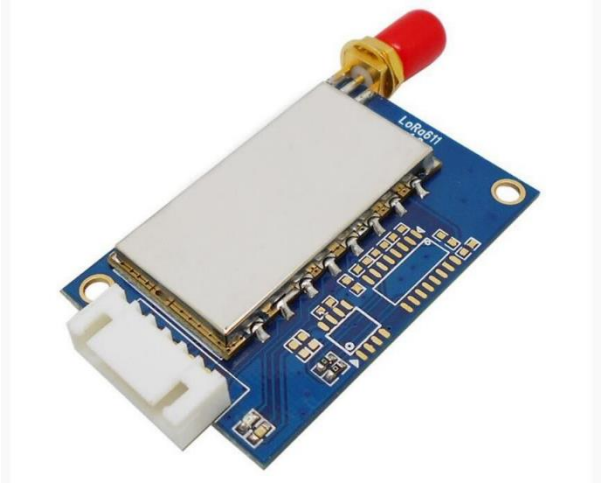


FIG 4.4 - LORA

NODE MCU:

ESP8266 is a WiFi Module chip that can be configured to connect to the Internet for Internet of Things(IoT) and similar Technology Projects. Basically, Your normal Electrical and Mechanical equipments cannot connect to the Internet on their own. They don’t have the in-built setup to do so.Now, Coming to your question about NodeMCU. NodeMCU is a Firmware on ESP8266. Its basically an SoC (System on

Chip) A System on a Chip or System on Chip(SoC) is an integrated circuit that integrates all components of a computer or other electronic systems.

FIG 4.5 - NODE MCU

EMBEDDED SYSTEMS

The embedded systems play a major role in this project management. Embedded System is also known as an integrated system due to its combination of hardware and software. All the individual components are connected together to perform dedicated functions one after the other without any disruption. Some embedded systems are being massively produced, benefiting from economies of scale.

The components such as sensors which are individually available are being grouped together as a single unit there by forming as an embedded system.

HARSH ENVIRONMENT

Most of the IoT related projects do not operate in the controlled environment. Excessive heat is often being dissipated in this kind of applications involving combustion of liquids or gases. In this project the combustion of fuel can be the major problem. As the entire body of the vehicle gets heated up, the equipment has to be placed in the depth of the car where there will be no risk of heating up or malfunctioning of the device. Additional problems can be caused for the devices that are being placed inside the vehicle, by a need for protection from the vehicle's vibrations, slight shocks, lightning, power flow or power supply fluctuations, water, fire, corrosion, and general physical abuse.

SYSTEM SAFETY AND RELIABLITLY

As the system complexity and the computing power continue to grow rapidly, they are starting to control more and more regarding the safety of the particular system. These safety measures can be in the form of software as well as hardware control. Mechanical safety backups are normally activated whenever the computer systems do not have control in order to safely shut down system operation. Software safety and reliability is a much bigger issue. Software doesn't normally "break" as that of the hardware components. However software can be so complex that a set of unexpected circumstances that can cause software failures leading to unsafe situations. Discussion of this topic is much more difficult, but the challenges the system designers include designing reliable software and building them cheap, available systems using the unreliable components. The main challenge for the system designers is to obtain low-cost reliability with minimal redundancy.

CONTROL OF PHYSICAL SYSTEMS

One of the main reasons for an embedded computer or a system is to interact with the environment. This is often being done by monitoring and controlling external machinery. The arduino UNO that is being placed inside of any of the automobile transforms the analog signals from sensors into digital form for processing. The outputs must be transformed back to analog signal levels. When controlling the physical equipment, current loads have to be switched in order to operate actuators and the sensors. To meet these kinds of needs, embedded systems may need large computer circuit boards with many non-digital components. The designers of the equipment must carefully balance the system’s analog components, network, power, mechanical, and digital hardware with the corresponding software.

SMALL AND LOW WEIGHT

The IoT depended systems mostly consists of rather bigger or larger systems as they include a lot of external devices which in turn makes the entire system to put on the

weight. But the devices that are being used in this project such as MEMS sensor, GPS and the LORA modules tend to weigh very less so that they can be easily be installed in a small area One of the challenges for the embedded systems designers is to develop non-rectangular geometries for certain solutions. Weight can also be one of the most critical constraints. Embedded automobile control systems, must be light weight for fuel economy. These devices are rather portable as they need a comparatively smaller space and can be installed very easily.

COST SENSITIVITY

Cost is a major issue in most of the IoT related systems, because it consists of a lot of different types of the equipments which perform different types of functionalities. But the sensitivity to cost changes can vary dramatically different in these types of systems. As most of the devices are being placed inside of an automobile, the sensors may bare a less amount but the LORA module which is capable of transmitting the information to rather farther places might be an implication. Apparently this system is rather cost bearable and is very much useful at the time of accidents.

POWER MANAGEMENT

Most of the IoT systems have strict constraints on power. The power input that is being supplied to the arduino have to be stabilized in order to reduce the fluctuations and hence to produce the desired output. Minimization of the heat production is another bigger aspect in all the IoT included systems.

INCOMPATIBLE CONNECTIONS

As the entire equipment is being placed inside a moving automobile or a vehicle, there will be a constant movement of the entire vehicle due to the massive engine inside it. Sometimes due to the jerks, misconnections can take place which in turn leads to the failure of device. Hence the device that is being installed had to be regularly monitored. Another condition where the device wouldn’t work is when the

arduino failure takes place. Dislocation of the chip or supply of improper voltage or overloading can be the reasons for the arduino failure.

HETEROGENEOUS ARCHITECHTURE

Most of the systems that are based on the IoT have a rather complex structure when compared to any other systems. It uses different kinds of processors for the the final output to be obtained.

The combination of input and output interfaces, local and external memories, sensors and actuators makes the embedded system design a truly unique one. All the individual components are being connected together doing different types of functionalities one after the other in their specific memory locations, forming components output as the other components input. This heterogenous architecture is being described very effortlessly in this system.

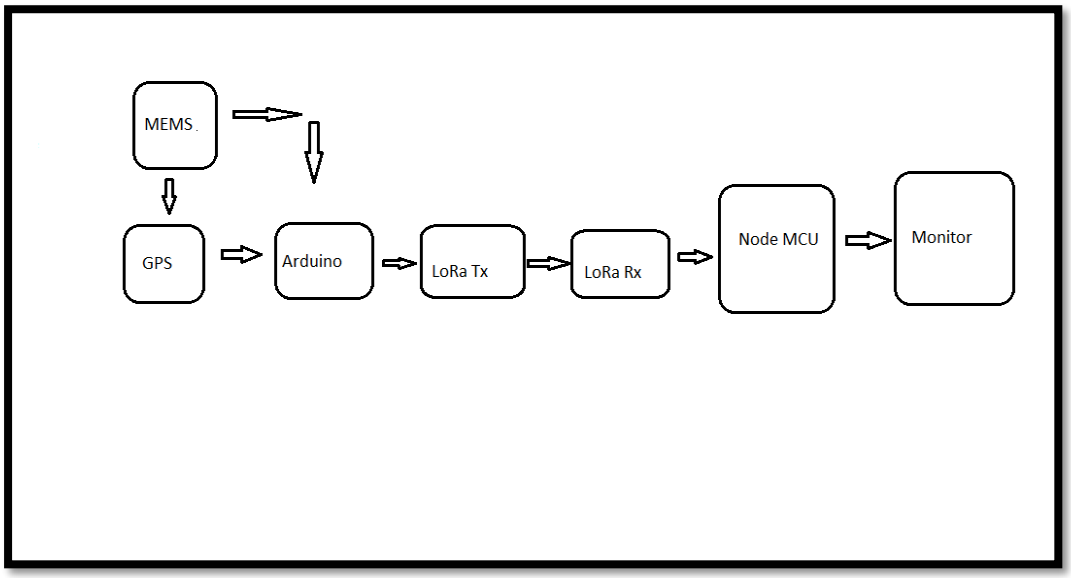


FIG 4.6 – WORK FLOW OF THE PROJECT

SCALABILITY OF BANDWIDTH

LoRa usually uses different types of license-free radio frequency bands like 433 MHz, 868 MHz in Europe, 915 MHz in Australia, North America and 923 MHz

in Asia. LoRa can send and also receive the information or the signals by two channels. LoRa enables a long range transmission as well as the receptions (more than 10 km of range in rural areas) with rather low power consumption. LoRa modulation is both bandwidth and also more frequency scalable. This technology can be used for both narrow band frequency hopping and also the wide band sequence applications. As they can transmit and share the information to far away distances, they are being used in military based applications.

HIGH ROBUTSNESS AND FADING RESISTANT

LoRa always has an asynchronous nature. Due to this nature the signals are very resistant to in-band and also out-band interference mechanisms. Its receiver can obtain out-of-channel selectively figures that are ranging from 90dB .The frequency or the pulse that is being produced by the LoRa is relatively broadband and hence it has the immunity to the multipath and also the fading, making it very much ideal for the urban and suburban environments.

HIGH NETWORK CAPACITY AND ITS RANGING

This technology mainly uses the orthogonal spreading of the factors rather than horizontal or the vertical manner. Hence it enables multiple spread signals to be transmitted at the same time and also on the same channel. Modulated signals at specific and different spreading factors appear as noise to the target receiver. LoRa can linearly discriminate between the frequency and the time errors. These devices have geo-location capabilities that are used for triangulation positions of devices via timestamps from gateways. LoRa is ideal for radar applications and also for the real-time location services.

DOPPLER RESISTANT

Doppler shift causes a small or a minute frequency shift in the LoRa pulse which introduces a relatively negligible shift in the time axis of the baseband signal. This might create a very small or very minute disturbance in the reception side of the device

4.4 DEFINE THE MODULES

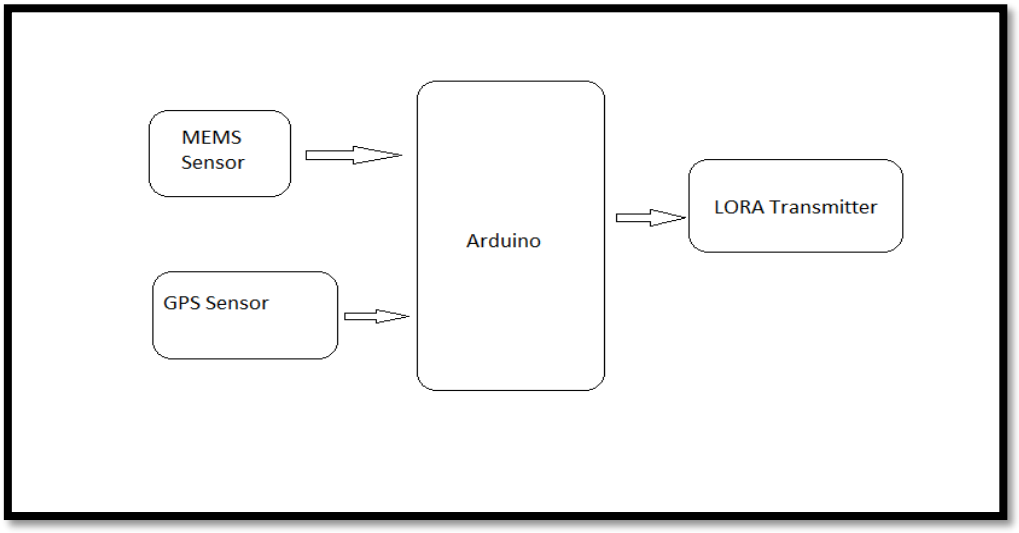


FIG 4.7 – TRANSMITTER BLOCK

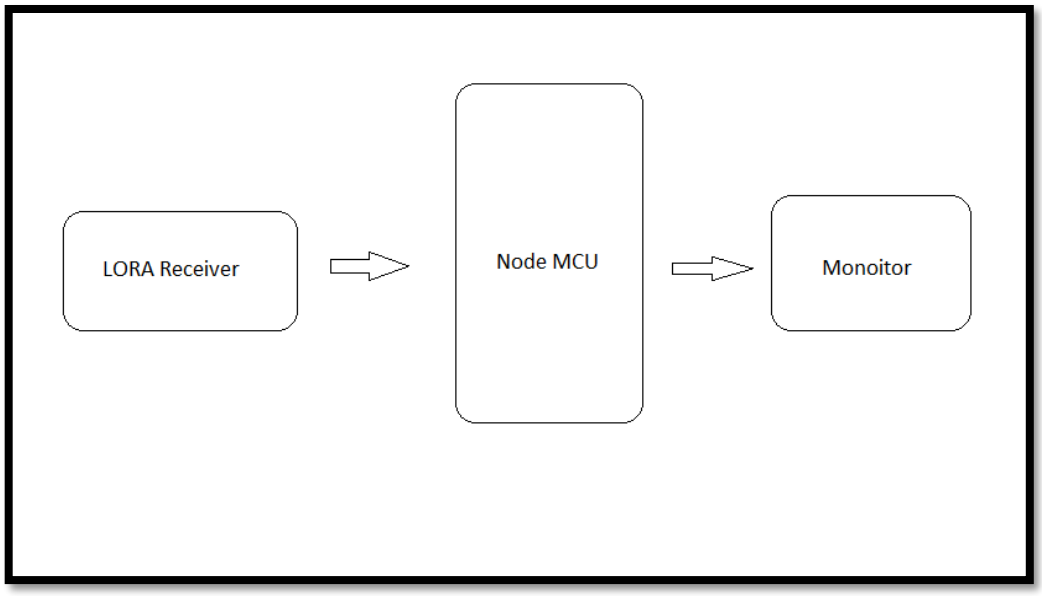


Fig 4.8 – RECEIVER BLOCK

MEMS SENSOR**:**

The cost of the mems sensor is not expensive. These can be used in the industrial applications. MEMS is referred as “Micro Electro-Mechanical System sensor.”

- It is a chip based technology. In this project we are using mems sensor to identify an accident.
- Whenever the mems sensor is tilted we can tell that an accident has occurred. The angle will be titled whenever there are speed breakers and sudden breaks.
- At this point of the time no message should be sent to the hospital. To avoid this we will be setting a value.
- If the angle is titled above the given value then the message will be sent to the hospital.

The mems sensor has ST, X,Y,Z,GND,VCC pins. We will be connecting the X to the A0 pin of the arduino. And the VCC and GND pins to the VCC and GND pins of the Arduino repectively. When tilt is applied to the sensor, there will be a difference in electric potential which is measured as a change in capacitance.

*GPS***:**

GPS refers to Global Positioning System. The system consists satellites and ground based control installations. GPS sensor contains a surface mount chip which is used to process the signals from GPS satellites using a rectangular antenna, which is mounted on the top of the GPS chip.

Whenever the accident takes place the GPS sensor will send the location where the accident has occurred. The gps is connected to the arduino board. The tx and rx pins of the gps are connected to the rx and tx pins of the arduino respectively. GPS sensor requires DC power supply. We usually use GPS to find our current location or to enroute to a new place. We can see the application of gps in various areas in our day to day lives, example: food delivery apps, bus tracking system and etc. GPS receiver can differentiate signals from at least four satellites by comparing their received pseudo random bit sequences and can calculate receiver's distance to each of these satellites by comparing arrival times of satellite signals.

It starts giving the data as soon as it identifies the satellites within its range. The transmission rate is either 4800 bps or 9600 bps and uses {8 bits, no parity, 1 stop bit} for decoding. The data block is known as a sentence which are of 80 characters in length. These gps sentences contain latitude, longitude, altitude. These sentences are decoded by connecting a microcontroller with GPS module and writing small program.

*ARDUINO***:**

- Arduino is an open-source platform which can be used for building electronics projects.
- Arduino contains a physical programmable circuit board.
- It is also called as a microcontroller where the programs can be executed by using software, or IDE (Integrated Development Environment).
- The software will run on your computer, and can be used to write and upload code to the physical board.
- The arduino board contains 14 digital I/O pins (six capable of PWM output), 6 analog I/O pins, and is programmable with the Arduino IDE (Integrated Development Environment)

To run the program on Arduino you should install the Arduino Desktop IDE .The Uno will be programmed using the Arduino Software (IDE), Integrated Development Environment which is common to all our boards. Connect your Uno board with an an USB cable; this cable is called as an USB printer cable

These days many people try to use the arduino because it makes things easier due to the simplified version of C++ and the already made Arduino microcontroller (atmega328 microcontroller) that you can program, erase and reprogram at any given time.

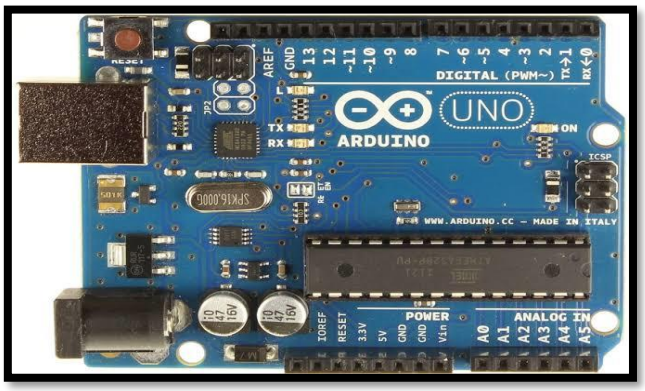


FIG 4.9 - ARDUINO BOARD

NODE MCU**:**

- ESP8266 is a WiFi Module chip that can shape to connect to the Internet for Internet of Things (IoT) and similar Technology Projects.
- NodeMCU is an open source based firmware developed for ESP8266 wifi chip.
- NodeMCU firmware comes with ESP8266 Development board or kit i.e. NodeMCU Development board.
- Basically, normal Electrical and Mechanical equipments cannot connect to the Internet on their own. They don’t have the in-built setup.
- NodeMCU is a Firmware on ESP8266. It’s basically an SoC (System on Chip).
- A System on a Chip (SoC) is a combined circuit that can integrate all components of a computer or other electronic systems.
- NodeMCU has Arduino like Analog (i.e. A0) and Digital (D0-D8) pins on its board.

LORA **:**

- LoRa is called as Long Range. It is a spread spectrum inflection technique which is derived from the chirp spread spectrum (CSS) technology.
- It has been used for communication in military and space applications for decades due to the long distance communication, robustness and interference.
- LoRa is a wireless radio frequency technology. which has a long range when compared to the Bluetooth, low power wireless platform that has become the most widely used technology for Internet of Things (IoT) networks worldwide.
- Direct Spread Spectrum Systems (DSSS) will adjust the carrier phase of the transmitter in accordance with a high frequency code called chip sequence. This code sequence will spread the signal bandwidth of the data signal by an amount referred to as processing gain.
- The receiver recovers the required data signal by re-multiplying with a locally generated replica of the spreading sequence. This will compress the spread signal back to the original bandwidth.
- Besides enhancing the transmission signal, interfering signals are also reduced by the processing gain of the receiver.
- These are dispersed beyond the desired bandwidth and can be easily removed by filtering. Direct Spread Spectrum Systems need highly accurate reference clock sources to process long code sequences.
- Power-constrained devices need to be repeatedly and rapidly synchronized.

4.5 MODULE FUNCTIONALITY

TRANSMITTER PHASE

In the transmitter phase we will be having one LoRa module (which will be used as a transmitter), to which we will be connecting a gps module and a mems sensor. These all will be connected to an Arduino board. Gps is used to detect the location (where the accident has occurred). whereas the mems sensor is used to detect the accident. The connections in the transmitter phase are as follows.

GPS TO ARDUINO

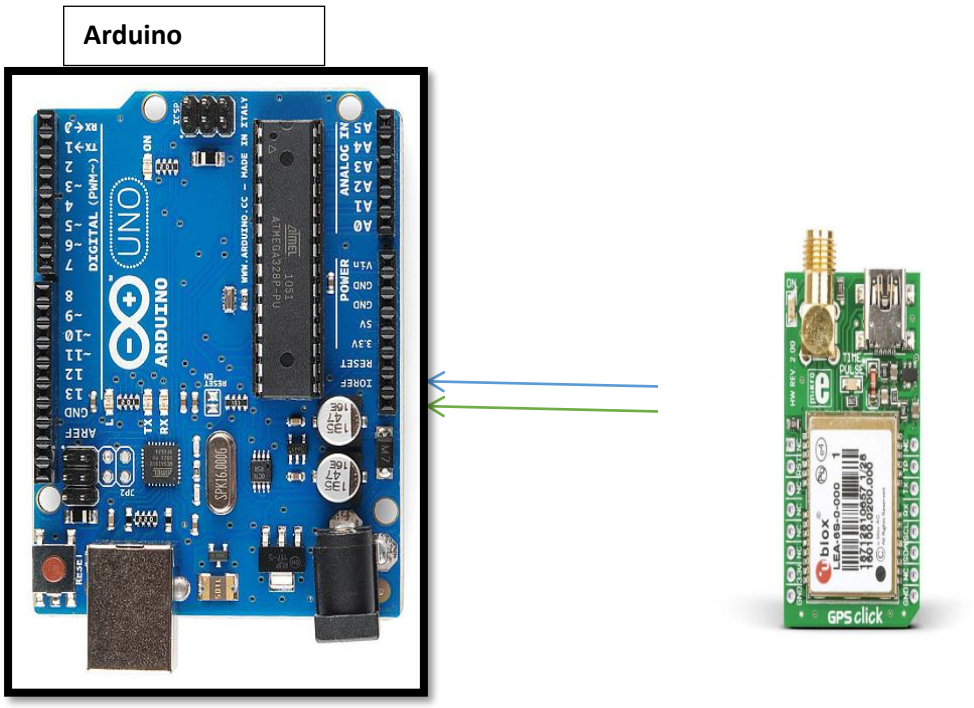


FIG 4.10 – GPS TO ARDUINO CONNECTION

The gps module is connected to the arduino. Usually the RX and TX pins are connected to the TX and RX pins of the arduino respectively. But in the transmitter side, the lora module should also be connected to the arduino's TX and RX pins. So we will be taking two pins 2 and 3 (of arduino) to which the TX and RX pins of gps will be connected.

THINGS TO REMEMBER ABOUT DIGITAL

- Digital Input/ Output uses the Digital pins, but whereas the Analog In pins can be used as Digital
- To receive a Digital signal use: digitalRead(pinNumber);
- To transmit a Digital signal we should use: digitalWrite(pinNumber, value); Digital Input and Output are either HIGH or LOW

LORA TO ARDUINO

The RX and TX pins of lora are connected to the TX and RX pins of the arduino. So whenever the accident occurs the lora transmitter transmits the required information to the lora receiver which is placed in the nearby hospitals.

Advantages of LoRa technology

- It can be utilized to communicate with thousands of IoT nodes and devices.
- It supports long-range communication between 5-15km.
- Power consumption is very low when compared with other devices. Hence, it is a best choice for battery critical applications.
- The frequency band used for LoRa is available in all countries.

MEMS SENSOR TO ARDUINO

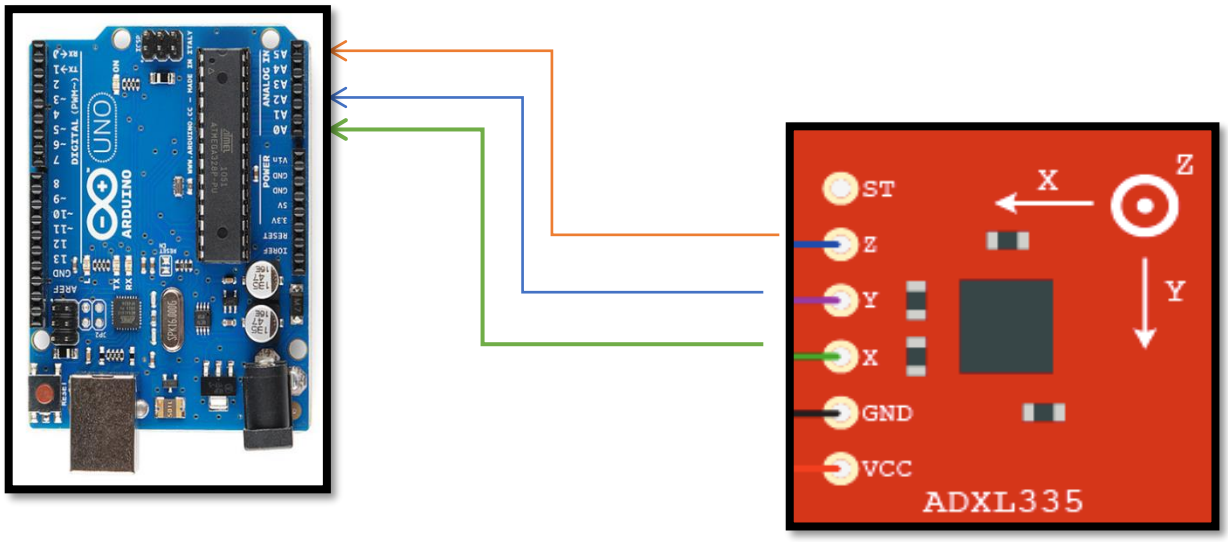


FIG 4.11 – MEMS SENSOR TO ARDUINO CONNECTION

- Meme sensor will be having ST, X,Y,Z,GND and VCC pins.
- The X pin of the sensor will be connected to the analog pin of the arduino that is A0. The GND and VCC of memes sensor will be connected to the GND and VCC of the arduino respectively.

THINGS TO REMEMBER ABOUT ANALOG:

- Analog Input will use the Analog In pins, Analog Output uses the PWM pins
- To receive an Analog signal use: analogRead(pinNumber);
- To send a PWM signal we should use: analogWrite(pinNumber, value);
- Analog Input values usually range from 0 to 1023 (1024 values because it uses 10 bits, 210)
- Whereas the PWM Output values range from 0 to 255 (256 values because it uses 8 bits, 28)

RECEIVER PHASE

Coming to the receiver's phase we will be having Node mcu , Lora receiver and an USB cable which is used to connect the node mcu to the monitor (for viewing the output). The connections are shown below.

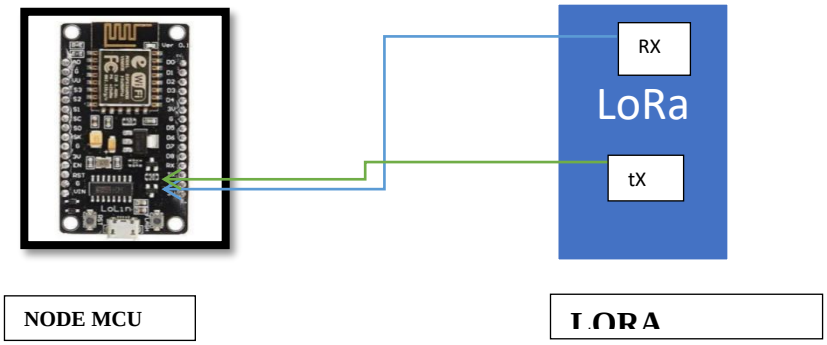


FIG 4.12 – NODE MCU TO LORA CONNECTION

In the receivers phase, the TX and RX pins of Lora receiver are connected to the RX and TX pins of the Node mcu respectively.

- Here we are using node MCU, which will be having an inbuilt wifi feature whereas in the arduino there is no inbuilt wifi.
- If at all we want to see the output through an app in our mobiles we can connect our phones to the node MCU with the help of wifi.
- But in our project we will be viewing the output on the monitor.
- So with help of an USB cable we can connect the node MCU to the monitor

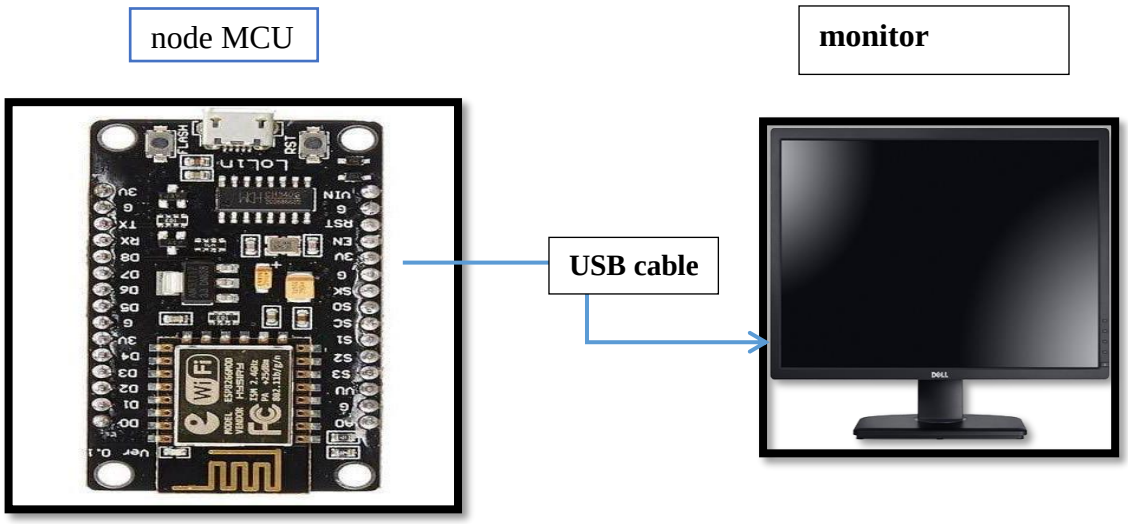


FIG 4.13 – NODE MCU TO MONITOR CONNECTION

MEMS SENSOR

- The mems sensor is connected to the arduino board.
- The x pin of mems sensor is connected to the A0 of the arduino.
- And the gnd and vcc of mems sensor are connected to the gnd and vcc of arduino respectively.
- Whenever the value of mems sensor is greater than 400 ,the location is captured by the gps module and it is sent through the Lora transmitter.
- The value of mems sensor ranges from 0 to 1024.
- Here the mems sensor value, that is the limit is being set to 400. All the situations like sudden breaks, going on the speed breakers can be avoided.
- Because in situations like these we don't need any medical support.
- The mems sensor value will be greater than 400 for the situations where an immediate action must be taken.

GPS

- The tx and rx pins of gps are connected to the tx and rx pins of the arduino.
- The gps location is taken based on the longitudes and latitudes.
- Here the user need not transmit any data, and it operates independently.
- We can use gps in any system, either in mobiles or laptops.
- The GPS is used to provide critical positioning capabilities to military, civil, and commercial users around the world.
- The gps calculates or identifies a location based on the information generated by the satellites.
- Generally Nine satellites are detectable from any point on the ground.
- The location is sent to the hospital through the Lora transmitter

LORA TRANSMITTER

- The main aim of our project is to reduce the delay in the communication when an accident occurs.
- Lora is the new technology. When compared to the Bluetooth the Lora allows a long range communication and it consumes less power.
- It also has a long battery life.
- Low Power Wide Area Network like LoRa WAN on the other hand is built from ground-up. LoRa is developed to work only with IoT devices which require best class battery life.
- Also, LoRa kind of network is great of IoT application as they need very minimal cost for deployment.
- The Lora transmitter transmits the location where the accident has occurred to the hospital and this information is received by the Lora receiver.
- Since the tx and rx pins are used by the other modules in the transmitter phase we have connected the tx and rx pins of the Lora transmitter to the #2 and #3 pins of the arduino.

ARDUINO

- The arduino takes the input from the gps.
- The arduino board has pins which will act as input and output pins.
- The board contains set of digital and analog input/output (I/O) pins. We can interface to various expansion boards (shields) and other circuits based on the requirements.
- The arduino has an inbuilt analog to digital convertor that is the ADC.
- This will convert the analog voltage signals to digital values.
- The digital signals will have two values high and low. And to read the value of an analog pin we use analogRead(pin).
- This function will convert the value of the voltage on an analog input pin

to digital value from 0 to 1023.

- To map an analog input value, which ranges from 0 to 1023 to a PWM output signal that ranges from 0 - 255, we can use the function map (value, fromLow, fromHigh, toLow, toHigh) .
- It has five constraints, one is the variable (where the analog value is stored) , while the rest are 0, 1023, 0 and 255 respectively
- The value of the voltage can be 5v or 3.3v.
- The arduino doesn't have an inbuilt digital to analog convertor (DAC). We can use PWM pins (pulse width modulation pins) in order to convert the digital signal to achieve few functions of analog signals.
- In the Arduino boards, the PWM function is accessible on pins 3, 5, 6, 9, 10, and 11.

Applications for further use are

LoRa, which is the latest technology is used these days to attain maximum efficiency in data transmission and reception. The applications of LoRa can be given in wide areas starting from military applications to space communicationStarting with military applications, LoRa can be used in the military base stations for a secured transmission of data or information between the stations. Due to huge distance between the various military bases, the data cannot be transmitted using Bluetooth, WIFI or ZigBee as their transmission range is very less. LoRa, having a transmission width of more than 100 km can be used to send the information in a safe and secured manner. The other main reason through which this application can be proven more feasible is its wireless transmission medium. LoRa supports the best wireless transmissions, even to very large distances. Loss of information is the most desired feature of LoRa.A human life is the most precious resource on this earth. It is this human life which gives life to many other man-made technologies like IoT etc. So, it the responsibility of us, the human beings to protect the lives of our race. Hence the utilization of LoRa can be used to send the information of the accident occurred and its location along with the details for the family members of the person in the car to the hospital as well as the nearby police station for immediate help. The help given by the hospital i.e. first aid or emergency checkup can save most of the lives in many situations.

[3]Lacuna Space, has successfully concluded its first phase of testing to provide complete global coverage for LoRa devices, anywhere in the world, no matter how remote. LoRa technology has started making its space even in the satellite IoT connectivity with low cost, high scalability and reliable global connections. The main aim of the Lacuna is to extend LoRa connectivity to enable direct communication from LoRa based devices to satellite gateways using the LoRaWAN protocol.

The following can be considered as the features for the development of the satellite communication using LoRa:

- Low-cost, battery-powered sensors transmit signal directly to a passing satellite.
- The signals use a long-range wide area network protocol called LoRaWAN which is specially designed to conserve battery.
- The satellites are low-cost (about the size of a shoe-box) and fly in a 500km orbit, circling the Earth fourteen times a day.

The satellites store the messages for a short period of time until they pass over the network of our ground stations.

5. PROJECT SYSTEM DESIGN

5.1 DFDS IN CASE OF DATABASE PROJECTS

Our project can use database to store the coordinates of the gps along with the dare and day of accident occurred. We can use cloud based data services to store the data and access it from anywhere. Lorawan gateway can be directly linked to database to reflect all the possible transactions between them. A few examples of the same are Mongodb or firebase etc. Today the IoT is enabling companies to blend the physical and digital worlds. Realizing the business value of connecting all of these “things” enables creation of new revenue models, improves productivity, and generates new insights that drive operational efficiencies. The IoT already connects billions of devices worldwide, and that number is growing daily. Many market observers predict that only by adopting IoT can organizations fully unlock the revenue opportunities promised by digital transformation. MongoDB can help you rapidly capture the most value from the IoT. The Database can also be a source of data analytics, which can seriously boost the development of the next level IoT based products on the same technology. IoT and data remain intrinsically linked together. Data consumed and produced keeps growing at an ever expanding rate. This influx of data is fueling widespread IoT adoption as there will be nearly 30.73 billion IoT connected devices by 2020. The Internet of Things (IoT) is an interconnection of several devices, networks, technologies, and human resources to achieve a common goal. There are a variety of IoT-based applications being used in different sectors and have succeeded in providing huge benefits to the users. The data generated from IoT devices turns out to be of value only if it gets subjected to analysis, which brings data analytics into the picture. Data Analytics (DA) is defined as a process, which is used to examine big and small data sets with varying data properties to extract meaningful conclusions and actionable insights. These conclusions are usually in the form of trends, patterns, and statistics that aid business organizations in proactively engaging with data to implement effective decision-making processes.

5.2 E-R DIAGRAMS

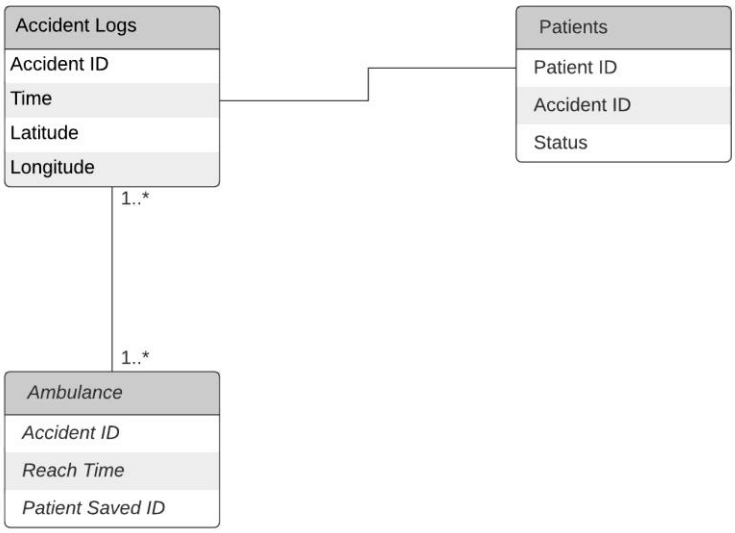


FIG 5.1- E-R diagram

5.3 UML DIAGRAMS

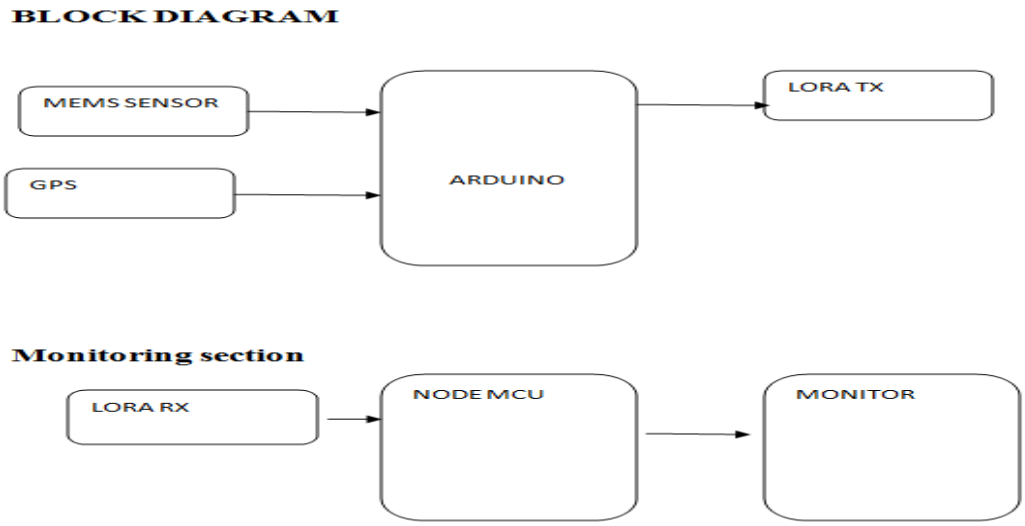


FIG 5.2 BLOCK DIAGRAM

The whole system is based on Lora and gps connected to arduino to monitor the vehicle movement. If any vehicle comes under accident then the arduino passes alert message to near by hospital through lora communication.

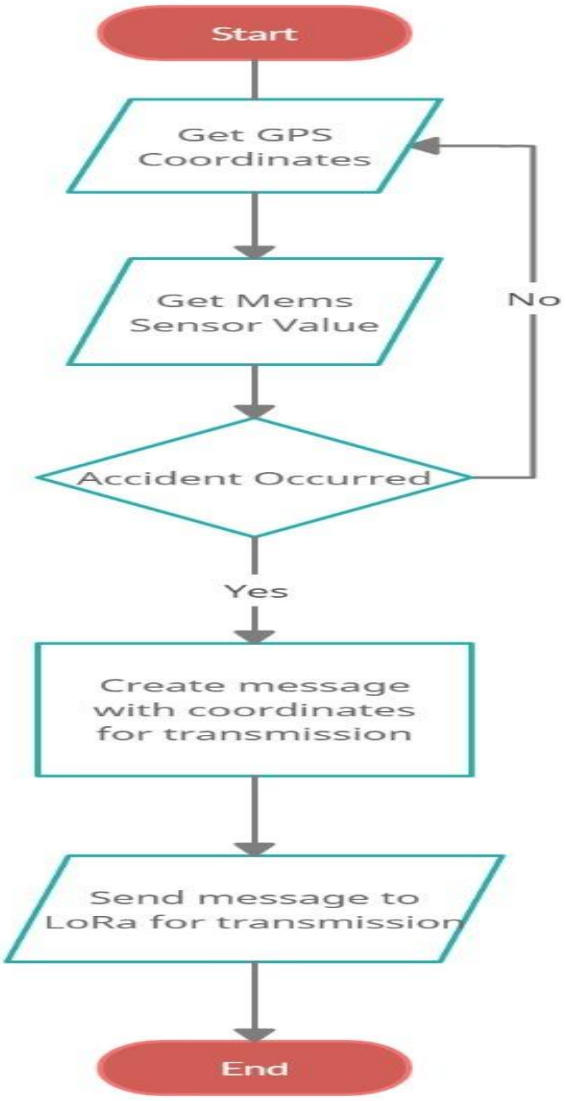


FIG 5.3 TRANSMITTER SIDE USE CASE DIAGRAM

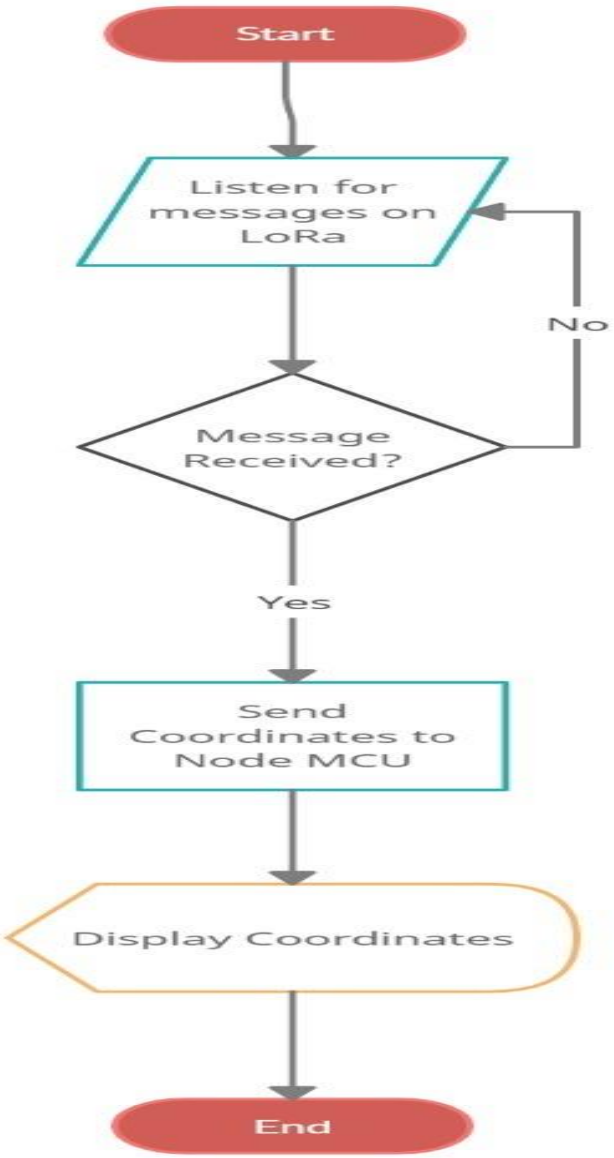


FIG 5.4 RECEIVER SIDE USE CASE DIAGRAM

6. PROJECT CODING

6.1 CODE TEMPLATES

Transmitter:

```
#include <SoftwareSerial.h>
SoftwareSerial gps(9,8); // RX, TX

char str[70];
String gpsString = "";
char *test = "$GPGGA";
String latitude = "No Range      ";
String longitude = "No Range     ";
int temp = 0, i;
boolean gps_status = 0;



int memspin = A0;
int mems;
int x;
int y;

void setup()
{
  Serial.begin(9600);
  gps.begin(9600);
  get_gps();
  pinMode(memspin, INPUT);
}

//long lastTime=millis();
```

```
void loop()
{

//Check for every second approach
//  if(millis()-lastTime>1000)
//  lastTime=millis();
//   }
//  Serial.println(latitude+" "+longitude);

  mems = analogRead(memspin);
  delay(1000);
  if ( mems<=370 || mems >= 420)
  {
    tracking1();
    delay(1000);

  }

}

void gpsEvent()
{
  gpsString = "";
  while (1)
  {
    while (gps.available() > 0)          //checking serial data from GPS
    {
      char inChar = (char)gps.read();
      gpsString += inChar;
```

```
    //store data from GPS into gpsString
    i++;
    if (i < 7)
    {
      if (gpsString[i - 1] != test[i - 1])    //checking for $GPGGA sentence
      {
        i = 0;
        gpsString = "";
      }
    }
    if (inChar == '\r')
    {
      if (i > 65)
      {
        gps_status = 1;
        break;
      }
      else
      {
        i = 0;
      }
    }
  }
  if (gps_status)
    break;
 }
}

void get_gps()
{
```

```
  gps_status = 0;
  int x = 0;
  while (gps_status == 0)
  {
    gpsEvent();
    int str_lenth = i;
    latitude = "";
    longitude = "";
    int comma = 0;
    while (x < str_lenth)
    {
      if (gpsString[x] == ',')
        comma++;
      if (comma == 2)     //extract latitude from string
        latitude += gpsString[x + 1];
      else if (comma == 4)     //extract longitude from string
        longitude += gpsString[x + 1];
      x++;
    }
    int l1 = latitude.length();
    latitude[l1 - 1] = ' ';
    l1 = longitude.length();
    longitude[l1 - 1] = ' ';
    i = 0; x = 0;
    str_lenth = 0;
    delay(100);
  }
}

void send_data(String message)
```

```
{
  Serial.println(message);
  delay(100);
}

void tracking1()
{

 String x=(latitude+'@'+longitude+'!');

  Serial.println(x);

 }
```

Receiver Side Code:

```
void setup()
{
  Serial.begin(9600);
}
void loop()
{
    while(1){
      if (Serial.available())
      {
        String latitude = Serial.readStringUntil('@');
        String longitude = Serial.readStringUntil('!');

        Serial.println("ACCIDENT DETECTED NEARBY\n");
        Serial.println("LATITUDE: "+latitude+" LONGITUDE: "+longitude);
       break;
      }
```

```
    }
}
```

6.2 OUTLINE FOR VARIOUS FILES

SerialSoftware:

The Arduino hardware has built-in support for serial communication on pins 0 and 1 (which also goes to the computer via the USB connection). The native serial support happens via a piece of hardware (built into the chip) called a UART. This hardware allows the Atmega chip to receive serial communication even while working on other tasks, as long as there is room in the 64 byte serial buffer.

The SoftwareSerial library allows serial communication on other digital pins of the Arduino, using software to replicate the functionality It is possible to have multiple software serial ports with speeds up to 115200 bps. A parameter enables inverted signaling for devices which require that protocol.

Arduino has just one Serial Port placed at pins 0 and 1. So, if you are having two or more serial modules, then there's a difficulty in adding two modules because we just have one serial port. So, in such cases there's a need to add one more serial port and that serial port can be created at any two pins of Arduino and that serial port is called software serial.Software Serial is also named as Virtual Serial Port. It's really convenient if one is working on serial modules. It is best to never use hardware serial port because pin 0 and pin 1 are also used for uploading code and debugging the code via Arduino Serial Monitor. So, one should connect the Serial modules via software serial and then check their output on Serial Monitor.

Other useful methods available in the SerialSoftware module:

- SoftwareSerial()
- available()
- begin()
- isListening()
- overflow()
- peek()
- read()
- print()

- println()
- listen()
- write()

6.3 CLASS WITH FUNCTIONALITY

Functions used:

setup():

The setup() function is called when a sketch starts. It is used to initialize variables, pin modes, start using libraries, etc. The setup() function will only run once, after each powerup or reset of the Arduino board.

In the project, it is used to setup the modules such as GPS pins and get the initial GPS latitude and longitude location. It is also used to setup the MEMS sensor connection so that the values of the sensor can be used later to detect the accident if and when it occurs and the GPS locations which are used to get the coordinates in the setup function are then transmitted to the receiver.

loop():

After creating a setup() function, which initializes and sets the initial values, the loop() function does precisely what its name suggests, and loops consecutively, allowing the program to change and respond. It is used to actively control the Arduino board.

In the project, inside the transmitter module, we use it to get the value of the mems sensor which is in turn used to determine if the accident has occurred. The MEMS sensor value is read continuously in the loop function and if the value falls out of the range of the minimum and maximum threshold values, it is noted that an accident has occurred. Now the real-time GPS value is fetched and is sent to the receiver. The loop function does this repeatedly for as long as the board is powered-up.

get_gps():

The main function that handles gps triangulation of the vehicle. This function takes the input from the GPS module which outputs NMEA sentences containing the location of the vehicle along with status codes describing the status of triangulation process. These sentences are then decoded and the latitude and longitude locations are obtained.

The function stores these coordinates in two global variables called latitude and longitude. These global variables are used in the process of transmission.

6.4 METHODS INPUT AND OUTPUT PARAMETERS.

Vehicle Side Input and Output:

Input:

The Arduino takes input from two modules, namely GPS and MEMS sensor. The inputs are then used to detect an accident and transmit the coordinates to the hospital .

GPS module uses GPS triangulation to locate the vehicle and returns the coordinates in the form of NMEA sentences.These coordinates are then taken as input in the Arduino and are decoded to obtain the latitude and longitude location.

The MEMS sensor detects movement in x,y,z coordinates which allow us to know if the vehicle has tilted. The MEMS sensor constantly sends the value as output, which is taken as input in the Arduino and is used to detect if an accident has occurred.

The output produced by these two modules are taken as input in the Arduino and are used to detect accident and locate the vehicle.

Output:

After locating the vehicle and detecting the coordinates, the Arduino uses it to build a message containing the latitude and longitude delimited by special characters. This message is sent as output to the LoRa module.

The LoRa module takes the message as input and transmits it using radio frequency.

Hospital Side Input and Output:

Input:

The message sent by the vehicle's LoRa module is captured and taken as input in the LoRa module at the hospital. The module then outputs the message to the NodeMCU.

The board continuously keeps listening for messages from LoRa. When the data is available to read, the board takes the mssage as input.

Output:

Node MCU, after receiving the message as input from the LoRa module, sends it onto the serial output port. This is taken as input by the serial monitor on the computer and is displayed on the screen, hence producing the final output which is the coordinates received from the vehicle.

7. PROJECT TESTING

7.1 VARIOUS TEST CASES

IOT testing is a type of testing to check IOT devices. Today there is increasing need to deliver better and faster services. There is a huge demand to access, create, use and share data from any device. The thrust is to provide greater insight and control, over various interconnected IOT devices. Hence, IOT testing framework is important.

Testing for IoT devices broadly revolves around Security, Analytics, Device, Networks, Processors, Operating Systems, Platforms and Standards.

Let's investigate the broad testing types

Usability Testing:

There are so many devices of different shape and form factors are used by the users. Moreover, the perception also varies from one user to other. That's why checking usability of the system is very important in IoT testing.

Compatibility Testing:

There are lots of devices which can be connected though IOT system. These devices have varied software and hardware configuration. Therefore, the possible combination are huge. As a result, checking the compatibility in IOT system is important.

Reliability and Scalability Testing:

Reliability and Scalablity is important for buidling an IOT test enviorment which involves simulation of sensors by utilzing virtualization tools and technolgoies.

Data Integrity Testing:

It's important to check the Data integrity in IOT testing as it involves large amount of data and its application.

Security testing:

In the IOT environment, there are many users are accessing a massive amount of data. Thus, it is important to validate user via authentication, have data privacy controls as part of security testing.

Performance Testing:

Performance testing is important to create strategic approach for developing and implementing an IOT testing plan.

In this project we use two particular test cases for calculating when an accident has occurred and when to report the same for the accidents. The Mems sensor values are responsible for deciding whether an accident has occurred or not. The two possible ways are like to either check if the difference between the two consecutively read mems value is high or the difference between them is substancial then we can say an accident has occurred. The other possible way is to put a range for the mems value at which we must report an accident has occurred. The threshold of the values for mems sensor are decided or based up on the type of the vehicle and its centre of mass etc. based on these testing we can successfully decide when an accident has occurred.

In this project we decided to use the range of values as they gave more precise results as compared to using difference of consecutive values. We used a range of 370-400 for a small vehicle and the values decrease as the size of the vehicle increases the movement in vehicle required to make a an accident reduces.

7.2BLACK BOX

Black Box Testing is a software testing method in which the functionalities of software applications are tested without having knowledge of internal code structure, implementation details and internal paths. Black Box Testing mainly focuses on input and output of software applications and it is entirely based on software requirements and specifications. It is also known as Behavioral Testing.

FIG 7.1 BLACK BOX TESTING

The above Black-Box can be any software system you want to test. For Example, an operating system like Windows, a website like Google, a database like Oracle or even your own custom application. Under Black Box Testing, you can test these applications by just focusing on the inputs and outputs without knowing their internal code implementation.

There are many types of Black Box Testing but the following are the prominent ones -

- Functional testing - This black box testing type is related to the functional requirements of a system; it is done by software testers.

- Non-functional testing - This type of black box testing is not related to testing of specific functionality, but non-functional requirements such as performance, scalability, usability.
- Regression testing - Regression Testing is done after code fixes, upgrades or any other system maintenance to check the new code has not affected the existing code.

7.3 WHITE BOX TESTING

White Box Testing is software testing technique in which internal structure, design and coding of software are tested to verify flow of input-output and to improve design, usability and security. In white box testing, code is visible to testers so it is also called Clear box testing, Open box testing, Transparent box testing, Code-based testing and Glass box testing.

It is one of two parts of the Box Testing approach to software testing. Its counterpart, Blackbox testing, involves testing from an external or end-user type perspective. On the other hand, White box testing in software engineering is based on the inner workings of an application and revolves around internal testing.

The term "WhiteBox" was used because of the see-through box concept. The clear box or WhiteBox name symbolizes the ability to see through the software's outer shell (or "box") into its inner workings. Likewise, the "black box" in "Black Box Testing" symbolizes not being able to see the inner workings of the software so that only the end-user experience can be tested.

White box testing involves the testing of the software code for the following:

- Internal security holes
- Broken or poorly structured paths in the coding processes
- The flow of specific inputs through the code
- Expected output
- The functionality of conditional loops
- Testing of each statement, object, and function on an individual basis

The testing can be done at system, integration and unit levels of software development. One of the basic goals of whitebox testing is to verify a working flow for an application. It involves testing a series of predefined inputs against expected or

desired outputs so that when a specific input does not result in the expected output, you have encountered a bug.

8. OUTPUT SCREENS

8.1 USER INTERFACES

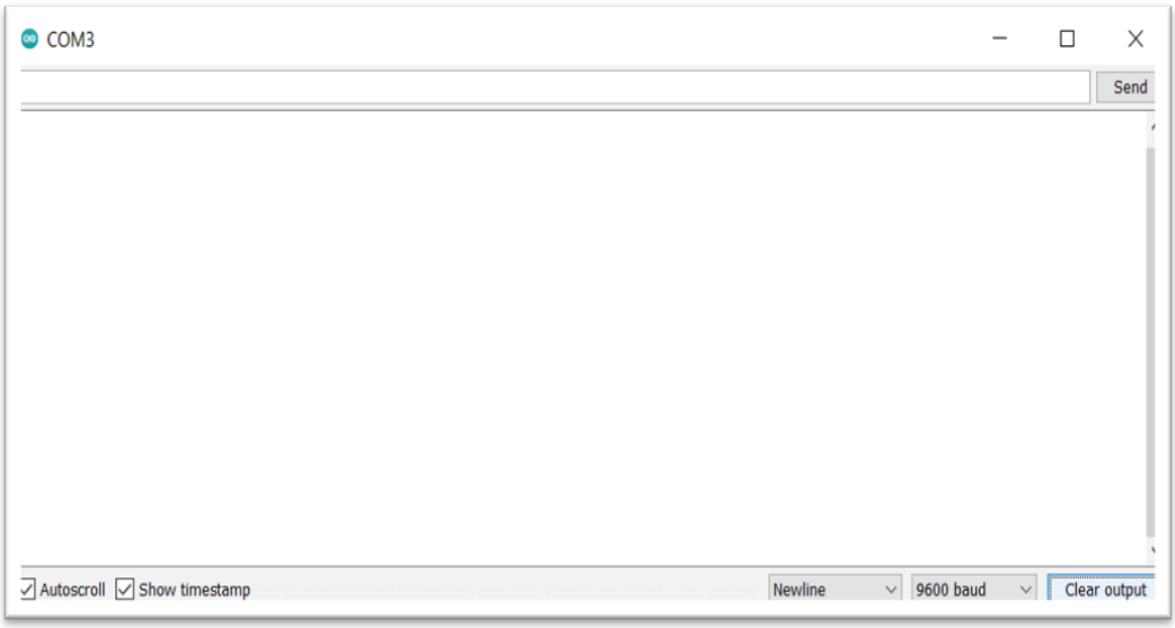


FIG 8.1 – SERIAL MONITER BEFORE ACCIDENT

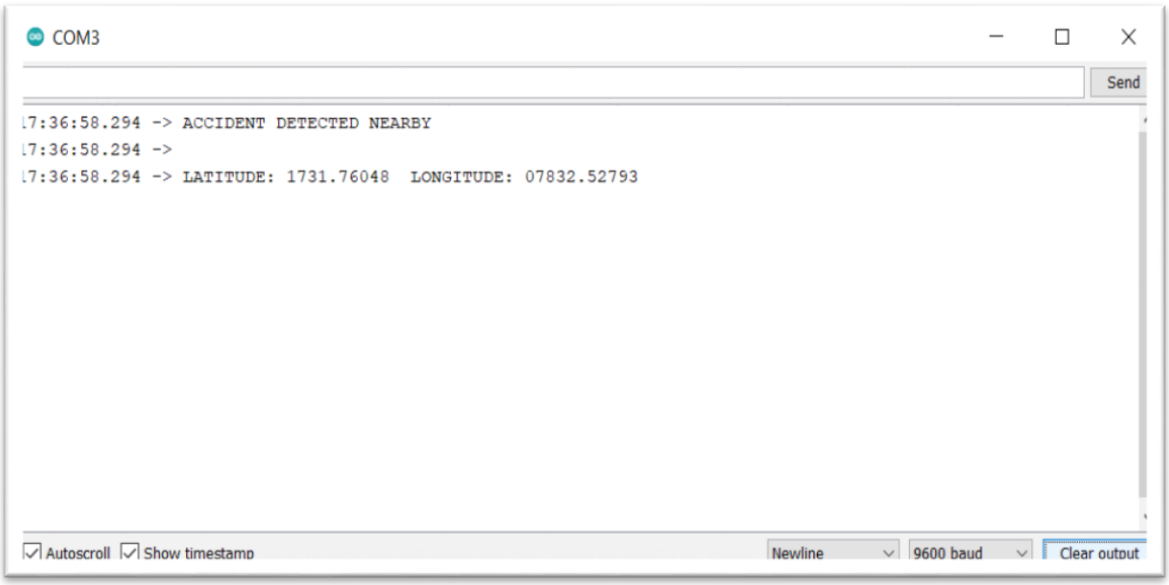


FIG 8.2 – SERIAL MONITER AFTER ACCIDENT

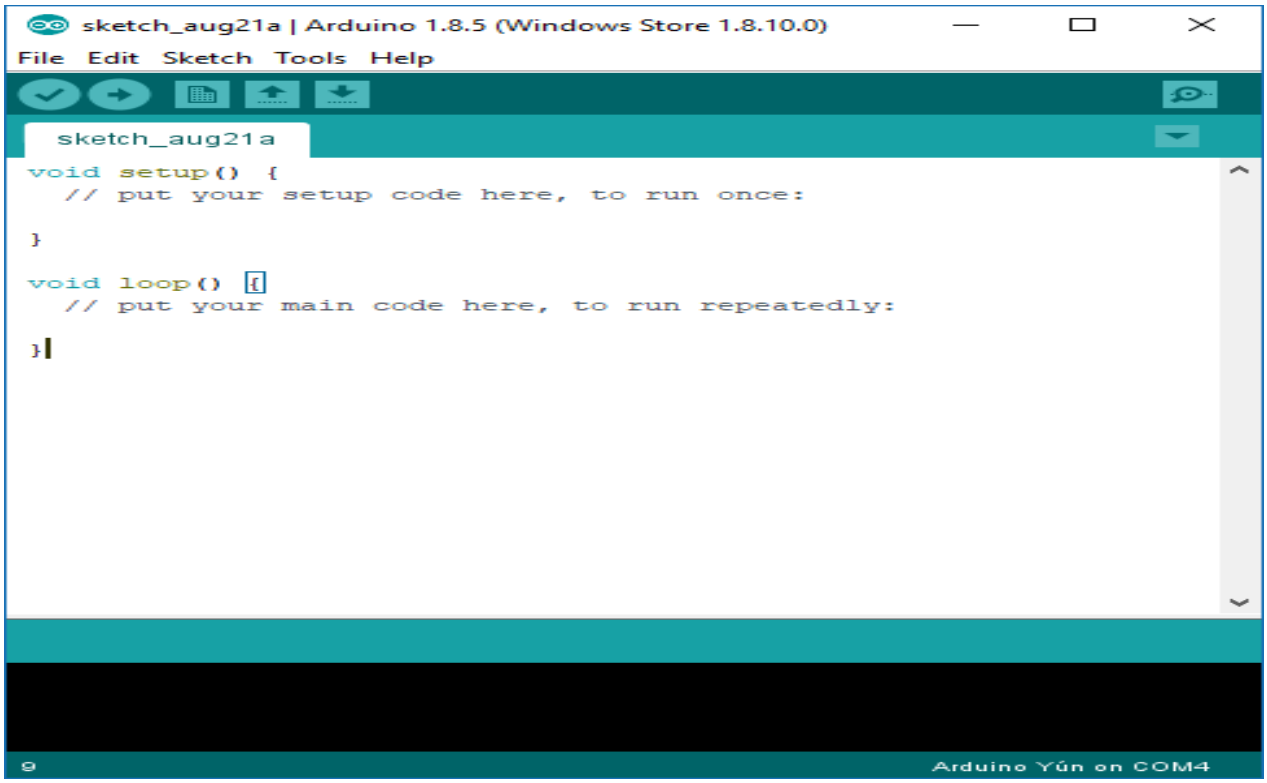


Fig 8.3 Arduino ide

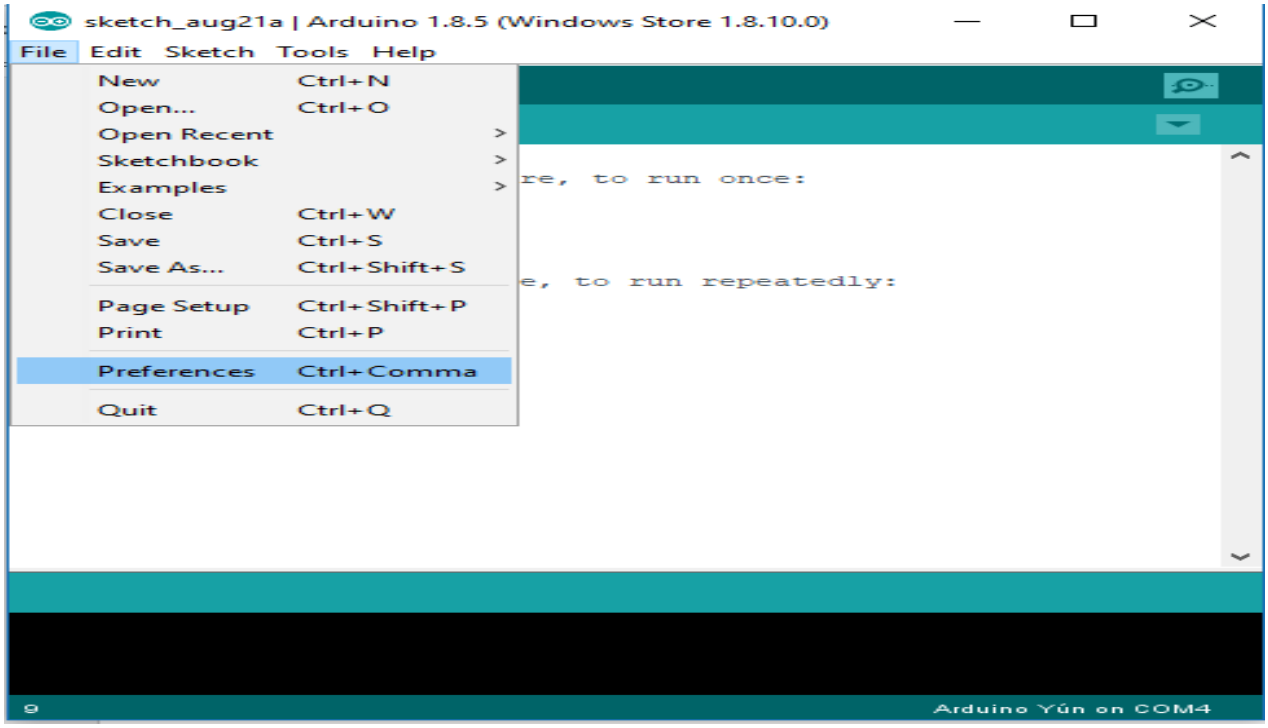


Fig 8.4 Select file and prefrences

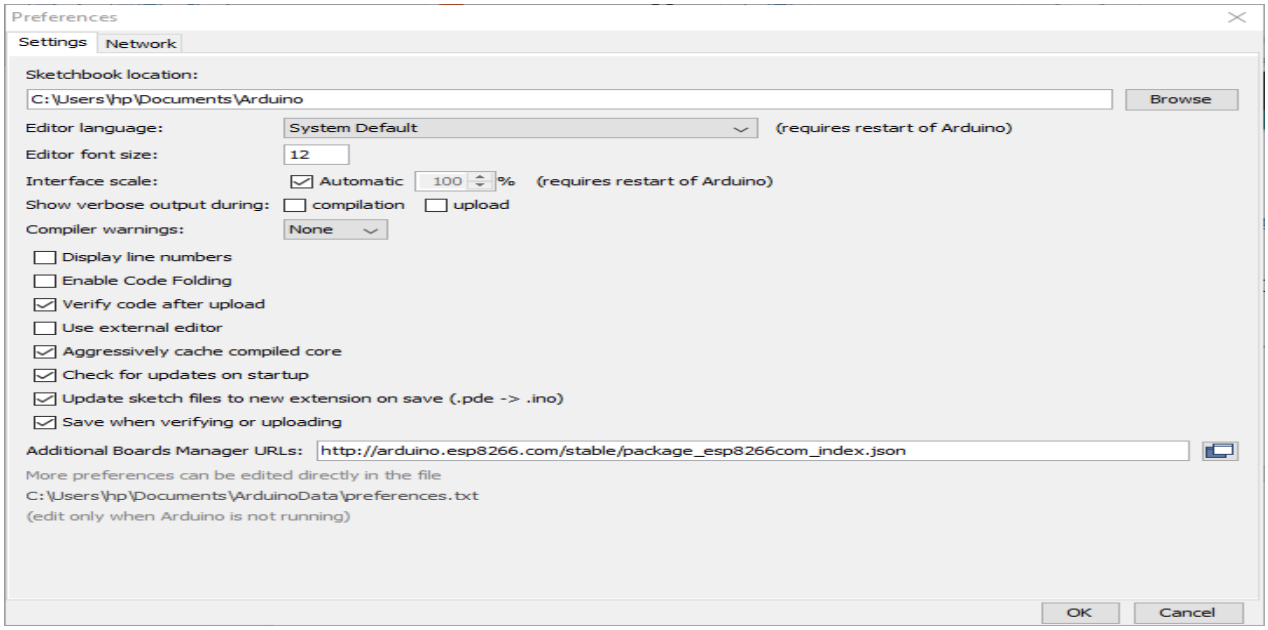


Fig 8.5 Additional Board Manager URLs:

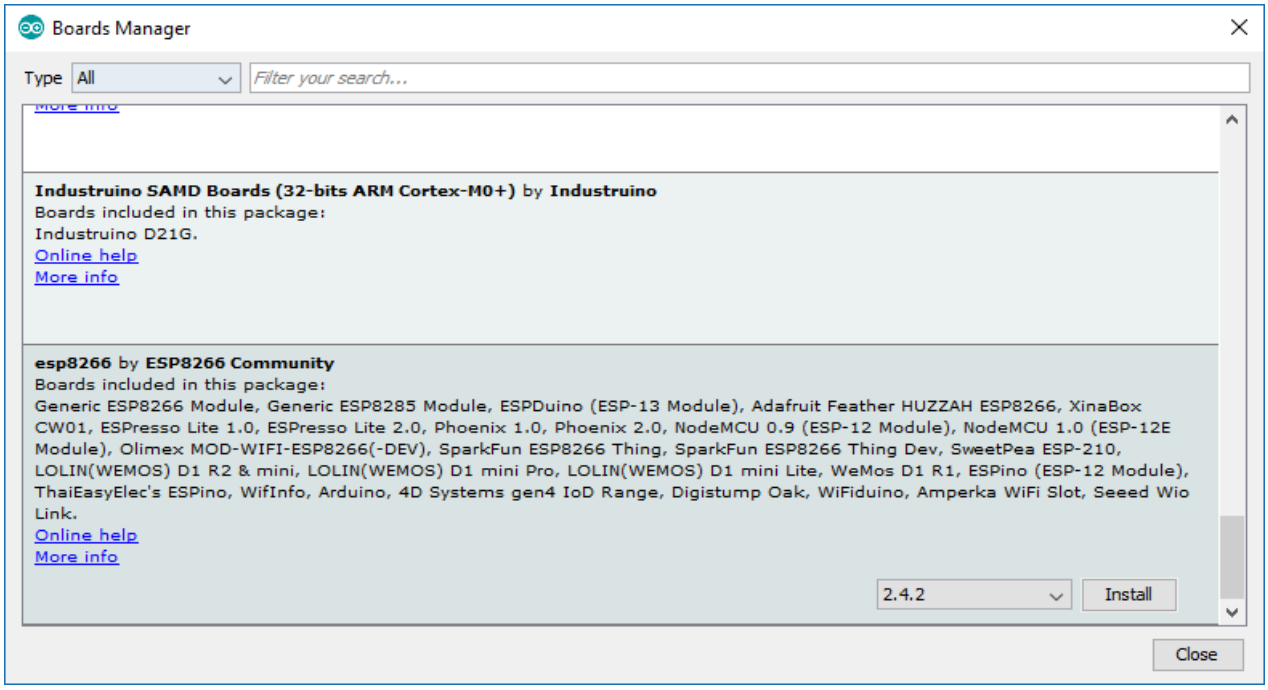


Fig 8.6 esp8266 by ESP8266 Community package installation:

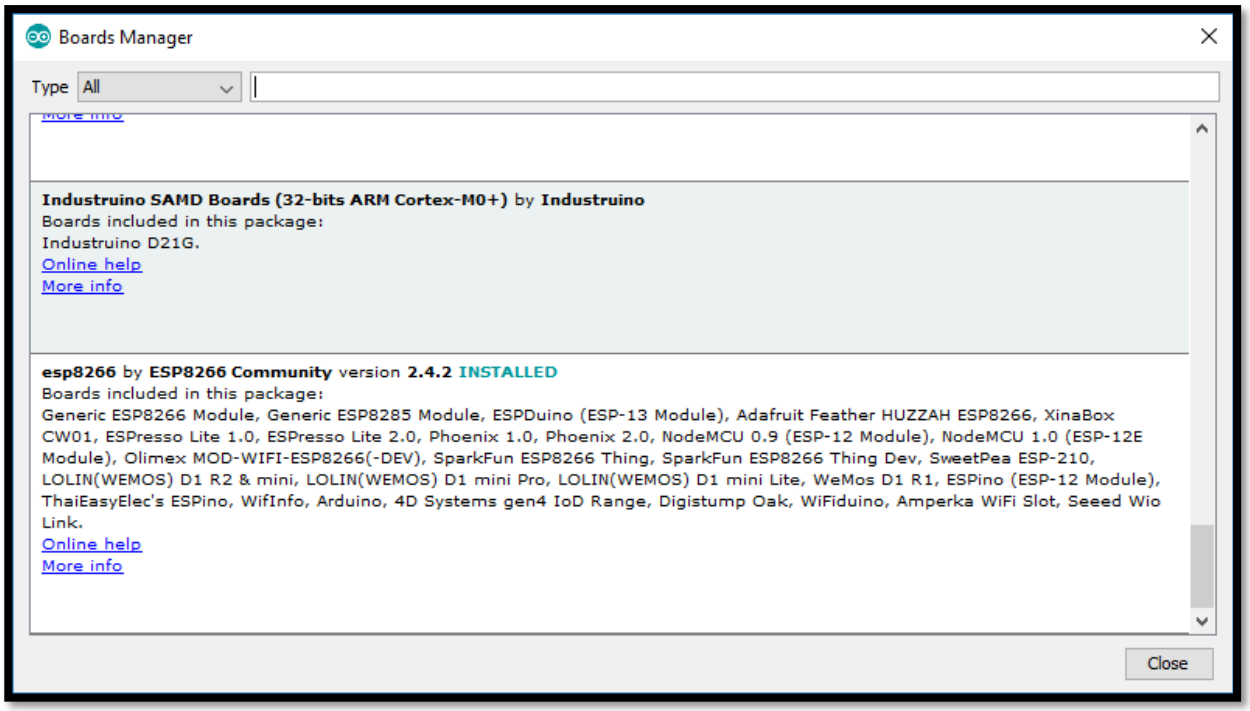


fig 8.7 ESP8266 board is successfully installed for Arduino IDE:

8.2OUTPUT SCREENS

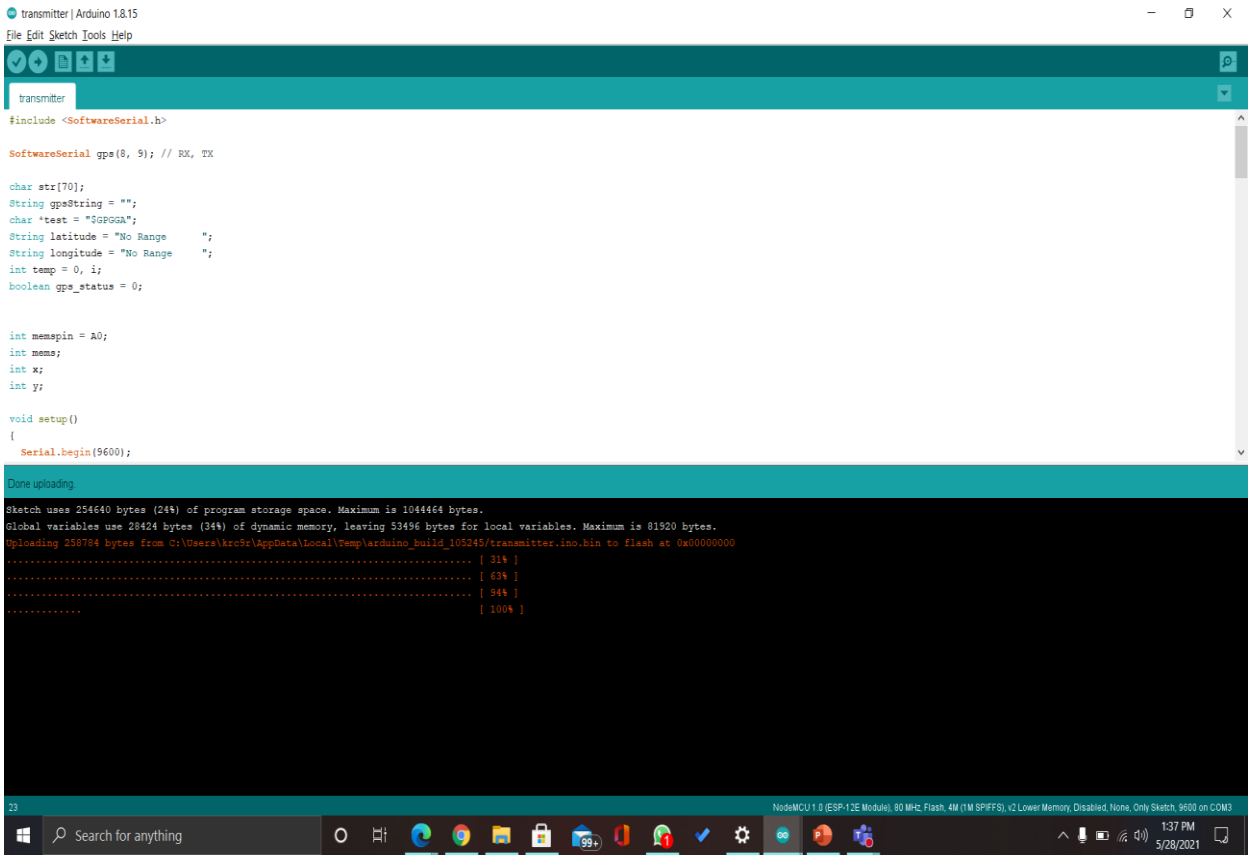


FIG 8.8 Uploading of Transmitter side code to Arduino

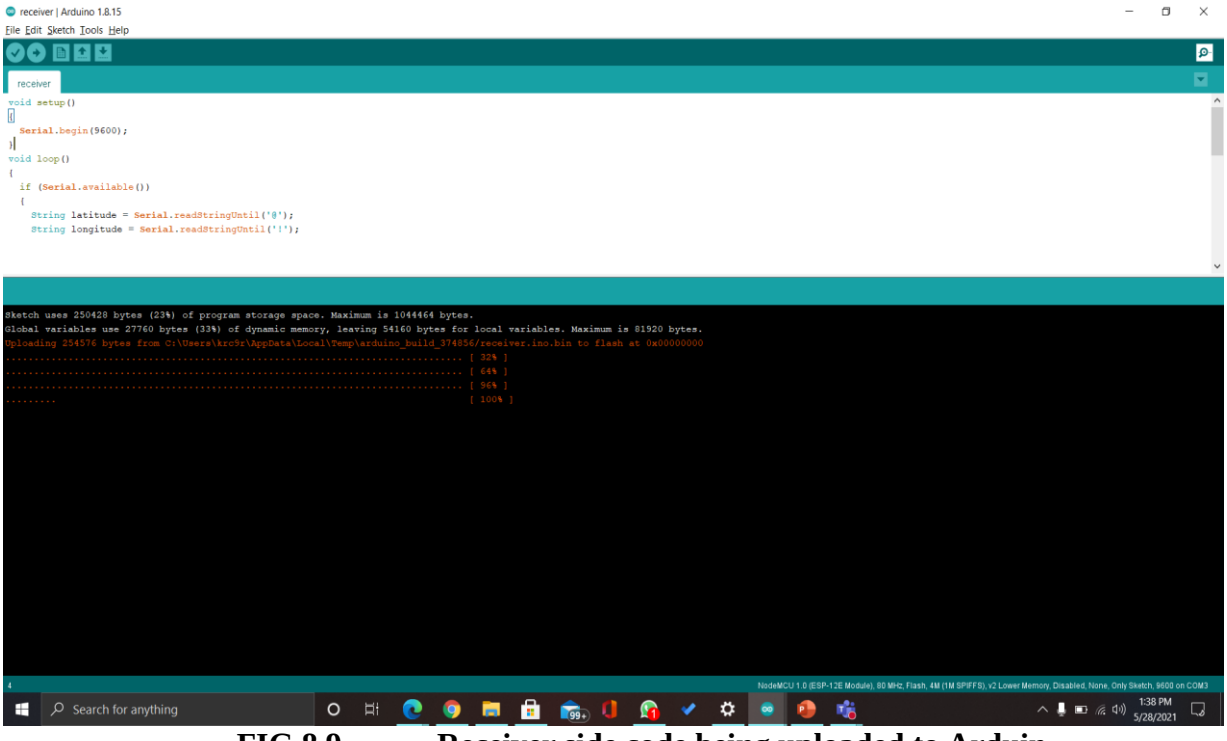


FIG 8.9 Receiver side code being uploaded to Arduin

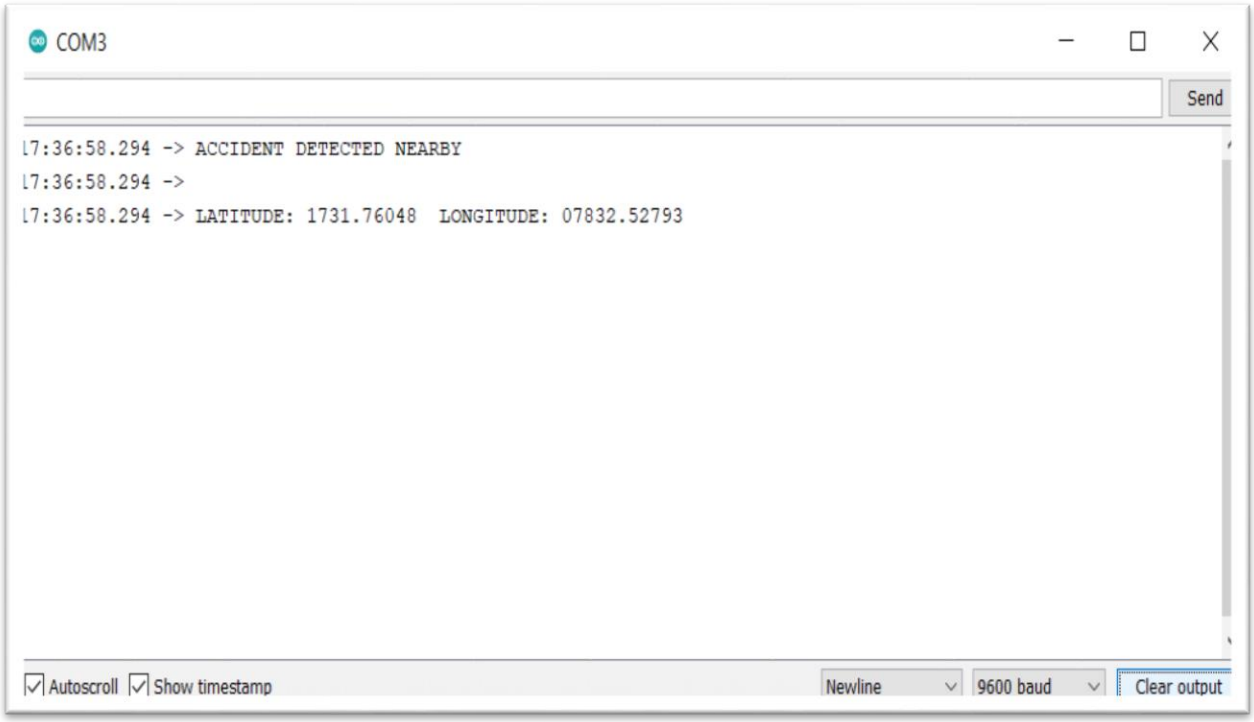


Fig 8.10 Serial monitor

9. EXPERIMENTAL RESULTS

The Experimental results were very satisfactory as we were able to successfully detect the ocuurence of a accident and this was clearly visible and understood through the demo and on the output serial monitor on the screen of the receiver side.

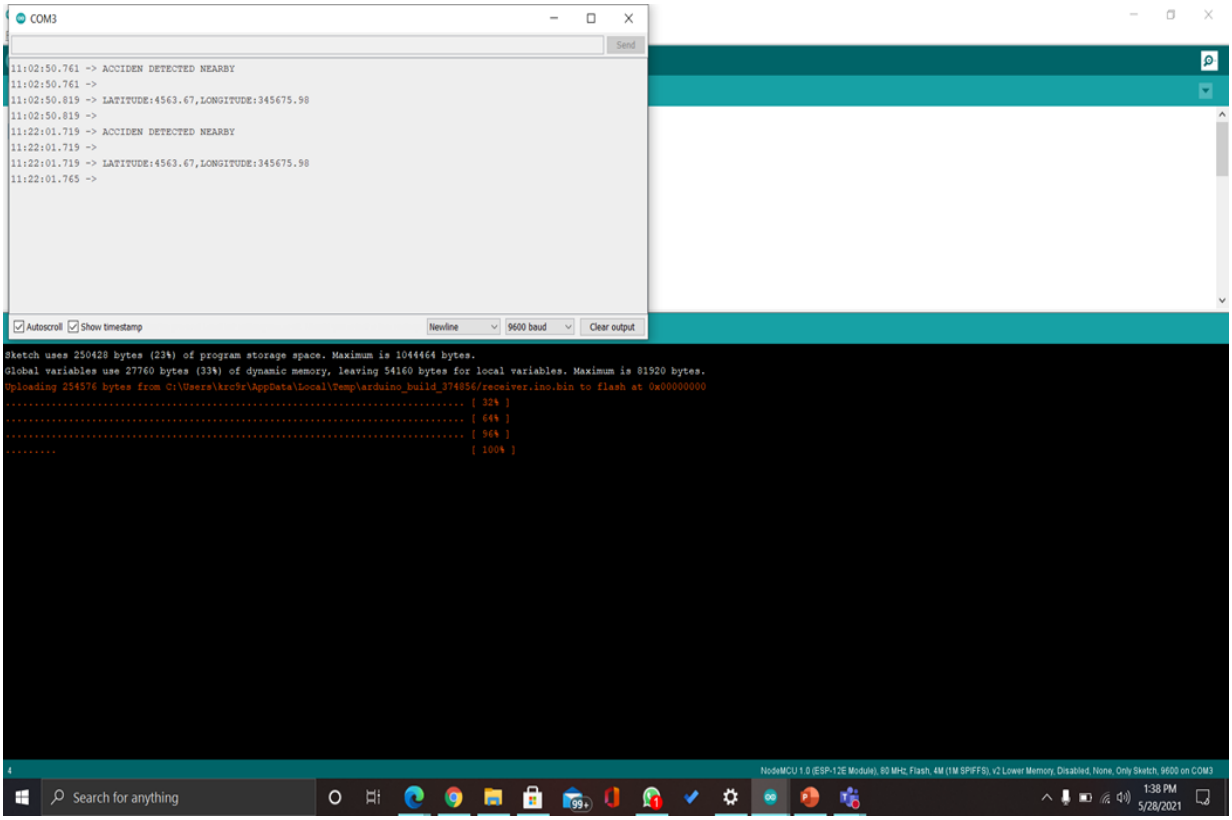


FIG 9.1 Result of coordinates on the serial monitor

10. CONCLUSION AND FUTURE ENHANCEMENT

CONCLUSION

About 1.25 million people are affected each year due to road accidents. The safety of the driver, the passengers and the public should be of primary concern in regard to various road safety measures that are being adopted by governments. Adopting a solution to prevent such fatal accidents due to inefficiency of human assisted driving is of utmost importance. With rise in technology and automation it has become possible to propose a plan to mitigate the adverse effects of Human assisted Driving.

This Project presents a new and innovative communication system between hospital and the vehicle to prevent loss of life. The proposed method in this paper relies on LoRa transmitter and receiver systems. We present a LoRa based open source system for tracking vehicles. We successfully built the prototype and proved the correctness of system operation. Future improvements can be done in various aspects This paper would encourage and motivate others to further explore the notion of LoRa Technology and hence, develop more efficient strategy to enhance the transmitting range of the LoRa system and propose innovative solutions to overcome the challenges.

FUTURE ENHANCEMENTS

For future enhancement, we may develop a vehicle tracking, monitoring and alerting system using combination of RFID system, GPS, GSM/GPRS with high speed processor which can make the smartest car possible. The system will have latest technology with moderate cost. The system may focus on accurate and real time position of vehicle. The system can be installed in cargo, trucks, buses and cars. Depending on the vehicle on which the safety device is inserted, the angle of the occurrence of the accident can be measured.

The node mcu can be WIFI enabled, which can also be facilitated in the development of an application which can be installed in the mobile phones for the easier monitoring and information access. By using the LoRaWAN or the LoRa gateway, the information and the location where the accident has occurred can be send to the nearest hospitals within the range of the location of the accident. The information can also be sent to the nearest police station which can help in any way possible. Hardware-wise, the LoRa transmitter and the receiver can be tested in more robust environmental conditions for their placement in any location of the vehicle. Software-wise, the scripts should be refactored to ensure better maintainability as the system grows.

11. REFERENCES

[1] Study on LTE-based V2X Services (Release 14), document TR 36.885 V14.0.0, TSG RAN, 3GPP, June 2016.

[2] Study on enhancement of 3GPP Support for 5G V2X Services (Release 15), document TR 22.886 V15.1.0, TSGS and SA, 3GPP, March 2017.

[3] “Timeline for deployment of LTE-V2X,” 5GAA, Munich, Germany, White Paper, Feb. 2018.

[4] K. Lee, J. Kim, Y. Park, H. Wang, and D. Hong, “Latency of Cellular-Based V2X: Perspectives on TTI-Proportional Latency and TTIIndependent Latency,” IEEE Access, vol. 5, pp. 15 800–15 809, July 2017.

[5] ”What is LoRa? | Semtech LoRa Technology | Semtech”. www.semtech.com. Retrieved 2021-01-21.

[6] Jump up to:a b"LoRa Modulation Basics" (PDF). Semtech. Archived from the original (PDF) on 2019-07-18. Retrieved 2020-02-05.

[7] ”Semtech Acquires Wireless Long Range IP Provider Cycleo". Design And Reuse. Retrieved 2019-10-17.

[8] Ramon Sanchez-Iborra; Jesus Sanchez-Gomez; Juan Ballesta-Viñas; Maria-Dolores Cano; Antonio F. Skarmeta (2018). "Performance Evaluation of LoRa Considering Scenario Conditions". Sensors. 18 (3): 772. doi:10.3390/s18030772. PMC 5876541. PMID 29510524

[9] Adelantado, Ferran; Vilajosana, Xavier; Tuset-Peiro, Pere; Martinez, Borja; Melia-Segui, Joan; Watteyne, Thomas (2017). "Understanding the Limits of LoRaWAN". IEEE Communications Magazine. 55 (9): 34–

12. CAPTURING PUBLICATIONS

- **UGC APPROVED JOURNAL**
- **International Conference on "Innovations in Computers Networks, Computational Intelligence and lOT" (ICICCI-21) (Online Mode)**

13. STUDENTS PROFILE

Kodali Sai Sumanth Chowdary is currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. He completed his intermediate from Sri Gayatri Junior College and 10th class from St. Mary's High School. He is one of the members of Coders Club in the college. His responsibilities in that group include mentoring and motivating students to take coding as a serious hobby. His technical skills include Java, ReactJS, Python, Django. He took part in Employability Skill development Program conducted by Zensar. He is also a student of Smart Interviews. He has participated in Smart India Hackathon, Project expo conducted by the college and has competed in various programming contests. His participations also include: National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020, "Know More - Teach More ", the Global Webinar on Cloud & Big Data – Changing the way we work which was conducted Global Education & Careers Forum (GECF) on 12th August 2020, a two day summit called e-summit hosted by Marri Laxman Reddy Institute Of Technology.He has completed internships at TDG Labs as a ReactJS Developer and at Apty Inc as a software development engineer. His areas of interest are Artificial Intelligence, Machine Learning and Deep Learning. He completed few certification courses from online platforms like Coursera, Solo Learn.

Kanugu Bhaskar is currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. He completed his intermediate from Sri Chaitanya College and 10th class from Sri Chaitanya School. His technical skills include C, Python and Java. He also has a basic understanding of C++. He took part in Employability Skill Development Program conducted by Zensar. He is also a student of Smart Interviews. His participations include: National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020, "Know More - Teach More ", the Global Webinar on Cloud & Big Data – Changing the way we work which was conducted Global Education & Careers Forum (GECF) on 12th August 2020, April to 22nd May 2020. His areas of interest are Python, Artificial Intelligence, Machine Learning and Deep Learning. He completed few certification courses from online platforms like Coursera.

Gaurav Karwa is currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. He completed his intermediate from Sri Chaitanya College and 10th class from Sister Nivedita School. His technical skills include Python , Java, React ,Nodejs, javascript. He also has a basic understanding of C. He took part in Employability Skill Development Program conducted by Zensar. He is also a participant of Smart Interviews. He has participated in Smart India Hackathon, Project expo conducted by the college and has competed in various programming contests. His participations include: National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020, "Know More - Teach More ", the Global Webinar on Cloud & Big Data – Changing the way we work which was conducted Global Education & Careers Forum (GECF) on 12th August 2020, April to 22nd May 2020. His areas of interest are Python, Artificial Intelligence, Machine Learning and Deep Learning. He completed few certification courses from online platforms like Coursera. His areas of interest are Artificial Intelligence, Machine Learning and Deep Learning. He completed few certification courses from online platforms like Edx,Courseera etc. He is currently hired by Brainscale to work as cloud developer in their company.

Nandyala Vedha Kalyan is currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. He completed his intermediate from Narayana Junior College and 10th class from Nalanda Vidya Bhavan High School. His technical skills include C, C++, Python and Java. He also has a basic understanding MYSQL. He took part in Employability Skill Development Program conducted by Zensar. He is also a student of Smart Interviews. His participations include: Attended the India First Leadership Talk webinar conducted by MHRD's Innovation Cell on 9th May 2020 , Workshop on HTML/CSS conducted by TAM from 5th Jan 2018 to 3rd Feb 2018. He completed few certification courses from online platforms like Coursera.

14. APPENDICES

An appendix contains supplementary material that is not an essential part of the text itself but which may be helpful in providing a more comprehensive understanding of the research problem and/or is information which is too cumbersome to be included in the body of the paper. A separate appendix should be used for each distinct topic or set of data and always have a title descriptive of its contents

APPENDIX A – DIGITAL HEALTH AGENCIES/ EHRS IN EACH PROVINCE/TERRITORY:

Web-based Testing Tools

Tools may be used during development as well as after deployment. Since standards and best practices change over time, it is a good idea to periodically re-test deployed applications.

Tool	Notes	Link(s)
Securityheaders.io	Website-based testing tool that checks a deployed website for HTTP header security best practices and provides a list of recommended standards and best practices.	https://securityheaders.io/
OWASP Security Headers Project	Explanations and implementation guidance for HTTP security headers.	https://www.owasp.org/index.php/OWASP_Secure_Headers_Project

Tool	Notes	Link(s)
SSL Labs Server Test	Website-based testing tool that checks if deployed website uses proper SSL/TLS certificates and configuration. The tool provides a grade for the website and appropriate recommendations if any tests are failed.	https://www.ssllabs.com/ssltest/
Badssl.com	Website-based testing tool for ensuring clients (not servers) are properly configured for using SSL/TLS. May not be necessary depending on the service.	https://badssl.com/
OWASP WAP	Tool for detecting vulnerabilities in PHP web	https://www.owasp.org/index.php/OWASP_WAP-Web_Application_Protection

Vulnerability Remediation Resources

Once the web application is deployed, the application, server, or other components may need updates, patches, or other changes to address vulnerabilities as they are discovered. In some cases, the system owner or their developer will need to develop the mitigation or remediation.

Resource	Notes	Link(s)
How to Win Friends and Remediate Vulnerabilities	Whitepaper from SANS Institute with advice on setting up a remediation capability.	https://www.sans.org/reading-room/whitepapers/application/win-friends-remediate-vulnerabilities-34530
Guide to Enterprise Patch Management Technologies	NIST report on managing patches for vulnerability remediation	http://dx.doi.org/10.6028/NIST.SP.800-40r3
SQL Injection Prevention	Advice on avoiding and fixing SQL injection vulnerabilities from OWASP	https://www.owasp.org/index.php/SQL_Injection_Prevention_Cheat_Sheet
Cross-Site Scripting	Advice on preventing and fixing cross-site scripting	https://www.google.com/about/appsecurity/learning/xss/#PreventingXSS

Resource	Notes	Link(s)
	(XSS) vulnerabilities from Google.	
Common Weakness Enumeration (CWE)	CWE provides a taxonomy of different types of vulnerabilities. Many CWE entries provide brief advice on potential mitigations.	https://cwe.mitre.org/
CWE/SANS Top 25 Most Dangerous Software Errors	List of dangerous vulnerabilities, the *Insecure Interaction Between Components* section.	https://www.sans.org/top25-software-errors/

A

PROJECT REPORT

On

Hand Gesture Recognition using Convolution Neural Network

Submitted by

1)Mr.A.Rajeshwar (17K81A05C2) 2)Mr.K.Santhosh Reddy(17K81A05F2)

3)Mr.K. Bharatvamsi (17K81A05C9) 4)Ms.P.Shivani(17K81A05G7)

in partial fulfillment for the award of

the degree of

BACHELOR OF TECHNOLOGY

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Under the Guidance of
Dr. M. Narayanan
Professor & HOD (CSE)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ST.MARTIN'S ENGINEERING COLLEGE
An Autonomous Institute
Dhulapally, Secunderabad – 500 100

JUNE 2021

BONAFIDE CERTIFICATE

This is to certify that the project entitled Hand Gesture Recognition using Convolution Neural Network, is being submitted by 1)**Mr.A. Rajeshwar (17K81A05C2), 2)Mr. K. Santhosh Reddy(17K81A05F2), 3) Mr.K. Bharatvamsi (17K81A05C9), 4) Ms. P.Shivani (17K81A05G7)** in partial fulfillment of the requirement for the award of the degree of **BACHELOR OF TECHNOLOGY IN Computer Science** is recorded of bonafide work carried out by them. The result embodied in this report have been verified and found satisfactory.

Dr. M.NARAYANAN
Professor & HOD
Department of CSE

Head of the Department
Dr. M. NARAYANAN
Department of CSE

Internal Examiner

External Examiner

Place:

Date:

DECLARATION

We, the student of **Bachelor of Technology** in Department of Computer Science and Engineering', session: 2017 – 2021, St. Martin's Engineering College, Dhulapally, Kompally, Secunderabad, hereby declare that work presented in this Project Work entitled Hand Gesture Recognition using Convolution Neural Network is the outcome of our own bonafide work and is correct to the best of our knowledge and this work has been undertaken taking care of Engineering Ethics. This result embodied in this project report has not been submitted in any university for award of any degree.

A.Rajeshwar (17K81A05C2)

K.Santhosh Reddy (17K81A05F2)

K.Bharatvamsi (17K81A05C9)

P.Shivani (17K81A05G7)

ABSTRACT

Hand Gesture Recognition (HGR) targets on interpreting the sign language into text or speech, so as to facilitate the communication between deaf-mute people and ordinary people. This task has broad social impact, but is still very challenging due to the complexity and large variations in hand actions. Existing methods for HGR use hand-crafted features to describe sign language motion and build classification models based on those features. However, it is difficult to design reliable features to adapt to the large variations of hand gestures. To approach this problem, we propose a novel convolutional neural network (CNN) which extracts discriminative spatial temporal features from raw video stream automatically without any prior knowledge, avoiding designing features. To boost the performance, multi-channels of video streams, including color information, depth clue, and body joint positions, are used as input to the CNN in order to integrate color, depth and trajectory information. We validate the proposed model on a real dataset collected with Microsoft Kinect and demonstrate its effectiveness over the traditional approaches based on hand-crafted features.

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible and whose encouragements and guidance have crowded effects with success.

We extended our deep sense of gratitude to Principal**,** **Dr. P. SANTOSH KUMAR PATRA**, St. Martin's Engineering College, Dhulapally, for permitting us to undertake this project.

We are also thankful to **Dr. M.NARAYANAN**, Head of the Department, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

We would like to thank **Dr. T. POONGOTHAI,** Department of Computer Science and Engineering (AI & ML) for her encouragement and insightful comments and as well as our project coordinators **Dr. B.RAJALINGAM**, Associate Professor and **Mr. J.SUDHAKAR**, Assistant Professor, in Department of Computer Science and Engineering for their valuable support.

We would like to express our sincere gratitude and indebtedness to our project supervisor **Dr. M.NARAYANAN**, Professor & HOD, Department of Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

Finally, we express thanks to all those who have helped us successfully to completing this project. Furthermore, we would like to thank our family and friends for their moral support and encouragement.

We express thanks to all those who have helped us in successfully completing the project.

A. Rajeshwar (17K81A05C2)

K. Santhosh Reddy (17K81A05F2)

K. Bharatvamsi(17K81A05C9)

P. Shivani (17K81A05G7)

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF OUTPUT SCREENS

LIST OF ACRONYMS

<AI>	Artificial Intelligence
<ANN>	Artificial Neural Network
<AVI>	Audio Video Interlace
<CPU>	Central Processing Unit
<CNN>	Convolutional Neural Network
<DCNN>	Deep Convolutional Neural Network
<DFD>	Data Flow Diagram
<GB>	Giga Bytes
<GMM>	Gaussian Mixture Model
<GUI>	Graphical User Interface
<HCI>	Human-Computer Interaction
<IR-UWB>	Impulsive-Radio Ultra-Wideband
<RAM>	Random Access Memory
<SSD>	Single-Shot Multi box Detector
<SVM>	Support Vector Machine
<UML>	Unified Modeling Language

1. INTRODUCTION

The communication or transfer of data in between human and human is really easy and understandable. But when it comes to human and machine it's really difficult because even machine knows all the languages humans can speak and understand, they cannot communicate with that knowledge or data. So for improving that communication features with machines we can develop some interactive techniques like gesture recognition. This is becoming a trending topic in the field of computer science and its applications related to deep learning technology.

Hand Gesture acknowledgment is essential for structuring no touch or control interfaces in vehicles. Such technologies enable drivers to drive while at that time connecting with different controls, e.g., sound and cooling, and also there are line enhance drivers' security and solace. In the recent decades, numerous vision-based powerful hand signal acknowledgment calculations were presented.

To perceive motions, distinctive highlights, for example, handmade spatiotemporal descriptors and enunciated models were utilized. As signal classifiers, concealed Markov models, contingent irregular fields and bolster vector machines (SVM) have been broadly utilized. Notwithstanding, vigorous order of signals under broadly fluctuating lighting conditions, and from various subjects is as yet a testing issue.

Computer-human communication refers to the way how the human communicate to the computer/machine, and since the machine is not useful until a human trains the machine for a particular task. There are mainly 2 characteristics that will be checked when developing a man-machine communication model as mentioned in: machine's performance and usage. The Model performance refers to how well the machines are performing to communicate with the human and usage refers to weather all the provided functionalities are performing according to the development.

Gestures can be in any form like hand image or pixel image or any human given pose that require less computational difficulty or power for making the devices required for the recognitions to make work. Different techniques are being proposed by the companies for gaining necessary information/data for recognition handmade gestures recognition models. Some models work with special devices such as data glove devices and color caps to develop a complex information about gesture provided by the user/human.

Recently, visual communications like augmented reality and virtual reality has attracted attention since many companies have introduced futuristic devices for the company's growth. For example, smart glasses for realistic features and many applications that have been developed in many companies. Virtual glass and augmented devices displays have also been introduced and are being used in futuristic

applications. Hand gesture recognition is widely used as an interface to give commands to these reality controlled devices.

1.1 PROJECT OVERVIEW

Hand gesture recognition from camera images is a topic for developing intelligent vision systems. We propose convolution neural network (CNN) method to recognize hand gestures of human task activities from a camera image.

To achieve the robustness performance, the skin model and the calibration of hand position and orientation are applied to obtain the training and testing data for the CNN.

The calibration of hand position and orientation aims at translating and rotating the hand image to a neutral pose.

1.2 PROJ ECT OBJECTIVES

The main objective of this project is to overcome drawbacks of existing system by using CNN algorithm in the proposed system to fetch effective results/outputs.

In order to overcome the drawback and fetch effective results we use convolution neural network (CNN)

Algorithm as this algorithm is the best on image recognition. The lighting conditions seriously effects the skin colour, in the existing system this criteria hasn’t been taken care of but in the proposed system to overcome the serious lighting conditions we use GMM.

Since the light condition seriously affects the skin colour, we adopt a Gaussian Mixture model (GMM) to train the skin model which is used to robustly filter out non-skin colours of an image.

1.3 ORGANIZATION OF CHAPTERS

Besides the introduction, the thesis is organized in other six chapters as follows:

Chapter 2, LITERATURE SURVEY: the review is made in the context of hand gesture recognition systems with a particular attention on those implementations that assess the scalability and performances or their implementations. Most of the related work is on convolution neural network, whereas a small part is on cloud solutions. It will be possible to notice that only a small subset of the literature actually focuses on the analysis of the systems in mass crises scenarios. Chapter 3, SOFTWARE AND HARDWARE REQUIREMENTS: this chapter discuss about the software and hardware required for the execution of the project. Chapter 4, SOFTWARE DEVELOPMENT ANALYASIS: this chapter explains the assumptions and technical specifications of the project. Chapter 5, PROJECT SYSTEM DESIGN: this chapter explains all the software development process with DFD and UML diagrams clearly. Chapter 6, PROJECT CODING: this chapter explains the design of the system, roles and responsibilities, as well as the requirements of a HGR management solution based on CNN. Chapter 7, PROJECT TESTING: this chapter explains various test cases to test the project working. Chapter 8, OUTPUT SCREENS: explains a step by step process of the project execution. Chapter 9, EXPERIMENTAL RESULTS: tests and results are shown and explained in this chapter. The results are analyzed in the context of the thesis project and followed by discussion on systems throughput and resiliency, as well as the approaches to testing and analysis. Chapter 10, CONCLUSION AND FUTURE ENHANCEMENT: the chapter ends the project with a short summary of the main concepts mentioned in the thesis as well as the relevant results.

2. LITERATURE SURVEY

2.1 SURVEY ON BACKGROUND

Detailed analysis of back-propagation learning and multi-layer perceptions. Explores the intricacies of the learning procession essential component for understanding neural networks. Considers recurrent networks, such as Hopfield [1] networks, Boltzmann machines, and mean field theory machines, as well as modular networks, temporal processing, and neuro dynamics. Integrates computer experiments throughout, giving the opportunity to see how neural networks are designed and perform in practice.

CNN model augmented by edit distance for the recognition of static and dynamic gestures of Pakistani sign language, and achieved 90.79% accuracy. Al- Hammadi et al. proposed a 3DCNN model to learn region-based spatiotemporal features for hand gestures the biggest challenge faced by the[2] researchers is designing a robust hand gesture recognition framework that overcomes the most typical problems with fewer limitations and gives an accurate and reliable result. Real-time processing of hand gestures also has some limitations, such as illumination variation, background problems, distance range, and multi-gesture problems

Human posture detection allows the capture of the kinematic parameters of the human body, which is [3] important for many applications, such as assisted living, healthcare, physical exercising and rehabilitation. We use five types of gestures, namely those for Stop, Forward, Backward, Turn Left, and Turn Right. Users will control devices through a camera connected to computers.

The algorithm will analyze gestures and take actions [4] to perform appropriate action according to user requests via their gestures. The results show that the average accuracy of proposal algorithm is 92.6 percent for images and more than 91 percent for video, which both satisfy performance requirements for real-world application, specifically for smart home services.

Control an industrial robot by means of [5] gestures and voice commands. It describes the elements of creating software for off-line and on-line robot control. The application for the Kinect module was developed in the C# language in the Visual Studio environment, while the industrial robot control program was developed in the RAPID language in the Robot Studio environment

Voice pathology disorders can be effectively [6] detected using computer-aided voice pathology classification tools. These tools can diagnose voice pathologies at an early stage and offering appropriate treatment. This study aims to develop a powerful feature extraction voice pathology detection tool based

on Deep Learning. In this paper, a pre-trained Convolutional Neural Network (CNN) was applied to a dataset of voice pathology to maximize the classification accuracy.

This study also proposes a distinguished training method combined with various training strategies in order to generalize the [7] application of the proposed system on a wide range of problems related to voice disorders. In the recent decade, a wide range of tools and methodologies have been introduced to support the design of for various domains, therefore it can be a challenging task to choose the most appropriate technique for the design process. Our research aims to present a classification to guide the identification of the most relevant and appropriate methodologies in the given scenario.

Neural [8] networks and evolutionary computing to solve real-world problems that cannot be satisfactorily solved using conventional crisp computing techniques. Representation and processing of human knowledge, qualitative and approximate reasoning, computational intelligence, computing with words, and biological models of problem solving and optimization form key characteristics of soft computing, and are directly related to intelligent systems and applications. Human Computer Interaction

(HCI) technologies are rapidly evolving the way we interact with computing devices [9] and adapting to the constantly increasing demands of modern paradigms. One of the most useful tools in this regard is the integration of Human-to-Human Interaction gestures to facilitate communication and expressing ideas. Gesture recognition requires the integration of postures, gestures, face expressions and movements for communicating or conveying certain messages.

This paper proposes a novel approach [10] to identify hand gestures in complex scenes by the Single-Shot Multi box Detector (SSD) deep learning algorithm with 19 layers of a neural network. A benchmark database with gestures is used, and general hand gestures in the complex scene are chosen as the processing objects. A real-time hand gesture recognition system based on the SSD algorithm is constructed and tested.

One of the most promising fields where big data can be applied to make a change is healthcare. Big healthcare [11-12] data has considerable potential to improve patient outcomes, predict outbreaks of epidemics, gain valuable insights, avoid preventable diseases, reduce the cost of healthcare delivery and improve the quality of life in general. However, deciding on the allowable uses of data while preserving security and patient’s right to privacy is a difficult task.

Big [13] data, no matter how useful for the advancement of medical science and vital to the success of all health care organizations, can only be used if security and privacy issues are addressed

Given that the existing HCI approaches exhibit various limitations, hand gesture recognition-based HCI may serve as a more natural mode [14] of man–machine interaction in many situations. predict outbreaks of epidemics, gain valuable insights, avoid preventable diseases, reduce the cost of healthcare delivery and improve the quality of life in general. However, deciding on the allowable uses of data while preserving security and patient's right to privacy is a difficult task Inspired by an inception module-based deep-learning network (GoogLeNet), this paper presents a novel hand gesture recognition technique for impulse-radio ultra-wideband (IR-UWB) radars which demonstrates a higher gesture recognition accuracy.

First, methodology to demonstrate radar signals as three-dimensional image patterns is presented and then, the inception module-based variant of GoogLeNet is used to analyze the pattern within the images for the [15] recognition of different hand gestures. The proposed framework is exploited for eight different hand gestures with a promising classification accuracy of 95%. To verify the robustness of the proposed algorithm, multiple human subjects were involved in data acquisition.

2.2 CONCLUSIONS ON SURVEY

This work is a CNN-based human hand gesture recognition system. CNN is a research branch of neural networks. Using a CNN to learn human gestures, there is no need to develop complicated algorithms to extract image features and learn them. Through the convolution and sub-sampling layers of a CNN, invariant features are allowed with little dislocation. To reduce the effect of various hand poses of a hand gesture type on the recognition accuracies, the principal axis of the hand is found to calibrate the image in this work. Calibrated images are advantageous to a CNN to learn and recognize correctly

3. SOFTWARE AND HARDWARE REQUIREMENTS

3.1 SOFTWARE REQIREMENTS

For developing the application the following are the Software Requirements:

- Python 3.4
- Django3.2.4
- Mysql 8.0.23
- Wampserver 3.2.0

3.2 HARDWARE REQIREMENTS

- System - Intel(R) Core(TM)
- Speed - 2.4GHz
- RAM - 4GB
- Hard Disk - 1TB
- Processor - i3
- Storage Type - 128GB SSD

4. SOFTWARE DEVELOPMENT ANALYASIS

4.1 OVERVIEW OF PROBLEM

Hand gesture recognition has been a rounded for a quiet period of time and is a common research subject.

In this project we use convolution neural network to solve the problem of achieving effective results of outputs.

We have use convolutional Neural Network because it is a type of artificial neural network used in image recognition. CNN are powerful image processing artificial intelligence (AI).

4.2 DEFINE THE PROBLEM

In the existing system, the data sets are very minimum and the hand gestures in the data sets are very less so the result obtained is not effective so, we used Convolution Neural Network (CNN) to overcome these problems**.**

By using CNN, effective results are obtained and as we increase our knowledge in sign language, we will add new hand gesture images in the data sets to that we can increase the span. In this project we give a image captured in computer and give the output in the text format**.**

4.3 MODULES OVERVIEW

1. TensorFlow
2. Pandas
3. Scikit – learn
4. Matplotlib
5. Numpy

4.4 DEFINE THE MODULES

Tensorflow:

- TensorFlow is a free and open-source software library for dataflow and differentiable programming across a range of tasks.
- It is a symbolic math library, and is also used for machine learning applications such as neural networks.
- It is used for both research and production at Google.
- TensorFlow was developed by the Google Brain team for internal Google use.
- It was released under the Apache 2.0 open-source license on November 9, 2015.
- Its particular focus is on training & interference of deep neural network.
- It is written in Python, C++.
- It works on Linux, macOs, Windows, Android, Javascript.

Pandas:

- Pandas is an open-source Python Library providing high-performance data manipulation and analysis tool using its powerful data structures.
- Python was majorly used for data munging and preparation.
- It had very little contribution towards data analysis. Pandas solved this problem.
- Using Pandas, we can accomplish five typical steps in the processing and analysis of data, regardless of the origin of data load, prepare, manipulate, model, and analyze.
- Python with Pandas is used in a wide range of fields including academic and commercial domains including
- finance, economics, Statistics, analytics, etc
- In particular, it offers data structures and operations for manipulating numeric tables and time series

- The name is derived from the term “panel data”, an econometrics term for data sets. It’s name is play on the phase “python data analysis” itself.
- It is written in Python, C.
- It is initially released on 11th January 2008.

Scikit – learn:

- Scikit-learn provides a range of supervised and unsupervised learning algorithms via a consistent interface in Python.
- It is licensed under a permissive simplified BSD license and is distributed under many Linux distributions, encouraging academic and commercial use.
- It is free software machine learning library for the python programming language.
- It is initially released in June 2007.

Matplotlib:-

- It is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms.
- Matplotlib can be used in Python scripts, the Python and IPython shells, the Jupyter Notebook, web application servers, and four graphical user interface toolkits.
- Matplotlib tries to make easy things easy and hard things possible.
- You can generate plots, histograms, power spectra, bar charts, error charts, scatter plots, etc., with just a few lines of code. For examples, see the sample plots and thumbnail gallery.
- For simple plotting the pyplot module provides a MATLAB-like interface, particularly when combined with IPython.
- For the power user, you have full control of line styles, font properties, axes properties, etc, via an object oriented interface or via a set of functions familiar to MATLAB users

- It is written in Python
- It is released in 2003
- It was originally Written by John D.Hunder

Numpy:

- Numpy is a general-purpose array-processing package.
- It provides a high performance multidimensional array object, and tools for working with these arrays.
- It is the fundamental package for scientific computing with Python. It contains various features including these important ones:

  A powerful N-dimensional array object

  Sophisticated (broadcasting) functions

  Tools for integrating C/C++ and Fortran code

  Useful linear algebra, Fourier transform, and random number capabilities
- Besides its obvious scientific uses, Numpy can also be used as an efficient multi-dimensional container of generic data.
- It is Written in Python & C.
- The ancestor of Numpy, Numeric was originally created by Jim Hugunin.
- It is initially released as Numeric in 1995, as Numpy in 2006

4.5 MODULE FUNCTIONALITY

1. Upload Hand Gesture Dataset – This is the first module of our project where we need to upload a folder that contains different images containing different hand postures. Which have been captured previously and will be used as references in this project. This folder which contains these images of hand gesture postures are uploaded first.

2. Train CNN –Convolution Neural Network (CNN) is an artificial neural network used in image recognition. It is the best algorithm which can be used in image recognition. In our project we first train the CNN on every different hand gesture individually before we further proceed in this project. This helps the algorithm to work more efficiently and produce effective results/outputs.

3. Classify Flower-CNN Classifies the images based on different hand posture. We all know that CNN is used in image recognition. In this project we use various hand gesture posture images. In this project, CNN is used to classify the images of every hand gesture posture individually.

4. Webcam Predict – In this project we take the captured image of the hand gesture from the web cam. The captured image from the webcam and the hand gesture in the captured image will be predicted by using CNN and the data set folders which we have uploaded in the first step and then the message will be displayed as the output in the form of text form.

5. Close- Exit from the project.

5. PROJECT SYSTEM DESIGN

5.1 DFDS IN CASE OF DATABASE PROJECTS

A data flow diagram shows the way information flows through a process or system. It includes data inputs and outputs, data stores, and the various sub processes the data moves through. DFDs are built using standardized symbols and notation to describe various entities and their relationships.

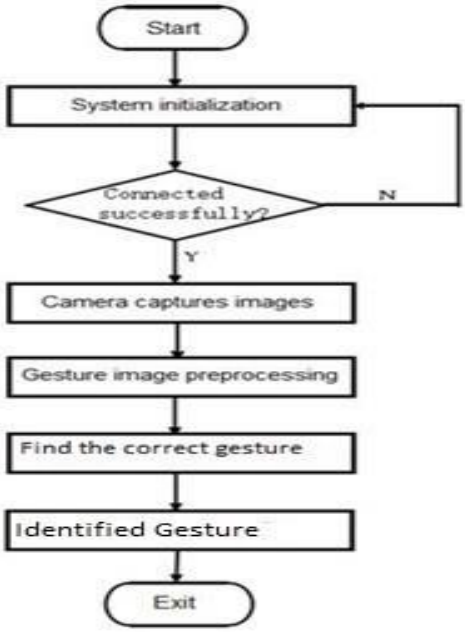


Fig 5.1: Data flow Diagram

Data flow diagrams visually represent systems and processes that would be hard to describe in a chunk of text. You can use these diagrams to map out an existing system and make it better or to plan out a new system for implementation. Visualizing each element makes it easy to identify inefficiencies and produce the best possible system.

5.2 UML DIAGRAMS

UML stands for Unified Modeling Language. UML is a standardized general-purpose modeling language in the field of object-oriented software engineering. The standard is managed, and was created by, the Object Management Group.

The goal is for UML to become a common language for creating models of object oriented computer software. In its current form UML is comprised of two major components: a Meta-model and a notation. In the future, some form of method or process may also be added to; or associated with, UML.

The Unified Modelling Language is a standard language for specifying, Visualization, Constructing and documenting the artifacts of software system, as well as for business modelling and other non-software systems.

The UML represents a collection of best engineering practices that have proven successful in the modelling of large and complex systems.

The UML is a very important part of developing objects oriented software and the software development process. The UML uses mostly graphical notations to express the design of software projects.

GOALS:

The Primary goals in the design of the UML are as follows:

1. Provide users a ready-to-use, expressive visual modeling Language so that they can develop and exchange meaningful models.
2. Provide extendibility and specialization mechanisms to extend the core concepts.
3. Be independent of particular programming languages and development process.
4. Provide a formal basis for understanding the modeling language.
5. Encourage the growth of OO tools market.
6. Support higher level development concepts such as collaborations, frameworks, patterns and components.
7. Integrate best practices.

USE CASE DIAGRAM

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.

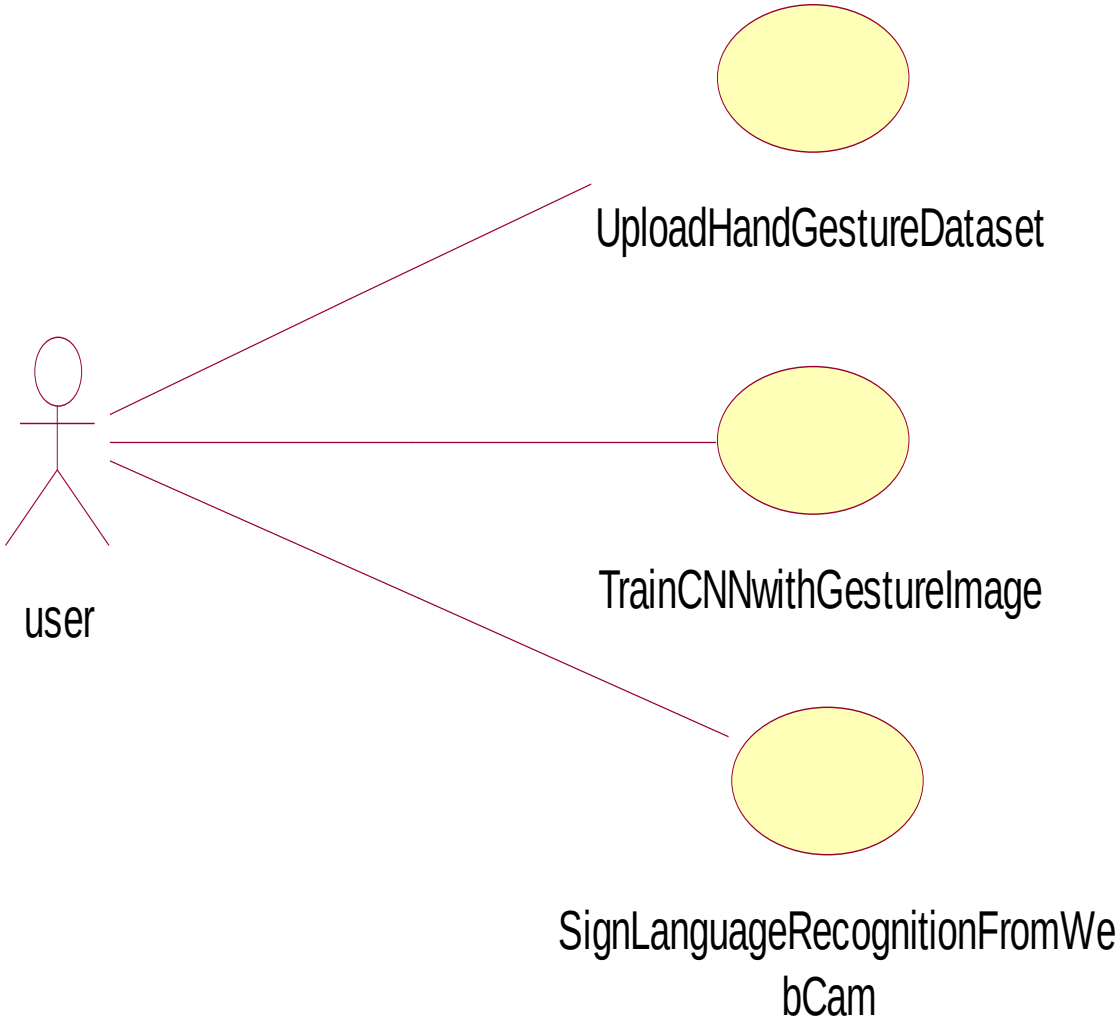


Fig 5.2: Use Case Diagram

CLASS DIAGRAM

In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.

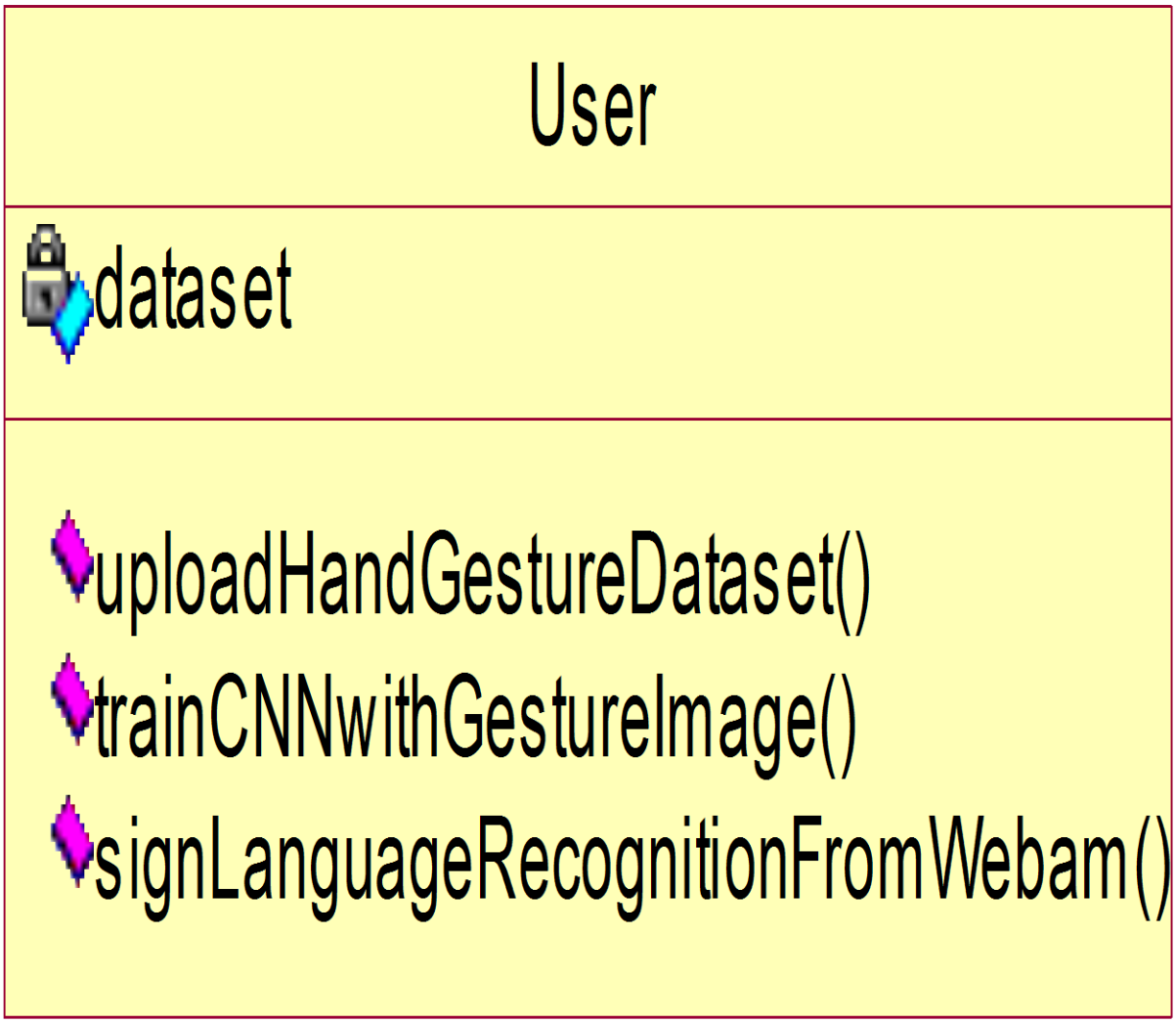


Fig 5.3: Class Diagram

SEQUENCE DIAGRAM

A sequence diagram in Unified Modelling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.

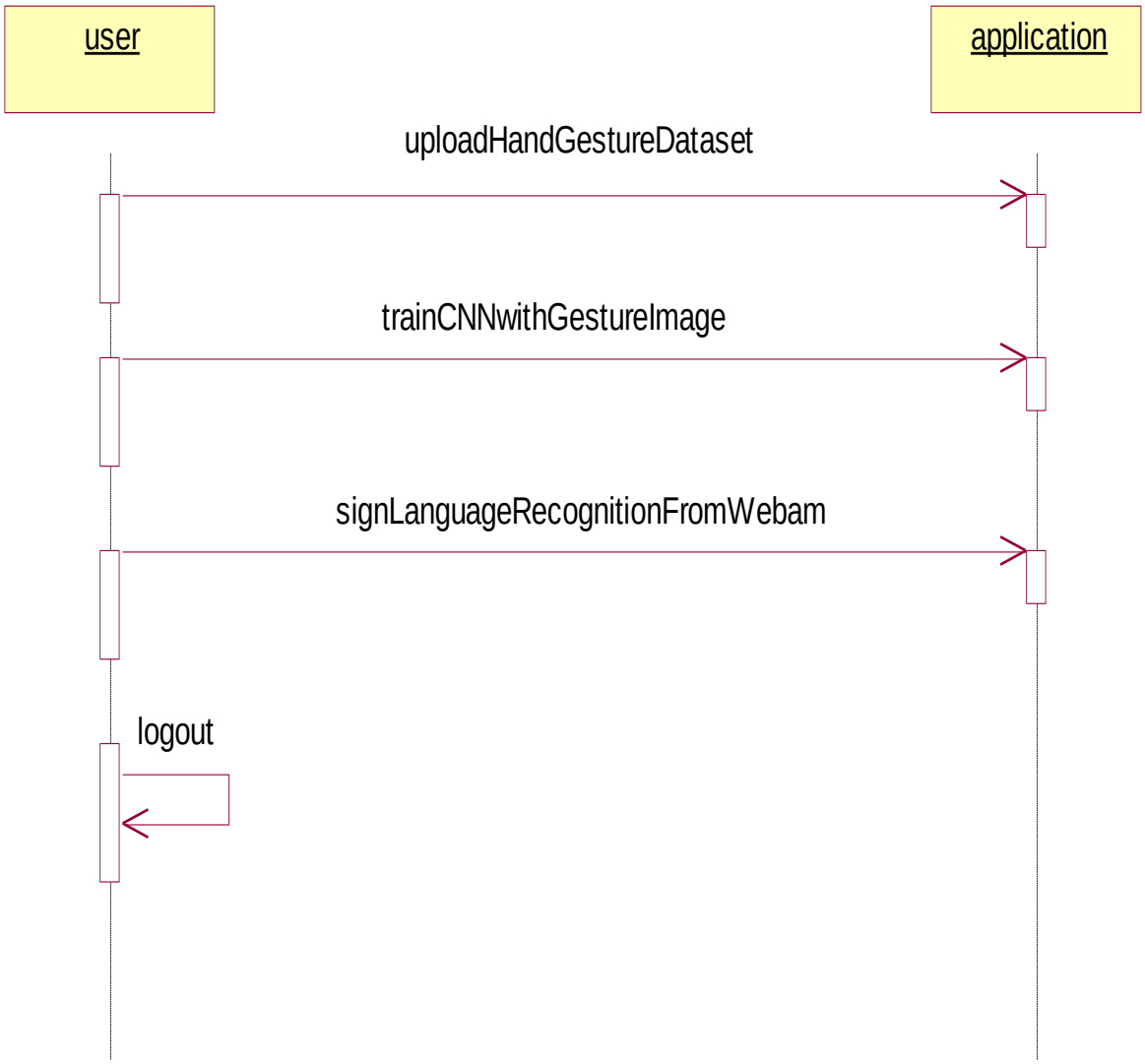


Fig 5.4: Sequence Diagram

COLLABORATION DIAGRAM

Collaboration diagrams (known as Communication Diagram in UML 2.x) are used to show how objects interact to perform the behavior of a particular use case, or a part of a use case. Along with sequence diagrams, collaboration are used by designers to define and clarify the roles of the objects that perform a particular flow of events of a use case. They are the primary source of information used to determining class responsibilities and interfaces.

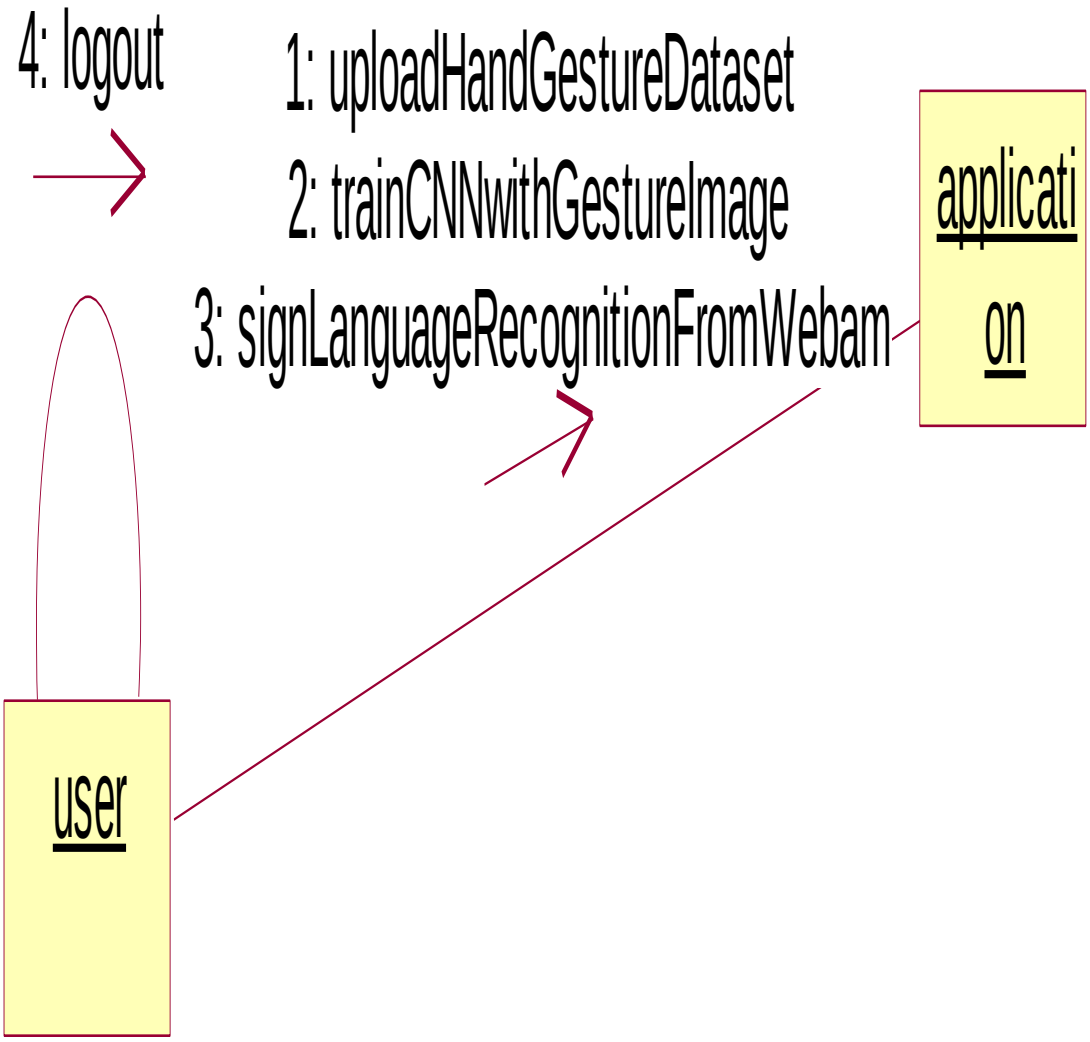


Fig 5.5: Collaboration Diagram

6. PROJECT CODING

6.1 CODE TEMPLATES

```
def trainCNN():
    global classifier
    text.delete('1.0', END)
    X_train = np.load('model/X.txt.npy')
    Y_train = np.load('model/Y.txt.npy')
    text.insert(END,"CNN is training on total images : "+str(len(X_train))+"\n")

    if os.path.exists('model/model.json'):
        with open('model/model.json', "r") as json_file:
            loaded_model_json = json_file.read()
            classifier = model_from_json(loaded_model_json)
        classifier.load_weights("model/model_weights.h5")
        classifier._make_predict_function()
        print(classifier.summary())
        f = open('model/history.pckl', 'rb')
        data = pickle.load(f)
        f.close()
        acc = data['accuracy']
        accuracy = acc[19] * 100
        text.insert(END,"CNN Hand Gesture Training Model Prediction Accuracy = "+str(accuracy))
```

```
else:
    classifier = Sequential()
    classifier.add(Convolution2D(32, 3, 3, input_shape = (64, 64, 3), activation = 'relu'))
    classifier.add(MaxPooling2D(pool_size = (2, 2)))
    classifier.add(Convolution2D(32, 3, 3, activation = 'relu'))
    classifier.add(MaxPooling2D(pool_size = (2, 2)))
    classifier.add(Flatten())
    classifier.add(Dense(output_dim = 256, activation = 'relu'))
    classifier.add(Dense(output_dim = 5, activation = 'softmax'))
    print(classifier.summary())
    classifier.compile(optimizer = 'adam', loss = 'categorical_crossentropy', metrics = ['accuracy'])
    hist = classifier.fit(X_train, Y_train, batch_size=16, epochs=10, shuffle=True, verbose=2)
    classifier.save_weights('model/model_weights.h5')
    model_json = classifier.to_json()
    with open("model/model.json", "w") as json_file:
        json_file.write(model_json)
    f = open('model/history.pckl', 'wb')
    pickle.dump(hist.history, f)
    f.close()
    f = open('model/history.pckl', 'rb')
    data = pickle.load(f)
    f.close()
    acc = data['accuracy']
    accuracy = acc[19] * 100
    text.insert(END,"CNN Hand Gesture Training Model Prediction Accuracy = "+str(accuracy))
```

6.2 OUTLINE FOR VARIOUS FILES

We used Python programming to implement our project. A single python file is used to implement our code. This file consists of various modules that we have used. Our project modules are - Upload Hand gesture Dataset, train CNN, classify Flower, webcam Predict. We also used various python modules like pandas, matplotlib, numpy, tensorflow, sklearn.

6.3 METHODS INPUT AND OUTPUT PARAMETERS

In our project code, we implemented six different methods. They are:

1. uploadDataset()

2. trainCNN():

3. classifyFlower():

4. webcamPredict():

5. close()

Our first method uploadDataset() doesn't take any input parameters but after successful execution, it displays a message "Hand Gesture dataset loaded". Second method train CNN() doesn't have any input parameters and after successful completion, it displays a message "CNN is training on total images : 2000".Third Method classify Flower() classifies the images based on different hand positions. After building the CNN algorithm, the accuracy of our project is displayed. Webcam predict() doesn't have any input parameters but upon successful completion, it displays the Hand Gesture name along with image. close() don't have any parameters but upon clicking this button, it will close the project window.

7. PROJECT TESTING

7.1 VARIOUS TEST CASES

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub-assemblies, assemblies and/or a finished product. It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

TYPES OF TESTS

Unit testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

Integration testing

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

Functional test

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items:

Valid Input : identified classes of valid input must be accepted.

Invalid Input : identified classes of invalid input must be rejected.

Functions : identified functions must be exercised.

Output : identified classes of application outputs must be exercised.

Systems/Procedures : interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

System Test

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

Unit Testing

Unit testing is usually conducted as part of a combined code and unit test phase of the software lifecycle, although it is not uncommon for coding and unit testing to be conducted as two distinct phases.

Test strategy and approach

Field testing will be performed manually and functional tests will be written in detail.

Test objectives

- All field entries must work properly.
- Pages must be activated from the identified link.
- The entry screen, messages and responses must not be delayed.

Features to be tested

- Verify that the entries are of the correct format
- No duplicate entries should be allowed
- All links should take the user to the correct page.

Integration Testing

Software integration testing is the incremental integration testing of two or more integrated software components on a single platform to produce failures caused by interface defects.

The task of the integration test is to check that components or software applications, e.g. components in a software system or – one step up – software applications at the company level – interact without error.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

Acceptance Testing

User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

7.2 BLACK BOX TESTING

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box .you cannot “see” into it. The test provides inputs and responds to outputs without considering how the software works.

The below Black-Box can be any software system you want to test. For Example, an operating system like Windows, a website like Google, a database like Oracle or even your own custom application. Under

Black Box Testing, you can test these applications by just focusing on the inputs and outputs without knowing their internal code implementation

Various approaches to black-box testing

There are a set of approaches for black-box testing.
Manual UI Testing: In this approach, a tester checks the system as a user. Check and verify the user data, error messages.

Automated UI Testing: In this approach, user interaction with the system is recorded to find errors and glitches. Testers can set record demand as per schedule.

Documentation Testing: In this approach, a tester purely checks the input and output of the software. Testers consider what system should perform rather than how. It is a manual approach to testing.

The tester doesn't need any technical knowledge to test the system. It is essential to understand the user's perspective.
Testing is performed after development, and both the activities are independent of each other.
It works for a more extensive coverage which is usually missed out by testers as they fail to see the bigger picture of the software.
Test cases can be generated before development and right after specification.
Black box testing methodology is close to agile.

7.3 WHITE BOX TESTING

The box testing approach of software testing consists of black box testing and white box testing. We are discussing here white box testing which also known as glass box is **testing, structural testing, clear box testing, open box testing and transparent box testing**.

It tests internal coding and infrastructure of a software focus on checking of predefined inputs against expected and desired outputs. It is based on inner workings of an application and revolves around internal structure testing. In this type of testing programming skills are required to design test cases. The primary goal of white box testing is to focus on the flow of inputs and outputs through the software and strengthening the security of the software.

The term 'white box' is used because of the internal perspective of the system. The clear box or white box or transparent box name denote the ability to see through the software's outer shell into its inner workings.

Developers do white box testing. In this, the developer will test every line of the code of the program. The developers perform the White-box testing and then send the application or the software to the testing team, where they will perform the black box testing and verify the application along with the requirements and identify the bugs and sends it to the developer.

The developer fixes the bugs and does one round of white box testing and sends it to the testing team. Here, fixing the bugs implies that the bug is deleted, and the particular feature is working fine on the application.

Here, the test engineers will not include in fixing the defects for the following reasons:

- Fixing the bug might interrupt the other features. Therefore, the test engineer should always find the bugs, and developers should still be doing the bug fixes.
- If the test engineers spend most of the time fixing the defects, then they may be unable to find the other bugs in the application.

8. OUTPUT SCREENS

8.1 USER INTERFACES

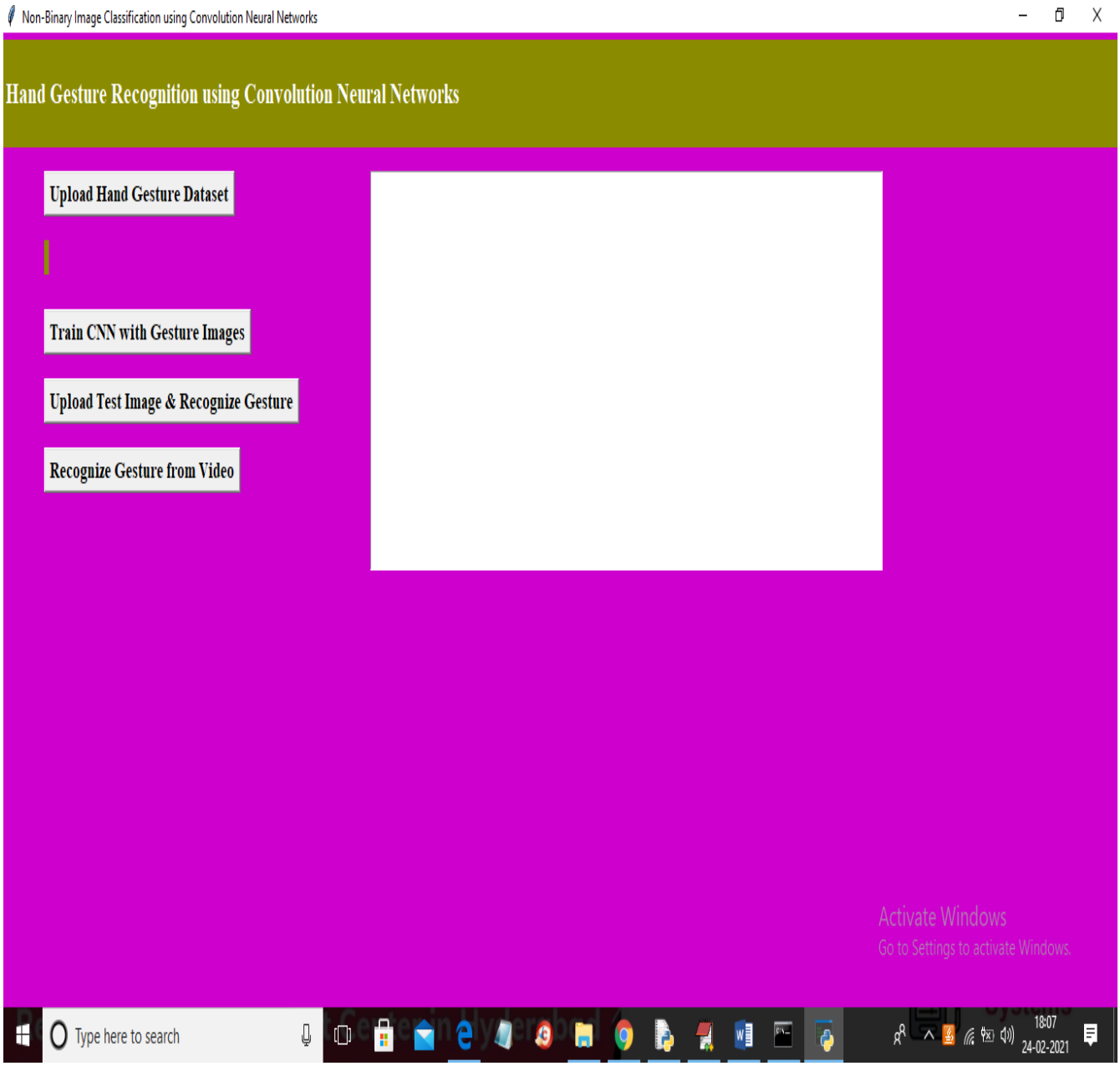


Fig 8.1 The figure shows user interface. The user interface has three modules which are Train CNN with gesture images, Upload test image and recognize gestures and recognize gesture from video

8.2 OUTPUT SCREENS

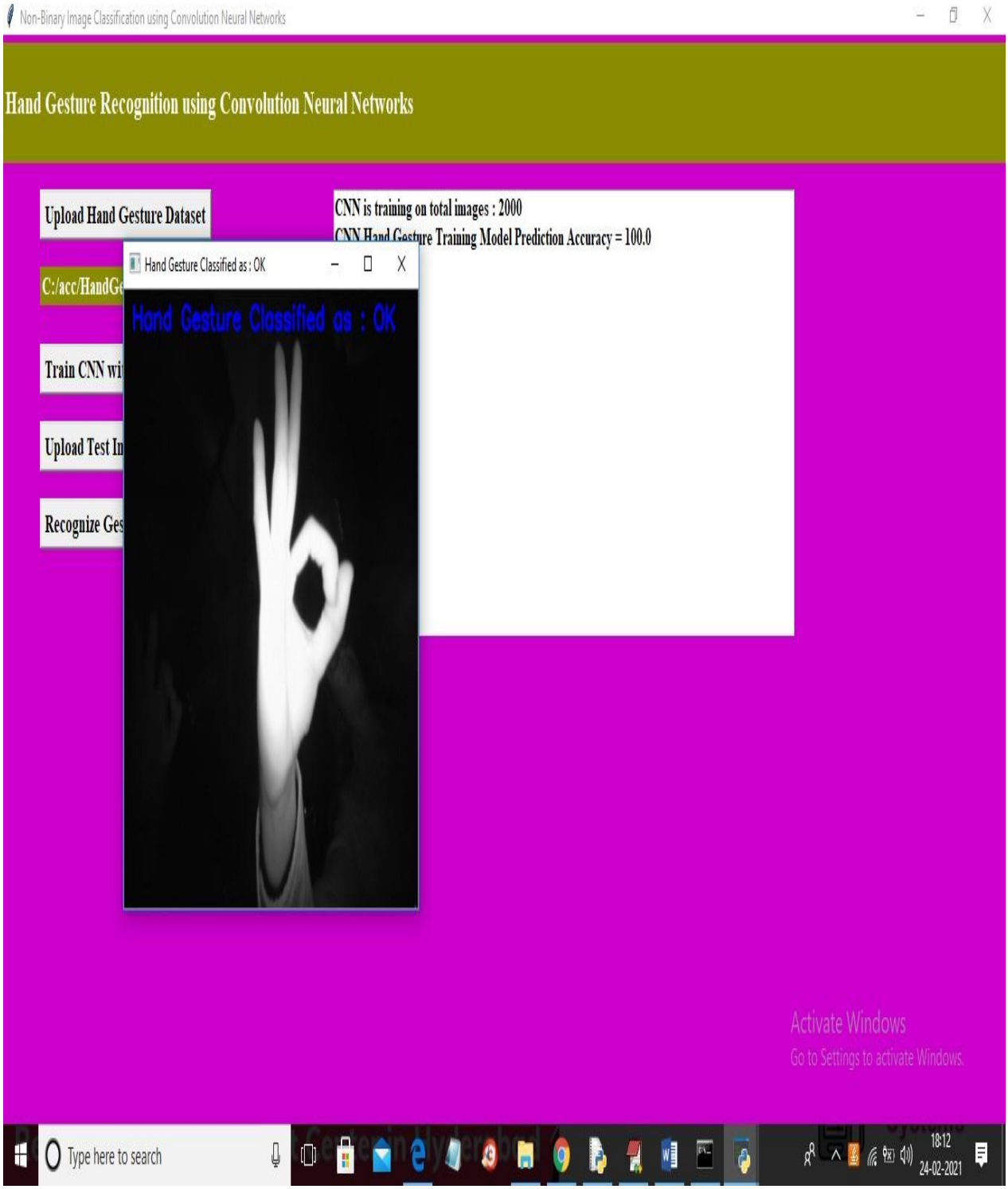


Fig 8.2 In above screen gesture recognised as OK.The output shows ok hand posture along with text – Hand Gesture classified as : OK.

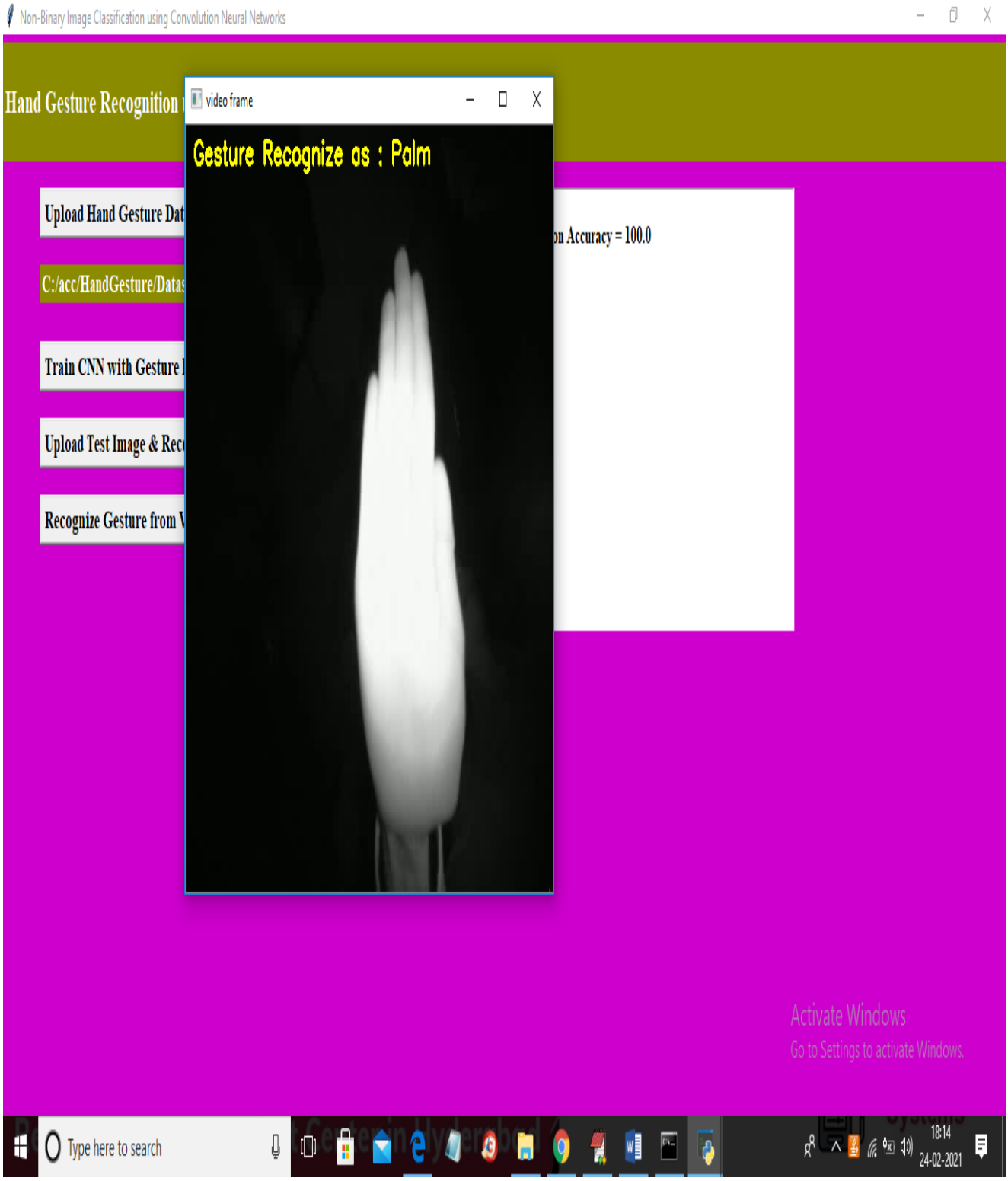


Fig 8.3 In above screen gesture recognised as Palm. The output shows palm hand posture along with text – Hand Gesture recognize as : Palm

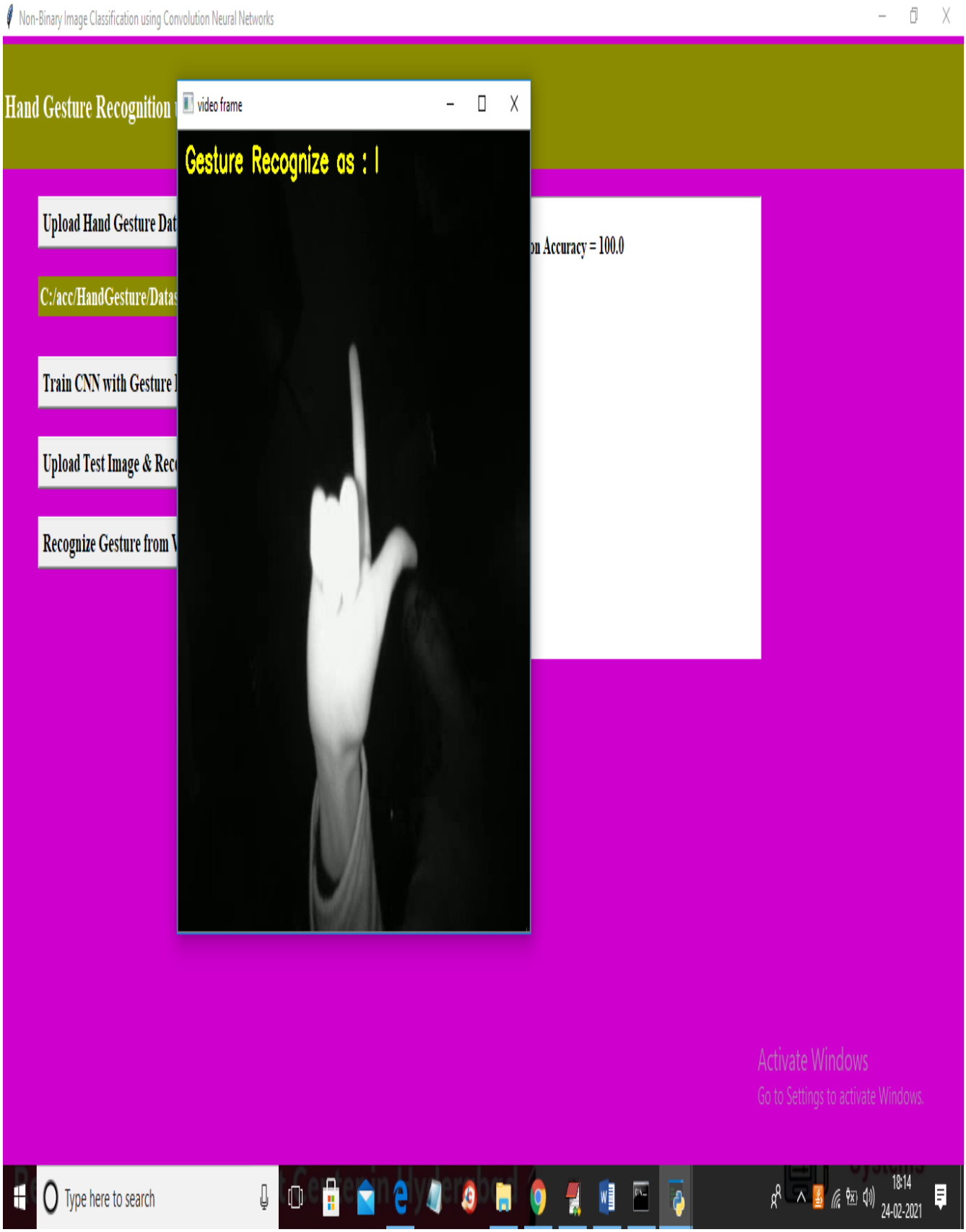


Fig 8.4 In above screen gesture recognised as I. The output shows I hand posture along with text – Hand Gesture recognize as : I.

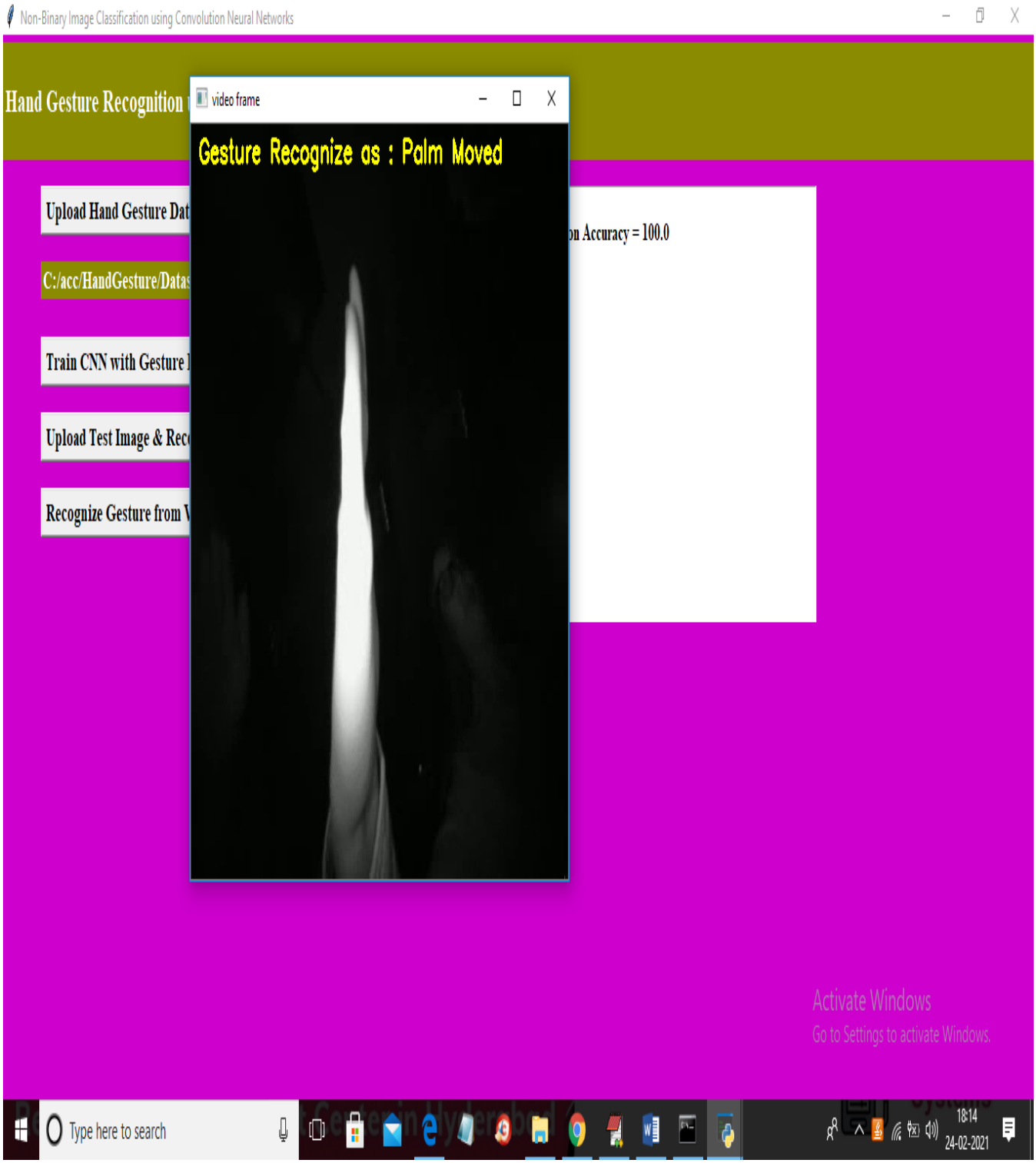


Fig 8.5 In above screen gesture recognised as palm moved. The output shows palm moved hand posture along with text – Hand Gesture recognize as : palm moved.

9. EXPERIMENTAL RESULTS

Homepage

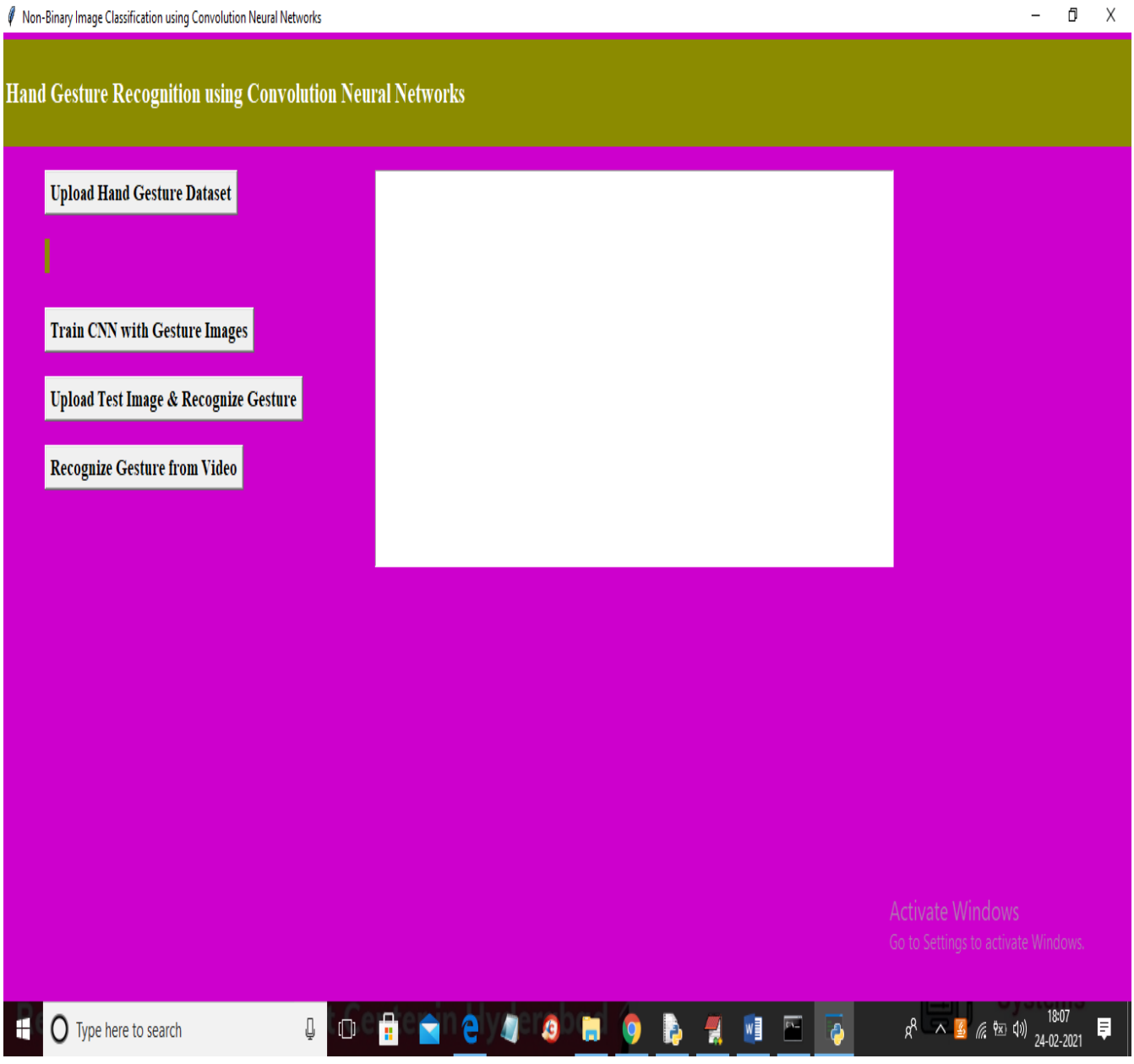


Fig 9.1: Home page

Uploading dataset:

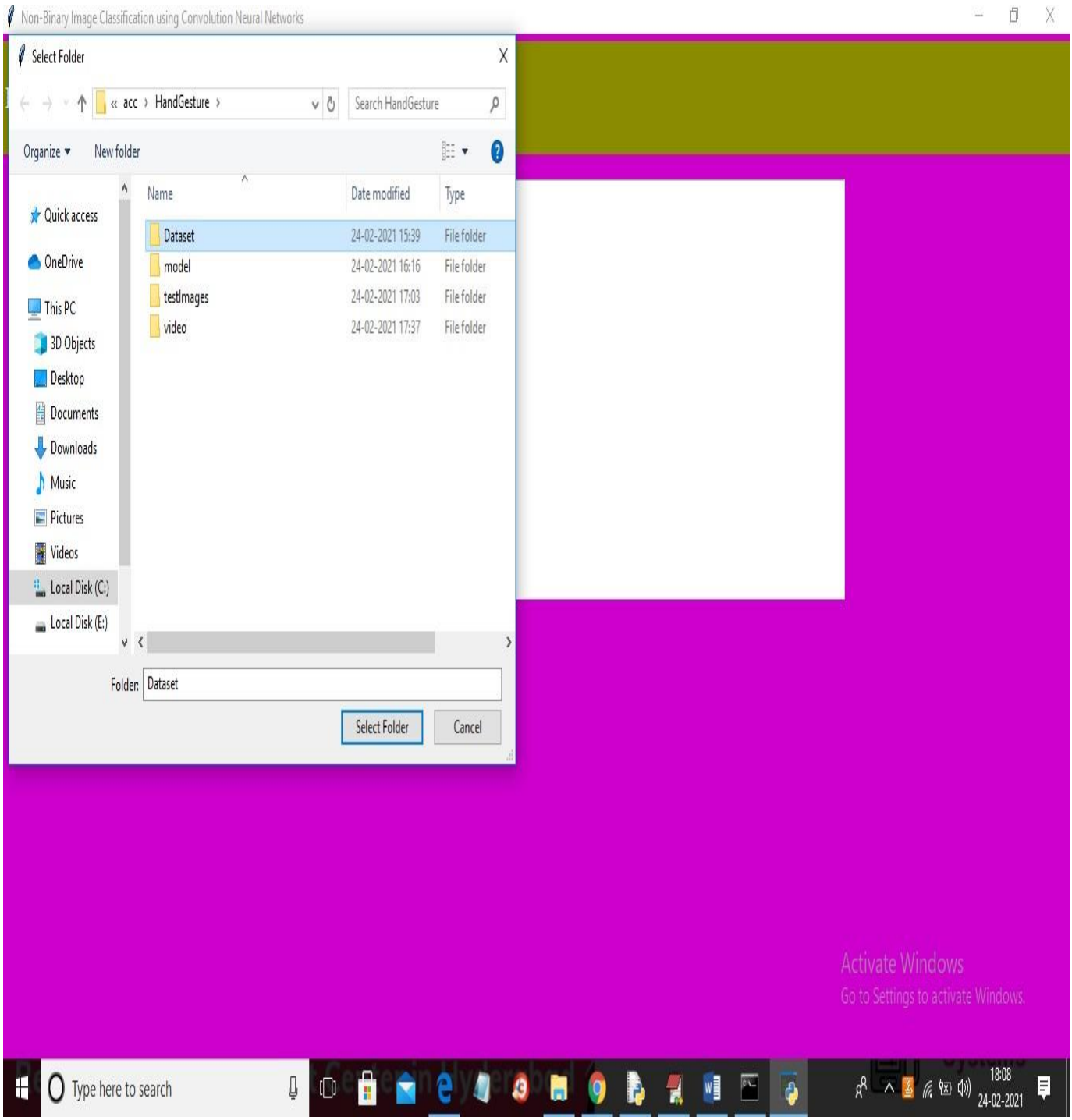


Fig 9.2

In above screen selecting and uploading ‘dataset’ folder and then click on ‘Select Folder’ button to load dataset

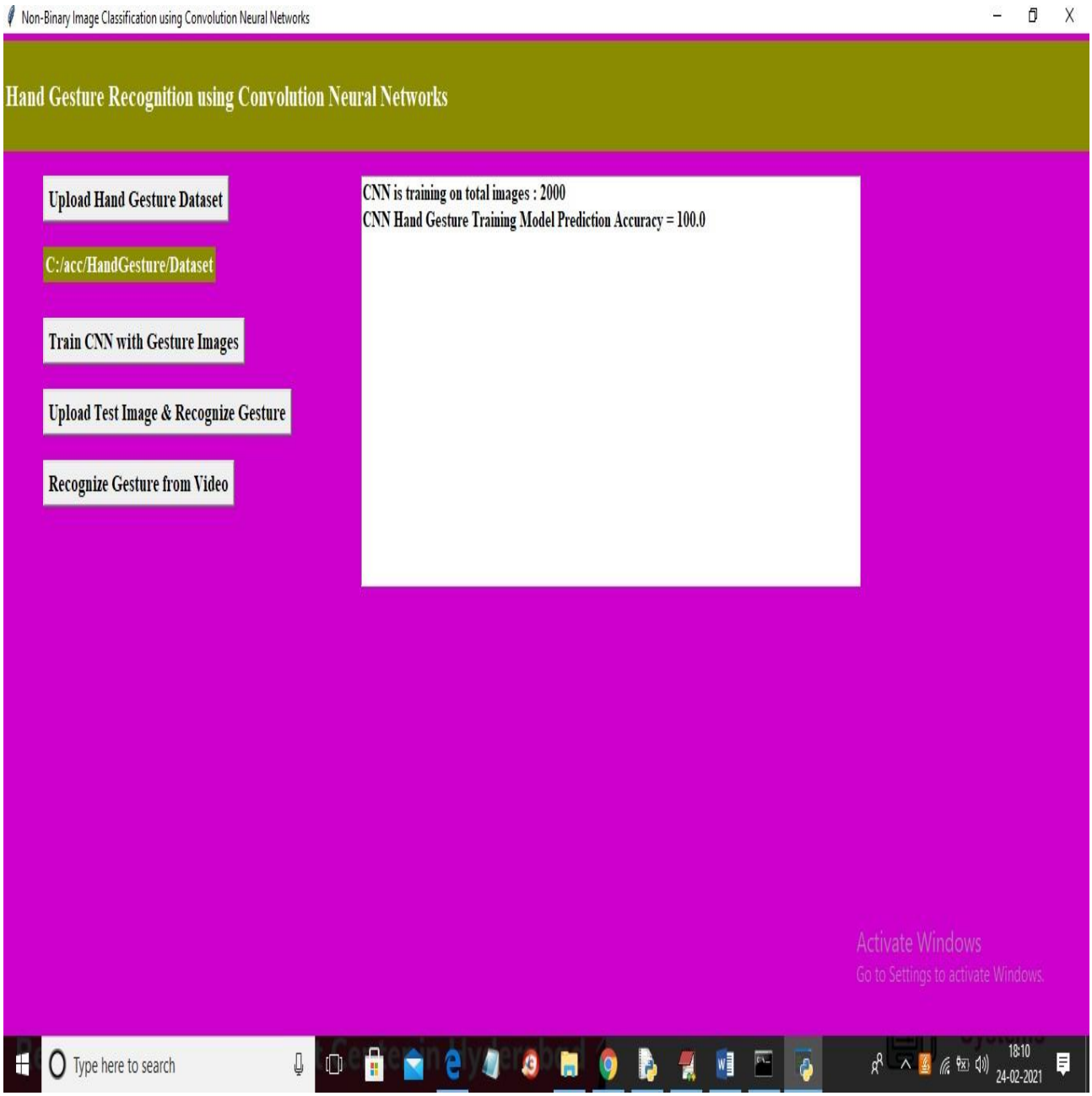


Fig 9.3

In above screen CNN model trained on 2000 images and its prediction accuracy we got as 100% now click on 'Upload Test Image & Recognize Gesture'

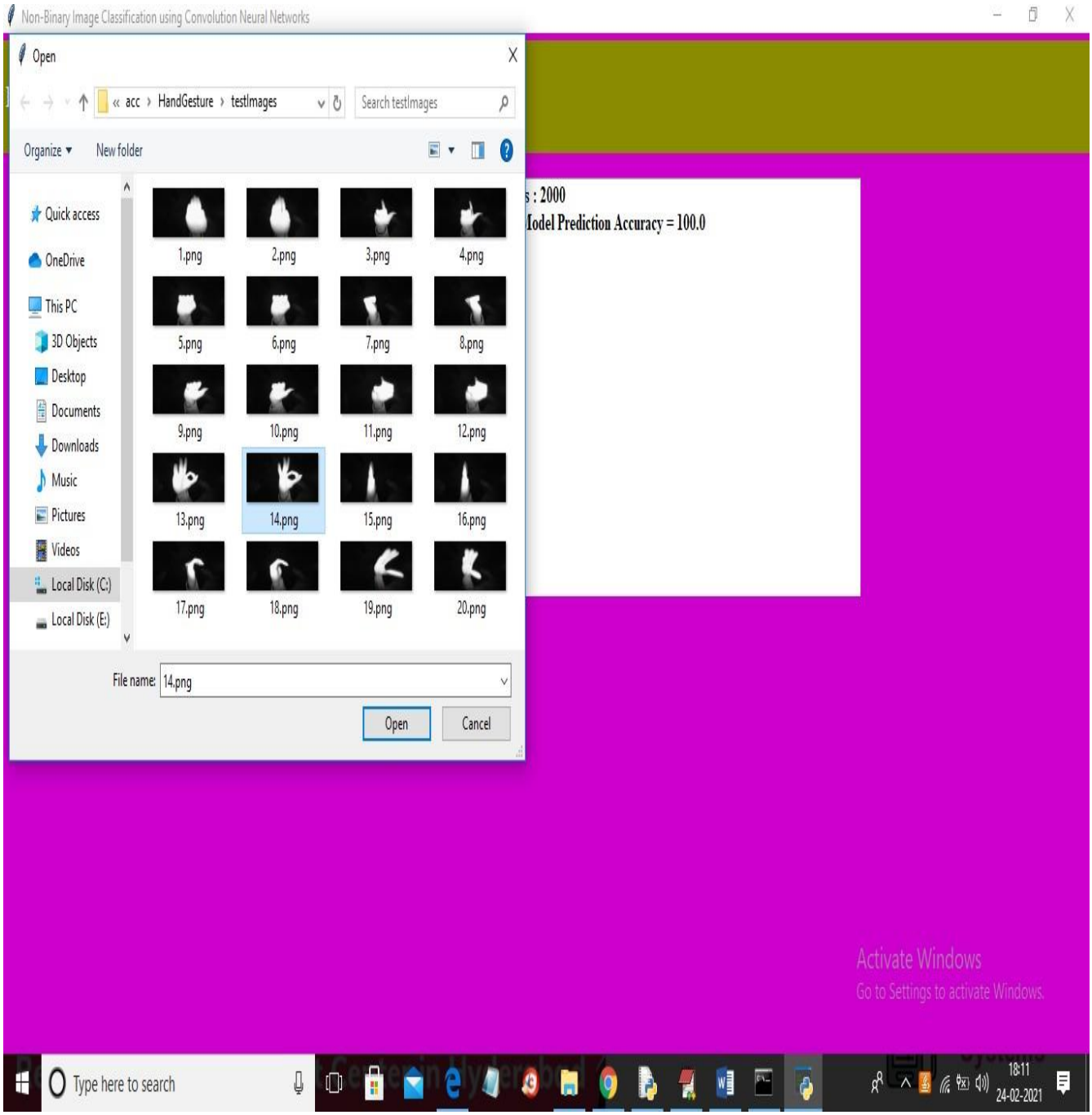


Fig 9.4

In above screen selecting and uploading ’14.png’ file and then click Open button to get below result

Results:

Upon running the algorithms the results are given below

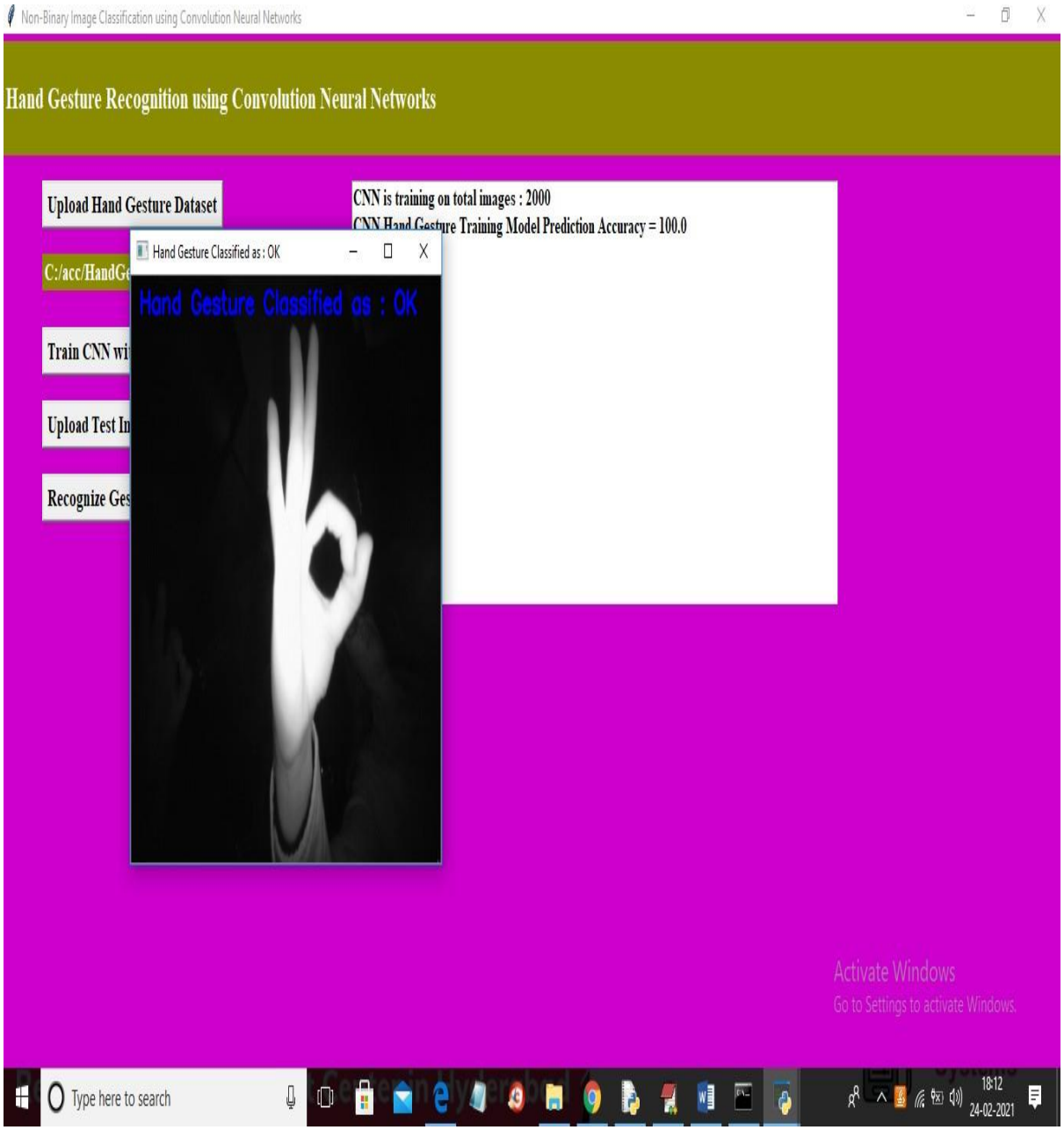


Fig 9.5:

In above screen gesture recognize as OK and similarly you can upload any image and get result and now click on ‘Recognize Gesture from Video’ button to upload video and get result

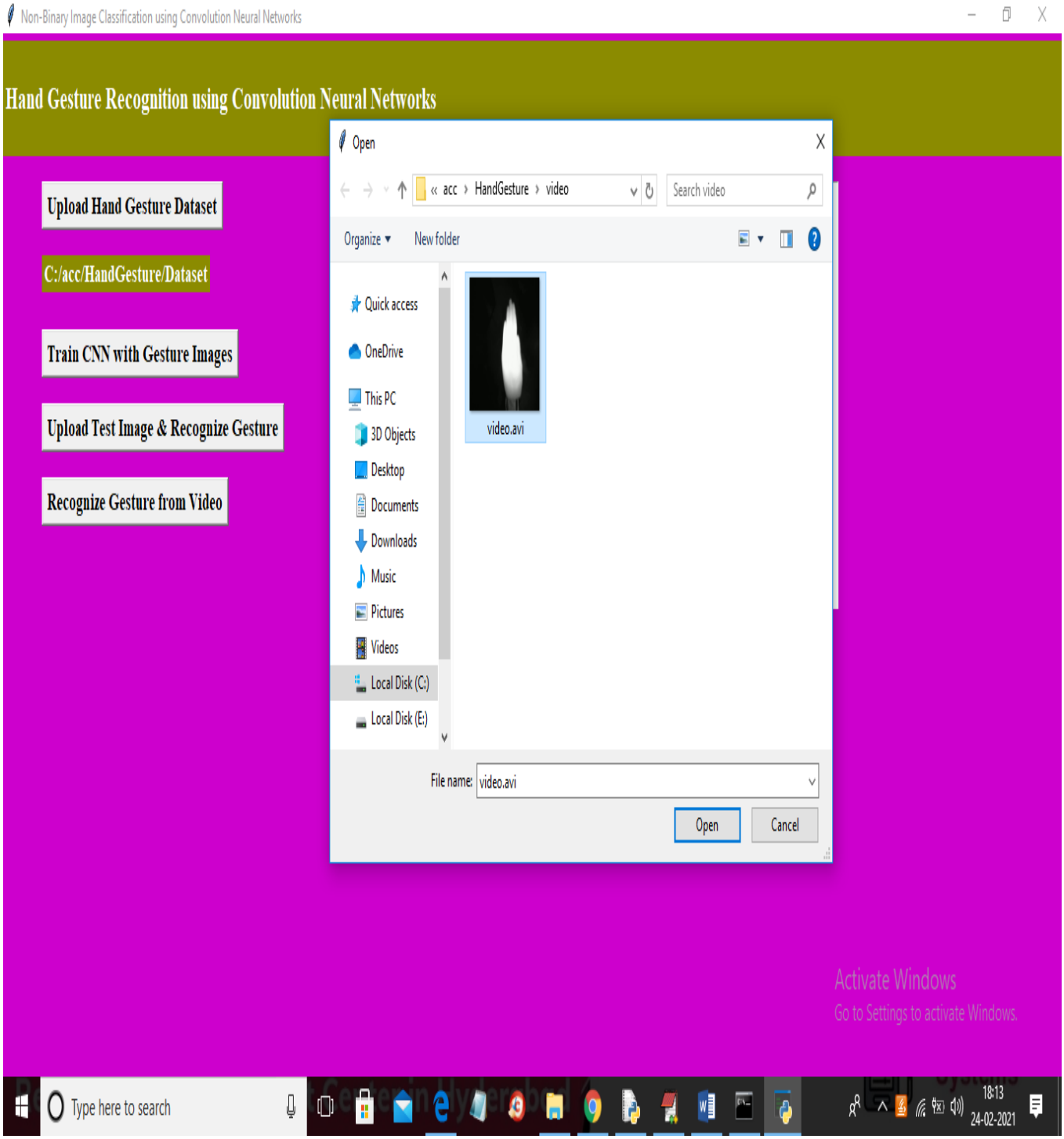


Fig 9.6:

In above screen selecting and uploading ‘video.avi’ file and then click on ‘Open’ button to get below result

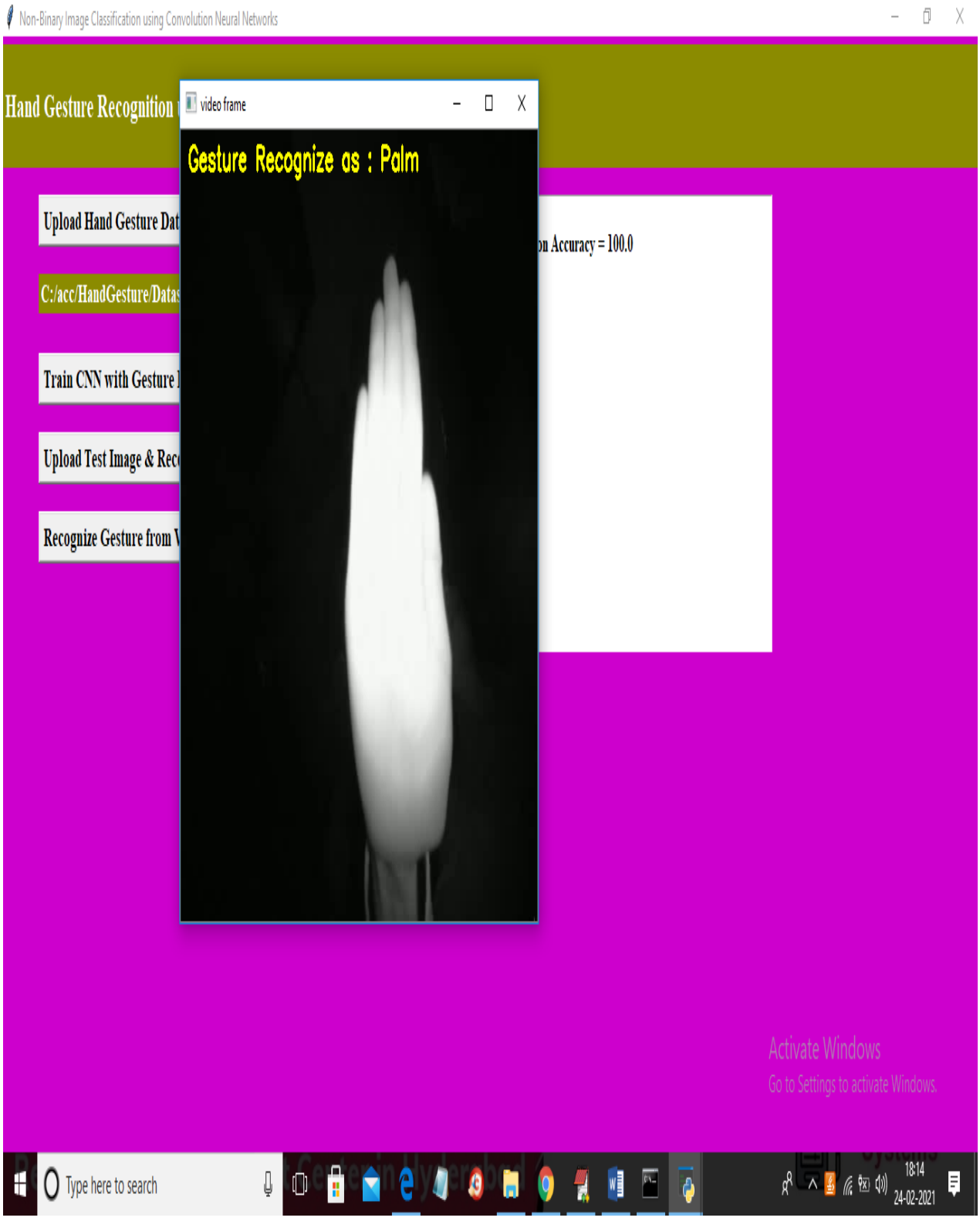


Fig 9.7

In above screen gesture recognised as Palm. The output shows palm hand posture along with text – Hand Gesture recognize as : Palm

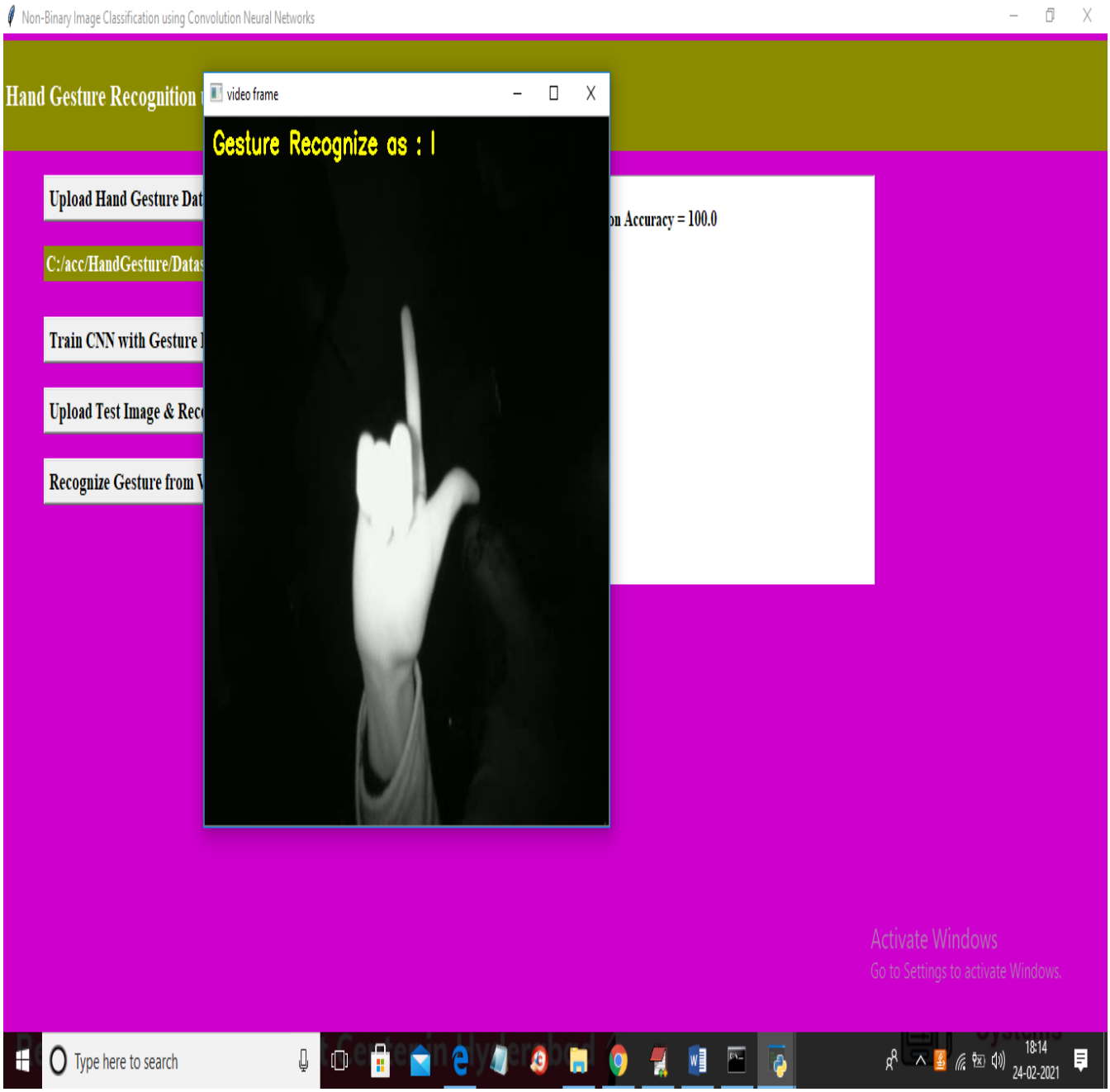


Fig 9.8:

In above screen gesture recognised as I. The output shows I hand posture along with text – Hand Gesture recognize as : I.

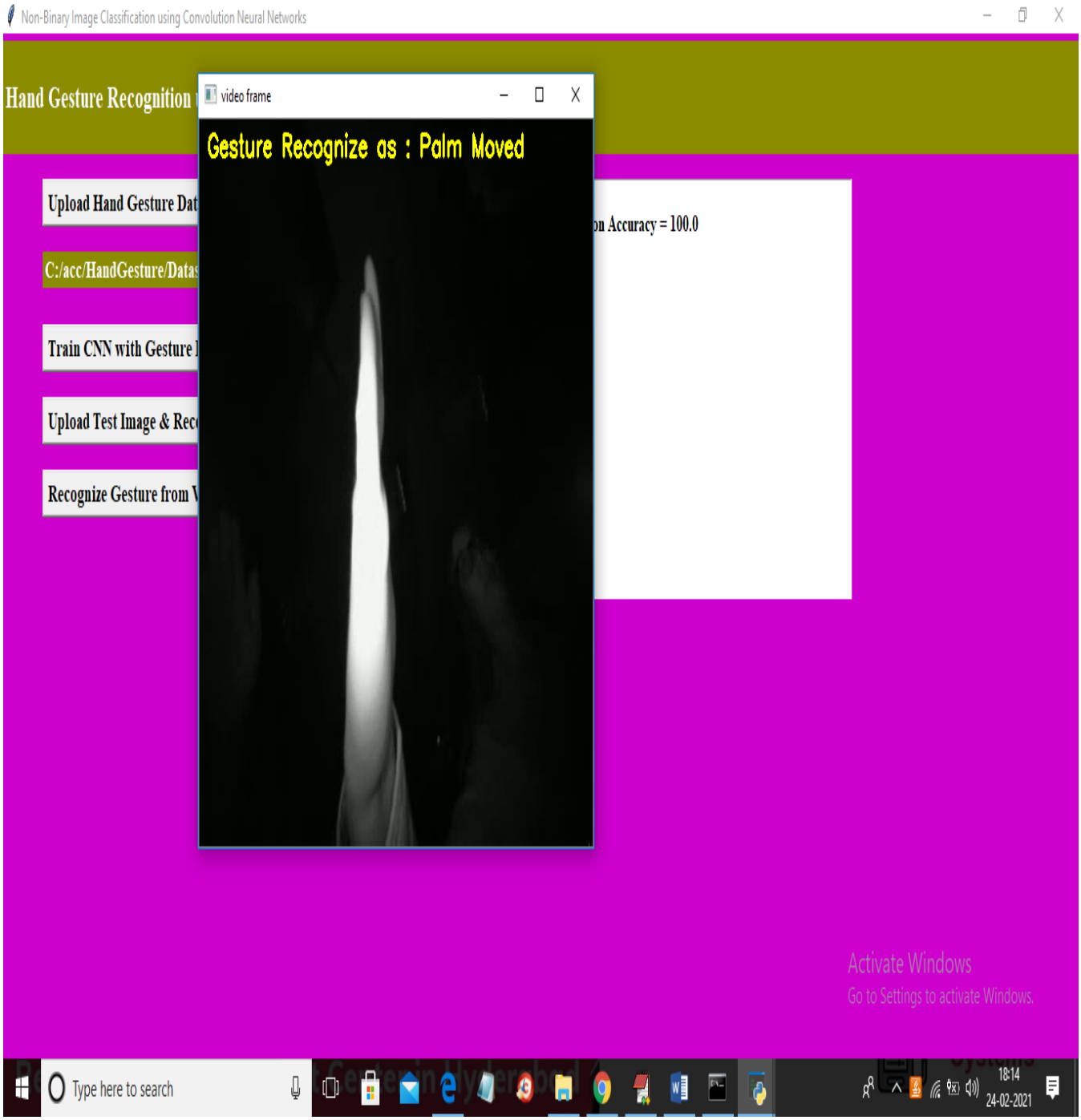


Fig 9.9

In the above screen Gesture Recognition is Palm Moved.. The output shows palm moved hand posture along with text – Hand Gesture recognize as : Palm Moved.

10. CONCLUSION AND FUTURE ENHANCEMENT

From the models that are developed we can conclude that it is able to handle some hand gestures provided by any person and help us to identify what the gesture is. So the main point which we can look into is that the machine is able to understand on the images and is able to identify what the images are that is really helpful in many ways

In this project, we developed a CNN-based human hand gesture recognition system. The salient feature of the system is that there is no need to build a model for every gesture using hand features such as fingertips and contours. To have robust performance, we applied a GMM to learn the skin model and segment the hand area for recognition.

Also, the calibration of the hand pose was used to rotate and shift the hand on the image to a neutral pose. Then, a CNN was trained to learn seven gesture types in this project. In the experiments, we conducted 4-fold cross-validation on the system where 600 and 200 images from a subject were used to train and test, respectively and the results showed that the average recognition rates of the seven gesture types were around 99%.

To test the proposed method on multiple subjects, we trained and tested the hand images of the seven gesture types from seven subjects. The average recognition rate was 95.96%. The proposed system also had the satisfactory results on the transitive gestures in a continuous motion using the proposed rules. In the future, a high-level semantic analysis will be applied to the current system to enhance the recognition capability for complex human tasks.

Our future enhancements are to improve our knowledge in sign language and frequently add images of new hand gestures in the data set so that the captured image from camera of that particular hand gesture can be interpreted and can produce the message in text format.

REFERENCES

[1] M.A.; Sharif, M.; Kadry, S.; Manogaran, G.; Saba, T. A framework of human action recognition using length control features fusion and weighted entropy-variances based feature selection. Image Vision Comput.2021

[2] Raudonis, V.; Damasevicius, R. Recognition of basketball referee signals from real-time videos.J.Ambient Intell. Humaniz. Comput. 2020

[3] Maskeliunas, R.; Damasevi cius, R. Detection of sitting posture using hierarchical image composition and deeplearning. PeerJ Comput. Sci. 2021

[4] Minh, Q.T. An ANN-based gesture recognition algorithm for smart-home applications. KSII Trans. Internet Inf. Syst. 2020

[5] Kaczmarek, W.; Panasiuk, J.; Borys, S.; Banach, P. Industrial robot control by means of gestures and voice commands in off-lineand on-line mode. Sensors 2020

[6] Mohammed, M.A.; Abdulkareem, K.H.; Mostafa, S.A. Voice pathology detection and classification using convolutional neural network model. Appl. Sci. 2020

[7] M.Mernik, et.al, (2018). Journal of visual languages and computing. ISSN: 1045-926X.

[8] Prof. Quan Min Zhu, (2018) International Journal of Computer Applications in Technology, Vol 57, No 4. pp 1-16.

[9] N.Paragios, (2018) Computer vision and understanding. Vol 116. pp 102-114

[10] Donsong Ya, Yongjun Xie, et.al, (2018) Design on Computer System based on hand gesture recognition. 2018 IEEE15th international conference on networking, Sensing and control, pp 1.

[11] G.M.Foddy, et.al, (2008), Multiclass and Binary SVM Classification: Implications for Training and Classification Users. IEEE Geoscience and Remote Sensing Letters. Vol: 5, Issue: 2. pp 241-245.

[12] Hong Dai, (2018). Research on SVM improves algorithm for large data classification. IEEE 3rd international conference on big data analytics, pp 181-185

[13] N.Paragios, (2018) Computer vision and understanding. Vol 116. pp 102-114

[14] Tan, Y.S.; Lim, K.M.; Lee, C.P. Hand gesture recognition via enhanced densely connected convolutional neural network. Expert Syst. Appl. 2021

[15] Ahmed, S. Hand Gesture Recognition Using an IR-UWB Radar with an Inception Module-Based Classifier. Sensors 2020.

PUBLICATIONS

JOURNAL (UGC approved Journal)

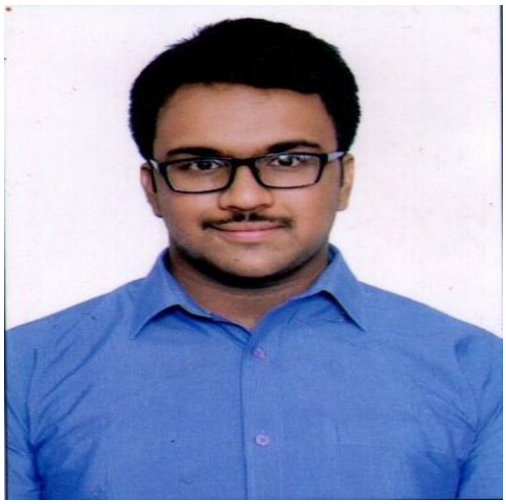
CONFERENCE (International Conference on “Innovations in Computers Networks, Computational Intelligence and IOT” [ICICCI-21]).

PAPER ID : ICICCI-21-0136

TOPIC : HAND GESTURE RECOGNITION USING CONVOLUTION NEURAL NETWORK.

STUDENT'S PROFILE

Adarla Rajeshwar is a Bachelor of Technology Student at St Martin's Engineering College in Computer Science Engineering stream. He finished his schooling till 10th grade in TSWR School,Armoor.and he completed his intermediate in TSWR Junior College. His technical skills include C, C++, Java, Python, MySQL. He has done an internship in Java at Electronics Corporation Of India Limited (ECIL). His participations include: a National Level Seminar on "Recent Trends in Cloud Computing, Fog, and Edge Computing" on 18th and 19th of June 2021 and a National Level Three Day Workshop on "AI & ML in Speech & Audio Processing" From 10th to 12th of December 2020. He also took part in the Employability Skill Development Program conducted by Zensar.



K. Bharatvamsi is a Bachelor Of Technology Student At St Martin's Engineering College in Computer Science Engineering stream. He finished his schooling till 10th grade in P. Obul Reddy Public School and Intermediate from Narayana Junior College. He was recruited in AIESEC which is the world's largest NGO , is in 126 countries & it's a non-profit organization recognised by UNESCO which provided young people with leadership development, cross cultural internships and global volunteers exchange experience from September 2017 to August 2018. In January 2019 he was selected to lead the financial team of the college branch of a NGO called Street Cause. He also did an internship in Java at Electronics Corporation Of India Limited (ECIL) in June 2019. He also took part in Employability Skill Development Program conducted by Zensar in 2019. His technical skill include C, C++, Java, MySQL & Python . He has keen interest in Creative and Content Writing, he has worked of StuMagz as a content writer in 2019 and also doing freelance writing for some clients.

Koluguri Santhosh Reddy is pursuing his Bachelor of Technology in the stream of Computer science and engineering at St.Martin’s Engineering College. He completed his intermediate from Narayana Junior College and schooling from Rao's My Techno School His technical skills include C, C++, Java and Python.He took part in Employability Skill development Program conducted by Zensar. His participations include: National Level Three Day Online Workshop on “AI & ML in Speech and Audio Processing” which was conducted from 10th to 12th December 2020. His areas of interest are Python, Artificial Intelligence, Machine Learning and Deep Learning. He completed few certification courses from online platforms like Coursera, CursaApp and SoloLearn.

Shivani Pagadala is pursuing her Bachelor of Technology in the stream of Computer science and engineering at St.Martin's Engineering College. She completed her intermediate from Toppers Junior college and schooling from Bhashyam School. Her technical skills include C++, Java, HTML ,CSS, Javascript and Python. Her area of interest is Data Science and web development. She has completed few certificate courses from online platforms like Coursera on Python Programming, Machine Learning , HTML, CSS, Managing Project Risks and Changes. Her participations include National Level Three Day Online Workshop on "AI & ML in speech and audio processing" which was conducted from 10th and 12th December, 2020, Leadership Talk with Mr.Mahesh Babu CEO Mahindra Electric Mobility Ltd. She worked as web developer at River Bend Data solutions for 1 month and as python tutor at Cuemath for 4 months.

APPENDICES

```
from tkinter import messagebox

from tkinter import *

from tkinter import simpledialog

import tkinter

from tkinter import filedialog

from tkinter.filedialog import askopenfilename

import cv2

import random

import numpy as np

from keras.utils.np_utils import to_categorical

from tensorflow.keras.layers import  MaxPooling2D

from tensorflow.keras.layers import Dense, Dropout, Activation, Flatten

from tensorflow.keras.layers import Convolution2D

from tensorflow.keras.models import Sequential

from tensorflow.keras.models import model_from_json

import pickle

import os

main = tkinter.Tk()

main.title("Non-Binary Image Classification using Convolution Neural Networks")

main.geometry("1300x1200")


global filename

global classifier

names = ['Palm','I','Fist','Fist Moved','Thumb','Index','OK','Palm Moved','C','Down']
```

```
bgModel = cv2.createBackgroundSubtractorMOG2(0, 50)


def remove_background(frame):
    fgmask = bgModel.apply(frame, learningRate=0)
    kernel = np.ones((3, 3), np.uint8)
    fgmask = cv2.erode(fgmask, kernel, iterations=1)
    res = cv2.bitwise_and(frame, frame, mask=fgmask)
    return res


def uploadDataset():
    global filename
    global labels
    labels = []
    filename = filedialog.askdirectory(initialdir=".")
    pathlabel.config(text=filename)
    text.delete('1.0', END)
    text.insert(END,filename+" loaded\n\n");




def trainCNN():
    global classifier
    text.delete('1.0', END)
    X_train = np.load('model/X.txt.npy')
    Y_train = np.load('model/Y.txt.npy')
```

```
text.insert(END,"CNN is training on total images : "+str(len(X_train))+"\n")


if os.path.exists('model/model.json'):
    with open('model/model.json', "r") as json_file:
        loaded_model_json = json_file.read()
        classifier = model_from_json(loaded_model_json)
    classifier.load_weights("model/model_weights.h5")
    classifier._make_predict_function()
    print(classifier.summary())
    f = open('model/history.pckl', 'rb')
    data = pickle.load(f)
    f.close()
    acc = data['accuracy']
    accuracy = acc[19] * 100
    text.insert(END,"CNN Hand Gesture Training Model Prediction Accuracy = "+str(accuracy))
else:
    classifier = Sequential()
    classifier.add(Convolution2D(32, 3, 3, input_shape = (64, 64, 3), activation = 'relu'))
    classifier.add(MaxPooling2D(pool_size = (2, 2)))
    classifier.add(Convolution2D(32, 3, 3, activation = 'relu'))
    classifier.add(MaxPooling2D(pool_size = (2, 2)))
    classifier.add(Flatten())
    classifier.add(Dense(output_dim = 256, activation = 'relu'))
    classifier.add(Dense(output_dim = 5, activation = 'softmax'))
    print(classifier.summary())
```

```
        classifier.compile(optimizer = 'adam', loss = 'categorical_crossentropy', metrics = ['accuracy'])
        hist = classifier.fit(X_train, Y_train, batch_size=16, epochs=10, shuffle=True, verbose=2)
        classifier.save_weights('model/model_weights.h5')
        model_json = classifier.to_json()
        with open("model/model.json", "w") as json_file:
            json_file.write(model_json)
        f = open('model/history.pckl', 'wb')
        pickle.dump(hist.history, f)
        f.close()
        f = open('model/history.pckl', 'rb')
        data = pickle.load(f)
        f.close()
        acc = data['accuracy']
        accuracy = acc[19] * 100
        text.insert(END,"CNN Hand Gesture Training Model Prediction Accuracy = "+str(accuracy))


def classifyFlower():
    filename = filedialog.askopenfilename(initialdir="testImages")
    img = cv2.imread(filename, cv2.IMREAD_COLOR)
    img = cv2.flip(img, 1)
    gray = cv2.cvtColor(img, cv2.COLOR_BGR2GRAY)
    blur = cv2.GaussianBlur(gray, (41, 41), 0)  #tuple indicates blur value
    ret, thresh = cv2.threshold(blur, 150, 255, cv2.THRESH_BINARY + cv2.THRESH_OTSU)
    thresh = cv2.resize(thresh, (224, 224))
    thresh = np.array(thresh)
```

```
    frame = np.stack((thresh,)*3, axis=-1)
    frame = cv2.resize(frame, (64, 64))
    frame = frame.reshape(1, 64, 64, 3)
    frame = np.array(frame, dtype='float32')
    frame /= 255
    predict = classifier.predict(frame)
    result = names[np.argmax(predict)]
    img = cv2.imread(filename)
    img = cv2.resize(img, (600,400))
    cv2.putText(img, 'Hand Gesture Classified as : '+result, (10, 25),
cv2.FONT_HERSHEY_SIMPLEX,0.7, (255, 0, 0), 2)
    cv2.imshow('Hand Gesture Classified as : '+result, img)
    cv2.waitKey(0)


def webcamPredict():
    videofile = askopenfilename(initialdir = "video")
    video = cv2.VideoCapture(videofile)
    while(video.isOpened()):
        ret, frame = video.read()
        if ret == True:
            img = frame
            img = cv2.flip(img, 1)
            gray = cv2.cvtColor(img, cv2.COLOR_BGR2GRAY)
            blur = cv2.GaussianBlur(gray, (41, 41), 0)  #tuple indicates blur value
            ret, thresh = cv2.threshold(blur, 150, 255, cv2.THRESH_BINARY + cv2.THRESH_OTSU)
```

```
            thresh = cv2.resize(thresh, (64, 64))
            thresh = np.array(thresh)
            img = np.stack((thresh,)*3, axis=-1)
            img = cv2.resize(img, (64, 64))
            img = img.reshape(1, 64, 64, 3)
            img = np.array(img, dtype='float32')
            img /= 255
            predict = classifier.predict(img)
            print(np.argmax(predict))
            result = names[np.argmax(predict)]
            cv2.putText(frame, 'Gesture Recognize as : '+str(result), (10, 25),
cv2.FONT_HERSHEY_SIMPLEX,0.7, (0, 255, 255), 2)
            cv2.imshow("video frame", frame)
            if cv2.waitKey(950) & 0xFF == ord('q'):
                break
        else:
            break
    video.release()
    cv2.destroyAllWindows()


font = ('times', 16, 'bold')
title = Label(main, text='Hand Gesture Recognition using Convolution Neural Networks',anchor=W,
justify=CENTER)
title.config(bg='yellow4', fg='white')
title.config(font=font)
title.config(height=3, width=120)
```

```
title.place(x=0,y=5)

font1 = ('times', 13, 'bold')
upload = Button(main, text="Upload Hand Gesture Dataset", command=uploadDataset)
upload.place(x=50,y=100)
upload.config(font=font1)

pathlabel = Label(main)
pathlabel.config(bg='yellow4', fg='white')
pathlabel.config(font=font1)
pathlabel.place(x=50,y=150)

markovButton = Button(main, text="Train CNN with Gesture Images", command=trainCNN)
markovButton.place(x=50,y=200)
markovButton.config(font=font1)

lexButton = Button(main, text="Upload Test Image & Recognize Gesture", command=classifyFlower)
lexButton.place(x=50,y=250)
lexButton.config(font=font1)

predictButton = Button(main, text="Recognize Gesture from Video", command=webcamPredict)
predictButton.place(x=50,y=300)
predictButton.config(font=font1)
```

```
font1 = ('times', 12, 'bold')

text=Text(main,height=15,width=78)

scroll=Scrollbar(text)

text.configure(yscrollcommand=scroll.set)

text.place(x=450,y=100)

text.config(font=font1)



main.config(bg='blue')

main.mainloop()
```

A

PROJECT REPORT

On

OPINION MINING FOR FEEDBACK MANAGEMENT SYSTEM

Submitted by

1)Mr.G.Suprabath Reddy (17K81A05D8)

2)Mr.E.Mahendra (17K81A05D6)

3)Mr.B.Ajay (17K81A05C8)

4)Mr.MD.Faisal (17K81A05G2)

in partial fulfillment for the award of the

degree of

BACHELOR OF TECHNOLOGY

IN

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Under the Guidance of

Mr. G.Venu Babu M.Tech

Assistant Professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ST.MARTIN'S ENGINEERING COLLEGE
An Autonomous Institute
Dhulapally, Secunderabad – 500 100
JUNE 2021

BONAFIDE CERTIFICATE

This is to certify that the project entitled Opinion Mining for Feedback Management System, is being submitted by **Mr.G.Suprabath Reddy (17K81A05D8), Mr.E. Mahendra(17K81A05D6),Mr.B.Ajay(17K81A05C8), Mr.MD.Faisal(17K81AO5G2)** in partial fulfillment of the requirement for the award of the degree of **BACHELOR OF TECHNOLOGY** in **Computer Science and Engineering** is recorded of bonafide work carried out by them. The result embodied in this report have been verified and found satisfactory.

Guide

Mr.G.Venu Babu

Assistant Professor

Department of CSE

Head of the Department

Dr.M.NARAYANAN

Department of CSE

Internal Examiner

External Examiner

Place:
Date:

DECLARATION

We, the student of **Bachelor of Technology** in Department of 'Computer Science and Engineering', session:<2017 – 2021>, St.Martin's Engineering College, Dhulapally, Kompally, Secunderabad, hereby declare that work presented in this Project Work entitled **Opinion Mining for Feedback Management System** is the outcome of our own bonafide work and is correct to the best of our knowledge and this work has been undertaken taking care of Engineering Ethics. This result embodied in this project report has not been submitted in any university for award of any degree.

Mr. G.Suprabath Reddy (17K81A05D8)
Mr.E.Mahendra (17K81A05D6)
Mr.B.Ajay (17K81A05C8)
Mr.MD.Faisal (17K81A05G2)

ABSTRACT

Academic industries used to collect feedback from the students on the main aspects of course such as preparations, contents, delivery methods, punctual, skills, appreciation, and learning experience. The feedback is collected in terms of both qualitative and quantitative cores.

Recent approaches for feedback mining use manual methods and it focus mostly on the quantitative comments. So the evaluation cannot be made through deeper analysis. In this paper, we develop a student feedback mining system (SFMS) which applies text analytics and sentiment analysis approach to provide instructors a quantified and deeper analysis of the qualitative feedback from students that will improve the students learning experience. We have collected feedback from the students and then text processing is done to clean the data. Features or topics are extracted from the pre-processed document. Feedback comments about each topic are collected and made as a cluster. Classify the comments using sentiment classifier and apply the visualization techniques to represent the views of students.

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible and whose encouragements and guidance have crowded effects with success.

We extended our deep sense of gratitude to Principal**, Dr. P. SANTOSH KUMAR PATRA**, St.Martin's Engineering College, Dhulapally, for permitting us to undertake this project.

We are also thankful to **Dr.M.NARAYANAN**, Head of the Department, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

We would like to thank **Dr.T.POONGOTHAI,** Department of Computer Science and Engineering (AI & ML) for her encouragement and insightful comments and as well as our project coordinators **Dr. B. RAJALINGAM**, Associate Professor and **Dr. N. SATHEESH**, Professor, in Department of Computer Science and Engineering for their valuable support.

We would like to express our sincere gratitude and indebtedness to our project supervisor **G.Venu babu**, Assistant Professor, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

Finally, we express thanks to all those who have helped us successfully to completing this project. Furthermore, we would like to thank our family and friends for their moral support and encouragement.

We express thanks to all those who have helped us in successfully completing the project.

Mr.G.Suprabath Reddy (17K81A05D8)

Mr.E.Mahendra (17K81A05D6)

Mr.B.Ajay (17K81A05C8)

Mr.MD.Faisal (17K81A05G2)

TABLEOFCONTENTS

LIST OF FIGURES

1. INTRODUCTION

Students provide feedback in quantitative ratings and qualitative comments related to preparation, contents, delivery methods, punctual, skills, appreciation, and learning experience. The delivery methods and preparation component refers to instructor's interaction, delivery style, ability to motivate students, out of class support, etc. The content refers to course details such as concepts, lecture notes, labs, exams, projects, etc. The preparation refers to student's learning experience such as understanding concepts, developing skills, applying acquired skills, etc. The paper correction refers to correction of mistakes and providing solutions to overcome it. The punctual refers to the class timing and assignment or record submission. The appreciation refers to the comments given when something is done perfectly. Analyzing and evaluating this qualitative data helps us to make better sense of student feedback on instruction and curriculum. Recent methods for analyzing student course evaluations are manual and it mainly focuses on the quantitative feedback. It does not support for deeper analysis. This paper focus on providing qualitative and quantitative feedback to analyze and provide better teaching to improve the student's performance. The paper will be structured as follows: Section 2 will review the techniques used in text processing and sentiment analysis approach in the background. Section 3 will describe the related works of the current research about the student feedback mining system. Section 4 will provide the proposed system of this paper. Section 5 will have experiments and future works to be implemented and finally we concluded in section 6

1.1 PROJECT OVERVIEW

Content growth in the Internet in recent years has made a huge volume of information available. This information is presented in different formats such as posts, news articles, comments, and reviews. Especially in the automotive, electronics and film sectors, customers have written reviews about products or their features. By collecting and analyzing these reviews, new customers find others' opinion about different features of the product. They can compare the products to each other to find the best one that meets their needs. Moreover, manufacturers will find out strengths and weaknesses of their products or those of their competitors.

1.2 PROJECT OBJECTIVES

Given an object and a collection of reviews on it, our objectives are

• Extract Nouns, Adjectives, Verbs and Adverbs on the remaining reviews by using dictionary approach.

• Identify frequent words by using Apriori frequent item set mining Algorithm.

• Perform Sentiment Analysis on the frequent words using SentiWordNet.

• Provide visualization.

As mentioned in many literatures and reports, the Web contains a wealth of product reviews by customers due to the ease of publishing online in the recent years. This leads to an explosion in the web content of customer reviews and opinions on products. It is an expensive and daunting task to sift through them. The good news is that this mass of data makes automation of online customer reviews collection and classification possible and worthwhile. The mining, classification and collection of information from these online customers are important to the product manufacturers, as it provides them information about any product defects at an early stage from customers complaints that could quickly proliferate through the Internet. Knowing products limitations and defects can be helpful in risk management and henceforth, reducing future liabilities, as well as to make sound marketing strategies. These online reviews are also interesting to existing and potential customers too. It could significantly influence their purchase decisions of a product or in hiring a service. These explain for the importance of opinion mining and sentiment classification of online customer reviews. To further explain, the goal of business intelligence in customer satisfaction studies of their opinions on products and/or services, and can often be explored using surveys and focus groups studies. However, these can incur high expenses associated with the design and the administration of the surveys, and also are time consuming in the process of the obtainment of the resulting data. A less formal or structured alternative is the collection of spontaneous customer reviews and feedbacks on the Web, such as blogs, newsgroups, feedback emails and review websites. These data can come in free text that is less structured compared to surveys and focus groups. However, without much effort, free-form customer opinions can also be processed more efficiently and effectively, using automatic text mining techniques in the market, such as clustering and key term extraction.

In this way, manufacturers will solve the reported problems and use the business intelligence behind the analysis for future investments. From the sentiment perspective, there are two kinds of textual information, namely, facts and opinions. While facts are the objective statements about the nature of a product, opinions describe attitudes, appraisals, and emotions regarding a product, service, topic, or an issue. Although the majority of research focuses on building applications around facts, the recent trend in the area of text mining has been focused on building applications around opinions.

1.3 ORGANIZATION OF CHAPTERS

Sentiment analysis is an interdisciplinary field that crosses natural language processing, artificial intelligence, and text mining. Since most opinions are available in the text format and its processing is easier than other formats, sentiment analysis has emerged as a subfield of

text mining. It generally recognizes opinions of people expressed in text. The opinions could be judgments, evaluations, affective (or emotional) states, beliefs, or wishes. Sentiment analysis appeared in the literature in 1990 for the first time and then it became a major research topic in 2000. Classifying the polarity of a given text as positive or negative is the basic task of sentiment analysis. Due to its many aspects it is often referred to with different names such as opinion mining, sentiment classification, sentiment analysis, and sentiment extraction. It is widely believed that Sentiment analysis is needed and useful. It is also widely accepted that extracting sentiment from text is a hard semantic problem even for human beings. Additionally, sentiment analysis is domain specific, therefore the polarity of some terms depends on the context in which they are used. For example, while “small” for “size” as a feature in the electronic products is positive, in agricultural products such as fruit it has a negative polarity. Sentiment analysis is used in different domains such as shopping, entertainment, politics, education, marketing, and research and development. This paper focuses on sentiment classification in social domains from the technical perspective, two main approaches for sentiment analysis are Bag Of Words (BOW) and Feature Based Sentiment (FBS). In the BOW approach, each document is seen as a set of words. As a result, the syntactic and semantic information between words are lost. The BOW approach is not useful when opinions about products and their features have to be . In such cases, it is required to extract features. FBS has emerged as an approach for analyzing the sentiments of products and their features. The results of sentiment classification are presented in various formats in different domains: positive/negative, like/dislike, recommended/not recommended, good/bad, buy / don’t buy, excellent/boring(film), support/against, favorable / unfavorable , bullish/bearish, or optimistic / pessimistic.

2. LITERATURE SURVEY

In this section, we present a review of the existing and related works on Opinion Mining (OM) and Sentiment Classification (SC) proposed in the literature. The state-ofthe-art theories and models in today's literature are also presented. This review is categorized in the following seven categories as shown in Figure 1. It outlines the various techniques used for Opinion Mining and Sentiment Classification from the existing literature. The different techniques used to mine opinions, classify sentiment of mined items and features, as well as the strength of the sentiment are reviewed; and compared and contrasted against each other.

Item Extraction: we analyze all the frameworks that are related to item extraction. Specifically two papers are of importance as they focus on this topic in detail. Item extraction is the process of extracting the subject matter where opinions have been expressed on in customer reviews. It is also commonly referred to as 'Opinion Extraction'. The term 'subject matter' is also commonly referred to as a situation or a product. It is an important task as it is the beginning stage in the task of OM. In contrast to sentiment classification, opinion extraction aims at producing useful richer information for in-depth analysis of opinions.

When evaluating these frameworks, we rely on the questions listed in Table 1. Answers to these questions reflect the advantages and disadvantages of the researched framework. Since this problem was discovered, limited research has been undertaken to attempt to solve it, for example the work by Kobayashi et al.and Gamon et al.

Kobayashi and his team presented a method for opinion extraction in a structured form. It also discussed the most effective way to structure customer reviews in web documents and focused on extracting subject/aspect evaluation relations, and extracting subject/aspect-aspect relations, using a machine learning-based method, which is portable across domains. It addressed the task of opinion extraction by combining contextual clues and context independent statistical clues using a machine-learning technique, 'boosting-based algorithm' by Kudo (04) implemented as the package BACT. Experiments were carried out and evaluation was conducted using 5 fold cross validation on all data in the aspects of recall and precision.

Compared the Kobayashi, the approach presented by Gamon and his team, is termed as Pulse, a prototype system for mining topics and sentiment orientation jointly from free text customer feedback. Pulse combines a clustering technique with a machine-learned sentiment classifier, allowing for a visualization of topic and associated customer sentiment. The Tree Map visualization is used to display clusters and their associated sentiment. It allowed the identification of the overall sentiment associated with product make/model, the most common topics that customers mentioned in reviews, as well as the most positive and the most negative topics, at a glance. Machine-learning techniques have been applied in both algorithms and experimental results have shown significant improvement over many existing methods such as Baseline A-E model and Context-only A-E model; where Baseline A-E (aspect-evaluation) model stimulates the algorithm proposed by (Tateishi 04) and Context-only A-E model uses contextual pattern-based clues but not statistical clues and is a boost-based algorithm proposed by Kudo.that Kobayashi used supervised learning and a combination of contextual

clues and context Flowchart of OM and SC Framework 7.Strength of Sentiments Opinion Mining and Sentiment Classification Opinion Extracted 1. Item Extraction 2.Feature Extraction 3. Comparison of Items and Features 4. Sentiment Classification on Items 6. Sentiment Classification on Features 5. Sentiment Classification on Items and Features Feature vs Feature Item vs Item Item A Feature A Item B Feature A vs 2009 3rd IEEE International Conference on Digital Ecosystems and Technologies 978-1-4244-2346-0/09/$25.00 ©2009 IEEE 397 independent statistical clues in their experiment; whilst Gamon [10] used a combination of clustering techniques and machine-learned sentiment classifier. We will explore another area of opinion mining, 'feature extraction', in the next section that can add more value to the existing knowledge about the item extracted from reviews.

B. Feature Extraction: Feature extraction is the identification of features of products which customers have expressed their opinions on their reviews and feedbacks. Features refer to product features, product attributes, and/or product functions like the picture quality of the Canon IXUS 10, or the interior design of a Ford territory, or the service of hotel staff. It is essential to readers that the features of the reviewed products are known as their areas of importance in different products may differ from people. For example, a reader might be more interested in the cleanliness of the hotel room, whilst the reviewer is more concerned with the quality of the customer service of the hotel staff. Having said, current work in the domain of Feature Extraction is still in its infancy. We have reviewed three papers and identified their main contributions, weaknesses, similarities and differences as shown in Table 2. An evaluation of Feature Extraction Methods Papers Reviewed Parameters for evaluation 1 6 8 Does the algorithm identify & classify opinion sentences in reviews? Y N N Does the algorithm provide a summarization of the results? Y N N Does the algorithm use POS Tagging? Y N Y Does the algorithm use WordNet? Y N Y Does the algorithm use Word n-grams? N N Y Can algorithm detect the weight of the opinion based on the opinion itself? N NN Can algorithm differentiate different features of the same product? Y YY Can feature hierarchy be constructed? Y N N Hu et al. [1] studied the problem of generating featurebased summaries (FBS) of customer reviews of products sold online and strived to mine and summarize all the customer reviews. Their task was performed in three main steps. Firstly, to mine product features that have been commented on by customers; secondly, to identify opinion sentences in each review and decide whether each opinion sentence is positive or negative; and finally to summarize the results. The paper also proposed several novel techniques to aid in the process of performing these tasks such as POS Tagging, Association Miner CBA, and WordNet. Their experimental results using a large number of customer reviews of 5 products sold online demonstrated the effectiveness of the Feature-Based Summarization (FBS) and its techniques against other existing methods used by other researchers such as the FASTR of Christian Jacquemin [14]. In this work, opinion classification was performed at the sentence level rather than at the document level. In contrast to related works like Dave, Lawrence and Pennock's work [2], they did not find features on which opinions have been expressed, or summarized similarities and differences of reviews, but instead aimed to find the key features that are talked about in multiple reviews. Their focus was on using association mining to find all frequent features. The next paper by Kobayashi et al. [6] not only extracted the explicit features in reviews, they also extracted the Subject and the Value of the reviews. They believed these details are essential for addressing the task. Assuming opinions can be represented as tuples (Subject, Attribute, Value), they addressed the task of opinions extraction by employing a computational method for tuples extraction. Machine-learning based techniques are then applied to the main task of opinion extraction, which was then decomposed into two subtasks: Extraction of attribute-value pairs related to a product (where an attribute represents one aspect of a subject and the

value is a specific language expression that qualifies or quantifies the aspect); and Determination of its subjectivity on the opinion as a whole. The proposed method had yield a better outcome. The next paper to be discussed is from Mishne who took a totally different and unique approach, the first in the database of literature on opinion mining at that time. Mishne's work as was more focused on the classification of the blog posts by different moods. In other words, the paper not only addressed the task of feature extraction, but continued to use the extracted features in assisting for their task of classifying blog posts by moods. They believed that mood classification is useful for various applications, such as assisting behavioral scientists and improving doctorpatient interaction. Their objective was to predict the reviewer's most likely state of mind when the post was written using a machine learning approach to identify a set of features to be used for the learning process. Their experimental results had showed a consistent modest improvement on the naïve baseline. Table 2 shows that Hu and Mishne had used POS Tagging and Word n-grams in their work. Hu also used Association Miner CBA, and WordNet; whilst Mishne used Naïve method, frequencies of word lemmas acquired from Tree Tagger, as well as SVMlight package in their work. Contrasted, Kobayashi applied a machine learning-based method used for anaphora resolution to the opinion extraction problem. Table 2 also shows that all the three algorithms can differentiate different features of the same product, but only Hu's work allowed feature hierarchy to be constructed as they focused on finding key features of the products. The tasks of item and feature extraction are covered in section A and B respectively. In section C, the task of sentiment classification is discussed. 2009 3rd IEEE International Conference on Digital Ecosystems and Technologies 978-1-4244-2346-0/09/$25.00 ©2009 IEEE 398

C. Sentiment Classification in General: Sentiment classification is the process of determining the subjectivity of a given text. In simple words, it is the task of deciding whether a given text expresses a positive or negative opinion about its 'subject matter' and 'subject attributes', which is also known as 'product' and 'features'. Table 3. An evaluation of Sentiment Classification Methods Papers reviewed Parameters for evaluation 2 7 11 12 Does the algorithm uses scoring methods from information retrieval for sentiment determination? Y N NN Does the algorithm use General Inquirer (GI) lexicon? N N Y Y Does the algorithm uses WordNet? N NN Y Does the algorithm use linguistic rules? N Y N N Does the algorithm use aggregation function? N Y N N Sentiment classification assessment on opinions is usually done on document level rather than on sentence level as in contrast to opinion mining. Table 3 presents the similarities and differences of the four papers relevant to the area of sentiment classification. Dave et al. applied various machine learning methods for the task of opinion extraction and sentiment classification and discovered several problems that have not been expected initially. They began by using structured reviews for testing and training and identifying appropriate features. Table 2 shows that scoring methods was used in information retrieval for determination of the sentiment of reviews. Experiments were conducted on user reviews on C|net and Amazon where authors provided quantitative or binary ratings, which were believed to be perfect for training and testing for sentiment orientation. Two tests were conducted and experimental results showed that their best methods performed as well as or better than traditional machine learning methods. Sentiment classification in general is covered and the next section is 'sentiment classifications on items' which provides more specified and targeted information.

D. Sentiment Classification on Item: This section is more focused on whether a given text has positive or negative connotation on its subject matter only. For example, Camera 1 has positive or negative feedback from users online. The paper by Esuli et al. is the extension of the authors' previous work: "Esuli & Sebastiani, 2005, 'Determining the semantic orientation

of terms through gloss analysis'. It confronted the task on the decision of whether a given term has a positive connotation, or a negative connotation, or has no subjective connotation at all; thus, this problem subsumed the problem of determining orientation. This problem was tackled by testing three different variants of a semi-supervised method previously proposed for orientation detection. Their results showed that determining subjectivity orientation was much of a harder problem than determining orientation alone. The benchmark that they had used for their experiments is the General Inquirer (GI) lexicon (Stone et al., 1996) as shown in Table 3. This is a lexicon of terms labled according to a large set of categories, each one denoting the presence of a specific trait in the term. Unfortunately, their results had shown that an algorithm that had shown excellent state-of-the-art performance in deciding term orientation , once modified for the purposes of deciding term subjectivity, performed more poorly. This had been shown by testing several variants of the basic algorithm, some of them involving radically different supervised learning policies. The next work by Esuli et al. used SentiWordNet, a lexical resource describing the degree of positivity and negativity of the extracted terms and features. It is not an extension of the previous research, although Table 3 has showed that both used GI lexicon. SENTIWORDNET, is a lexical resource in which each WORDNET synset, s, is associated to three numerical scores Obj (s), Pos (s) and Neg (s), describing how objective, positive, and negative the terms contained in the synset are. Based on observations in their previous works, they had decided to combine different configurations of training set and learner into a committee to produce the final SENTIWORDNET scores. They believed SentiWordNet is a useful tool in opinion mining applications, because of its wide coverage and its fine grain properties, obtained by qualifying the labels by means of numerical scores. The next section is sentiment classification on features, which is the next stage in sentiment classification on items.

E. Sentiment Classification on Features: We are discussing sentiments on features in this section, as in deciding whether a given text has a Positive or Negative opinion on its 'subject attributes', which we also commonly referred to as 'product attributes' and/or 'product features'. For example, the features of Camera A being 'buttons placement on the camera' and 'size of screen'. Of particular relevance in this topic is the work by Ding et al.'s. Ding discussed the problem of determining semantics of opinions expressed on product features in customer reviews, rather than on the products (items) mentioned in the reviews compared to Esuli et al. in and. The objective of Ding's work was to use linguistic rules together with a new opinion aggregation function to address the problem of sentiment classification on features as shown in Table 3. This approach used context to infer the orientations of opinions on a product feature. The 'S system', called Opinion Observer, was also presented in this research. In this work, the product features are assumed to be given or discovered before determining whether an opinion 2009 3rd IEEE International Conference on Digital Ecosystems and Technologies 978-1-4244-2346-0/09/$25.00 ©2009 IEEE 399 is positive or negative and results of experiments on Opinion Observer showed that it had outperformed both FBS and OPINE . Additional experiments had also showed that both the new opinion aggregation function and linguistic rules had contributed roughly equally to the improved results of Opinion Observer over FBS with the recall factor improving dramatically without much loss in the precision rating. We have covered the different aspects of sentiment classifications in section C, D and E, as well as the different methods used to accomplish the task effectively. In the next section, we touched on the 'strengths of the sentiments' to get more detailed information and a better understanding on the extracted sentiments.

F. Strength of Sentiments: Determining the strength of sentiments is the process of deciding whether a Positive opinion expressed by a text on its subject matter is Weakly Positive,

Mildly Positive, or Strongly Positive, and/or whether a Negative opinion expressed is Weakly Negative, Mildly Negative, or Strongly Negative. Table 4. An evaluation of determining strength of sentiments methods Papers reviewed Parameters for evaluation 3 4 Is algorithm portable across domain (Domain independent)? Y N Does the algorithm categorize the sentiment strength using objective measures? Y Y In this section, the two papers of particular relevance are Popescu et al. and Wilson & Wiebe et al., whose works discuss the strength of the sentiments of customer reviews and feedbacks. The paper by Popescu et al. introduced OPINE, a more in-depth and detailed unsupervised information extraction system which extracted fine-grained features, and associated opinions, from reviews. OPINE mined reviews for the purpose of building a model of important product features, their evaluation by reviewers, and their relative quality across products. It focused on the extraction of explicit features, identifying corresponding customer opinions about these features and determining their polarity. It differed from method used in as instead of only finding the key features in multiple reviews, it also determines the polarity and strengths of the sentiments of the reviews. It is also portable across domain as shown in Table 4. OPINE extracts product features such as properties, parts, features of product parts, related concepts, and parts and properties of related concept, and also opinion phrases, which are adjective, noun, verb or adverb phrases representing customer opinions using WordNet's IS-A hierarchy and morphological cues from a set of reviews from Hu & Liu's publicly available data sets. After various tests and experiments, results had confirmed the performance superiority of OPINE in its tasks as compared to similar review-mining system, FBS. Its novel use of relaxation labeling technique for determination of semantic orientation of words in context of given product features and sentences had also led to better performance on the tasks of customer opinions identification and their polarity. OPINE was literally an improvisation of existing Hu's FBS on feature extraction. The work by Wilson and Wiebe et al. was the first research that touches on automated opinion and sentiment classification. Its main objective was to automate the task of distinguishing between objective and subjective languages as well as the strength of the sentiments of each sentence in the corpus of reviews used for their paper, which was very different to OPINE by Popescu et al. Experiments were conducted and they had achieved significant improvements in mean-squared error over baseline using three machine learning algorithms, which are BoosTexter (Schapire and Singer, 2000) AdaBoost.HM for boosting, Ripper (Cohen, 1995) for rule learning, and SVMlight (Joachims, 1999) for support vector regression. In the next section, the comparison of items and features is discussed.

G. Comparison of Items and Features: The comparison of items and features is the process of comparing products in the same product groups (e.g. Canon camera and Sony camera) mentioned in customer reviews in terms of their features. E.g. "the battery life of camera A is much longer than the battery life of camera B", or "I like the lens of camera A but at the same time I prefer the battery life of Camera B". The purpose of the comparison is to allow readers to quickly and clearly identify the more superior product out of many in terms of different features. The work by Liu et al. addressed this topic by proposing an analysis system, Opinion Observer, with a visual component for comparison of consumer opinions on different products. This system compared and contrasted features of various products in many forms of visual diagrams allowing users to clearly see the strengths and weaknesses of products based on features at a single glance. Their work was related but quite different from sentiment classification; as its purpose was to classify reviews as positive or negative and did not identify product features. Experimental results on the evaluation of their proposed technique in identification of product features from Pros and Cons had shown that the technique was highly effective and outperformed existing methods significantly. We can see that the Opinion

Observer is a very powerful and useful analysis system created by Liu et al. to allow readers to quickly absorbed information in numerous reviews at a single glance to make informed purchase decision. Without it, readers will have to spend many hours going through mountainous of online reviews which are tedious and unconventional in our today's time-pressed society.

3.SOFTWARE AND HARDWARE REQUIREMENTS

3.1 SOFTWARE REQUIREMENTS

- Python
- HTML & CSS
- MySQL

Operating System:

- Windows 10

3.2 HARDWARE REQUIREMENTS

- System : Intel core i3 (2.4 GHz)
- Hard Disk : 1TB
- Mouse : Logitech
- Ram : 8 GB

4. SOFTWARE DEVELOPMENT ANALYSIS

4.1 OVERVIEW OF PROBLEM

Academic industries used to collect feedback from the students on the main aspects of course such as preparations, contents, delivery methods, punctual, skills, appreciation, and learning experience. The feedback is collected in terms of both qualitative and quantitative scores. Recent approaches for feedback mining use manual methods and it focus mostly on the quantitative comments. So the evaluation cannot be made through deeper analysis.

4.2 DEFINE THE PROBLEM

Recent approaches for feedback mining use manual methods and it focusmostly on the quantitative comments. So the evaluation cannot be made through deeper analysis. In this paper, we develop a student feedback mining system (SFMS) which applies text analytics and sentiment analysis approach to provide instructors a quantified and deeper analysis of the qualitative feedback from students that will improve the students learning experience. We have collected feedback from the students and then text processing is done to clean the data. Features or topics are extracted from the pre-processed document. Feedback comments about each topic are collected and made as a cluster. Classify the comments using sentiment classifier and apply the visualization techniques to represent the views of students

4.3 MODULES OVERVIEW

- **Admin Login : -** Admin login's to the system using his Admin ID and password.

- **Add Post : -** Admin can post topics.
- **Add Keywords:** - Admin add keywords in database so that system will match the comment with the keywords in database and will rank the topic.
- **User Login : -** User login's to the system using his user ID and password.
- **Comment :** - User will post comment on the topic.
- **View Comment:** - User can view comment of other user's.
- **Rating Calculation:** - System will match the comment with the keywords in database and will rate the topic.
- **Edit Profile :** - User can edit his profile and can change his profile picture.
- **Status :-** User can view status and can change his status.

5. PROJECT SYDTEM DESIGN

5.1 DFDS IN CASE OF DATABASE PROJECTS

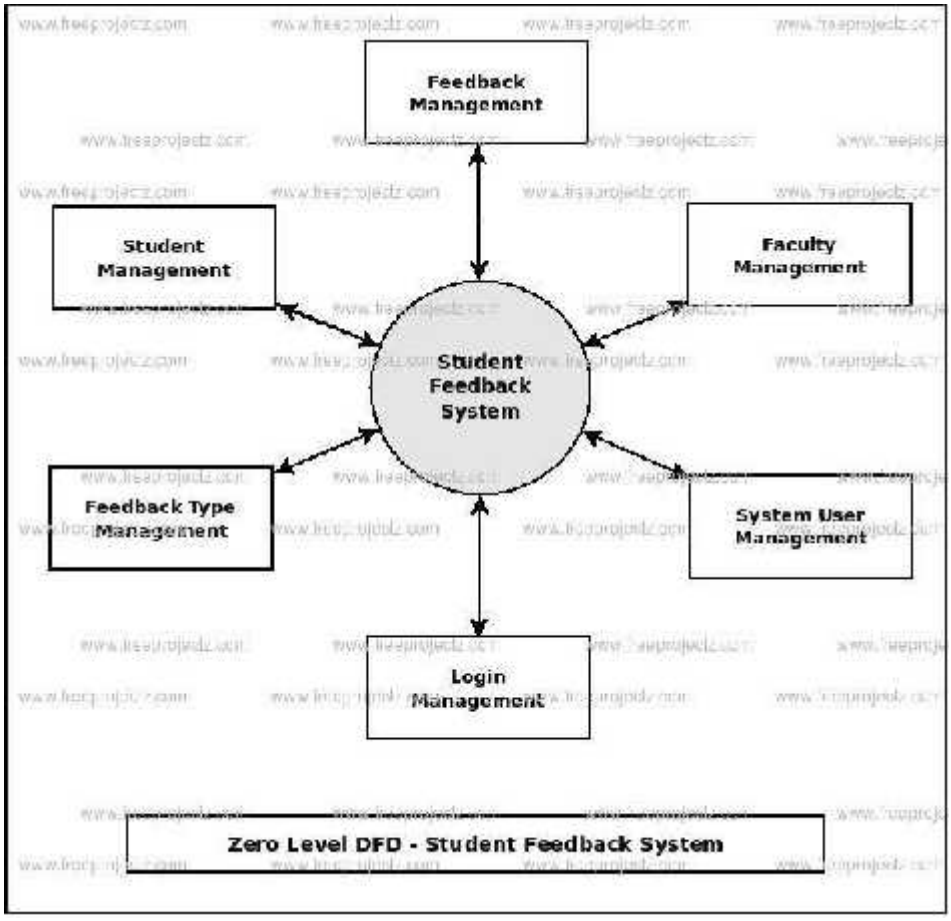


Figure 1: DFDS IN CASE OF DATABASE PROJECTS

5.2 E-R DIAGRAMS

a.User

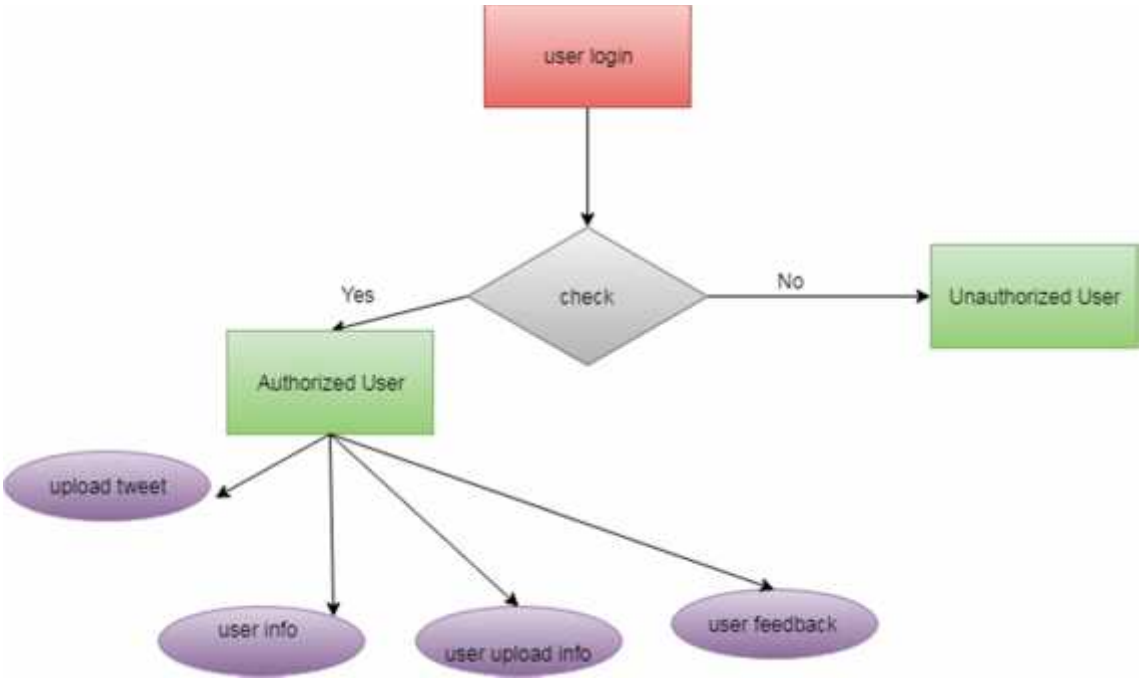


Figure 2: ER diagram for user

b. Admin

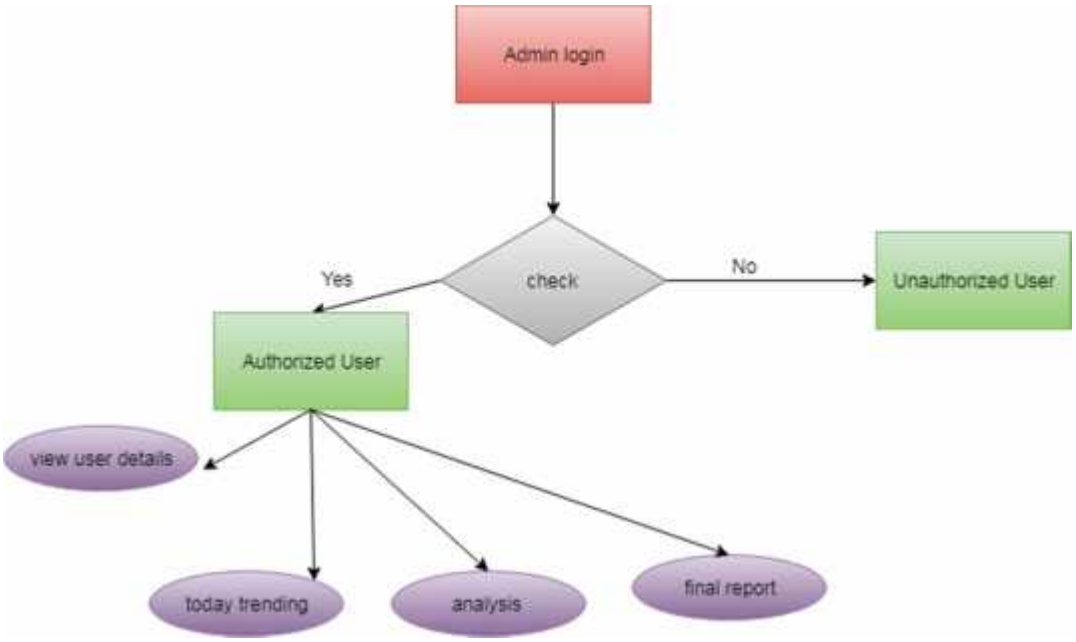


Figure 3: ER diagram for Admin

5.3 UML DIAGRAMS

CLASS DIAGRAM

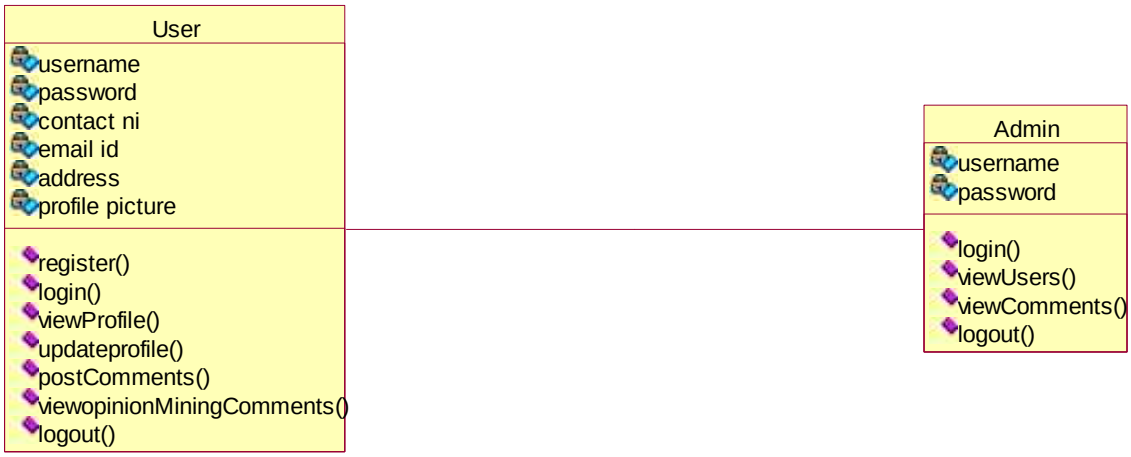


Figure 4: Class diagram

USECASE DIAGRAM

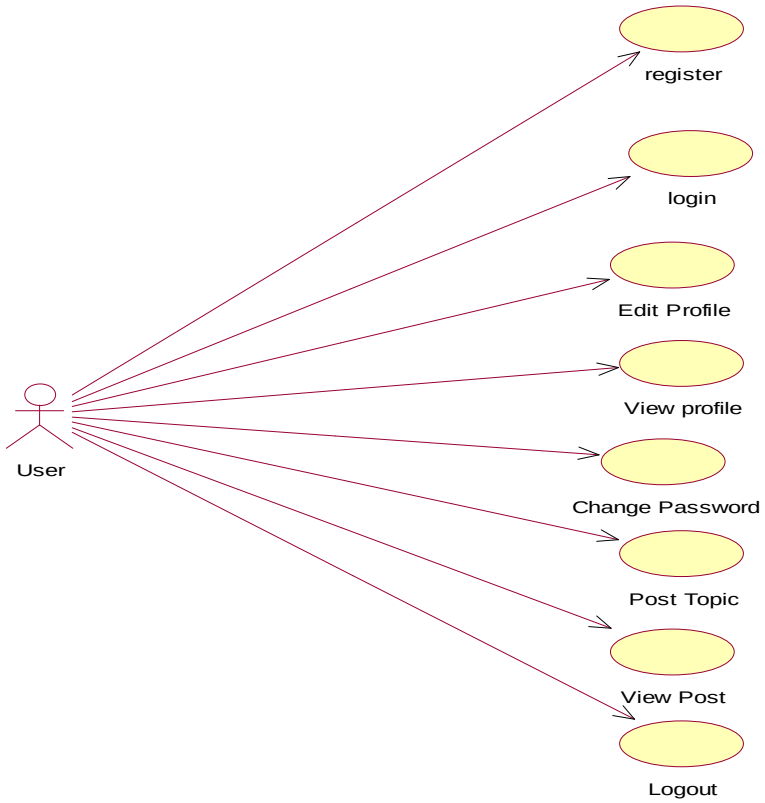


Figure 5: Use Case diagram

SEQUENCE DAIGRAM

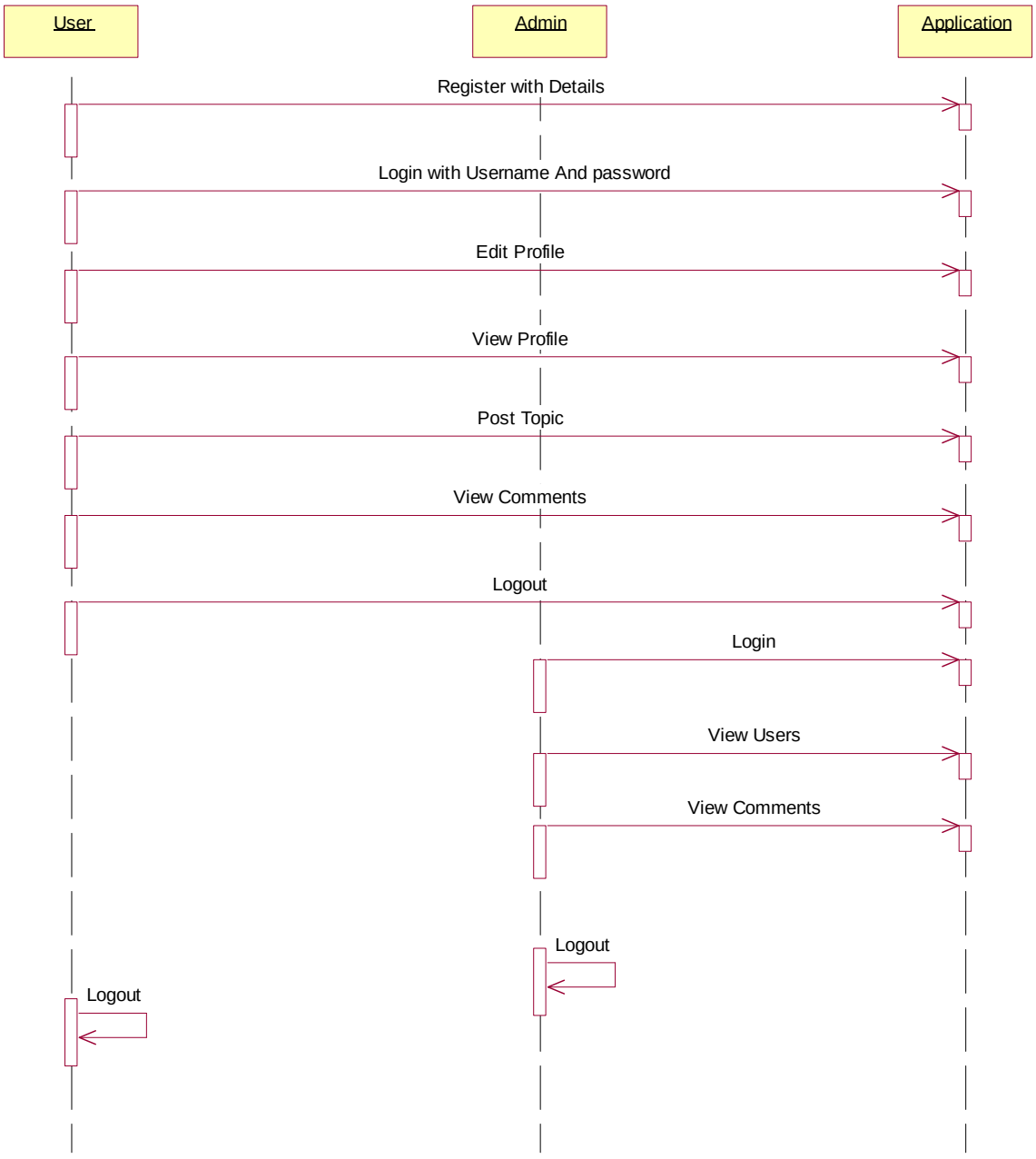


Figure 6: Sequence diagram

COLLABORATION DIAGRAM

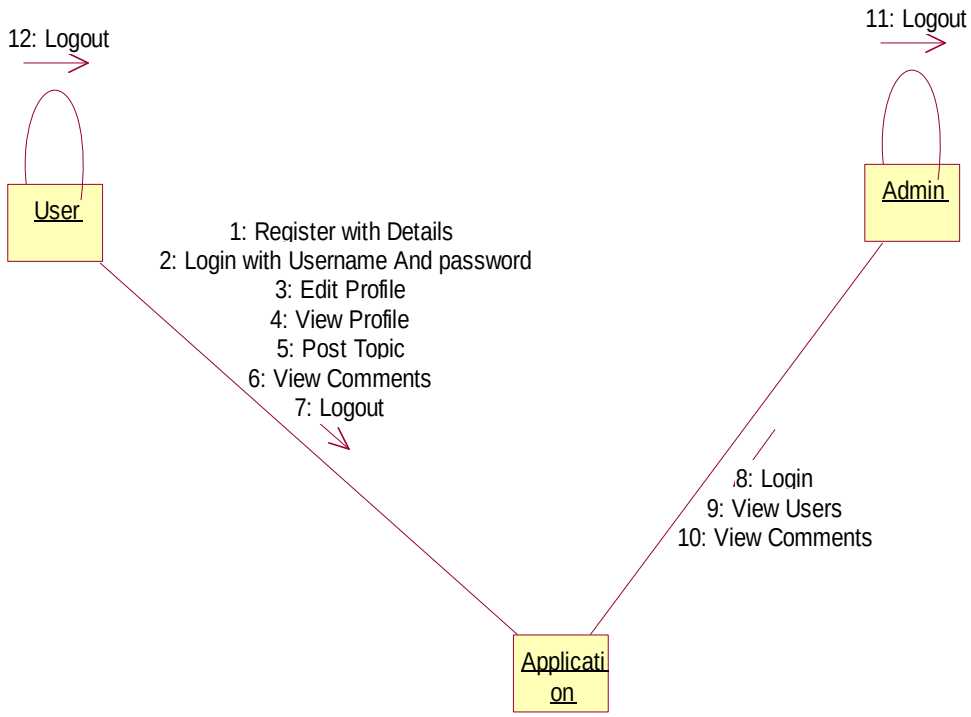


Figure 7: Collaboration diagram

6.PROJECT CODING

6.1 CODE TEMPLATE

Index page code

```
{% load static %}
<html>
<head>
<title>Opinion Mining</title>
<meta http-equiv="content-type" content="text/html; charset=utf-8" />
<link rel="stylesheet" type="text/css" href="{% static 'style.css' %}"/>
</head>
<body>
<div id="wrapper">
<div id="header">
<div id="logo">

<center><font size="4" color="yellow">Opinion Mining For Social Networking
Site</font></center>
</div>
<div id="slogan">

</div>
</div>
<div id="menu">
  <ul><center>
 <li><a href="{% url 'index' %}">Home</a></li>
      <li><a href="{% url 'Login' %}">Login</a></li>
      <li><a href="{% url 'Register' %}">Register Here</a></li>

</center></ul>
<br class="clearfix" />
</div>
```

```
<div id="splash">
<img class="pic" src="{% static 'images/investor.jpg' %}" width="870" height="230" alt=""
/>
</div>
<br/>
  <p align="justify"><font size="3" style="font-family: Comic Sans MS">
                                        Opinion Mining For Social Networking Site.

</p>
</body>
</html>
```

Login Page code

```
{% load static %}
<html>
<head>
<title>Opinion Mining</title>
<meta http-equiv="content-type" content="text/html; charset=utf-8" />
<link rel="stylesheet" type="text/css" href="{% static 'style.css' %}"/>
 <script LANGUAGE="Javascript" >
function validate(){
                var x=document.forms["f1"]["username"].value;
                var y=document.forms["f1"]["password"].value;

                        if(x == null || x==""){
                                window.alert("Username must be enter");
                                        document.f1.username.focus();
                                                return false;
                                }
                        if(y == null || y==""){
                                window.alert("Password must be enter");
                                        document.f1.password.focus();
                                                return false;
```

```
                                        }
                                    return true;
}
</script>
</head>
<body>
<div id="wrapper">
<div id="header">
<div id="logo">

<center><font size="4" color="yellow">Opinion Mining For Social Networking
Site</font></center>
 </div>
<div id="slogan">
</div>
</div>
<div id="menu">
<ul><center>
  <li><a href="{% url 'index' %}">Home</a></li>
 <li><a href="{% url 'Login' %}">Login</a></li>
 <li><a href="{% url 'Register' %}">Register Here</a></li>

 </center></ul>
<br class="clearfix" />
</div>
<div id="splash">
<img class="pic" src="{% static 'images/investor.jpg' %}" width="870" height="230" alt=""
/>
</div><center>
<form name="f1" method="post" action={% url 'UserLogin' %} OnSubmit="return
validate()">
                            {% csrf_token %}<br/>
   <h3><b>User Login Screen</b></h3>
```

```
<font size="" color="black"><center>{{ data }}</center></font>


<table align="center" width="80" >
<tr><td><b>Username</b></td><td><input type="text" name="username" style="font-family: Comic Sans MS" size="30"/></td></tr>

<tr><td><b>Password</b></td><td><input type="password" name="password" style="font-family: Comic Sans MS" size="30"/></td></tr>

<tr><td></td><td><input type="submit" value="Login">
</td>
</table>
        <br/><br/><br/><br/><br/><br/><br/><br/><br/>
</div>
</div>
</body>
</html>
```

Edit Profile code

```
{% load static %}
<html>
<head>
<title>Opinion Mining</title>
<meta http-equiv="content-type" content="text/html; charset=utf-8" />
<link rel="stylesheet" type="text/css" href="{% static 'style.css' %}"/>
 <script LANGUAGE="Javascript" >
function validate(){
                var x=document.forms["f1"]["username"].value;
                var y=document.forms["f1"]["password"].value;
                  var c=document.forms["f1"]["contact"].value;
                   var e=document.forms["f1"]["email"].value;
                  var a=document.forms["f1"]["address"].value;
```

```
var image=document.forms["f1"]["image"].value;
                if(x == null || x==""){
                                window.alert("Username must be enter");
                                                document.f1.username.focus();
                                                                return false;
                                }
                if(y == null || y==""){
                                window.alert("Password must be enter");
                                                document.f1.password.focus();
                                                                return false;
                                }
                if(c == null || c==""){
                                window.alert("Contact No must be enter");
                                                document.f1.contact.focus();
                                                                return false;
                                }
                    if(isNaN(c)){
                                window.alert("Please enter valid contact number");
                                                document.f1.contact.focus();
                                                                return false;
                                }
                if(e == null || e==""){
                                window.alert("Email ID must be enter");
                                                document.f1.email.focus();
                                                                return false;
                                }
var filter = /^([a-zA-Z0-9_\.\-])+\@(gmail+\.)+(com)+$/;
                  if (!filter.test(e)) {
                                window.alert('enter a valid email address');
                                                document.f1.email.focus();
                                                                return false;
                                }
                if(a == null || a==""){
```

```
                                        window.alert("Address must be enter");
                                                document.f1.address.focus();
                                                        return false;
                                                }
                                if(image == null || image==""){
                                                window.alert("profile image must be upload");
                                                        document.f1.image.focus();
                                                                return false;
                                                }
                                        return true;
}


                                        </script>
</head>
<body>
<div id="wrapper">
                                <div id="header">
                                                                <div id="logo">

<center><font size="4" color="yellow">Opinion Mining For Social Networking Site</font></center>
                                                                        </div>
                                                                <div id="slogan">

                                                                        </div>
                                        </div>
                                <div id="menu">
                                                                <ul><center>
    <li><a href="{% url 'HomePage' %}">Home</a></li>
            <li><a href="{% url 'EditProfile' %}">Edit Profile</a></li>
        <li><a href="{% url 'UpdateStatus' %}">Update Status</a></li>
   <li><a href="{% url 'ChangePassword' %}">Change Password</a></li>
```

```
            <li><a href="{% url 'PostTopic' %}">Post Topic</a></li>
      <li><a href="{% url 'index' %}">Logout</a></li>
                                                                    </center></ul>
<br class="clearfix" />

                                        </div>

                                    <div id="splash">
                                            <img class="pic" src="{% static 'images/investor.jpg'
%}" width="870" height="230" alt="" />
                                        </div>

                                        <center>
<form name="f1" method="post" action={% url 'EditMyProfile' %}
enctype="multipart/form-data" OnSubmit="return validate()">
                                    {% csrf_token %}<br/>
  <h3><b>Edit Profile Screen</b></h3>

  <table align="center" width="80" >

  {{ data|safe }}


<tr><td><b>Profile Image</b></td><td><input type="file" name="image"
style="font-family: Comic Sans MS" size="60"/></td></tr>
        <tr><td></td><td><input type="submit" value="Edit Profile">
</td>
</table>
 <br/><br/><br/><br/><br/><br/><br/><br/><br/>
</div>
</div>
</body>
</html>
```

Register page code

```
{% load static %}
<html>
<head>
<title>Opinion Mining</title>
<meta http-equiv="content-type" content="text/html; charset=utf-8" />
<link rel="stylesheet" type="text/css" href="{% static 'style.css' %}"/>
 <script LANGUAGE="Javascript" >
function validate(){
                var x=document.forms["f1"]["username"].value;
                var y=document.forms["f1"]["password"].value;
                 var c=document.forms["f1"]["contact"].value;
                  var e=document.forms["f1"]["email"].value;
                 var a=document.forms["f1"]["address"].value;
               var image=document.forms["f1"]["image"].value;
                              if(x == null || x==""){
                                                  window.alert("Username must be enter");
                                                                document.f1.username.focus();
                                                                                return false;
                                              }
                              if(y == null || y==""){
                                                  window.alert("Password must be enter");
                                                                document.f1.password.focus();
                                                                                return false;
                                              }
                              if(c == null || c==""){
                                                  window.alert("Contact No must be enter");
                                                                document.f1.contact.focus();
                                                                                return false;
                                              }
                                  if(isNaN(c)){
                                                  window.alert("Please enter valid contact number");
                                                                document.f1.contact.focus();
```

```
                                                    return false;
                                }
                        if(e == null || e==""){
                                        window.alert("Email ID must be enter");
                                                document.f1.email.focus();
                                                    return false;
                                }
            var filter = /^([a-zA-Z0-9_\.\-])+\@(gmail+\.)+(com)+$/;
                            if (!filter.test(e)) {
                                        window.alert('enter a valid email address');
                                                document.f1.email.focus();
                                                    return false;
                                }
                        if(a == null || a==""){
                                        window.alert("Address must be enter");
                                                document.f1.address.focus();
                                                    return false;
                                }
                    if(image == null || image==""){
                                        window.alert("profile image must be upload");
                                                document.f1.image.focus();
                                                    return false;
                                }
                                return true;
}


                                </script>
</head>
<body>
<div id="wrapper">
                        <div id="header">
                                                    <div id="logo">
```

```
<center><font size="4" color="yellow">Opinion Mining For Social Networking
Site</font></center>
                                                                    </div>
                                                                <div id="slogan">

                                                                    </div>
                                        </div>
                                    <div id="menu">
                                                                    <ul><center>
                <li><a href="{% url 'index' %}">Home</a></li>
    <li><a href="{% url 'Login' %}">Login</a></li>
    <li><a href="{% url 'Register' %}">Register Here</a></li>

                                                                    </center></ul>
<br class="clearfix" />

                                        </div>

                                    <div id="splash">
                                            <img class="pic" src="{% static 'images/investor.jpg'
%}" width="870" height="230" alt="" />
                                        </div>

                                    <center>
<form name="f1" method="post" action={% url 'Signup' %} enctype="multipart/form-data"
OnSubmit="return validate()">
                                {% csrf_token %}<br/>
  <h3><b>New User Signup Screen</b></h3>

      <font size="" color="black"><center>{{ data }}</center></font>
```

```
                    <table align="center" width="80" >

<tr><td><b>Username</b></td><td><input type="text" name="username" style="font-family: Comic Sans MS" size="30"/></td></tr>


                                        <tr><td><b>Password</b></td><td><input type="password" name="password" style="font-family: Comic Sans MS" size="30"/></td></tr>


                                    <tr><td><b>Contact No</b></td><td><input type="text" name="contact" style="font-family: Comic Sans MS" size="20"/></td></tr>

                                    <tr><td><b>Email ID</b></td><td><input type="text" name="email" style="font-family: Comic Sans MS" size="40"/></td></tr>

                                    <tr><td><b>Address</b></td><td><input type="text" name="address" style="font-family: Comic Sans MS" size="60"/></td></tr>

<tr><td><b>Profile Image</b></td><td><input type="file" name="image" style="font-family: Comic Sans MS" size="60"/></td></tr>
        <tr><td></td><td><input type="submit" value="Register">
</td>
 </table>
<br/><br/><br/><br/><br/><br/><br/><br/><br/>
</div>
</div>
</body>
</html>
```

6.2 OUTLINE FOR VARIOUS FILES

Post comment page code

```
{% load static %}
<html>
<head>
<title>Opinion Mining</title>
<meta http-equiv="content-type" content="text/html; charset=utf-8" />
<link rel="stylesheet" type="text/css" href="{% static 'style.css' %}"/>
</head>
<body>
<div id="wrapper">
                                        <div id="header">
                                                                        <div id="logo">

<center><font size="4" color="yellow">Opinion Mining For Social Networking
Site</font></center>
                                                                        </div>
                                                                    <div id="slogan">

                                                                        </div>
                                            </div>
                                        <div id="menu">
                                                                        <ul><center>
    <li><a href="{% url 'HomePage' %}">Home</a></li>
          <li><a href="{% url 'EditProfile' %}">Edit Profile</a></li>
        <li><a href="{% url 'UpdateStatus' %}">Update Status</a></li>
  <li><a href="{% url 'ChangePassword' %}">Change Password</a></li>
          <li><a href="{% url 'PostTopic' %}">Post Topic</a></li>
    <li><a href="{% url 'index' %}">Logout</a></li>
                                                                        </center></ul>
                                                                    <br class="clearfix" />
                                            </div>
                                        <div id="splash">
```

```
                                        <img class="pic" src="{% static 'images/investor.jpg'
%}" width="870" height="230" alt="" />
                                    </div>
<form name="f1" method="post" action={% url 'PostMyComment' %} OnSubmit="return
validate()">
                                {% csrf_token %}<br/>
                                    {{ data|safe }}


</body>
</html>
```

Post Topic Page

```
{% load static %}
<html>
<head>
<title>Opinion Mining</title>
<meta http-equiv="content-type" content="text/html; charset=utf-8" />
<link rel="stylesheet" type="text/css" href="{% static 'style.css' %}"/>
 <script LANGUAGE="Javascript" >
function validate(){
                    var x=document.forms["f1"]["name"].value;
                     var y=document.forms["f1"]["topic"].value;
                  var c=document.forms["f1"]["description"].value;
                    var e=document.forms["f1"]["image"].value;

                                if(x == null || x==""){
                                                window.alert("name must be enter");
                                                        document.f1.name.focus();
                                                                return false;
                                        }
                                if(y == null || y==""){
                                                window.alert("Topic must be enter");
                                                        document.f1.topic.focus();
                                                                return false;
```

```
                }
            if(c == null || c==""){
                window.alert("Description No must be enter");
                    document.f1.description.focus();
                        return false;
                }

            if(e == null || e==""){
                window.alert("post image must be upload");
                    document.f1.image.focus();
                        return false;
                }
            return true;
}


                </script>
</head>
<body>
<div id="wrapper">
                <div id="header">
                        <div id="logo">

<center><font size="4" color="yellow">Opinion Mining For Social Networking
Site</font></center>
                        </div>
                    <div id="slogan">

                        </div>
                </div>
                <div id="menu">
                        <ul><center>
    <li><a href="{% url 'HomePage' %}">Home</a></li>
```

```
            <li><a href="{% url 'EditProfile' %}">Edit Profile</a></li>
          <li><a href="{% url 'UpdateStatus' %}">Update Status</a></li>
    <li><a href="{% url 'ChangePassword' %}">Change Password</a></li>
            <li><a href="{% url 'PostTopic' %}">Post Topic</a></li>
      <li><a href="{% url 'index' %}">Logout</a></li>
                                                                    </center></ul>
<br class="clearfix" />

                                        </div>

                                    <div id="splash">
                                            <img class="pic" src="{% static 'images/investor.jpg'
%}" width="870" height="230" alt="" />
                                        </div>

                                        <center>
<form name="f1" method="post" action={% url 'PostMyTopic' %} enctype="multipart/form-
data" OnSubmit="return validate()">
                                {% csrf_token %}<br/>
   <h3><b>Post Topic Screen</b></h3>

        <font size="" color="black"><center>{{ data }}</center></font>
<table align="center" width="80" >

<tr><td><b>Name</b></td><td><input type="text" name="name" style="font-family:
Comic Sans MS" size="30"/></td></tr>


                                        <tr><td><b>Topic</b></td><td><input type="text"
name="topic" style="font-family: Comic Sans MS" size="40"/></td></tr>
```

<tr><td>Description</td><td><input type="text" name="description" style="font-family: Comic Sans MS" size="60"/></td></tr>

<tr><td>Post Image</td><td><input type="file" name="image" style="font-family: Comic Sans MS" size="60"/></td></tr>

<tr><td></td><td><input type="submit" value="Submit">
</td>
</table>

</body>
</html>

Update status page

```
{% load static %}
<html>
<head>
<title>Opinion Mining</title>
<meta http-equiv="content-type" content="text/html; charset=utf-8" />
<link rel="stylesheet" type="text/css" href="{% static 'style.css' %}"/>
 <script LANGUAGE="Javascript" >
function validate(){
                var x=document.forms["f1"]["status"].value;

                        if(x == null || x==""){
                                        window.alert("Status must be enter");
                                                document.f1.status.focus();
                                                        return false;
                        }
```

```
                return true;
}


                </script>
</head>
<body>
<div id="wrapper">
                <div id="header">
                                <div id="logo">

<center><font size="4" color="yellow">Opinion Mining For Social Networking Site</font></center>
                                </div>
                                <div id="slogan">

                                </div>
                </div>
                <div id="menu">
                                <ul><center>
    <li><a href="{% url 'HomePage' %}">Home</a></li>
        <li><a href="{% url 'EditProfile' %}">Edit Profile</a></li>
      <li><a href="{% url 'UpdateStatus' %}">Update Status</a></li>
  <li><a href="{% url 'ChangePassword' %}">Change Password</a></li>
         <li><a href="{% url 'PostTopic' %}">Post Topic</a></li>
    <li><a href="{% url 'index' %}">Logout</a></li>
                                </center></ul>
<br class="clearfix" />

                </div>

                <div id="splash">
```

```
                                        <img class="pic" src="{% static 'images/investor.jpg'
%}" width="870" height="230" alt="" />
                                        </div>
                                        <center>
<form name="f1" method="post" action={% url 'UpdateMyStatus' %} OnSubmit="return
validate()">
                                {% csrf_token %}<br/>
  <h3><b>Status Update Screen</b></h3>

      <font size="" color="black"><center>{{ data }}</center></font>
<table align="center" width="80" >

<tr><td><b>Status</b></td><td><input type="text" name="status" style="font-family:
Comic Sans MS" size="30"/></td></tr>
      <tr><td></td><td><input type="submit" value="Update Status">
</td>
</table>

          <br/><br/><br/><br/><br/><br/><br/><br/><br/><br/>
</div>

</div>
</body>
</html>
```

6.3CLASS WITH FUNCTIONALITY

Settings code:

```
import os

# Build paths inside the project like this: os.path.join(BASE_DIR, ...)
BASE_DIR = os.path.dirname(os.path.dirname(os.path.abspath(__file__)))



# Quick-start development settings - unsuitable for production
```

```
# See https://docs.djangoproject.com/en/2.1/howto/deployment/checklist/

# SECURITY WARNING: keep the secret key used in production secret!
SECRET_KEY = 'yh88d6ur$x-xu3lgduzeypz-@5&v6l*&x*_47sm#te0$!fuzlx'

# SECURITY WARNING: don't run with debug turned on in production!
DEBUG = True

ALLOWED_HOSTS = []


# Application definition

INSTALLED_APPS = [
    'django.contrib.admin',
    'django.contrib.auth',
    'django.contrib.contenttypes',
    'django.contrib.sessions',
    'django.contrib.messages',
    'django.contrib.staticfiles',
    'OpinionApp'
]

MIDDLEWARE = [
    'django.middleware.security.SecurityMiddleware',
    'django.contrib.sessions.middleware.SessionMiddleware',
    'django.middleware.common.CommonMiddleware',
    'django.middleware.csrf.CsrfViewMiddleware',
    'django.contrib.auth.middleware.AuthenticationMiddleware',
    'django.contrib.messages.middleware.MessageMiddleware',
    'django.middleware.clickjacking.XFrameOptionsMiddleware',
]
```

```
ROOT_URLCONF = 'Opinion.urls'

TEMPLATES = [
    {
        'BACKEND': 'django.template.backends.django.DjangoTemplates',
        'DIRS': [
                                            os.path.join('C:/Python/OpinionMining/OpinionApp',
'templates'),
                                            ],
        'APP_DIRS': True,
        'OPTIONS': {
            'context_processors': [
                'django.template.context_processors.debug',
                'django.template.context_processors.request',
                'django.contrib.auth.context_processors.auth',
                'django.contrib.messages.context_processors.messages',
            ],
        },
    },
]

WSGI_APPLICATION = 'Opinion.wsgi.application'


# Database
# https://docs.djangoproject.com/en/2.1/ref/settings/#databases

DATABASES = {
    'default': {
        'ENGINE': 'django.db.backends.mysql',
        'NAME': 'OpinionMining',
        'HOST': '127.0.0.1',
        'PORT': '3306',
```

```
        'USER': 'root',
        'PASSWORD': '',
                                        'OPTIONS': {
          'autocommit': True,
        },
    }
}


# Password validation
# https://docs.djangoproject.com/en/2.1/ref/settings/#auth-password-validators

AUTH_PASSWORD_VALIDATORS = [
    {
        'NAME': 'django.contrib.auth.password_validation.UserAttributeSimilarityValidator',
    },
    {
        'NAME': 'django.contrib.auth.password_validation.MinimumLengthValidator',
    },
    {
        'NAME': 'django.contrib.auth.password_validation.CommonPasswordValidator',
    },
    {
        'NAME': 'django.contrib.auth.password_validation.NumericPasswordValidator',
    },
]


# Internationalization
# https://docs.djangoproject.com/en/2.1/topics/i18n/

LANGUAGE_CODE = 'en-us'
```

```
TIME_ZONE = 'UTC'

USE_I18N = True

USE_L10N = True

USE_TZ = True

# Static files (CSS, JavaScript, Images)
# https://docs.djangoproject.com/en/2.1/howto/static-files/

STATIC_URL = '/static/'
```

URL Code

```
from django.contrib import admin
from django.urls import path, include

urlpatterns = [
    path('admin/', admin.site.urls),
    path('', include('OpinionApp.urls')),
]
```

WSGI code

```
import os

from django.core.wsgi import get_wsgi_application

os.environ.setdefault('DJANGO_SETTINGS_MODULE', 'Opinion.settings')

application = get_wsgi_application()
```

6.4 METHODS OF INPUT AND OUTPUT PARAMETERS

INPUT DESIGN

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the following things:

Design considered the following things:

- What data should be given as input?
- How the data should be arranged or coded?
- The dialog to guide the operating personnel in providing input.

Methods for preparing inp ut validations and steps to follow when error occur.

OBJECTIVES

1. Input Design is the process of converting a user-oriented description of the input into a computer-based system. This design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerized system.

2. It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The goal of designing input is to make data entry easier and to be free from errors. The data entry screen is designed in such a way that all the data manipulates can be performed. It also provides record viewing facilities.

3. When the data is entered it will check for its validity. Data can be entered with the help of screens. Appropriate messages are provided as when needed so that the user will not be in maize of instant. Thus the objective of input design is to create an input layout that is easy to follow

OUTPUT DESIGN

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making.

Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is designed so that people will find the system can use easily and effectively. When analysis design computer output, they should Identify the specific output that is needed to meet the requirements.

Select methods for presenting information.

Create document, report, or other formats that contain information produced by the system.

The output form of an information system should accomplish one or more of the following objectives.

- Convey information about past activities, current status or projections of the
- Future.
- Signal important events, opportunities, problems, or warnings.
- Trigger an action.

7. PROJECT TESTING

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the

Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

7.1 VARIOUS TESTCASES

Unit testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

Integration testing

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

Functional test

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items:

Valid Input : identified classes of valid input must be accepted.

Invalid Input : identified classes of invalid input must be rejected.

Functions : identified functions must be exercised.

Output : identified classes of application outputs must be exercised.

Systems/Procedures: interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for

testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

System Test

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

7.2 BLACKBOX TESTING

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box .you cannot “see” into it. The test provides inputs and responds to outputs without considering how the software works.

7.3 WHITEBOX TESTING

White Box Testing is a testing in which in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is purpose. It is used to test areas that cannot be reached from a black box level.

Unit Testing:

Unit testing is usually conducted as part of a combined code and unit test phase of the software lifecycle, although it is not uncommon for coding and unit testing to be conducted as two distinct phases.

Test strategy and approach

Field testing will be performed manually and functional tests will be written in detail.

Test objectives

- All field entries must work properly.
- Pages must be activated from the identified link.
- The entry screen, messages and responses must not be delayed.

Features to be tested

- Verify that the entries are of the correct format
- No duplicate entries should be allowed
- All links should take the user to the correct page.

Integration Testing

Software integration testing is the incremental integration testing of two or more integrated software components on a single platform to produce failures caused by interface defects.

The task of the integration test is to check that components or software applications, e.g. components in a software system or – one step up – software applications at the company level – interact without error.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

Acceptance Testing

User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

8. OUTPUT SCREENS

8.1 USER INTERFACES

In this project users can register with the application and then login and post their topics and then other users may view that post topic and write their description/comment. User's description/comment will rate using SVM classifier.

Create database by copying content from 'DB.txt' file and paste in MYSQL

Create 'Python' folder in C directory and then put 'OpinionMining' folder inside C:/Python folder and then start django server and run code in browser by using below URL

http://127.0.0.1:8000/index.html and press enter key to get below screen

In above screen click on ‘Register Here’ link to add new user

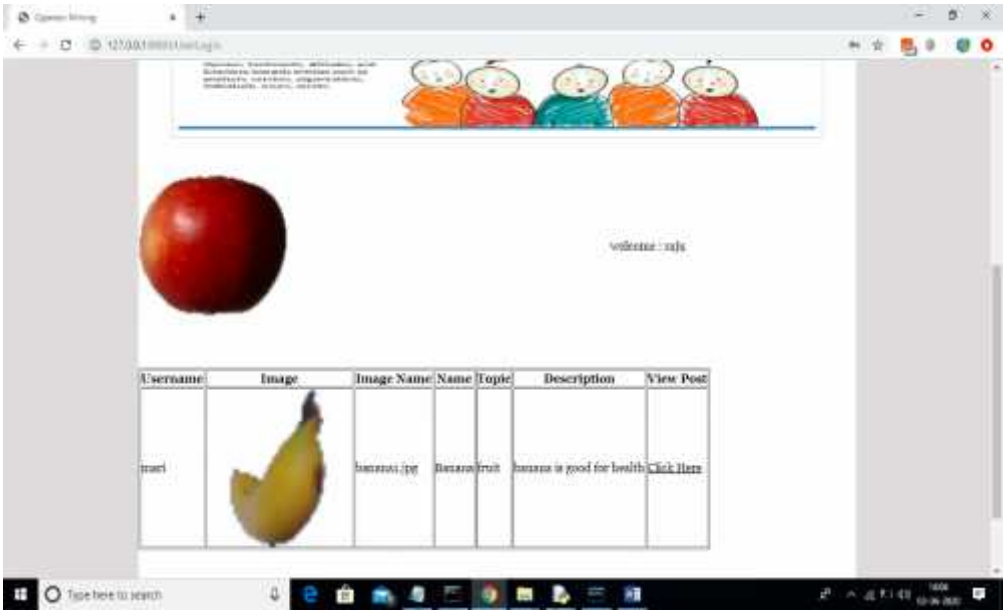
In above screen creating user profile with uploading image and then click on ‘Open’ button and then click on ‘Register’ button to complete registration process

In above screen signup process completed and login as raju user to upload post topic

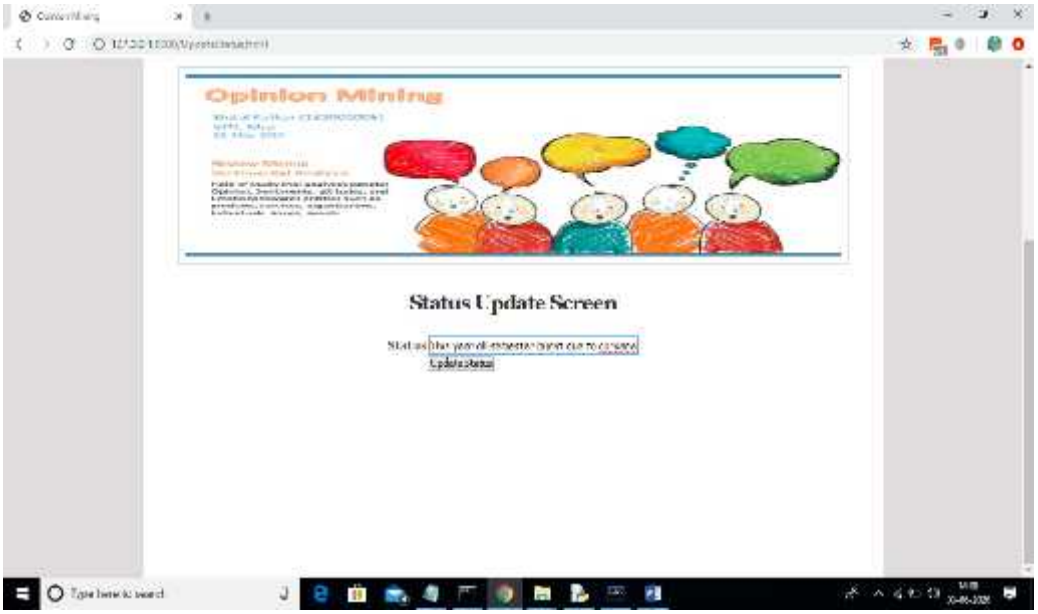
8.2 OUTPUT SCREENS

After login will get below screen

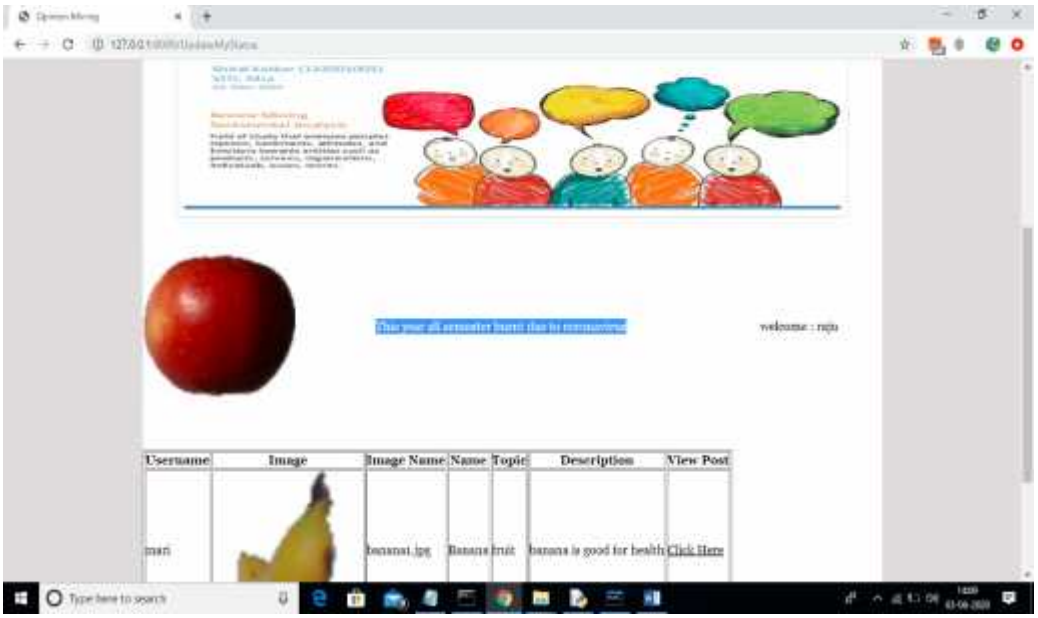
In above screen by clicking on Home Page user will get below screen and by clicking on ‘Edit Profile’ user may change profile image and other details and by using ‘Update Status’ link user can update his status text and by using Post Topic link he will post new topic.



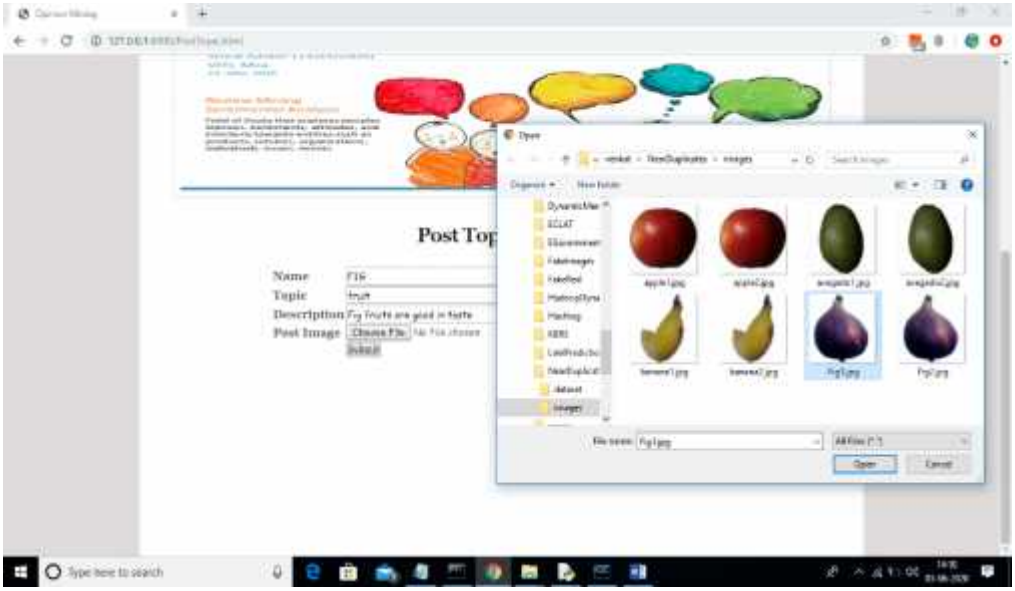
In above HomePage screen we can see raju user profile image and his name and below we can see all topics posted by different users. Now click on ‘status update’ link to add status



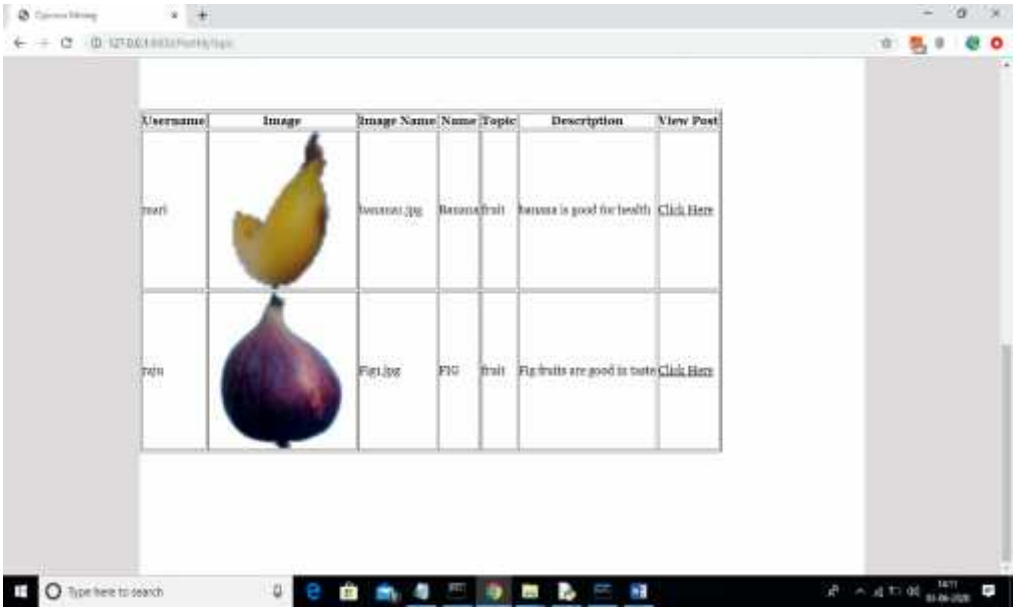
In above screen I am adding some status and now click on ‘Update Status’ button to get below screen



In above screen selected text we can see updated user status. Now user can post topic by clicking on post topic link



In above screen user is uploading some topics and now click ‘Open’ button and then press submit button to upload that topic



In this screen we can see 2 topics from 2 users mari and raju. Now both can comments on their topic. Now I will logout and login as mari and comment on raju topic. To comment on any topic click on ‘Click Here’ option to get below screen

In above screen writing come comments and press submit button to get below screen

In above screen mari wrote some comments on raju post and then using SVM classifier we rate each comment.

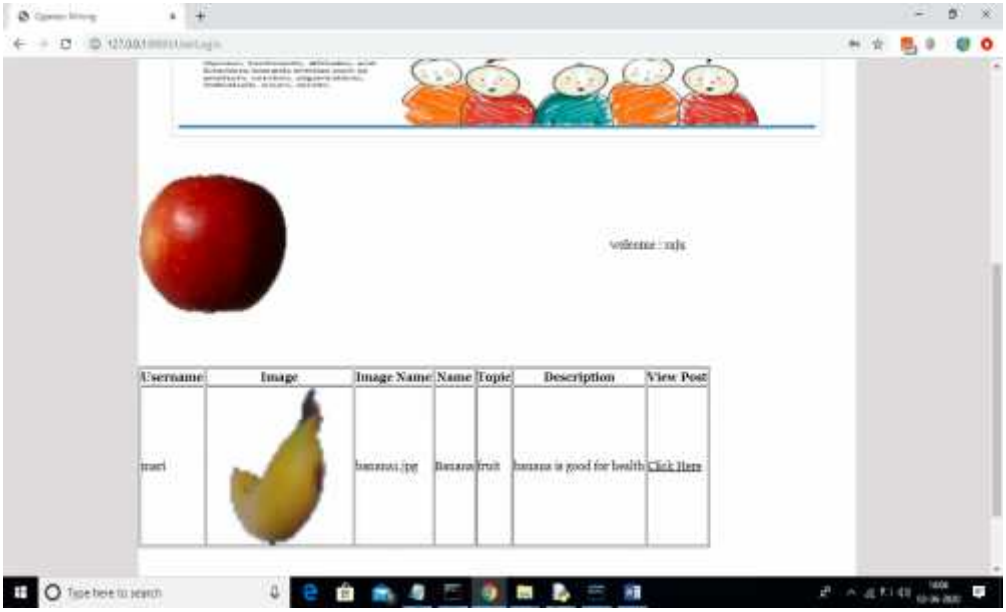
Note: after starting server each time SVM classifier will take some loading time for first commentand from next comment it will run faster.

After editing profile picture then it will get update from next login as browser maintain its cache.

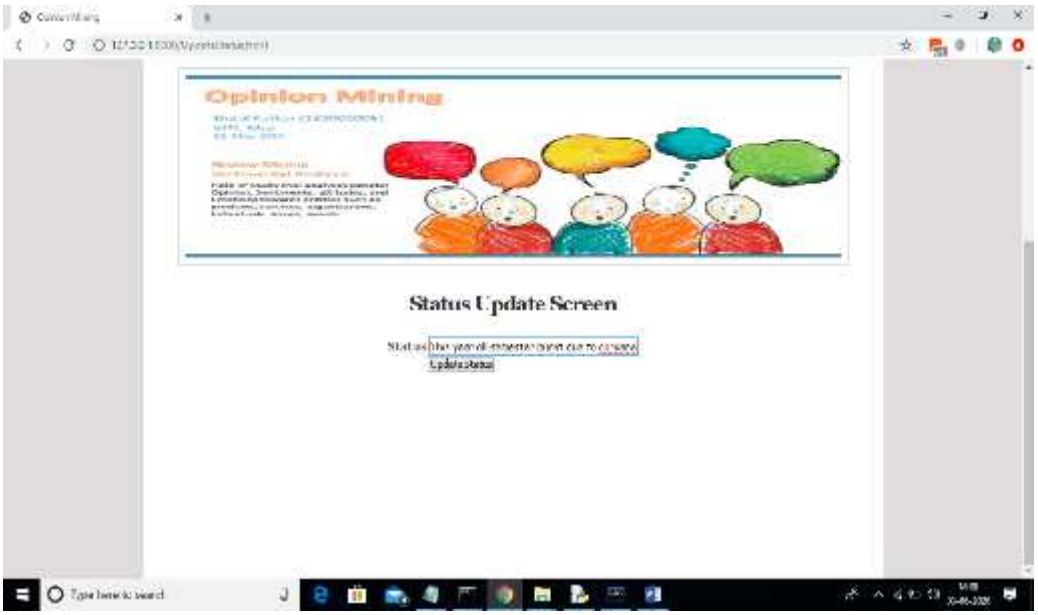
9.EXPERIMENTAL RESULTS

After login will get below screen

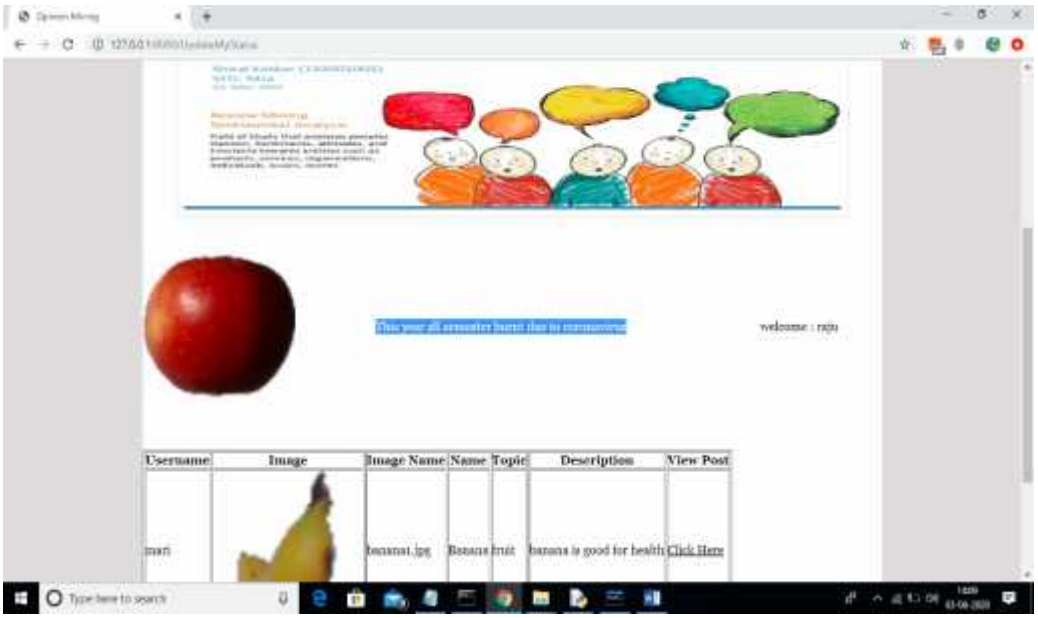
In above screen by clicking on Home Page user will get below screen and by clicking on 'Edit Profile' user may change profile image and other details and by using 'Update Status' link user can update his status text and by using Post Topic link he will post new topic.



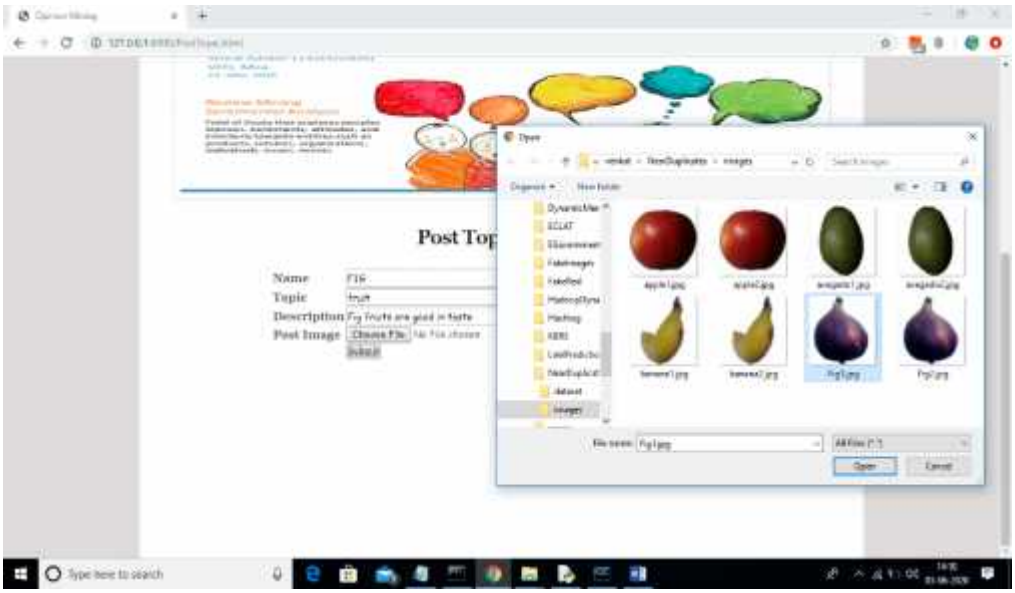
In above HomePage screen we can see raju user profile image and his name and below we can see all topics posted by different users. Now click on 'status update' link to add status



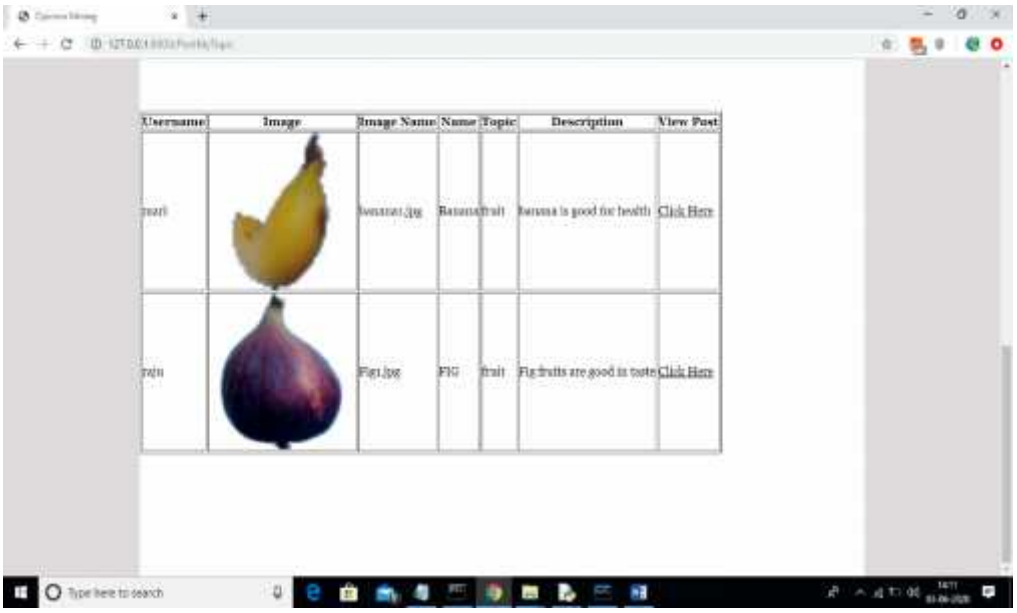
In above screen I am adding some status and now click on 'Update Status' button to get below screen



In above screen selected text we can see updated user status. Now user can post topic by clicking on post topic link



In above screen user is uploading some topics and now click ‘Open’ button and then press submit button to upload that topic



In this screen we can see 2 topics from 2 users mari and raju. Now both can comments on their topic. Now I will logout and login as mari and comment on raju topic. To comment on any topic click on 'Click Here' option to get below screen

In above screen writing come comments and press submit button to get below screen

In above screen mari wrote some comments on raju post and then using SVM classifier we rate each comment.

Note: after starting server each time SVM classifier will take some loading time for first commentand from next comment it will run faster.

After editing profile picture then it will get update from next login as browser maintain its cache.

10. CONCLUSION AND FUTURE ENHANCEMENT

In this paper, a Student Feedback Mining System is build to analyze topics and their sentiments from student generated feedback. This system uses preprocessing, topic extraction, clustering, classification to represent the student views in a graphical way. This system will be useful to improve the students learning and instructor's methods of delivery.

Thus, Opinion Mining and Sentiment analysis has wide area of applications and it also facing many research challenges. Since the fast growth of internet and internet related applications, the Opinion Mining and Sentiment Analysis become most interesting research area among natural language processing community. A more innovative and effective techniques required to be invented which should overcome the current challenges faced by Opinion Mining.

11. REFERENCES

[1]. Abdulla, M.H. The use of an online student response system to support learning of physiology during lectures to medical students. Education and Information Technologies (2018)

[2]. Balahadia, F.F., Fernando, M.C.G., &Juanatas, I.C. Teacher's performance evaluatio IEEE region 10 symposium (2016)

[3]. Brennan, J.& Williams, R. (2010) Collecting and Using Student Feedback. A Guide to Good Practice (LTSN,York)

[4]. Dhanalakshmi V in Opinion mining from student feedback data using supervised learning algorithms.

[5].Donovan, J.,Mader, C.E., &Shinsky, J.(2008). Constructive Student Feedback: Online vs. Traditional Course Evaluations.

[6]. Elaine Keane & Iain Mac Labhrainn, Obtaining Student Feedback on Teaching & Course Quality, CELT, April 2005

[7]. Francis F.Balahadia; Ma.Corazon G.Fernando; Irish C.Juanatas in Teacher's performance evaluation tool using opinion mining with sentiment analysis

[8].GokarnIlaNitin,Asst.Prof.Gottipati Swapna, Prof.VenkyShankararaman in Analyzing Educational Comments for Topics and Sentiments: A Text Analytics Approach.

[9].HarshaliP.Patil,Mohammad Atiquein Sentiment analysis for Social media.

[10].K.P.Mohanan, the place of student feedback in teaching evaluation http://www.cdtl.nus.edu.sg/publications/studfeedback/StudFeedback_Teach Quality.pdf

[11]. Mark McGuire,ConstanceKampf Aarhus University in Using Social Media Sentiment Analysis to Understand Audiences: A New Skill for Technical Communicators?

[12].M.S.Neethu,R.Rajasree in Sentiment analysis in twitter using machine learning techniques.

[13].NabeelaAltrabsheh; Mihaela Cocea; SanazFallahkhair in Sentiment Analysis: Towards a Tool for Analysing RealTime Students Feedback

[14]. Tan Li Im, PhangWai San, Chin Kim On Center of Excellence in Semantic Agents Universities Malaysia in Rule-based Sentiment Analysis for Financial News

[15].Yao, Y., & Grady, M.L. (2005). How do faculty make formative use of student evaluation feedback?: A multiple case study. Journal of Personnel Evaluation in Education,

18(2), 107-126. 16. Zhao, Y.,Karypis, G., & Du, D.Z.(2005). Criterion functions for document clustering (Doctoral dissertation, University of Minnesota.).

[16].Zhao, Y., Karypis, G., & Du, D. Z. (2005). Criterion functions for document clustering (Doctoral dissertation, University of Minnesota.).

[17].Object Refinery Limited. "JFreeChart". JFreeChart. Jan 2004. Web. 17 Apr. 2004 http://www.jfree.org/jfreechart/

[18].The Apache Software Foundation."Apache Poi Project". Apache POI.Jan 2003. Web. 17 Apr. 2003. https://poi.apache.org

12.PUBLICATIONS

International Conference on “Innovations in Computers Networks, Computational Intelligence and IOT” (ICICCI-21)

Paper ID: ICICCI-21-0058

PROJECT CONTRIBUTION

	Name of the student	Contribution
Planning	G.Suprabath Reddy, B.Ajay, E. Mahendra, MD. Faisal	Analysing the project idea, Creating project schedule, Defining project timelines, roles and responsibilities.
Requirements	G .Suprabath Reddy, B. Ajay	Gathering necessary input datasets, pictures, analysing software and hardware requirements.
Design	E .Mahendra, MD. Faisal	UML Diagrams, basic prototype.
Development	G. Suprabath Reddy B.Ajay E.Mahendra MD.Faisal	1.Front end development-CSS 2.Front end development HTML 3. Database Management 4. Implementing ML Algorithms

G.Suprabath Reddy is currently pursuing her Bachelor of Technology with specialization in Computer Science Engineering at St. Martin‟s Engineering College. He completed his completed his intermediate from Sri Gayatri junior college and 10th from Pallavi model school. His technical skills include C, C++, Python and Java. He worked with Introduction and Programming with IOT Board Intern for 2 months and also participated in the online training provided by AWS Fundamentals: Going Cloud•Native. Participated in Managing Project Risks and Changes conducted by University of COURSE California, Irvine , Software testing for begginers by Guru99 and also HTML by EJ Media, He spends his free time taking online certification courses related to his field of study such as Word by GCFLearnFree , as well as personal interests from platform such as Coursera and Cursa.

Mohammed Faisal is currently pursuing his Bachelor of Technology with specialisation in computer science Engineering at St.Martin Engineering Collage.he completed his 12th from Sri Chaitanya junior collage and 10th from Nishant high school.His technical skills include Aws,python,mysql,managing project risk and changes ,machine learning and data analyst.Also have a basic level of knowledge about C and C++.he trained in aws devops of all tools like Linux,git, GitHub ,chef,ansible,docker,kubernet. Participated in HTML and CSS workshop from 5 January to 3rd February 2018 organised by Tam.

He is also participated in street cause for people help by providing there needs. He completed few certifications courses from online platforms like udemy, coursera, coursaApp.

E.Mahendra is currently pursuing her Bachelor of Technology with specialization in Computer Science Engineering at St. Martin's Engineering College. He completed his 12th class from Sri chaithanya junior college and 10th class from St.Anns High School . His technical skills include AWS, Digital marketing, Machine Learning and Deep Learning. Also has an intermediary knowledge about C and Java. He worked with DATAI Analytics India Pvt Ltd. as a Data Science and Machine Learning Intern for 6 months from Nov 1, 2020 to April 30, 2021 and also participated in the online training provided by IIT Khanpur. He worked as a volunteer during 2017-18 in Street Cause SMEC division. Participated in Machine Learning workshop conducted on 8th and 9th May 2019 by TAM and also in National Level Project Expo and Competition "Technovation-2018" organized by Mechanical and Computer Science department of SMEC on 28th March 2018. His participations include: Online Two Day National Level Seminar on "Recent Trends in Cloud Computing Fog and Edge Computing" from 18th to 19th June, 2021. He completed few certification courses from online platforms like Udemy, Cousera, CursaApp.

B. Ajay is currently pursuing his Bachelor of Technology with specialization in Computer Science Engineering at St. Martin’s Engineering College. He completed his 12th class from Narayana Junior College and 10th class from Geethanjali High school. His technical skills include Python, MySQL,HTML & CSS. Also has an intermediary knowledge about C, C++ and Java. and also participated in 5- days online International Hands on Certification Training in 'Python Programming' conducted by St Martin's Engineering College in association with LEBANESE FRENCH UNIVERSITY (Iraq) from 20th to 24th August 2020. and also Participated in Machine Learning workshop conducted on 8th and 9th May 2019 by TAM. His main interests include Web Development and Python Programming. He completed few certification courses from online platforms like Cousera,Cursa, Edapt, Sololearn etc.

APPENDICES

DATABASE

```
create database opinionmining;
use opinionmining;

create table register(username varchar(30) primary key,
password varchar(30),
contact varchar(12),
email varchar(30),
address varchar(40),
status varchar(200));

create table post(username varchar(30),
post_id varchar(50),
image varchar(100),
name varchar(100),
topic varchar(100),
description varchar(100));

create table comment(post_id varchar(30),
username varchar(30),
comment varchar(200),
rate varchar(10));
```

Styling template

```
/*
CSS Credit: http://www.templatemo.com
*/

body {
    margin: 0;
    padding: 0;
    line-height: 1.5em;
    font-family: Arial, Helvetica, sans-serif;
    font-size: 15px;
    color:black;
    background: #afa87d;
}

a:link, a:visited { color: #d46528; text-decoration: none; font-weight: bold; }
a:active, a:hover { color: #2da3e9; }

p {
    margin: 0px;
    padding: 0px;
}

img {
    margin: 0px;
    padding: 0px;
    border: none;
}
```

```
.cleaner { clear: both; width: 100%; height: 0px; font-size: 0px;  }

.margin_bottom_10 { float: left; width: 100%; height: 10px; font-size: 1px;  }
.margin_bottom_20 { clear: both; width: 120%; height: 25px; font-size: 19px;  }
.margin_bottom_30 { clear: both; width: 100%; height: 30px; font-size: 1px;  }
.margin_bottom_40 { clear: both; width: 100%; height: 40px; font-size: 1px;  }
.margin_bottom_50 { clear: both; width: 100%; height: 50px; font-size: 1px;  }
.margin_bottom_60 { clear: both; width: 100%; height: 60px; font-size: 1px;  }

.margin_right_10 { margin-right: 10px; }
.margin_right_20 { margin-right: 20px; }
.margin_right_50 { margin-right: 50px; }

.border_bottom {
                                                    border-bottom: 1px
solid #CCC;
}

#templatemo_container {
                                                    width: 1040px;
                                                    margin: 0 auto;
}

#templatemo_header {
                                                    width: 1040px;
                                                    height: 250px;
                                                    background:
url(images/templatemo_header_bg.jpg) no-repeat;
}

#templatemo_header #site_logo {
```

```
                                                                    float: left;
                                                                    margin: 80px 0 0
40px;
                                                                    width: 410px;
                                                                    height: 75px;
                                                                    background:url(image
s/logo1.png) bottom  no-repeat;
}

.rc_btn_01 a{
                                                                    float: right;
                                                                    clear: both;
                                                                    display: block;
                                                                    width: 80px;
                                                                    height: 15px;
                                                                    text-align: center;
                                                                    padding: 10px 0 10px
10px;
                                                                    background:
url(images/templatemo_buttom_01.jpg) bottom right no-repeat;
                                                                    color: #d46528;
                                                                    font-weight: bold;
                                                                    text-decoration: none;
}
/* menu */

#templatemo_menu {
                                                                    clear: both;
                                                                    width: 920px;
                                                                    height: 50px;
                                                                    padding: 0 80px 0
40px;
```

```
                                                                    background:
url(images/templatemo_menu_bg.jpg) no-repeat;
}

#templatemo_menu ul {
                                                                    margin: 0px;
                                                                    padding: 0px;
                                                                    list-style: none;
}

#templatemo_menu ul li {
                                                                    display: inline;
}

#templatemo_menu ul li a {
                                                                    float: left;
                                                                    padding: 20px 40px 0
0;
                                                                    text-align: center;
                                                                    font-size: 12px;
                                                                    text-align: center;
                                                                    text-decoration: none;
                                                                    color: #2aa3e8;
                                                                    font-weight: bold;
                                                                    outline: none;
}

#templatemo_menu li a:hover, #templatemo_menu li .current {
                                                                    color: #000000;
}
```

```
#templatemo_menu li .last {
                                                    background: none;
}

/* end of menu*/

/* content */

.header_01 {
                                                    font-size: 20px;
                                                    padding-bottom:
10px;
                                                    margin-bottom: 20px;
                                                    font-weight: bold;
                                                    color: #d46528;
}

#templatemo_content {
                                                    clear: both;
                                                    width: 920px;
                                                    padding: 10px 80px
30px 40px;
                                                    background:
url(images/templatemo_content_bg.jpg) repeat-y;
}

#templatemo_content #content_left {
                                                    float: left;
                                                    width: 600px;
                                                    padding: 20px 0 0 0;
                                                    /* background:
#a4ddfe; */
```

```
}

#templatemo_content #content_right {
                                        float: right;
                                        width: 270px;
}

#content_left .left_column_section {
                                        margin: 0;
                                        padding: 0;
}

.left_column_section p {
                                        text-align: justify;
                                        margin-bottom: 10px;
}

.image_box {
                                        float: left;
                                        width: 280px;
                                        height: 120px;
                                        background: #ffffff;
                                        border: 1px solid
#999;
                                        padding: 4px;

}

.image_box img {
                                        width: 280px;
                                        height: 120px;
```

```
}

#content_right .right_column_section {
                                                    clear: both;
                                                    position: relative;
                                                    background:
url(images/templatemo_section_01_content_bg.jpg) repeat-y;
}

.right_column_section .header_02 {
                                                    width: 270px;
                                                    height: 30px;
                                                    font-size: 16px;
                                                    font-weight: bold;
                                                    padding: 40px 0 0 0;
                                                    text-align: center;
                                                    background:
url(images/templatemo_section_01_header_bg.jpg) no-repeat;
}


.right_column_section  .header_03 {
                                                    font-size: 12px;
                                                    margin-bottom: 5px;
                                                    font-weight: bold;
                                                    color: #333333;
}

.right_column_section .content {
                                                    padding: 10px 20px 0
20px;
```

```
                                                    background:
url(images/templatemo_section_01_content_bg.jpg) repeat-y;
}

.right_column_section span {
                                                    position: absolute;
                                                    width: 270px;
                                                    height: 60px;
                                                    background:
url(images/templatemo_section_01_bottom_bg.jpg);
}

/* bottom panel */

#templatemo_bottom_panel {
                                                    clear: both;
                                                    width: 600px;
                                                    height: 210px;
                                                    padding: 0px 400px
30px 40px;
                                                    background:
url(images/templatemo_bottom_panel_bg.jpg) no-repeat;
}

.content_panel_section {
                                                    float: left;
                                                    width: 275px;
}

.content_panel_section ul {
                                                    margin: 0px;
                                                    padding: 0px;
```

```
                                                    list-style: none;
}

.content_panel_4_col li {
                                                    margin: 0px;
                                                    padding: 0px;
                                                    color: #2da3e9;
                                                    padding-bottom: 5px;
                                                    margin-bottom: 5px;
                                                    border-bottom: 1px
solid #CCC;
}

.content_panel_4_col li a {
                                                    color: #2da3e9;
}

.content_panel_4_col li a:hover {
                                                    color: #d46528;
}

.content_panel_4_col li span {
                                                    clear: both;
                                                    display: block;
                                                    color: #333;
                                                    font-weight: normal;
}


/* end of bottom panel*/
```

```
/* footer */
#templatemo_footer {
                                                                clear: both;
                                                                width: 920px;
                                                                padding: 20px 80px
30px 40px;
                                                                text-align: left;
                                                                color: #000;
                                                                background: #afa87d
url(images/templatemo_footer.jpg) top center  no-repeat;
}

#templatemo_footer a{
                                                                color: #000;
                                                                font-weight: bold;
}
```

A

PROJECT REPORT

ON

CHARACTERIZING AND PREDICTING EARLY REVIEWERS FOR EFFECTIVE PRODUCT MARKETING ON E-COMMERCE WEBSITES

Submitted by

1) Mr. Neti Rohit Kumar (16K81A05F5)

2) Mr. Neeradi Sunil Raj (16K81A05F4)

3) Ms. Rimmala Prerana (16K81A05G2)

4) Mr. D Chaitanya (16K81A05C6)

In partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Under the Guidance of

Mr. CH. Malleswar Rao

Assistant Professor, Dept. of CSE

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ST.MARTIN'S ENGINEERING COLLEGE

An Autonomous Institute

Dhulapally, Secunderabad – 500 100

JUNE 2021

BONAFIDE CERTIFICATE

This is to certify that the project entitled **CHARACTERIZING AND PREDICTING EARLY REVIEWERS FOR EFFECTIVE PRODUCT MARKETING ON E-COMMERCE WEBSITES**, is being submitted by 1. **Mr. Neti Rohit Kumar 16K81A05F5, 2. Neeradi Sunil Raj 16K81A05F4, 3. Miss. Rimmala Prerana 16K81A05G2, 4. D. Chiatanya 16K81A05C6,** in partial fulfillment of the requirement for the award of the degree of **BACHELOR OF TECHNOLOGY IN Computer Science and Engineering** is recorded of bonafide work carried out by them. The result embodied in this report have been verified and found satisfactory.

Mr. CH. Malleshwar Rao
Department of CSE

Head of the Department
Dr. M.NARAYANAN
Department of CSE

Internal Examiner

External Examiner

Place:

Date:

DECLARATION

We, the student of **Bachelor of Technology** in Department of Computer Science and Engineering', Session: 2016 – 2021, St. Martin's Engineering College, Dhulapally, Kompally, Secunderabad, hereby declare that work presented in this Project Work entitled **Characterizing And Predicting Early Reviewers For Effective Marketing Using E-Commerce Websites** is the outcome of our own bonafide work and is correct to the best of our knowledge and this work has been undertaken taking care of Engineering Ethics. This result embodied in this project report has not been submitted in any university for award of any degree.

Neti Rohit Kumar 16K81A05F5
Neeradi Sunil Raj 16K81A05F4
Rimmala Prerana 16K81A05G2
D Chaintanya 16K81A05C6

ABSTRACT

Online reviews have become an important source of information for users before making an informed purchase decision. Early reviews of a product tend to have a high impact on the subsequent product sales. In this paper, we take the initiative to study the behaviour characteristics of early reviewers through their posted reviews on two real-world large e-commerce platforms, i.e., E - Commerce and Yelp. In specific, we divide product lifetime into three consecutive stages, namely early, majority and laggards. A user who has posted a review in the early stage is considered as an early reviewer. We quantitatively characterize early reviewers based on their rating behaviours, the helpfulness scores received from others and the correlation of their reviews with product popularity. We have found that an early reviewer tends to assign a higher average rating score, and an early reviewer tends to post more helpful reviews. Our analysis of product reviews also indicates that early reviewers' ratings and their received helpfulness scores are likely to influence product popularity. By viewing review posting process as a multiplayer competition game, we propose a novel margin-based embedding model for early reviewer prediction. Extensive experiments on two different e-commerce datasets have shown that our proposed approach outperforms a number of competitive baselines.

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible and whose encouragements and guidance have crowded effects with success.

We extended our deep sense of gratitude to Principal**, Dr. P. SANTOSH KUMAR PATRA**, St. Martin's Engineering College, Dhulapally, for permitting us to undertake this project.

We are also thankful to **Dr. M.NARAYANAN**, Head of the Department, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

We would like to thank **Dr. T. POONGOTHAI,** Department of Computer Science and Engineering (AI & ML) for her encouragement and insightful comments and as well as our project coordinators **Dr. B.RAJALINGAM**, Associate Professor and **Mr. J.SUDHAKAR**, Assistant Professor, in Department of Computer Science and Engineering for their valuable support.

We would like to express our sincere gratitude and indebtedness to our project supervisor <**CH. Malleshwar Rao**, Professor>, Computer Science and Engineering,St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

Finally, we express thanks to all those who have helped us successfully to completing this project. Furthermore, we would like to thank our family and friends for their moral support and encouragement.

We express thanks to all those who have helped us in successfully completing the project.

Neti Rohit Kumar 16K81A05F5

Neeradi Sunil Raj 16K81A05F4

Rimmala Prerana 16K81A05G2

D Chaitanya 16K81A05C6

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

<AVI>	Audio Video Interlace
<BMP>	Bitmap
<CPU>	Central Processing Unit
<GB>	Giga Bytes
<GUI>	Graphical User Interface

INTRODUCTION

The emergence of e-commerce websites has enabled users to publish or share purchase experiences by posting product reviews, which usually contain useful opinions, comments and feedback towards a product. As such, a majority of customers will read online reviews before making an informed purchase decision. It has been reported about 71% of global online shoppers read online reviews before purchasing a product. Product reviews, especially the early reviews (i.e., the reviews posted in the early stage of a product), have a high impact on subsequent product sales. We call the users who posted the early reviews early reviewers. Although early reviewers contribute only a small proportion of reviews, their opinions can determine the success or failure of new products and services. It is important for companies to identify early reviewers since their feedbacks can help companies to adjust marketing strategies and improve product designs, which can eventually lead to the success of their new products. For this reason, early reviewers become the emphasis to monitor and attract at the early promotion stage of a company. The pivotal role of early reviews has attracted extensive attention from marketing practitioners to induce consumer purchase intentions. For example, E - Commerce, one of the largest e-commerce companies in the world, has advocated the Early Reviewer Program1, which helps to acquire early reviews on products that have few or no reviews. With this program, E - Commerce shoppers can learn more about products and make smarter buying decisions. As another related program, E - Commerce Vine2 invites the most trusted reviewers on E - Commerce to post opinions about new and pre release items to help their fellow customers make informed purchase decisions.

PROJECT OVERVIEW

Business Case:

The Early Reviewer program allows a seller to submit one of their product SKUs (stock keeping unit) to be promoted by E - Commerce for review by a specific, pre-vetted reviewer.

The program costs $60 per SKU. However, you aren't charged until you get one review or one year has passed, whichever comes first. The product should receive between 1-5 reviews from reviewers, who have been handpicked by E- Commerce site. Reviewers are chosen for the program because they have "no history of abusive or dishonest reviews" and they meet all of E-Commerce's "eligibility criteria." Once program reviewers leave feedback, E- Commerce marks the review with an orange badge denoting its early reviewer status.

Reviewers are rewarded for their participation in the program with a gift card typically valued at $1-$3, regardless of the review, be it a 1-star or 5-star rating. (Some sellers may feel this contradicts E- Commerce's policy as sellers themselves aren't allowed to incentivize reviews. To learn more about E- Commerce's rules/regulations regarding reviews, be sure to read our article outlining E - Commerce's terms of service.)

Opportunities:

E - Commerce Early Reviewer Program, as the name suggests, was introduced by E - Commerce to help its sellers get reviews for new products. It came into effect shortly after the e-commerce giant banned incentivized reviews as they made it quite tricky to judge the authenticity of said reviews.

Sellers can choose the products they want to enrol in the program. After that, E - Commerce will contact the customers who have purchased these items, at random, to ask them to leave an authentic review and share their experience. In return for their testimonials, these customers will be rewarded a small E - Commerce.com gift card, ranging from $1 to $3. It doesn't matter whether the product is worthy of 1-star or 5-stars; as long as the customers have purchased any of the participating items, they will be eligible to leave a review under the program.

E - Commerce Early Reviewer Program is a win-win situation for both shoppers and sellers. Sellers get more exposure for their products and brands, while buyers get a small reward for sharing their opinion. Reviews from this early reviewer program are distinguished and identified by an orange badge.

Problems:

Along with the Early Reviewer Program, E - Commerce also added functionality — in the form of the "Request A Review" button — for sellers to solicit a review from customers who have purchased their products. That triggers an email that E - Commerce sends to customers, telling them that the seller wants their feedback. E - Commerce also lowered the barrier to leaving a review in the first place. Starting in late 2019, customers have been able to leave only a rating, rather than writing out a full review, through the company's new One-Tap Rating option.

The end of the Early Reviewer Program appears to be a sign that these new features are gaining traction. "They threw some spaghetti at the wall to see what stuck," said Mercer. Regarding the demise of the program, he said, "I don't think this is going to be something that negatively affects sellers much."

That's not to say fake reviews aren't still a problem for E - Commerce. E - Commerce groups that trade free products in exchange for review, sometimes under the guise of giving products to "influencers," continue to proliferate, and so do websites selling fake reviews on E - Commerce. In February, a British consumer group identified 10 websites devoted to selling fake reviews on E - Commerce, and Modern Retail has previously reported on how the ecosystem works. Some third-party services also offer services that they say will encourage customers to leave honest reviews of products on E - Commerce, but Mercer said that "they're pretty much all outside the bounds" of E - Commerce's terms of service.

Still, if E - Commerce's goal has been to boost the overall share of customers who review products on its marketplace, the efforts seem to be working. Mercer said that, as E - Commerce has rolled out new features encouraging E - Commerce customers to leave reviews, the purchase-to-review ratio has steadily climbed.

Fred Ruckel, who sells a popular cat mat on E - Commerce, said that "the response rate on the 'ask for a review button' has been pretty good." Before it debuted, Ruckel remembers that he used

to send messages to every customer, asking them to leave an honest review one by one. “I would manually send as many as 200 in a half hour,” he said. Ruckel said that he never used the Early Reviewer program, in part because he “always considered that akin to a discount for a review and felt it cheapened the value of the review system.” But at least in the abstract, these newest features have streamlined the process of getting a review even further.

Still, no matter how easy — and closely regulated — the process has become, Ruckel said that only a small portion of customers ever leave reviews at all. “Based on sales over time, I would say only around 5% leave reviews,” he said.

PROJECT OBJECTIVES

Goals:

The main Goal of the project is to retrieve individual reviewer data and use it as a trust factor for future customer. The final goal of the project is to affect the product performance and increase the usage of E - Commerce site. To developed a competitive environment for enthusiastic participation of early reviewers in future enhancement.

We can see that early reviewers are extremely important for product marketing. we take the initiative to study the behaviour characteristics of early reviewers through their posted reviews on representative e-commerce platforms, e.g Amazon. We aim to conduct effective analysis and make accurate prediction on early reviewers. This problem is strongly related to the adoption of innovations.

The Scope of this project is to help the new brands and products built their reputation and gain an audience before entering the actual market with the feedback of early reviewers. By using reviewer data and creating a competitive game like review environment the early reviews could be used to

- Increase the reach of the Product and Brand.
- Target the right audience.
- Helps a product skip its growth stage in the beginning.
- Increase the revenue of both E commerce and individual brands respectively.

2 LITERATURE SURVEY

When you go to make an online purchase, what's the first thing you do? In an ecommerce-driven world where customers can't physically experience products before purchasing, many consumers turn to online product reviews. As online review sites such as Yelp! And Facebook have expanded, finding an opinion on just about anything is only a few clicks away. The proliferation of reviews has even gone so far as to shape how businesses are perceived online.

In today's web-based world, virtually everyone is reading online reviews. In fact, 91% of people read them and 84% trust them as much as they would a personal recommendation. The effects of reviews are measurable, too. The average customer is willing to spend 31% more on a retailer that has excellent reviews.

Negative reviews can carry as much weight as positive ones. One study found that 82% of those who read online reviews specifically seek out negative reviews. That may sound alarming — this stat only emphasizes that negative reviews aren't going unnoticed — but there are some benefits: Research indicates that users spend five times as long on sites when interacting with negative reviews, with an 85% increase in conversion rate.

Nearly nine out of ten (89 percent) consumers worldwide make the effort to read reviews before buying products (Trustpilot, 2020). There doesn't seem to be a big difference in reviews consumption between men and women. Just slightly more women (90 percent) read reviews compared to men (88 percent). Google is by far the most popular channel people turn to for reviews, with approximately 57 percent of shoppers using it (Bizrate Insights, 2019). This is followed by a business' own website at just over 40 percent and Yelp and Face book at around 20 percent each.

NAVI BAYE'S ALGORITHM

Naïve Bayes Algorithm: In machine learning, naive Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naive) independence assumptions between the features. Naive Bayes has been studied extensively since the 1950s. It was introduced under a different name into the text retrieval community in the early 1960s, and remains a popular (baseline) method for text categorization, the problem of judging documents as belonging to one category or the other (such as spam or legitimate, sports or politics, etc.) with word frequencies as the features. With appropriate pre-processing, it is competitive in this domain with more advanced methods including support vector machines. It also finds application in automatic medical diagnosis.[3] Naive Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem. Maximum-likelihood training can be done by evaluating a closed-form expression, 718 which takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers. In the statistics and computer science literature, naive Bayes models are known under a variety of names, including simple Bayes
and independence Bayes. All these names reference the use of Bayes' theorem in the classifier's decision rule, but naive Bayes is not (necessarily) a Bayesian method.

Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred. Bayes' theorem is stated mathematically as the following equation:

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

where A and B are events and P(B) ? 0.

Basically, we are trying to find probability of event A, given the event B is true. Event B is also termed as evidence.
P(A) is the priori of A (the prior probability, i.e. Probability of event before evidence is seen). The evidence is an attribute value of an unknown instance(here, it is event B).
P(A|B) is a posterior probability of B, i.e. probability of event after evidence is seen.
Now, with regards to our dataset, we can apply Bayes' theorem in following way:

$$P(y|X) = \frac{P(X|y)\,P(y)}{P(X)}$$

where, y is class variable and X is a dependent feature vector (of size n) where:

$$X = (x_1,x_2,x_3,.....,x_n)$$

Just to clear, an example of a feature vector and corresponding class variable can be: (refer 1st row of dataset)

X = (Rainy, Hot, High, False)

y = No

So basically, P(y|X) here means, the probability of “Not playing golf” given that the weather conditions are “Rainy outlook”, “Temperature is hot”, “high humidity” and “no wind”.

3. SOFTWARE AND HARDWARE REQUIREMENTS

REQUIREMENT ANALYSIS:

The project involved analyzing the design of few applications so as to make the application more users friendly. To do so, it was really important to keep the navigations from one screen to the other well ordered and at the same time reducing the amount of typing the user needs to do. In order to make the application more accessible, the browser version had to be chosen so that it is compatible with most of the Browsers.

PYTHON:

Python is a general-purpose interpreted, interactive, object-oriented, and high-level programming language. An interpreted language, Python has a design philosophy that emphasizes code readability (notably using whitespace indentation to delimit code blocks rather than curly brackets or keywords), and a syntax that allows programmers to express concepts in fewer lines of code than might be used in languages such as C++or Java. It provides constructs that enable clear programming on both small and large scales. Python interpreters are available for many operating systems. CPython, the reference implementation of Python, is open source software and has a community-based development model, as do nearly all of its variant implementations. CPython is managed by the non-profit Python Software Foundation. Python features a dynamic type system and automatic memory management. It supports multiple programming paradigms, including object-oriented, imperative, functional and procedural, and has a large and comprehensive standard library

DJANGO:

Django is a high-level Python Web framework that encourages rapid development and clean, pragmatic design. Built by experienced developers, it takes care of much of the hassle of Web development, so you can focus on writing your app without needing to reinvent the wheel. It’s free and open source.

Django's primary goal is to ease the creation of complex, database-driven websites. Django emphasizes reusability and "plug ability" of components, rapid development, and the principle of don't repeat yourself. Python is used throughout, even for settings files and data models.

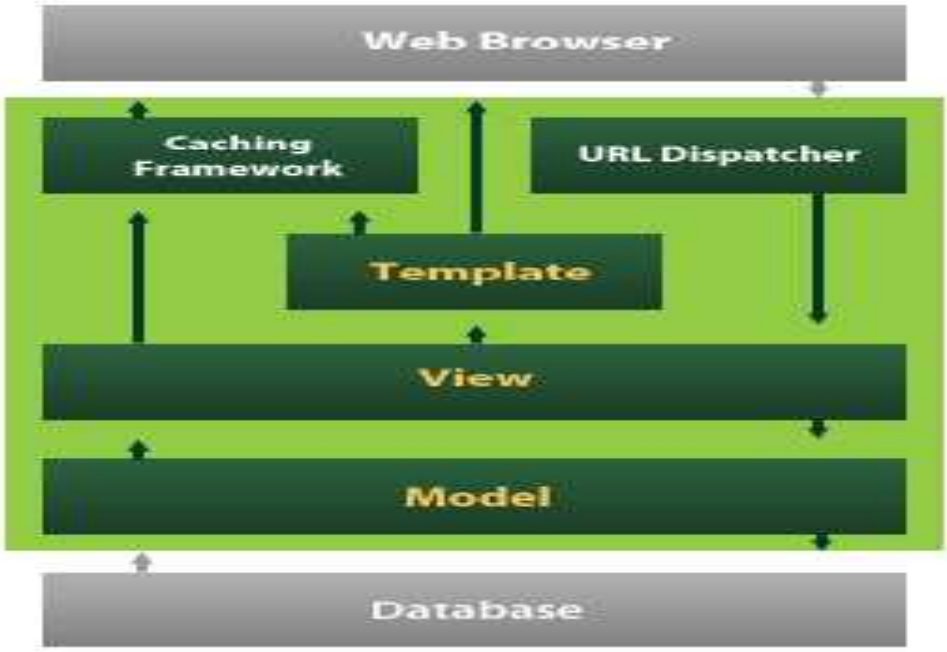


Fig. 1 Django Framework.

Django also provides an optional administrative create, read, update and delete interface that is generated dynamically through introspection and configured via admin models

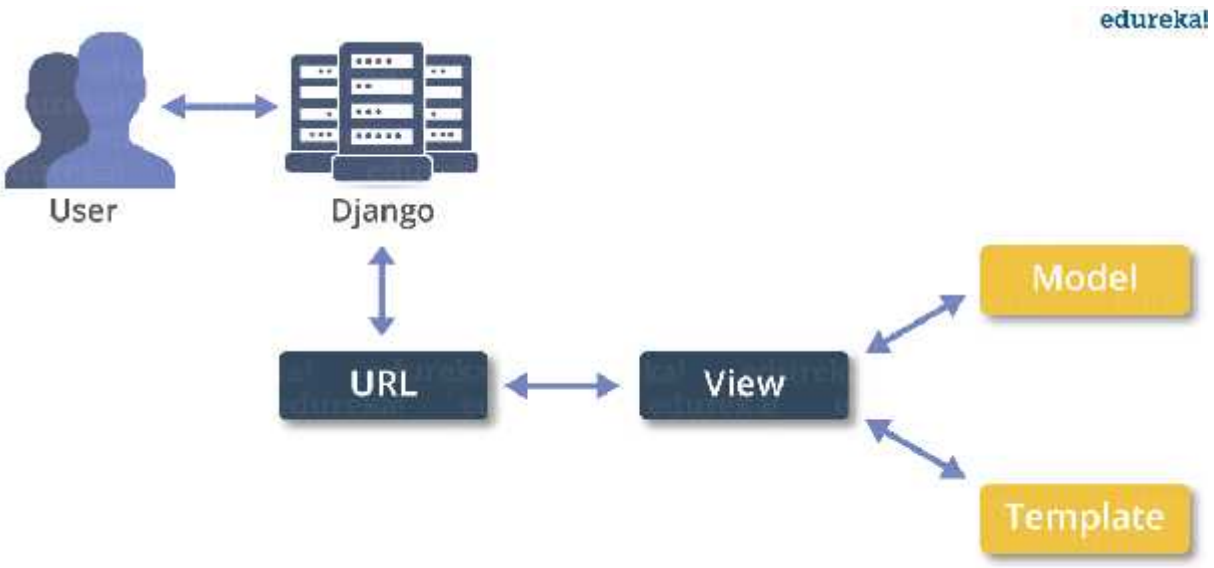


Fig. 2 Django Framework

3.1 HARDWARE REQUIREMENTS

Required Processor : Pentium - Intel Core 10th Gen (latest Version)

Processor Speed : 1.1 GHz

RAM : 8 GB

Hard Disk : 1 TB

3.2 SOFTWARE REQUIREMENTS

Programming Language : Python

Python web framework : Django

Database : Mysql

Web Server : WAMP Server

SOFTWARE DEVELOPMENT ANALYSIS

4.1 OVERVIEW OF PROBLEM:

Early prediction model are not accurate. These prediction models may vary depending on the result outcome and other factors may affect the accurate prediction.

Existing works depends on extracting opinions or identifying opinion targets (or holders) from review data. This extraction is performed by analyzing user data and appropriate results.

Most of these studies are theoretical analysis at the macro level and there is a lack of quantitative investigations.

Given a product p and a candidate user set Up: {u1, u2, ..., uNp }, the task of predicting early reviewers aims to produce a top-K list of users from Up, who would post reviews on p at the early stage of product p in market. Producing a top-K list can be formulated as a ranking problem. We propose to use a ranking function S(p, u) to select users, which measures the likelihood that user u becomes an early reviewer of product p. To learn such a function, we assume that a training set of past early adoption records is available, i.e., {hpi ,Lii} . Each training instance consists of a product pi with a complete lifetime, and Li : hu (i) 1 , s(i) 1 i →hu (i) 2 , s(i) 2 i... →hu (i) Np , s(i) Npi i is an ordered list of reviewers{u (i) j } on pi by the timestamps{s(i) j } when publishing the reviews. A major challenge is that our task is a cold-start ranking problem. Since we are interested in the early reviewers of a product, the predictions should be made when a new product is just released. We will have very little and sometimes even no observed user behavior data at the early stage of a new product. Inspired by previous cold-start recommendation algorithms, we propose to utilize side information to help with this ranking problem. We assume that a product p is with a category label cp and a title description tp and use the two types of side information to learn product representations or embeddings as will be discussed.

A competition-based viewpoint to the ranking task. To address the ranking problem, we draw

our inspiration from multiplayer competition to develop our approach. Generally speaking, given a product p and two candidate user's u and u 0, we seek to model the partial order between them. We consider the review posting process as multiplayer competition: only the most competitive users can become the early reviewers with respect to a product. The competition process can be further decomposed into multiple pair wise comparisons between two players. A competition is carried out between two users given a product. In a two-player competition, the winner will beat the loser with an earlier timestamp. Formally, we use u p u 0 denote that user u has an earlier review timestamp than u 0 for product p. Competition-based ranking has been previously explored for community question answering and player ranking. However, to the best of our knowledge, it has never been explored for early reviewer or early adopter prediction.

A Margin-based Embedding Model for Predicting Early Reviewers:

The essence of this task is to model the partial order between two candidate users' u and u 0 given a product p. Hence, we can cast the total order ranking problem into a pair wise comparison problem. Inspired by the recent progress in distributed representation learning we propose to use an embedding model for this task. We assume that both users and products are mapped into a latent space. In this way, a user u is modelled with a low-dimensional representation vector vu, and a product p is modelled with a low-dimensional dense representation vector vp. In the embedding space, we can reconstruct the partial order relations in the training set and learn the model parameters.

Modelling the Pair wise Comparison:

Based on the embedding representation, we can define the objective function S(p, u) as an inner product between user and product embeddings, i.e.,

$$S(p,u) - \boldsymbol{v}_p^\top \cdot \boldsymbol{v}_u. \qquad (1)$$

In the embedding space, it is expected that v > p · vu > v > p · vu0 when u p u 0 . Given the original training set A = {hpi ,Lii}, we first transform them into a set of partial order pairs T = {u p u 0 |u, u0 ∈ Lp}, where Lp is the reviewer list of product p. To learn such embeddings, we minimize a margin-based ranking criterion over the training set T

$$\begin{aligned} \mathcal{E}(T) &= \sum_{u \succ_p u' \in U} [m + S(p,u') - S(p,u)]_+ \qquad (2) \\ &= \sum_{u \succ_p u' \in U} [m + \boldsymbol{v}_{u'}^\top \cdot \boldsymbol{v}_p - \boldsymbol{v}_u^\top \cdot \boldsymbol{v}_u]_+, \end{aligned}$$

4.2 DEFINE THE PROBLEM

Early Reviewer Program is no longer being accepted. Major companies confirm the discontinuation of the program within the Early Reviewer Program FAQs. The Early Review Program proved to be a reliable, trustworthy way to generate authentic reviews by incentivizing customers to leave a review. In 2016, A leading e-commerce company updated its policy on reviews in response to a significant uptick in fake reviews. As a result, review generation became difficult, especially for new products entering the marketplace. In response, Amazon created the Early Reviewer Program to provide an option for new sellers to generate their first five reviews, with customers receiving small gift cards once they've left a review for a purchased product.

4.3 MODULES OVERVIEW

DATA PREPARATION:

In this module, we use the Amazon datasets. Amazon dataset originally contains 142:8 million product reviews ranging from May 1996 to July 2014 and Yelp dataset contains 4:7 million product reviews ranging from July 2004 to January 2017. Each review is a textual comment posted by a user on a product, and is accompanied with it's publish timestamp which accurate to days in our study. A review is associated with a rating score in a five-star scale. Each product is associated with a category label and a textual description. Given a review, other Amazon users can vote on its helpfulness using a binary choice of Yes or No button. The number of votes on positive attitude (i.e., Yes) and negative attitude (i.e., No) can be recorded. While in Yelp dataset, other users can only vote on the helpfulness of a review by clicking the Useful button, and explicit negative attitude on the helpfulness is not recorded.

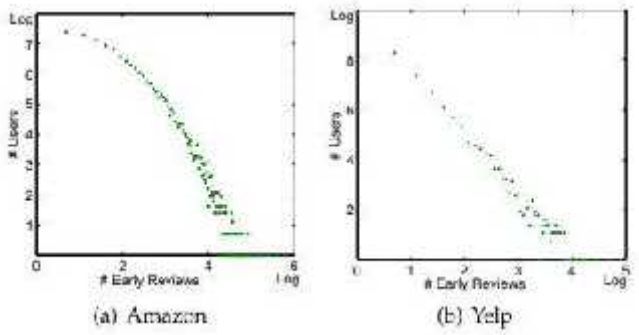


Fig. 3. #early reviews v.s. #users. #early reviews is the number of early reviews that a user has posted, i.e., the number of times that a user has acted as an early reviewer of a product.

DATA CLEANING:

Our data cleaning contains two main steps as follows.

Preprocessing

Review Spammer Detection and Removal

PREPROCESSING:

We first remove reviews from anonymous users, since we would like to associate each review with a unique user. We then remove duplicate reviews often caused by multiple versions of the same product. We also remove inactive users and unpopular products: we only keep the users who have posted at least ten and five reviews, and products which have received at least ten and five reviews in Amazon and Yelp datasets respectively. For review text, we remove stop words and very infrequent words.

REVIEW SPAMMER DETECTION AND REMOVAL:

Our focus is to study the early adoption behaviors of genuine Amazon and Yelp users. However, the number of spam reviews has increasingly grown on ecommerce websites, and it was found that about 10% to 15% of reviews echoed earlier reviews and might be posted by review spammers. It is possible that spam reviews are posted to give biased or false opinions on some products so as to influence the consumers' perception of the products by directly or indirectly inflating or damaging the product's reputation. The existence of spam reviews could lead to erroneous conclusions in our study. Therefore, we need to remove review spammers as part of our data cleaning process.

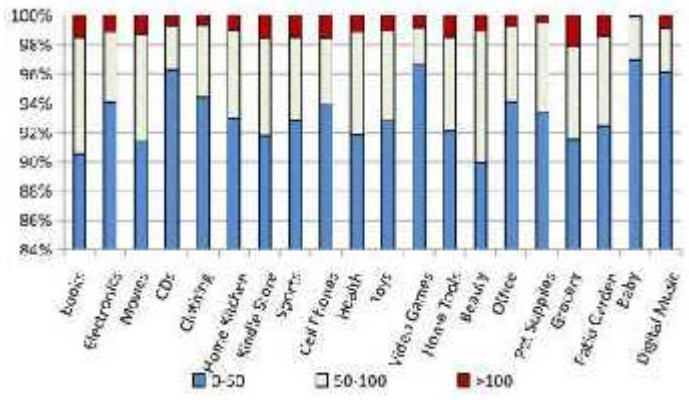


Fig. 4. The percentage of Amazon users posting early reviews in different bins by product categories.Three bins are considered, i.e., [0, 50], (50, 100] and (100, +∞).

CHARACTERISTICS OF EARLY REVIEWERS:

It has been reported that early adopters are important to the diffusion of innovations. Hence, we hypothesize that early reviewers play a key role in future product adoptions. There has been a lack of quantitative analysis of the correlations between the early reviewers and product adoptions on large datasets, i.e., Amazon and Yelp. In this section, we study how early reviewers are different from others and how they impact product popularity

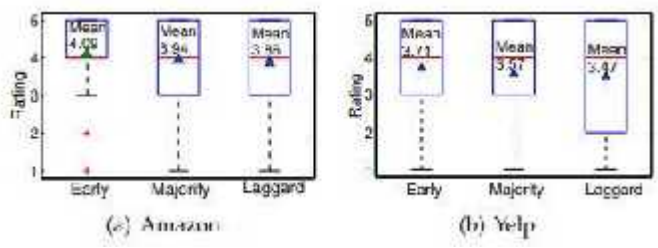
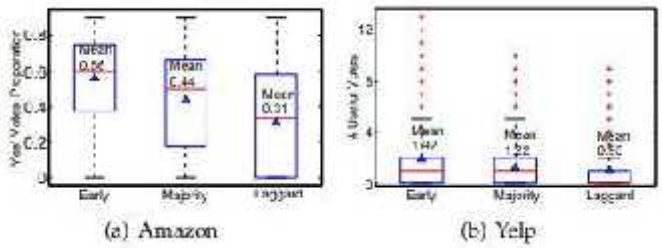


Fig. 5. Comparisons of the rating scores by the three categories of reviews.

Fig.6 Comparisons of the helpfulness scores by the three categories of reviews.



In this module, to understand how early reviewers are different from others; we start with an analysis of their posted early reviews by looking into average ratings of the reviews and helpfulness scores voted by others. Using the categorization method discussed in Section 2, we assign each review into one of the three categories defined in Figure 2. Recall that each review is associated with a rating score and votes on its helpfulness. The rating score is in a five-star scale. For helpfulness, in Amazon dataset, we count the number of Yes and No votes respectively and then normalize them to the range of [0; 1]. While in Yelp dataset, users vote on the helpfulness of a review by clicking the Useful button. We count the number of Useful as the review's helpfulness score. Given the three categories of reviews, we compute the average ratings and helpfulness scores in each review category.

Categories	No. of 'Yes'	No. of 'No'	Normalized 'Yes'	Normalized 'No'
Early	15.5 ± 0.21	3.28 ± 0.05	0.72 ± 0.002	0.28 ± 0.002
Majority	4.98 ± 0.03	2.28 ± 0.02	0.68 ± 0.001	0.32 ± 0.001
Laggards	2.23 ± 0.04	1.44 ± 0.03	0.67 ± 0.003	0.33 ± 0.003

Table: 1 Comparisons of the helpfulness scores by the three categories of reviews in the Amazon dataset.

4.3 DEFINE THE MODULES

USER MODULE:

User Login and Registration:

The user module allows users to register, log in, and log out. Users benefit from being able to sign on because this associates content they create with their account and allows various permissions to be set for their roles.

The user module supports user roles, which can be set up with fine-grained permissions allowing each role to do only what the administrator permits. Each user is assigned one or more roles. By default there are three roles: *anonymous* (a user who has not logged in) and *authenticated* (a user who is registered), and *administrator* (a signed in user who will be assigned site administrator permissions). Users can use their own name or handle and can fine tune some personal configuration settings through their individual *my account* page. Registered users need to authenticate by supplying their username and password, or alternately an Open ID login.

A visitor accessing your website is assigned a unique ID, the so-called session ID, which is stored in a cookie. For security's sake, the cookie does not contain personal information but acts as a key to retrieving the information stored on your server.

You can:

- view your user page at http://127.0.0.1:8000/ when you're logged in
- Administer users at People (***Admin*** **>** ***User***) in Login Page (***Administer*** **>** ***User***) in previous versions.
- Create new users on the People page through the *add user* link in Drupal 7 and 8, and (***Administer*** **>** ***User*** **>** ***Add user***) in previous versions.

- Configure user registration, user email, and user picture settings on the Account settings page (***Administer > Configuration > People > Account settings***) in Drupal 7 and 8 and (***Administer > Settings > User***) in previous versions.
- Allow users to select themes in versions prior to Drupal 7 from their user account by enabling themes in (***Administer > Themes***).
- Read user profile help at (***Administer > Help > User***).
- Configure access permissions at (***Administer > People > Permissions***) in Drupal 7 and 8, and (***Administer > Access control***) in previous versions.

Home:

A home page is the main web page of a website. The term also refers to one or more pages always shown in a web browser when the application starts up. In this case, it is also known as the start page. The word "home" comes from the use of the Home key on a keyboard to return to the start page at any time.

View Product and Cart:

A product description is the marketing copy used to describe a product's value proposition to potential customers. A compelling product description provides customers with details around features, problems it solves and other benefits to help generate a sale.

It's no wonder they are worried — the quality of a product description can make or break a sale, especially if it doesn't include the information a shopper needs to make a purchase decision. Providing key product details is critical if you want the shopper to click "Add to Cart" and differentiate your ecommerce website from the competition.

Whether your products have a specific function, like a camera, or a personal purpose, like fashion, all products exist to enhance or improve the purchaser's quality of life in one way or another. As the shopper browses, they instinctively imagine having each product in hand, using it and enjoying it.

The more powerful the customer's fantasy of owning the product, the more likely they are to buy it. Therefore, I like to think of product descriptions as storytelling and psychology, incorporating the

elements of both prose writing and journalism. A “good” product description will not do. Competition is getting too fierce. It must be great!

Below examples highlighting how improving product descriptions improve conversion rates as well as tips to help you craft the perfect copy.

View Ratings and Add Ratings:

There are many users who purchase products through E-commerce websites. Through online shopping many E-commerce enterprises were unable to know whether the customers are satisfied by the services provided by the firm. This boosts us to develop a system where various customers give reviews about the product and online shopping services, which in turn help the E-commerce enterprises and manufacturers to get customer opinion to improve service and merchandise through mining customer reviews.

An algorithm could be used to track and manage customer reviews, through mining topics and sentiment orientation from online customer reviews. In this system user will view various products and can purchase products online. Customer gives review about the merchandise and online shopping services. Certain keywords mentioned in the customer review will be mined and will be matched with the keywords which are already exist in the database based on the comparison, system will rate the product and services provided by the enterprise. This system will use text mining algorithm in order to mine keywords.

The System takes review of various users, based on the review, system will specify whether the products and services provided by the E-commerce enterprise is good, bad, or worst. We use a database of sentiment based keywords along with positivity or negativity weight in database and then based on these sentiment keywords mined in user review is ranked. This system is a web application where user will view various products and purchase products online and can give review about the merchandise and online shopping services. This system will help many E-commerce enterprises to improve or maintain their services based on the customer review as well as to improve the merchandise based on the customer review.

ADMIN MODULE:

The Admin module allows project administrators to manage products, upload products, analyze user data and manages products add, edit, modifications and services, as well as edit the user profile.

Admin login:

The Admin module allows Admin to register, log in, and log out. Admin benefit from being able to sign on because this associates content they create with their account and allows various permissions to be set for their roles.

The Admin module supports Admin roles, which can be set up with fine-grained permissions allowing each role to do only what the administrator permits. Each Admin is assigned one or more roles. By default there are three roles: *anonymous* (an Admin who has not logged in) and *authenticated* (a Admin who is registered), and *administrator* (a signed in Admin who will be assigned site administrator permissions).

Admin can use their own name or handle and can fine tune some personal configuration settings through their individual *my account* page. Registered Admin need to authenticate by supplying their Admin name and password, or alternately an Open ID login.

A visitor accessing your website is assigned a unique ID, the so-called session ID, which is stored in a cookie. For security's sake, the cookie does not contain personal information but acts as a key to retrieving the information stored on your server.

Home:

A home page is the main web page of a website. The term also refers to one or more pages always shown in a web browser when the application starts up. In this case, it is also known as the start page. The word "home" comes from the use of the Home key on a keyboard to return to the start page at any time.

Upload Products:

Upload Products in Marketplace, upload marketplace add-on allows the sellers to add products to the store using Local files on local devices. The upload marketplace add-on supports simple, configurable, virtual, and downloadable types of products. Mention all the details such as name, category, price, stock, description, tax, etc. in the file including the images.

The admin can also upload the products for the seller from the back-end.

Charts:

A Product Knowledge Graph is an e-commerce specific form of knowledge graph built to improve product find ability and end-user experiences by enriching a brand’s content with data. It consists of data about products, brands, product categories, product features, reviews, hi-res images, shipping data, FAQs and a lot more. Made of structured data and extended product mark-up, injected across both editorial and product content, a product knowledge graph is built on top of the product database to link

all data together combining both structured information, (for instance, the list of products for a brand) or unstructured (for example the descriptions related to a collection of products).

Simply put, a Product Knowledge graph bridges the currently existing gap (hindering Joe's smooth experience and our brand's opportunities to connect to Joe where he most needs it) between product content and editorial content

4.4 MODULES FUNCTIONALITY

MODULES:

There are three modules can be divided here for this project they are listed as below

- Upload products
- Product Review Based Order
- Rating and Reviews
- Data Analysis

From the above three modules, project is implemented. Bag of discriminative words are achieved

MODULE DESCRIPTION:

1. **UPLOAD PRODUCTS**

Uploading the products is done by admin. Authorized person is uploading the new arrivals to system that are listed to users. Product can be uploaded with its attributes such as brand, colour, and all other details of warranty. The uploaded products are able to block or unblock by users.

2. **PRODUCT REVIEW BASED ORDER**

The suggestion to user's view of products is listed based on the review by user and rating to particular item. Naïve bayes algorithm is used in this project to develop the whether the sentiment of given review is positive or negative. Based on the output of algorithm suggestion

to users is given. The algorithm is applied and lists the products in user side based on the positive and negative.

3. RATINGS AND REVIEWS

Ratings and reviews are main concept of the project in order to find effective product marketing. The main aim of the project is to get the user reviews based on how they purchased or whether they purchased or not. The major find out of the project is when they give the ratings and how effective it is. And this will helpful for the users who are willing to buy the same kind of product.

4. DATA ANALYSIS

The main part of the project is to analysis the ratings and reviews that are given by the user. The products can be analysis based on the numbers which are given by user. The user data analysis of the data can be done by charts format. The graphs may vary like pie chart, bar chart or some other charts.

5. PROJECT SYSTEM DESIGN

EXISTING SYSTEM:

Previous studies have highly emphasized the phenomenon that individuals are strongly influenced by the decisions of others, which can be explained by herd behaviour. The influence of early reviews on subsequent purchase can be understood as a special case of herding effect. Early reviews contain important product evaluations from previous adopters, which are valuable reference resources for subsequent purchase decisions. As shown in, when consumers use the product evaluations of others to estimate product quality on the Internet, herd behaviour occurs in the online shopping process. Different from existing studies on herd behaviour, we focus on quantitatively analyzing the overall characteristics of early reviewers using large-scale real-world datasets. In addition, we formalize the early reviewer prediction task as a competition problem and propose a novel embedding based ranking approach to this task. To our knowledge, the task of early reviewer prediction itself has received very little attention in the literature. Our contributions are summarized as follows:

We present a first study to characterize early reviewers on an e-commerce website using two real-world large datasets. We quantitatively analyze the characteristics of early reviewers and their impact on product popularity. Our empirical analysis provides support to a series of theoretical conclusions from the sociology and economics. We view review posting process as a multiplayer competition game and develop an embedding-based ranking model for the prediction of early reviewers. Our model can deal with the cold-start problem by incorporating side information of products. Extensive experiments on two real-world large datasets, i.e., Amazon and Yelp have demonstrated the effectiveness of our approach for the prediction of early reviewers.

PROPOSED SYSTEM:

To predict early reviewers, we propose a novel approach by viewing review posting process as a multiplayer competition game. Only the most competitive users can become the early reviewer's w.r.t. to a product. The competition process can be further decomposed into multiple pairwise comparisons between two players. In a two-player competition, the winner will beat the loser with an earlier timestamp. Inspired by the recent progress in distributed representation learning, we propose to use a margin-based embedding model by first mapping both users and products into the same embedding space, and then determining the order of a pair of users given a product based on their respective distance to the product representation.

5.1 E R DIAGRAMS:

a. User

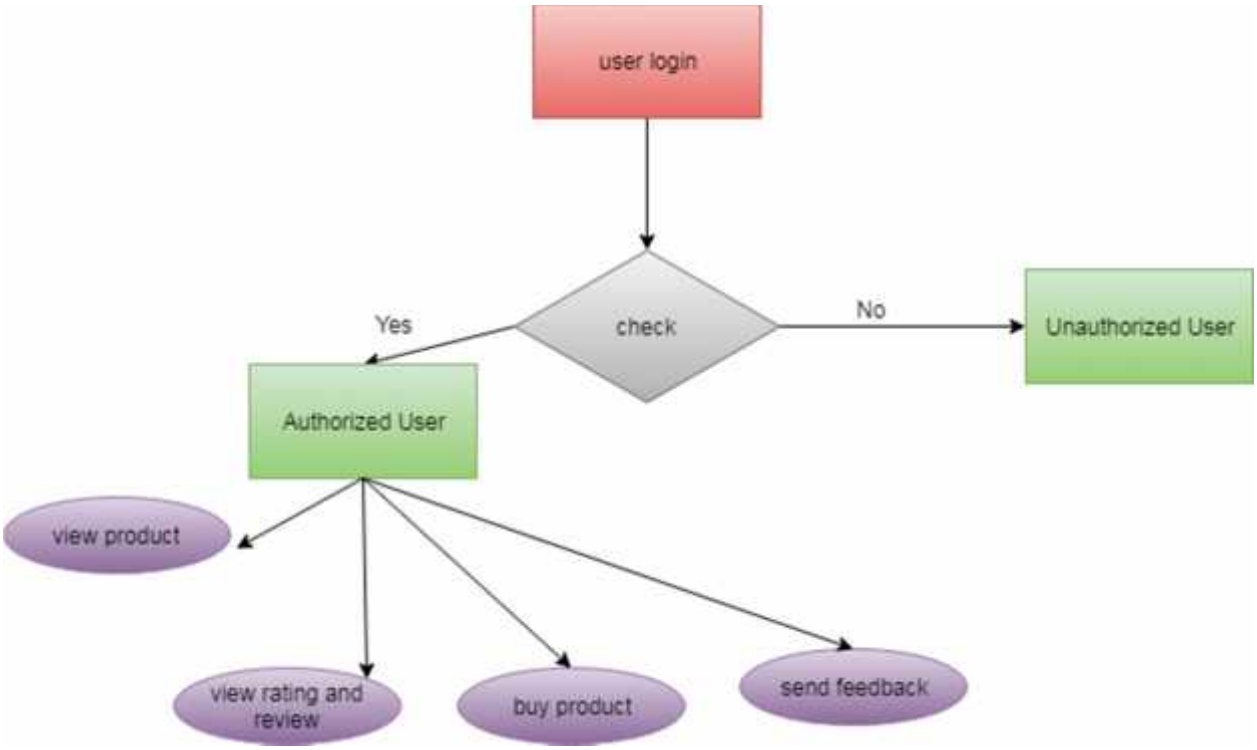


Fig. 7 User Login Flow Chart

b. Admin

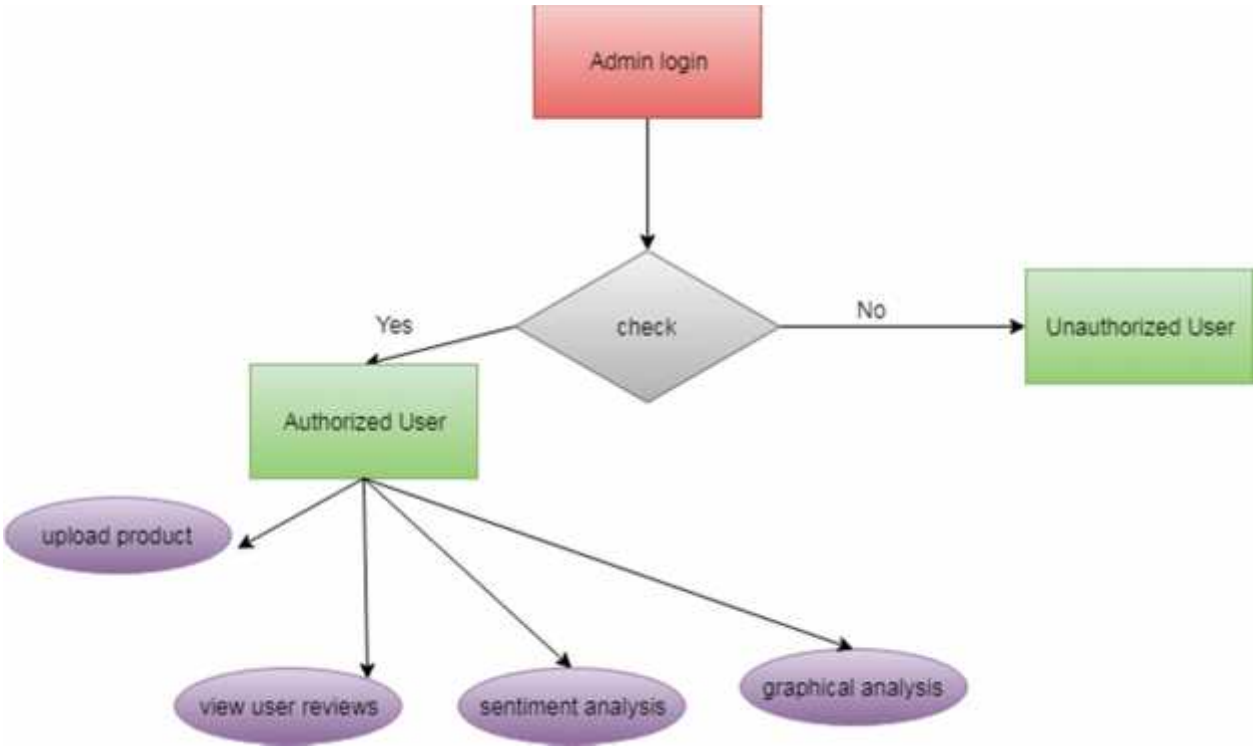


Fig. 8 Admin Login Flow Chart

5.1 UML DIAGRAMS:

2. ARCHITECTURE DIAGRAM

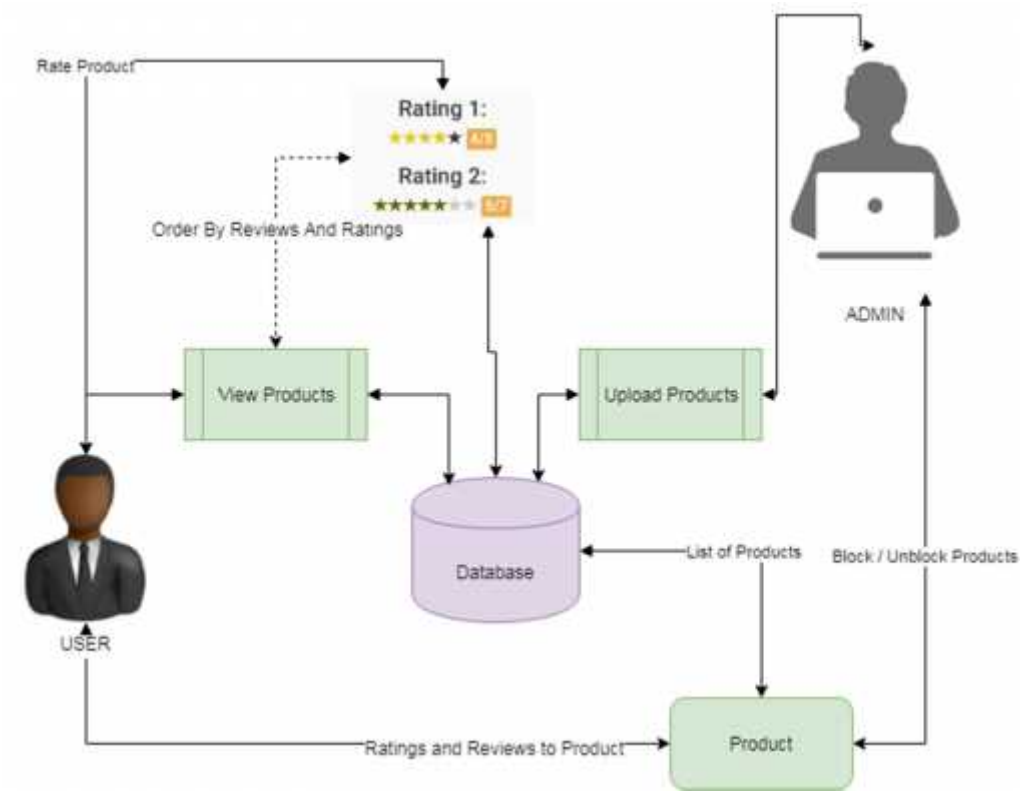


Fig. 9 Architecture Flow Chart

3. COMPONENT DIAGRAM:

a. User

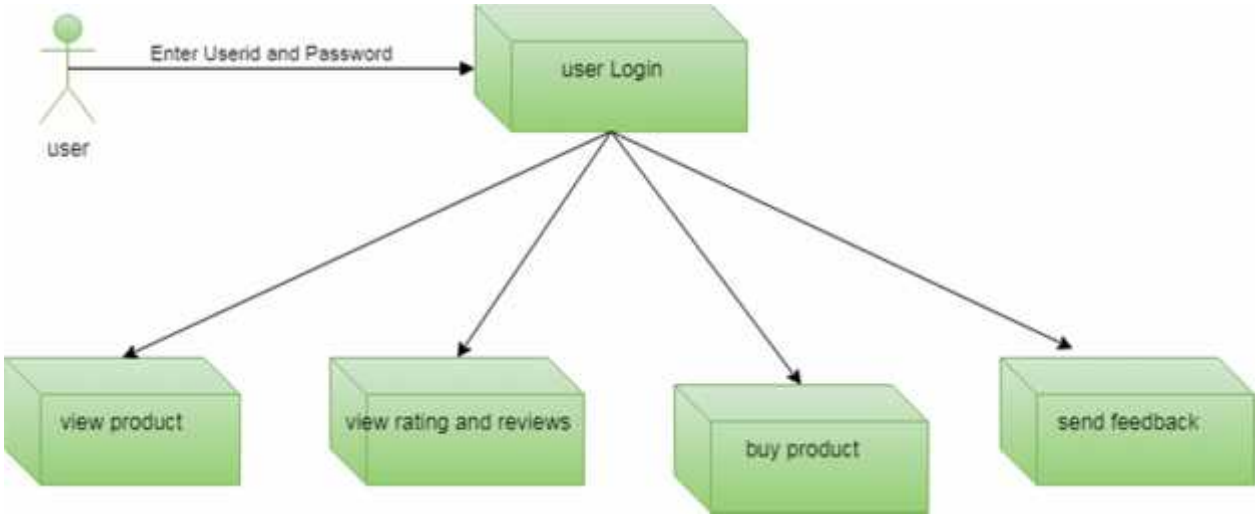


Fig. 9 User Component Diagram

b. Admin

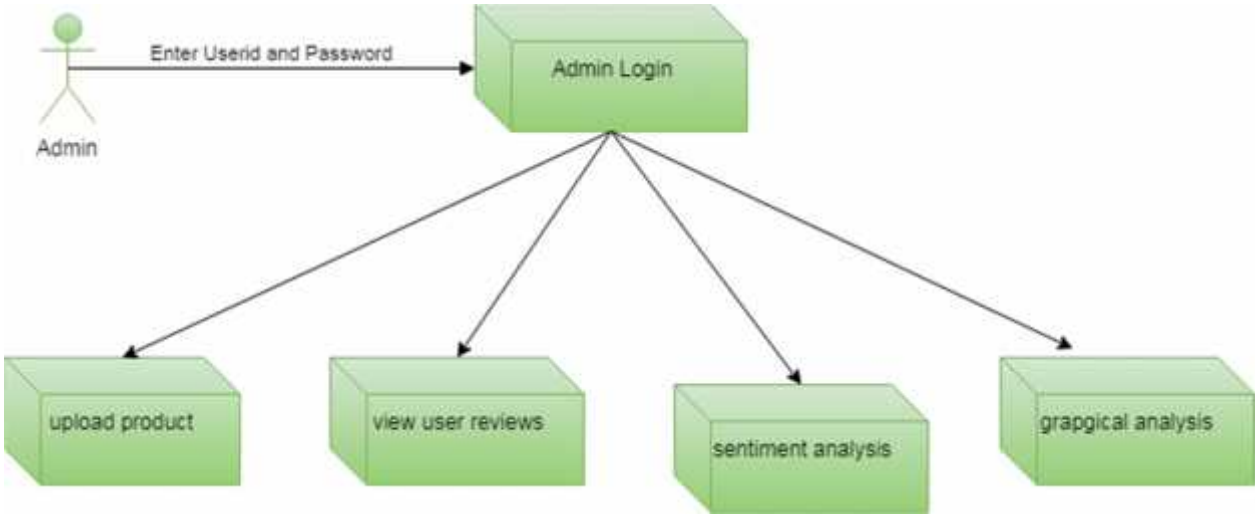


Fig. 10. User Component Diagram

4. USE CASE DIAGRAM

a. User

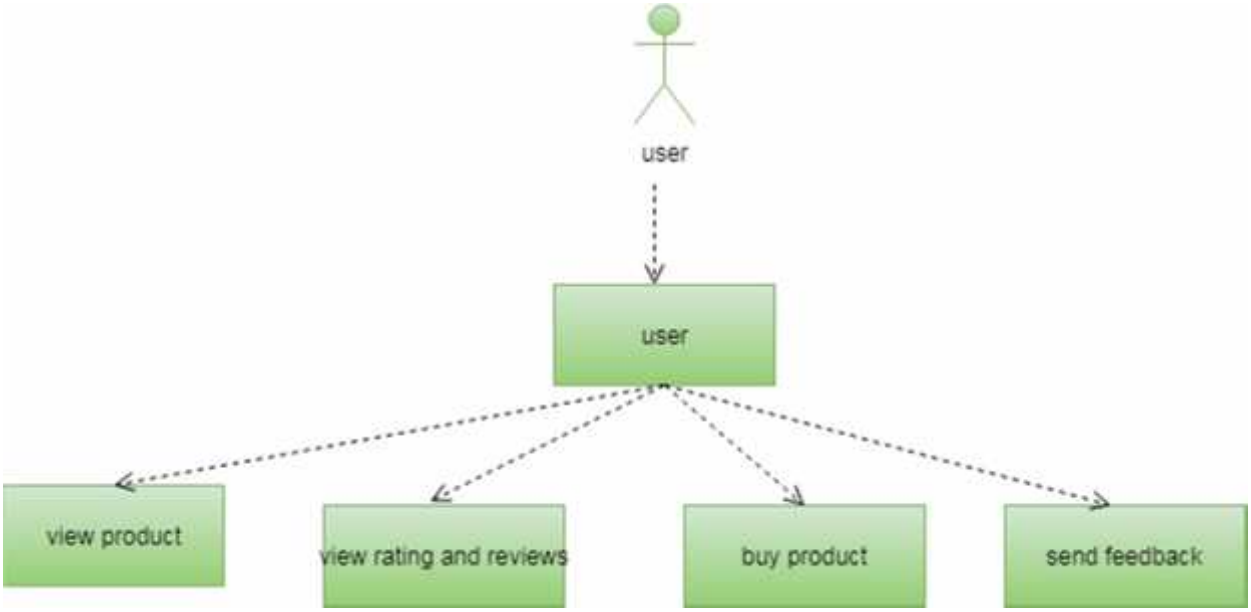


Fig. 11 User Side Use Case Diagram

b. Admin

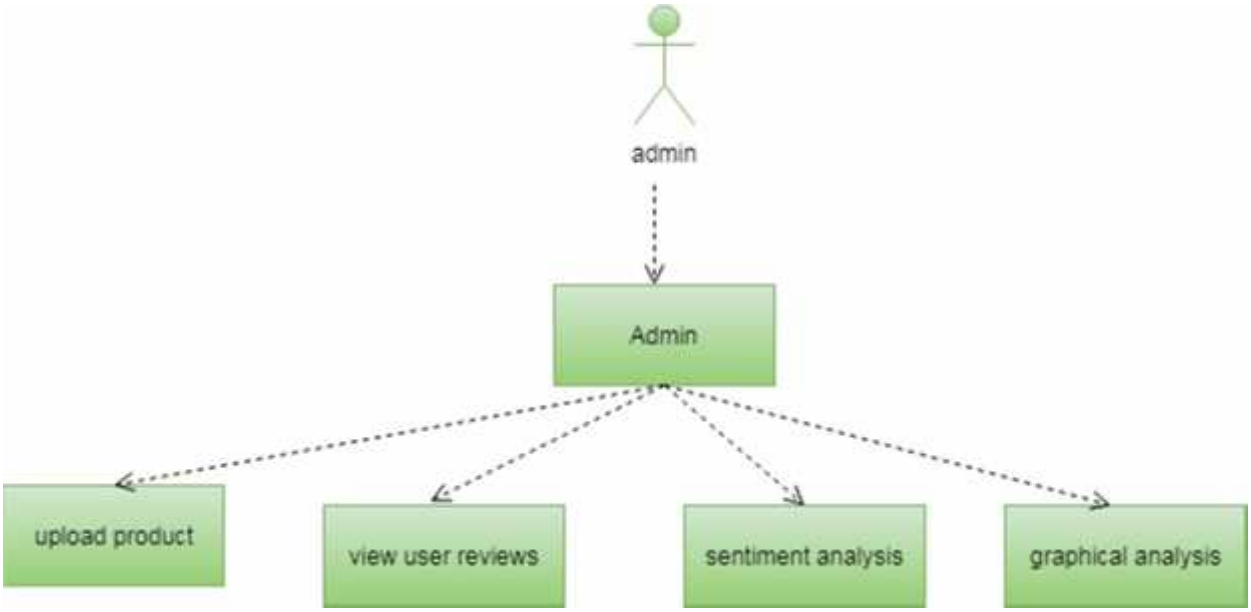


Fig. 12. Admin Side Use Case Diagram

5. CLASS DIAGRAM

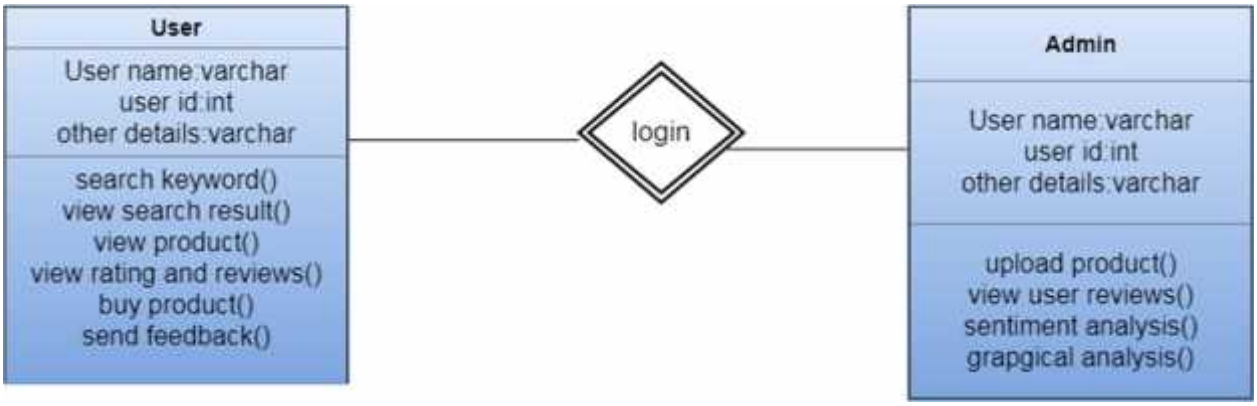


Fig. 13. Class Diagram

6. DATA FLOW DIAGRAM

a. User

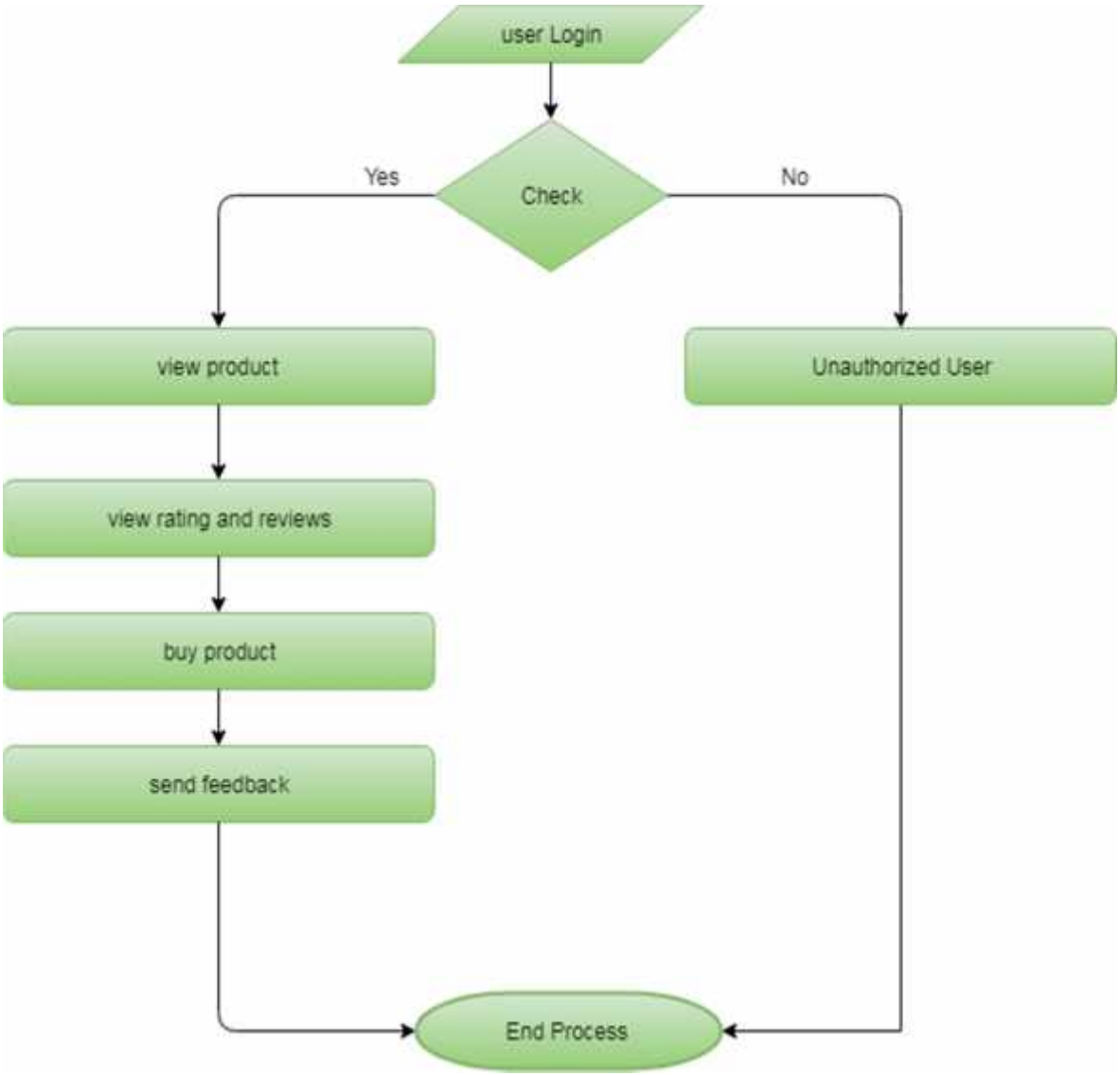


Fig. 14. User Data Overflow

b. Admin

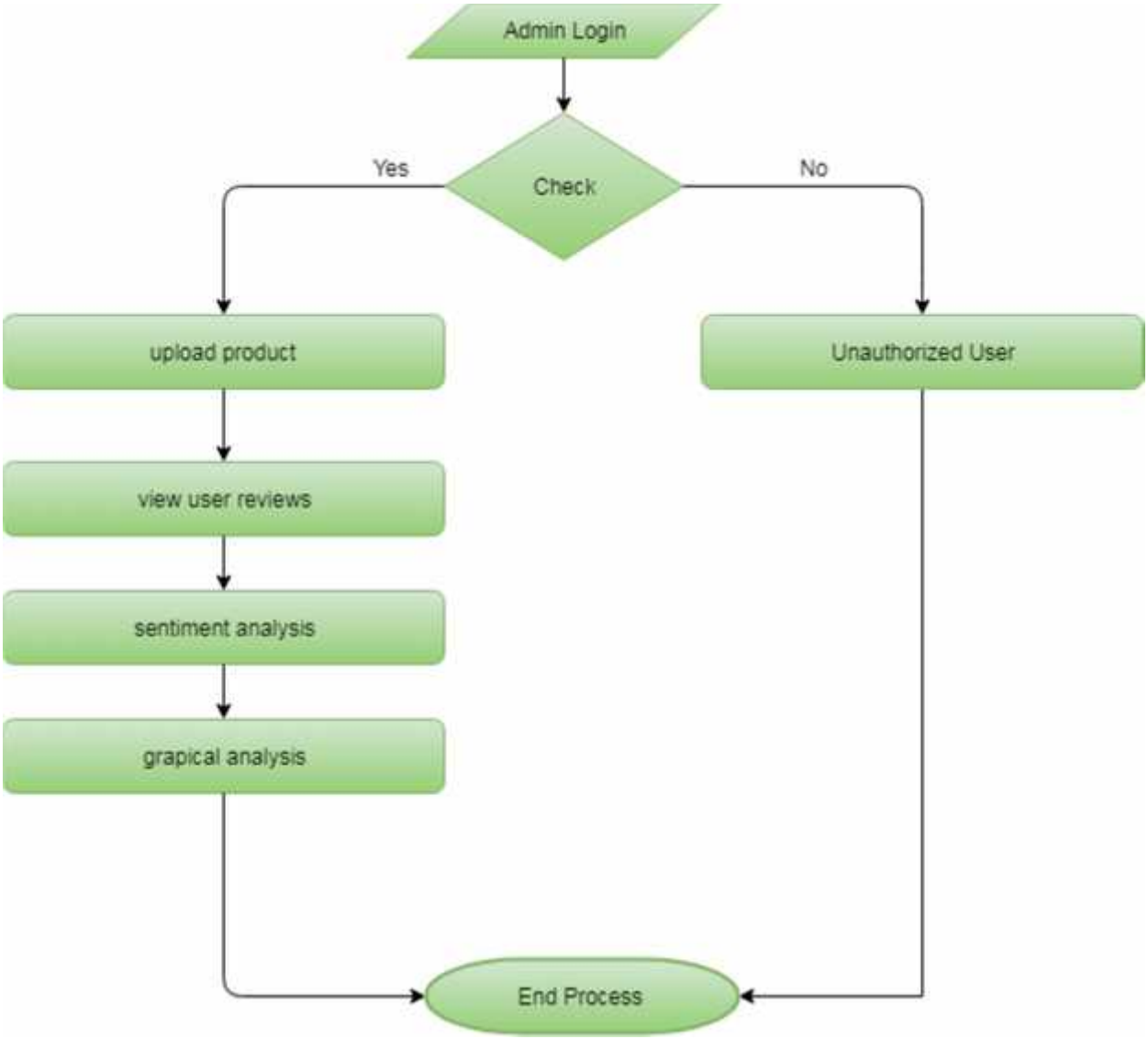


Fig. 15. Admin Data Overflow

7. ACTIVITY DIAGRAM

a. User

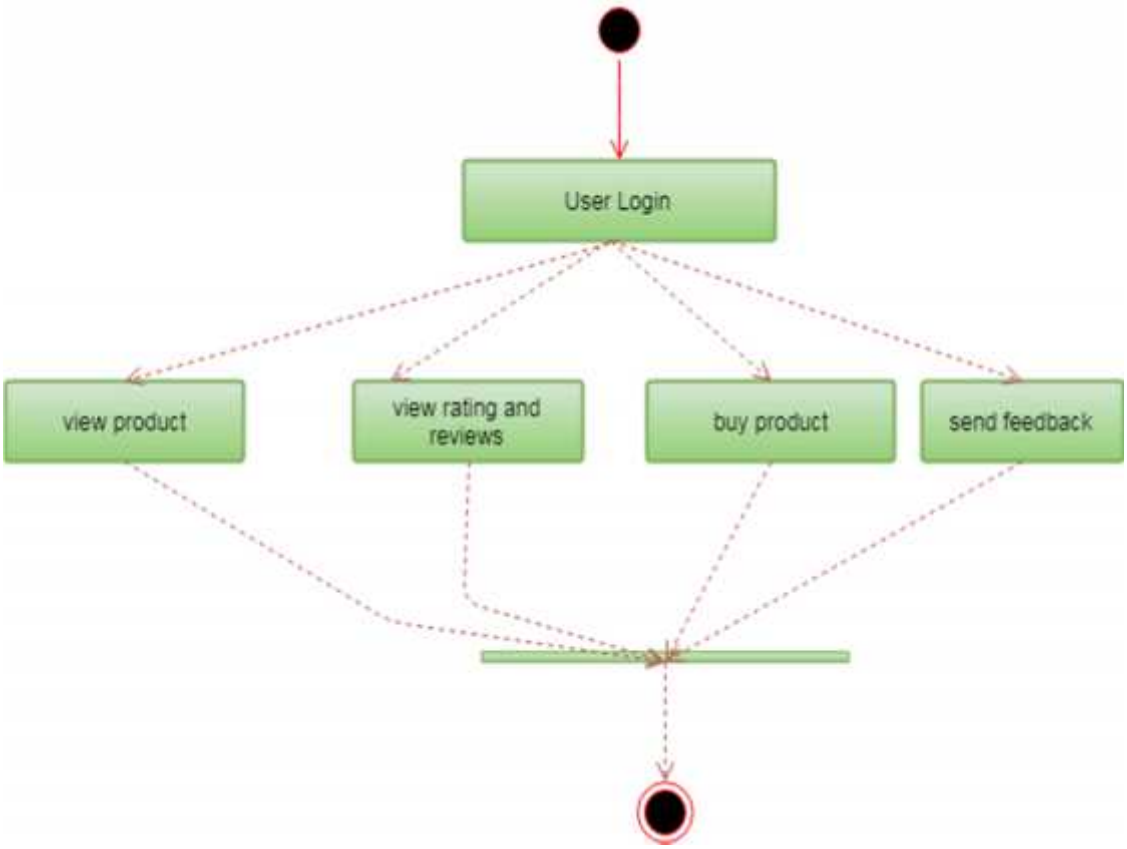


Fig. 16. User Activity Diagram

b. Admin

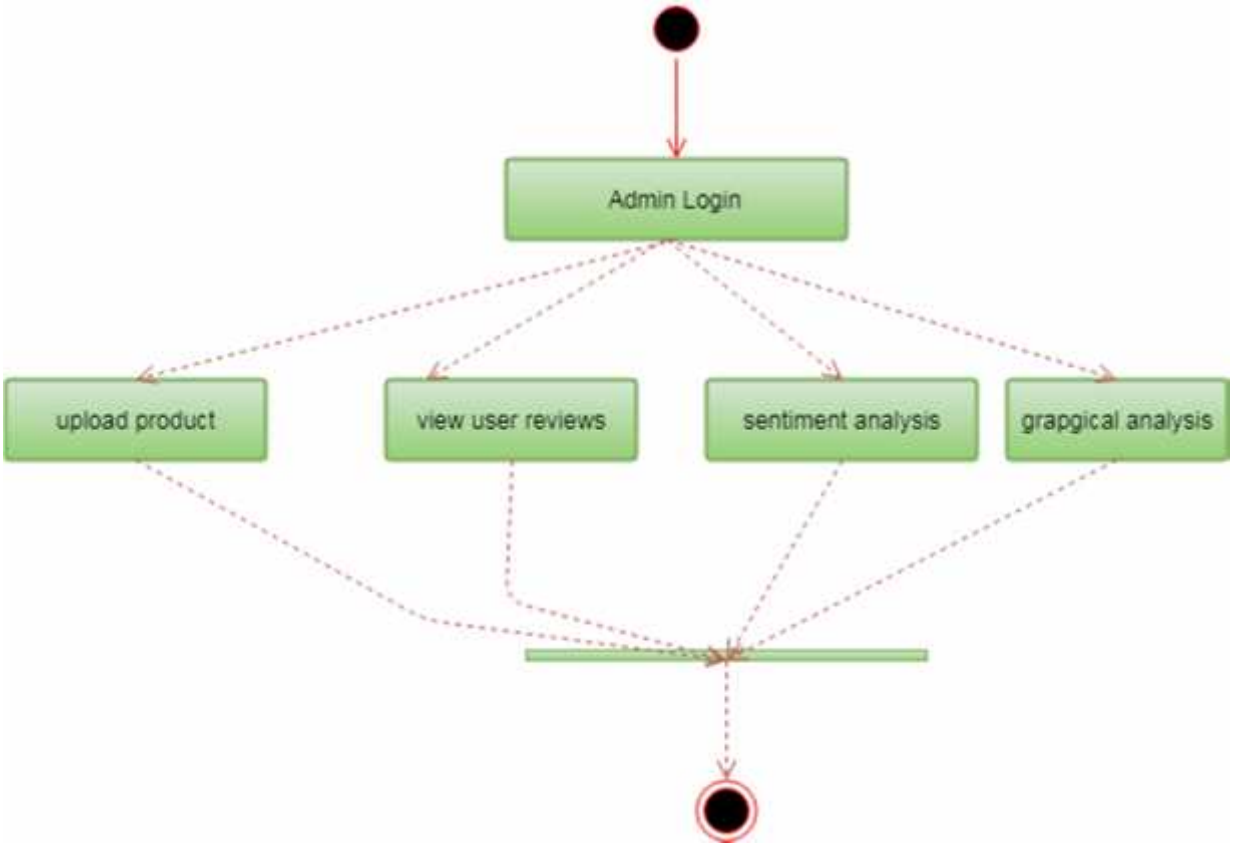


Fig. 17. Admin Activity Diagram

8. SEQUENCE DIAGRAM

a. User

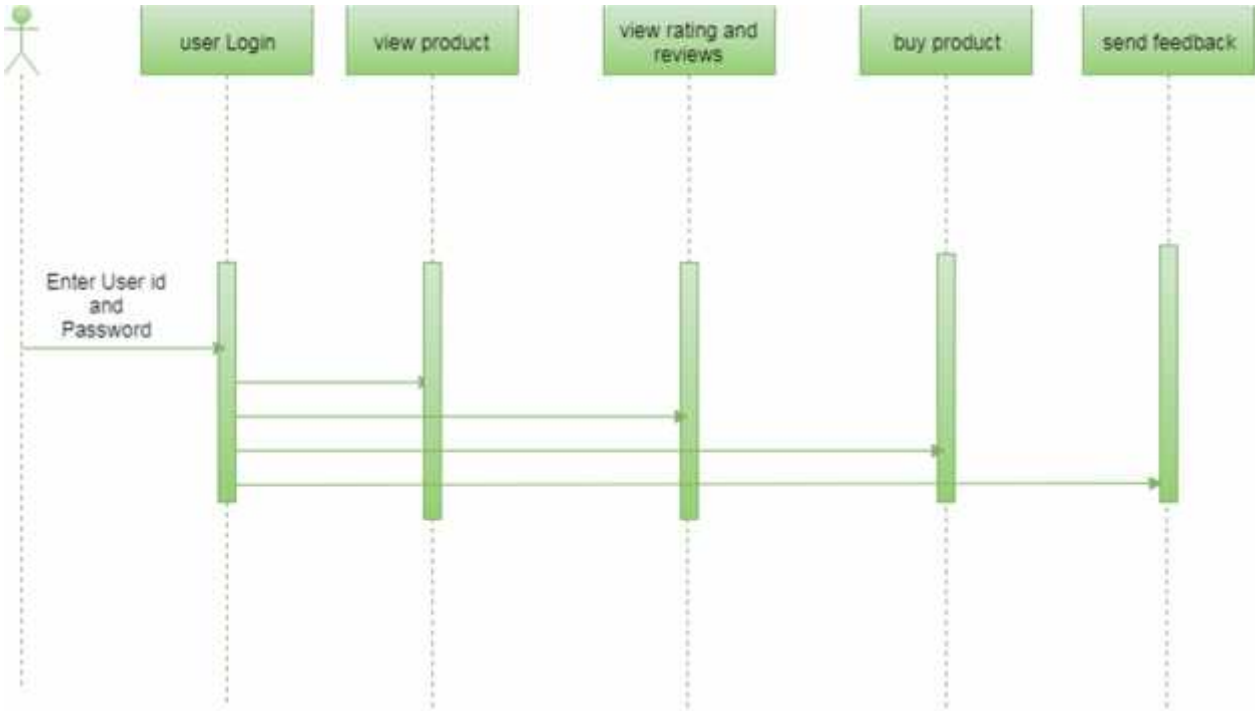


Fig. 18. User Sequence Diagram

b. Admin

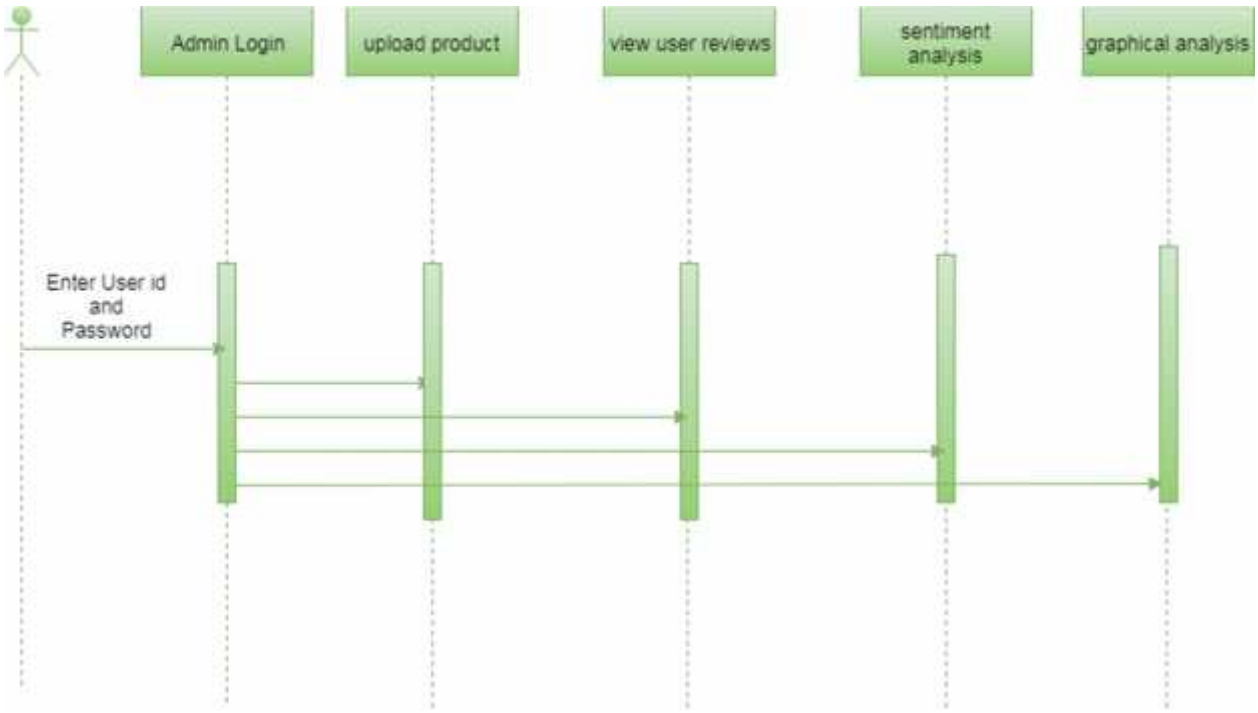


Fig. 19. Admin Sequence Diagram

6. PROJECT CODING

6.1 CODE TEMPLATES:

Admin Templates:

Import required modules

Declare global variables

def index (request):

Syntax: def index(request):
    return render(request, 'index.html', {})

Here, **request** is the url request mapping and calling the view function. **render** combines a given template with a given context dictionary. **{}** denotes the dictionary of values that can be added to the template context.

def home( ):

Syntax: def home(request):
    return render(request, 'home.html', {})

Here, **request** is the URL request mapping and calling the view function. **render** combines a given template with a given context dictionary. **{}** denotes the dictionary of values that can be added to the template context.

def uploadproducts( ):

Syntax: def home(request):
return render(request, 'uploadproducts.html', {})

Uploading the products is done by admin. Authorized person is uploading the new arrivals to system that are listed to users. Product can be uploaded with its attributes such as brand, colour, and all other details of warranty. The uploaded products are able to block or unblock by users.

def charts ( ):

Syntax: def charts(request):
return render(request, 'charts.html', {})

A Product Knowledge Graph is an e-commerce specific form of knowledge graph built to improve product find ability and end-user experiences by enriching a brand's content with data. It consists of data about products, brands, product categories, product features, reviews, hi-res images, shipping data, FAQs and a lot more. Made of structured data and extended product mark-up, injected across both editorial and product content, a product knowledge graph is built on top of the product database to link all data together combining both structured information, (for instance, the list of products for a brand) or unstructured (for example the descriptions related to a collection of products).

User Templates:

Import required modules

Declare global variables.

def index( ):

Syntax: def index(request):
return render(request, 'index.html', {})

Here, **request** is the url request mapping and calling the view function. **render** combines a given template with a given context dictionary. **{}** denotes the dictionary of values that can be added to the template context.

def home( ):

Syntax: def home(request):
return render(request, 'home.html', {})

Here, **request** is the URL request mapping and calling the view function. **render** combines a given template with a given context dictionary. **{}** denotes the dictionary of values that can be added to the template context.

def viewproducts ( ):

Syntax: def charts(request):
return render(request, 'viewproduct.html', {})

def cart( ):

Syntax: def charts(request):
return render(request, 'cart.html', {})

def addratings ( ):

Syntax: def charts(request):
return render(request, 'addratings.html', {})

def viewratings ( ):

Syntax: def charts(request):
return render(request, 'viewproduct.html', {})

There are many users who purchase products through E-commerce websites. Through online shopping many E-commerce enterprises were unable to know whether the customers are satisfied by the services provided by the firm. This boosts us to develop a system where various customers give reviews about the product and online shopping services, which in turn help the E-commerce enterprises and manufacturers to get customer opinion to improve service and merchandise through mining customer reviews.

6.2 CLASS WITH FUNCTIONALITY

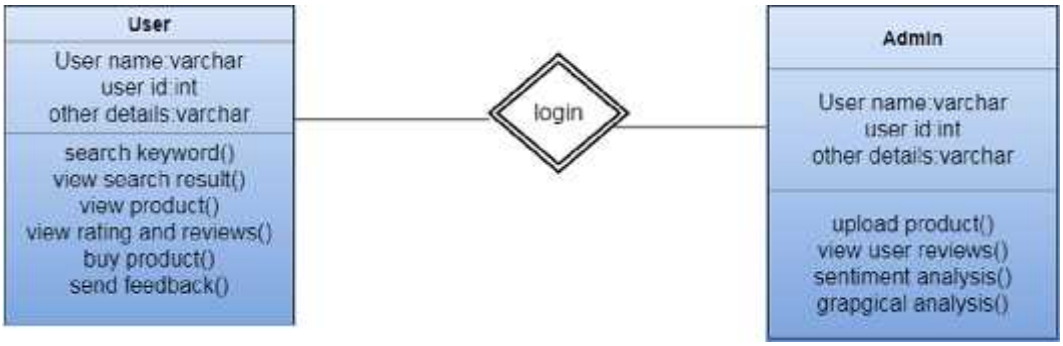


Fig.20. Class with Functionality Diagram

6.4 METHODS INPUT AND OUTPUT PARAMETERS.

def index( ):

def home( ):

def viewproducts ( ):

def cart( ):

def addratings ( ):

def viewratings ( ):

def uploadproducts( ):

def charts ( ):

NAVI BAYE'S ALGORITHM

Naïve Bayes Algorithm: In machine learning, naive Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naive) independence assumptions between the features. Naive Bayes has been studied extensively since the 1950s. It was introduced under a different name into the text retrieval community in the early 1960s, and remains a popular (baseline) method for text categorization, the problem of judging documents as belonging to one category or the other (such as spam or legitimate, sports or politics, etc.) with word frequencies as the features. With appropriate pre-processing, it is competitive in this domain with more advanced methods including support vector machines. It also finds application in automatic medical diagnosis.[3] Naive Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem. Maximum-likelihood training can be done by evaluating a closed-form expression, 718 which takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers. In the statistics and computer science literature, naive Bayes models are known under a variety of names, including simple Bayes
and independence Bayes. All these names reference the use of Bayes' theorem in the classifier's decision rule, but naive Bayes is not (necessarily) a Bayesian method

algorithm is used in this project to develop the whether the sentiment of given review is positive or negative. Based on the output of algorithm suggestion to users is given. The algorithm is applied and lists the products in user side based on the positive and negative.

Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred. Bayes' theorem is stated mathematically as the following equation:

$$P(A|B) = \frac{P(B|A) P(A)}{P(B)}$$

where A and B are events and P(B) ? 0.

Basically, we are trying to find probability of event A, given the event B is true. Event B is also termed as evidence.
P(A) is the priori of A (the prior probability, i.e. Probability of event before evidence is seen). The evidence is an attribute value of an unknown instance(here, it is event B).
P(A|B) is a posterior probability of B, i.e. probability of event after evidence is seen.
Now, with regards to our dataset, we can apply Bayes' theorem in following way:

$$P(y|X) = \frac{P(X|y) P(y)}{P(X)}$$

where, y is class variable and X is a dependent feature vector (of size n) where:

$$X = (x_1,x_2,x_3,.....,x_n)$$

Just to clear, an example of a feature vector and corresponding class variable can be: (refer 1st row of dataset)

X = (Rainy, Hot, High, False)
y = No
So basically, P(y|X) here means, the probability of "Not playing golf" given that the weather conditions are "Rainy outlook", "Temperature is hot", "high humidity" and "no wind".

7. PROJECT TESTING

Project Testing:

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

7.1 VARIOUS TEST CASES:

TYPES OF TESTS:

Unit testing:

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

Integration testing:

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

Functional test:

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is cantered on the following items:

Valid Input : identified classes of valid input must be accepted.

Invalid Input : identified classes of invalid input must be rejected.

Functions : identified functions must be exercised.

Output : identified classes of application outputs must be exercised.

Systems/Procedures: interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

System Test:

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

7.2 BLACK BOX

Black Box Testing:

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box .you cannot “see” into it. The test provides inputs and responds to outputs without considering how the software works.

Unit Testing:

Unit testing is usually conducted as part of a combined code and unit test phase of the software lifecycle, although it is not uncommon for coding and unit testing to be conducted as two distinct phases.

Test strategy and approach:

Field testing will be performed manually and functional tests will be written in detail.

Test objectives

- All field entries must work properly.
- Pages must be activated from the identified link.
- The entry screen, messages and responses must not be delayed.

Features to be tested

- Verify that the entries are of the correct format
- No duplicate entries should be allowed
- All links should take the user to the correct page.

Integration Testing:

Software integration testing is the incremental integration testing of two or more integrated software components on a single platform to produce failures caused by interface defects.

The task of the integration test is to check that components or software applications, e.g. components in a software system or – one step up – software applications at the company level – interact without error.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

Acceptance Testing:

User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

7.3 WHITE BOX TESTING

White Box Testing:

White Box Testing is a testing in which in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is purpose. It is used to test areas that cannot be reached from a black box level.

8. OUTPUT SCREENS

8.1 USER INTERFACES:

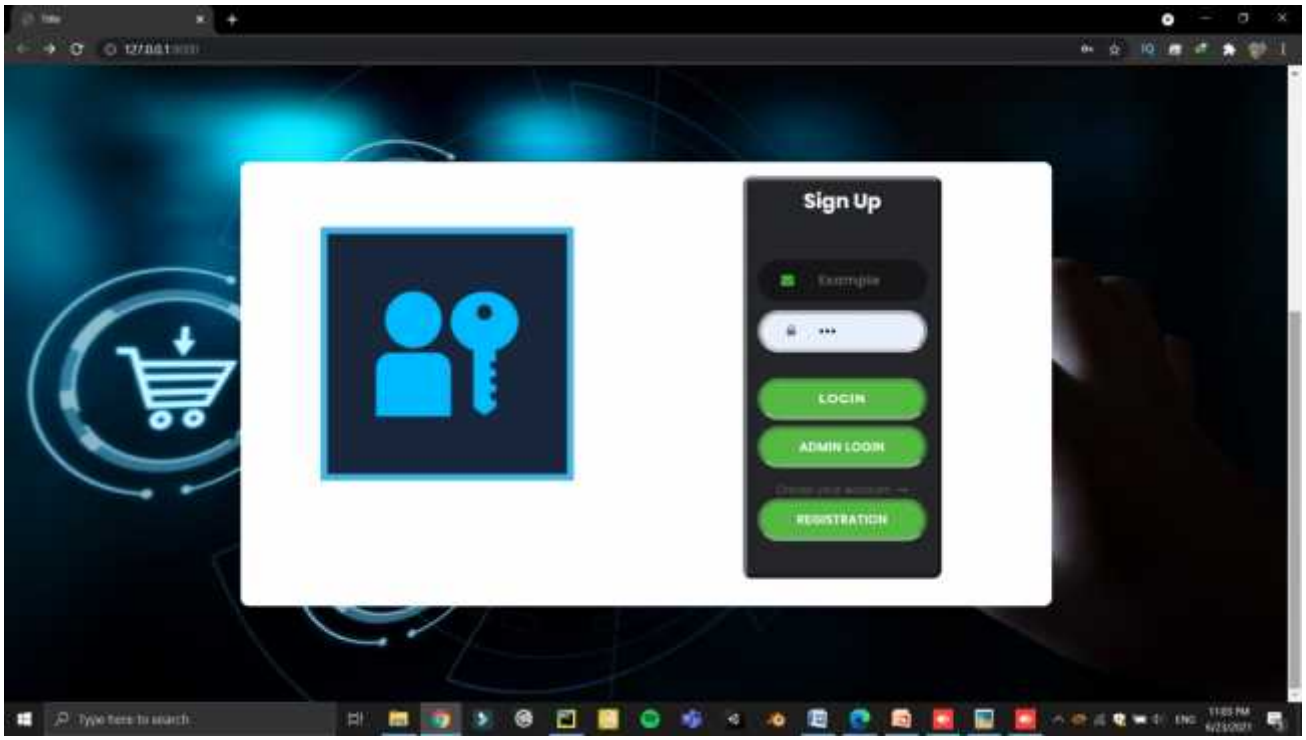


Fig.21. User Interface Login Page

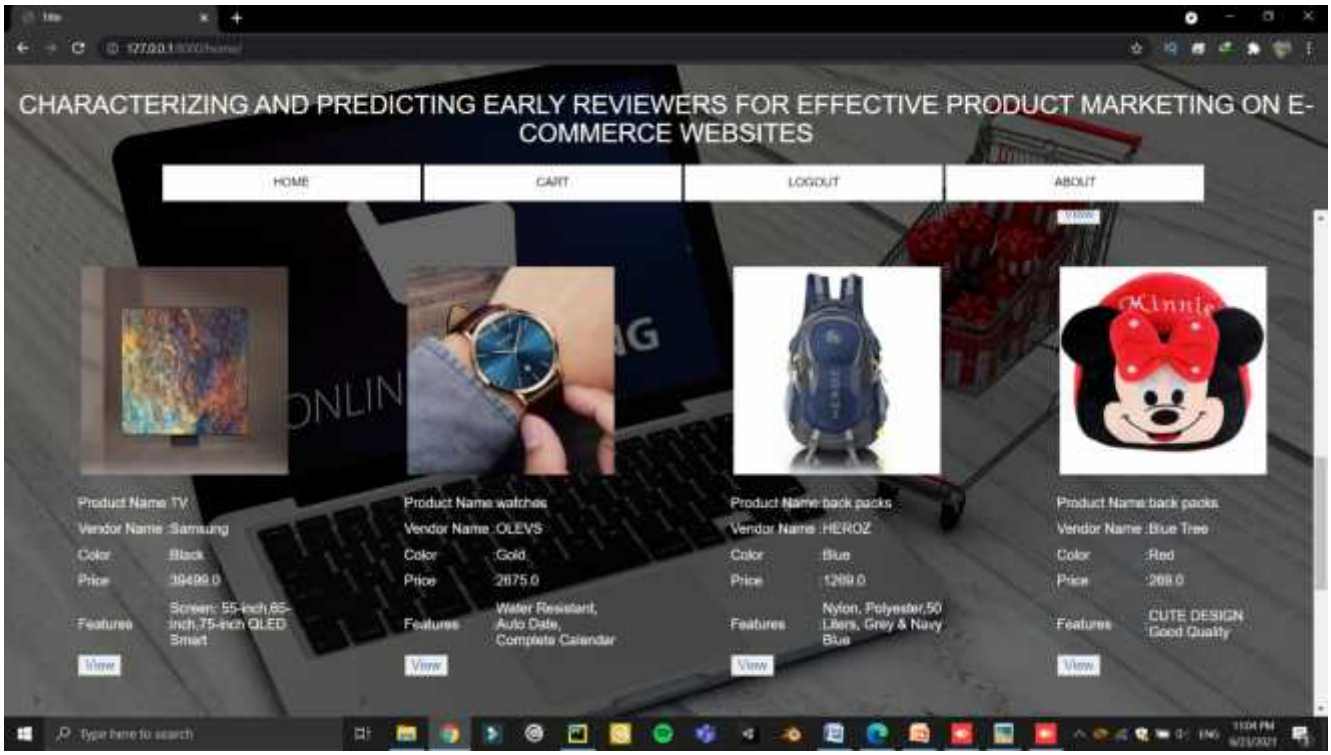


Fig.22. User Interface Home Page

Fig.23. User Interface Cart Page

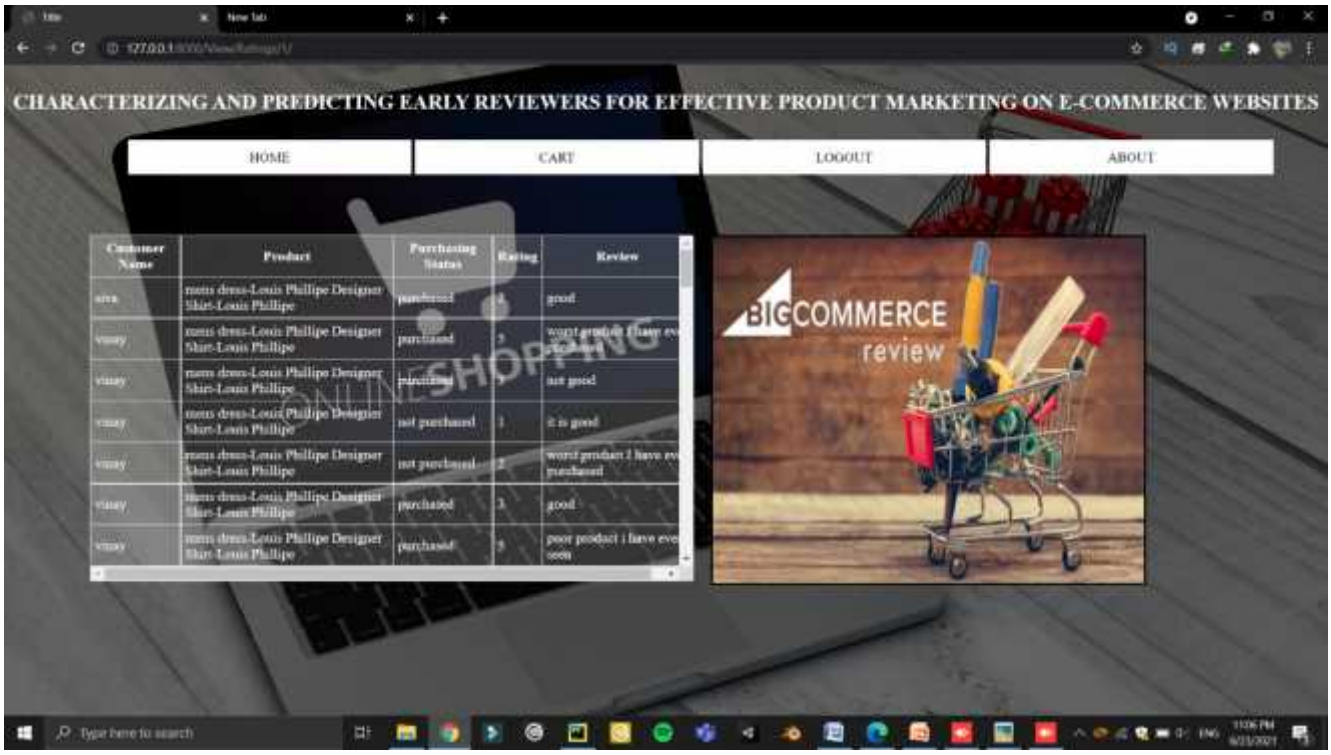


Fig.24. User Interface View Ratings Page

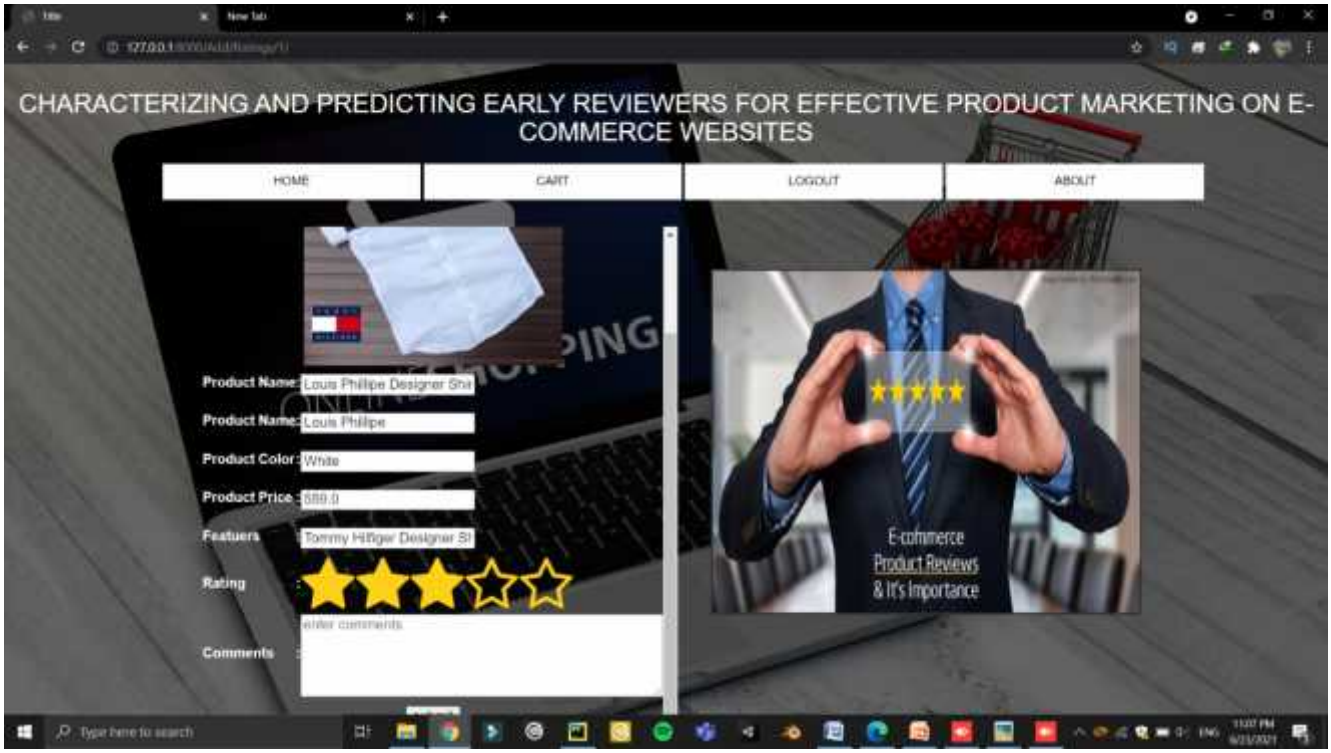


Fig.25. User Interface Add Ratings Page

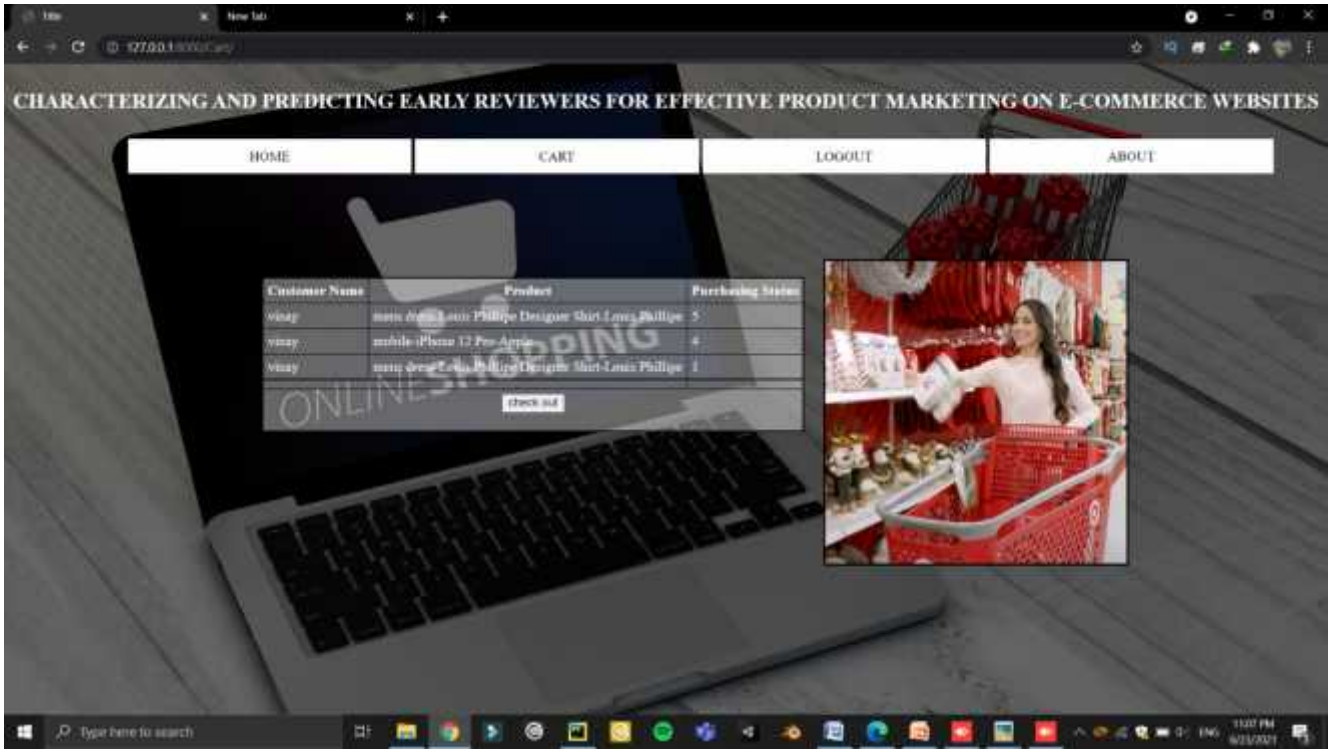


Fig.26. User Interface Items on Cart Page

8.2 OUTPUT SCREENS:

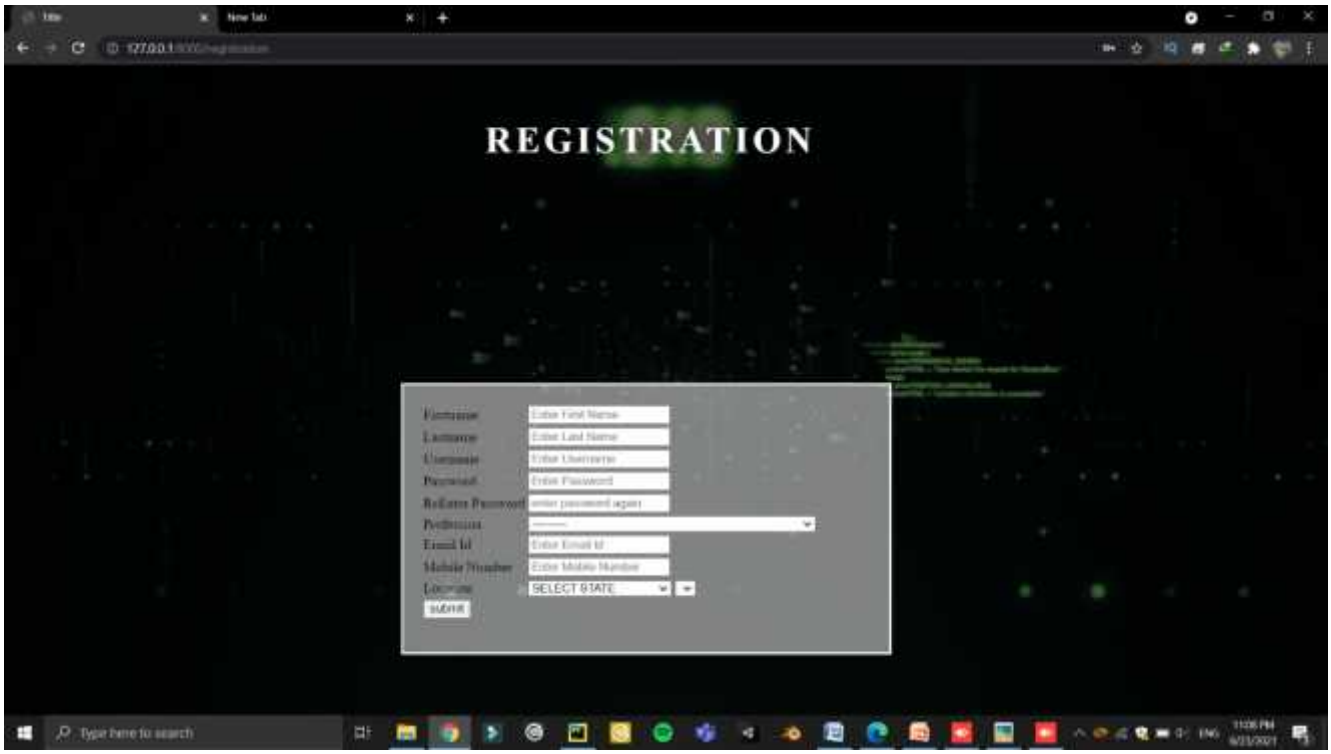


Fig.27. User Registration Form

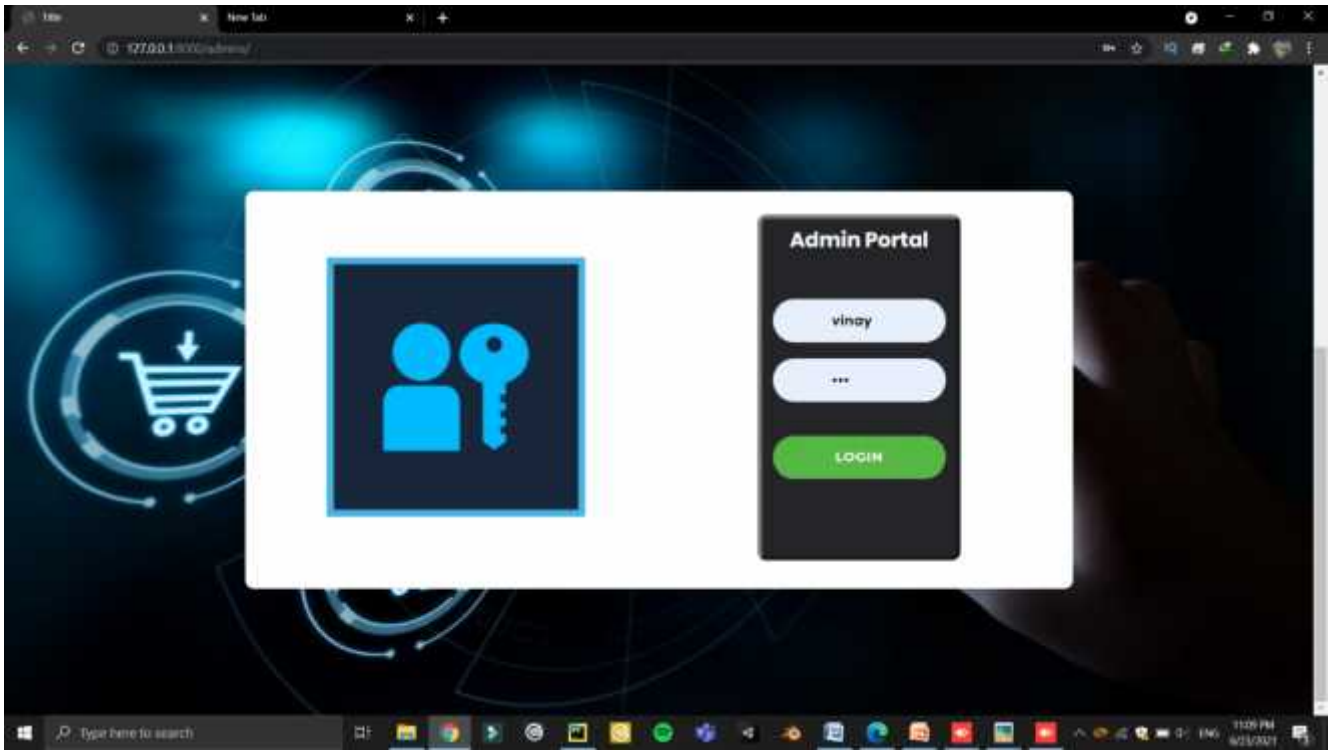


Fig.28. Admin Login

Fig.29. Upload Products

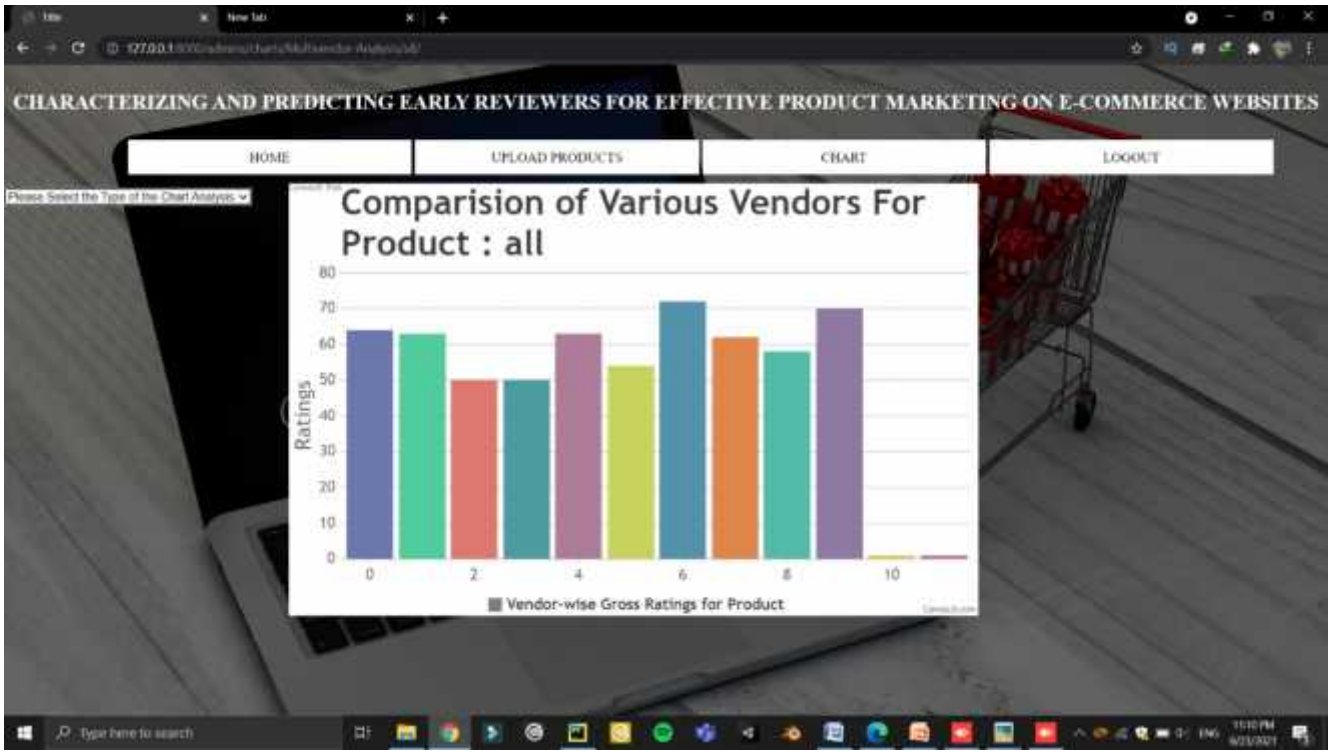


Fig.30. Comparison Various vendors for product chart

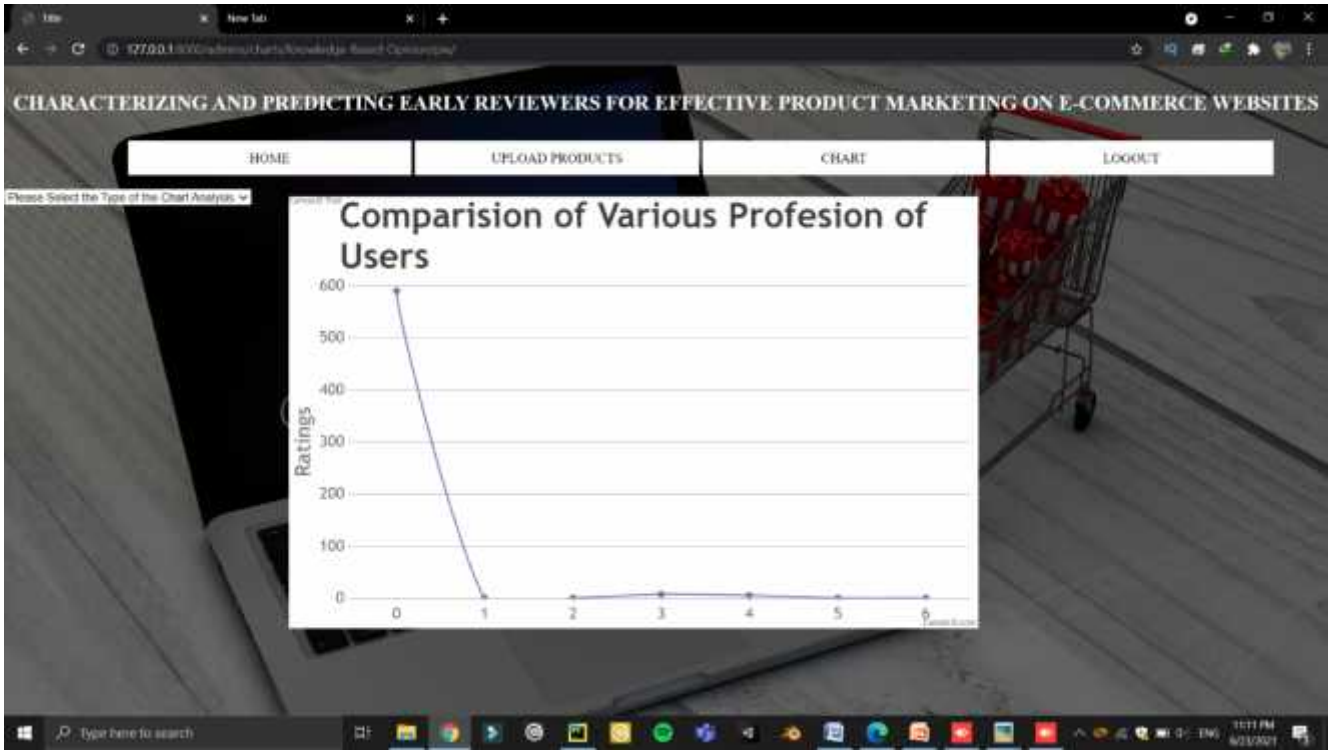
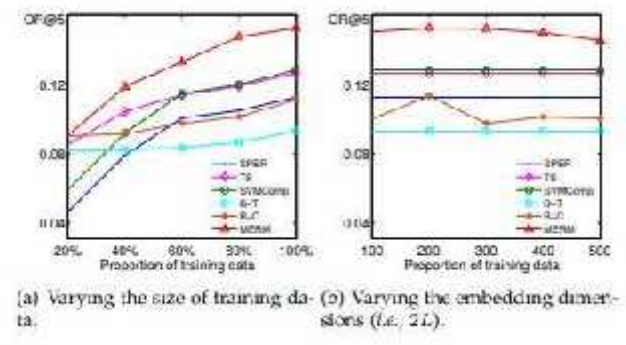


Fig.31. Comparison Various vendors profession of Users

Fig.32. Comparison of products Based on Sentiments

9. EXPERIMENTAL RESULTS

We present the results on early reviewer prediction in Table 4. It can be observed that the simplest baseline of ranking users based on the number of reviews posted before (NR) performs the worst. It indicates that users posted a large number of reviews are not necessarily active in early adoption of products. NER improves over NR, which shows that a user who has acted as an early reviewer for other products before is more likely to adopt new products in the future. PER, outperforms NER in Amazon dataset, while underperforms NER in Yelp dataset. The smoothed PER, i.e., SPER, performs better than PER. The two comparison based baselines B−T and B−C outperform the statistics−based methods only in some cases, and do not yield significant improvement. These results are consistent with the finding previously reported in [27] that a simple ratio based method works well when the training data is sufficiently large. Overall, B−C performs better than B−T. Instead of using a single value, B−C adopts a vectorized representation for modelling the player strength. Furthermore, the two competitions based methods TS and SVM Comp improve upon all the above baselines. Although SVM Comp is slightly better than TS, there is no significant difference between them. TS is a classic competition model for characterizing the player strength, while SVMComp has been shown to be effective in QA expert finding task [27]. These two methods perform best among our baselines. Our proposed model MERM achieves significant improvement in comparison to all the baselines. Compared with other baselines which only measure the earliness level of a user with a single value, MERM learns the multidimensional representation of users from comparative pairs. Although B−C also adopts a multi−dimensional representation for modelling player strength, it does not perform very well in our task. A possible reason is that B−C needs to learn more parameters (i.e., both blade vectors and chest vectors); while, in our datasets, the comparison pairs for training are sparse. The key difference of MERM is that it learns product embeddings also based on the side information involving both the title and category information of products. It effectively projects both product and user embeddings into the same continuous space for direct comparison and ranks users by optimizing a margin−based ranking objective function in a product dependent manner. In our second sets of experiments, we further examine 20% 40% 60% 80% 100% 0.04 0.08 0.12 SPER TS SVMComp B−T B−C MERM OR@5 Proportion of training data



(a) Varying the size of training data. (b) Varying the embedding dimensions (i.e., 2L).

Fig.33 Early reviewer prediction performance with different sizes of training set or embedding dimensions in Amazon dataset

(a) Varying the size of training data. 100 200 300 400 500 0.04 0.08 0.12 SPER TS SVMComp B-T B-C MERM OR@5 Proportion of training data (b) Varying the embedding dimensions (i.e., 2L).

Fig.11. Early reviewer prediction performance with different sizes of training set or embedding dimensions in Amazon dataset.

The impact of the amount of training data on the results of early reviewer prediction. We present the results of Amazon dataset; the results of Yelp dataset are similar and are omitted here. By fixing the test data at 20%, we vary the remaining 80% training data at five different splits: {20%, 40%, 60%, 80%, 100%}. The results are presented in Figure 11(a). Overall, we observe that all the methods suffer from performance drop with the decrement of training data. Our method MERM performs generally better than other methods with any amount of training data. We also vary the number of dimensions (i.e., 2L) for user and product representation in B-C and MERM, and report the results in Figure 11(b). It can be observed that the dimensionality of 200 yields the best performance.

Dataset	#Product	#User	#Pairs	ANRU	ANRP
Amazon	12,814	16,355	3,122,797	18	23
Yelp	2,545	3,912	282,718	14	22

Table:2 Statistics of the evaluation sets in early reviewer prediction. ANRU and ANRP are the abbreviations of Average Number of Reviews posted by each User and Average Number of Reviews received by each Product.

Datasets	Amazon					Yelp				
Models	OR@5	OR@10	Hit@5	Hit@10	RCCP	OR@5	OR@10	Hit@5	Hit@10	RCCP
NR	0.0910	0.1416	0.1105	0.2088	53.15%	0.0704	0.1187	0.0605	0.1110	55.26 %
NER	0.1018	0.1516	0.1260	0.2131	61.17%	0.0810	0.0982	0.1134	0.2052	60.53%
PER	0.1114	0.1577	0.1334	0.2218	64.56%	0.0738	0.0896	0.0971	0.1734	56.21%
SPER	0.1125	0.1614	0.1353	0.2261	65.31%	0.0763	0.1025	0.1063	0.2149	57.27%
BT	0.0931	0.1437	0.1120	0.2150	64.31%	0.0864	0.0939	0.1044	0.1859	59.89%
BC	0.1132	0.1635	0.1361	0.2290	62.23%	0.0931	0.1016	0.1120	0.1952	59.36%
TS	0.1265	0.1720	0.1450	0.2465	67.54%	0.0904	0.1013	0.1353	0.2300	59.82%
SVMComp	0.1283	0.1747	0.1483	0.2503	67.97%	0.0955	0.1045	0.1341	0.2201	60.13%
MERM	**0.1524***	**0.2273***	**0.1665***	**0.2823***	**69.25%***	**0.1212***	**0.1333***	**0.1462***	**0.2360***	**68.57%***

Note: " * " indicates the statistically significant improvements (i.e., two side t-test with $p < 0.01$) over the best baseline.

Table:3 Performance comparison on the results of early reviewer prediction.

10. CONCLUSION

In this paper, we have studied the novel task of early reviewer characterization and prediction on two real-world online review datasets. Our empirical analysis strengthens a series of theoretical conclusions from sociology and economics. We found that (1) an early reviewer tends to assign a higher average rating score; and (2) an early reviewer tends to post more helpful reviews. Our experiments also indicate that early reviewers' ratings and their received helpfulness scores are likely to influence product popularity at a later stage. We have adopted a competition-based viewpoint to model the review posting process, and developed a margin based embedding ranking model (MERM) for predicting early reviewers in a cold-start setting. In our current work, the review content is not considered. In the future, we will explore effective ways in incorporating review content into our early reviewer prediction model. Also, we have not studied the communication channel and social network structure in diffusion of innovations partly due to the difficulty in obtaining the relevant information from our review data. We will look into other sources of data such as Flixster in which social networks can be extracted and carry out more insightful analysis. Currently, we focus on the analysis and prediction of early reviewers, while there remains an important issue to address, i.e., how to improve product marketing with the identified

early reviewers. We will investigate this task with real e-commerce cases in collaboration with e-commerce companies in the future.

11. REFERENCES

[1] J. McAuley and A. Yang, “Addressing complex and subjective product-related queries with customer reviews,” in WWW, 2016, pp. 625–635.

[2] N. V. Nielsen, “E-commerce: Evolution or revolution in the fastmoving consumer goods world,” nngroup. com, 2014.

[3] W. D. J. Salganik M J, Dodds P S, “Experimental study of inequality and unpredictability in an artificial cultural market,” in ASONAM, 2016, pp. 529–532.

[4] R. Peres, E. Muller, and V. Mahajan, “Innovation diffusion and new product growth models: A critical review and research directions,” International Journal of Research in Marketing, vol. 27, no. 2, pp. 91 – 106, 2010.

[5] L. A. Fourt and J. W. Woodlock, “Early prediction of market success for new grocery products.” Journal of Marketing, vol. 25, no. 2, pp. 31 – 38, 1960.

[6] B. W. O, “Reference group influence on product and brand purchase decisions,” Journal of Consumer Research, vol. 9, pp. 183–194, 1982.

[7] J. J. McAuley, C. Targett, Q. Shi, and A. van den Hengel, “Image based recommendations on styles and substitutes,” in SIGIR, 2015, pp. 43–52.

[8] E. M.Rogers, Diffusion of Innovations. New York: The Rise of HighTechnology Culture, 1983.

[9] K. Sarkar and H. Sundaram, “How do we find early adopters who will guide a resource constrained network towards a desired distribution of behaviors?” in CoRR, 2013, p. 1303.

[10] D. Imamori and K. Tajima, “Predicting popularity of twitter accounts through the discovery of link-propagating early adopters,” in CoRR, 2015, p. 1512.

[11] X. Rong and Q. Mei, “Diffusion of innovations revisited: from social network to innovation network,” in CIKM, 2013, pp. 499– 508.

[12] I. Mele, F. Bonchi, and A. Gionis, “The early-adopter graph and its application to web-page recommendation,” in CIKM, 2012, pp. 1682–1686.

[13] Y.-F. Chen, “Herd behaviour in purchasing books online,” Computers in Human Behaviour, vol. 24(5), pp. 1977–1992, 2008.

[14] Banerjee, “A simple model of herd behaviour,” Quarterly Journal of Economics, vol. 107, pp. 797–817, 1992.

[15] A. S. E, “Studies of independence and conformity: I. a minority of one against a unanimous majority,” Psychological monographs: General and applied, vol. 70(9), p. 1, 1956.

[16] T. Mikolov, K. Chen, G. S. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” in ICLR, 2013.

[17] A. Bordes, N. Usunier, A. Garc´ıa-Duran, J. Weston, and ´ O. Yakhnenko, “Translating embeddings for modeling multirelational data,” in NIPS, 2013, pp. 2787–2795.

[18] A. S. E, “Studies of independence and conformity: I. a minority of one against a unanimous majority,” Psychological monographs: General and applied, vol. 70(9), p. 1, 1956.

[19] M. L. S. D. X. W. L. S. Mingling Chen, Qingguo Ma, “The neural and psychological basis of herding in purchasing books online: an event-related potential study,” Cyber psychology, Behaviour, and Social Networking, vol. 13(3), pp. 321–328, 2010.

[20] V. G. D. W. Shih-Lun Tseng, Shuya Lu, “The effect of herding behaviour on online review voting participation,” in AMCIS, 2017.

10. PUBLICATIONS

JOURNAL (UGC approved Journal)

CONFERENCE (International Conference on “Innovations in Computers Networks, Computational Intelligence and IOT” [ICICCI-21]).

PAPER ID: ICICCI-21-01128

TOPIC: CHARACTERIZING AND PREDICTING EARLY REVIEWERS FOR EFFECTIVE PRODUCT MARKETING ON E-COMMERCE WEBSITES.

.

NETI ROHIT KUMAR

Neti Rohit Kumar is pursuing his Bachelor of Technology in the stream of Computer science and engineering at St. Martin's Engineering College. He completed his intermediate from Sri Gayatri Junior college and schooling from Abhyudaya High School. His technical skills include C, C++, Java, HTML, CSS, JavaScript and Python. His area of interest is Data Science and web development. He has completed few certificate courses from online platforms like Coursera on Python Programming, Machine Learning, HTML, CSS, Leadership and Emotional Intelligence. His participations include National Level Three Day Online Workshop on "AI & ML in speech and

audio processing" which was conducted from 10th and 12th December, 2020, Leadership Talk with Mr. Mahesh Babu CEO Mahindra Electric Mobility Ltd. Leadership Talk With Dr. Nilesh N Oak, Expert (Indian Civilization & History). Participated in the IINNOVATE 1 MILLION SECONDS ONLINE HACKATHON by (TSIC) and Telangana Information Technology Association (TITA).

3

NEERADI SUNIL RAJ

Neeradi Sunil raj is pursuing his Bachelor of Technology in the stream of Computer science and engineering at St. Martin's Engineering College. He completed his intermediate from Sri Chaitanya Junior college and schooling from Little flower High School. His technical skills include C, C++, Java, HTML, CSS, JavaScript and Python. His area of interest is Data Science and web development. He has completed few certificate courses from online

platforms like Coursera on Python Programming, Machine Learning, HTML, CSS, Leadership and Emotional Intelligence. His participations include National Level Three Day Online Workshop on “AI & ML in speech and audio processing” which was conducted from 10th and 12th December, 2020, Leadership Talk with Mr. Mahesh Babu CEO Mahindra Electric Mobility Ltd.

RIMMALA PRERANA

Prerana Rimmala is pursuing his Bachelor of Technology in the stream of Computer science and engineering at St. Martin’s Engineering College. She completed his intermediate from Narayana Junior college and schooling from St. Peter’s High School. His technical skills include C, C++, Java, HTML, CSS, JavaScript and Python. She area of interest is Data Science and web development. She has completed few certificate courses from online

platforms like both Cursa and Coursera on Python Programming, Machine Learning, HTML EJ Media, Data Journalism and Fundamentals, Data Science maths skills, Managing Project Risks and Changes, AWS Fundamentals going cloud native, AI for Everyone, Technical Analysis Leadership and Emotional Intelligence, Womens in Cyber Security. She participations include National Level Three Day Online Workshop on “AI & ML in speech and audio processing” which was conducted from 10th and 12th December, 2020.

D CHAITANYA

D Chaitanyais pursuing his Bachelor of Technology in the stream of Computer science and engineering at St. Martin’s Engineering College. He completed his intermediate from Sri Gayatri Junior college and schooling from Kabson’s ZP High School. His technical skills include C, C++, Java, HTML, CSS, JavaScript and Python. His area of

interest is Data Science and web development. He has completed few certificate courses from online platforms like Coursera on Python Programming, Machine Learning, HTML, CSS, Leadership and Emotional Intelligence. His participations include National Level Three Day Online Workshop on “AI & ML in speech and audio processing” which was conducted from 10th and 12th December, 2020, Leadership Talk with Mr. Mahesh Babu CEO Mahindra Electric Mobility Ltd. He completed few certification courses from online platforms like Coursera, CursaApp and SoloLearn.

APPENDICES

A

PROJECT REPORT

On

Agri Seva App : platform for farmers, agriculture, farming information and services providing app

Submitted by

1)Mr. Shubham Agarwal (17K81A05H6)
2)Mr. Koushik Bhargava (17K81A05H3)
3)Mr. Surja Sharma (17K81A05H8)
4)Mr. Akash Singh Rawat (17K81A05C3)

in partial fulfillment for the award of

the degree of

BACHELOR OF TECHNOLOGY

IN

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Under the Guidance of

Dr. N. Satheesh

B.E., M.E., Ph.D.

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ST.MARTIN'S ENGINEERING COLLEGE
An Autonomous Institute
Dhulapally, Secunderabad – 500 100

JUNE 2021

BONAFIDE CERTIFICATE

This is to certify that the project entitled **Agri Seva App : platform for farmers, agriculture, farming information and services providing app**, is being submitted by 1.**Mr. Shubham Agarwal 17K81A05H6, 2.Mr. Koushik Bhargava 17K81A05H3, 3.Mr. Suraj Sharma 17K81A05H8, 4.Mr. Akash Singh Rawat 17K81A05C3** in partial fulfillment of the requirement for the award of the degree of **BACHELOR OF TECHNOLOGY IN Computer Science and Engineering** is recorded of bonafide work carried out by them. The results embodied in this report have been verified and found satisfactory.

Signature

Dr. N. Satheesh

Department of CSE

Head of the Department

Dr.M.NARAYANAN

Department of CSE

Internal Examiner

External Examiner

Place: Secunderabad

Date: -06-2021

DECLARATION

We, the student of **Bachelor of Technology** in Department of **'Computer Science and Engineering'**, session: **2017 – 2021**, **St. Martin's Engineering College**, Dhulapally, Kompally, Secunderabad, hereby declare that work presented in this Project Work entitled **Agri Seva App : platform for farmers, agriculture, farming information and services providing app** is the outcome of our own bonafide work and is correct to the best of our knowledge and this work has been undertaken taking care of Engineering Ethics. This result embodied in this project report has not been submitted in any university for award of any degree.

Shubham Agarwal 17K81A05H6
Koushik Bhargava 17K81A05H3
Suraj Sharma 17K81A05H8
Akash Singh Rawat 17K81A05C3

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible and whose encouragement and guidance have crowded effects with success.

We extended our deep sense of gratitude to Principal, Dr. P. SANTOSH KUMAR PATRA, St. Martin's Engineering College, Dhulapally, for permitting us to undertake this project.

We are also thankful to **Dr. M.NARAYANAN**, Head of the Department, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

We would like to thank **Dr. T. POONGOTHAI,** Department of Computer Science and Engineering (AI & ML) for her encouragement and insightful comments and as well as our project coordinators **Dr. B.RAJALINGAM**, Associate Professor and **Dr. N. SATHEESH**, Professor, in the Department of Computer Science and Engineering for their valuable support.

We would like to express our sincere gratitude and indebtedness to our project supervisor **Dr. N. Satheesh, B.E., M.E., Ph.D.**, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

Finally, we express thanks to all those who have helped us successfully to complete this project. Furthermore, we would like to thank our family and friends for their moral support and encouragement.

We express thanks to all those who have helped us in successfully completing the project.

Shubham Agarwal 17K81A05H6

Koushik Bhargava 17K81A05H3

Suraj Sharma 17K81A05H8

Akash Singh Rawat 17K81A05C3

ABSTRACT

A project work was undertaken under rural India agricultural development with an initiative and intention of providing easily accessible informational resources, services and to heighten awareness, Agri Seva App is a platform for farmers and for any other people working in agri and farming sector, it has various features through which it aim to provide agricultural information and services using technology, the main aim of the research is to facilitate the farmers by educating them by using a web-app or a mobile-app or a kiosk machine that will provide them information on farming, right usage of related commodities, suitable weather and climatic conditions and other services like, real-time pricing, expert consultation, and agri store.

In the present scenario with advent in technology where different sectors in the world are experiencing an era of advancement in technology along with ease of access to information and services yet the farming and agri sector is lagging behind even though being one of the most important sectors where humans serve for humans and is a major sector supporting the means of human life, survival and social prosperity, besides in a country like INDIA where agriculture supports the economy and where share of agriculture in GDP is continuously increasing ie. 20.19 Percent in jun, 2021 from 19.9 percent in 2020-21 and 17.8 percent in 2019-20, we think we have a lot more to offer to the agriculture-farming sector and vice versa.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF OUTPUT SCREENS

LIST OF ACRONYMS

CNN	Convolutional Neural Networks
LR	Logistic Regression
CPU	Central Processing Unit
RAM	Random Access Memory
GB	Giga Bytes
IE	Internet Explorer(browser)
WWH	Why? What? How?
UI	User Interface
MBPS	Megabyte Per Second
ATM	Automated Teller Machine
R&D	Research and Development
DFD	Data Flow Diagram
JS	Java Script
HTML	Hypertext Markup Language
CSS	Cascading Style Sheets
API	Application Programming Interface
AI	Artificial Intelligence
CAGR	Compound Annual Growth Rate

1. INTRODUCTION

1.1 PROJECT OVERVIEW

Along with a rapid rise in consumers and increasing complexities and problems related to farming there is a need for enhancement in the existing agricultural sector. Therefore, the project Agri Seva App was undertaken with an initiative and intention of providing easily accessible informational resources, services and to heighten awareness.

After understanding the problem statement and a strict adherence to the same plus with principled use of technology, methodology the project “Agri Seva App” aims to solve current and upcoming problems in farming sector and not just invent a solution but also focuses on timely delivery of information and services to its users, and to make this possible it makes use of cutting edge technologies in the development of product while maintaining simplicity in design, architecture part and a user-friendly and handy interface.

Next challenge after inventing a proper solution was to deliver the product and to make sure that the product has some relevance, can create impact. The usage of agri apps and smart mobile phones in India is uninterruptedly growing(as seen in below graphs), consequently delivering the Agri Seva App to the user via a web-app and smart mobile devices is the finest method calculated.

.

Figure 1 : Compound annual growth rate of precision farming
(image source : mobindustry)

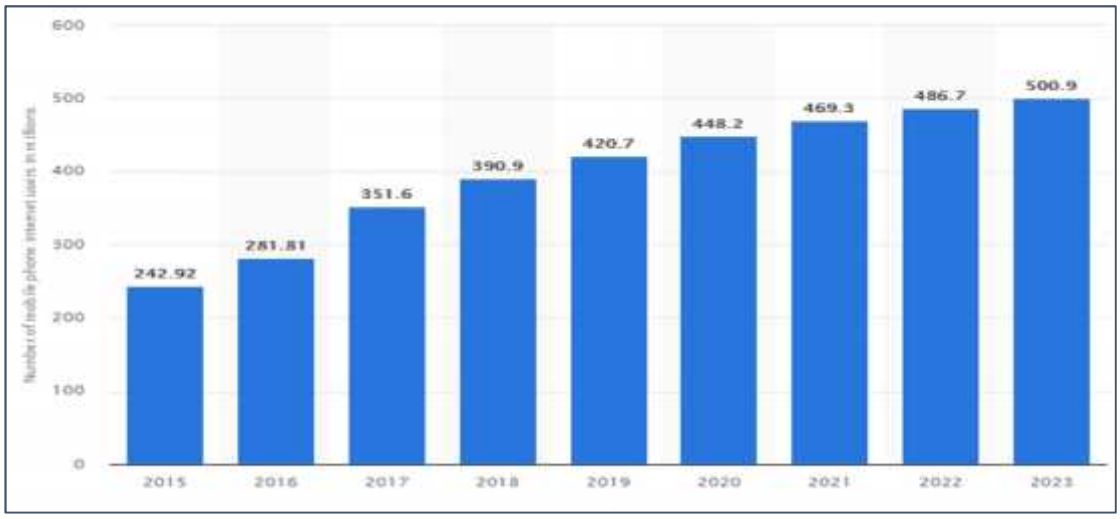


Figure 2 : Mobile phone internet users in India 2015-2023
(image source:statista.com)

1.2 PROJECT OBJECTIVE

Agri Seva App is a platform for farmers and for any other people working in agri and farming sector, it has various features through which it aim to provide agricultural information and services by means of internet and use of technology the main aim of the research is to facilitate the farmers by educating them by using a web-app on a smart mobile device

1.3 ORGANIZATION OF CHAPTERS

Step by step simplification on why Agri Seva App was developed, what was the goal and motivation, how the idea was invented and developed, what are the technologies employed, lastly classification on a variety of current and future problems it could solve is covered in detail in following parts of this project document.

1. Problem overview and definition.
2. A catalog of different modules, features and functionalities.
3. major requirements, system design, algorithms.
4. Code templates, pictorial representation, methods/functions.
5. Maintenance and testing.

2. LITERATURE SURVEY

2.1 SURVEY ON BACKGROUND

Agri Seva App was purely categorized and developed based on WWH divided into three categories. why? what? and how? thorough research was done so that to get the best out of available resources, some of the major findings are discussed below,

Why? - Problem identification, scope, need for Agri Seva App.
What? - proposed project, problems resolved, features and functionalities, product compliance.
How? - Technologies employed, deployment and delivery strategy, reliability and sustainability.

Proper identification and understanding of problem statement, what problems we wanted to solve were some if the challenges we faced at initial stages, we identified an urgent need for a portal that could not only act as a reliable and accurate source of information and services but to deliver the information and services to the user in time, as late delivery would degrade the importance of information at times, hence providing a the product as a web-app which is accessible on a smart mobile device was the finest option found, then the selection of options and features on the portal has to be sought after in order to maintain the complexity in using the portal at minimum levels, then the technologies, frameworks, programming languages, and services were finalized keeping the need for design changes, easy modifiability, upgradeability in mind, as a result we could come up with a product like Agri Seva App loaded with all its unique and essential features are discussed in further segments.

2.2 CONCLUSION ON SURVEY

In pursuit of developing Agri Seva App we have referred and studied fifteen plus research papers, journals, we have gone through a variety of portals and other similar projects by peer members before and during the project development, some explained the problem statement precisely while others contained good knowledge on future scope, however, most of them required a proper strategy on how the product will be delivered to user and in what ways a user will be able to access the product, lack of sustainable outreach plan was an additional drawback identified, nevertheless we as Agri Seva Team has tried our best to overcome mentioned drawbacks and have come up with a quintessential product like Agri Seva App.

3. SOFTWARE AND HARDWARE REQUIREMENTS

3.1 SOFTWARE REQUIREMENTS

Desktop devices(non-mobile devices):

Operating System : Windows(latest is preferred)
Browser: chrome, IE, Firefox

Other smart devices(mobile devices):

Operating System : Android or iOS
Browser: chrome, Safari

3.2 HARDWARE REQUIREMENTS

Desktop devices(non-mobile devices):

Processor : 2 core CPU
RAM : 4 gb
Storage : 2 gb(considering downloadable content: Pdfs, images, catalogs)
Internet connection : 8 mbps(the higher the better)

Other smart devices(mobile devices):

Processor : 4 core CPU
RAM : 4 GB
Storage : nill for the portal, 2 gb(for browser and downloadable content)
Internet connection : 8 mbps(the faster the better)

4. SOFTWARE DEVELOPMENT ANALYSIS

4.1 OVERVIEW OF PROBLEM

Following paragraph covers basic problem statements and the need for undertaking following project work, though many NGOs, innovators/ entrepreneurs, research institutes and krushi vikas organizations are trying hard to provide farming, agricultural information and services still the information and services are not easily accessible to the people in that sector. Hence, there is a strong need for projects like Agri Seva App.

4.2 DEFINE THE PROBLEM

Some of the major reasons we identified during this project and research is lack of proper medium and channel, reliability, and accessibility via which the information cannot only be delivered but delivered in time, as old and expired information would be worthless, some other problems that we identified are lack of standardized pricing, lack of guidance and consultation when needed, lack of genuine sources for purchasing agri commodities like seeds, fertilizers, pesticides, and majorly no platform which provides all mentioned services combined in one app or portal. Here Agri Seva App comes into the picture and aims to solve mentioned problems by sustainable, user-friendly use of technology and by means of the internet.

4.3 MODULES OVERVIEW

Unique features of Agri seva App are its major strength, the goal of implementing such features is to make the farmer self sustainable by not only providing information by also to educate them and making them self-reliant to overcome future challenges, features namely: crop information, agri news and schemes, weather forecast, pricing, agri store, chat with an expert.

4.4 DEFINE THE MODULES

In the following paragraph, let us now understand all the modules, features and functionalities available on Agri Seva App and get ourselves familiar to different terminologies related to it, hence, the various features are: Crop information, Agri-news and Schemes, Weather forecast, Real time standardized pricing, Agri-store, Expert consultation with Artificially intelligent voice assistant and Drop query in any language to hear back from an agri expert option, all in all the features and information will be made available via a web-app, a handy mobile app and also via a kiosk machine which is very similar to an ATM Machine, the option of kiosk machine is added for a strong reason, being that there are a number of farmers who are still very below the poverty level and are minimum wage earners, how could we expect them to have a smart device with an internet subscription, hence, as this particular category of farmers would find it difficult to access the services we have whole unique concept of kiosk machines for them which they can use and gain all that insightful information and educate themselves equally to any other farmer. Therefore, the first person to get benefited by the proper use of Agri Seva App would be the farmer, and using such latest tools, techniques and methodologies adequately will lead to holistic development of farmers and a gradual increase in profitability and overall crop quality will be experienced.

4.5 MODULE FUNCTIONALITY

In following segments a brief description of all the features and functionalities are discussed,

Homepage is the first and main screen of the Agri Seva App which will be displayed when the user logs on to the portal, using this screen farmer will be able to scroll through the different feature of the app, such as Crop information/ selection/ information, Agri- News and Schemes, Weather forecast, Current market pricing, Agri store, Expert support and the other features of the app.

Module no 1 : Crop Information - this is a page which will cover all informational aspects, information on prevailing practices and methodologies, from sowing a crop to bowing or harvesting it followed by an informative and insightful tutorial video as we believe pictorial and added visual oriented method is the most effective in not only transferring the knowledge but also in retaining knowledge for longer time, particularly details like: Seed Information, Cultivation procedure, Crop Nutrition, Climatic Requirements, use of right fertilizers with quantity and information Crop Protection and quality improvement

Module no 2 : Agri News and Schemes - using this feature, a farmer will be able to receive information on latest schemes and policies introduced by the government and news related to the agri sector so that the farmer can keep itself up to date at any point in time.

Module no 3 : Simple weather forecast, as we see, weather and climatic conditions these days are not stable and, in fact, are very fluctuating. In this situation it becomes difficult to make right decisions about the right selection of crops and

whether the upcoming climatic conditions will be suitable for that particular crop or not. Hence, Agri Seva App makes it possible to understand and make right selection of crop, and it makes it possible using the weather forecast feature which provides a list of suitable crops and guides the farmer with the best selection of crop, this feature is making use of Global Forecast System for its first part of data on which the model will perform analysis, Convolutional Neural Network(CNN) as core algorithm

Module no 4 : pricing, with the help of pricing options, farmers can know the latest pricing of any crop selected, and this in turn will help the farmer to sell their crops, grains or pulses at the right price put down by the government. Pricing feature is still under R&D as we are looking for reliable sources for real time pricing data, the current plan is to modify the pricing manually in a cycle of 24 hours until the automation is done.

Module no 5 : Agri Store, this feature is meant to support both farmers and Agri commodity vendors, sellers, for farmers agri store will act as a reliable source to purchase high quality seeds, fertilizers etc, and for vendors it will act as a marketplace to sell heir agri products, only products which are of high quality and gauged to produce top quality crop will be available to maintain overall quality aspect of Agri Store.

Module no 6 : Expert consultation, this feature enables user to drop their queries to get in touch with an agri consultant, with the use of mentioned feature farmer can seek help from an agri expert associated with Agri Seva App, time being this option comes with limited language support, but plans have been made to overcome this limitation with addition of AI assistant in future which could enable us to provide quick and immediate resolution to user queries. In current version of feature farmer can record its query as a message in any language and sent it along with its name, mobile number and query type using the form available within the option on the portal in addition to this an id will be generated which shall be preserved by the user for future references or follow ups related to previous queries, after filling and submitting the form, as a result the farmer will get a call back from an agri expert as early as possible, he/she will resolve all the queries, educate, guide farmer and provide solution to the problems that the farmer might have or might encounter in future. Resolution to queries and providing genuine guidance is the major goal of Expert consultation/ chat with an expert option.

5. PROJECT SYSTEM DESIGN

5.1 Data Flow Diagrams

Data flow diagram/process flow 1 : Above Data Flow Diagram gives us an idea about how crop information can be accessed.

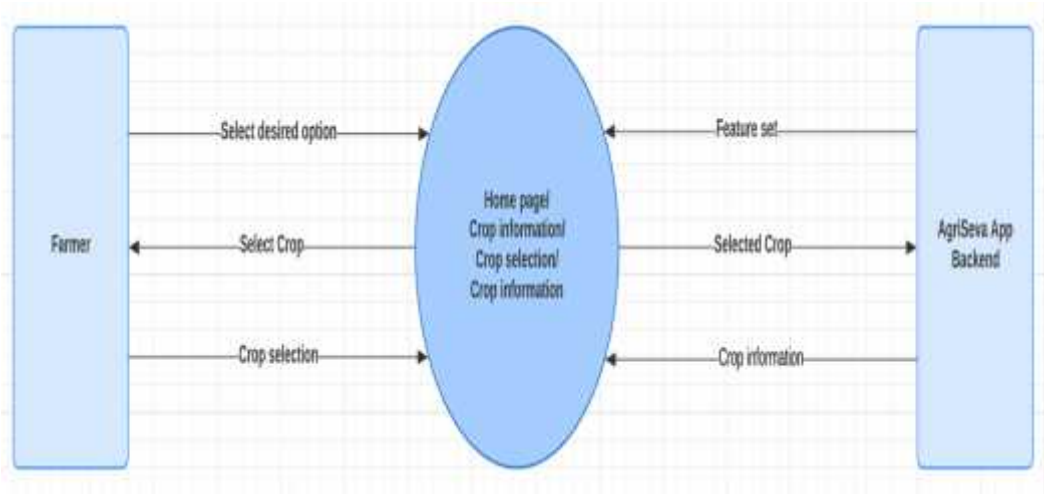


Figure 3 : DFD 1 - process of fetching crop information.

Data Flow Diagram/process flow 2 : Above diagram gives us an idea about how a user is able to file a query request and generate a reference id.

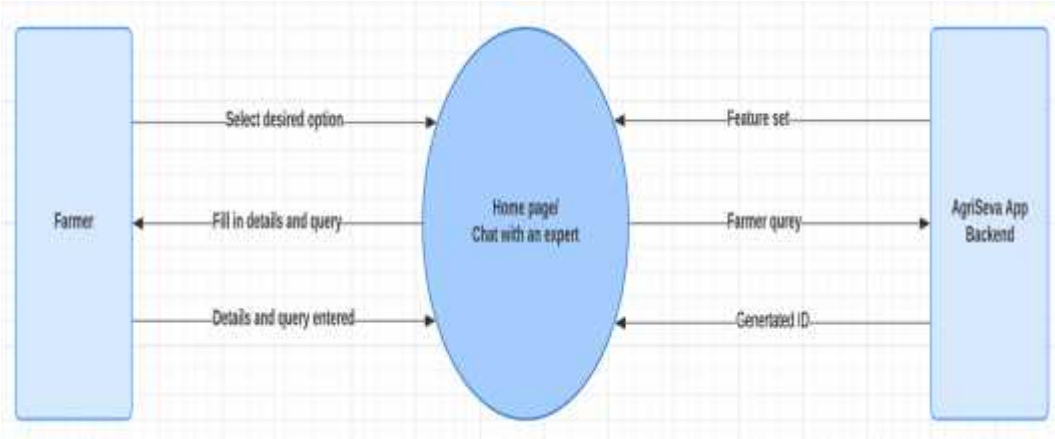


Figure 4 : DFD 2 - process involved in filing a query - chat with an expert option.

5.2 E-R DIAGRAM

Following E-R diagram helps us understand the databases associated with Agri Seva App and the flow of data within the backend system.

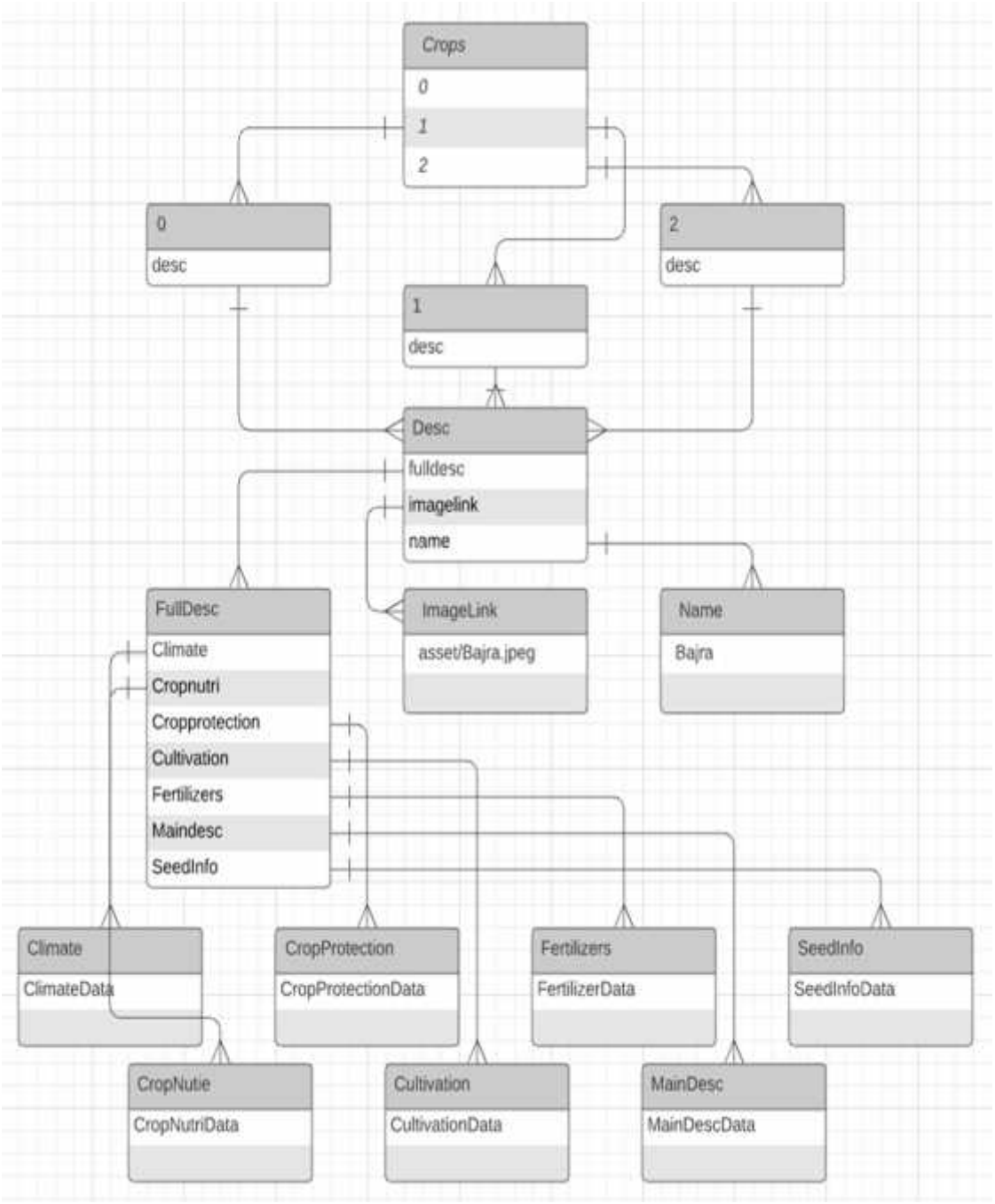


Figure 5 : E-R diagram 1 - Database structure.

5.3 UML DIAGRAMS

Use case diagram: Following diagram helps us understand how a user can obtain agri related information using Agri Seva App.

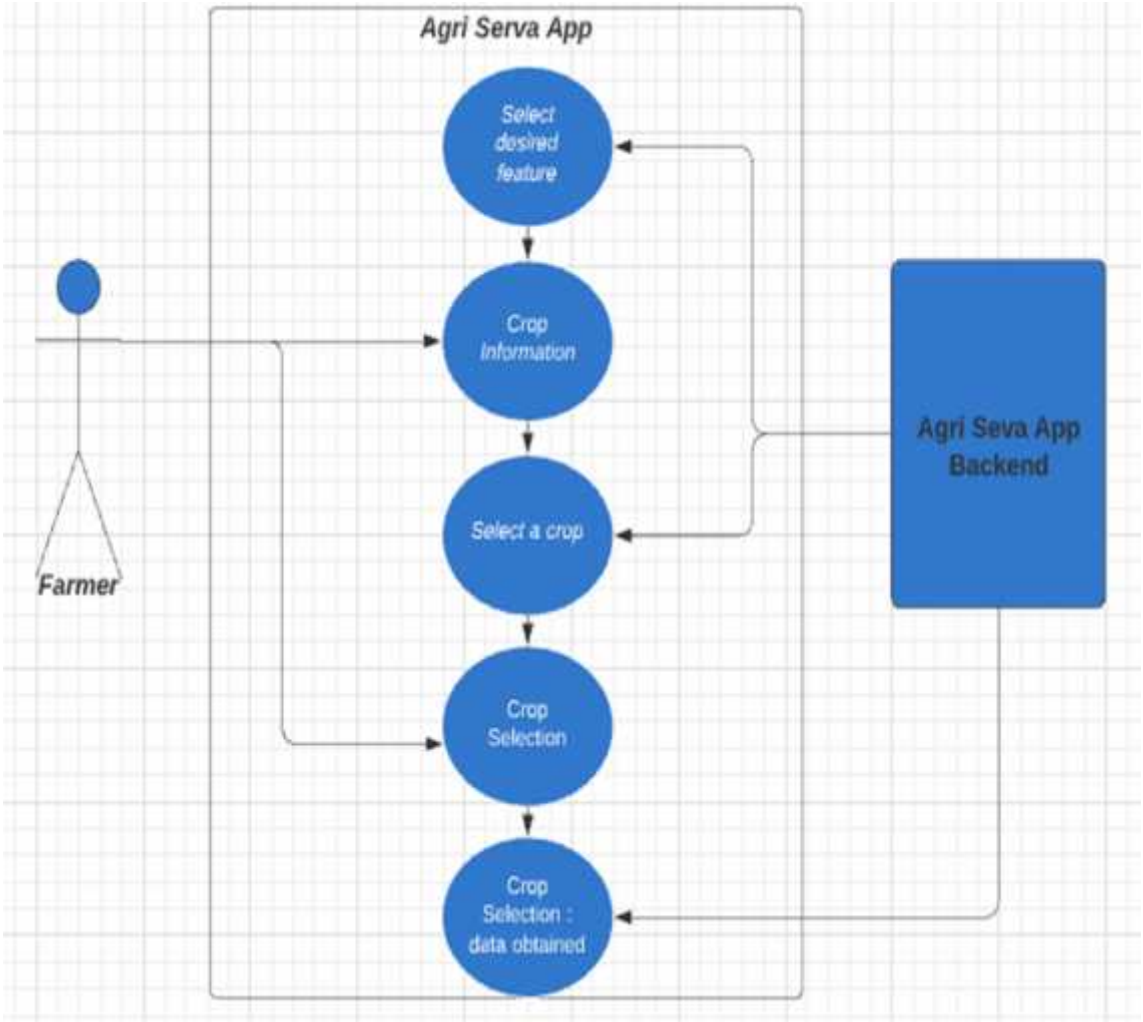


Figure 6 : UML diagram 1 - use case of fetching crop information.

Project overview flowchart : Below flowchart includes all modules and features of Agri Seva App.

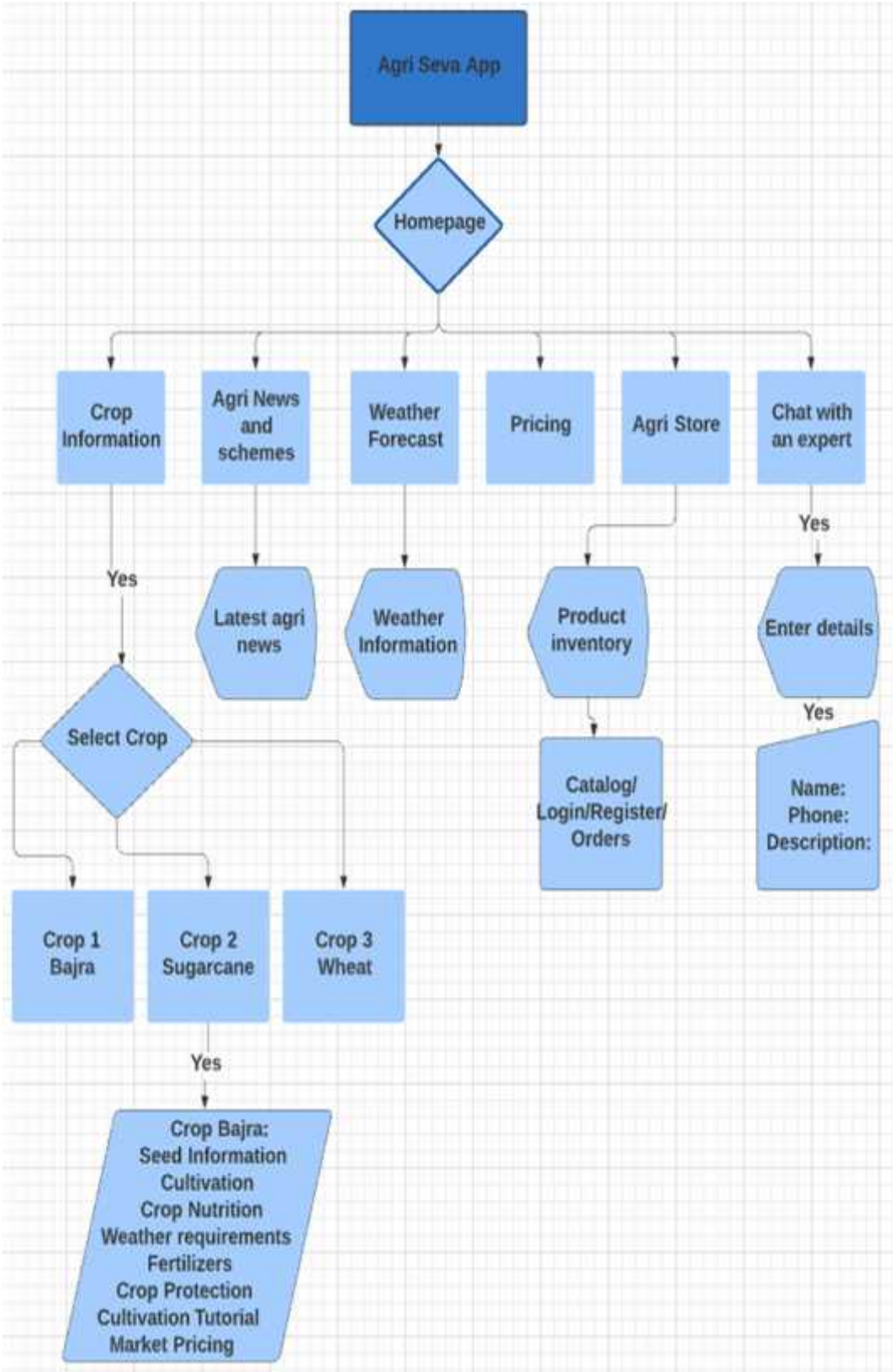


Figure 7 : UML diagram 2 - flowchart covering all modules and basic architecture.

6. PROJECT CODE

6.1 CODE TEMPLATES

Front End Code: Index.ejs - Homepage - HTML and CSS

```
<!DOCTYPE html>

<html lang="en">

<head>
   <meta charset="utf-8">
   <meta http-equiv="X-UA-Compatible" content="IE=edge">
   <meta name="viewport" content="width=device-width, initial-scale=1">
   <!-- The above 3 meta tags must come first in the head; any other head
content must come after these tags -->

   <title><%=title%></title>

   <!-- Bootstrap Core CSS -->
   <link href="/css/bootstrap.min.css" rel="stylesheet">

   <!-- Custom CSS: You can use this stylesheet to override any Bootstrap
styles and/or apply your own styles -->
   <link href="/css/custom.css" rel="stylesheet">


</head>

<body class="well-lg">

   <!-- Navigation -->
   <div class="navbar">
      <a href="#">Home</a>
      <a href="/crops">Crop information</a>
      <a href="/news">News</a>
      <a href="/weather">Weather</a>
      <a href="#">Pricing</a>
      <a href="https://shop.iffcobazar.com/en/plant-growth-
promoters.html">Agri Store</a>
      <a href="/contact">Expert Help</a>
      <a href="/about">About developers</a>
   </div>
 <div class="footer-blurb jumbotron">
      <div class="container">
```

```
        <h1 class="text-uppercase text-center bold"> Agri Seva Kiosk</h1>
        <h4 class="text-center text-lowercase">Agriculture one stop kiosk for
all Farmer's needs.</h4>
      </div>
   </div>

   <!-- Content -->
   <div class="container">

      <!-- Heading -->
      <div class="row">
         <div class="col-lg-12">
            <h1 class="page-header text-center"><strong>    
     Please Select an option
                                 
  </strong></h1>

         </div>
      </div>
      <!-- /.row -->

      <!-- Feature Row -->
      <div class="row">
         <article class="col-md-4 article-intro">
            <a href="/crops">
               <img class="img-responsive img-rounded"
                  src="assets/cultivos-cultures-crops-coltura-marruecos-maroc-
morocco-marocco-3-700x300.jpg"
                  alt="">
            </a>
            <h3><a href="/crops">Crop Information</a></h3>
            <p>Find information about various crops found in your area, popular
crops, cash crops, fertilizers and
               Precautions</p>
         </article>
         <article class="col-md-4 article-intro">
            <a href="/news">
               <img class="img-responsive img-rounded"
src="assets/15192783_1138458692876353_7817023225401072078_n-
700x300.jpg" alt="">
            </a>
```

```
            <h3>
                <a href="/news">Agri-News and Schemes</a>
            </h3>
            <p>Find news about various schemes and yojanas introduced by the
Government for the welfare of Farmers.
            </p>
          </article>

          <article class="col-md-4 article-intro">
            <a href="/weather">
                <img src="assets/download.jpg" alt="" width="343" height="147"
class="img-responsive img-rounded">
            </a>
            <h3><a href="/weather">Weather Forecast</a></h3>
            <p>Find out about various weather related information that could
help you plan out your crop cultivation
                planning.</p>
          </article>
        </div>
        <!-- /.row -->

    </div>
    <!-- /.container -->

    <footer>
        <div class="feature">
          <div class="container">
            <div class="row">
                <article class="col-md-4 article-intro"> <a href="#"> <img
class="img-responsive img-rounded"
                        src="assets/pthomeandgarden_6-Vegetables-To-Plant-Late-
Summer-700x300 (1).jpg" alt="">
                    </a>
                    <h3><a href="#">Pricing</a></h3>
                    <p>Know about various local and standardized pricing of
various crops around your locality.</p>
                </article>
                <article class="col-md-4 article-intro"> <a
href="https://shop.iffcobazar.com/en/plant-growth-promoters.html"> <img
class="img-responsive img-rounded"
```

```
src="assets/essential-oil-brands-and-companies-700x300.jpg" alt=""> </a>
                    <h3><a href="https://shop.iffcobazar.com/en/plant-growth-
promoters.html">Agri- Store</a></h3>
                    <p>Quick store for purchasing pesticides and various agriculture
related goods. The store is
                        subsidized by Government of India.</p>
                </article>
                <article class="col-md-4 article-intro"> <a href="/contact"> <img
class="img-responsive img-rounded"
                        src="assets/pthomeandgarden-Importance-of-Using-
Organic-Products-Gardening-700x300.jpg"
                        alt=""> </a>
                    <h3> <a href="contact">Chat with an expert</a></h3>
                    <p>Talk to an expert and know about the various problems you
will be facing.</p>
                </article>
            </div>
            <!-- /.row -->
        </div>
    </div>

    <div class="small-print well-sm"> </div>
  </footer>

  <!-- jQuery -->
  <script src="/js/jquery-1.11.3.min.js"></script>

  <!-- Bootstrap Core JavaScript -->
  <script src="/js/bootstrap.min.js"></script>

  <!-- IE10 viewport bug workaround -->
  <script src="/js/ie10-viewport-bug-workaround.js"></script>

  <!-- Placeholder Images -->
  <script src="/js/holder.min.js"></script>

</body>

</html>
```

Major Backend Code: Node JS

```
const express = require('express');
```

```
const path = require('path');
const firebase = require('firebase');
const admin = require('firebase-admin');
const bodyParser = require('body-parser');

const app = express();
const PORT = process.env.PORT || 3000;

app.use(express.static(path.join(__dirname, '/public')));
app.use(bodyParser.urlencoded());
app.use(bodyParser.json())
app.set('views', './src/views');
app.set('view engine', 'ejs');

const crops = [];
const news = [];

const firebaseConfig = {
   apiKey: 'AIzaSyAh_1zYTaP3jeDjJzC5LM_Bc2ZeLWlGYR8',
   authDomain: 'agriseva-fb852.firebaseapp.com',
   databaseURL: 'https://agriseva-fb852.firebaseio.com',
   projectId: 'agriseva-fb852',
   storageBucket: 'agriseva-fb852.appspot.com',
   messagingSenderId: '165433753207',
   appId: '1:165433753207:web:2e386cbb7fe411d5e51072',
   measurementId: 'G-JHG3CK6CEE',
};
const serviceAccount = require('./src/config/agriseva-fb852-firebase-adminsdk-
psr8o-2ebb9d9dc7');

admin.initializeApp({
   credential: admin.credential.cert(serviceAccount),
   databaseURL: 'https://agriseva-fb852.firebaseio.com',
});
firebase.initializeApp(firebaseConfig);

const db = firebase.database();
const ref = db.ref('/crops');
ref.once('value', (snapshot) => {
   snapshot.forEach((childSnapshot) => {
      crops.push(childSnapshot.val());
   });
});
```

```
const newsRef = db.ref('/news');
newsRef.once('value', (snapshot) => {
```

```
    snapshot.forEach((childSnapshot) => {
        news.push(childSnapshot.val());
    });
});

const cropRouter = require('./src/routes/cropRouter')(crops);
const contactRouter = require('./src/routes/contactRouter')(db);
const adminRouter = require('./src/routes/adminRouter')(serviceAccount);
const newsRouter = require('./src/routes/newsRouter')(news);
const weatherRouter = require('./src/routes/weatherRouter')();
const aboutRouter = require('./src/routes/aboutRouter')();

app.post('/submit-form', (req, res) => {
    const { name } = req.body;
    const { phone } = req.body;
    const { desc } = req.body;

    const send = {
        name,
        phone,
        desc,
    };

    const formRef = db.ref('/contact');
    formRef.push(send);
    //debug(`${send}updated`);
    res.redirect('/contact');
    res.end();
    // debug(send);
});
app.use('/crops', cropRouter);
app.use('/contact', contactRouter);
app.use('/admin', adminRouter);
app.use('/news', newsRouter);
app.use('/weather', weatherRouter);
app.use('/about', aboutRouter);
app.get('/', (req, res) => {
    res.render(
        'indexn',
        { title: 'AgriSeva' },
    );
});

app.listen(PORT, () => {
    console.log("App running on port: "+PORT);
});
```

API Calls: forecast7

```
const firebaseConfig = {
   apiKey: 'AIzaSyAh_1zYTaP3jeDjJzC5LM_Bc2ZeLWlGYR8',
   authDomain: 'agriseva-fb852.firebaseapp.com',
   databaseURL: 'https://agriseva-fb852.firebaseio.com',
   projectId: 'agriseva-fb852',
   storageBucket: 'agriseva-fb852.appspot.com',
   messagingSenderId: '165433753207',
   appId: '1:165433753207:web:2e386cbb7fe411d5e51072',
   measurementId: 'G-JHG3CK6CEE',
};

const serviceAccount = require('./src/config/agriseva-fb852-firebase-adminsdk-
psr8o-2ebb9d9dc7');

admin.initializeApp({
   credential: admin.credential.cert(serviceAccount),
   databaseURL: 'https://agriseva-fb852.firebaseio.com',
});

firebase.initializeApp(firebaseConfig);

const db = firebase.database();
const ref = db.ref('/crops');
ref.once('value', (snapshot) => {
   snapshot.forEach((childSnapshot) => {
      crops.push(childSnapshot.val());
   });
});

const newsRef = db.ref('/news');
newsRef.once('value', (snapshot) => {
   snapshot.forEach((childSnapshot) => {
      news.push(childSnapshot.val());
   });
});
```

Crop Page : cropPage.ejs - HTML and CSS

```
<!DOCTYPE html>
<!-- Template by Quackit.com -->
<html lang="en">

<head>
    <meta charset="utf-8">
    <meta http-equiv="X-UA-Compatible" content="IE=edge">
    <meta name="viewport" content="width=device-width, initial-scale=1">
    <!-- The above 3 meta tags must come first in the head; any other head
content must come after these tags -->

    <title>Crops</title>

    <!-- Bootstrap Core CSS -->
    <link href="/css/bootstrap.min.css" rel="stylesheet">

    <!-- Custom CSS: You can use this stylesheet to override any Bootstrap
styles and/or apply your own styles -->
    <link href="/css/custom.css" rel="stylesheet">

    <!-- HTML5 Shim and Respond.js IE8 support of HTML5 elements and
media queries -->
    <!-- WARNING: Respond.js doesn't work if you view the page via file:// -->
    <!--[if lt IE 9]>
        <script
src="https://oss.maxcdn.com/libs/html5shiv/3.7.0/html5shiv.js"></script>
        <script
src="https://oss.maxcdn.com/libs/respond.js/1.4.2/respond.min.js"></script>
    <![endif]-->

</head>

<body class="well-lg">

    <!-- Navigation -->
    <div class="navbar">
        <a href="/../">Home</a>
        <a href="#">Crop information</a>
        <a href="/../news">News</a>
        <a href="/../weather">Weather</a>
```

```
      <a href="#">Pricing</a>
      <a href="https://shop.iffcobazar.com/en/plant-growth-
promoters.html">Agri Store</a>
      <a href="/../contact">Expert Help</a>
      <a href="/about">About developers</a>
   </div>

   <div class="footer-blurb jumbotron">
      <div class="container">
         <h1 class="text-uppercase text-center bold"> Agri Seva Kiosk</h1>
         <h4 class="text-center text-lowercase">Agriculture one stop kiosk for
all Farmer's needs.</h4>
      </div>
   </div>

   <!-- Content -->
   <div class="container">

      <!-- Heading -->
      <div class="row">
         <div class="col-lg-12 glyphicon">
            <h1 class="page-header text-center"><strong class="font">
&amp;nbsp;Select Crop &amp;nbsp; &amp;nbsp; &amp;nbsp; &amp;nbsp;
                  &amp;nbsp; &amp;nbsp; &amp;nbsp; &amp;nbsp; </strong></h1>
         </div>
         <!-- <div class="col-lg-12 glyphicon">
            <div class="input-group mb-3">
               <div class="input-group-prepend">
                  <span class="input-group-text">Search</span>
               </div>
               <input type="text" class="form-control">
            </div>
         </div> -->
      </div>
      <!-- /.row -->
```

```
      <!-- Feature Row -->
       <div class="row">
          <article class="col-md-4 article-intro">
             <a href="#">
             </a> </article>
       </div>
       <div class="row">
          <%for(let i=0;i<crops.length;i++) {%>
             <div class="card col-sm-6 panel-body" style="width: 28rem;">
                <a href="/crops/<%=crops[i].name%>"><img
src="<%=crops[i].imageLink%>" alt="Card image cap" style="width: 28rem;"
                   class="card-img-top img-responsive"></a>
                <div class="card-body">
                   <h2 class="card-title bold"><%=crops[i].name%></h2>
                   <p class="card-text">
                      <%=crops[i].desc%>
                   </p>
                   <a href="/crops/<%=crops[i].name%>" class="btn btn-
primary">Get info</a>
                </div>
             </div>
             <%}%>
       </div>
    </div>
    <!-- /.container -->

    <footer class="text-center">Made by SMEC </footer>


    <!-- jQuery -->
    <script src="/js/jquery-1.11.3.min.js"></script>

    <!-- Bootstrap Core JavaScript -->
    <script src="/js/bootstrap.min.js"></script>

    <!-- IE10 viewport bug workaround -->
    <script src="/js/ie10-viewport-bug-workaround.js"></script>

    <!-- Placeholder Images -->
    <script src="/js/holder.min.js"></script>
</body>

</html>
```

Crop Info Page: cropInfoPage.ejs - HTML and CSS

```
<!DOCTYPE html>

<html lang="en">

<head>
    <meta charset="utf-8">
    <meta http-equiv="X-UA-Compatible" content="IE=edge">
    <meta name="viewport" content="width=device-width, initial-scale=1">
    <!-- The above 3 meta tags must come first in the head; any other head
content must come after these tags -->

    <title>Info</title>

    <!-- Bootstrap Core CSS -->
    <link href="/css/bootstrap.min.css" rel="stylesheet">

    <!-- Custom CSS: You can use this stylesheet to override any Bootstrap
styles and/or apply your own styles -->
    <link href="/css/custom.css" rel="stylesheet">


</head>

<body class="well-lg">

    <!-- Navigation -->
    <div class="navbar">
        <a href="/../../">Home</a>
        <a href="/../">Crop information</a>
        <a href="/../../news">News</a>
        <a href="/../../weather">Weather</a>
        <a href="#">Pricing</a>
        <a href="https://shop.iffcobazar.com/en/plant-growth-
promoters.html">Agri Store</a>
        <a href="/../../contact">Expert Help</a>
        <a href="/about">About developers</a>
    </div>
```

```
<div class="footer-blurb jumbotron">
        <div class="container">
            <h1 class="text-uppercase text-center bold"> Agri Seva Kiosk</h1>
```

```
        <h4 class="text-center text-lowercase">Agriculture one stop kiosk for
all Farmer's needs.</h4>
      </div>


   </div>
   <div class="feature"> </div>


   <div class="container">
      <h1 class="center-block text-center text-
uppercase"><strong><%=title%></strong></h1>
      <img src="/<%=cropInfo.imageLink%>" alt="" width="320"
height="320"
         class="img-rounded img-thumbnail center-block" />
      <p><%=cropInfo.fullDesc.seedInfo%></p>
      </p>
   </div>

   <div class="container">
      <h1 class="container-fluid container text-center text-
uppercase"><strong>Seed Information</strong></h1>
      <p><%=cropInfo.fullDesc.seedInfo%></p>
      <p> </p>
   </div>
   <div class="container">
      <h1 class="container text-center text-
uppercase"><strong>Cultivation</strong></h1>
      <p><%=cropInfo.fullDesc.cultivation%></p>
   </div>
   <div class="container">
      <h1 class="container text-center text-uppercase"><strong>Crop
Nutrition</strong></h1>
      <p><%=cropInfo.fullDesc.cropNutri%></p>
   </div>
```

```
 <div class="container">
      <h1 class="text-center bold text-uppercase"><strong>Climatic
Requirements</strong></h1>
```

```
      <p> <%=cropInfo.fullDesc.climate%>
      </p>
   </div>
   <div class="container">
      <h1 class="text-center bold text-
uppercase"><strong>Fertilizers</strong></h1>
      <p> <%=cropInfo.fullDesc.fertilizers%> </p>

   </div>
   <div class="container">
      <h1 class="text-center bold text-uppercase"><strong>Crop
Protection</strong></h1>
      <p> <%=cropInfo.fullDesc.crop Protection%>
      </p>

   </div>
   <div class="container">
      <h1 class="container text-center text-uppercase"><strong>Cultivation
tutorial</strong></h1>
      <video controls class="center-block">
         <source src="/assets/paddy cultivaton_Trim.mp4">
      </video>
   </div>

   <table width="200" border="1">

      </tbody>
   </table>
   <div class="container center-block">
      <div class="center-block">
         <h1 class="text-center text-uppercase"><strong>Market
Pricing</strong></h1>
         <table width="200" border="1">
            <tbody>
               <tr> </tr>
               <tr> </tr>
               <tr> </tr>
               <tr> </tr>
            </tbody>
         </table>
```

```
 <p class="container container-fluid text-justify text-uppercase">
            <tbody>
               <table width="1200" border="1">
```

```
                <tbody>
                  <tr>
                    <td>Variety </td>
                    <td>MTU-1001 </td>
                  </tr>
                  <tr>
                    <td>Minimum price </td>
                    <td> 1047</td>
                  </tr>
                  <tr>
                    <td>Average price </td>
                    <td>1470 </td>
                  </tr>
                  <tr>
                    <td>Maximum price </td>
                    <td>1470 </td>
                  </tr>
                </tbody>
              </table>
          </p>
        </div>
    </div>
    <div class="feature jumbotron"> </div>
    <!-- jQuery -->
    <script src="/js/jquery-1.11.3.min.js"></script>

    <!-- Bootstrap Core JavaScript -->
    <script src="/js/bootstrap.min.js"></script>

    <!-- IE10 viewport bug workaround -->
    <script src="js/ie10-viewport-bug-workaround.js"></script>
    <!-- Placeholder Images -->
    <script src="js/holder.min.js"></script>
</body>

</html>
```

Chat With An Expert Page : contactPage.ejs - HTML and CSS

```
<!DOCTYPE html>

<html lang="en">

<head>
    <meta charset="utf-8">
    <meta http-equiv="X-UA-Compatible" content="IE=edge">
    <meta name="viewport" content="width=device-width, initial-scale=1">
    <!-- The above 3 meta tags must come first in the head; any other head
content must come after these tags -->

    <title>Contact</title>

    <!-- Bootstrap Core CSS -->
    <link href="css/bootstrap.min.css" rel="stylesheet">

    <!-- Custom CSS: You can use this stylesheet to override any Bootstrap styles
and/or apply your own styles -->
    <link href="css/custom.css" rel="stylesheet">

    <!-- HTML5 Shim and Respond.js IE8 support of HTML5 elements and
media queries -->
    <!-- WARNING: Respond.js doesn't work if you view the page via file:// -->
    <!--[if lt IE 9]>
        <script
src="https://oss.maxcdn.com/libs/html5shiv/3.7.0/html5shiv.js"></script>
        <script
src="https://oss.maxcdn.com/libs/respond.js/1.4.2/respond.min.js"></script>
    <![endif]-->

</head>

<body class="well-lg">
```

```
    <!-- Navigation -->
    <div class="navbar">
```

```
      <div class="navbar">
         <a href="/../">Home</a>
         <a href="/../crops">Crop information</a>
         <a href="/../news">News</a>
         <a href="/../weater">Weather</a>
         <a href="#">Pricing</a>
         <a href="https://shop.iffcobazar.com/en/plant-growth-
promoters.html">Agri Store</a>
         <a href="#">Expert Help</a>
         <a href="/about">About developers</a>
      </div>
   </div>

   <div class="footer-blurb jumbotron">
      <div class="container">
         <h1 class="text-uppercase text-center bold"> Agri Seva Kiosk</h1>
         <h4 class="text-center text-lowercase">Agriculture one stop kiosk for
all Farmer's needs.</h4>
      </div>
   </div>

   <!-- Content -->
   <div class="container">

      <!-- Heading -->
      <div class="row">
         <div class="col-lg-12 container">
            <h1 class="page-header text-center"><strong class="center-block">
Enter Details      
                              </strong></h1>
         </div>
      </div>
      <!-- /.row →
<!-- Feature Row -->
      <div class="row"> </div>
      <!-- /.row -->

   </div>
   <!-- /.container -->
```

```
<footer>
      <form method="POST" action="/submit-form">
         <div class="form-group container">
```

```
            <label for="exampleInputEmail2">Name</label>
            <input type="text" class="form-control" name="name"
placeholder="Name">
        </div>
        <div class="form-group container">
            <label for="exampleInputPassword1">Phone</label>
            <input type="number" name="phone" class="form-control"
placeholder="Phone number">
        </div>
        <div class="form-group container">
            <label for="exampleInputEmail2">Description</label>
            <input type="text" name="desc" class="form-control form-group
form-group-lg" placeholder="Description">

        </div>
        <!-- <div class="checkbox container">
            <label>
                <input type="checkbox">
                Urgent matter </label>
        </div> -->
        <h3 class="container center-block text-center text-uppercase"><strong>
                <button type="submit" class="btn btn-default"
id="submit">Submit</button>
                <span class="badge badge-
success">Success</span></strong></h3>
      </form>
      <div class="col-lg-12">
          <h1 class="page-header text-center"><strong> ID Generated : 001
           
                      </strong></h1>
      </div>
   </footer>
 <!-- jQuery -->
   <script src="js/jquery-1.11.3.min.js"></script>

   <!-- Bootstrap Core JavaScript -->
   <script src="js/bootstrap.min.js"></script>
```

```
 <!-- IE10 viewport bug workaround -->
   <script src="js/ie10-viewport-bug-workaround.js"></script>
```

```
   <!-- Placeholder Images -->
   <script src="js/holder.min.js"></script>

   <!-- The core Firebase JS SDK is always required and must be listed first -->
   <!-- <script src="https://www.gstatic.com/firebasejs/7.5.2/firebase-
app.js"></script>

  <script src="https://www.gstatic.com/firebasejs/7.5.2/firebase-
analytics.js"></script>

   <script>
      // Your web app's Firebase configuration
      var firebaseConfig = {
         apiKey: "AIzaSyAh_1zYTaP3jeDjJzC5LM_Bc2ZeLWlGYR8",
         authDomain: "agriseva-fb852.firebaseapp.com",
         databaseURL: "https://agriseva-fb852.firebaseio.com",
         projectId: "agriseva-fb852",
         storageBucket: "agriseva-fb852.appspot.com",
         messagingSenderId: "165433753207",
         appId: "1:165433753207:web:2e386cbb7fe411d5e51072",
         measurementId: "G-JHG3CK6CEE"
      };
      // Initialize Firebase
      firebase.initializeApp(firebaseConfig);
      firebase.analytics();
   </script>
   <script src="js/contactScript.js"></script> -->
   <div class="col-lg-12 feature">
      <h1 class="dark"><strong>          
             
                           
             

                 An Expert will get back to you shortly  
         
                  </strong></h1>
   </div>
</body>

</html>
```

Agri News And Schemes Page: agrinews .ejs - HTML and CSS

```
<!DOCTYPE html>
```

```
<html lang="en">

<head>
   <meta charset="utf-8">
   <meta http-equiv="X-UA-Compatible" content="IE=edge">
   <meta name="viewport" content="width=device-width, initial-scale=1">
   <!-- The above 3 meta tags must come first in the head; any other head
content must come after these tags -->

   <title>News</title>

   <!-- Bootstrap Core CSS -->
   <link href="/css/bootstrap.min.css" rel="stylesheet">

   <!-- Custom CSS: You can use this stylesheet to override any Bootstrap
styles and/or apply your own styles -->
   <link href="/css/custom.css" rel="stylesheet">


   <meta name="viewport" content="width=device-width, initial-scale=1">
   <link rel="stylesheet" href="https://cdnjs.cloudflare.com/ajax/libs/font-
awesome/4.7.0/css/font-awesome.min.css">
</head>

<body class="well-lg">
   <!-- Navigation -->
   <div class="navbar">
      <a href="/../">Home</a>
      <a href="/../crops">Crop information</a>
      <a href="#">News</a>
      <a href="/../weather">Weather</a>
      <a href="#">Pricing</a>
      <a href="https://shop.iffcobazar.com/en/plant-growth-
promoters.html">Agri
   </div>
```

```
<div class="footer-blurb jumbotron">
      <div class="container">
         <h1 class="text-uppercase text-center bold"> Agri Seva Kiosk</h1>
```

```
        <h4 class="text-center text-lowercase">Agriculture one stop kiosk for
all Farmer's needs.</h4>
      </div>
   </div>

   <!-- Content -->
   <h1 class="center-block text-justify text-uppercase"><strong>  
             
                          
           
                          
           
                          
    Agri-News</strong>
   </h1>
   <picture alt="description of image">
      <!-- Default low resolution image -->
      <source src="small.jpg">
      <!-- Medium resolution image -->
      <source src="medium.jpg" media="(min-width: 768px)">
      <!--High resolution Image -->
      <source src="large.jpg" media="(min-width: 1200px)">
      <!-- Fallback image incase the custom picture element is not supported -->
   </picture>
   <div class="container">
      <!-- Heading -->
      <div class="row"></div>
      <!-- /.row -->
      <!-- Feature Row -->
      <!-- /.row -->
   </div>
   <!-- /.container →
   <div class="feature">
      <div class="row">
         <%for(let i=0;i<news.length;i++) {%>
         <div class="col-sm-6 col-md-4 jumbotron">
            <div class="thumbnail"> <img src="assets/download (1).jpg"
alt="...">
```

```
               <div class="caption" max-lines="3">
                  <h3><%=news[i].title%></h3>
```

```
                <p class="news-page"><%=news[i].desc%></p>
                <!-- <p><a href="#" class="btn btn-primary" role="button">See
more</a> </p> -->
              </div>
            </div>
          </div>
          <%}%>
          <!-- <div class="col-sm-6 col-md-4 jumbotron">
            <div class="thumbnail"> <img src="assets/download (1).jpg"
alt="...">
              <div class="caption">
                <h3>News2</h3>
                <p>Content</p>
                <p><a href="#" class="btn btn-primary" role="button">See
more</a> </p>
              </div>
            </div>
          </div>
          <div class="col-sm-6 col-md-4 jumbotron">
            <div class="thumbnail"> <img src="assets/download (1).jpg"
alt="...">
              <div class="caption">
                <h3>News3</h3>
                <p>Content</p>
                <p><a href="#" class="btn btn-primary" role="button">See
more</a> </p>
              </div>
            </div>
          </div>
          <div class="col-sm-6 col-md-4 jumbotron">
            <div class="thumbnail"> <img src="assets/download (1).jpg"
alt="...">
              <div class="caption">
                <h3>News4</h3>
                <p>content</p>
                <p><a href="#" class="btn btn-primary" role="button">See
more</a> </p>
              </div>
            </div>
          </div>
```

```
          <div class="col-sm-6 col-md-4 jumbotron">
            <div class="thumbnail"> <img src="assets/download (1).jpg"
alt="...">
```

```
                <div class="caption">
                    <h3>News5</h3>
                    <p>Content</p>
                    <p><a href="#" class="btn btn-primary" role="button">See
more</a> </p>
                </div>
            </div>
        </div>
        <div class="col-sm-6 col-md-4 jumbotron">
            <div class="thumbnail"> <img src="assets/download (1).jpg"
alt="...">
                <div class="caption">
                    <h3>News6</h3>
                    <p>Content</p>
                    <p><a href="#" class="btn btn-primary" role="button">See
more</a> </p>
                </div>
            </div>
        </div> -->
      </div>
  </div>

  <footer>
  </footer>


  <!-- jQuery -->
  <script src="/js/jquery-1.11.3.min.js"></script>

  <!-- Bootstrap Core JavaScript -->
  <script src="/js/bootstrap.min.js"></script>

  <!-- IE10 viewport bug workaround -->
  <script src="/js/ie10-viewport-bug-workaround.js"></script>

  <!-- Placeholder Images -->
  <script src="/js/holder.min.js"></script>

</body>
</html>
```

Weather Forecast Page : weatherPage.ejs - HTML and CSS

```
<!DOCTYPE html>
```

```
<html lang="en">

<head>
    <meta charset="utf-8">
    <meta http-equiv="X-UA-Compatible" content="IE=edge">
    <meta name="viewport" content="width=device-width, initial-scale=1">
    <!-- The above 3 meta tags must come first in the head; any other head
content must come after these tags -->

    <title>Weather</title>

    <!-- Bootstrap Core CSS -->
    <link href="css/bootstrap.min.css" rel="stylesheet">

    <!-- Custom CSS: You can use this stylesheet to override any Bootstrap
styles and/or apply your own styles -->
    <link href="css/custom.css" rel="stylesheet">

    <!-- HTML5 Shim and Respond.js IE8 support of HTML5 elements and
media queries -->
    <!-- WARNING: Respond.js doesn't work if you view the page via file:// -->
    <!--[if lt IE 9]>
        <script
src="https://oss.maxcdn.com/libs/html5shiv/3.7.0/html5shiv.js"></script>
        <script
src="https://oss.maxcdn.com/libs/respond.js/1.4.2/respond.min.js"></script>
    <![endif]-->

</head>

<body class="well-lg">
```

```
    <!-- Navigation -->
    <div class="navbar">
        <a href="/../">Home</a>
```

```
      <a href="/../crops">Crop information</a>
      <a href="/../news">News</a>
      <a href="#">Weather</a>
      <a href="#">Pricing</a>
      <a href="https://shop.iffcobazar.com/en/plant-growth-
promoters.html">Agri Store</a>
      <a href="/../contact">Expert Help</a>
      <a href="/about">About developers</a>
   </div>


   <div class="footer-blurb jumbotron">
      <div class="container">
         <h1 class="text-uppercase text-center bold"> Agri Seva Kiosk</h1>
         <h4 class="text-center text-lowercase">Agriculture one stop kiosk for
all Farmer's needs.</h4>
      </div>
   </div>

   <!-- Content -->
   <div class="container">
      <a class="weatherwidget-io"
href="https://forecast7.com/en/17d3978d49/hyderabad/" data-
label_1="HYDERABAD"
         data-label_2="WEATHER" data-font="Open Sans" data-
icons="Climacons Animated" data-theme="original">HYDERABAD
         WEATHER</a>
      <script>
         !function (d, s, id) { var js, fjs = d.getElementsByTagName(s)[0]; if
(!d.getElementById(id)) { js = d.createElement(s); js.id = id; js.src =
'https://weatherwidget.io/js/widget.min.js'; fjs.parentNode.insertBefore(js, fjs);
} }(document, 'script', 'weatherwidget-io-js');
      </script>
      <div class="row"> </div>
      <!-- /.row →
   </div>
   <!-- /.container -->
```

```
   <!-- jQuery -->
   <script src="js/jquery-1.11.3.min.js"></script>

   <!-- Bootstrap Core JavaScript -->
```

```
    <script src="js/bootstrap.min.js"></script>

    <!-- IE10 viewport bug workaround -->
    <script src="js/ie10-viewport-bug-workaround.js"></script>

    <!-- Placeholder Images -->
  <script src="js/holder.min.js"></script>

    <div class="col-lg-12"> </div>

</body>

</html>
```

About Developers Page : aboutDevs.ejs - HTML and CSS

```
<!DOCTYPE html>

<html lang="en">

<head>
    <meta charset="utf-8">
    <meta http-equiv="X-UA-Compatible" content="IE=edge">
    <meta name="viewport" content="width=device-width, initial-scale=1">
    <!-- The above 3 meta tags must come first in the head; any other head
content must come after these tags -->

    <title>About</title>

    <!-- Bootstrap Core CSS -->
    <link href="css/bootstrap.min.css" rel="stylesheet">



    <link href="css/custom.css" rel="stylesheet">
    <style type="text/css">
```

```
        body {
            background-color: #FFFFFF;
        }
    </style>
```

```
</head>

<body class="well-lg">

   <!-- Navigation -->
   <div class="navbar">
      <a href="/../">Home</a>
      <a href="/../crops">Crop information</a>
      <a href="/../news">News</a>
      <a href="/../weather">Weather</a>
      <a href="#">Pricing</a>
      <a href="https://shop.iffcobazar.com/en/plant-growth-
promoters.html">Agri Store</a>
      <a href="/../contact">Expert Help</a>
      <a href="#">About developers</a>
   </div>

   <div class="footer-blurb jumbotron">
      <div class="container">
         <h1 class="text-uppercase text-center bold"> Agri Seva Kiosk</h1>
         <h4 class="text-center text-lowercase">Agriculture one stop kiosk for
all Farmer's needs.</h4>
      </div>
   </div>
   <div class="container center-block container-fluid container_custom">

      <div class="container">
         <img src="assets/Logo final.png" width="1100" height="115"
class="img-thumbnail">
      </div>
      <div class="container">
         <h2 class="center-block text-center"> About the Developers </h2>
      </div>
      <img src="assets/kiosk.jpg" width="1200" height="500" class="img-
thumbnail container center-block">
```

```
 <div class="container">
        <div class="container center-block">
           <h3 class="text-uppercase"><strong class="center-block">  
    R.KOUSHIK BHARGAVA  
```

```
                          Suraj Prasad Sharma    
SHUBHAM AGARWAL</strong></h3>
            <h3> </h3>
          </div>
        </div>
    </div>
    <footer>
        <div class="feature jumbotron">
            <div class="container">
                <!-- /.row -->
            </div>
        </div>
    </footer>


    <!-- jQuery -->
    <script src="js/jquery-1.11.3.min.js"></script>

    <!-- Bootstrap Core JavaScript -->
    <script src="js/bootstrap.min.js"></script>

    <!-- IE10 viewport bug workaround -->
    <script src="js/ie10-viewport-bug-workaround.js"></script>

    <!-- Placeholder Images -->
    <script src="js/holder.min.js"></script>

</body>

</html>
```

About aboutRouter.js - Javascript

```
const express = require('express');
```

```
const debug = require('debug')('app:cropPageRouter');

const aboutRouter = express.Router();

function router() {
  aboutRouter.route('/')
    .get((req, res) => {
      res.render('aboutDevs');
      debug('About page loads and working');
    });
  return aboutRouter;
}

module.exports = router;
```

ContactRouter.js - Javascript

```
const express = require('express');
const debug = require('debug')('app:cropPageRouter');

constant contactRouter = express.Router();
express().use(express.urlencoded());

function router() {
  contactRouter.route('/')
    .get((req, res) => {
      res.render('contact Page');
      debug('Contact page loads and working');
    });
  return contactRouter;
}

module.exports = router;
```

CropRouter.js - Javascript

```
const express = require('express');
```

```
const debug = require('debug')('app:cropPageRouter');

const cropRouter = express.Router();


function router(crops) {
  cropRouter.route('/')
    .get((req, res) => {
      res.render(
        'cropPagen',
        {
          title: 'Crops',
          crops,
        },
      );
      debug('Connected to cropPage');
    });

  cropRouter.route('/:id')
    .get((req, res) => {
      const { id } = req.params;
      const index = crops.findIndex((x) => x.name === id);
      res.render(
        'cropInfoPagen',
        {
          title: id,
          cropInfo: crops[index],
        },
      );
    });
  return cropRouter;
}

module.exports = router;
```

NewsRouter.js - Javascript

```
const express = require('express');
const debug = require('debug')('app:newsRoutes');

const adminRouter = express.Router();

function router(news) {
  adminRouter.route('/')
    .get((req, res) => {
      res.render(
        'agrinews',
        { news },
      );
      debug('agrinews Page loaded');
    });

  return adminRouter;
}

module.exports = router;
```

WeatherRouter.js - Javascript

```
const express = require('express');
const debug = require('debug')('app:newsRoutes');

const weatherRouter = express.Router();

function router() {
  weatherRouter.route('/')
    .get((req, res) => {
      res.render('weatherPage');
      debug('Connected to weather page');
    });

  return weatherRouter;
}

module.exports = router;
```

6.2 OUTLINE FOR VARIOUS FILES

Following are the major files in development of Agri Seva App,

1. **Index.js -** includes main server code.
2. **Routes folder -** files in route folder are used to handle navigation within different pages.
3. **Views folder -** files in views folder holds the Frontend.
4. **Index.ejs -** includes frontend code for complete homepage.

Following outline image can be used to understand the order of above files and folders

Figure 8 : Code files.

6.3 CLASS/ FUNCTION/ METHOD WITH FUNCTIONALITY

Functionality and usage of all major methods are discussed below,

Function/Method no 1 : admin.initializeApp() - is used to initialize the database.

Function/Method no 2 : const db = firebase.database(); and const newsRef = db.ref('/news'); - are used to access data from the database.

Function/Method no 3 : const cropRouter = require('./src/routes/cropRouter')(crops); - is used for page redirection operation.

Function/Method no 4 : app.post('/submit-form', (req, res) and const send = name,phone,desc }; - these are important in sending data entered by user using chat with expert form to Agri Seva App backend database.

6.4 METHODS INPUT AND OUTPUT PARAMETERS

Following are the major methods used for routing operations with their actual and formal parameters,

1. const cropRouter = require('./src/routes/cropRouter')(crops);
2. const contactRouter = require('./src/routes/contactRouter')(db);
3. const adminRouter = require('./src/routes/adminRouter')(serviceAccount);
4. const newsRouter = require('./src/routes/newsRouter')(news);
5. const weatherRouter = require('./src/routes/weatherRouter')();
6. const aboutRouter = require('./src/routes/aboutRouter')();

7. PROJECT TESTING

7.1 VARIOUS TEST CASES

Table 1 : test case scenario #1

Test case description	Test case	Pre conditions	Post Condition	Test steps
Check login functionality.	check response on entering valid username and password.	Latest browser must be installed, Chrome, Safari preferred.	Time, date of login along with username and password is stored in the database.	1. Launch application using valid web-link. 2. Enter username. 3. Enter Password. 4. Click on the submit button.

Test data	Expected results	Actual results	Pass/Fail
Username : Farmer, farmer, farmer100 Password : Pass_farmer	Login must be successful, the user must be taken to the homepage.	Login Successful	Pass

Table 2 : test case scenario #2

Test case description	Test case	Pre conditions	Post Condition	Test steps
Check	Check	Login to the	Fetch and	1. Launch

webpage navigation functionality and accuracy in information retrieval.	response on selecting particular crop for information retrieval.	application using the valid web-link.	display crop information from the database.	application. 2. Login using valid username and password. 3. Select crop information option. 4. Select preferred crop.

Test data	Expected results	Actual results	Pass/Fail
Crop selection (bajra)	Users must be navigated to the selected crop page (ie.,agri-seva-app /crops/bajra) and information retrieval must be successful.	Information retrieval and page navigation was successful.	Pass

Table 3 : test case scenario #3

Test case description	Test case	Pre conditions	Post Condition	Test steps
Check Chat with an expert/ expert consultation functionality.	a) check response on entering valid name, contact number, and description of problem. b) Check response on entering invalid name, phone or description.	Minimum 4-8 mbps of bandwidth must be available, Plain alphabets or numbers must be used.	Name of the user, contact number along with query description is sent to Agri Seva App team, User ID generated.	1. Launch application. 2. Login using valid username and password(optional). 3. Select Chat with an expert/ expert consultation option 4. enter a valid set of details and query.

Test data	Expected results	Actual results	Pass/Fail
Name : Farmer, farmer,farmer99 (valid samples) Contact Number: 9999999999 (valid sample) Query-descriptio : Need help in improving overall crop quality.	e-form submission must be successful and id must be generated.	Form submission and id generation is successful.	Pass

7.2 BLACK BOX TESTING

Functional Testing :

Black Box Testing : Case #1 - Buttons and options response check

BlackBox Testing : Case #2 - Page routing and formatting check

BlackBox Testing :Case#3 - Form submission and input validation

System Testing : System testing has been performed to check and validate if all the components, functions are properly streamlined and has no variations at design, architecture level.

Regression Testing : to ensure any changes made at the interface level, after code fixes, upgrades, during addition of new features or any other system maintenance is not affecting other components of the app or other functionalities regression testing was helpful.

7.3 WHITE BOX TESTING

Unit Testing : all major components, functions and unit fragments are tested here

for major bug identification at early stages of development of Agri Seva App.

Table 4 : white box testing : test case #1 and #2

Test Case #	Test Case Description	Unit to be tested	Result

1	To test whether the app.post function is successfully able to send contact form data to the Agri Seva app associated database.	app.post('/submit-form', (req, res) => {
		const { name } = req.body;
		const { phone } = req.body;
		const { desc } = req.body;
		
		const send = {
		name,
		phone,
		desc,
		};
		
		const formRef = db.ref('/contact');
		formRef.push(send);
		//debug(`${send}updated`);
		res.redirect('/contact');
		res.end();
		// debug(send);
		});
			app.post successfully sends user contact data to the database, function has no bugs and is working fine.

2	To test whether routing and page navigation is working correctly and redirection is taking place as prescribed.	const cropRouter = require('./src/routes/cropRouter')(crops);
		const contactRouter = require('./src/routes/contactRouter')(db);
		const adminRouter = require('./src/routes/adminRouter')(serviceAccount);
		const newsRouter = require('./src/routes/newsRouter')(news);
		const weatherRouter = require('./src/routes/weatherRouter')();
		const aboutRouter = require('./src/routes/aboutRouter')();
			Page navigations are happening as needed and have no issues.

Table 5 : white box testing : test case #3 and #4

Test Case	Test Case Description	Unit to be tested	Result

#			

3	Following functions are tested in order to make sure firebase.database(); and db.ref('/news'); functions are able to fetch data from its associated database upon user request.	a) const db = firebase.database();
		const ref = db.ref('/crops');
		ref.once('value', (snapshot) => {
		snapshot.forEach((childSnapshot) => {
		crops.push(childSnapshot.val());
		});
		});
		
		b) const newsRef = db.ref('/news');
		newsRef.once('value', (snapshot) => {
		snapshot.forEach((childSnapshot) => {
		news.push(childSnapshot.val());
		});
		});
			Both functions a and b are working correctly and are efficient in fetching data in correct format.

4	Check and test database initialization function to make sure creation of databases and tables are happening based on the data model defined.	admin.initializeApp({ credential: admin.credential.cert(serviceAccount), databaseURL: 'https://agriseva-fb852.firebaseio.com', });	The database initializer function is working error free.

In addition to above testing techniques Agri Seva App is tested for compliance,

Validation testing: test has been performed to make sure the product is compliant and can actually create some impact, is performed during and after the development stage.

Majorly tested for :

1. Business requirements and sector relevance validation.
2. Product relevance and compliance.
3. Impact of solution.

8. OUTPUT SCREENS

8.1 USER INTERFACES

Homepage: Homescreen

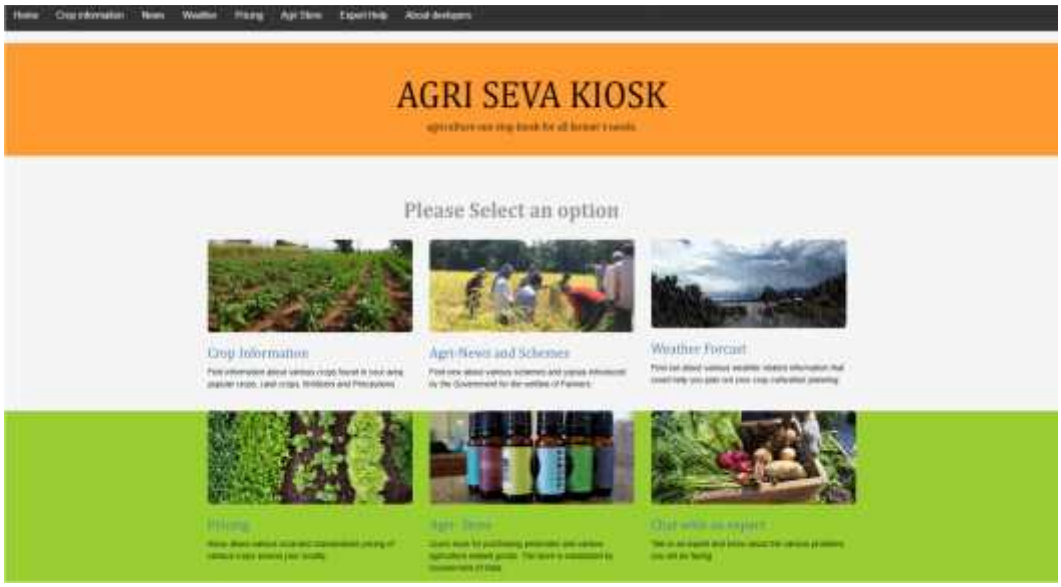


Figure 9 : Homepage- Agri Seva App

Weather forecast page: Homepage/Weather Forecast

Figure 10 : Weather forecast - Agri Seva App

Crop selection page: Homepage/Crop information/Crop selection

Figure 11 : Crop Selection page

8.2 OUTPUT SCREENS

Crop information page: Homepage/Crop information/Crop selection/ Bajra(sample crop)

Figure 12 : Crop information page

Chat with an expert page: Homepage/Chat with an expert

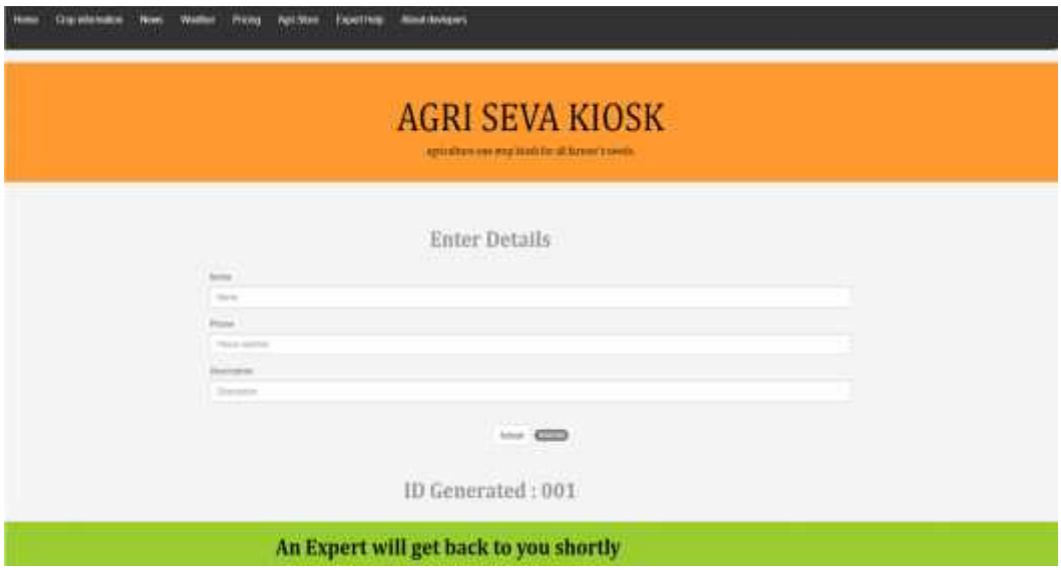


Figure 13 : Chat with expert - query form.

Algorithm Output: Sample output on running logistic regression model for crop recommendation.

```
[57] from sklearn.metrics import classification_report
     print(classification_report(y_test,predictions))
```

	precision	recall	f1-score	support
0	0.00	0.00	0.00	12
1	1.00	0.11	0.20	9
2	0.20	0.10	0.13	10
3	0.00	0.00	0.00	5
4	0.33	0.15	0.21	13
5	0.00	0.00	0.00	11
6	0.07	0.17	0.10	6
7	0.00	0.00	0.00	11
8	0.18	0.44	0.26	9
9	0.00	0.00	0.00	11
10	0.00	0.00	0.00	8
accuracy			0.09	105
macro avg	0.16	0.09	0.08	105
weighted avg	0.17	0.09	0.08	105

Figure 14 : Output screen on performing regression using a set of sample data.

9. EXPERIMENTAL RESULTS

Experiment 1: User trying to fetch “crop information” related to “wheat”

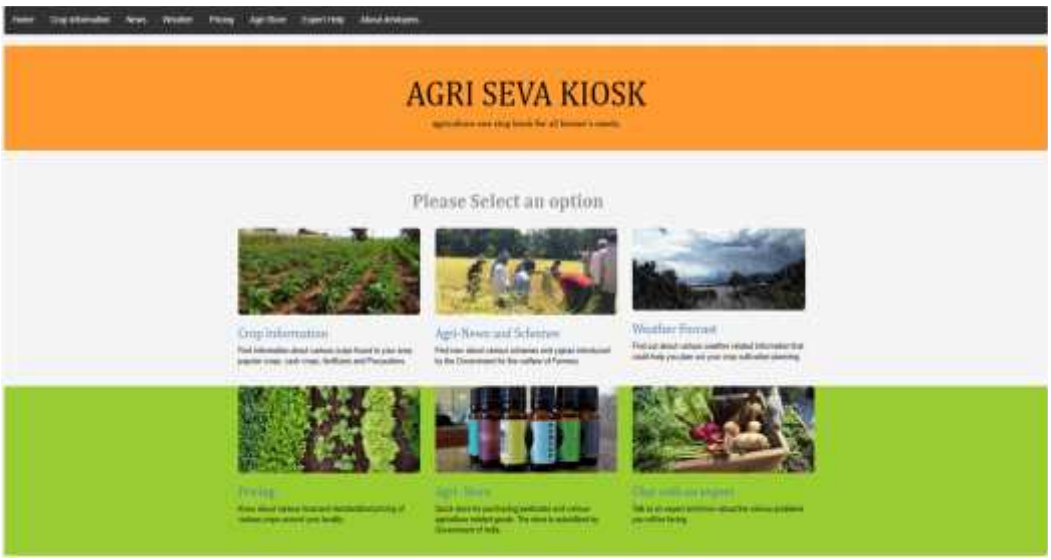


Figure 15 : Experimental result 1 - MainScreen

HomeScreen - Select desired option (Crop information/Agri-news/Weather forecast, Agri-store, Chat with expert)

Next, **Crop selection page** - user will have to select desired crop, selecting wheat for this experiment

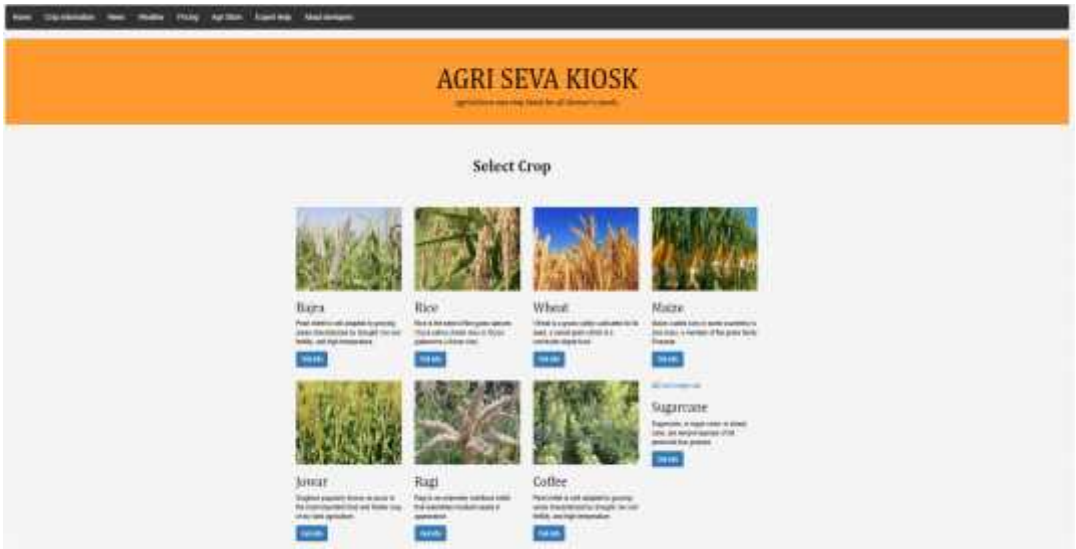


Figure 16 : Experimental result 1 - Crop selection page

Finally, **Crop information page** - users will be displayed with vital information related to the selected crop, and it can make use of this information and could improve its farming capacity and quality.

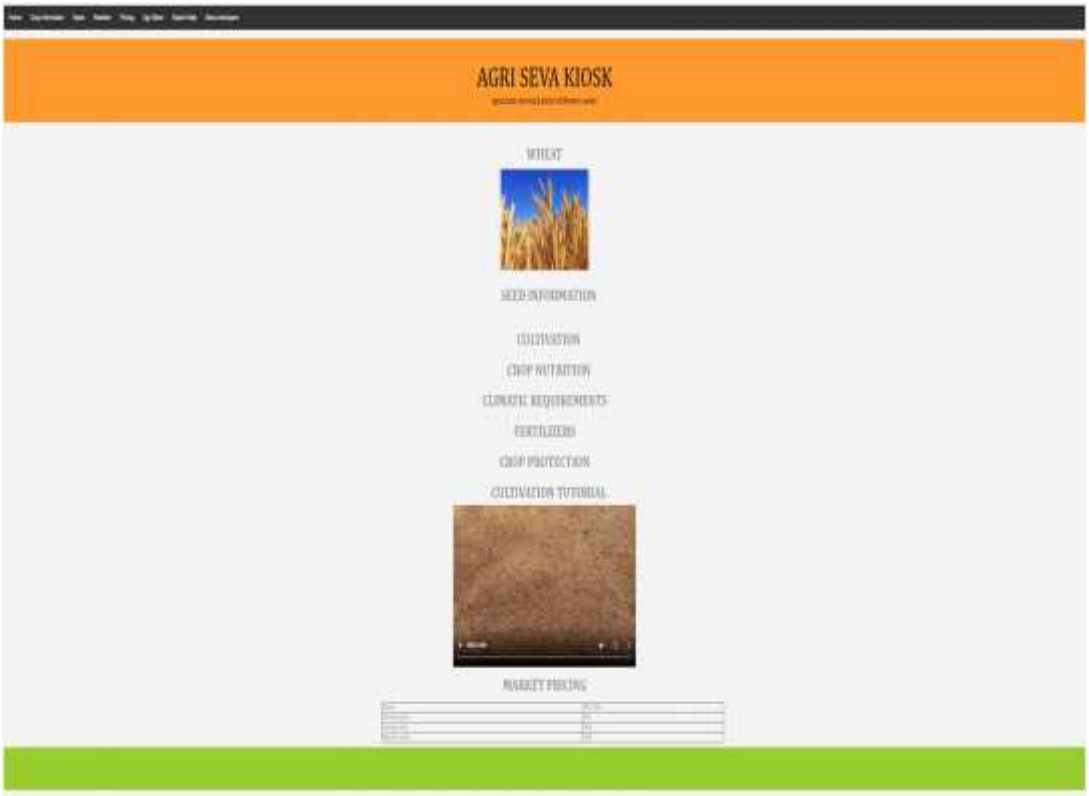


Figure 17 : Experimental result 1 - detailed crop info.

Experiment 2 : A farmer is in need of some additional help and let's see how "chat with an expert" is useful

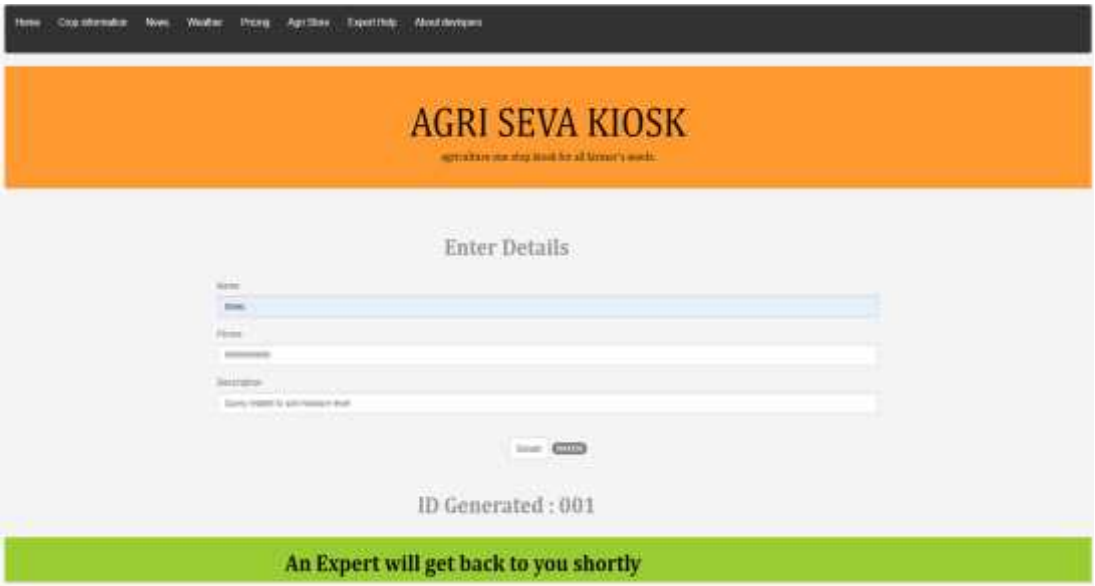


Figure 18 : Experimental result 2 - Query form.

Next, the farmer will have to enter its name, contact number, and a simple description of the query , then after clicking on submit the request will be sent to Agri Seva App team and an id will be generated which the user needs to preserve for future references and follow ups.

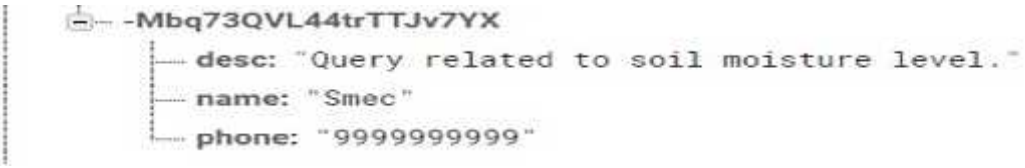


Figure 19 : Experimental result 2- storing and fetching query at backend.

At our end the user query will be saved in the above format, thereafter an agri expert will get in touch with the user as soon as possible and try to resolve the query to the fullest, and close the request if resolved.

10. CONCLUSION AND FUTURE ENHANCEMENT

Agri Seva App is a transparent, unbiased and theocratic approach towards the welfare of farmers and the agri sector on the whole. The principal objective of this research work, design and application development is to facilitate a smart, reliable and sustainable solution to problems related to the farming sector, agri sector and for the welfare of rural India development.

Future goal:

- To deploy Agri Seva App and remain widely available and accessible, install kiosks and make farmers aware about the various benefits of it.
- To provide slip style printout facility to resolve availability in kiosk model.
- To implement AI Assistant for immediate user query resolution and add multi language support.

Future outreach strategy/activity:

- Kiosk will be set up, internet and electrical connections will be provided. Demos and workshops will be conducted at the initial stages to make farmers acquainted with it.
- The kiosk machines are placed at various places in villages such as Post offices, Banks, VR offices, panchayats etc. to facilitate the farmer.

REFERENCES

[1] Barh, A., & Balakrishnan, M. (2018). Smartphone applications: Role in agri-information dissemination Smart phone applications: Role in agri-information dissemination.

[2] Kumar, Y. P., Shivacharan, G., Raghuveer, M., Poshadri, A., Kumar, M. S., & Ramadevi, A. (2020). Evaluation of Mobile Based Advisory Services on Agri and Allied Sectors in Adilabad District. *Int. J. Curr. Microbiol. App. Sci*, *9*(7), 1191-1196.

[3] Pal, S., Marwaha, S., Arora, A., Choubey, A. K., Singh, A. K., Poswal, R. S., ... & Kumar, S. (2019). KVK Mobile App: An ICT tool to empower farmers.

[4] Raj, S., & Darekar, A. Farming 2.0: Digitising Agri Value Chain.

[5] Waiker, V., Dongre, P., & Lohi, R. (2016). Mobile Apps: Next revolution in Business Economy. *International Journal in Management & Social Science*, *4*(3), 207-211.

[6] Pawar, H. Y., Kapse, P. S., & Puri, S. G. (2020). Utility of AgroTech VNMKV Mobile App as Perceived by the User Farmers. *Int. J. Curr. Microbiol. App. Sci*, *9*(4), 1280-1286.

[7] Balkrishna, A., Sharma, J., Sharma, H., Mishra, S., Singh, S., Verma, S., & Arya, V. (2021). Agricultural Mobile Apps used in India: Current Status and Gap Analysis. *Agricultural Science Digest*, *41*(1).

[8] Kumar, A., Kumar, S., Khan, N., Singh, C. B., & Singh, S. Weather based agromet advisory bulletin to the farmers under Gramin Krishi Mausam Seva (GKMS) project during lockdown period (Covid-19) at Kanpur region of Uttar Pradesh.

[9] Qiang, C. Z., Kuek, S. C., Dymond, A., & Esselaar, S. (2012). Mobile applications for agriculture and rural development.

[10] Mohan Kumar, S., Suman, S., Kulkarni, U. P., & Siddalingaswamy, N. H. (2019). Feasibility study of effective usage of available Agricultural Information System for various Village Boundaries of India. *J Robot Mech Eng Resr*, *3*(2), 1-7.

[11] Sharma, A., & Kumar, S. Information and Communication Technology (ICT) in Indian Agriculture. *Agri Mirror: Future India ISSN: 2582-6980*.

[12] SARMA, V. S., & KATTA, P. K. (2020). SMART FARMING USING AGRI AIUTO APP IN INDIA. *Journal of Natural Remedies*, *21*(4), 239-243.

CAPTURING PUBLICATIONS

JOURNAL : UGC Journals

CONFERENCE : International conference on “Innovation in Computer Networks, Computational Intelligence and IOT " **[ICICCI-21], C7 batch.**

Paper title : Agri Seva App : platform for farmers, agriculture, farming information and services providing app

PAPER ID : “ICICCI-21-0059"

PRESENTED AND PUBLISHED PAPER

Agri Seva App : platform for farmers, agriculture, farming information and services providing app

Shubham Agarwal[1], Suraj Sharma[2], Koushik Bhargava[3], Akash Singh Rawat[4], Dr. N. Satheesh[5]

[1,2,3,4] UG Scholar, [5] Professor

Department of Computer Science and Engineering,

St. Martin's Engineering College,

Near Forest Academy, Dulapally, Kompally, Secunderabad, Telangana 500 014, India

E-Mail: agarwal717@gmail.com[1], suraj10620@gmail.com[2], koushik.bhargava@gmail.com[3], rawatakash678@gmail.com[4] , drsatheeshcse@smec.ac.in[5]

***Abstract*- A research work was undertaken under rural India agricultural development with an initiative and intention of providing easily accessible informational resources, services and to heighten awareness, Agri Seva App is a platform for farmers and for any other people working in agri and farming sector, it has various features through which it aim to provide agricultural information and services using technology, the main aim of the research is to facilitate the farmers by educating them by using a web-app or a mobile-app or a kiosk machine that will provide them information on farming, right usage of related commodities, suitable weather and climatic conditions and other services like, real-time pricing, expert consultation, and agri store.**

Keywords: Agri information and services, Agri Store, AI voice assistant, climate and crop suitability analysis, Rural development.

I. INTRODUCTION

In the present scenario with advent in technology where different sectors in the world are experiencing an era of advancement in technology along with ease of access to information and services yet the farming and agri sector is lagging behind even though being one of the most important sectors where humans serve for humans and is a major sector supporting the means of human life, survival and social prosperity, besides in a country like INDIA where agriculture supports the economy and where share of agriculture in GDP is continuously increasing ie. to 19.9 percentage in 2020-21 from 17.8 percent in 2019-20, we think we have a lot more to offer to the agriculture-farming sector and vice versa.

Problem identified during research and Literature Survey:
Following paragraph covers basic problem statement and need for undertaking following research work, though many NGOs, innovators/ entrepreneurs, research institutes and krushi vikas organizations are trying hard to provide farming, agricultural information and services still the information and services are not easily accessible to the people in that sector, one of the major reasons we identified during this research is lack of proper medium and channel, reliability, and accessibility via which the information cannot only be delivered but delivered in time, as old and expired information would be worthless, some other problems that we identified are lack of standardized pricing, lack of guidance and consultation when needed, lack of genuine sources for purchasing agri commodities like seeds, fertilizers, pesticides, and majorly no platform which provides all mentioned services combined in one app or portal. Here Agri Seva App comes into the picture and aims to solve mentioned problems by sustainable, user-friendly use of technology and by means of the internet.

Proposed research work and solution in depth:
Firstly, let's discuss all the features and functionalities available on agri seva kiosk and get ourselves familiar to different terminologies related to it, hence, the various features are: Crop information, Agri-news and Schemes, Weather forecast, Real time standardized pricing, Agri-store, Expert consultation with Artificially intelligent voice assistant and Drop query in any language to hear back from an agri expert option, all in all the features and information will be made available via a web-app, a handy mobile app and also via a kiosk machine which is very similar to an ATM Machine, the option of kiosk machine is added for a strong reason, being that there are a number of farmers who are still very below the poverty level and are minimum wage earners, how could we expect them to have a smart device with an internet subscription, hence, as this particular category of farmers would find it difficult to access the services we have whole unique concept of kiosk machines for them which they can use and gain all that insightful information and educate themselves equally to any other farmer. Therefore, the first person to get benefited by the proper use of Agri Seva App would be the farmer, and using such latest tools, techniques and methodologies adequately will lead
to holistic development of farmers and a gradual increase in profitability and overall crop quality will be experienced.

Implementation, features, functionalities and working of Agri Seva App in depth:
let's deep dive into implementation of functionality and working of each feature mentioned hand in hand algorithms, techniques and technologies used, with rapid increase in use of mobile applications and web applications among almost all age group of people, making agri seva app available via a mobile application is must in addition to the mobile application Agri Seva App is made available via web-application and kiosk machine, this was about the infrastructure as a service part.

Coming to the services and functionalities part of Agri Seva App in detail, The very first is **Crop Information** - this is a page which will cover all informational aspects, information on prevailing practices and methodologies, from sowing a crop to bowing

or harvesting it followed by an informative and insightful tutorial video as we believe pictorial and added visual oriented method is the most effective in not only transferring the knowledge but also in retaining knowledge for longer time, particularly details like: Seed Information, Cultivation procedure, Crop Nutrition, Climatic Requirements, use of right fertilizers with quantity and information Crop Protection and quality improvement.

Coming to the next feature which is **Agri News and Schemes** - using this feature, a farmer will be able to receive information on latest schemes and policies introduced by the government and news related to the agri sector so that the farmer can keep itself up to date at any point in time.

Moving to the next feature which is **weather forecast**, as we see, weather and climatic conditions these days are not stable and, in fact, are very fluctuating. In this situation it becomes difficult to make right decisions about the right selection of crops and whether the upcoming climatic conditions will be suitable for that particular crop or not. Hence, Agri Seva App makes it possible to understand and make right selection of crop, and it makes it possible using the weather forecast feature which provides a list of suitable crops and guides the farmer with the best selection of crop, this feature is making use of Global Forecast System for its first part of data on which the model will perform analysis, Convolutional Neural Network(CNN) as core algorithm, lastly IBM Watson-Cognitive Recognition Engine for working AI assistant, moving forward reader will be femalarised with the implementation and working of each algorithm in detail.

Next, is **pricing**, with the help of pricing options, farmers can know the latest pricing of any crop selected, and this in turn will help the farmer to sell their crops, grains or pulses at the right price put down by the government.

Moving forward we have **Agri Store**, this feature is meant to support both farmers and Agri commodity vendors, sellers, for farmers agri store will act as a reliable source to purchase high quality seeds, fertilizers etc, and for vendors it will act as a marketplace to sell heir agri products, only products which are of high quality and gauged to produce top quality crop will be available to maintain overall quality aspect of Agri Store.

Last but not the least we have **Expert consultation with Artificially intelligent voice assistant** and Drop your query in any language to hear back from an agri expert option, with the use of mentioned feature farmer can seek help from an agri expert associated with Agri Seva App, there are two ways of using this option, one is one-on-one chat with the embedded AI Assistant as get immediate resolution to queries but this option comes with limited language support and to overcome this limitation we have the second mode that is a farmer can record its query as a message in any language and sent it along with its name, mobile number and query type using the form available within the option on the portal, as a result the farmer will get a call back for an agri expert and he/she will resolve all the queries and provide solution the problems that the farmer might encounter.

Lastly, during the research we identified some **additional possibilities of complications** following is the probable solution after performing the research and analysing the situation in depth, in case a farmer does not have access to a smart mobile phone or internet subscription and selects to use the kiosk model of Agri Seva App there is a chance that it might face difficulty in remembering all the new knowledge learnt as the information would include numbers example; for different quantities and furthermore availability issues due to long queues as the kiosk machine vs farmer ratio would be very vast for obvious reasons, hence, to resolve both the issues kiosk machine will have a print option using which farmer can get printouts of any sections selected on the crop information tab and it back to their farming space with them.

1. Convolutional Neural Networks.
2. Global Forecast System.
3. IBM Watson-Cognitive Recognition Engine.

Convolutional Neural Networks(CNN)

Convolutional Neural Networks, also called ConvNets, is a very mature deep learning algorithm. Iit was first introduced in the 1980s by a postdoctoral computer science researcher Yann LeCun. With time CNN was improved and numerous changes were

made in order to overcome all associated drawbacks and make it quintessential. The structure of CNN consists of 3 layers,

Let's start with the top layer

(a) The Math Layer - The top layer is considered as the mathematical layer or feature extraction layer. Math layer is convolutional layer which helps in understanding and deals with the number pattern it sees, the layer is consists of filter also referred as neurons or kernel which helps in reading the data and forms conclusion in the form of array of numbers, then it uses the generated array performs multiplication operation on the array and deduces a single number which is also the outcome of this stage.

(b) The Poling Layer - the primary aim of the pooling layer is to decrease the size of the convolved feature map and reduce the computational cost. This tries to reduce the connections between layers, depending upon pooling operation selected among max, average or sum pooling.

(c) The Fully Connected Layer - it is the final layer encompassing all the interior complexities, it is also the completion layer in a CNN, the fully connected feature will help in understanding high-level outcomes and thus providing output of classification.

(d) Usage and application strategy of CNN in Agri Seva App - working of CNN here will depend of two categories of data, one being the weather data coming from Global Forecast System which will be further discussed in coming sections of this research, other being the data set gathered related to crop and its suitable weather conditions, the model will be trained to generate an AI model which will take also take newly generated data and will predict the best crop option for the current and expected weather conditions.

(e) Pictorial representation of working of CNN -

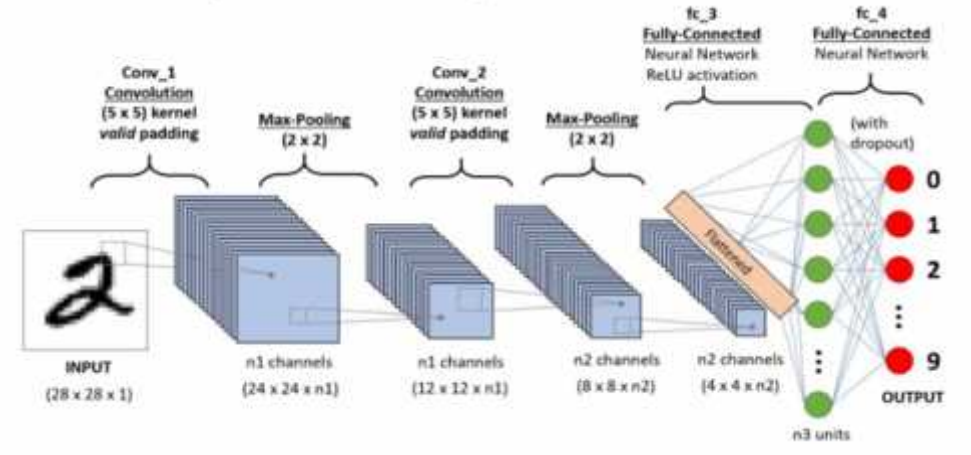


Figure : CNN and its layer(Source of image: TowardsDataScience)

Global Forecast System

GFS is a global numerical weather prediction system containing a global computer model and variational analysis run by the United States national weather service(NWS). The mathematical model runs four times a day, and produces forecasts for up to 16 days in advance, but with decreased spatial resolution after 10 days. The forecast skill generally decreases with time (as with any numerical weather prediction model) and for longer-term forecasts, only the larger scales retain significant accuracy. It is one of the predominant synoptic scale medium-range models in general use.

Moving on to the **usage of GFS in Agri Seva App**, the app will make API calls to GFS and rely on it for the current and past weather data, obtained data will act as first part of data and will be used in training of model, then using the second set which will contain data related to crop and its suitable weather the algorithm will do the job of decision making and finally provide us with the list of best suitable crops.

IBM watson cognitive recognition engine using clustering of intents and recognised instruction set for AI assistant

Watson Assistant is a conversational AI platform that leverages customer support, some advantages are - its fast, straightforward and accurate answers to queries, across any application or device. It is generally used to create AI-driven conversational flows, to embed existing help content, to bring the assistant to your customers andwhere they are,

used to track customer engagement and satisfaction.

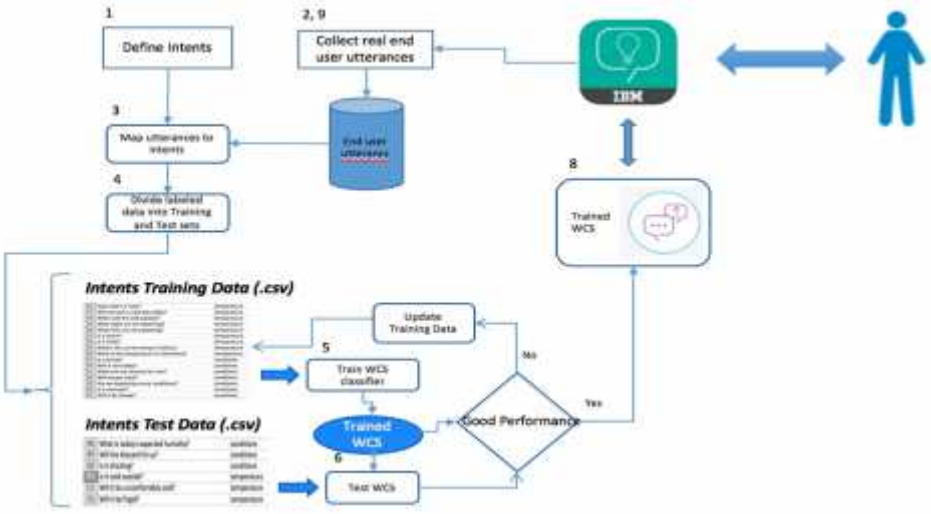


Figure 2 : IBM Watson-cognitive recognition Engine using Clustering of intents and recognised instruction sets.

Results and Discussion:

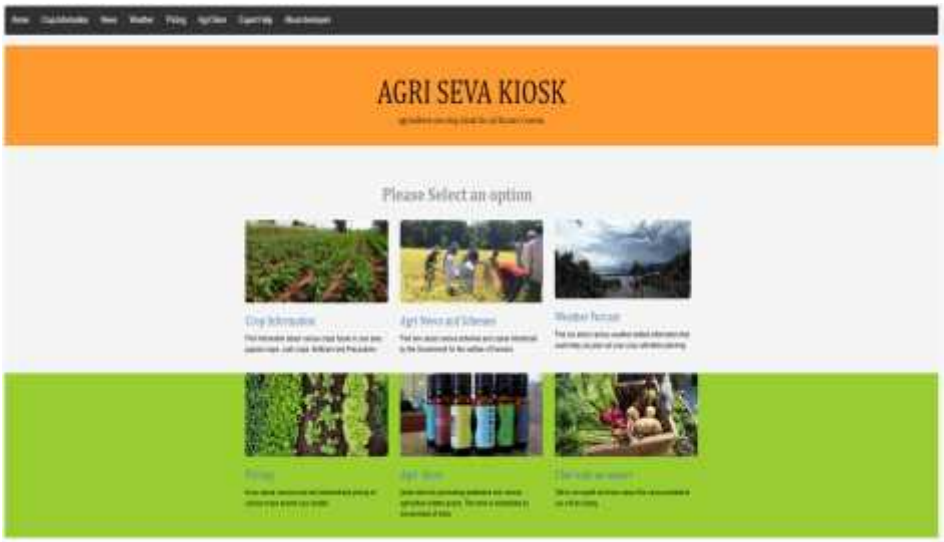


Figure 3 : Agri Seva App - Home Screen

A working prototype of Agri Seva App is available @ : https://agri-seva-kiosk.herokuapp.com for better understanding of reader and to have a glimpse of the app,

the prototype includes most of the findings, features, functionalities and methodologies discussed in this research work, using above prototype we have made it possible for peer members, scholars and other researchers to get an idea of what the app does, how it does and its probable impacts.

Conclusion

Agri Seva App is a transparent, unbiased and theocratic approach towards the welfare of farmers and the agri sector on the whole. The main goal of this research work is to provide a smart, reliable and sustainable solution to problems related to the farming sector, agri sector and for the welfare of rural India development.

STUDENT PROFILE

Shubham Agarwal is currently pursuing his Bachelors of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. He has completed his intermediate from Narayana Junior College and 10th class from Gowtham Model School. He is the president of Cyber Security Club and a member of Coders Club at our college. His responsibilities included mentoring, creating awareness, motivating students and helping with their projects, profil building and to find reliable and faster solutions. During his engineering he has been awarded a memento for actively working for the development of CSE department, adding to this he has received certificates and letters of appreciations from college for planning, conducting and managing a number of technical events and hackathons for the students of computer science, to name few, "International Conference-2020 - www.icrcsit.com" (june 2020), 2-days National Hackathon conducted on 7th and 8th february 2020, "Industrial Visit to BSNL for the students of CSE department" (Sep,2019), planning and conducting college level "Project Expo"(Aug,2019), and has helped the college in organizing "Employability Skill development Program conducted by Zensar". Coming to his skills he holds good managemental, interpersonal and technical skills some of his top skills are Python programming, data management and visualization, cloud computing, embedded AI, theoretical ML, CRM tools, project management etc, he has always implemented his skills in real life as an outcome some of his other projects and research works includes, Code Ease - a self learning user friendly platform for learning programming in less time(with 1 other member, Jan, 2021), research and solution - Wearable IOT with complex Artificial perception embedding for Alzheimer patients - probable and portable solution(Apr, 2020). moving on to his professional exposure and experiences, he has worked as a Business Process Manager in an Electrical Manufacturing Plant located at Balanagar-Hyderabad(jun-Aug 2019), as Associate-Operations and a member of Analytics Team at Uable a startup located in bangalore(Aug-Sep,2020), and as Operations Associate with CampK12 EdTech Startup from Gurugram(May-Jul, 2020).

Koushik Bhargava Rupanakuntla is currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. He completed his intermediate from Vijaya Ratna Junior College and 10th class from Oxford High School. He is one of the members of Coders Club and Network Security club in our college. His responsibilities in that group include mentoring and motivating students to take coding as a serious hobby. He is also a part of the organizing team of Technology Awareness Month which is a month-long fest conducted by the students for the development of other students. He is part of the Technical Department in the fest. His technical skills include C, C++, Python, JavaScript, Android, Nodejs and ReactJs. He took part in the Employability Skill development Program conducted by Zensar. He is also a student of Smart Interviews. His areas of interest are Python, Artificial Intelligence, Machine Learning and Deep Learning. He has experience of Teaching and creating coding curriculum for various companies like Camp K12, Toppr and Cuemath. He is currently working as a Software Developer and Instructor at Smart Interviews.

Suraj Prasad Sharma is currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin’s Engineering College. He completed his 12th from Prabhuji English Medium and 10th class from Loyola Public School. He is a member of the Cybersecurity club in our college. His responsibilities in that group include mentoring and motivating students to take careers in Cybersecurity. His technical skills include Python and Java. He also has a fundamental experience of Web Development and AI. He took part in the Employability Skill development Program conducted by Zensar. He is a Google Certified Educator. His participations include: National Level Three Day Online Workshop on “AI & ML in Speech and Audio Processing” which was conducted from 10th to 12th December 2020, “Know More - Teach More “, the Global Webinar on Cloud & Big Data – Changing the way we work which was conducted Global Education & Careers Forum (GECF) on 12th August 2020, Google IO Summit, Google Developers conference .and IIC Online Sessions conducted by Institution's Innovation Council (IIC) of MHRD's Innovation Cell, New Delhi to promote Innovation, IPR, Entrepreneurship, and Start-ups among HEIs from 28th April to 22nd May 2020.His flagship project also got selected for Dronacharya Awards 2019 in state level. Her areas of interest are Web Technologies, Applications of Machine Learning and Agile Methodologies in Product Management. He is a Google Certified Educator, Hackerrank certified Javascript professional and has linkedin verified skills for Agile, Javascript and Python

Akash Singh Rawat is currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin’s Engineering College. He completed his intermediate from Army Public School , Jammu And Kashmir and 10th class from Kendriya Vidyalaya Sunjuwan . He is one of the members of Coders Club in our college. His responsibilities in that group include mentoring and motivating students to take coding as a serious hobby. His technical skills include C++, Python , Java , Node.js , React.js . He took part in the Employability Skill development Program conducted by Zensar. He is also a student of Smart Interviews. He also participated in the National Level Hackathon conducted by JNTU. His areas of interest are Python, Artificial Intelligence, Machine Learning and Deep Learning .

APPENDICES

A working prototype of Agri Seva App is available @ : https://agri-seva-kiosk.herokuapp.com for better understanding of reader and to have a glimpse of the app, the working model includes most of the findings, features, functionalities and methodologies mentioned in this project documentation, using above working model we have made it possible for peer members, scholars and other researchers to get an idea of what the app does, how it does and its probable impacts.

A

PROJECTREPORT

On

CREDIT CARD FRAUD DETECTION USING RANDOM FOREST AND CART ALGORITHM

Submitted by

1)Ms.A.Madhuri(17K81A05C1) 2)Ms.G.Mounika(17K81A05D7)

3)Mr.M.Satya(17K81A05F6) 4)Ms.K.Pranavi(17K81A05H1)

in partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Under the Guidance of

Mruthyunjayam.A

Asst.Professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ST.MARTIN'S ENGINEERING COLLEGE
An Autonomous Institute
Dhulapally, Secunderabad – 500100

JUNE 2021

BONAFIDE CERTIFICATE

This is to certify that the project entitled Credit Card Fraud Detection using Random Forest and Cart Algorithm, is being submitted by**1.Ms.A.Madhuri (17K81A05C1), 2.Ms.G.Mounika (17K81A05D7), 3. Mr. M.Satyanarayana (17K81A05F6), 4. Ms. K.Pranavi (17K81A05H1)**in partial fulfillment of the requirement for the award of the degree of **BACHELOR OF TECHNOLOGY IN Computer Science** is recorded of bonafide work carried out by them. The result embodied in this report havebeen verified and found satisfactory.

Mruthyunjayam. A
Department of
CSE

Head of the Department
Dr.M.NARAYANAN
Department of CSE

Internal Examiner

ExternalExaminer

Place:

Date:

TUESDAY, 15 JUNE 2021

INTERNSHIP CERTIFICATE

THIS IS TO CERTIFY THAT **MADHURI** WITH ROLL NO.**17K81A05C1, MOUNIKA** WITH ROLL NO.**17K81A05D7**, **PRANAVI** WITH ROLL NO.**17K81A05H1**, **SATYA** WITH ROLL NO.**17K81A05F6**, OF B.TECH – IV YEAR, **COMPUTER SCIENCE ENGINEERING DEPARTMENT** OF **ST. MARTIN'S ENGINEERING COLLEGE**, KOMPALLY, SECUNDERABAD HAVE COMPLETED ONE MONTH INTERNSHIP PROGRAM AT **LASYA IT SOLUTION PVT. LTD,KOMPALLY.**

DURING THE PERIOD, THEY HAVE SUCCESSFULLY COMPLETED MAJOR PROJECT TITLED "**CREDIT CARD FRAUD DETECTION USING RANDOM FOREST & CART ALGORITHM**" AT OUR DEVELOPMENT CENTER,KOMPALLY.

WE WISH THEM SUCCESS IN THEIR FUTURE ENDEVOUR.

ORUGANTI VENKAT
DIRECTOR
TRAININGS & PLACEMENTS LASYA IT SOLUTIONS
PVT LTD.

LasyaITSolutionsPvtLtd,BehindCinePlanet,Kompally,MedchalRoad,Secunderabad500014 Email : contact@lasyainfotech.com,v@lasyainfotech.com
Website :www.lasyainfotech.com| contact: 7330666881/82/83/84/86

DECLARATION

We, the student of **Bachelor of Technology** in Department of Computer Science and Engineering', session: 2017 – 2021, St. Martin's Engineering College, Dhulapally, Kompally, Secunderabad, hereby declare that work presented in this Project Work entitled CREDIT CARD FRAUD DETECTION USING RANDOM FOREST AND CART ALGORITHM is the outcome of our own bonafide work and is correct to the best of our knowledge and this work has been undertaken taking care of Engineering Ethics. This result embodied in this project report has not been submitted in any university for award of any degree.

A.Madhuri (17K81A05C1)
G.Mounika (17K81A05D7)
M.Satyanarayana (17K81A05F6)
K.Pranavi (17K81A05H1)

ABSTRACT

The project is mainly focussed on credit card fraud detection in real world. A phenomenal growth in the number of credit card transactions, has recently led to a considerable rise in fraudulent activities. The purpose is to obtain goods without paying, or to obtain unauthorized funds from an account. Implementation of efficient fraud detection systems has become imperative for all credit card issuing banks to minimize their losses. One of the most crucial challenges in making the business is that neither the card nor the cardholder needs to be present when the purchase is being made. This makes it impossible for the merchant to verify whether the customer making a purchase is the authentic cardholder or not. With the proposed scheme, using random forest algorithm the accuracy of detecting the fraud can be improved can be improved. Classification process of random forest algorithm to analyse data set and user current dataset. Finally optimize the accuracy of the result data. The performance of the techniques is evaluated based on accuracy, sensitivity, and specificity, and precision. Then processing of some of the attributes provided identifies the fraud detection and provides the graphical model visualization. The performance of the techniques is evaluated based on accuracy, sensitivity, and specificity, and precision.

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible and whose encouragements and guidance have crowded effects with success.

We extended our deep sense of gratitude to Principal**, Dr. P. SANTOSH KUMARPATRA**, St. Martin's Engineering College, Dhulapally, for permitting us to undertake this project.

We are also thankful to **Dr. M.NARAYANAN**, Head of the Department, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

We would like to thank **Dr. T. POONGOTHAI,** Department of Computer Science and Engineering (AI & ML) for her encouragement and insightful comments and as well as our project coordinators **Dr. B.RAJALINGAM**, Associate Professor and **Dr. N. SATHEESH**, Professor, in Department of Computer Science and Engineering for their valuable support.

We would like to express our sincere gratitude and indebtedness to our project supervisor Mruthyunjayam.A Asst.professor , Computer Science and Engineering,St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

Finally, we express thanks to all those who have helped us successfully to completing this project. Furthermore, we would like to thank our family and friends for their moral support and encouragement.

We express thanks to all those who have helped us in successfully completing the project.

A.Madhuri (17K81A05C1)
G.Mounika (17K81A05D7)
M.Satyanarayana (17K81A05F6)
K. Pranavi (17K81A05H1)

TABLEOFCONTENTS

LISTOF FIGURES

LIST OF OUTPUT SCREENS

LIST OF ACRONYMS

EMV	Europay-MasterCard-VISA
RF	Random Forest
SVM	Support Vector Machine
LOR	Logistic Regression
AIRS	Artificial Immune Recognition System
GP	Genetic Programming
PNN	Probabilistic Neural Network
DT	Decision Trees
BNN	Bayesian Belief Networks
CFDM	Computational Fraud Detection
SOM	Self Organizing Map
POS	Point Of Sale

1.INTRODUCTION

There are various fraudulent activities detection techniques has implemented in credit card transactions have been kept in researcher minds to methods to develop models based on artificial intelligence, data mining, fuzzy logic and machine learning. Credit card fraud detection is significantly difficult, but also popular problem to solve. In our proposed system we built the credit card fraud detection using Machine learning. With the advancement of machine learning techniques. Machine learning has been identified as a successful measure for fraud detection. A large amount of data is transferred during online transaction processes, resulting in a binary result: genuine or fraudulent. Within the sample fraudulent datasets, features are constructed. These are data points namely the age and value of the customer account, as well as the origin of the credit card. There are hundreds of features and each contributes, to varying extents, towards the fraud probability. Note, the level in which each feature contributes to the fraud score is generated by the artificial intelligence of the machine which is driven by the training set, but is not determined by a fraudanalyst. So, in regards to the card fraud, if the use of cards to commit fraud is proven to be high, the fraud weighting of a transaction that uses a credit card will be equally so. However, if this were to shrink, the contribution level would parallel. Simply make, these models self-learn without explicit programming such as with manual review. Credit card fraud detection using Machine learning is done by deploying the classification and regression algorithms. We use supervised learning algorithm such as Random forest algorithm to classify the fraud card transaction in online or by offline. Random forest is advanced version of Decision tree. Random forest has better efficiency and accuracy than the other machine learning algorithms. Random forest aims to reduce the previously mentioned correlation issue by picking only a subsample of the feature space at each split. Essentially, it aims to make the trees de-correlated and prune the trees by fixing a stoppingcriteria for nodesplits.

Protection of Purchases Credit cards may also offer customers, additional protection if the purchased merchandise becomes lost, damaged, or stolen. Both the buyers credit card statement and company can confirm that the customer has bought if the original receipt is lost or stolen. In addition, some credit card companies provide insurance for large purchases.Year after year, the damages inflicted by the credit card fraud problem are growing rapidly. In the year 2014 alone, it was estimated that the total global monetary loss was in 16.31 billion dollars – accumulation of damages from card issuers, acquiring banks and merchants. Although new technology is introduced to the public and being mandated by the government agency to replace the old magnetic stripe to EMV (Europay-MasterCard-VISA), some group of individuals are starting to challenge its design and implementation. The advancements in our technology and ease of availmentopened more opportunities for fraudsters' ability to speed-up their execution plan and retain their anonymity at the same time. Surely, a layer of security will not be enough to protect the card holders, merchants, and issuing banks from a possible attack. There should be another layer that must be available to proactively detect these anomalies. Credit card fraud transpires when a perpetrator uses somebody's credit card for personal gain and sometimes in absolute secrecy or anonymity; even the issuing banks are unconscious that the card is being utilized. Moreover, the perpetrator has no relationship with the cardholder or issuer, and has no intention of informing the card owner of the lost card and making repayments for the transactions made. Evaluation of credit card related fraud cases in the past two decades reveals that the top five modusoperandi performed by the fraud stersare

1.counterfeit creditcards,

2. lost or stolen,

3. no-card fraud (e.g., giving card information to non-legitimatetelemarketer),

4. stolen cards during mailing fraud, andlastly

5. identity-theftfraud.

1.1 PROJECT OVERVIEW

In this proposed project we designed a protocol or a model to detect the fraud activity in credit card transactions.This system is capable of providing most of the essential features required to detect fraudulent and legitimate transactions. As technology changes, it becomes difficult to track the behaviour and pattern of fraudulent transactions.With the upsurge of machine learning, artificial intelligence and other relevant fields of information technology, it becomes feasible to automate the process and to save some of the effective amount of labor that is put into detecting credit card fraudulent activities.

The algorithm heuristically searches for the best attribute combination through cross-over, mutation then evaluated by a fitness function. Another method used an outlier data mining technique to identify fraudulent activities by evaluating the account and transaction data of the card holders. In retrospect, data-mining techniques are now well established. Nevertheless, researches pertaining to this area are very limited due to privacy issues. Bank customers are well protected by several laws that prohibit the disclosure of their personal information without proper consent.However, these methods were able to acquire data – their strategy is to combine information from customers, accounts, cards, and transaction datasets. This study will focus primarily on the transaction details made by the card holder.

The intention of thisstudy is to fully explore the effectiveness of utilizing the credit card transaction logs to differentiate anomalous from legitimate transactions. With this, various learning algorithms available in Weka will be evaluated by measuring their effectiveness in predicting the correct classification of the input dataset.

1.2 PROJECT OBJECTIVES

There are various fraudulent activities detection techniques has implemented in credit card transactions have been kept in researcher minds to methods to develop models based on artificial intelligence , data mining, fuzzy logic and machine learning. Simply make, these models self-learn without explicit programming such as with manual review. Credit card fraud detection using Machine learning is done by deploying the classification and regression algorithms.

We use supervised learning algorithm such as Random forest algorithm to classify the fraud card transaction in online or by offline. Random forest is advanced version of Decision tree. Random forest has better efficiency and accuracy than the other machine learning algorithms. Random forest aims to reduce the previously mentioned correlation issue by picking only a subsample of the feature space at each split.

1.3 ORGANIZATION OF CHAPTERS

Introduction – This chapter covers the overview of our project and its objectives. Literature Survey – This includes the details of our survey, Software and Hardware Requirements – We specify our software and hardware requirements here. Software Development Analysis – This section includes the problem definition and details of the modules we used in our project. Project System Design – This chapter includes the design part of our project which includes uml diagrams. Project Coding – This section contains the details of our project code. Project Testing – The details of test cases and testing are included in this chapter. Output Screens – This contains the screenshots of how our project looks like when executed. Experimental Results – This chapter contains the screenshots of our results. Conclusion and Future Enhancements – This covers the conclusion of our project and the possible future development.

2.LITERATURESURVEY

2.1 SURVEY ONBACKGROUND

A. SINGLE MODELS For credit card fraud detection, Random Forest (RF), Support Vector Machine, (SVM) and Logistic Regression (LOR) were examined in [6]. The data set consisted of one-year transactions. Data under-sampling was used to examine the algorithm performances, with RF demonstrating a better performance as compared with SVM and LOR [6]. An Artificial Immune Recognition System (AIRS) for credit card fraud detection was proposed in [7]. AIRS is an improvement over the standard AIS model, where negative selection was used to achieve higher precision. This resulted in an increase of accuracy by 25% and reduced system response time by 40% [7]. A credit card fraud detection system was proposed in [8], which consisted of a rule-based filter, Dumpster–Shafer adder, transaction history database, and Bayesian learner. The Dempster–Shafer theory combined various evidential information and created an initial belief, which was used to classify a transaction as normal, suspicious, or abnormal. If a transaction was suspicious, the belief was further evaluated using transaction history from Bayesian learning [8]. Simulation results indicated a 98% true positive rate [8]. A modified Fisher Discriminant function was used for credit card fraud detection in [9]. The modification made the traditional functions to become more sensitive to important instances. A weighted average was utilized to calculate variances, which allowed learning of profitable transactions. The results from the modified function confirm it can eventuate more profit [9]. Association rules are utilized for extracting behavior patterns for credit card fraud cases in [10]. The data set focused on retail companies in Chile. Data samples were defuzzified and processed using the Fuzzy Query 2+ data mining tool [10]. The resulting output reduced excessive number of rules, which simplified the task of fraud analysts [10]. To improve the detection of credit card fraud cases, a solution was proposed in [11]. A data set from a Turkish bank was used. Each transaction was rated as fraudulent or otherwise. The misclassification rates were reduced by using the Genetic Algorithm (GA) and scatter search. The proposed method doubled the performance, as compared with previous results [11]. Another key financial loss is related to financial statement fraud. A number of methods including SVM, LOR, Genetic Programming (GP) and Probabilistic Neural Network (PNN) were used to identify financial statement fraud [12]. A data set involving 202 Chinese companies was used. The t-statistic was used for feature subset selection, where 18 and 10 features were selected in two cases. The results indicated that the PNN performed the best, which was followed by GP [12]. Decision Trees (DT) and Bayesian Belief Networks (BNN) were used in [13] to identify financial statement fraud. The input comprised the ratios taken from financial statements of 76 Greek manufacturing firms. A total of 38 financial statements were verified to be fraud cases by auditors. The BBN achieved the best accuracy of 90.3% accuracy, while DT achieved 73.6% [13]. A computational fraud detection model (CFDM) was proposed in [14] to detect financial reporting fraud. It utilized textual data for fraud detection. Data samples from 10-K filings at Security and Exchange Commission were used. The CDFM model managed to distinguish fraudulent filings from non-fraudulent ones [14]. Afraud

detection method based on user accounts visualization and threshold-type detection was proposed in [15]. The Self-Organizing Map (SOM) was used as a visualization technique. Real-world data sets related to telecommunications fraud, computer network intrusion, and credit card fraud were evaluated.
The results were displayed with visual appeal to data analysts as well as non-experts, as high-dimensional data samples were projected in a simple 2- dimensional space using the SOM [15]. Fraud detection and understanding spending patterns to uncover potential fraud cases was detailed in [16]. It used the SOM to interpret, filter, and analyze fraud behaviors. Clustering was used to identify hidden patterns in the input data. Then, filters were used to reduce the total cost and processing time. By setting appropriate numbers of neurons and iteration steps, the SOM was able to converge fast. The resulting model appeared to be an efficient and a cost-effective method [16]

TITLE: A COST-SENSITIVE DECISION TREE APPROACH FOR FRAUD DETECTION

AUTHOR'S:

Y. SAHIN

S. BULKAN

E. DUMAN

With the developments in the information technology, fraud is spreading all over the world, resulting in huge financial losses. Though fraud prevention mechanisms such as CHIP&PIN are developed for credit card systems, these mechanisms do not prevent the most common fraud types such as fraudulent credit card usages over virtual POS (Point Of Sale) terminals or mail orders so called online credit card fraud.

TITLE: A SURVEY OF MACHINE-LEARNING AND NATURE-INSPIRED BASED CREDIT CARD FRAUD DETECTION TECHNIQUES

AUTHOR'S:

O. ADEWUMI

A. AKINYELU

Credit card is one of the popular modes of payment for electronic transactions in many developed and developing countries. Invention of credit cards has made online transactions seamless, easier, comfortable and convenient. However, it has also provided new fraud opportunities for criminals, and in turn, increased fraud rate. The global impact of credit card fraud is alarming, millions of US dollars have been lost by many companies and individuals.

TITLE: CREDIT CARD FRAUD DETECTION USING HIDDEN MARKOV MODEL

AUTHOR'S:

SRIVASTAVA

A. KUNDU

S. SURAL

A. MAJUMDAR

The most accepted payment mode is credit card for both online and offline in today's world, it provides cashless shopping at every shop in all countries. It will be the most convenient way to do online shopping, paying bills etc. Hence, risks of fraud transaction using credit card has also been increasing.

TITLE: REAL-TIME CREDIT CARD FRAUD DETECTION USING COMPUTATIONAL INTELLIGENCE

AUTHOR'S:

J. T. QUAH

M. SRIGANESH

Online banking and e-commerce have been experiencing rapid growth over the past few years and show tremendous promise of growth even in the future. This has made it easier for fraudsters to indulge in new and abstruse ways of committing credit card fraud over the Internet. This paper focuses on real-time fraud detection and presents a new and innovative approach in understanding spending patterns to decipher potential fraud cases.

2.2 CONCLUSIONS ONSURVEY

The performances of machine learning algorithm are effective approaches,In general, they rely on the performance efficiency,precision and relaibility characteristics. The machine learning approaches in the literature are under two main categories: unsupervised approaches supervised approaches. Depending on the datasets used, unsupervised approaches often suffer from high false positive rate and supervised approach can handle the data set negotiations . Therefore, the need of both, supervised and unsupervised approaches arises to overcome FraudDetectio.

3. SOFTWARE AND HARDWAREREQUIREMENTS

3.1 SOFTWAREREQUIREMENTS

- Operatingsystem : Windows 10.
- CodingLanguage : Python.
- Front-End : Python.
- Designing : Html,css,javascript.
- DataBase : MySQL

3.2 HARDWAREREQUIREMENTS

- System : Pentium IV,Intel
- HardDisk : 40 GB
- Mouse : OpticalMouse.
- Ram : 8GB

4. SOFTWARE DEVELOPMENTANALYASIS

4.1 OVERVIEW OFPROBLEM

Billions of dollars of loss are caused every year by the fraudulent credit card transactions. Fraud is old as humanity itself and can take an unlimited variety of different forms. One of the most crucial challenges in making the business is that neither the card nor the cardholder needs to be present when the purchase is being made. This makes it impossible for the merchant to verify whether the customer making a purchase is the authentic cardholder or not. The PwC global economic crime survey of 2017 suggests that approximately 48% of organizations experienced economic crime. Therefore, there is definitely an urge to solve the problem of credit card fraud detection. Moreover, the development of new technologies provides additional ways in which criminals may commit fraud. The use of credit cards is prevalent in modern day society and credit card fraud has been kept on growing in recent years. Hugh Financial losses has been fraudulent affects not only merchants and banks, but also individual person who are using the credits. Fraud may also affect the reputation and image of a merchant causing non-financial losses that, though difficult to quantify in the short term, may become visible in the long period. For example, if a cardholder is victim of fraud with a certain company, he may no longer trust their business and choose a contender.

4.2 DEFINE THEPROBLEM

To detect fraudulent activities we use a Machine learning model in online financial transactions. Analyzing fake transactions manually is impracticable due to vast amounts of data and its complexity. However, adequately given informative features, could make it is possible using Machine Learning. This hypothesis will be explored in the project. To classify fraudulent and legitimate credit card transaction by supervised learning Algorithm such as Random forest. To help us to get awareness about the fraudulent and without loss of any financially.

Packages

Which are being used for data exploration, pro processing and for using random forest algorithm are:

- NumPy: For simplearrays.
- Pandas: For reading thefile.

Random forest is a supervised machine learning algorithm based on ensemble learning. Ensemble learning is an algorithm where the predictions are derived by assembling or bagging different models or similar model multiple times. The random forest algorithm works in a similar way and uses multiple algorithm i.e. multiple decision trees, resulting in a forest of trees, hence the name "Random Forest". The random forest algorithm can be used for both regression and classification tasks.

Advantages of using random forest

- The random forest algorithm is not biased and depends on multiple trees where each tree istrained separately based on the data, therefore biasedness is reducedoverall.
- It's a very stable algorithm. Even if a new data point is introduced in the dataset it doesn't affect the overall algorithm rather affect the only a singletree.

4.3 MODULESOVERVIEW

In order to determine the tasks we use modules as source.In this modules the data form datasets is taken and this helps to train and test the data separately which produces frauds.

i. Datacollection

ii. Datapre-processing

iii. Featureextraction

iv. Evaluationmodel

4.4 DEFINE THEMODULES

1.DATACOLLECTION

Data used in this paper is a set of product reviews collected from credit card transactions records. This step is concerned with selecting the subset of all available data that you will be working with. ML problems start with data preferably, lots of data (examples or observations) for which you already know the target answer. Data for which you already know the target answer is called labelled data.

2. DATAPREPROCESSING

Pre-processing is the process of three important and common

steps as follows:

Formatting: It is the process of putting the data in a legitimate way that it would be suitable to work with. Format of the data files should be formatted according to the need. Most recommended format is .csv files.

Cleaning: Data cleaning is a very important procedure in the path of data science as it constitutes the major part of the work. It includes removing missing data and complexity with naming category and so on. For most of the data scientists,Data Cleaning continues of 80% of work.

Sampling: This is the technique of analyzing the subsets from whole large datasets, which could provide a better result and help in understanding the behaviour and pattern of data in an integrated way.You can take a smaller representative sample of the selected data that may be much faster for exploring and prototyping solutions before considering the wholedataset.

3. FEATURE EXTRACTION:

Feature extraction is the process of studying the behaviour and pattern of the analyzed data and draw the features for further testing and training. Finally, our models are trained using the Classifier algorithm. We use classify module on Natural Language Toolkit library on Python. We use the labelled dataset gathered. The rest of our labelled data will be used to evaluate the models. Some machine learning algorithms were used to classify pre-processed data. The chosen classifiers were Random forest. These algorithms are very popular in text classificationtasks.

4. EVALUATION MODEL:

Model Evaluation is an essential part of the model development process. It helps to find the best model that represents our data and how well the selected model will work in the future. Evaluating model performance with the data used for training is not acceptable in data science because it can effortlessly generate overoptimistically and over fitted models. To avoid overfitting, evaluation methods such as hold out and cross-validations areused to test to evaluate model performance. The result will be in the visualized form. Representation of classified data in the form of graphs. Accuracy is well-defined as the proportion of precise predictions for the test data. It can be calculated easily by mathematical calculation i.e.dividing the number of correct predictions by the number of total predictions.

4.5 MODULEFUNCTIONALITY

Data Collection

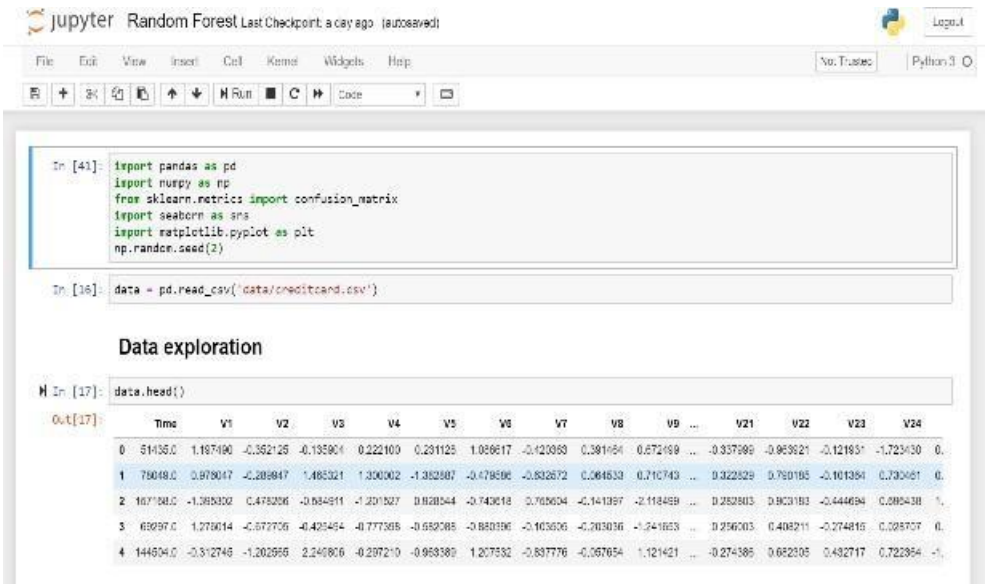


Fig4.1: Importing python packages for data exploration, preprocessing and for using random forest algorithm.

Data exploration is an approach similar to initial data analysis, whereby a data analyst uses visual exploration to understand what is in a dataset and the characteristics of the data, rather than through traditional data management systems.

Data Preprocessing

Pre-processing with python commands

Step - 1

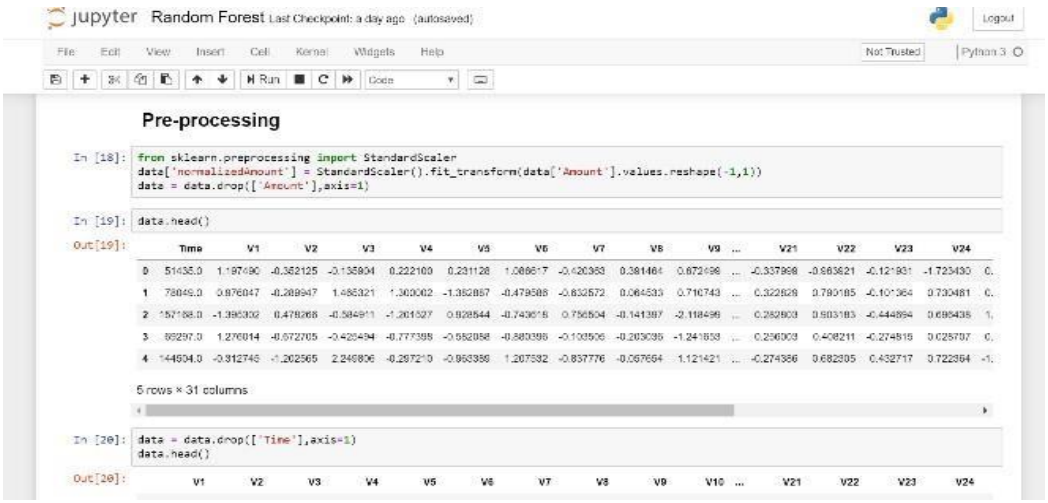


Fig 4.2: Pre processing

Step2:

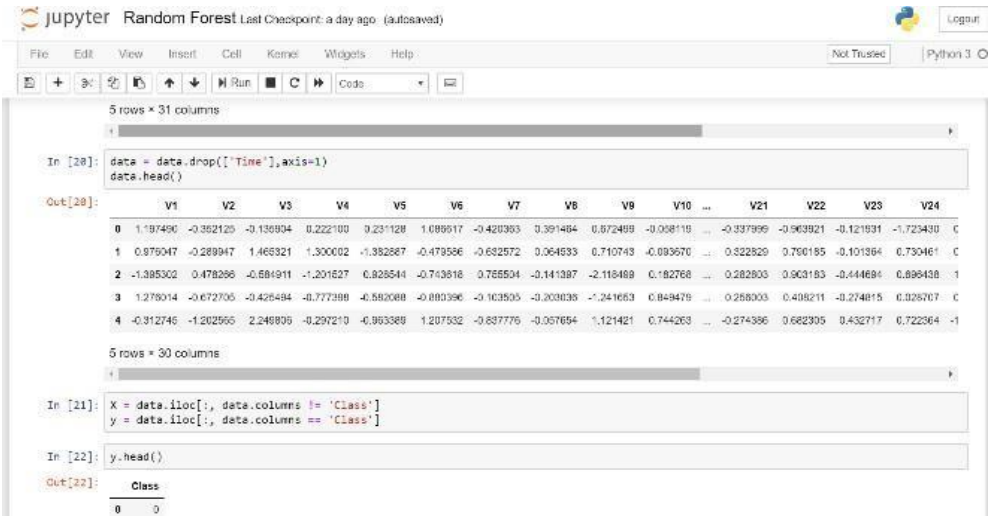


Fig4.3: Preprocessing Step 2

Step 3:

Acquired trained and testing dataset from the large dataset

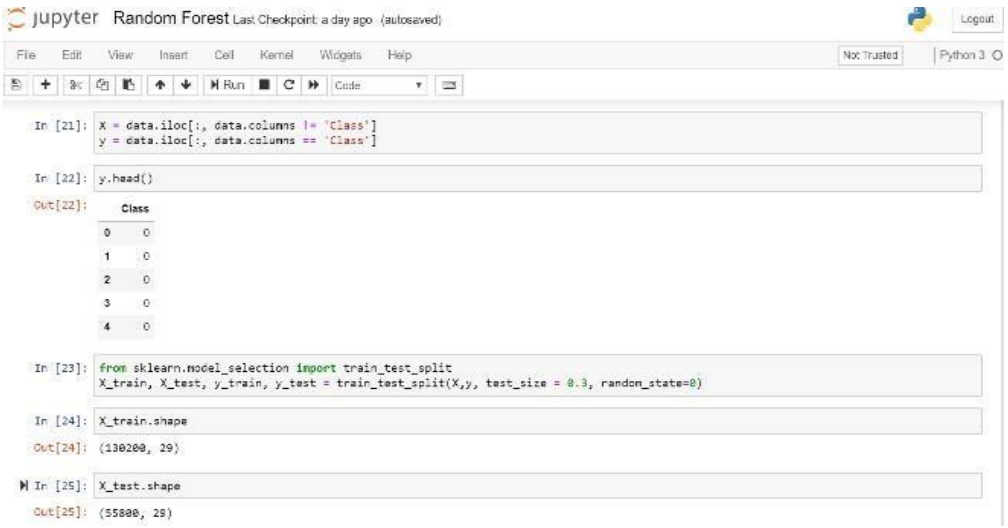


Fig 4.4: Training and testing data

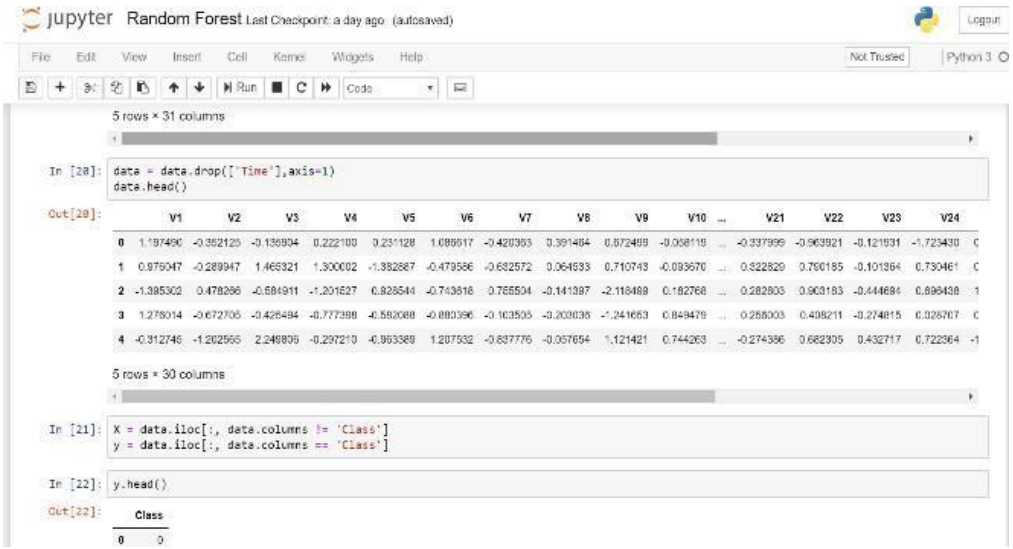


Fig 4.5: Process of training and testing data extraction

Training Set : A training dataset is a dataset of examples used during the learning process and is used to fit the parameters (e.g., weights) of, for example, a classifier.

• For classification tasks, a supervised learning algorithm looks at the training dataset to determine, or learn, the optimal combinations of variables that will generate a good predictivemodel.

• The goal is to produce a trained (fitted) model that generalizes well to new, unknowndata.

Test Set : A test dataset is a dataset that is independent of the training dataset, but that follows the same probability distribution as the training dataset. If a model fit to the training dataset also fits the test dataset well, minimal overfitting has taken place (see figure below). A better fitting of the training dataset as opposed to the test dataset usually points to overfitting.

• A test set is therefore a set of examples used only to assess the performance (i.e. generalization)of a fully specified classifier.

5. PROJECT SYSTEM DESIGN

5.1 DFDS IN CASE OF DATABASE PROJECTS

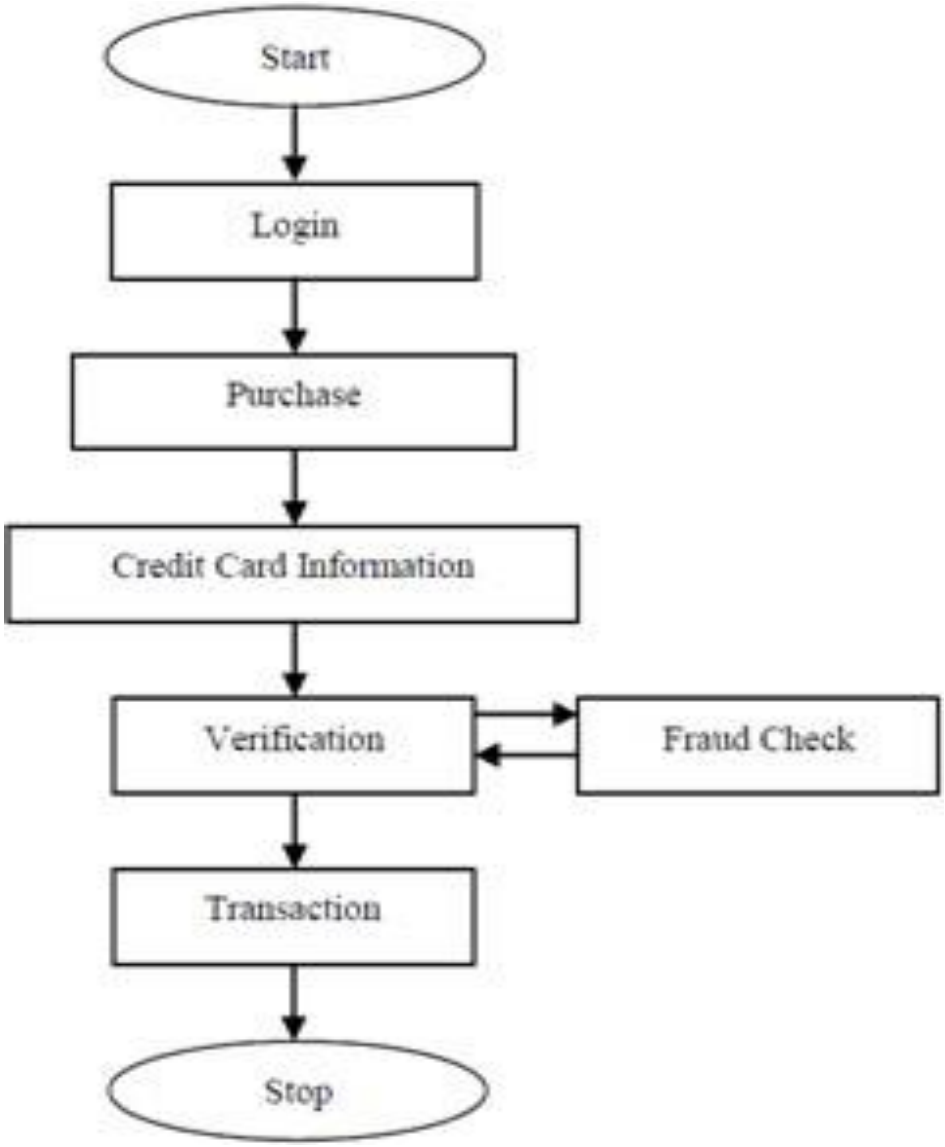


Fig 5.1: Data Flow Diagram For Fraud Detection

Data flow diagrams are used to graphically represent the flow of data in a business information system. DFD describes the processes that are involved in a system to transfer data from the input to the file storage and reports generation.

Data flow diagrams can be divided into logical and physical. The logical data flow diagram describes flow of data through a system to perform certain functionality of a business. The physical data flow diagram describes the implementation of the logical data flow.

In this DFD Diagram it shows the process of how the user will login with the account details and the purchases made by him. And this information will be stored in the database which helps to check whether any fraudulent activities are imposed on a card.The Database shows the limit of the particular card that has set default. If the limit exceeds it predicts fraud and this can be detected using the Random Forest Algorithm.This Algorithm helps to detect and prevent from fraudulent activities.

5.2 E-RDIAGRAMS

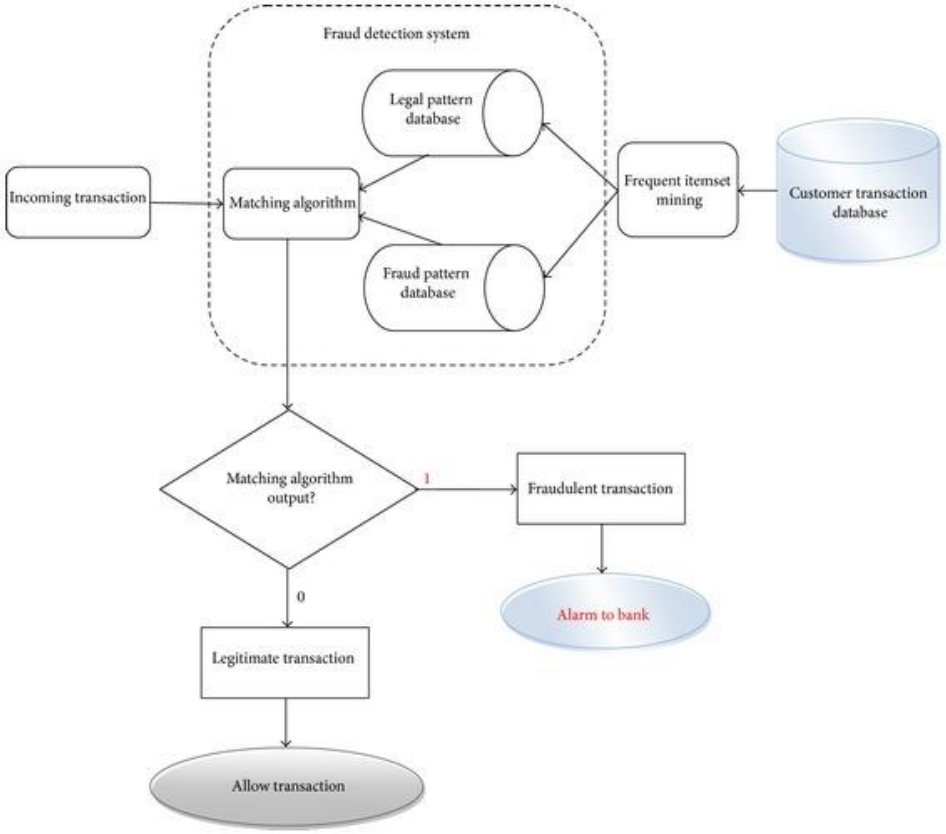


Fig 5.2 ER DIAGRAM for credit card system

An Entity Relationship (ER) Diagram is a type of flowchart that illustrates how “entities” such as people, objects or concepts relate to each other within a system. ER Diagrams are most often used to design or debug relational databases in the fields of software engineering, business information systems, education and research.

ER diagrams are related to data structure diagrams (DSDs), which focus on the relationships of elements within entities instead of relationships between entities themselves. ER diagrams also are often used in conjunction with data flow diagrams (DFDs), which map out the flow of information for processes or systems.

5.3 UMLDIAGRAMS

UML stands for Unified Modeling Language. UML is a standardized general-purpose modeling language in the field of object-oriented software engineering. The standard is managed, and was created by, the Object Management Group.

The goal is for UML to become a common language for creating models of object oriented computer software. In its current form UML is comprised of two major components: a Meta-model and a notation. In the future, some form of method or process may also be added to; or associated with, UML.

The Unified Modeling Language is a standard language for specifying, Visualization, Constructing and documenting the artifacts of software system, as well as for business modeling and other non-software systems.

The UML represents a collection of best engineering practices that have proven successful in the modeling of large and complexsystems.

The UML is a very important part of developing objects oriented software and the software development process. The UML uses mostly graphical notations to express the design of software projects.

GOALS :

The Primary goals in the design of the UML are as follows:

1. Provide users a ready-to-use, expressive visual modeling Language so that they can develop and exchange meaningfulmodels.
2. Provide extendibility and specialization mechanisms to extend the coreconcepts.
3. Be independent of particular programming languages and developmentprocess.
4. Provide a formal basis for understanding the modelinglanguage.
5. Encourage the growth of OO toolsmarket.
6. Support higher level development concepts such as collaborations, frameworks, patterns and components.
7. Integrate bestpractices.

USECASE DIAGRAM

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.

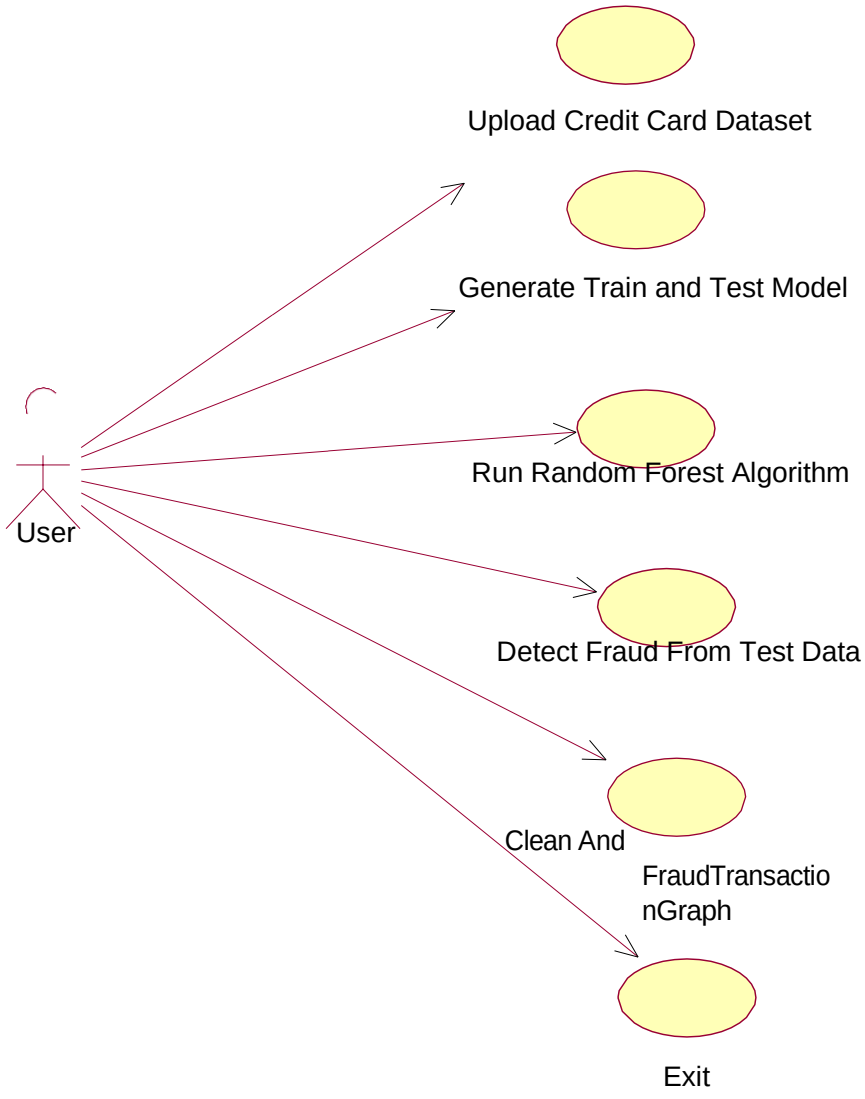


Fig 5.3: Usecase Diagram

In this usecase diagram as it shows user will first login to his account and the admin will be monitoring all the login credentials required for the user. Then,the user make online purchase and they does the transaction,if the transaction get delayed it will generate in valid transaction and the user need to do the transaction again and the admin observes if any fraudulent activities are detecting by running the random forest algorithm.If it generates any fraud it will find the fraud with highest accuracy rate or else it accepts the transaction and exist thepage.

CLASS DIAGRAM

In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.

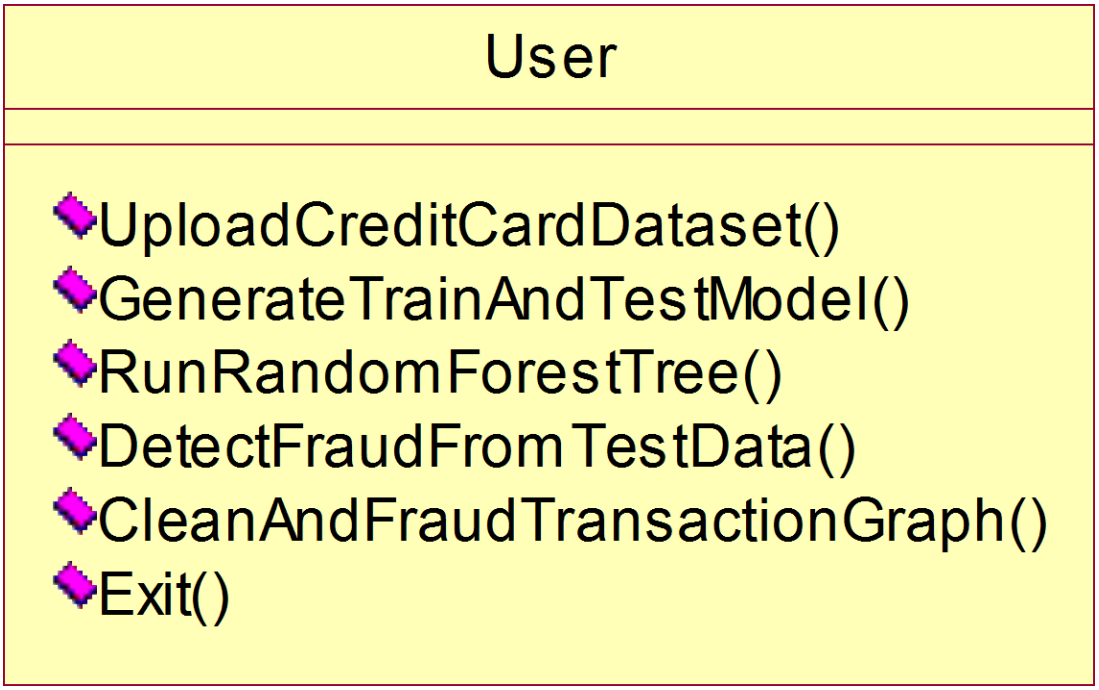


Fig 5.4: Class Diagram

In this class diagram it creates subclass for each instance of the process such as user class, database class,admin class. In each of this separate class it creates required objects for particular field.For suppose in order to login to the portal the user needs username and password which can be stored in string methods so the user class creates object for the username with string method.In admin class the admin needs to access the online purchases,transactions so admin Create these classes.Whenever they require any step they just use the created class to view the information.

SEQUENCE DIAGRAM

A sequence diagram in Unified Modeling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.

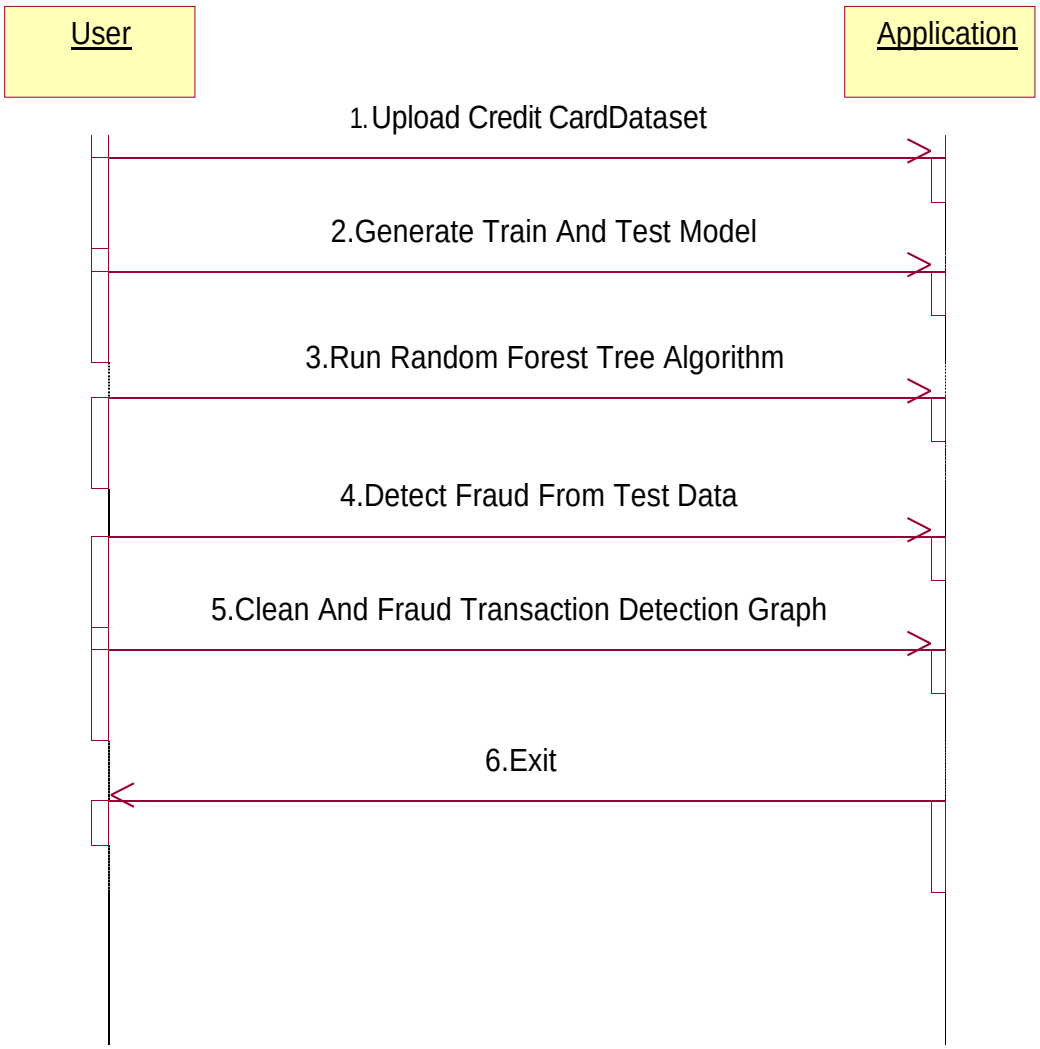


Fig 5.5: Sequence Diagram

In this sequence diagram the user and application sequences are created and from them the information will be passed to each other such as the user need to upload the dataset and after uploading the data will be trained and tested using the random forest classifier and then the fraud will be detected from the test modules if any fraud detected it cleans the transaction data or else the application will send the exit message to theuser.

COLLABORATION DIAGRAM

Collaboration diagrams (**known as Communication Diagram in UML 2.x**) are used to show how objects interact to perform the behaviour of a particular use case, or a part of a use case. Along with sequence diagrams, collaboration are used by designers to define and clarify the roles of the objects that perform a particular flow of events of a use case. They are the primary source of information used to determining class responsibilities and interfaces.

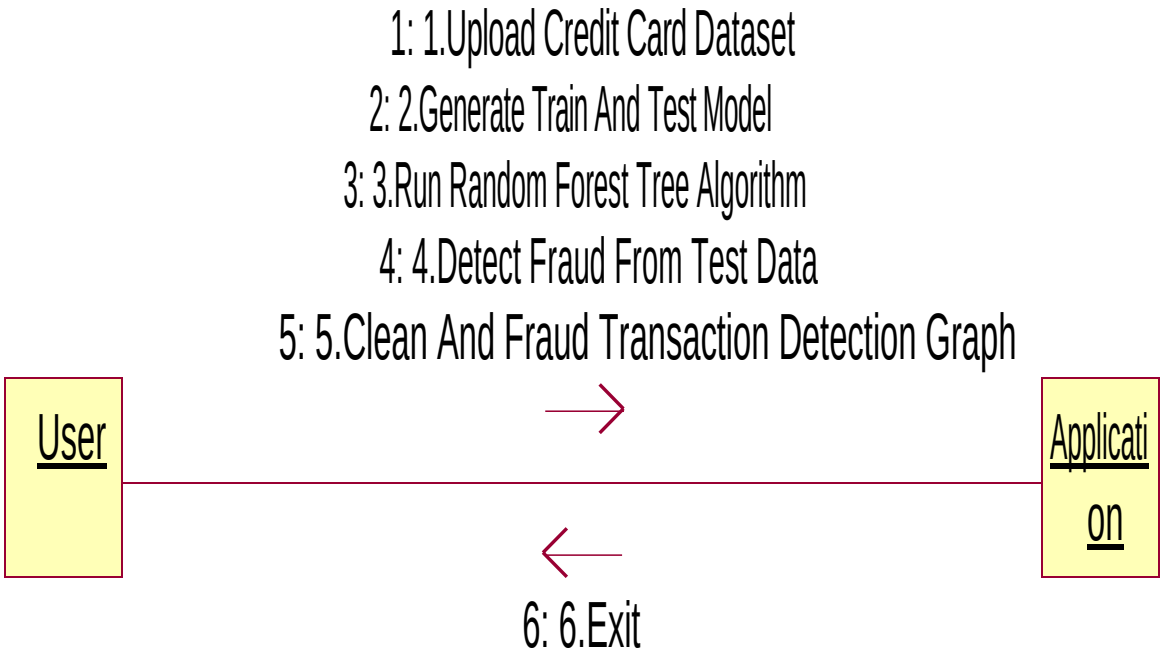


Fig 5.6: Collaboration Diagram

In this collaboration diagram each step is done Continously like uploading dataset,generating train and test modules,running random forest classifier and detection of fraud after the analysis of whole data process the application sends the exit message to the user.

6. PROJECT CODING

6.1 CODE TEMPLATES

```
# Import required modules

# Declare global variables

def uploadDataset():

 # This is the first module of our project where we need to upload a folder that contains different images containing different hand postures.

def PrintStats()

 #In our project we first try to print the statistical values and make the components separate from the confusion matrix . This helps the training and testing modules to generate the F1 Score,Cohen kappa efficiency model and produce effective results/outputs.

def RunModel():

 #In this process it runs the train and test model created from the dataset by sending them to the random forest classifier.
This gives the prediction of the output based on the data values that bestfits to run the algorithm.
It finally generates predictions through the matrix.

df=pd.read()

#In this it imports the datasets and stores all the values if f1 score, Cohen kappa and the predictions from the Matrix.

#importRandomForestClassifier

#After importing all the values then we apply the random forest classifier algorithm to the values generated which produces the efficiency of the model and detects the fraud from the values . It generates the sample models for the frauds and the correct ones.

def close():

# Exit from the project.

main.close()
```

6.2 OUTLINE FOR VARIOUS FILES

We used Python programming to implement our project. A single python file is used to implement our code. This file consists of various modules that we have used. Our project modules are - Uploading Dataset,RandomForestClassifier,ConfusionMatrix,Decision Trees. We also used various python modules like pandas,numpy,sklearn.

6.3 METHODS INPUT AND OUTPUTPARAMETERS

In our project code, we implemented four different methods. They are:

1. uploadDataset():
2. RunModel():
3. RandomForestClassifier():
4. Close()

Our first method uploadDataset() doesn't take any input parameters but after successful execution, it displays "Credit Card DataSets". Second method runModel() it takes train and test parameters to divide the dataset values and produces the efficiency and coefficient values. Third Method RandomForestClassifier() classifies the fraud and the true values based on the data set provided. close() don't have any parameters but upon clicking this button, it will close the project window.

INPUT DESIGN

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the followingthings:

- What data should be given asinput?
- How the data should be arranged orcoded?
- The dialog to guide the operating personnel in providinginput.
- Methods for preparing input validations and steps to follow when erroroccur.

OBJECTIVES

1. Input Design is the process of converting a user-oriented description of the input into a computer-based system. This design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerizedsystem.

2. It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The goal of designing input is to make data entry easier and to be free from errors. The data entry screen is designed in such a way that all the data manipulates can be performed. It also provides record viewingfacilities.

3. When the data is entered it will check for its validity. Data can be entered with the help of screens. Appropriate messages are provided as when needed so that the user will not be in maize of instant. Thus the objective of input design is to create an input layout that is easy tofollow.

OUTPUT DESIGN

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making.

1. Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is designed so that people will find the system can use easily and effectively. When analysis design computer output, they should Identify the specific output that is needed to meet therequirements.

2. Select methods for presentinginformation.

3. Create document, report, or other formats that contain information produced by the system.

The output form of an information system should accomplish one or more of thefollowing objectives.

- Convey information about past activities, current status or projections of the
- Future.
- Signal important events, opportunities, problems, or warnings.
- Trigger anaction.
- Confirm anaction.

7. PROJECT TESTING

7.1 VARIOUS TEST CASES

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, subassemblies, assemblies and/or a finished product. It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

TYPES OF TESTS

Unit testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

Integration testing

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination ofcomponents.

Functional test

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items:

Valid Input : identified classes of valid input must be accepted.

Invalid Input : identified classes of invalid input must be rejected.

Functions : identified functions must be exercised.

Output : identified classes of application outputs must be exercised.

Systems/Procedures : interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

System Test

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

Unit Testing

Unit testing is usually conducted as part of a combined code and unit test phase of the software lifecycle, although it is not uncommon for coding and unit testing to be conducted as two distinct phases.

Test strategy and approach

Field testing will be performed manually and functional tests will be written in detail.

Test objectives

- All field entries must workproperly.
- Pages must be activated from the identifiedlink.
- The entry screen, messages and responses must not bedelayed.

Features to be tested

- Verify that the entries are of the correctformat.
- No duplicate entries should beallowed.
- All links should take the user to the correctpage.

Integration Testing

Software integration testing is the incremental integration testing of two or more integrated software components on a single platform to produce failures caused by interface defects. The task of the integration test is to check that components or software applications, e.g. components in a software system or – one step up – software applications at the company level – interact without error.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

Acceptance Testing

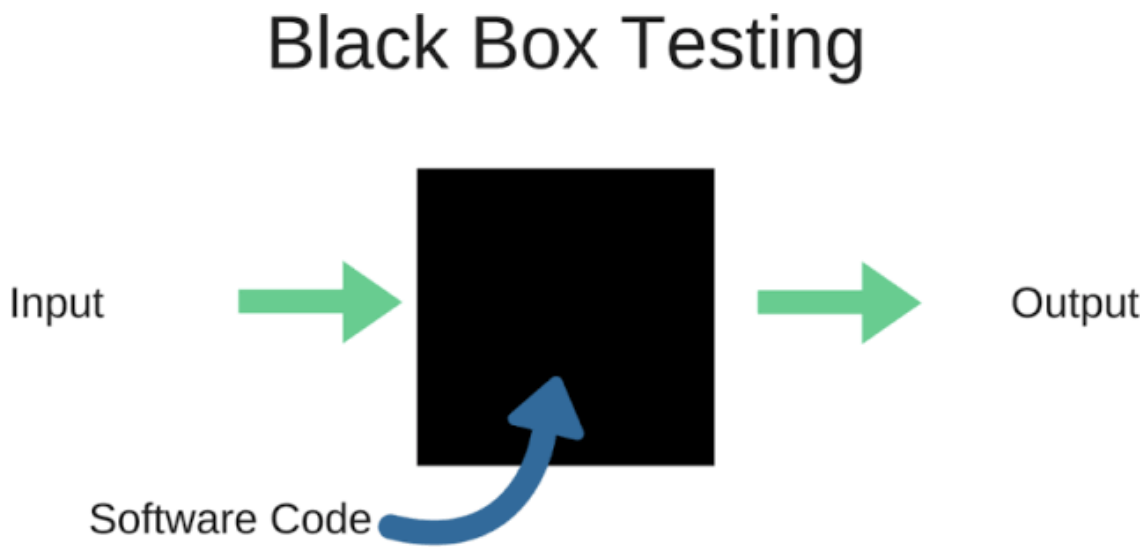
User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

7.2 BLACK BOXTESTING

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box .you cannot “see” into it. The test provides inputs and responds to outputs without considering how the software works.

The below Black-Box can be any software system you want to test. For Example, an operating system like Windows, a website like Google, a database like Oracle or even your own custom application. UnderBlack Box Testing, you can test these applications by just focusing on the inputs and outputs without knowing their internal code implementation.



Various approaches to black-box testing

There are a set of approaches for black-box testing.

Manual UI Testing: In this approach, a tester checks the system as a user. Check and verify the user data, error messages.

Automated UI Testing: In this approach, user interaction with the system is recorded to find errors and glitches. Testers can set record demand as per schedule.

Documentation Testing: In this approach, a tester purely checks the input and output of the software. Testers consider what system should perform rather than how. It is a manual approach to testing.

The tester doesn't need any technical knowledge to test the system. It is essential to understand the user's perspective.

Testing is performed after development, and both the activities are independent of each other. It works for a more extensive coverage which is usually missed out by testers as they fail to see the bigger picture of the software.

Test cases can be generated before development and right after specification. Black box testing methodology is close to agile.

7.3 WHITE BOXTESTING

The box testing approach of software testing consists of black box testing and white box testing. We are discussing here white box testing which also known as glass box is testing, structural testing, clear box testing, open box testing and transparent box testing. It tests internal coding and infrastructure of a software focus on checking of predefined inputs against expected and desired outputs. It is based on inner workings of an application and revolves around internal structure testing. In this type of testing programming skills are required to design test cases. The primary goal of white box testing is to focus on the flow of inputs and outputs through the software and strengthening the security of the software.

The term 'white box' is used because of the internal perspective of the system. The clear box or white box or transparent box name denote the ability to see through the software's outer shell into its inner workings.

Developers do white box testing. In this, the developer will test every line of the code of the program. The developers perform the White-box testing and then send the application or the software to the testing team, where they will perform the black box testing and verify the application along with the requirements and identify the bugs and sends it to the developer.

The developer fixes the bugs and does one round of white box testing and sends it to the testing team. Here, fixing the bugs implies that the bug is deleted, and the particular feature is working fine on the application.

Here, the test engineers will not include in fixing the defects for the followingreasons:

- Fixing the bug might interrupt the other features. Therefore, the test engineer should always find the bugs, and developers should still be doing the bugfixes.

- If the test engineers spend most of the time fixing the defects, then they may be unable to find the other bugs in theapplication.

8. OUTPUT SCREENS

8.1 USER INTERFACES

To run project double click on ‘run.bat’ file to get below screen

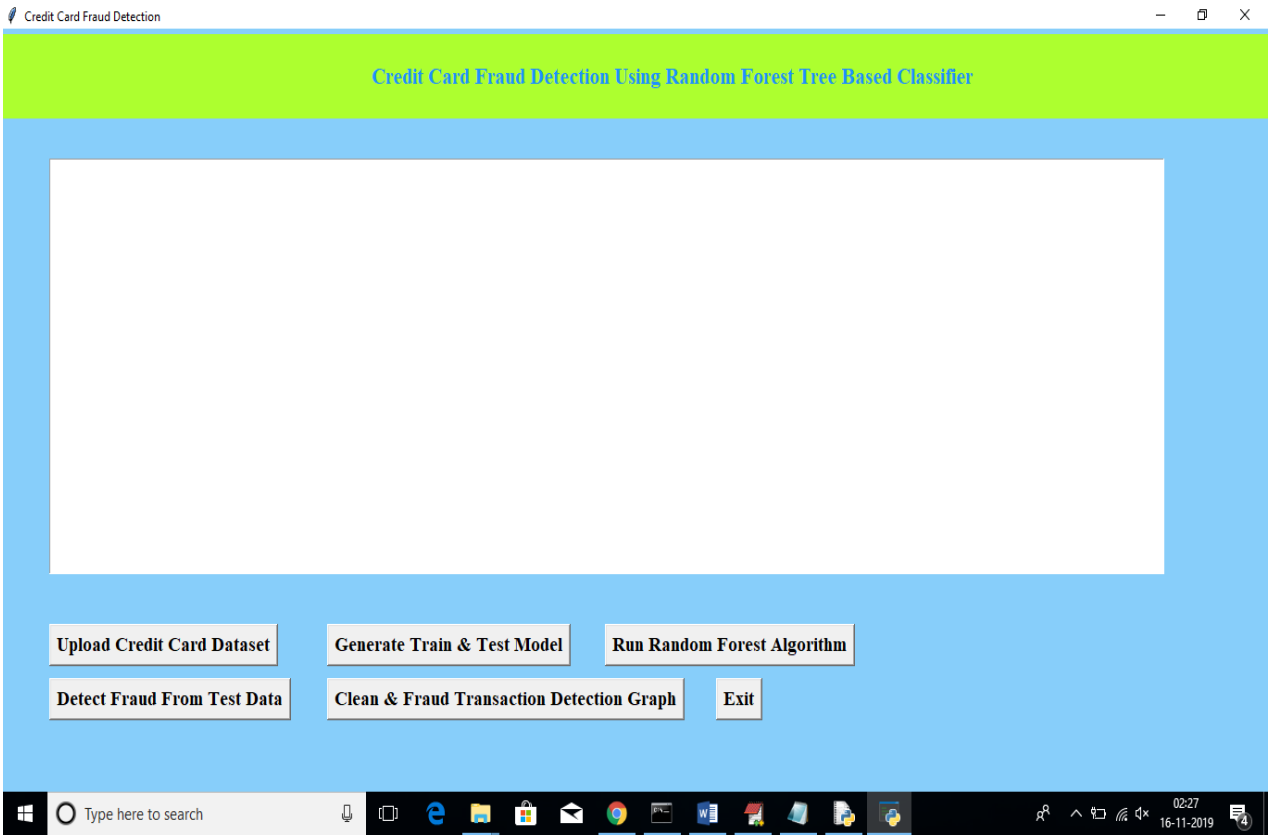


Fig 8.1 The figure shows user interface.

In the above screen we get all the buttons required for the execution such as Uploading Dataset,Training and Test, and Random forest algorithm,fraud detection.

8.2 OUTPUT SCREENS

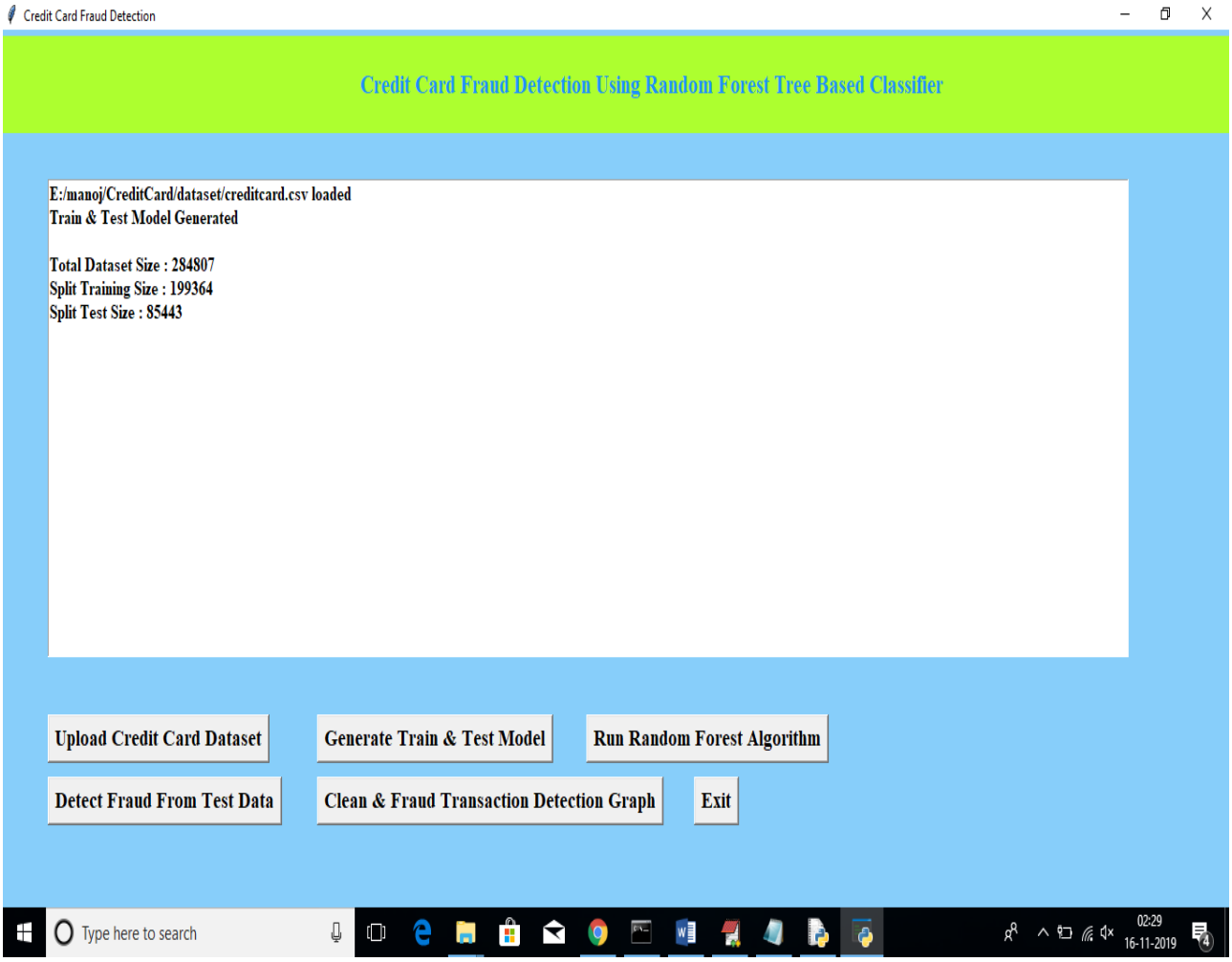


Fig 8.2: Generate Train and Test data

In above Screen,After generating model we can see total records available in dataset and then application using how many records for training and how many for testing. Now click on “Run Random Forest Algorithm” button to generate Random Forest model on train and test data.

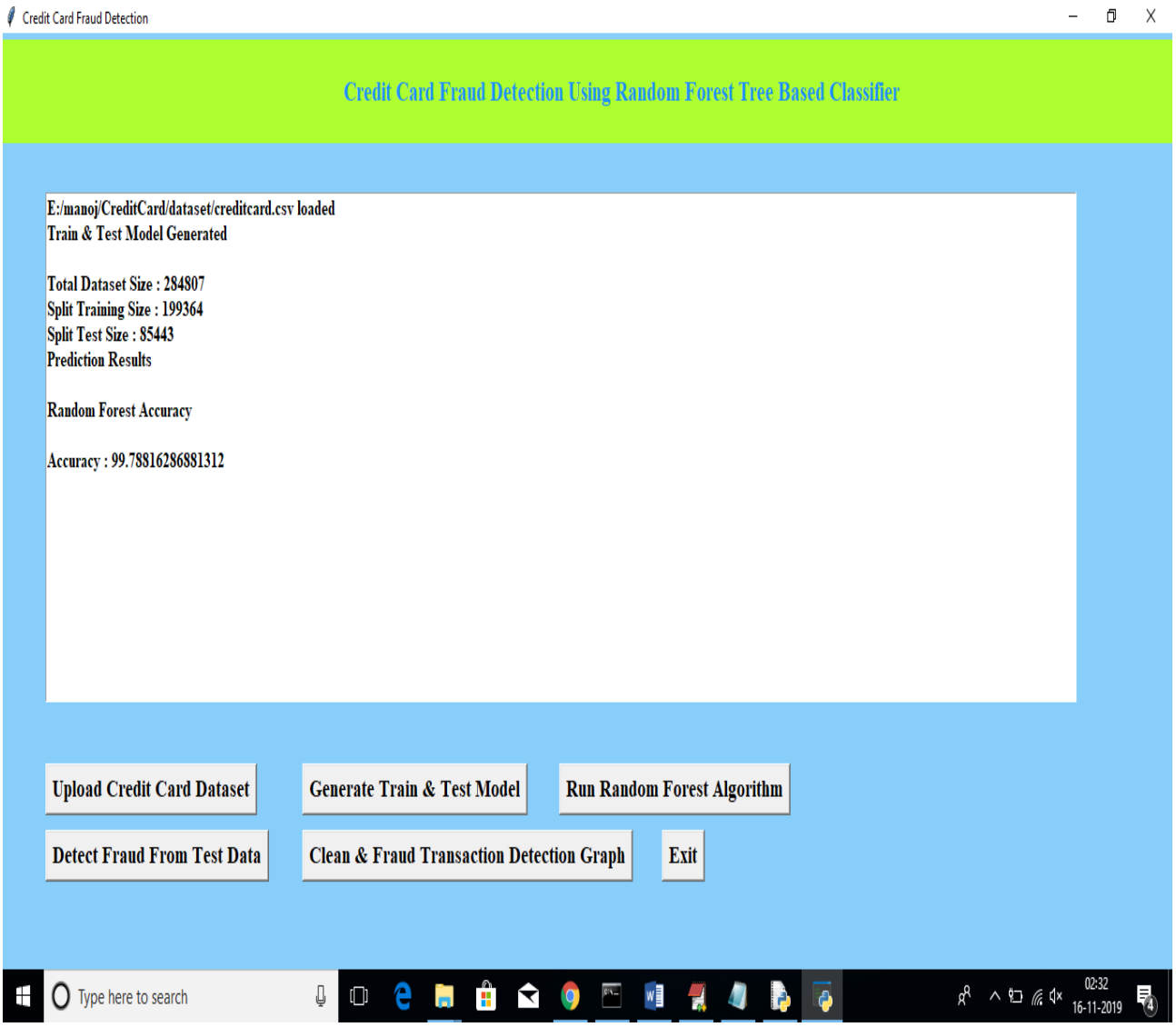


Fig 8.3: Run the Random Forest Algorithm

After Generating Random forest we get as In above screen Random Forest generate 99.78% percent accuracy while building model on train and test data. Now click on 'Detect Fraud From Test Data' button to upload test data and to predict whether test data contains normal or fraud transaction.

9. EXPERIMENTAL RESULTS

MainPage

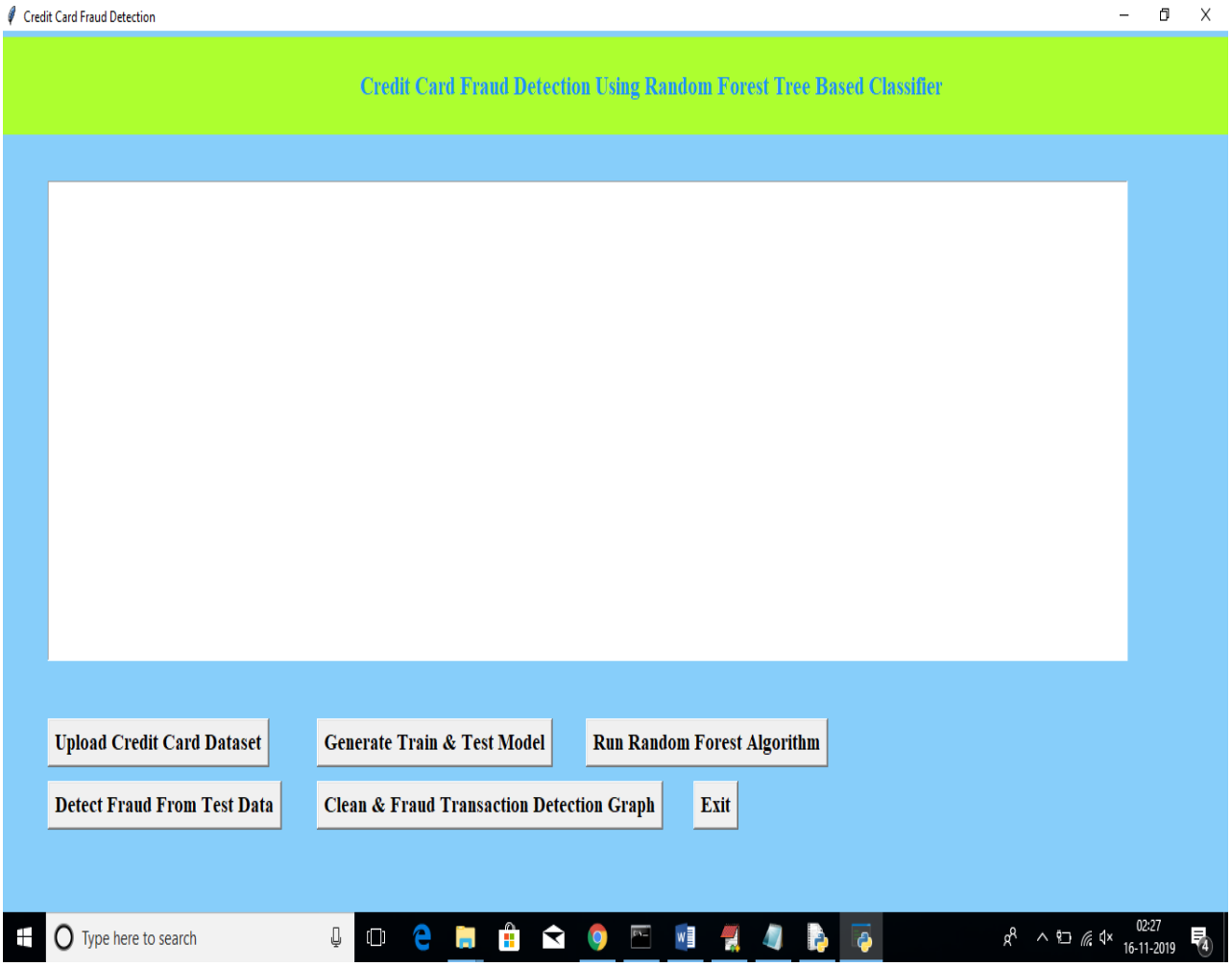


Fig 9.1: Main Page

To run project double click on ‘run.bat’ file to get below screen

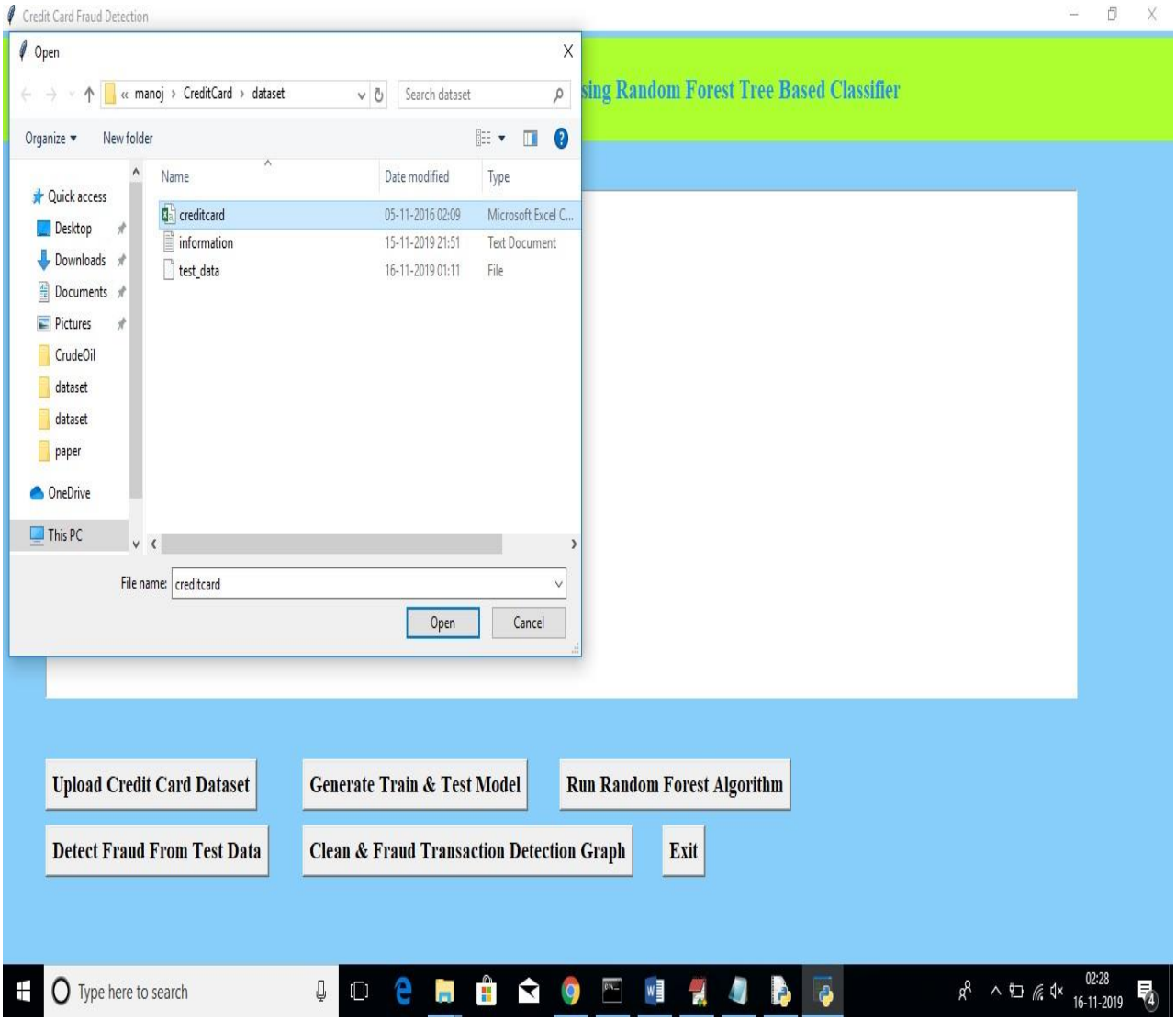


Fig 9.2: Uploading Dataset

In above screen click on ‘Upload Credit Card Dataset’ button to upload dataset.

After uploading dataset will get below screen

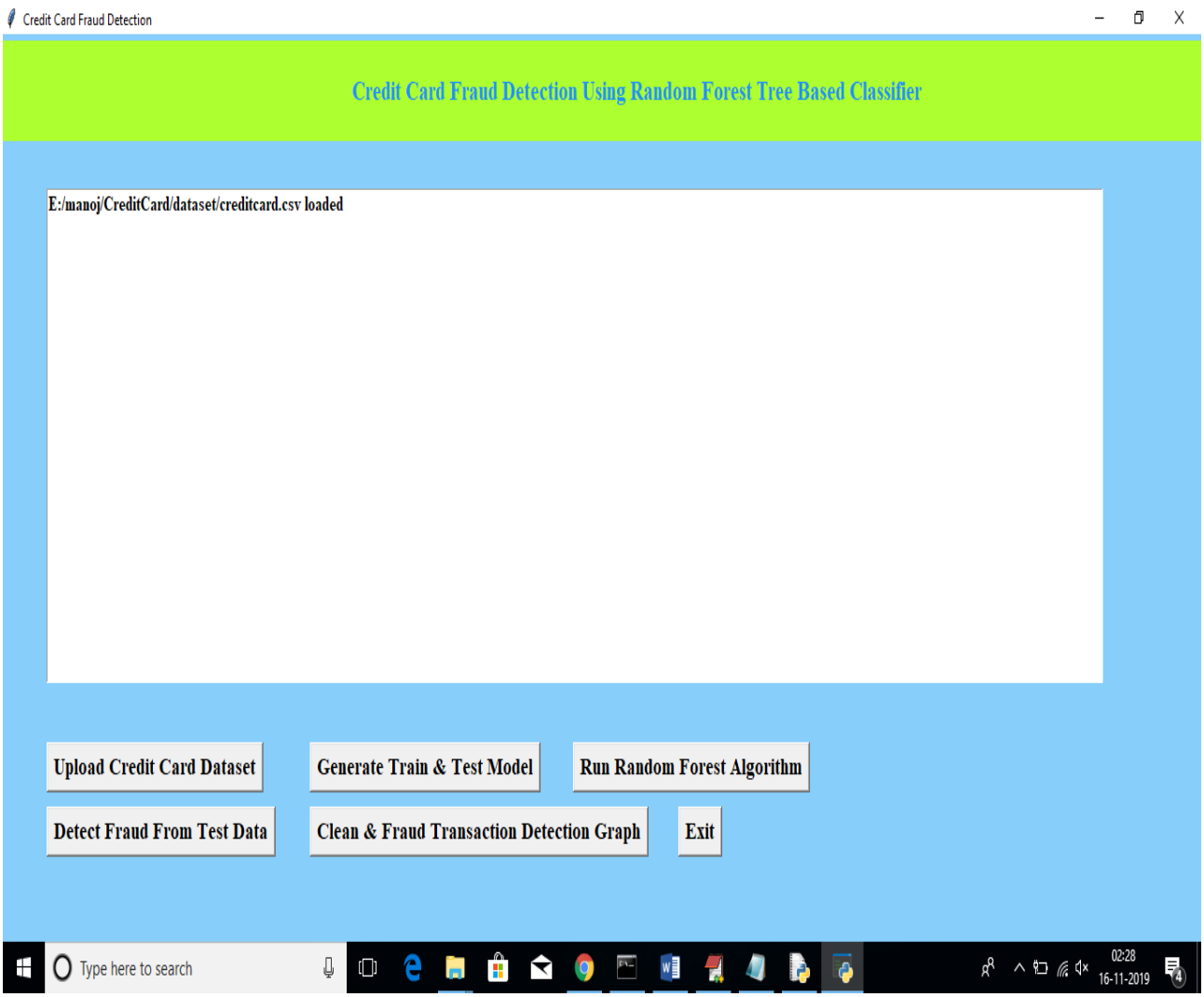


Fig 9.3: Train and Test models

Now click on ‘Generate Train & Test Model’ to generate training model for Random Forest Classifier.

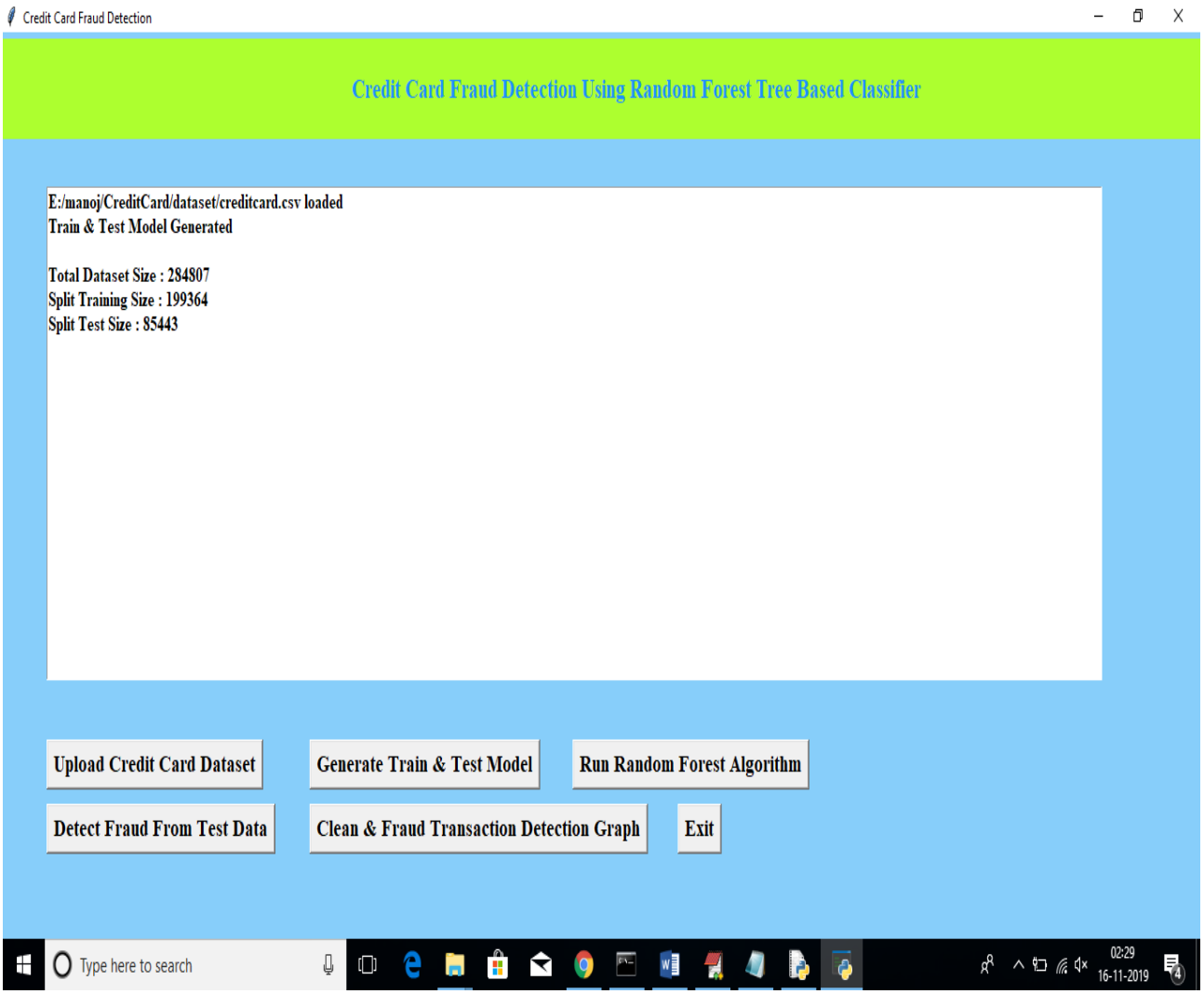


Fig 9.4: Generating Train and Test models

In above screen after generating model we can see total records available in dataset and then application using how many records for training and how many for testing. Now click on “Run Random Forest Algorithm’ button to generate Random Forest model on train and test data.

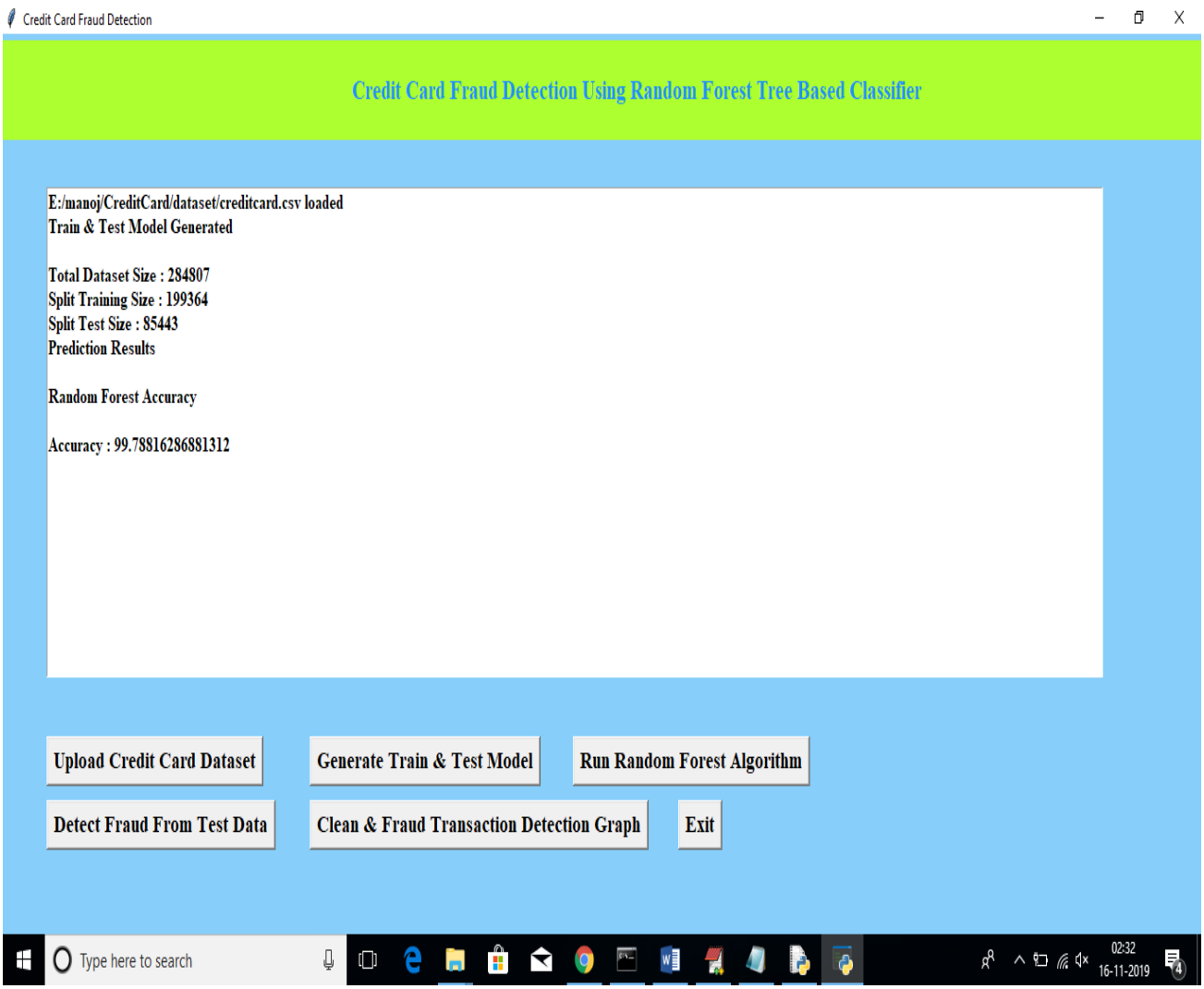


Fig 9.5: Run the Random Forest Algorithm

In above screen we can see Random Forest generate 99.78% percent accuracy while building model on train and test data. Now click on 'Detect Fraud From Test Data' button to upload test data and to predict whether test data contains normal or fraud transaction.

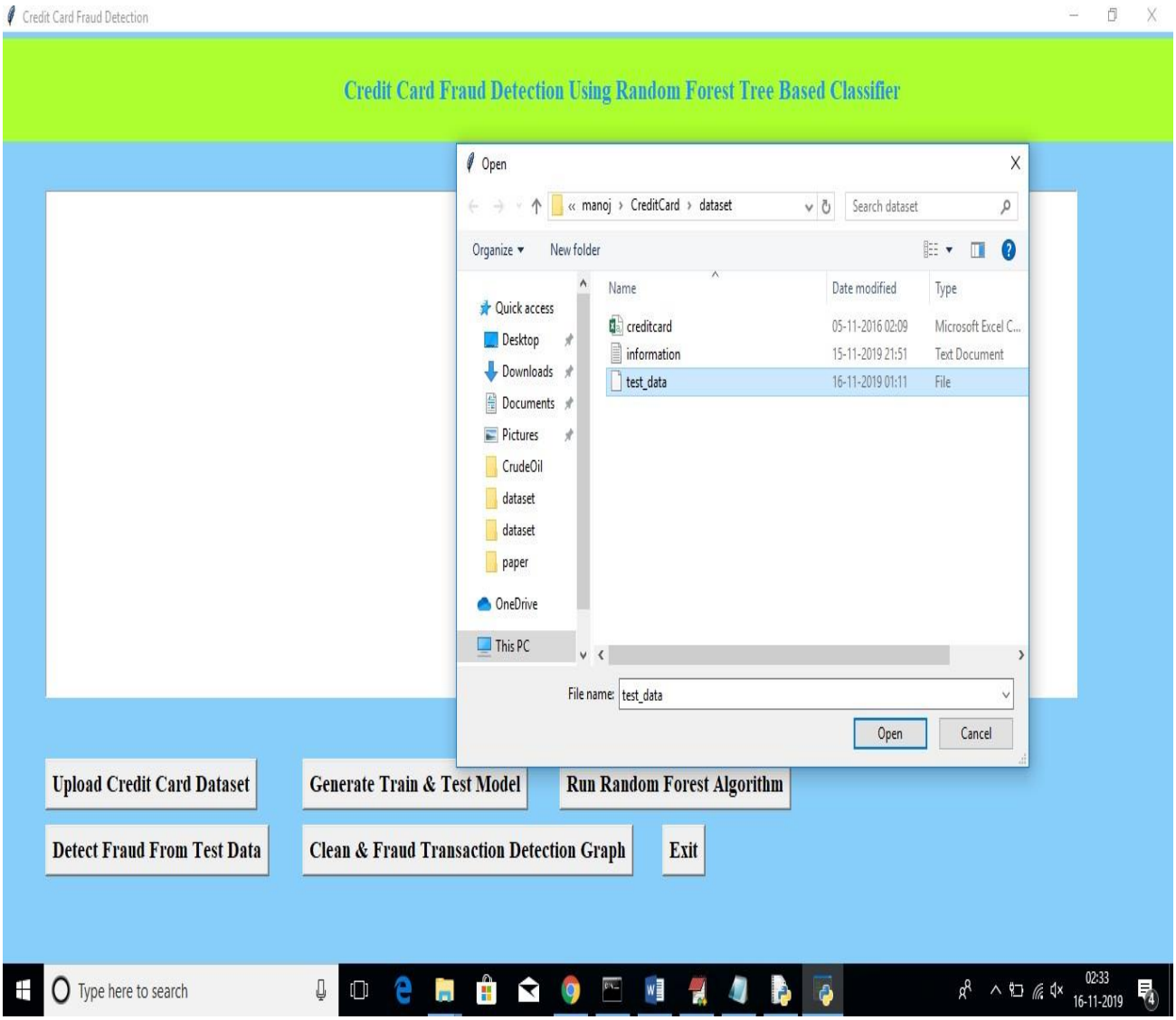


Fig 9.6: Uploading Test Dataset

In above screen I am uploading test dataset and after uploading test data will get below prediction details.

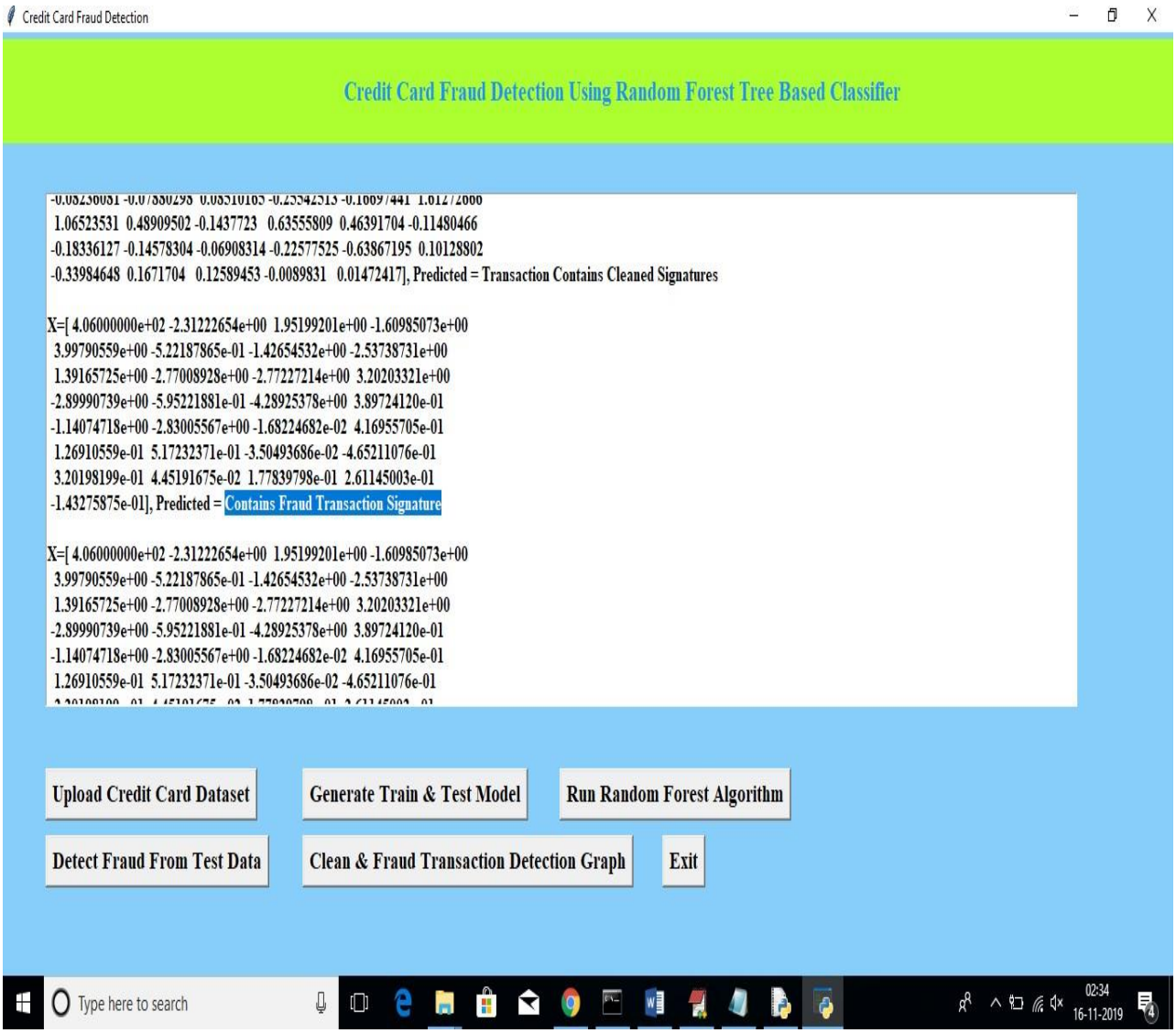


Fig 9.7: Fraud Signatures

In above screen beside each test data application will display output as whether transaction contains cleaned or fraud signatures.

9. CONCLUSION AND FUTURE ENHANCEMENT

Hence, we have acquired the result of an accurate value of credit card fraud detection i.e. 0.9994802867383512 (99.93%) using a random forest algorithm with new enhancements. In comparison to existing modules, this proposed module is applicable for the larger dataset and provides more accurate results. The Random forest algorithm will provide better performance with manytraining data, but speed during testing and application will stillsuffer. Usage of more pre-processing techniques would also assist. Our future work will try to represent this into a software application and provide a solution for credit card fraud using the new technologies like Machine Learning, Artificial Intelligence and Deep Learning.The Random forest algorithm will perform better with a larger number of training data, but speed during testing and application will suffer. Application of more pre-processing techniques would also help. The SVM algorithm still suffers from the imbalanced dataset problem and requires more preprocessing to give better results at the results shown by SVM is great but it could have been better if more preprocessing have been done on the data.

In the near future, this paper will be used as a reference by some banks or individuals to implement fraud detection system in the financial sector. Benefits of implementing such detection system will lessen the phone and SMS costs shouldered by the banks; instead of sending SMS transaction notifications to all customers, message will be sent to those customers with detected anomalous transaction.

REFERENCES

[1] KosemaniTemitayo Hafiz, Dr. Shaun Aghili, Dr.PavolZavarsky Research on "The Use of Predictive Analytics Technology" to Detect Credit Card Fraud in Canada2021.

[2] Amlan Kundu, SuvasinPanigrahi, Shamik Sural, Senior Member, IEEE, and Arun K. Majumdar "BLAST-SSAHA Hybridization" for Credit Card Fraud Detection in2019.

[3] W. Yu and N. Wang, "Research on Credit Card Fraud Detection Model Based on Distance Sum", 2009 International Joint Conference on Artificial Intelligence in2019.

[4] Vijayshree B. Nipane, Poonam S. Kalinge, DipaliVidhate, Kunal War, Bhagyashree P. Deshpande, "Fraudulent Detection in Credit Card System Using SVM & Decision Tree" in2020.

[5] Dahee Choi and Kyungho Le "Machine Learning Based Approach" to Financial Fraud Detection Process in Mobile Payment System in2021.

[6] Sitaram patel, Sunita Goud research on "Supervised Machine (SVM) Learning"for Credit CardFraud Detection in2018.

[7] Y. Sahin and E. Duman research on "Detecting Credit Card Fraud by Decision Trees and Support Vector Machines" in2019.

[8] Rong-Chang Chen and Taishi Chen research on "A new binary support vector system" for increasing detection rate of credit card fraud in2018.

[9] Mohammad Behdad and Tim French research on"Nature-Inspired Techniques" in the Context of Fraud Detection in2019.

[10] Shalini Gupta and Rahul, Johari research on "A New Framework for Credit Card Transactions Involving Mutual Authentication between Cardholder and Merchant" in2017.

[11] Olivier Caelen,TinaEliassi-Rad researchon "APATE" : A Novel Approach for Automated Credit Card Transaction Fraud Detection using Network-Based Extensions in 2016.

[12] Leo Breiman,Charles J. Stone research on "Classification and Regression Trees" in Dec2015.

[13] Leila Seyedhossein,Mahmoud Reza Hasheresearch on "Mining information from credit card time series for timelier fraud detection"in2016.

[14] Tong Seng Quah,Sriganesh Srihari research on "Real-time credit card fraud detection using computational intelligence"in2018.

[15] Richard J. Bolton,David Hand research on "Statistical Fraud Detection" A Review,Stat.Sci.,2015.

PUBLICATIONS

JOURNAL (UGC approved Journal)

CONFERENCE (International Conference on “Credit Card Fraud Detection using Random Forest” [ICICCI-20-02]).

PAPER ID : ICICCI-21-0060

TOPIC : CREDIT CARD FRAUD DETECTION USING RANDOM FOREST AND CART ALGORITHM

STUDENT'SPROFILE

Madhuri Adarapuis currently pursuing her Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. She completed her intermediate from AbhyassJunior College and 10th class from Sri Chaitanya Techno School.Her technical skills include C,C++,Python and MySQL. She also has a basic understanding of Java.She is also a student of Smart Interviews. Her participations include: National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing"which was conducted from 10th to 12th December 2020,"MHRDInnovation cell On Startup policy",Women online workshop on "Women in Cyber Security and Privacy in 2020" which was conducted from 6th to 10th July 2020.Her areas of interest areJavaScirpt, Python ,Machine Learning.She completed few certification courses from online platforms like Coursera, CursaApp and SoloLearn.

Mounika Gandeis currently pursuing her Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. She completed her intermediate from AbhyassJunior College and 10th class from Sri Chaitanya Techno School. Her technical skills include C,Java,Python and MySQL. She also has a basic understanding of C++.She is also a student of Smart Interviews. Her participations include: National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020,"MHRD Innovation cell On Startup policy",Women online workshop on "Women in Cyber Security and Privacy in 2020" which was conducted from 6th to 10th July 2020.Her areas of interest are Python, IoT, ArtificalIntelligence.She completed few certification courses from online platforms like Coursera, CursaApp and SoloLearn.

Pranavikattais currently pursuing her Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. She completed her intermediate from Sri Chaitanya Junior College and 10th class from krishnaveni High School. She is one of the members of Coders Club in our college. Her responsibilities in that group include mentoring and motivating students to take coding as a serious hobby. Her technical skills include C, Python and Java. She also has a basic understanding of C++. She took part in Employability Skill development Program conducted by Zensar. She is also a student of Smart Interviews. Her participations include: National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing"which was conducted from 10th to 12th December 2020, "Know More - Teach More ", the Global Webinar on Cloud & Big Data – Changing the way we work which was conducted Global Education & Careers Forum(GECF) on 12th August 2020, Women online workshop on "Women in Cyber Security and Privacy in 2020" which was conducted from 6thto 10th July 2020, "Know More - Teach More ", the Global Webinar on Cyber Threats and Defense Techniques conducted by GECF on 22nd July 2020, "One Day Webinar on Internet of Things and Its Applications" conducted by Anand Institute of Higher Technology on 21st May 2020 and IIC Online Sessions conducted by Institution's Innovation Council (IIC) of MHRD's Innovation Cell, New Delhi to promote Innovation, IPR, Entrepreneurship, and Start-ups among HEIs from 28thApril to 22nd May 2020. Her areas of interest are Python, Artificial Intelligence, Machine Learning and Deep Learning. She completed few certification courses from online platforms like Coursera, CursaAppandSoloLearn.

Satyanarayana is currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. He completed her intermediate from Kakatiya Junior College and 10th class from Kakatiya HighSchool. His technical skills include C, C++. He also has a basic understanding ofPython. His participations include: National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing"which was conducted from 10th to 12th December 2020,"MHRDInnovation cell On Startup policy".His areas of interest areProfessional in Marketing business. He completed few certification courses from online platforms like Coursera, CursaApp and SoloLearn.

APPENDICES

```
import numpy as np
import pandas as pd
from sklearn.metrics import confusion_matrix, cohen_kappa_score
from sklearn.metrics import f1_score, recall_score
def PrintStats(cmat, y_test, pred):
  # separate out the confusion matrix components
tpos =cmat[0][0]
fneg = cmat[1][1]
fpos = cmat[0][1]
tneg =cmat[1][0]
  # calculate F!, Recall scores
  f1Score = round(f1_score(y_test, pred), 2)
recallScore = round(recall_score(y_test, pred), 2)
  # calculate and display metrics
  print(cmat)
print( 'Accuracy: '+ str(np.round(100*float(tpos+fneg)/float(tpos+fneg + fpos + tneg),2))+'%')
print( 'Cohen Kappa: '+ str(np.round(cohen_kappa_score(y_test, pred),3)))
print("Sensitivity/Recall for Model : {recall_score}".format(recall_score = recallScore))
print("F1 Score for Model : {f1_score}".format(f1_score = f1Score))
def RunModel(model, X_train, y_train, X_test, y_test):
model.fit(X_train, y_train.values.ravel())
  pred = model.predict(X_test)
  matrix = confusion_matrix(y_test, pred)
  return matrix, pred
```

```
df = pd.read_csv(r'C:\Users\acss\OneDrive\Desktop\CREDIT CARD FRAUD DETECTION USING RANDOM FOREST & CART ALGORITHM\creditcard.csv')

class_names = {0:'Not Fraud', 1:'Fraud'}

print(df.Class.value_counts().rename(index = class_names))

feature_names = df.iloc[:, 1:30].columns

target = df.iloc[:1, 30: ].columns


data_features = df[feature_names]

data_target = df[target]


from sklearn.model_selection import

train_test_splitnp.random.seed(123)

X_train, X_test, y_train, y_test=train_test_split(data_features,      data_target, train_size=0.70, test_size=0.30,random_state=1)


from sklearn.ensemble import RandomForestClassifier

rf = RandomForestClassifier(n_estimators = 100, n_jobs =4)

cmat, pred = RunModel(rf, X_train, y_train, X_test, y_test)

PrintStats(cmat, y_test, pred)

fraud_records = len(df[df.Class == 1])

# pull the indicies for fraud and valid rows

fraud_indices = df[df.Class == 1].index

normal_indices = df[df.Class == 0].index

# randomly collect equal samples of each type

under_sample_indices = np.random.choice(normal_indices, fraud_records, False)

df_undersampled = df.iloc[np.concatenate([fraud_indices,under_sample_indices]),:]

X_undersampled = df_undersampled.iloc[:,1:30]
```

```
Y_undersampled = df_undersampled.Class

X_undersampled_train, X_undersampled_test, Y_undersampled_train, Y_undersampled_test =
train_test_split(X_undersampled,Y_undersampled,test_size = 0.3)

lr_undersampled = LogisticRegression(C=1)

# run the new model

cmat, pred = RunModel(lr_undersampled, X_undersampled_train, Y_undersampled_train,
X_undersampled_test, Y_undersampled_test)

PrintStats(cmat, Y_undersampled_test, pred)

cmat, pred = RunModel(lr_undersampled, X_undersampled_train, Y_undersampled_train, X_test, y_test)

PrintStats(cmat, y_test, pred)
```

A

PROJECT REPORT

On

USING ARTIFICIAL NEURAL NETWORKS TO IDENTIFY FAKE PROFILES

Submitted by

1)Anshuman Singh(17K81A05C6) **2)S Raja Lingam Seth(17K81A05H4)**

3)Md Afroz (17K81A05G1) **4)Asad Ul Islam(18K85A0506)**

5)Rohan Mahesh Katkam(18K85A0507)

in partial fulfillment for the award of the

degreeof

BACHELOR OF TECHNOLOGY

IN

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Under the Guidance of

Mrs. Manu Hajari

Assistant Professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ST.MARTIN'S ENGINEERING COLLEGE
An Autonomous Institute

Dhulapally, Secunderabad – 500 100

JUNE 2021

BONAFIDE CERTIFICATE

This is to certify that the project entitled Using artificial neural networks to identify fake profiles, is being submitted by **1.Mr. Anshuman Singh 17K81A05C6, 2.Mr. S Raja Lingam Seth 17K81A05H4, 3.Mr. Md Afroz 17K81A05G1, 4.Mr. Asad Ul Islam 18K85A0506, 5.Mr. Rohan Mahesh Katkam** in partial fulfillment of the requirement for the award of the degree of **BACHELOR OF TECHNOLOGY IN Computer Science And Engineering** is recorded of bonafide work carried out by them. The result embodied in this report have been verified and found satisfactory.

Signature

Mrs. Manu Hajari

Department of CSE

Head of the Department

Dr.M.NARAYANAN

Department of CSE

Internal Examiner

External Examiner

Place:

Date:

DECLARATION

We, the student of **Bachelor of Technology** in Department of Computer Science and Engineering', session: 2017 – 2021, St. Martin's Engineering College, Dhulapally, Kompally, Secunderabad, hereby declare that work presented in this Project Work entitled Using artificial neural networks to identify fake profiles is the outcome of our own bonafide work and is correct to the best of our knowledge and this work has been undertaken taking care of Engineering Ethics. This result embodied in this project report has not been submitted in any university for award of any degree.

Anshuman Singh	17K81A05C6
S Raja Lingam Seth	17K81A05H4
Md Afroz	17K81A05G1
Asad Ul Islam	18K85A0506
Rohan Mahesh Katkam	18K85A0507

ABSTRACT

There is a tremendous increase in technologies these days.. Mobiles are becoming smart. Technology is associated with online social networks which has become a part in every one's life in making new friends and keeping friends , their interests are known easier. But this increase in networking online make many problems like faking their profiles, online impersonation having become more and more in present days. Users are fed with more unnecessary knowledge during surfing which are posted by fake users. Researches have observed that 20% to 40% profiles in online social networks like facebook are fake profiles. Thus this detection of fake profiles in online social networks results into solution using frameworks.

We use machine learning, namely an artificial neural network to determine what are the chances that Facebook friend request is authentic or not. We also outline the classes and libraries involved. Furthermore, we discuss the sigmoid function and how the weights are determined and used. Finally, we consider the parameters of the social network page which are utmost important in the provided solution.

The other dangers of personal data being obtained for fraudulent purposes is the presence of bots and fake profiles. Bots are programs that can gather information about the user without the user even knowing. This process is known as web scraping. What is worse, is that this action is legal. Bots can be hidden or come in the form of a fake friend request on a social network site to gain access to private information.

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible and whose encouragements and guidance have crowded effects with success.

We extended our deep sense of gratitude to Principal**, Dr. P. SANTOSH KUMAR PATRA**, St. Martin's Engineering College, Dhulapally, for permitting us to undertake this project.

We are also thankful to **Dr. M.NARAYANAN**, Head of the Department, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

We would like to thank **Dr. T. POONGOTHAI,** Department of Computer Science and Engineering (AI & ML) for her encouragement and insightful comments and as well as our project coordinators **Dr. B.RAJALINGAM**, Associate Professor and **Dr. N. SATHEESH**, Professor, in Department of Computer Science and Engineering for their valuable support.

We would like to express our sincere gratitude and indebtedness to our project supervisor **Mrs. Manu Hajari**, Assistant Professor Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

Finally, we express thanks to all those who have helped us successfully to completing this project. Furthermore, we would like to thank our family and friends for their moral support and encouragement.

We express thanks to all those who have helped us in successfully completing the project.

Anshuman Singh **17K81A05C6**

S Raja Lingam Seth **17K81A05H4**

Md Afroz **17K81A05G1**

Asad Ul Islam **18K85A0506**

Rohan Mahesh Katkam **18K85A0507**

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

<AVI>	Audio Video Interlace
<BMP>	Bitmap
<CPU>	Central Processing Unit
<GB>	Giga Bytes
<GUI>	Graphical User Interface

CHAPTER 1

INTRODUCTION

1.1 PROJECT OVERVIEW

Each input neuron would be a different, previously chosen feature of each profile converted into a numerical value (e.g., gender as a binary number, female 0 and male 1) and if needed, divided by an arbitrary number (e.g., age is always divided by 100) to minimize one feature having more influence on the result than the other. The neurons represent nodes. Each node would be responsible for exactly one decision-making process

1.2 PROJECT OBJECTIVES

In our solution, we use machine learning, namely an artificial neural network to determine what are the chances that a friend request is authentic or not.

We utilize Microsoft Excel to store old and new fake data profiles. The algorithm then stores the data in a data frame. This collection of data will be divided into a training set and a testing set. We would need a data set from the social media sites to train our model.

For the training set, the features that we use to determine a fake profile are Account age, Gender, User age, Link in the description, Number of messages sent out, Number of friend requests sent out, Entered location, Location by IP, Fake or Not. Each of these parameters is tested and assigned a value. For example, for the gender parameter if the profile can be determined to be a female or male a value of (1) is assigned to the training set for Gender. The same process is applied to other parameters

CHAPTER 2

LITERATURE SURVEY

Online social media is the place each person has a outlook then be able to keep connecting their relations, transfer their updates, join with the people having same likes. Online Social Networks makes use of front end technologies, which permits permanency accounts in accordance with to know each other. Facebook, Twitter are developing along with humans to maintain consultation together with all others. The online accounts welcome people including identical hobbies collectively who makes users easier after perform current friends. Gaming and entertaining web sites which have extra followers unintentionally that means more fan base and supreme ratings. Ratings drives online account holders to understand newer approaches not naturally or manually to compete more with their neighbours. By these analogies, the maximum famous candidate in an election commonly gets a greater number of votes. Happening of fake social media accounts and interests may be known. Instance is fake online account being sold on-line at a online market places for minimum price, brought from collaborative working offerings.

More often feasible to have Twitter fans and Facebook media likes in online. Fake user accounts may be created by humans or computers like bots, cyborgs. Cyborg is half bot and half human account. These accounts are usually opened by human, but their actions are made by bots. Another reason for people to create fake profiles for defaming accounts they dislike. This type of users creates accounts with the username of the people they hate and post irrelevant stories and snap shots on their accounts to redirect everybody so that they assume that particular person is awful and make their reputation low.

Most attackers are in it to make money. They make money by distributing unwanted ads (spam) or capturing accounts they can reuse or resale (phishing). Spammers gather resources to know fake and real users, email ids, IP locations and computing knowledge power. Every one of these advantages can have a huge expense related with them, and an assault, similar to any business adventure, needs benefit to continue onward. Attackers more often use Facebook logins, applications, Events, Group users to gather login credentials, spam users, and ultimately gain profits. They need email records, treats, and a wide scope of IP delivers to go around notoriety based protections. Moreover, they use telephone numbers, taken charge cards, and CAPTCHA arrangements trying to go around validation checks.

Facebook security privileges its system to gather users to prevent spams and fishing accounts. Facebook Immune System does continuous minds all gather and each its activity made by it. Social bot is a known that stops and controls social online accounts. Bots socially is an auto generated software. Precised way a social account duplicates relies upon at the social media, also in contrast to general bot, a social bot interacting more in different customers that the social bot is a actual man or woman. More auto generated programs or semi generated computer programs that duplicate the human behavior in Social-media. So, to use them hackers attack online social networks.

Cyborg bots appear like accounts of human from random calls of human, often selected human users image and user records published more often from collective accounts to be prepare from before online account attackers. Cyborg bots ship gather random users. If a person acknowledges the request from user, ship to get request of the account agree request, will increase popularity price because of lifestyle of mutual friend’s request, will increase popularity price because of lifestyles of mutual friends.

CHAPTER 3

SOFTWARE AND HARDWARE REQUIREMENTS

All computer software needs certain hardware components or other software resources to be present on a computer. These prerequisites are known as (computer) system requirements and are often used as a guideline as opposed to an absolute rule. Most software defines two sets of system requirements: minimum and recommended. With increasing demand for higher processing power and resources in newer versions of software, system requirements tend to increase over time. Industry analysts suggest that this trend plays a bigger part in driving upgrades to existing computer systems than technological advancements.

Hardware requirements: The most common set of requirements defined by any operating system or software application is the physical computer resources, also known as hardware, a hardware requirements list is often accompanied by a hardware compatibility list (HCL), especially in case of operating systems. An HCL lists tested, compatible, and sometimes incompatible hardware devices for a particular operating system or application

3.1 SOFTWARE REQUIREMENTS

For developing the Application

1. Python
2. Django
3. Mysql
4. Mysqlclient
5. WampServer 2.4

Technologies and Languages used to Develop
Python

1. HOME PAGE:-

- XML

- JAVA

2. ADMIN LOGIN PAGE:-

- XML

- JAVA

3. USER LOGIN PAGE

- XML
- JAVA

4. GENERATE ANN MODEL PAGE

- XML

- JAVA

5. VIEW TRAINED DATASET PAGE:-

- XML

- JAVA

6. RESULT PAGE:-

- XML

- JAVA

3.2 HARDWARE REQUIREMENTS

- Operating System supported by
  1. Windows 7
  2. Windows XP
  3. Windows 8
- Processor – Pentium IV or higher
- RAM -- 256 MB
- Space on Hard Disk -- Minimum 512 MB

CHAPTER 4

SOFTWARE DEVELOPMENT ANALYASIS

4.1 OVERVIEW OF THE PROBLEM

The other dangers of personal data being obtained for fraudulent purposes is the presence of bots and fake profiles. Bots are programs that can gather information about the user without the user even knowing. This process is known as web scraping. What is worse, is that this action is legal. Bots can be hidden or come in the form of a fake friend request on a social network site to gain access to private information, the ever-increasing dependency on computer technology has left the average citizen vulnerable to crimes such as data breaches and possible identity theft.

Malicious users create fake profiles to phish login information from unsuspecting users. A fake profile will send friend requests to many users with public profiles. These counterfeit profiles bait unsuspecting users with pictures of people that are considered attractive. Once the user accepts the request, the owner of the phony profile will spam friend requests to anyone this user is a friend.

4.2DEFINE THE PROBLEM

We use machine learning, namely an artificial neural network to determine what are the chances that Facebook friend request is authentic or not. We also outline the classes and libraries involved. Furthermore, we discuss the sigmoid function and how the weights are determined and used. Finally, we consider the parameters of the social network page which are utmost important in the provided solution.

The dangers of personal data being obtained for fraudulent purposes is the presence of bots and fake profiles. Bots are programs that can gather information about the user without the user even knowing. This process is known as web scraping. What is worse, is that this action is legal. Bots can be hidden or come in the form of a fake friend request on a social network site to gain access to private information. Thus, this detection of fake profiles in online social networks results into solution using frameworks.

4.3 MODULES OVERVIEW

Admin Module: Admin will login to application by using username as ‘admin’ and password as ‘admin’ and then perform below actions.

a) **Generate ANN Train Model:** Admin will upload profile dataset to ANN algorithm to build train model. This train model can be used to predict fake or genuine account by taking new account test data. ANN algorithm will be trained with all previous users fake and genuine account data and then whenever we gave new test data then that ANN train model will be applied on new test data to identify whether given new account details are from genuine or fake users.

b) **View ANN Train Dataset:** Using this module admin can view all dataset used to train ANN model. The dataset will be displayed in the format of rows and columns, so that it can be easily understood by

the corresponding user or the admin. The dataset can only be read by the user but can be modified only by the administrator or the developer working on the algorithm.

c) User Module: Any user can use this application and enter test data of new account and call ANN algorithm. ANN algorithm will take new test data and applied train model to predict whether given test data contains fake or genuine details.

4.4 MODULE FUNCTIONALITY

1.Collect Data and pre-process the data

2.Generate fake accounts.

3.Data Validation to find fake and real .

4.Create new features.

5.Apply neural networks, random forest.

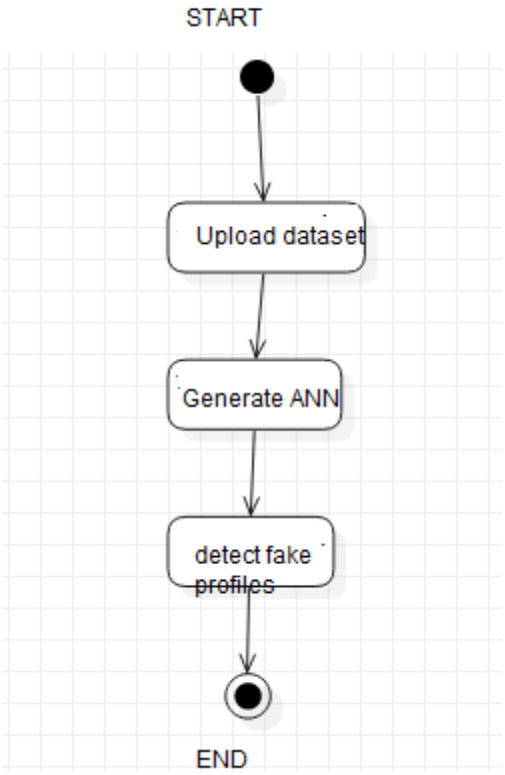
6.Evaluate results of accuracy, recall etc parameters. Thus these steps are implemented for detecting fake profiles.

CHAPTER 5

PROJECT SYSTEM DESIGN

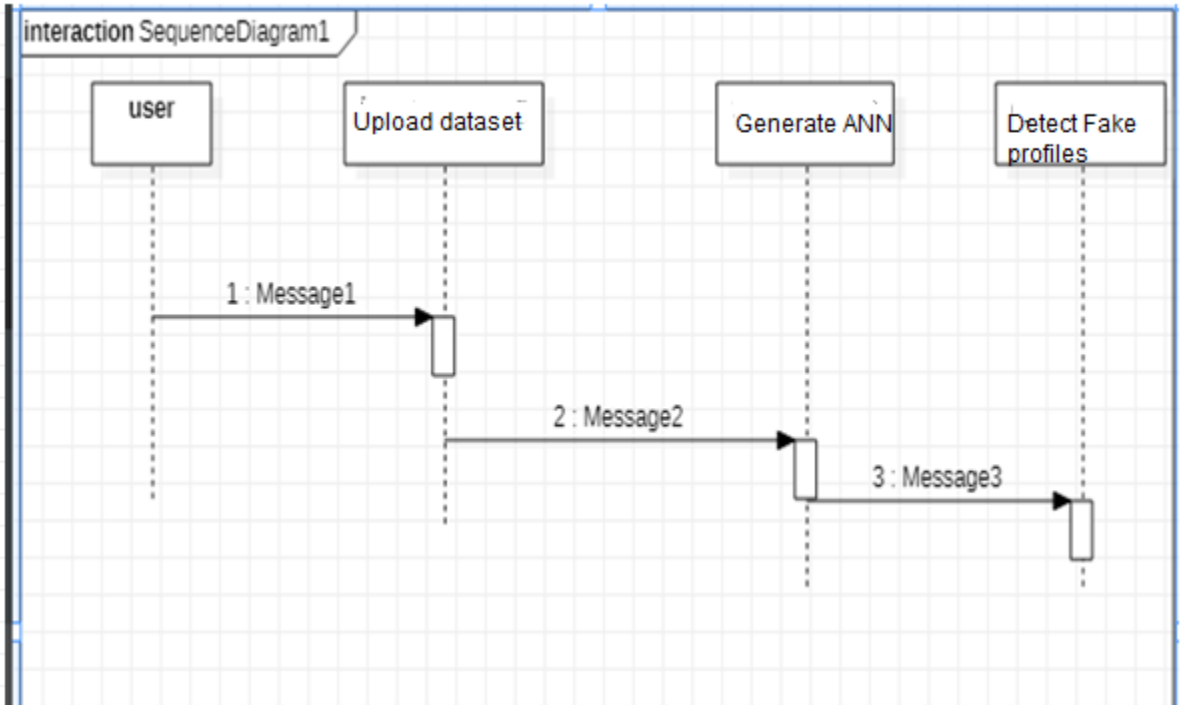
5.1 DATA FLOW DIAGRAMS

5.1.1 ACTIVITY DIAGRAM



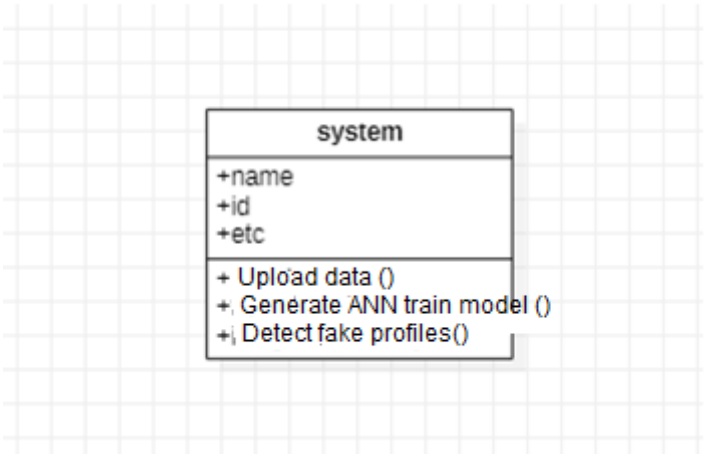
The activity diagram as in Fig-5.1.1 represents the flow of the execution process of the program from uploading the dataset to the prediction or detection of profiles in a continuous flow process.

5.1.2 SEQUENCE DIAGRAM



The sequence diagram as in Fig-5.1.2 represents the sequence of the operations that are carried out in the program i.e., from uploading the dataset to the prediction or detection of profiles in a continuous flow process.

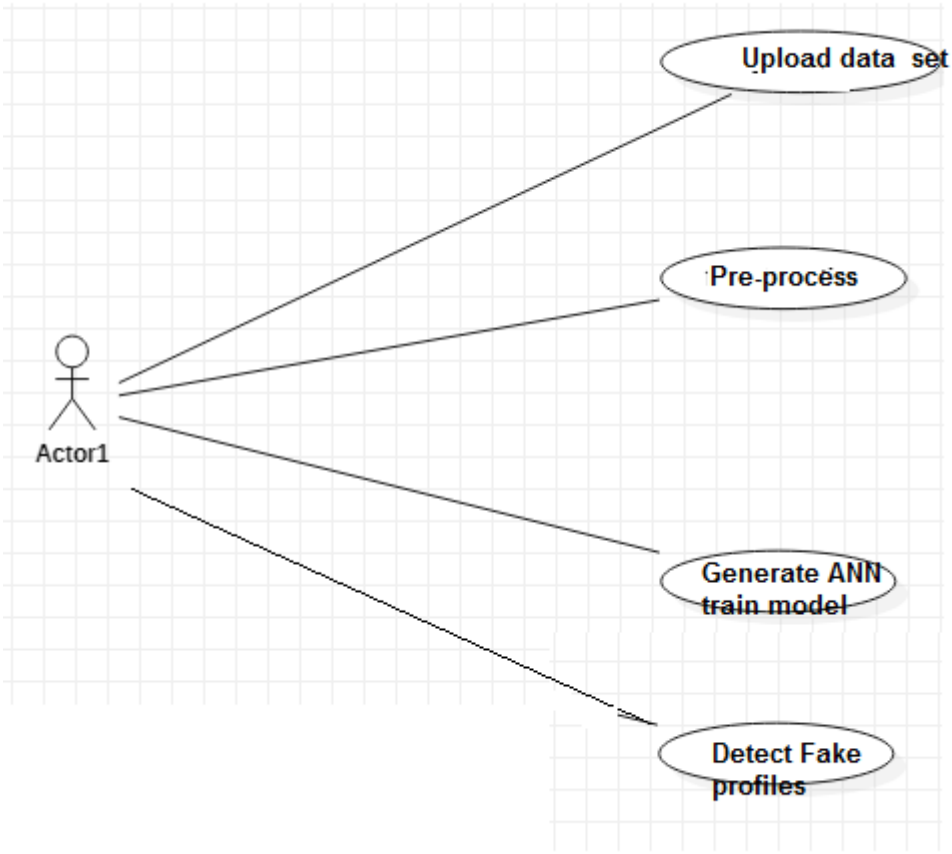
5.3 E-R DIAGRAM



5.3.1 CLASS DIAGRAM

The Class diagram as in Fig. – 5.3.1 represents the generic structure of the class being used to describe the structure of the system. This includes the system variables i.e., the attributes in the dataset that are to be used in order to the working of the algorithm and the methods that are being used to carry out the necessary functions.

5.3.2USE CASE DIAGRAM



The use case diagram as in Fig. – 5.3.2 demonstrates the use cases for the general user who operates on the algorithm to identify the fake profiles. This includes the necessary steps that are carried out by the user and have access to.

CHAPTER 6

PROJECT CODING

6.1 CODE TEMPLATES

1)admin

2)admin screen

3)Index

4)user

5)view data

```
from django.shortcuts import render
from django.template import RequestContext
from django.contrib import messages
from django.http import HttpResponse
import pandas as pd
from sklearn.model_selection import train_test_split
from keras.models import Sequential
from keras.layers.core import Dense,Activation,Dropout
from keras.callbacks import EarlyStopping
from sklearn.preprocessing import OneHotEncoder
from keras.optimizers import Adam

global model

def index(request):
    if request.method == 'GET':
        return render(request, 'index.html', {})

def User(request):
    if request.method == 'GET':
        return render(request, 'User.html', {})


def Admin(request):
    if request.method == 'GET':
        return render(request, 'Admin.html', {})

def AdminLogin(request):
    if request.method == 'POST':
        username = request.POST.get('username', False)
        password = request.POST.get('password', False)
        if username == 'admin' and password == 'admin':
            context= {'data':'welcome '+username}
            return render(request, 'AdminScreen.html', context)
```

```
    else:
     context= {'data':'login failed'}
     return render(request, 'Admin.html', context)


def importdata():
    balance_data = pd.read_csv('C:/FakeProfile/Profile/dataset/dataset.txt')
    balance_data = balance_data.abs()
    rows = balance_data.shape[0]  # gives number of row count
    cols = balance_data.shape[1]  # gives number of col count
    return balance_data

def splitdataset(balance_data):
    X = balance_data.values[:, 0:8]
    y_= balance_data.values[:, 8]
    y_ = y_.reshape(-1, 1)
    encoder = OneHotEncoder(sparse=False)
    Y = encoder.fit_transform(y_)
    print(Y)
    train_x, test_x, train_y, test_y = train_test_split(X, Y, test_size=0.2)
    return train_x, test_x, train_y, test_y

def UserCheck(request):
    if request.method == 'POST':
      data = request.POST.get('t1', False)
      input = 'Account_Age,Gender,User_Age,Link_Desc,Status_Count,Friend_Count,Location,Location_IP\n';
      input+=data+"\n"
      f = open("C:/FakeProfile/Profile/dataset/test.txt", "w")
      f.write(input)
      f.close()
      test = pd.read_csv('C:/FakeProfile/Profile/dataset/test.txt')
      test = test.values[:, 0:8]
      predict = model.predict_classes(test)
      print(predict[0])
      msg = ''
      if str(predict[0]) == '0':
        msg = "Given Account Details Predicted As Genuine"
      if str(predict[0]) == '1':
        msg = "Given Account Details Predicted As Fake"
      context= {'data':msg}
      return render(request, 'User.html', context)

def GenerateModel(request):
    global model
    data = importdata()
    train_x, test_x, train_y, test_y = splitdataset(data)
    model = Sequential()
    model.add(Dense(200, input_shape=(8,), activation='relu', name='fc1'))
    model.add(Dense(200, activation='relu', name='fc2'))
    model.add(Dense(2, activation='softmax', name='output'))
    optimizer = Adam(lr=0.001)
```

```
    model.compile(optimizer, loss='categorical_crossentropy', metrics=['accuracy'])
    print('CNN Neural Network Model Summary: ')
    print(model.summary())
    model.fit(train_x, train_y, verbose=2, batch_size=5, epochs=200)
    results = model.evaluate(test_x, test_y)
    ann_acc = results[1] * 100
    context= {'data':'ANN Accuracy : '+str(ann_acc)}
    return render(request, 'AdminScreen.html', context)

def ViewTrain(request):
    if request.method == 'GET':
        strdata = '<table border=1 align=center width=100%><tr><th><font size=4 color=white>Account Age</th><th>
<font size=4 color=white>Gender</th><th><font size=4 color=white>User Age</th><th><font size=4 color=white>
Link Description</th> <th><font size=4 color=white>Status Count</th><th><font size=4 color=white>Friend Count
</th><th><font size=4 color=white>Location</th><th><font size=4 color=white>Location IP</th><th><font size=4
color=white>Profile Status</th></tr><tr>'
        data = pd.read_csv('C:/FakeProfile/Profile/dataset/dataset.txt')
        rows = data.shape[0]  # gives number of row count
        cols = data.shape[1]  # gives number of col count
        for i in range(rows):
            for j in range(cols):
                strdata+='<td><font size=3 color=white>'+str(data.iloc[i,j])+'</font></td>'
            strdata+='</tr><tr>'
        context= {'data':strdata}
        return render(request, 'ViewData.html', context)


from django.urls import path

from . import views

urlpatterns = [path("index.html", views.index, name="index"),
           path("Admin.html", views.Admin, name="Admin"),
           path("AdminLogin", views.AdminLogin, name="AdminLogin"),
           path("GenerateModel", views.GenerateModel, name="GenerateModel"),
           path("ViewTrain", views.ViewTrain, name="ViewTrain"),
           path("User.html", views.User, name="User"),
           path("UserCheck", views.UserCheck, name="UserCheck"),
]
```

6.2 OUTLINE FOR VARIOUS FILES

ADMIN-ADMIN SCREEN-INDEX-USER-VIEW DATA-END

6.3 CLASS WITH FUNCTIONALITY

In software engineering, a functional requirement defines a system or its component. It describes the functions a software must perform. A function is nothing but inputs, its behavior, and outputs. It can be a calculation, data manipulation, business process, user interaction, or any other specific functionality which defines what function a system is likely to perform.

Functional software requirements help you to capture the intended behavior of the system. This behavior may be expressed as functions, services or tasks or which system is required to perform.

6.4 METHODS INPUT AND OUTPUT PARAMETERS

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the following things:

- What data should be given as input?
- How the data should be arranged or coded?
- The dialog to guide the operating personnel in providing input.
- Methods for preparing input validations and steps to follow when error occur.

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making.

1. Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is designed so that people will find the system can use easily and effectively. When analysis design computer output, they should Identify the specific output that is needed to meet the requirements.

2.Select methods for presenting information.

3.Create document, report, or other formats that contain information produced by the system.

The output form of an information system should accomplish one or more of the following objectives.

- Convey information about past activities, current status or projections of the
- Future.
- Signal important events, opportunities, problems, or warnings.
- Trigger an action.

- Confirm an action.

CHAPTER 7

PROJECT TESTING

7.1VARIOUS TEST CASES

Test Case 1	
Test Case Name	Empty login fields testing
Description	In the login screen if the username and password fields are empt
Output	Login fails showing an alert box asking to enter username and password.

TABLE 7.1.1 TEST CASE 1

Test Case 2	
Test Case Name	Wrong login fields testing
Description	A unique username and password are set by administrator. On ng wrong username or password gives.
Output	Login fails showing an alert box username or password incorrect.

TABLE 7.1.2 TEST CASE 2

Test Case 3	
Test Case Name	User Signup Fails.
Description	User signup need to provide all data.
Output	Signup Fails and an alert message appears asking to enter valid and name.

TABLE 7.1.3 TEST CASE 3

7.2 BLACK BOX TESTING

Black box testing treats the software as a "black box"—without any knowledge of internal implementation. Black box testing methods include equivalence partitioning, boundary value analysis, all-pairs testing, fuzz testing, model-based testing, traceability matrix, exploratory testing,and specification-based testing.

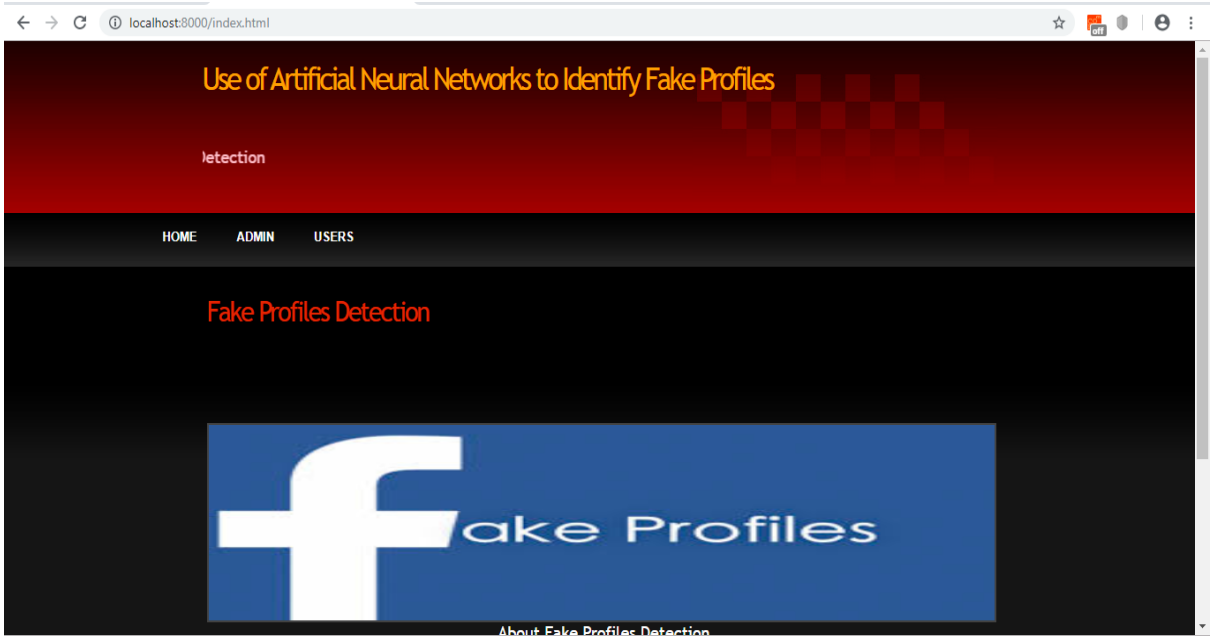
7.3 WHITE BOX TESTING

White box testing is when the tester has access to the internal data structures and algorithms including the code that implement these.

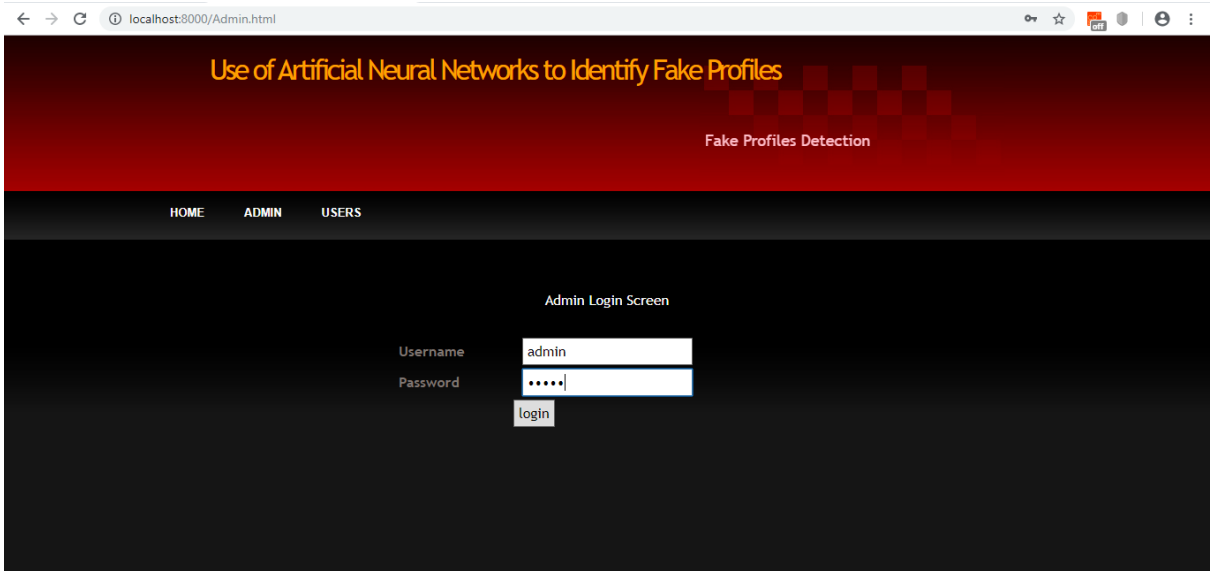
CHAPTER 8

OUTPUT SCREENS

8.1 USER INTERFACES

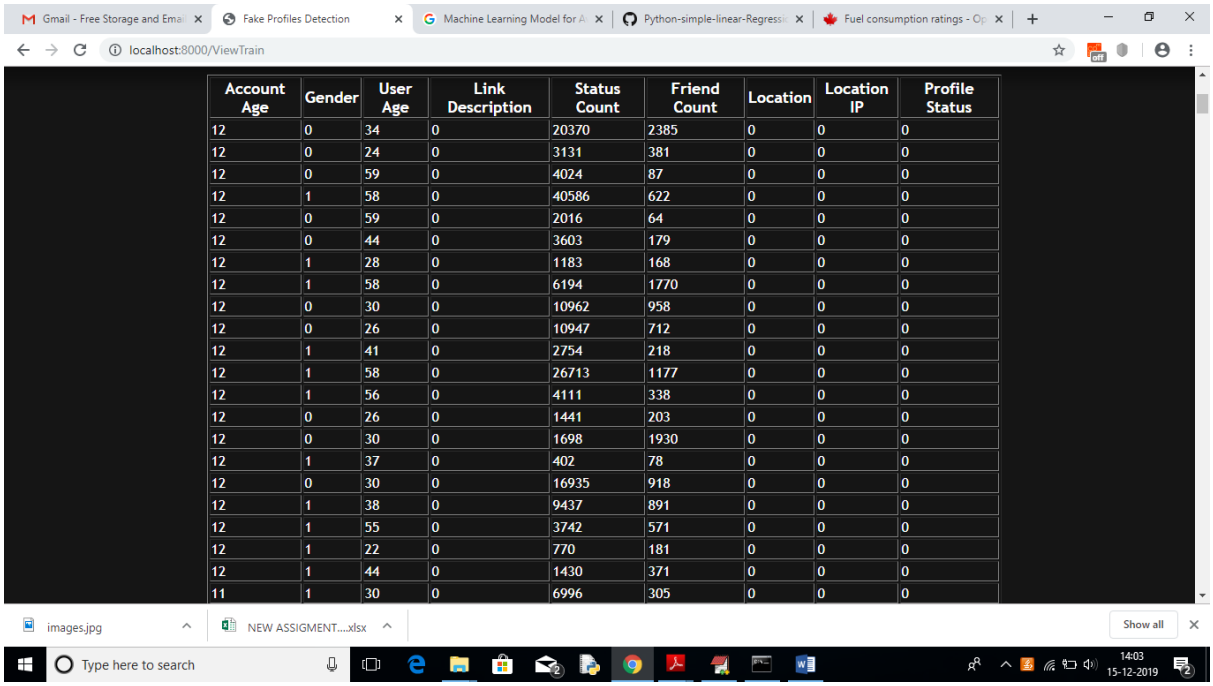


8.1.1. Home Screen



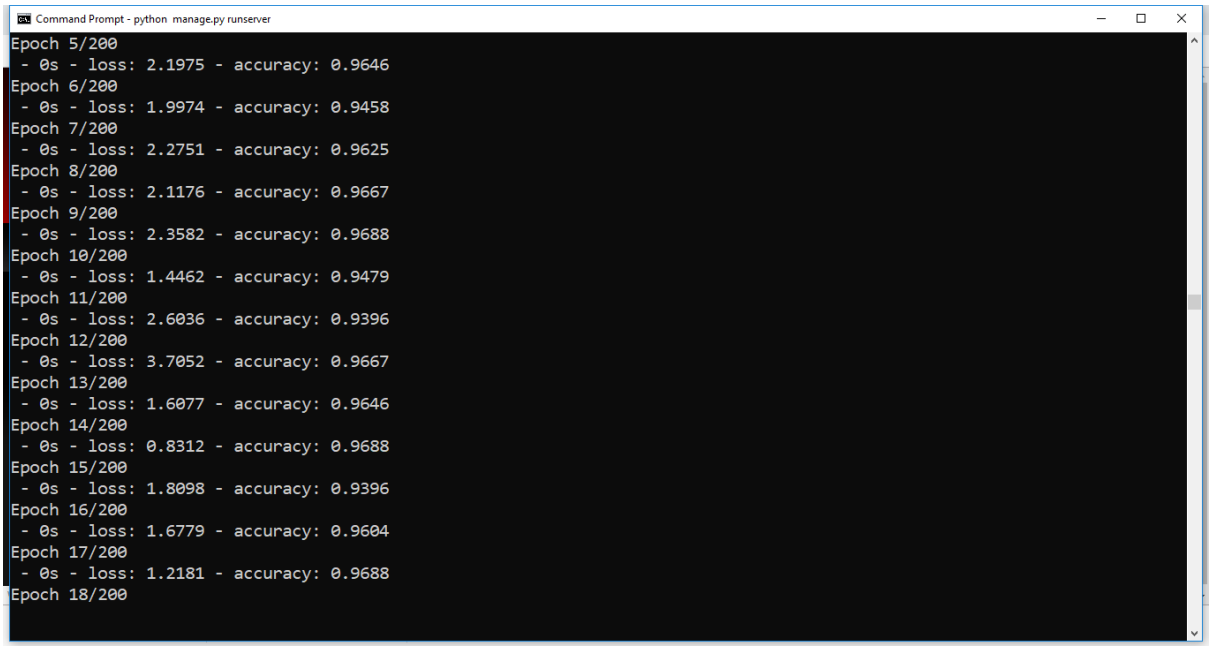
8.1.2 Admin Login

8.2 OUTPUT SCREENS



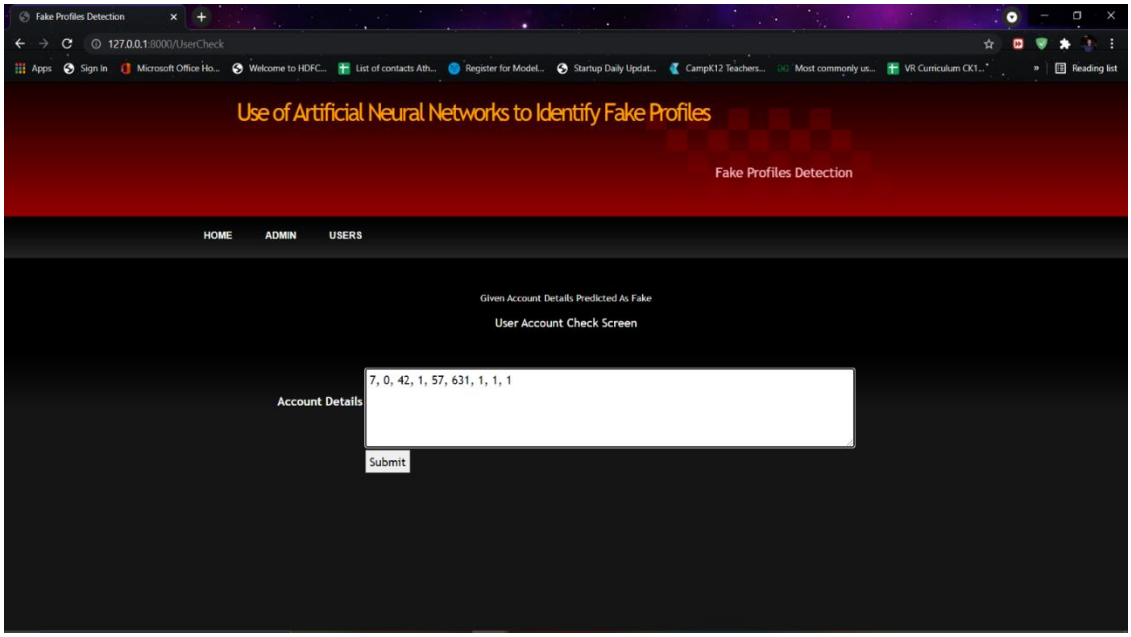
Account Age	Gender	User Age	Link Description	Status Count	Friend Count	Location	Location IP	Profile Status
12	0	34	0	20370	2385	0	0	0
12	0	24	0	3131	381	0	0	0
12	0	59	0	4024	87	0	0	0
12	1	58	0	40586	622	0	0	0
12	0	59	0	2016	64	0	0	0
12	0	44	0	3603	179	0	0	0
12	1	28	0	1183	168	0	0	0
12	1	58	0	6194	1770	0	0	0
12	0	30	0	10962	958	0	0	0
12	0	26	0	10947	712	0	0	0
12	1	41	0	2754	218	0	0	0
12	1	58	0	26713	1177	0	0	0
12	1	56	0	4111	338	0	0	0
12	0	26	0	1441	203	0	0	0
12	0	30	0	1698	1930	0	0	0
12	1	37	0	402	78	0	0	0
12	0	30	0	16935	918	0	0	0
12	1	38	0	9437	891	0	0	0
12	1	55	0	3742	571	0	0	0
12	1	22	0	770	181	0	0	0
12	1	44	0	1430	371	0	0	0
11	1	30	0	6996	305	0	0	0

8.2.1. Data Set

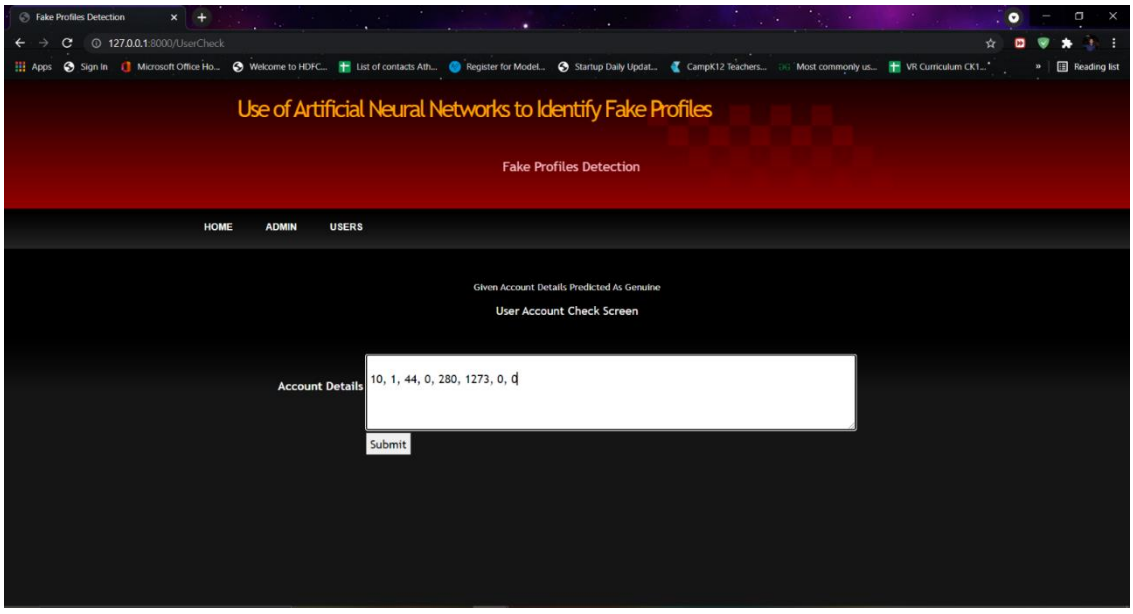


```
Command Prompt - python manage.py runserver
Epoch 5/200
 - 0s - loss: 2.1975 - accuracy: 0.9646
Epoch 6/200
 - 0s - loss: 1.9974 - accuracy: 0.9458
Epoch 7/200
 - 0s - loss: 2.2751 - accuracy: 0.9625
Epoch 8/200
 - 0s - loss: 2.1176 - accuracy: 0.9667
Epoch 9/200
 - 0s - loss: 2.3582 - accuracy: 0.9688
Epoch 10/200
 - 0s - loss: 1.4462 - accuracy: 0.9479
Epoch 11/200
 - 0s - loss: 2.6036 - accuracy: 0.9396
Epoch 12/200
 - 0s - loss: 3.7052 - accuracy: 0.9667
Epoch 13/200
 - 0s - loss: 1.6077 - accuracy: 0.9646
Epoch 14/200
 - 0s - loss: 0.8312 - accuracy: 0.9688
Epoch 15/200
 - 0s - loss: 1.8098 - accuracy: 0.9396
Epoch 16/200
 - 0s - loss: 1.6779 - accuracy: 0.9604
Epoch 17/200
 - 0s - loss: 1.2181 - accuracy: 0.9688
Epoch 18/200
```

8.2.2. Model Training



8.2.3. Prediction result-1



8.2.4. Prediction result - 2

CHAPTER 9

EXPERIMENTAL RESULTS

• The experimental results were accordant to the algorithm used.

• The accuracy of the algorithm trained on the given dataset turned out to be as expected.

• The output was tested for several values for which the algorithm gave accurate results based on the trained data.

• The algorithm is able to predict whether the account is genuine or fake and display

it to the user.

• At the time of submission of my application was capable of doing the following:

• Displaying the home screen with different fragments.

• Authentication of user by using login screen through MySQL Database.

• Home screen to display based on user.

• After successful login of admin,we can generate and view train model.

• The algorithm accuracy and training status is displayed on Console/Terminal.

• After successful login of user, they can provide account details.

• The user gets the message if the account is Fake or Genuine.

• Logout and end the session.

CHAPTER 10

CONCLUSION AND FUTURE ENHANCEMENTS

Conclusion

we use machine learning, namely an artificial neural network to determine what are the chances that a friend request is authentic are or not. Each equation at each neuron (node) is put through a Sigmoid function. We use a training data set by Facebook or other social networks. This would allow the presented deep learning algorithm to learn the patterns of bot behavior by back propagation, minimizing the final cost function and adjusting each neuron's weight and bias.

Scope for future work

• Each input neuron would be a different, previously chosen feature of each profile converted into a numerical value (e.g., gender as a binary number, female 0 and male 1) and if needed, divided by an arbitrary number (e.g., age is always divided by 100) to minimize one feature having more influence on the result than the other. The neurons represent nodes. Each node would be responsible for exactly one decision-making process

CHAPTER 11

REFRENCES

[1] Yadongzhou, Daewookkim, Junjiezhang, (Member, Ieee), Lili Liu1, Huanjin3, "(IEEE) ProGuard: Detecting Malicious Accounts in Social Network-Based Online Promotions(2019).

[2] Mauro Conti University of Padua, Radha Poovendran University of Washington, Marco Secchiero University of Padua,“ FakeBook: Detecting Fake Profiles in online Social Networks(2019)", ACM /IEEE International Conference on Advances in Social Networks Analysis and Mining

[3] Jain, Y., NamrataTiwari, S., Jain, S.: A comparative analysis of various credit card fraud detection techniques. Int. J. Recent. Technol. Eng. (2277–3878), 7(5S2), 402–407 (2019)

[4] Dr. Narsimha.G, Dr. Jayadev Gyani, P. Srinivas Rao ,"Fake Profiles Identification in Online Social Networks Using Machine Learning and NLP", International Journal of Applied Engineering Research ISSN 0973-4562, Number 6, Volume 13.

[5] D. Rajeswara Rao & V. Pellakuri. Training and development of artificial neural network models: Single layer feedforward and multi layer feedforward neural network(2019). Journal of Theoretical and Applied Information Technology, 150-156,84(2).

[6] Challa, N.,. Pasupuleti, S. K, & Chandra, J. V. A practical approach to E-mail spam filters to protect data from advanced persistent threat.(2018) Paper presented at the Proceedings of IEEE International Conference on Circuit, Power and Computing Technologies.

[7] D. Rajeswara Rao , & P. Vidyullatha. Machine learning techniques on multidimensional curve fitting data based on r_ square and chi_square methods(2019). International Journal of Electrical and Computer Engineering

[8] Pradeepini, G., Patil, S. T. , & Bangare, S. L (2019). Brain tumor classification using mixed method approach. Paper presented at the International Conference on Information Communication and Embedded Systems, ICICES

[9] Kutub Thakur; Thaier Hayajneh; Jason Tseng, Cyber Security in Social Media: Challenges and the Way Forward(2019).

[10] Jianhui Qiu; Xingfa Shen; Yan Guo; Jian Yao; Renzhe Fang, Detecting Malicious Users in Online Dating Application IEEE (2019)

[11] Reddy, A. V. N., & Phanikrishna, C. Contour tracking based knowledge extraction and object recognition using deep learning neural networks(2019).

[12] Sucharitha, G., &Senapati, R. K. Local extreme edge binary patterns for face recognition and image retrieval(2018). Journal of Advanced Research in Dynamical and Control Systems_10, 644-654

[13] Koh, Pang Wei, et al. "Concept bottleneck models." Proceedings of the 37th International Conference on Machine Learning (2020).

[14] Mothilal, Ramaravind K., Amit Sharma, and Chenhao Tan. "Explaining machine learning classifiers through diverse counterfactual explanations." Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 2020

[15] Su, Jiawei, Danilo Vasconcellos Vargas, and Kouichi Sakurai. "One pixel attack for fooling deep neural networks." IEEE Transactions on Evolutionary Computation (2019).

CHAPTER 12

PUBLICATION

Fake Profile Identification in Online Social Networks
St. Martin's Engineering College

S Raja Lingam Seth [Student1], Anshuman Singh [Student2], Rohan Katkam[Student3], Md. Afroz [Student4], Asad ul Islam [Student5], Manu Hajari Assistant Professor, Dr. M. Narayanan HOD CSE

Abstract: There is a tremendous increase in technologies these days. Mobiles are becoming smart. Technology is associated with online social networks which has become a part in every one's life in making new friends and keeping friends, their interests are known easier. But this increase in networking online makes many problems like faking their profiles, online impersonation having become more and more in present days. Users are fed with more unnecessary knowledge during surfing which are posted by fake users. Researches have observed that 20% to 40% profiles in online social networks like Facebook are fake profiles. Thus, this detection of fake profiles in online social networks results into solution using frameworks.

Keywords: Online Social Networks, Fake profiles, Classification, Neural Network.

I. INTRODUCTION

Online social media is the place each person has a outlook then be able to keep connecting their relations, transfer their updates, join with the people having same likes. Online Social Networks makes use of front-end technologies, which permits permanency accounts in accordance with to know each other. Facebook, Twitter is developing along with humans to maintain consultation together with all others. The online accounts welcome people including identical hobbies collectively who makes users easier after perform current friends. Gaming and entertaining web sites which have extra followers unintentionally that means more fan base and supreme ratings. Ratings drives online account holders to understand newer approaches not naturally or manually to compete more with their neighbours. By these analogies, the maximum famous candidate in an election commonly gets a greater number of votes. Happening of fake social media accounts and interests may be known. Instance is fake online account being sold on-line at a online market places for minimum price, brought from collaborative working offerings.

More often feasible to have Twitter fans and Facebook media likes in online. Fake user accounts may be created by humans or computers like bots, cyborgs. Cyborg is half bot and half human account. These accounts are usually opened by human, but their actions are made by bots. Another reason for people to create fake profiles for defaming accounts they dislike. This type of users creates accounts with the username of the people they hate and post irrelevant stories and snap shots on their accounts to redirect everybody so that they assume that particular person is awful and make their reputation low. Most attackers are in it to make money. They make money by distributing unwanted ads (spam) or capturing accounts they can reuse or resale (phishing). Spammers gather resources to know fake and real users, email ids, IP locations and computing knowledge power. Every one of these advantages can have a huge expense related with them, and an assault, similar to any business adventure, needs benefit to continue onward. Attackers more often use Facebook logins, applications, Events, Group users to gather login credentials, spam users, and ultimately gain profits. They need email records, treats, and a wide scope of IP delivers to go around notoriety-based protections. Moreover, they use telephone numbers, taken charge cards, and CAPTCHA arrangements trying to go around validation checks.

Facebook security privileges its system to gather users to prevent spams and fishing accounts. Facebook Immune System does continuous minds all gather and each its activity made by it. Social bot is a known that stops and controls social online accounts. Bots socially is an auto generated software. Précised way a social account duplicates relies upon at the social media, also in contrast to general bot, a social bot interacting more in different customers that the social bot is a actual man or woman. More auto generated programs or semi generated computer programs that duplicate the human behavior in Social media. So, to use them hackers attack online social networks.

Cyborg bots appear like accounts of human from random calls of human, often selected human users' image and user

records published more often from collective accounts to be prepare from before online account attackers. Cyborg bots ship gather random users. If a person acknowledges the request from user, ship to get request of the account agree request, will increase popularity price because of lifestyle of mutual friend's request, will increase popularity price because of lifestyles of mutual friends.

II. RELATED WORK

Accounts in online social media have heaps of input data like name, sexual orientation, companions, devotees, preferences, area numbers. Half part of this input data are both of public and private. We have to use input that are public to know profiles which are phony for interpersonal organization as data from private is unavailable.

In any case, on the off chance that our proposed plan is utilized by the interpersonal interaction organizations itself, at that point they can utilize the private data of the users to know not from abusing from security issues. Considered data is highlights for profiles to classify of phony and genuine profiles.

For detecting fake profiles, we followed these steps:

1. Functions are to be selected after choice of attributes, the data set of profiles which are already classified as fake or real are wanted for the schooling motive of the classification algorithm. We have used a publicly available dataset of 1337 fake customers and 1481 actual users which includes numerous attributes consisting of call, status count, number of friends, fans depend, favorites, languages regarded and so forth.
2. The selected attributes are extracted from profile for the purpose of type.
3. After this the dataset of fake and actual seasoned files are prepared. From this dataset, 80% of both seasoned files (authentic and pretend) are used to prepare a schooling dataset and 20% of both profiles are used to put together a testing dataset.
4. The schooling dataset is then fed to the classification set of rules. It learns from the education dataset and is predicted to offer correct elegance labels for the testing dataset.
5. The labels from the testing dataset are eliminated and are left for determination by the educated classifier.
6. The result of classification algorithm is shown in 4.4. we've got used two classification algorithms and have compared the efficiency of these algorithms.
7. The proposed structure in the figure 1 shows the succession of procedures that should be pursued for persistent location of phony profiles with dynamic gaining from the input of the outcome given by the arrangement calculation.

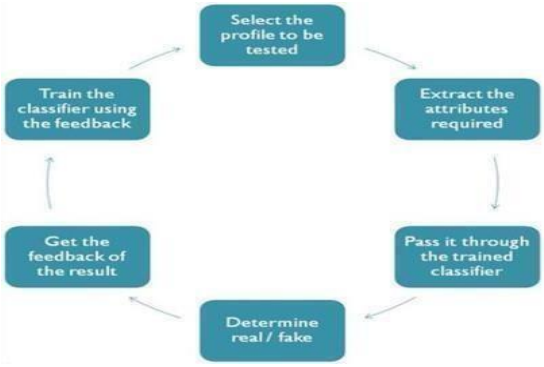


Fig 1. Cycle for Detection

The structure that can without much of a stretch be executed by long range informal communication organizations as they approach client data.

1. Order begins from the determination of profile that should be characterized.
2. When the profile is chosen, the helpful highlights are separated for the reason for order.
3. The separated highlights are then encouraged to prepared classifier.
4. Classifier is prepared routinely as new information is nourished into the classifier.
5. Classifier at that point decides if the profile is veritable or counterfeit.
6. The consequence of order calculation is then checked and input is sustained over into the classifier.
7. As the quantity of preparing information builds the classifier turns out to be more and increasingly precise in foreseeing the phony profiles.

III. METHODOLOGY

Implementation is a technique of categorizing an object into a particular class based on the training data set that was used to train the classifier. We feed the classifier with data set so that we can train it to identify related objects with as best accuracy as possible. Classifier is an algorithm used for classification. In this project we have used two classifiers namely Neural Networks and Support Vector Machines and have thereby compared their efficiencies.

1. Neural Network:

The conventional method by which a computer works is that you provide instructions or algorithms to the computer and it generates output based on it. But what if you do not know the algorithm to solve a problem? Will your computer still be able to provide solutions? If we use conventional techniques, then the computer will not be able to solve the problem unless you provide some instructions. Here comes the concept of Neural Networks. We can still solve such a problem by training a network as such our program will learn on its own and will provide solution close to a certain accuracy. The term Neural Networks was coined in 1943 but could not be implemented then due to lack of technology. Neural Networks learn by example. Neural Networks are based on biological neurons i.e. brain cells and the way information

is processed inside the brain. There are mainly two types of neural networks:

1. Single layer.
2. Multi-layer.

2. Random Forest Classification Technique:

This classifier classifies collection of decision trees to subset of randomly generated training set. Then it augments the likes from decision sub trees to know subclass of handling object for tests. Random forest will generate NA missing values for attributes increase accuracy for larger sets of data. If more number of tress, it doesn't allow to trees to fit model.

Table.1 Comparison of accuracy for different algorithms

Algorithm	Precision	Recall	Accuracy
Decision Tree Network (Twitter and face book)	0.999	0.991	99.9%
Neural Networks Network (Twitter)	1	0.417	-
Naïve Bayes Network (Email and Twitter)	0.778	0.444	94.5%

IV. IMPLEMENTATION

1.Collect Data and preprocess the data.
2.Generate fake accounts.
3.Data Validation to find fake and real.

4.Create new features, Apply neural networks, random forest.

5. Evaluate results of accuracy, recall etc parameters. Thus, these steps are implemented for detecting fake profiles.

Data set:

We needed dataset of fake and genuine profiles. Various attributes to include in the dataset are number of friends, followers, status count. Dataset is resulting to training and testing data. Classification algorithms are trained using training dataset and testing dataset is used to determine efficiency of algorithm. From the dataset used, more than 80 percent of accounts are used to train the data, 20 percent of accounts to test the data.

Attribute	Explanation
Post Count	The average number of posts created by users are expected to have a low count when the account is fake.
Comment Count	Fake accounts share and post unwanted links and advertisements which make a lower count.
Followers Count	Usually, fake profiles have low count but there is high follower count then they may belong to the same group.
Events	They won't add or share any event, live locations frequently.
Location	Fake profiles have irrelevant study and work locations.
Tagged Post	The number of tagged posts is comparatively less for fake users.
Created at	From the creation date, they use the timeline for less period of time.
Description	They make a description to advertise and connect with more number of people.
URL	The display name and URL don't match mostly.

Table.2
Description of attributes in data sets.

V. PERFOMANCE MEASURE

Efficiency = Count of correct predictions to that of total count of predictions. Percent Error = (1-Efficiency) * 100

Confusion Matrix: It is a way for summarizing the overall performance of a classification algorithm. Calculating a confusion matrix can come up with a better concept of what your category version is getting proper and what kinds of mistakes it is making.

TPR-True Positive Rate

TPR=TP/(TP+FN)

FPR- False Positive Rate

FPR=FP/(FP+TN)

TNR-True Negative Rate

TNR=TN/(FP+TN)

FNR- False Negative Rate

FNR=1-TPR

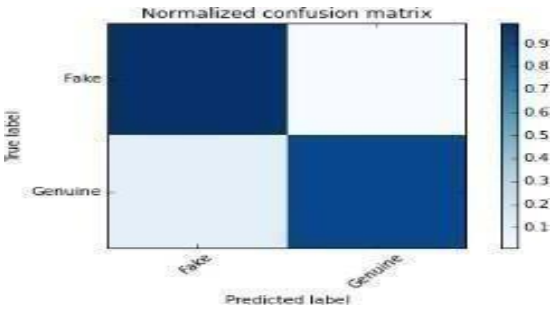
Recall – Number of the true positives were done, i.e. what number of the right hits were likewise found. Recall = TP /(TP+FN)

Precision- Precision is how many hits are returned to true positive i.e. what number of the found were right hits. Precision - TP/(TP+FP)

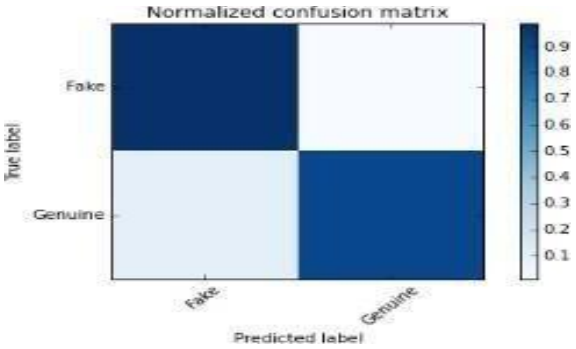
F1 score measure of accuracy for tests. It accept exactness the review p,r of the test scoring the figure. ROC Curve is the plot of FPR versus TPR. ROC used to differentiate the performance measurement of different classifying techniques.

1.Neural Networks Confusion Matrix:

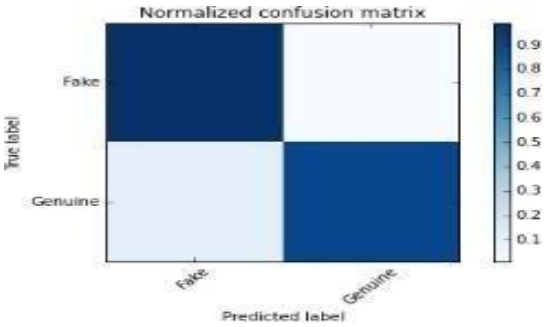
2. Random Forest Confusion Matrix:



Normalized Confusion Matrix for Neural Networks



Normalized Confusion Matrix for Random Forest

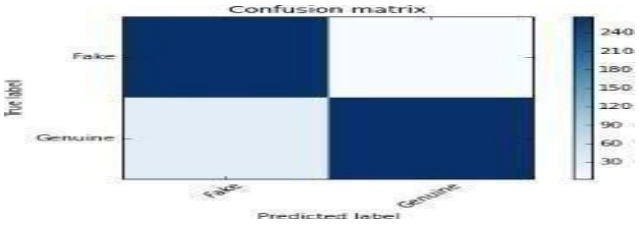


1. Confusion Matrix:

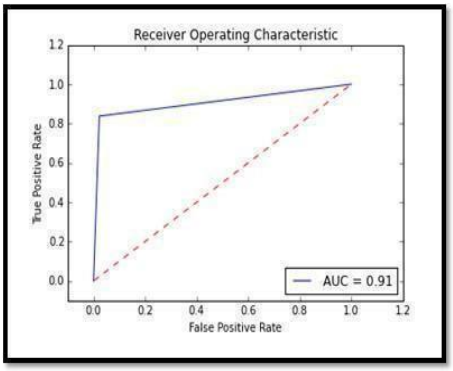
A confusion matrix is a summary of prediction outcomes on a classification problem. The number of accurate and incorrect predictions are summarized with depend values and damaged down by each elegance. that is the key to the confusion matrix.

The confusion matrix suggests the methods in which your classification model is confused while it makes predictions. It gives us perception now not only into the mistakes being made by a classifier but extra importantly the forms of mistakes which can be being made.

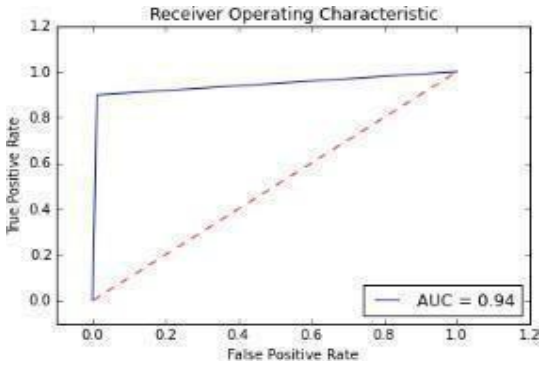
2. Random Forest Confusion Matrix:



ROC CURVE For Neural Networks:



ROC CURVE for Random Forest:



ROC CURVES:

Efficiency of neural neural network in classifying data is 91%. We have taken 80% of data for training neural network and 20% for classification.

Efficiency of random forest in classifying data is 91%. We have taken 80% of data for training random forest and 20% for classification.

V. CONCLUSION

Fake profiles are created in social networks for various reasons by individuals or groups. The results are about detecting the account is fake or genuine by using engineered features and trained using machine learning models like neural networks and random forest. The predictions indicate that the algorithm

neural network produced 93% accuracy. In the future, there is a hope that new features make to detect and identify easily like implementing skin detection can be done by using natural language processing techniques more accurate. When Facebook introduces new features then it will be easy to identify fake accounts easily.

VI. FUTURE WORK

Main problem is that a person can have multiple Facebook accounts which makes them an advantage of creating fake profiles and accounts in online social networks. The idea is of attaching Aadhar card number when signing up an account so that we can restrict to create a single account and there is no chance of fake profiles at any moment.

VII. REFERENCES

[1] Yadongzhou, Daewookkim, Junjiezhang, (Member, Ieee), Lili Liu1, Huanjin3, "(IEEE) ProGuard: Detecting Malicious Accounts in Social Network-Based Online Promotions(2019).

[2] Mauro Conti University of Padua, Radha Poovendran University of Washington, Marco Secchiero University of Padua,“ FakeBook: Detecting Fake Profiles in online Social Networks(2019)", ACM /IEEE International Conference on Advances in Social Networks Analysis and Mining

[3] Jain, Y., NamrataTiwari, S., Jain, S.: A comparative analysis of various credit card fraud detection techniques. Int. J. Recent. Technol. Eng. (2277–3878), 7(5S2), 402–407 (2019)

[4] Dr. Narsimha.G, Dr. Jayadev Gyani, P. Srinivas Rao ,"Fake Profiles Identification in Online Social Networks Using Machine Learning and NLP", International Journal of Applied Engineering Research ISSN 0973-4562, Number 6, Volume 13.

[5] D. Rajeswara Rao & V. Pellakuri. Training and development of artificial neural network models: Single layer feedforward and multi layer feedforward neural network(2019). Journal of Theoretical and Applied Information Technology, 150-156,84(2).

[6] Challa, N.,. Pasupuleti, S. K, & Chandra, J. V. A practical approach to E-mail spam filters to protect data from advanced persistent threat.(2018) Paper presented at the Proceedings of IEEE International Conference on Circuit, Power and Computing Technologies.

[1] D. Rajeswara Rao , & P. Vidyullatha. Machine learning techniques on multidimensional curve fitting data based on r_ square and chi_square methods(2019). International Journal of Electrical and Computer Engineering

[2] Pradeepini, G., Patil, S. T. , & Bangare, S. L (2019). Brain tumor classification using mixed method approach. Paper presented at the International Conference on Information Communication and Embedded Systems, ICICES

[3] Kutub Thakur; Thaier Hayajneh; Jason Tseng, Cyber Security in Social Media: Challenges and the Way Forward(2019).

[10] Jianhui Qiu; Xingfa Shen; Yan Guo; Jian Yao; Renzhe Fang, Detecting Malicious Users

in Online Dating Application IEEE (2019)

[11] Reddy, A. V. N., & Phanikrishna, C. Contour tracking based knowledge extraction and object recognition using deep learning neural networks(2019).

[12]Sucharitha, G., &Senapati, R. K. Local extreme edge binary patterns for face recognition and image retrieval(2018). Journal of Advanced Research in Dynamical and Control Systems_10, 644-654

[13] Koh, Pang Wei, et al. "Concept bottleneck models." Proceedings of the 37th International Conference on Machine Learning (2020).

[14] Mothilal, Ramaravind K., Amit Sharma, and Chenhao Tan. "Explaining machine learning classifiers through diverse counterfactual explanations." Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 2020

[15] Su, Jiawei, Danilo Vasconcellos Vargas, and Kouichi Sakurai. "One pixel attack for fooling deep neural networks." IEEE Transactions on Evolutionary Computation (2019).

CHAPTER 13

STUDENTS PROFILE

S Raja Lingam Seth is currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. He completed his intermediate from Sri Chaitanya Junior College and 10th standard from Triveni Talent School. His technical skills include C, C++, Python, Java and MySQL. He also has a good grip over machine learning. His participations include "2 – day machine learning workshop" conducted in St. Martin's Engineering College in the month of February in 2019 and a 3-day online workshop on "AI & ML in Speech and Audio Processing" conducted by St. Martins Engineering College in the month of December in 2020. He has worked on various projects such as Contacts Management System, Mini Assistant using python as a part of the academics. He has also completed various certification courses from platforms such as Udemy, Coursera and Sololearn.

Anshuman Singh is currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College.He has completed his 12^{th} and 10^{th} from KV AFS Bidar. His technical skills include C, C++, Python, Java and MySQL. His participations include "2 – day machine learning workshop" conducted in St. Martin's Engineering College in the month of February in 2019 and a 3-day online workshop on "AI & ML in Speech and Audio Processing" conducted by St. Martins Engineering College in the month of December in 2020.

Md Afroz is Currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martins Engineering College. He Completed his intermediate from Narayan Junior College and 10^{th} from Father's Model High School. His Technical Skills include C, Python and basic java. His Participations include "2 day Machine Learning Workshop conducted in St. Martins Engineering College in the month of February in 2019.He has worked on one such project is TAM(application) as a part of academics.

Rohan Mahesh Katkam is currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. He has completed his Diploma from Larsen and Toubro Institute of Technology, Mumbai University and his schooling from Andhra Education Society High School. His technical skills include C, C++, Python, Java and MySQL. He has worked on MEAN stack, SEO technologies from Allure Medical Spa Company as Intern for one year. His participations include "2 – day machine learning workshop" conducted in St. Martin's Engineering College in the month of February in 2019. He has also completed various certification courses from platforms such as Udemy, Coursera and Sololearn.

Asad Ul Islam is currently pursuing his bachelor of technology in the stream of Computer Science and Engineering at St. Martins Engineering College. he completed his diploma from Government Polytechnic Warangal and 10th standard from SR High School. His technical skills are Python, C, C++, MySQL and HTML. His participations include 3-day online workshop on "AI & ML in Speech and Audio Processing" conducted by St. Martins Engineering College in the month of December in 2020.

A

PROJECT REPORT

On

BIRD SPECIES IDENTIFICATION USING DEEP LEARNING

Submitted by

1) M.Aishwarya (17K81A05F8) 2) Y.SreeVishnuPriya (17K81A05J0)

3) P.Beulah (17K81A05G8) 4) G.HariTeja (17K81A05E4)

In partial fulfillment for the award of the

degreeof

BACHELOR OF TECHNOLOGY

IN

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Under the Guidance of

N.KrishnaVardhan

ASSISTANT PROFESSOR

B.TECH M.TECH

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ST.MARTIN'S ENGINEERING COLLEGE
An Autonomous Institute

Dhulapally, Secunderabad – 500 100

JUNE 2021

BONAFIDE CERTIFICATE

This is to certify that the project entitled Bird Species Identification Using Deep Learning, is being submitted by 1.**Ms. Mannelli Aishwarya 17K81A05F8, 2.Ms.Yamanamada SreeVishnuPriya 17K81A05J0, 3. Ms. Pasam Beulah 17K81A05G8, 4. Mr. Galinki HariTeja 17K81A05E2** in partial fulfillment of the requirement for the award of the degree of **BACHELOR OF TECHNOLOGY IN Computer Science Department** is recorded of bonafide work carried out by them. The result embodied in this report have been verified and found satisfactory.

<Signature>
N.KrishnaVardhan
Department of CSE

Head of the Department
Dr.M.NARAYANAN
Department of CSE

Internal Examiner

External Examiner

Place:

Date:

DECLARATION

We, the student of **Bachelor of Technology** in Department of Computer Science and Engineering', session: 2017 – 2021, St. Martin's Engineering College, Dhulapally, Kompally, Secunderabad, hereby declare that work presented in this Project Work entitled Bird Species Identification Using Deep Learning is the outcome of our own bonafide work and is correct to the best of our knowledge and this work has been undertaken taking care of Engineering Ethics. This result embodied in this project report has not been submitted in any university for award of any degree.

Ms. Mannelli Aishwarya	17K81A05F8
Ms.YamanamandaSreeVishnupriya	17K81A05J0
Ms. Pasam Beulah	17K81A05G8
Mr. Galinki Hariteja	17K81A05E2

ABSTRACT

The human knowledge over the bird species isn't enough to identify a species of bird accurately, as it requires lot of expertise in the field of Ornithology. We can automate this task. The existing system takes inputs in the form of an audio or video. However the inputs in the form of audio or video may not give accurate results due to some disturbances involved in it. So, an approach to classify bird using an image over audio or video is preferred. The proposed system works on the principle based on detection of a part and extracting CNN features from multiple convolutional layers. These features are aggregated and then given to the classifier for classification purpose. On basis of the results which has been produced, the system has provided the 80% accuracy in prediction of finding bird species.

Birds are the warm-blooded vertebrates constituting of class Aves, there are nearly 10 thousand living species of birds in the world with multifarious characteristics and appearances. Bird watching is often considered to be an interesting hobby by human beings in the natural environment. The human knowledge over the species isn't enough to identify a species of bird accurately, as it requires lot of expertise in the field of Ornithology. This project provides an automated model based on the deep neural networks which automatically identifies the species of a bird given as the test data set.

The model was trained and tested for 20 species of birds with the total images 7637 and 1853 images for train and test respectively and the model has shown a promising accuracy of 98% when tested with the test datasets.

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible and whose encouragements and guidance have crowded effects with success.

We extended our deep sense of gratitude to Principal**,** **Dr. P. SANTOSH KUMAR PATRA**, St. Martin's Engineering College, Dhulapally, for permitting us to undertake this project.

We are also thankful to **Dr. M.NARAYANAN**, Head of the Department, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

We would like to thank **Dr. T. POONGOTHAI,** Department of Computer Science and Engineering (AI & ML) for her encouragement and insightful comments and as well as our project coordinators **Dr. B.RAJALINGAM**, Associate Professor and **Dr. N. SATHEESH**, Professor, in Department of Computer Science and Engineering for their valuable support.

We would like to express our sincere gratitude and indebtedness to our project supervisor N.KrishnaVardhan, Assistant Professor, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

Finally, we express thanks to all those who have helped us successfully to completing this project. Furthermore, we would like to thank our family and friends for their moral support and encouragement.

We express thanks to all those who have helped us in successfully completing the project.

Ms. Mannelli Aishwarya	17K81A05F8
Ms.YamanamandaSreeVishnupriya	17K81A05J0
Ms. Pasam Beulah	17K81A05G8
Mr. Hariteja	17K81A05E2

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

CNN	Convolution Neural Network
DCCN	Deep Convolution Neural Network
SIFT	Scale invariant feature transform
MFCC	Mel frequency cepstral centroids
HSV	Hue Saturation Value

1. INTRODUCTION

Bird watching is one of the recreational activity that pro- videos more relaxation and enjoyment to the minds of human beings. Identification of bird species becomes challenging task due to the interclass and interclass similarities existing in between them. Based on the physical characteristics, colour, and shape the bird species are categorized into different classes. Due to the observer constraints such as location, distance, and equipment used to identify birds, recognizing birds with the naked eye of human being is based on basic characteristic features, and appropriate classification based on distinct features is often seen as tedious [1]. Nowadays a number of techniques are available to identify the species of birds. Deep learning is one of the emerging technology that can be used to recognize the birds. The convolutional neural network (CNN) is a category of deep learning neural networks. Convolutional neural networks constitute a big step forward in image recognition. They're maximum normally used to investigate visual imagery and are regularly operating behind the scene in image classification. In the past, computer vision [2], [3] and its subcategory of recognition, which use strategies along with machine learning, had been notably researched to delineate the specific features of objects, consisting of veggies and fruits, landmarks, clothing, cars, plants, and birds, inside a selected cluster of scenes.

The capability of convolutional neural networks to extract various features from captured image is utilized in many scenarios. In the past years, bird sound classification has received attention increasingly. Therefore, it is becoming ever more necessary to protect bird biodiversity, where monitoring bird population is the first step for the protection. CNN is mainly designed to recognize visual features from images and it requires a minimum level of pre-processing. In addition, the exponentially expanded amount of online data has gotten to be less demanding to gather as the learning information for the neural networks, and the refined information has been easily shared for the convolutional neural network learning. These intuitive driven to a convolutional neural network approach beyond the capabilities of the existing approaches. The convolutional neural network has become one of the leading architecture for most of the image recognition, classification, and prediction processes.

In image recognition, video analysis, natural language processing, and drug discovery applications the convolutional neural network can be used for better performances. And the CNNs performance rates are progressing yearly. In this proposed work we are making use of CNNs in the field of bird image recognition. From the captured image of the bird the CNN model could identify the species of that particular bird and produces prediction results accordingly. The main objectives of the proposed work are to develop a deep learning model by making use of train and test colored images of birds in order to identify/ classify the bird species into particular classes of its species according to the classification results obtained from the skip connection oriented CNN.

1.1 PROJECT OVERVIEW

This project uses concept to identify species of birds by using python TENSORFLOW and Deep Learning algorithm. Earlier technique were using birds voice or videos to predict it species but this technique will not give accurate result as audio may contains background or other animal voices. So images can be best option to identify species of birds. Deep Learning is a subset of machine learning comprising of various algorithms and was inspired by the human neural networks, the algorithms imitates the working of human brains in processing of the data and produces a pattern of data for decision making. This project gives an approach of convolution neural network model for identifying the species of birds.In this project, instead of recognizing a large number of disparate categories, the problem of recognizing a large number of classes within one category is investigated that of birds.

Classifying birds pose an extra challenge over categories, because of the large similarity between classes. In addition, birds are non-rigid objects that can deform in many ways, and consequently there is also a large variation within classes. Previous work on bird classification has deal with a small number of classes, or through voice.

To implement this technique we need to train all birds species and generate a model and then by uploading any image deep learning algorithm will convert uploaded image into gray scale format and apply that image on train model to predict best match species name for uploaded image.

To develop such system a trained dataset is required to classify an image. Trained dataset consists of two parts trained result and test result. The dataset has to be retrained to achieve higher accuracy in identification using retrain.py in Google collab. The training dataset is made using 50000 steps taking into consideration that higher the number of steps higher is its accuracy. The accuracy of training dataset is 93%. The testing dataset consists of nearly 1000 images with an accuracy of 80%. Dataset is validated with an accuracy of 75% to increase the performance of system. Whenever a user will upload an input file on website, the image is temporarily stored in database. This input file is then feed to system and given to CNN where CNN is coupled with trained dataset.

A CNN consists of various convolutional layers. Various alignments/features such as head, body, color, beak, shape, entire image of bird are considered for classification to yield maximum accuracy. Dataset is validated with an accuracy of 75% to increase the performance of system. This study developed a platform that uses deep learning for image processing to identify bird species from digital images uploaded by an end-user. The proposed system could detect and differentiate uploaded images as birds. With an overall accuracy is high for the training dataset using CNN model. This project ultimately aimed to design an automatic system for differentiating among bird images with shared fundamental characteristics but minor variations in appearance.

1.2 PROJECT OBJECTIVES

Identification of species requires the assistance of manual bird books. So, it also requires expertise in the field to identify the species accurately. Few Species of Birds look very familiar in their appearances thus identifying the exact species by humans may be error prone. The main aim of the proposed work is to develop an automated model which has capability of identifying the species of the bird where bird image is given as a test image from the dataset The main objectives are to develop an automated model by making use of train and test colored images of birds in order to identify/ classify the bird species to particular class of its species. From the captured image of the bird the CNN model could identify the species of that particular bird and produces prediction results accordingly.

The main objectives of the proposed work are to develop a deep learning model by making use of train and test colored images of birds in order to identify and classify the bird species into particular classes of its species according to the classification results obtained from the skip connection oriented CNN model. To improve the accuracy by 80-90% in identifying the birds and classifying according to their species to improve the performance of overall system.

2. LITERATURE SURVEY

The earlier approaches for the species identification involved the bird songs where audio feature extraction was based on marsyas framework [1] and the classical Machine learning algorithms for classification, the visual features i.e. SIFT (Scale invariant feature transform) [2] from bird images and acoustic features both were used to train a standard SVM for classification. The fine-grained visual categorization [3] have shown great results of classification. The trajectory features, turn based features and shape movement features and wing beat frequencies [4] were considered from the video captures of bird moments and a combined naive Bayes classifier and SVM was used. The MFCC (Mel frequency cepstral centroids) [5] formed a feature matrix for class model and SVM was used to test the samples. The mean standard deviation and skewness of the RGB planes [6] of bird images have helped in classifying the species. The ratio of distance of eye to the root of beak and the distance of width of the beak were also primarily considered for classification. An HSV model [7] (which is a combination of RGB and CMY) features were considered for color-based species identification. A transfer learning-based method with multistage learning [8] was used to mine both micro and macro level features from the bird images for classification in the recent years.

2.1 SURVEY ON BACKGROUND

Basically bird identification is done visually or acoustically. The main visual components comprise of birds shape, its wings, size, pose, color, etc. However, while considering the parameters time of year must be taken into consideration because birds wings changes according to their growth. The acoustics components comprise the songs and call that birds make [7]. The marks that distinguish one bird from another are also useful, such as breast spots, wing bars which are described as thin lines along the wings, eye rings, crowns, eyebrows. The shape of the beak is often an important aspect as a bird can recognized uniquely. The characteristics of bird such as shape and posture are the mostly used to identify birds. Mostly experts can identify a bird from its silhouette because this characteristic is difficult to change. A bird can also be differentiated using its tail. The tail can be recognized in many ways such as notched, long and pointed, or rounded. Sometimes legs are also used for recognizing an image in format long, or short [10].

By considering a single parameter will not yield an accurate result. So, multiple parameters are to be considered in order to get appropriate output. The size of a bird in an image varies depending upon factors such as the resolution, distance between the birds and the capturing device, and the focal distance of the lens. Therefore, based on a practical observation for large number of images, images are differentiated on the basis of color which consists of various pixel. In depth it is found that greater the image quality greater is its accuracy.

The automatic bird species identification for bird images project present a series of comparison conducted in a CUB- 200 dataset composed of more than 6,000 images with 200 different category [6]. In this paper, they have considered two different color spaces, RGB and HSV, and a different number of species to be classified. If the image consists of more than 70% of the pixels the accuracy of output was ranging from 8.82% to 0.43%.

2.2 CONCLUSIONS ON SURVEY

Classifying birds pose an extra challenge over categories, because of the large similarity between classes. Also, birds are flexible objects that can disfigure in many ways, and at the same time there is also a large variation within classes. So, the current study finds a method to identify the bird species using Deep learning algorithm on the dataset for classification of image. It has 200 categories or 11,788 photos. The generated system is connected with a user-friendly website where user will upload photo and it will give the desired output. The proposed model works on the principle of detection of a part and extracting CNN feature from multiple convolutional layers. These features are extracted and then given to the classifier for classification purpose. CNN architectures consist of convolutional and pooling layers.

CNN architecture is used for image classification task. An image, input to the network, is followed by convolution and pooling. Thereafter, these operations feed one or more fully connected layers. At last fully connected layer outputs the class label. First, raw input data of numerous semantic parts of a bird were gathered and localized. Second, the features of each generic part were detected and filtered based on shape, size, and color. Third, a CNN model was trained with the pictures in a graphics processing unit for feature extraction with the previously mentioned characteristics, and the classified, trained data were stored in a server to target object. Then, information obtained from an image uploaded by an end-user, captured using a camera, can be navigated to retrieve information and predict from the trained model.

3. HARDWARE AND SOFTWARE REQUIREMENTS

3.1 HARDWARE REQUIREMENTS

- Operating System supported by
    1. Windows 7
    2. Windows XP
    3 .Windows 8
- Processor: Pentium IV or higher
- RAM : 256 MB
- Space on Hard Disk : Minimum 512 MB

3.2 SOFTWARE REQUIREMENTS

- For developing the Application

1. Python
2. Django
3. MYSQL
4. MySQL client
5. WampServer 2.4

- Technologies and Languages used to Develop : Python

4. SOFTWARE DEVELOPMENT ANALYSIS

4.1 OVERVIEW OF PROBLEM

Bird behaviour and population trends have become an important issue now a days. Birds help us to detect other organisms in the environment (e.g. insects they feed on) easily as they respond quickly to the environmental changes. But, gathering and collecting information about birds requires huge human effort as well as becomes a very costlier method. In such case, a reliable system that will provide large scale processing of information about birds and will serve as a valuable tool for researchers, governmental agencies, etc. is required. So, bird species identification plays an important role in identifying that a particular image of bird belongs to which species. The identification can be done through image, audio or video. An audio processing technique makes it possible to identify by capturing the audio signal of birds. But, due to the mixed sounds in environment such as insects, objects from real world, etc. processing of such information becomes more complicated. Usually, human beings find images more effective than audios or videos. So, an approach to classify bird using an image over audio or video is preferred.

4.2DEFINE THE PROBLEM

It describes a concept to identify species of birds by using python tensor flow and Deep Learning algorithm. Earlier technique were using birds voice or videos to predict it species but this technique will not give accurate result as audio may contains background or other animal voices. So images can be best option to identify species of birds. Unfortunately a number of challenges have made this task extremely difficult to tackle.

Most prominent are:

a. Background noise
b. Multiple birds singing at the same time (multi-label)
c. Difference between mating calls and songs
d. Inter-species variance
e. Variable length of sound recordings
f. Large number of different species

Because of these, most systems are developed to deal with only a small number of species and require a lot of re-training and fine-tuning for each new species.

4.3 MODULES OVERVIEW

There are 4 types of modules

1. Data collection
2. Data Pre-Processing
3. Feature Extraction
4. Evaluation Model

4.4 DEFINE THE MODULES

4.4.1 DATA COLLECTION

Data collection is defined as the procedure of collecting, measuring and analyzing accurate insights for research using standard validated techniques. A researcher can evaluate their hypothesis on the basis of collected data. In most cases, data collection is the primary and most important step for research, irrespective of the field of research.

4.4.2 DATA PRE-PROCESSING

Data pre-processing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, lacking in certain behaviours or trends, and is likely to contain many error. Data pre-processing is a proven method of resolving such issues.

Three common data pre-processing steps are:

- Formatting
- Cleaning
- Sampling

4.4.3 FEATURE EXTRACTION

Feature extraction involves reducing the number of resources required to describe a large set of data. When performing analysis of complex data one of the major problems stems from the number of variables involved. Analysis with a large number of variables generally requires a large amount of memory and computation power, also it may cause classification algorithm to overfit to training samples and generalize poorly to new samples. Feature extraction is a general term for methods of constructing combinations of the variables to get around these problems while still describing the data with sufficient accuracy. Many machine learning practitioners believe that properly optimized feature extraction is the key to effective model construction.

4.4.4 EVALUATION MODULE

Model Evaluation is an integral part of the model development process. It helps to find the best model that represents our data and how well the chosen model will work in the future.

4.5 MODULE FUNCTIONALITY

4.5.1 DATA COLLECTION

Data used in this paper is a set of product reviews collected from credit card transactions records. This step is concerned with selecting the subset of all available data that you will be working with. ML problems start with data preferably, lots of data (examples or observations) for which you already know the target answer. Data for which you already know the target answer is called labelled data.

4.5.2 DATA PRE-PROCESSING

The data you have selected may not be in a format that is suitable for you to work with. The data may be in a relational database and you would like it in a flat file, or the data may be in a proprietary file format and you would like it in a relational database or a text file. Cleaning: Cleaning data is the removal or fixing of missing data. There may be data instances that are incomplete and do not carry the data you believe you need to address the problem. These instances may need to be removed. Additionally, there may be sensitive information in some of the attributes and these attributes may need to be removed from the data entirely. Sampling: There may be far more selected data available than you need to work with. More data can result in much longer running times for algorithms and larger computational and memory requirements. You can take a smaller representative sample of the selected data that may be much faster for exploring and prototyping solutions before considering the whole dataset.

4.5.3 FEATURE EXTRACTION

Feature extraction is an attribute reduction process. Unlike feature selection, which ranks the existing attributes according to their predictive significance, feature extraction actually transforms the attributes. The transformed attributes, or features, are linear combinations of the original attributes. Finally, our models are trained using Classifier algorithm. We use classify module on Natural Language Toolkit library on Python. We use the labeled dataset gathered. The rest of our labeled data will be used to evaluate the models. Some machine learning algorithms were used to classify pre-processed data. The chosen classifiers were Random forest. These algorithms are very popular in text classification tasks.

4.5.4 EVALUATION MODEL

Evaluating model performance with the data used for training is not acceptable in data science because it can easily generate overoptimistic and over fitted models. There are two methods of evaluating models in data science, Hold-Out and Cross-Validation. To avoid over fitting, both methods use a test set (not seen by the model) to evaluate model performance. Performance of each classification model is estimated base on its averaged. The result will be in the visualized form. Representation of classified data in the form of graphs. Accuracy is defined as the percentage of correct predictions for the test data. It can be calculated easily by dividing the number of correct predictions by the number of total predictions.

5. PROJECT SYSTEM DESIGN

5.1 E-R DIAGRAMS

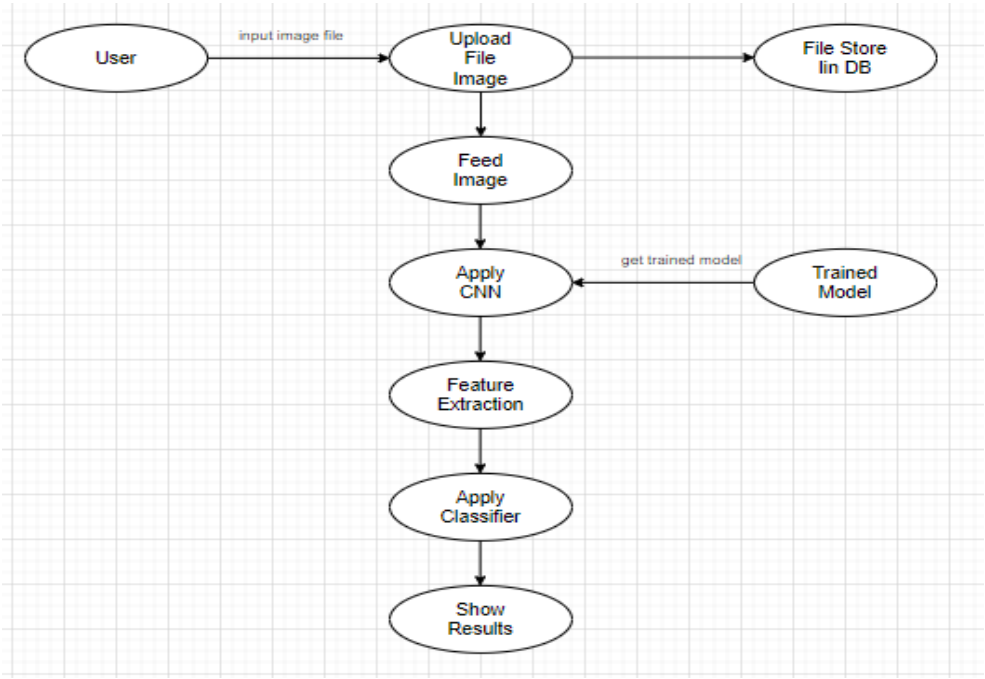


Fig 5.1.1- Workflow diagram.

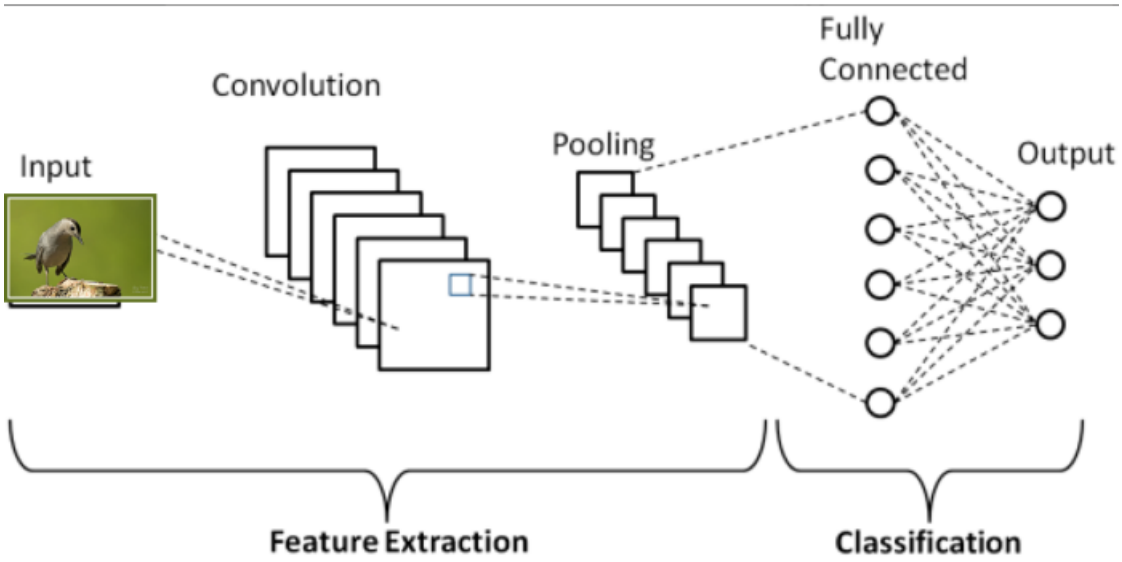


Fig 5.1.2-The Convolution neural network model for bird's species classification.

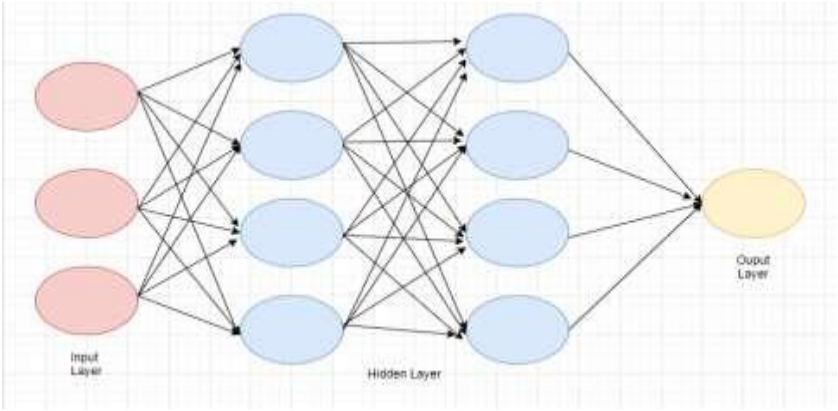


Fig 5.1.3 - Neural networks.

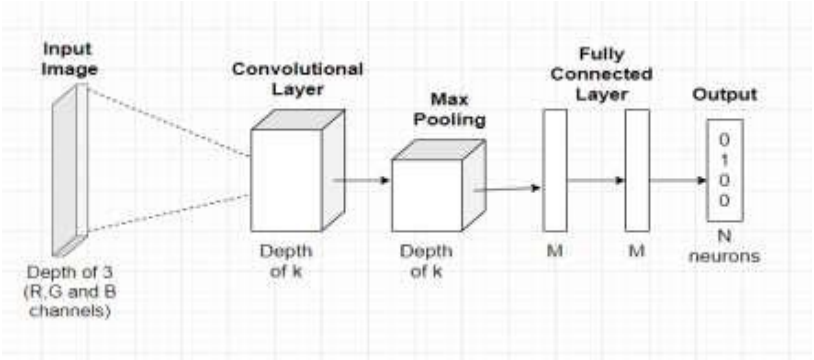


Fig 5.1.4 - Schematic representation of the architecture of Convolutional Neural Network Related Work

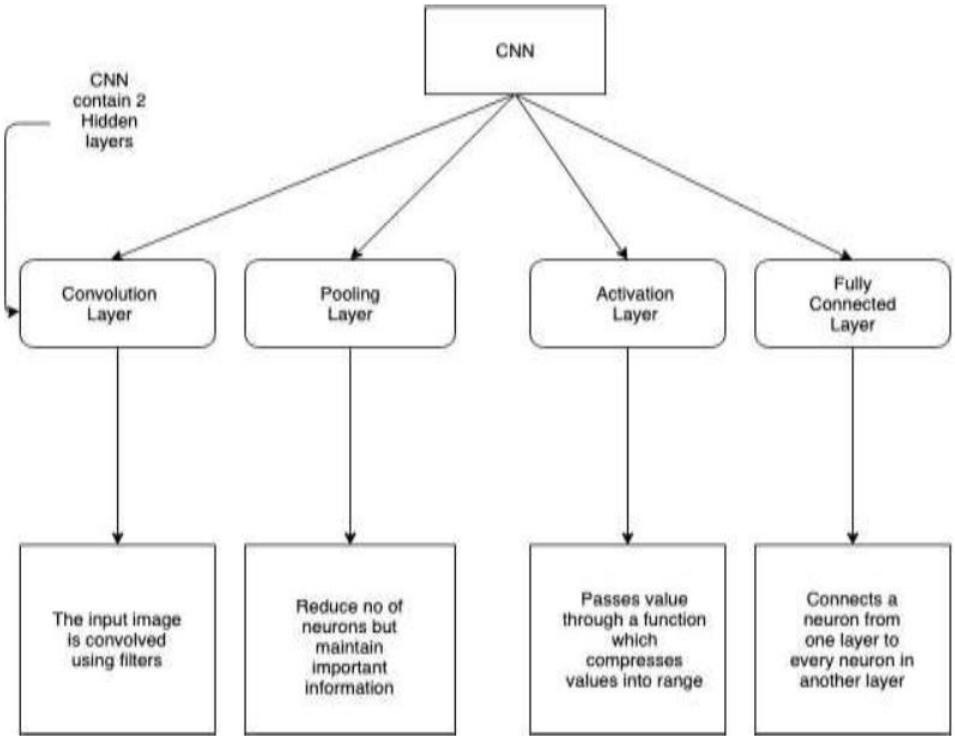


Fig 5.1.5 - CNN block diagram.

5.2 UML DIAGRAMS

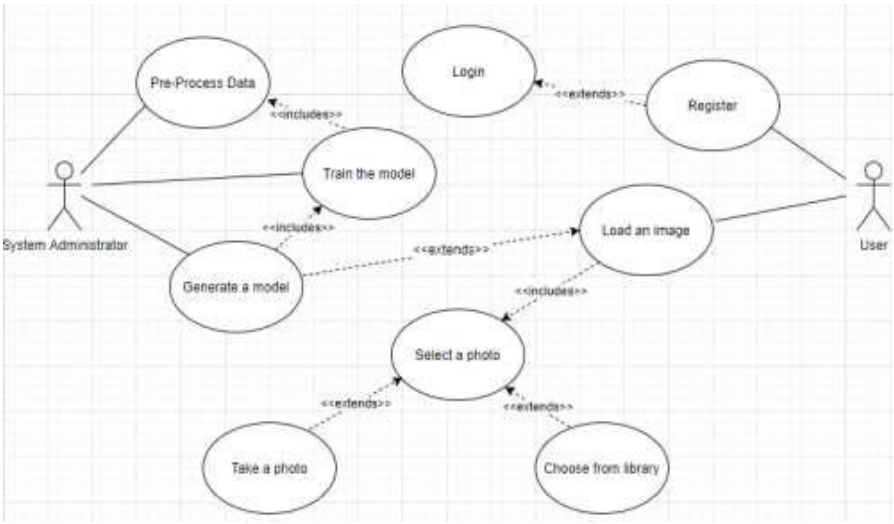


Fig 5.2.1- Use case diagram.

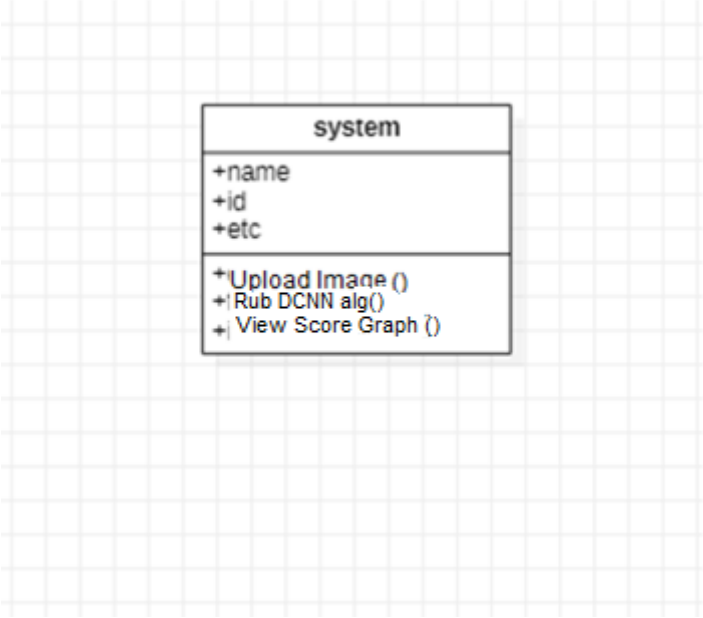


Fig 5.2.2 – Class diagram

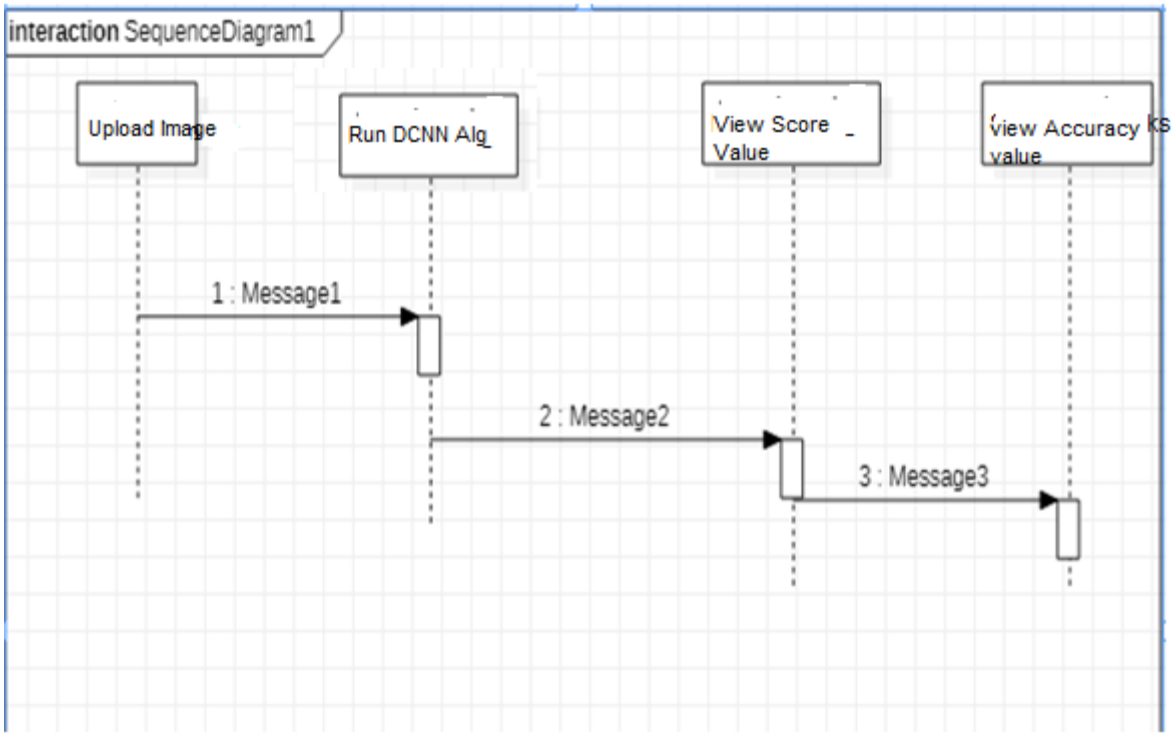


Fig 5.2.3 – Sequence diagram.

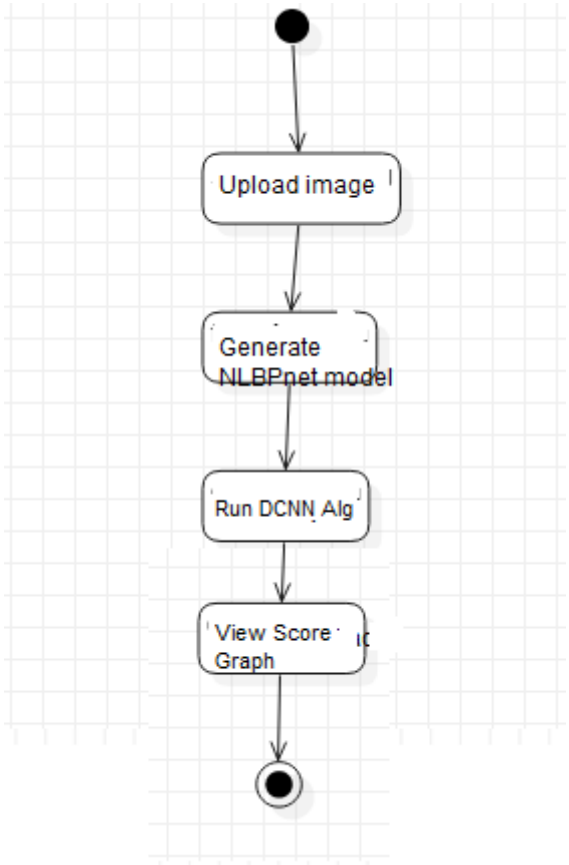


Fig 5.2.4 – Activity diagram

6 PROJECT CODING

6.1 CODETEMPLATES

```
        def DCNN():
                name = os.path.basename(filename)
                arr = name.split(".")
                kdt = KDTree(features_train, leaf_size=30, metric='euclidean')
                K = 5
                q_ind = int(arr[0])
                box = bounding_box[val_list[q_ind][1]+1].split(" ")
                print(features_val[q_ind:q_ind+1])
                dist, ind = kdt.query(features_val[q_ind:q_ind+1], k=K)
                print('Query image from validation set:')
                I = io.imread(filename)
                fig,ax = plt.subplots(1)
                plt.axis('off')
                plt.imshow(I)
                plt.suptitle("query image : "+os.path.basename(filename), fontsize=10)
                #rect =
        patches.Rectangle((float(box[1]),float(box[2])),float(box[3]),float(box[4]),linewidth=1,edgecolor='r',fac
        ecolor='none')
                #ax.add_patch(rect)
                plt.show()


        values = []
        x = []
        for i in range(K):
        plt.figure(figsize=(30,30))
        plt.subplot(1, K, i+1)
        values.append(0.5+1)
        I = io.imread(train_list[ind[0,i]][0])
        '''
        train_list[ind[0,i]][0] - img
        '''
        img_name = train_list[ind[0,i]][0]
        x.append(img_name[47:72])
        print(img_name[47:72])
        box = bounding_box[train_list[ind[0,i]][1]+1].split(" ")
        fig,ax = plt.subplots(1)
        plt.axis('off')
        plt.imshow(I)
        plt.suptitle(os.path.basename(train_list[ind[0,i]][0]), fontsize=10)
        rect =
patches.Rectangle((float(box[1]),float(box[2])),float(box[3]),float(box[4]),linewidth=1,edgecolor='r',fac
ecolor='none')
        ax.add_patch(rect)
        plt.show()
```

6.2 OUTLINE FOR VARIOUS FILES

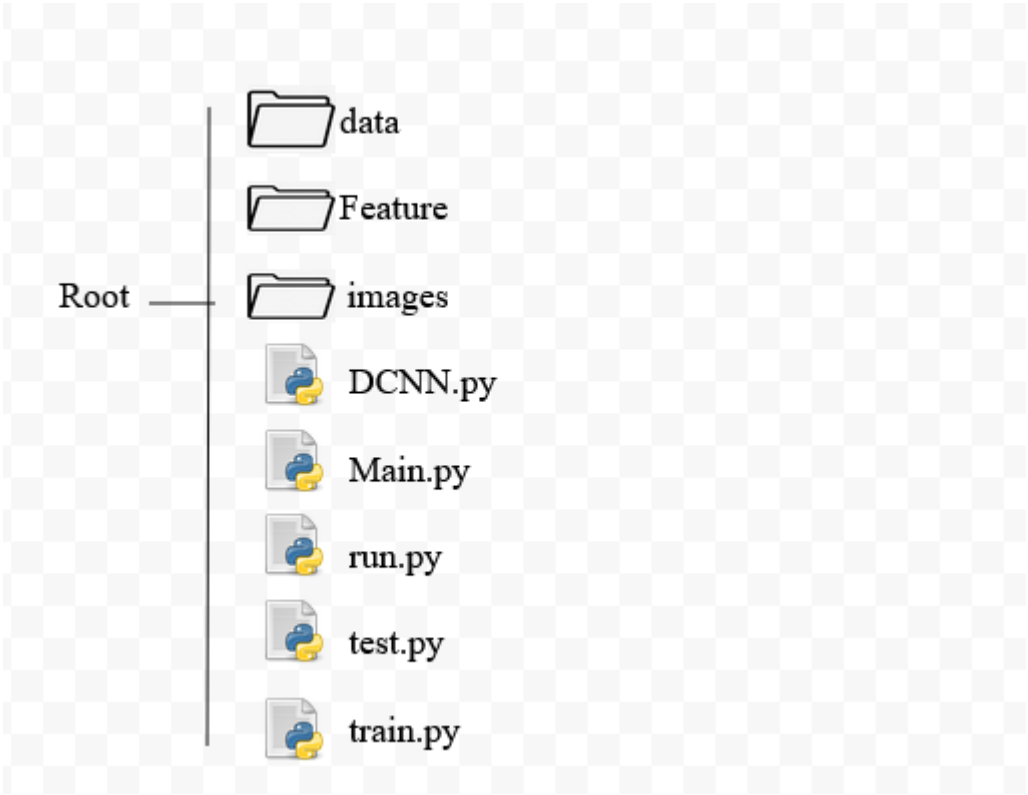


Fig 6.2.1 Outline for Various Files

6.3 CLASS WITH FUNCTIONALITY

1. tkinter.TK (*screenName=None,baseName=None,className='Tk', useTk=1*)

TheTKclass is instantiated without arguments. This creates a toplevel widget of Tk which usually is the main window of an application. Each instance has its own associated Tcl interpreter.

2. Numpy

NumPy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays. It is the fundamental package for scientific computing with Python. It contains various features including these important ones:

- A powerful N-dimensional array object
- Sophisticated (broadcasting) functions
- Tools for integrating C/C++ and Fortran code
- Useful linear algebra, Fourier transform, and random number capabilities

3. Matplotlib

Matplotlib is a cross-platform, data visualization and graphical plotting library for Python and its numerical extension NumPy. As such, it offers a viable open source alternative to MATLAB. Developers can also use matplotlib's APIs (Application Programming Interfaces) to embed plots in GUI applications.

4. Sklearn

Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python. This library, which is largely written in Python, is built upon NumPy, SciPy and Matplotlib.

6.4 METHODS INPUT AND OUTPUT PARAMETERS

A **Convolutional Neural Network (ConvNet/CNN)** is a Deep Learning algorithm which can take in an input image, assign importance (learnable weights and biases) to various aspects/objects in the image and be able to differentiate one from the other. The pre-processing required in a ConvNet is much lower as compared to other classification algorithms. While in primitive methods filters are hand-engineered, with enough training, ConvNets have the ability to learn these filters/characteristics.

7 PROJECT TESTING

7.1 VARIOUS TEST CASES

Test Case 1

Fig 7.1.1 Input 1

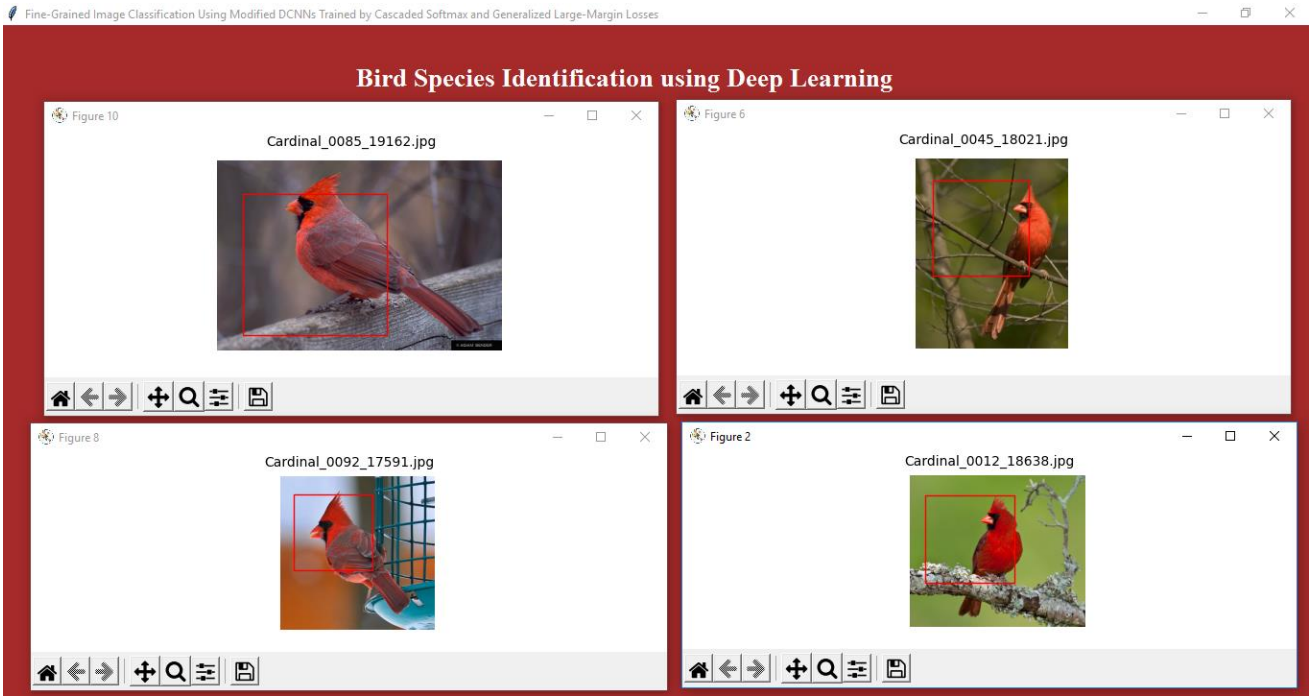


Fig 7.1.2 Output for input1.

Test Case 2

Fig 7.2.1 Input 2

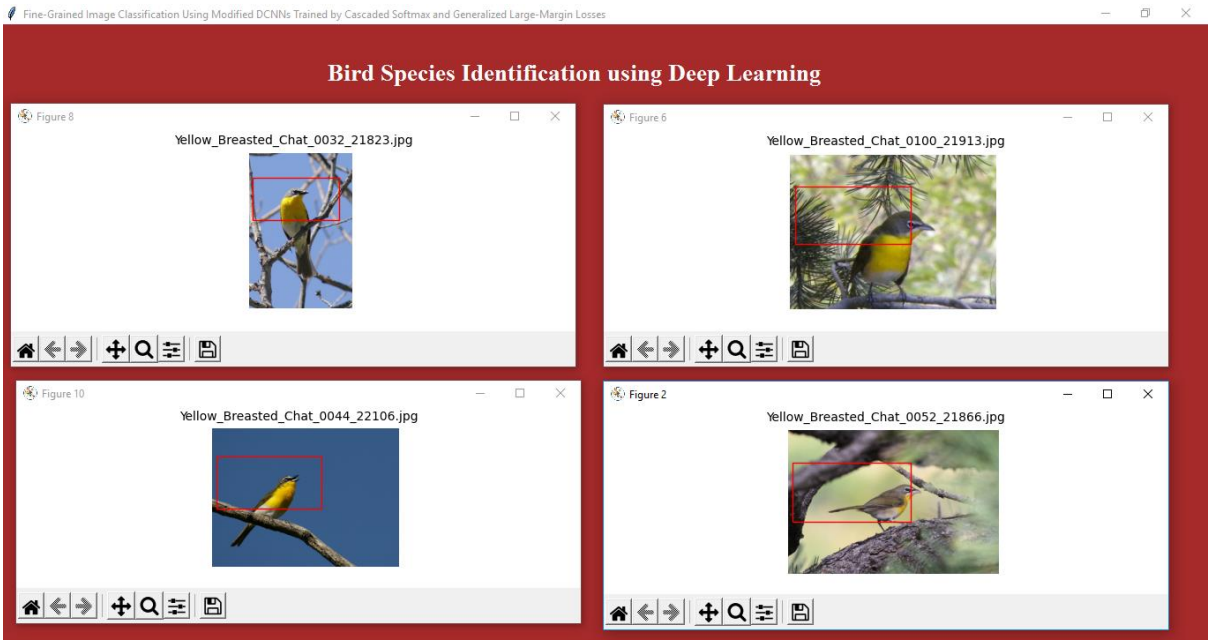


Fig 7.2.2 Output 2

7.2 BLACK BOX TESTING

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box .you cannot "see" into it. The test provides inputs and responds to outputs without considering how the software works.

Black-Box can be any software system you want to test. For Example, an operating system like Windows, a website like Google, a database like Oracle or even your own custom application. Under Black

Box Testing, you can test these applications by just focusing on the inputs and outputs without knowing their internal code implementation.

7.3 WHITE BOX TESTING

White Box Testing is a testing in which in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is purpose. It is used to test areas that cannot be reached from a black box level. The term "White Box" was used because of the see-through box concept. The clear box or White Box name symbolizes the ability to see through the software's outer shell (or "box") into its inner workings. Likewise, the "black box" in “Black box testing” symbolizes not being able to see the inner workings of the software so that only the end-user experience can be tested.

8 OUTPUT SCREENS

8.1 USER INTERFACES

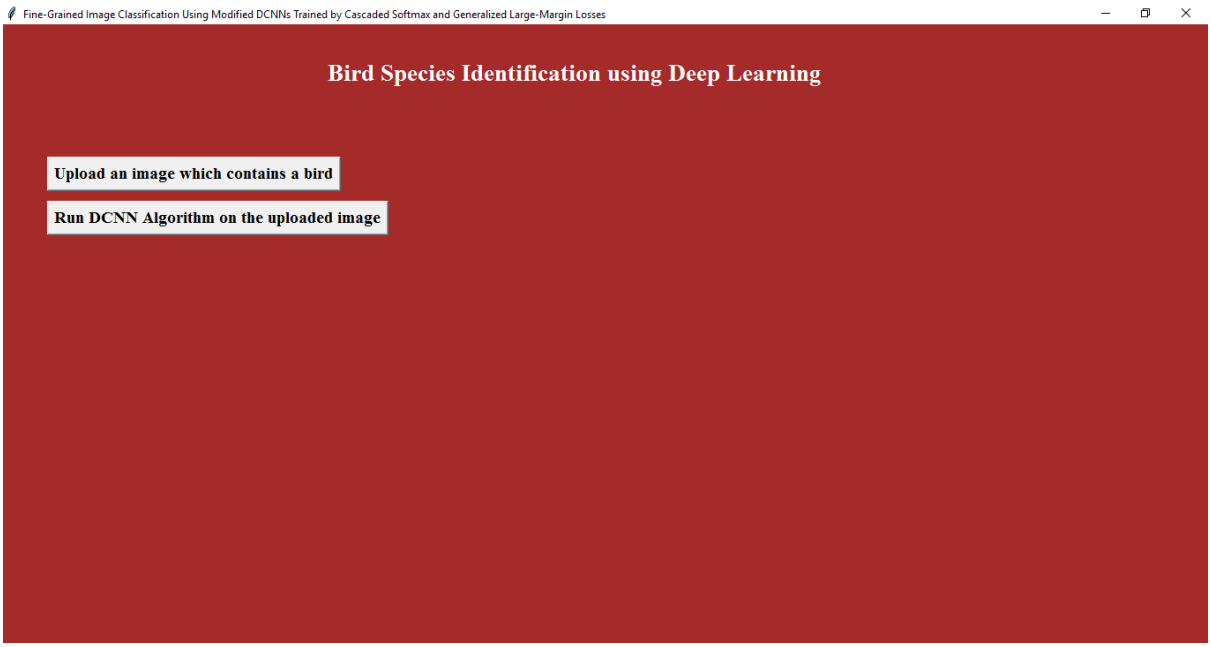


Fig 8.1.1 Home Screen.

In above screen click on ‘Upload an image which contains a bird’ button to upload input image

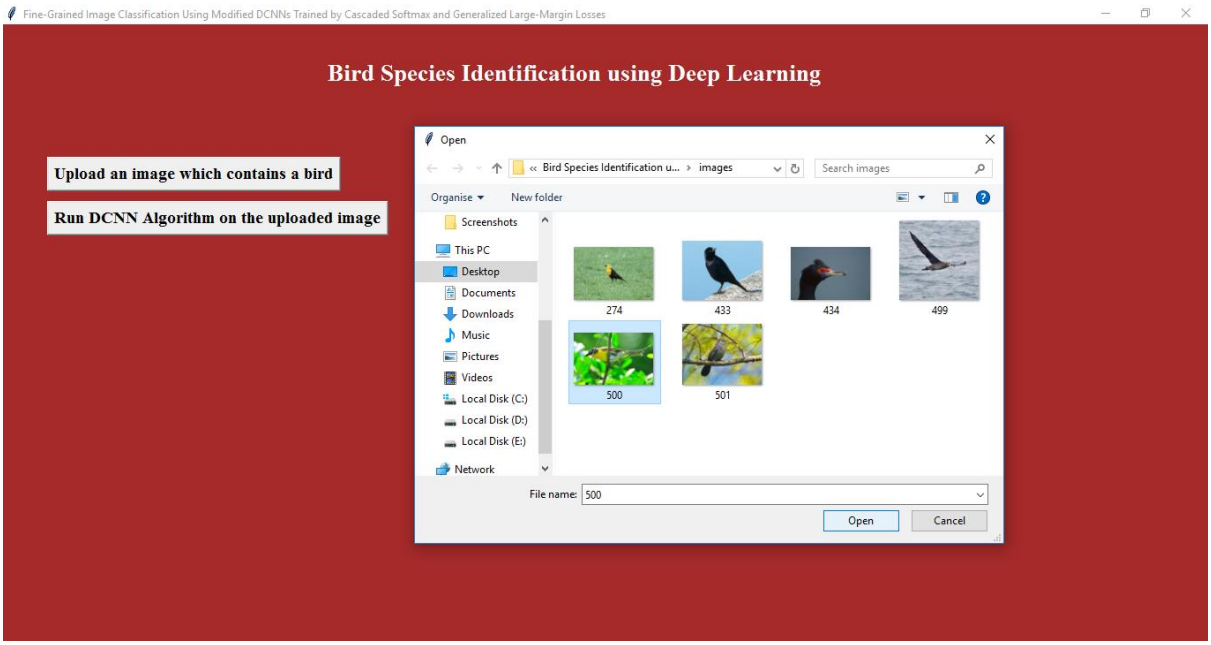


Fig 8.1.2 to upload the image.

In the above screen, select one image from the collection and click “Open”.

Now click on “Run DCNN Algorithm on the uploaded image”.

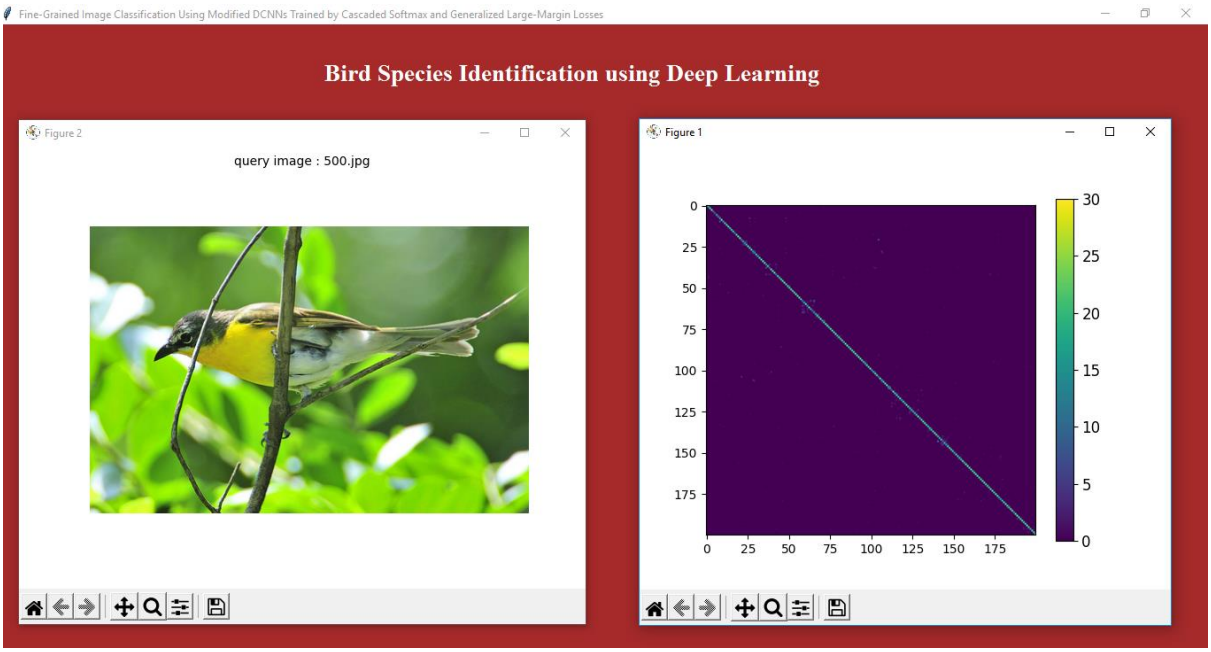


Fig 8.1.3 Query Image.

In above screen we are seeing the uploaded query image. Now we get all similar images related to query.

8.2 OUTPUT SCREENS

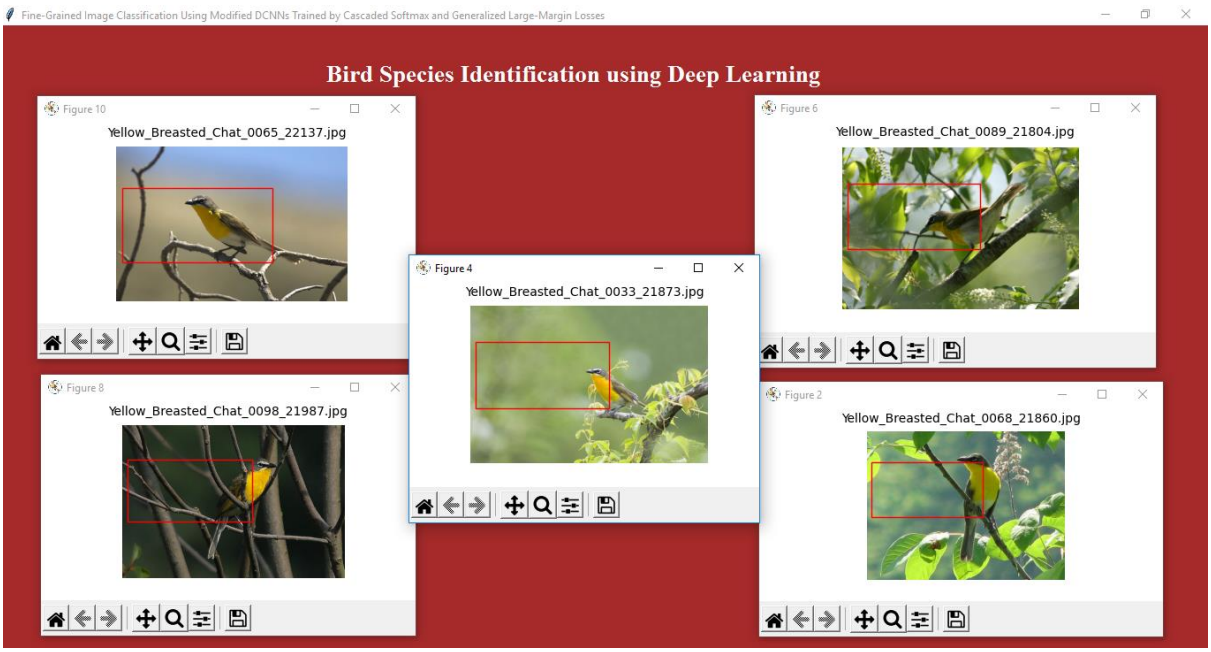


Fig 8.2.1 Output of Query Images

In above screen we can see five images as search result as all images are different but they got searched based

9. EXPERIMENTAL RESULTS

ANALYSING EXPERIMENTAL DATA

To implement this technique we need to train all birds species and generate a model and then by uploading any image deep learning algorithm will convert uploaded image into gray scale format and apply that image on train model to predict best match species name for uploaded image.

To train bird species we are using 'Caltech-UCSD Birds 200(CUB-200-2011)' dataset which contains 200 species or categories of birds. Model will be built using that dataset and tensor flow deep learning algorithm.

INTERPRETING EXPERIMENTAL RESULTS

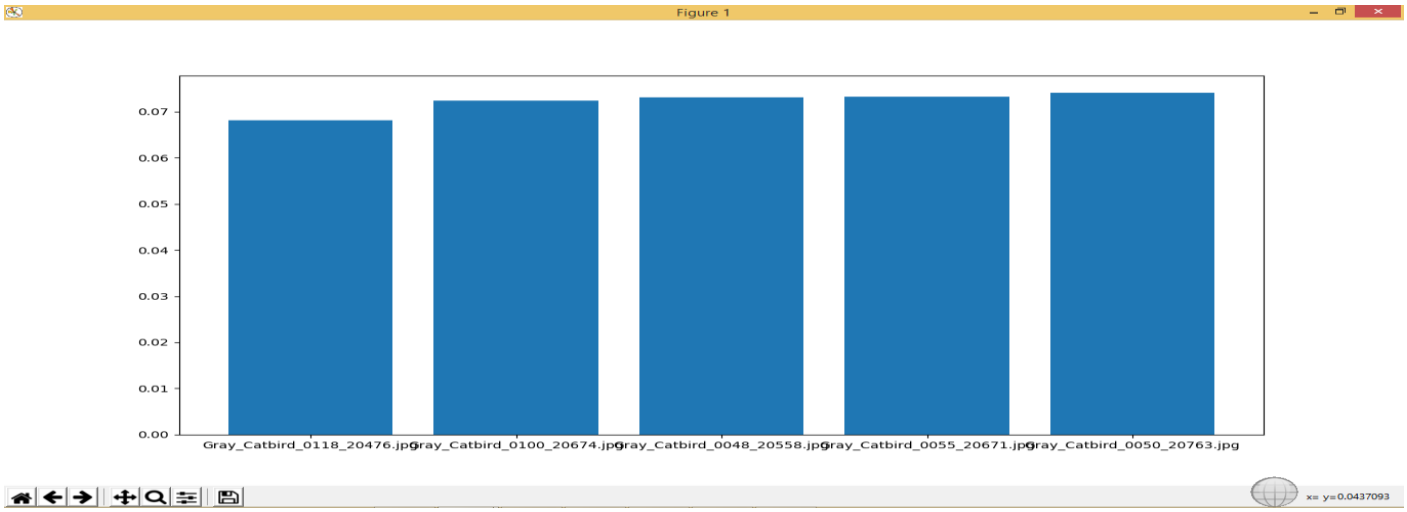


Fig 9.1 Score Sheet

X-axis represents name of the bird.

Y-axis represents matching score.

10. CONCLUSION AND FUTURE ENHANCEMENTS

This model helps building applications that helps tourist who go onto bird sanctuaries identify the bird species by just capturing a picture of a bird and uploading it as input to the model. As many species of birds have become endangered and are near to extinction many people have no knowledge about the species which are few in number, Thus application built using this model may be helpful in identifying the endangered species and help society in spreading awareness about the need of all the species for balance in the nature. As the model implies the knowledge of Deep Convolution neural networks, we can infer that the CNN is the best algorithm for analyzing the visual imagery and image Classification.

The biggest disadvantage of all these algorithms is that the accuracy of these algorithms is dependent on the quality of camera and view angle between camera and the target object. It is also noticed that at some angles the results were not accurate beyond a certain range of camera.

FUTURE ENHANCEMENTS

System can be implemented using cloud which can store large amount of data for comparison and provide high computing power for processing (in case of Neural Networks)

Create an android/ios app instead of website which will be more convenient to user**.** The future of image processing involves new intelligent, digital automated robots made by research scientists in various parts of the world. It includes development in various image processing applications. Due to changes in image processing and other related technologies, there will be millions of robots in the world in a few, transforming the way of living. Researches in image processing and artificial intelligence will involve voice commands, anticipating the information requirements of governments.

REFERENCES

[1]Image Recognition with Deep Learning Techniques ANDREIPETRU BĂRAR, VICTOR-EMIL NEAGOE, NICU SEBE Faculty of Electronics, Telecommunications &Information Technology Polytechnic University of Bucharest.

[2] Bo Zhao, Jiashi Feng Xiao Wu Shuicheng Yan A Survey on Deep Learning-based Fine-grained Object Classification and Semantic Segmentation.

[3]Zagoruyko, S. and Komodakis, N., 2016. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. arXiv preprint arXiv:1612.03928

[4]Andr´eia Marini, Jacques Facon and Alessandro L. Koerich Postgraduate Program in Computer Science (PPGIa) Pontifical Catholic University of Paran´a (PUCPR) Curitiba PR, Brazil 80215–901 Bird Species Classification Based on Color Features

[5]Roslan, R., Nazery, N. A., Jamil, N., &Hamzah, R. (2017). Color-based bird image classification using Support Vector Machine. 2017 IEEE 6th Global Conference on Consumer Electronics (GCCE). (2017)

[6] Bird Species Classification using Transfer Learning with Multistage Training" SouryaDipta Das and Akash Kumar (2018)

[7]Roslan, R., Nazery, N. A., Jamil, N., &Hamzah, R. (2017). Color-based bird image classification using Support Vector Machine. 2017 IEEE 6th Global Conference on Consumer Electronics (GCCE). (2017)

[8]Atanbori, J., Duan, W., Murray, J., Appiah, K., & Dickinson, P. (2016). Automatic classification of flying bird species using computer vision techniques. Pattern Recognition Letters, 81, 53–62." (2016).

[9]"Bird Species Categorization Using Pose Normalized Deep Convolutional Net" Steve Branson, Grant Van Horn, Serge Belongie ,Pietro Peron (2015)

[10] Marini, A., Turatti, A. J., Britto, A. S., &Koerich, A. L. (2015). Visual and acoustic identification of bird species. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2015).

[11]Tóth, B.P. and Czeba, B., 2016, September. Convolutional Neural Networks for Large-Scale Bird Song Classification in Noisy Environment. In CLEF (Working Notes) (pp. 560-568).

[12]Fagerlund, S., 2007. Bird species recognition using support vector machines. EURASIP Journal on Applied Signal Processing, 2007(1), pp.64-64.

[13]Pradelle, B., Meister, B., Baskaran, M., Springer, J. and Lethin, R., 2017, November. Polyhedral Optimization of TensorFlow Computation Graphs. In 6th Workshop on Extreme-scale Programming Tools (ESPT-2017) at The International Conference for High Performance Computing, Networking, Storage and Analysis (SC17).

[14]Cireşan, D., Meier, U. and Schmidhuber, J., 2012. Multi-column deep neural networks for image classification. ArXiv preprint arXiv: 1202.2745

[15] Stefan Kahl, Thomas Wilhelm-Stein, Hussein Hussein, HolgerKlinck, Danny Kowerko, Marc Ritter, and Maximilian Eibl Large-Scale Bird Sound Classification using Convolutional Neural Networks.

Mannelli Aishwarya is currently pursuing her Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. She completed her intermediate from Sri Gayatri Junior college and 10th class from St. Isaac Advent High. She is one of the members of Coders Club in our college. Her technical skills include C, C++, Python and Java (J2EE).She has experience working with frameworks like spring, NodeJS and front-end library ReactJS. She is also a student of Smart Interviews. Her participations include: Women online workshop on "Women in Cyber Security and Privacy in 2020" which was conducted from 6th to 10th July 2020, "Know More - Teach More ", the Global Webinar on Cyber Threats and Defence Techniques conducted by GECF on 22nd July 2020,"Machine learning workshop" three days' workshop on ML. National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" from 10th to 12th December 2020 which was hosted by St. Martin's Engineering College. Her areas of interest are Artificial Intelligence, Machine Learning, Deep Learning, and Neural Networks. She completed few certification courses from online platforms like Coursera, CursaApp, and Udemy.

Yamanamada Sreevishnupriya is currently pursuing her Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. She completed her intermediate from Sri Chaitanya Mahila Kalasala and SSC class from Dr.K.K.R's Gowtham Concept School .She is one of the members of Coders Club in our college. Her technical skills include C, C++, Python and Java .She has experience working with frameworks like spring, NodeJS, Django and front-end library ReactJS. She is also a student of Smart Interviews. Her participations include: Women online workshop on "Women in Cyber Security and Privacy in 2020" which was conducted from 6th to 10th July 2020, "Know More - Teach More ", the Global Webinar on Cyber Threats and Defence Techniques conducted by GECF on 22nd July 2020,"Machine learning workshop" three days' workshop on ML. Her areas of interest are Artificial Intelligence, Machine Learning and Deep Learning. She completed few certification courses from online platforms like Coursera, CursaApp, and Udemy.

Beulah Pasam is currently pursuing her Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. She completed her intermediate from Sri Chaitanya Junior college and 10th class from Saint Paul's High School. She completed her 15 days of internship on "Machine learning through python" at code mania. Her technical skills include C, C++ and Python. She also has a basic understanding of Java. She took part in Employability Skill development Program conducted by Zensar. She is also a student of Smart Interviews. Her participations include: National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020, the Global Webinar on Cloud & Big Data – Changing the way we work which was conducted Global Education & Careers Forum (GECF) on 12th August 2020, Women online workshop on "Women in Cyber Security and Privacy in 2020" which was conducted from 6th to 10th July 2020.

Her areas of interest are Python, Machine Learning and Deep Learning. She completed few certification courses from online platforms like Coursera, CursaApp, Udemy and Solo Learn.

Galinki Hari Teja is currently pursuing her Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. He completed his intermediate and 10th class from Kendriya Vidyalaya No. 1 airforce academy, Dundigal. His technical skills include C, Python and Java. He also has a basic understanding of C++. He took part in Employability Skill development Program conducted by Zensar. He is also a student of Smart Interviews. His participations include: National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020. His areas of interest are Python, Artificial Intelligence, Machine Learning and Deep Learning. He completed few certification courses from online platforms like Coursera, CursaApp and SoloLearn.

A

PROJECT REPORT

On

BIG MART SALES USING MACHINE LEARNING WITH DATA ANALYSIS

Submitted by

1)Ms. RapetiNavya Rani (17K81A05H2)
2)Ms. Chitlapally Bhavani (17K81A05D3)
3)Ms. Ganji Krishna Sri (17K81A05D9)
4)Mr. Krishna Teja (16K81A05E2)

in the partial fulfillment for the award of the

degree of

BACHELOR OF TECHNOLOGY

IN

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Under the Guidance of

Mr. D. Krishna, B.Tech,M.Tech(Ph.D)

Assistant Professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ST.MARTIN'S ENGINEERING COLLEGE
An Autonomous Institute

Dhulapally, Secunderabad – 500 100

JUNE 2021

BONAFIDE CERTIFICATE

This is to certify that the project entitled **Bigmart Sales using Machine Learning with Data Analysis**, is being submitted by 1.**Ms.Rapeti Navya rani(17K81A05H2), 2.Ms.Chitlapally Bhavani (17K81A05D3), 3.Ms.Ganji Krishna Sri (17K81A05D9), 4.Mr.Krishna Teja (16K81A05E2)** in partial fulfillment of the requirement for the award of the degree of **BACHELOR OF TECHNOLOGY IN COMPUTER SCIENCE AND ENGINEERING** is recorded of bonafide work carried out by them. The result embodied in this report have been verified and found satisfactory.

<signature>

Mr. D. Krishna
Department of CSE

Head of the Department

Dr.M.NARAYANAN

Department of CSE

Internal Examiner

External Examiner

Place: Hyderabad
Date:

DECLARATION

We, the student of **Bachelor of Technology** in Department of Computer Science and Engineering', session: <2017 – 2021>, St. Martin's Engineering College, Dhulapally, Kompally, Secunderabad, hereby declare that work presented in this Project Work entitled **Big Mart Sales using Machine Learning with Data Analysis** is the outcome of our own bonafide work and is correct to the best of our knowledge and this work has been undertaken taking care of Engineering Ethics. This result embodied in this project report has not been submitted in any university for award of any degree.

Ms.Rapeti Navya Rani 17K81A05H2

Ms.Chitlapally Bhavani 17K81A05D3

Ms.Ganji Krishna Sri 17K81A05D9

Mr.Krishna Teja 16K81A05E2

ABSTRACT

Everybody wants to know how to buy goods cheaper or how to advertise them at low cost. Here is the answer. That is Big Mart. Big Mart is online one stop marketplace where you can buy or sell or advertise your merchandise at low cost. The goal is to make Big Mart the shopping paradise for buyers and the marketing solutions for the sellers. The ultimate goal is to prosper with customers. The project BIGMART SALES DATASET aims to build a predictive model and find out the sales of each product at a particular store. Big Mart will use this model to understand the properties of products and stores which play a key role in increasing sales. This can also be done based on the hypothesis that should be done before looking at the data.Estimating future sales is the major aspect of the numerous distributions, manufacturing, marketing and wholesaling companies involved. This helps businesses to allocate capital effectively, to forecast realistic sales revenues as well as to prepare a better plan for potentially increasing the business. In this paper, estimating product sale from a single outlet is carried out using a random forest regression approach, XG booster approach which provides better predictive results compared to a linear regression model. This approach is carried out on data from Big-Mart Sales where data discovery, processed and sufficient relevant data is extracted which play a vital role in predicting accurate outcome.

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible and whose encouragements and guidance have crowded effects with success.

We extended our deep sense of gratitude to Principal**, Dr. P. SANTOSH KUMAR PATRA**, St. Martin's Engineering College, Dhulapally, for permitting us to undertake this project.

We are also thankful to **Dr. M.NARAYANAN**, Head of the Department, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

We would like to thank **Dr. T. POONGOTHAI,** Department of Computer Science and Engineering (AI & ML) for her encouragement and insightful comments and as well as our project coordinators **Dr. B.RAJALINGAM**, Professor and **Dr. N. SATHEESH**, Professor, in Department of Computer Science and Engineering for their valuable support.

We would like to express our sincere gratitude and indebtedness to our project supervisor Mr.D.Krishna, Assistant Professor, Computer Science and Engineering,St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

Finally, we express thanks to all those who have helped us successfully to completing this project. Furthermore, we would like to thank our family and friends for their moral support and encouragement.

We express thanks to all those who have helped us in successfully completing the project.

Ms. Navya RaniRapeti (17K81A05H2)

Ms. Chitlapally Bhavani (17K81A05D3)

Ms. Ganji Krishna Sri (17K81A05D9)

Mr. Krishna Teja (17K81A05E2)

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF OUTPUT SCREENS

1.INTRODUCTION

1. INTRODUCTION

1.1 PROJECT OVERVIEW

With the rapid development of global malls and stores chains and the increase in the number of electronic payment customers, the competition among the rival organizations is becoming more serious day by day. Each organization is trying to attract more customers using personalized and short-time offers which makes the prediction of future volume of sales of every item an important asset in the planning and inventory management of every organization, transport service, etc. Due to the cheap availability of computing and storage, it has become possible to use sophisticated machine learning algorithms for this purpose. In this paper, we are providing forecast for the sales data of big mart in a number of big mart stores across various location types which is based on the historical data of sales volume. According to the characteristics of the data, we can use the method of multiple linear regression analysis and random forest to forecast the sales volume.The project "BIGMART SALES DATASET" aims to build a predictive model and find out the sales of each product at a particular store. Big Mart will use this model to understand the properties of products and stores which play a key role in increasing sales**.** This can also be done based on the hypothesis that should be done before looking at the data. The data is preprocessed with the removal of noise , filling missing values. Feature engineering is performed to convert the data understandably. Then the data mining algorithms are applied.

1.2 PROJECT OBJECTIVES

- The project "BIGMART SALES DATASET" aims to build a predictive model and find out the sales of each product at a particular store.
- Big Mart will use this model to understand the properties of products and stores which play a key role in increasing sales**.**

- This can also be done based on the hypothesis that should be done before looking at the data. The data is preprocessed with the removal of noise , filling missing values.

- Feature engineering is performed to convert the data understandably. Then the data mining algorithms are applied.

1.3 ORGANIZATION OF CHAPTERS

The thesis is organized in the following chapters:

Chapter 1: Introduction

This chapter covers the overview of our project and its objectives. Literature Survey – This includes the details of our survey, Software and Hardware Requirements – We specify our software and hardware requirements here. Software Development Analysis – This section includes the problem definition and details of the modules we used in our project. Project System Design – This chapter includes the design part of our project which includes uml diagrams. Project Coding – This section contains the details of our project code. Project Testing – The details of test cases and testing are included in this chapter. Output Screens – This contains the screenshots of how our project looks like whenexecuted. Experimental Results – This chapter contains the screenshots of our results. Conclusion and Future Enhancements – This covers the conclusion of our project and the possible future developments.

Chapter 2: Literature Survey

In this section, we analyze the existing result which were done on linear regression model individually. We progress in the view of developing a project by merging both the technologies using Healthcare Domain as a challenge.

Chapter 3: Software and Hardware Requirements

In this chapter, we specified the Software and Hardware components required to develop our project.The Softwareand Hardware requirements specify the intended purpose, requirements, and nature of software/application/project to be developed. By selecting the dataset that most resembles the usage requirements in our environment, we can use the recommended topology and associated hardware requirements for our topology as a starting point when we plan for hardware of our project.. Requirements may vary based on utilization and observing performance of pilot projects is recommended prior to scale out.

Chapter 4: Software Development Analysis

In this project, we discussed about development and implementation of the project in detail. Considering the security of one's data, we developed in our roles as front-end, back-end and database administrator by collecting relevant data and testing it in required cases.

Chapter 5: Project System Design

This chapter reports on the analysis and design of our proposed application. This chapter describes the system design architecture and database design and is organized in a sequence included with data gathering and system design. Stakeholders will discuss factors such as risk levels, team composition, applicable technologies, time, budget, project limitations, method and architectural design.

Chapter 6: Project Coding

This chapter is a system implementation of the project. We will discuss briefly the implementation of our project.This section describes some of the coding templates, outline of various files, class with functionalities, the various methods of input and output parameters.

Chapter 7: Project Testing

In this chapter, we will discuss briefly the testing of each functionality of our proposed application in the project. We performed various testings like whitebox, blackbox, unit testing, integration testing and many more to check the accuracy and performance of our output. They notify developers of defects in the code. If developers confirm the flaws are valid, they improve the program, and the testers repeat the process until the software is free of bugs and behaves according to requirements.

Chapter 8: Output screens

In this chapter, we captured the screenshots of our project output. We considered few sample inputs and obtained desired outputs for our data with related database.

Chapter 9: Experimental Results

In this chapter, we conclude the performance analysis of our proposed project by comparing it with the existing project. In this chapter, we discuss briefly the conclusion of each chapter with the progress of our proposed system.

2.LITERATURE SURVEY

2.LITERATURE SURVEY

2.1 SURVEY ON BACKGROUND

This section of the literature survey eventually reveals some facts based on thoughtful analysis of many authors work as follows.

Machine Learning is defined as the computer program which learns by itself from its experience without any human interference. Research on sales prediction has been done and some of them has been discussed below:

[1], general linear approach, decision tree approach and good gradient approach were used to predict sales. The initial data set considered included many entries, but the final data set which is used for analyzing was much smaller than the original as it consists of non-usable data, redundant entries and insignificant sales data.

[2], linear regression method has been organized into structured data. Then it involves modeling data for predictions using machine learning techniques where the expected accuracy was 84%.

[3], they used linear regression and XG booster algorithm to forecast sales that included data collection and translation into processed data. Ultimately, they predicted which model would produce the better outcome.

[4],sales were predicted using three modules that are hive, R programming and tableau. By analysing the stores history which helps get an understanding of the store's revenue to make some improvements to the target so it can be more successful. Within the diagram, key values are obtained to reduce all intermediate values by reducing the intermediate key feature to obtain the results. Mohit Gurnani in his research proves that composite models achieve good results in comparison to individual models. He also stated that decomposition mechanisms are far better than hybrid mechanisms

[5]. J. Scott Armstrong in his research discussed about predicting solutions to interesting and difficult sales forecasting problems

[6]. Samaneh Beheshti-Kashi in his research reviewed different Various approaches on the predictive potential of consumer-generated content and search queries

[7]. Gopal Behera has done effective study on Big mart sales prediction and has given prediction metrics for various existing models

[8]. In this paper, we use random forest and XG booster methodology in which raw data obtained at large mart will be pre-processed for missing data, anomalies and outliers.

Then an algorithm will be used to predict the final results. ETL stands for Extract, Transform and Load and finally we compare all the models and predict which model gives accurate result

2.2 CONCLUSIONS ON SURVEY

This framework is to predict the future sales from given data of the previous year's using machine Learning techniques. In this paper, we have discussed how different machine learning models are built using different algorithms like Linear regression, Random forest regressor, and XG booster algorithms. These algorithms have been applied to predict the final result of sales. We have addressed in detail about how the noisy data is been removed and the algorithms used to predict the result. Based on the accuracy predicted by different models we conclude that the random forest approach and XG Booster approach are best models. Our predictions help big marts to refine their methodologies and strategies which in turn helps them to increase their profit.Each organization is trying to attract more customers using personalized and short-time offers which makes the prediction of future volume of sales of every item an important asset in the planning and inventory management of every organization, transport service, etc. Due to the cheap availability of computing and storage, it has become possible to use sophisticated machine learning algorithms for this purpose. In this paper, we are providing forecast for the sales data of big mart in a number of big mart stores across various location types which is based on the historical data of sales volume. According to the characteristics of the data, we can use the method of multiple linear regression analysis and random forest to forecast the

sales volume.Big Mart will use this model to understand the properties of products and stores which play a key role in increasing sales. This can also be done based on the hypothesis that should be done before looking at the data.Estimating future sales is the major aspect of the numerous distributions, manufacturing, marketing and wholesaling companies involved. This helps businesses to allocate capital effectively, to forecast realistic sales revenues as well as to prepare a better plan for potentially increasing the business. In this paper, estimating product sale from a single outlet is carried out using a random forest regression approach, XG booster approach which provides better predictive results compared to a linear regression model. This approach is carried out on data from Big-Mart Sales where data discovery, processed and sufficient relevant data is extracted which play a vital role in predicting accurate outcome.

3.SOFTWARE AND HARDWARE REQUIREMENTS

3. SOFTWARE AND HARDWARE REQUIREMENTS

The Softwareand Hardware requirements specify the intended purpose, requirements, and nature of software/application/project to be developed. These system requirements are the configuration that our system must have in order for a hardware or software application to run smoothly and effectively. These requirements at the system level describes the functions which the system as a whole should fulfil to satisfy the stakeholder needs and requirements.

3.1 SOFTWARE REQUIREMENTS

Operating System	**Windows 10**
Platform	**Python**

Table 3.1 Software Requirements

3.2 HARDWARE REQUIREMENTS

Processor	**Intel I3**
Hard Disk	**160GB**
RAM	**4GB**

Table 3.2 Hardware Requirements

4. SOFTWARE DEVELOPMENT ANALYSIS

4. SOFTWARE DEVELOPMENT ANALYSIS

4.1 OVERVIEW OF PROBLEM

In today's modern world, huge shopping centers such as big malls and marts are recording data related to sales of items or products with their various dependent or independent factors as an important step to be helpful in prediction of future demands and inventory management. The dataset built with various dependent and independent variables is a composite form of item attributes, data gathered by means of customer, and also data related to inventory management in a data warehouse. The data is thereafter refined in order to get accurate predictions and gather new as well as interesting results that shed a new light on our knowledge with respect to the task's data. This can then further be used for forecasting future sales by means of employing machine learning algorithms such as the random forests and simple or multiple linear regression model.

4.2 DEFINE THE PROBLEM

In this project, we use random forest regressor and XG-booster approach to predict sales where data mining techniques such as discovery, data transformation, feature development, model creation and testing are used. In this technique raw data collected by a big mart will be pre-processed for missing data, anomalies and outlier. An algorithm will then be trained to construct a model on that data. We will use this model to forecast the end results. It is a system in which three functions are combined. It is used to extract and transform the data from one database into an appropriate format.

In this proposed system we have used Random Forest Algorithm to incorporate predictions from multiple decision trees into a single model.

RANDOM FOREST MODEL

Random forest is a supervised machine learning algorithm based on ensemble learning. Ensemble learning is an algorithm where the predictions are derived by assembling or bagging different models or similar model multiple times. The random forest algorithm works in a similar way and uses multiple algorithm i.e. multiple decision trees, resulting in a forest of trees, hence

the name "Random Forest". The random forest algorithm can be used for both regression and classification tasks

4.3 MODULES OVERVIEW

In order to determine the tasks we use modules as source.In this modules the data form datasets is taken and this helps to train and test the data separately which produces frauds.

Dataset collection

Data exploration

Data cleaning

Feature engineering

Model building

4.4 DEFINE THE MODULES

In order to find a decent model to predict sales we performed an extensive search of various machine learning models available in R, in particular of those accessible through the caret wrapper. In the end, however, models from the h2o package yielded the best results for the task. In particular, deep learning neural networks h2o.deeplearning and gradient boosting regression trees h2o.gbm performed particularly well. An ensemble of various such models, constructed in h2oEnsemble.R forms the basis of our submission. Here, we used only the 12 most important predictors to avoid over-fitting. To include some features we may have missed with this rather small sub set of predictors we supplemented the ensemble with a deep learning neural net using 23 predictors.

4.5 MODULE FUNCTIONALITY

1. DATASET COLLECTION

We used wide market sales data as a dataset in our work where the dataset consists of 12 attributes. These 12 attributes define the basic features of the data which is being forecasted. These attributes are divided into Answer Variable and predictors. Here we use dataset which contains 8523 items spanning various locations as well as cities. Store-and product-level

hypotheses are the main factors on which our dataset focuses on. Attributes such as area, population density, capability of the store, location etc have been included in store level. At last the dataset is divided as training and test dataset.

2. DATA EXPLORATION

Valuable data information is drawn-out from the dataset in this step. Outlet year of establishment ranges from 1985 to 2009. These Values in this form may not be sufficient. There are 1559 different items present in the dataset and 10 different outlets Here we classify the data from the hypothesis vs available evidence which indicates that the size of the outlet attribute and the weight of the object faces the question of missing values, as well as the least value of Object view is Zero which is not feasible. The Item type attribute contains 16 specific values.

3. DATA CLEANING

Here in place of missing value for outlet size, we replace with mode value of that attribute and in place of missing values of that particular attribute of object weight, we substitute by mean value. The missing attributes are numerical, where correlation between the imputed attributes decreases as well as the mean and mode replacement decreases. we believe that there is no relation among the attributes calculated and the attribute imputed in our model.

4. FEATURE ENGINEERING

Feature engineering is all about converting cleaned data into predictive models to present the available problem in a better way. During data exploration, some noise was observed. In this phase, this noise is resolved and the data is used for building appropriate model. New features are created to make the model work precisely and effectively. A few created features can be combined for the model to work better. Feature engineering phase converts data into a form understandable by the algorithms.

5.MODEL BUILDING

After Feature engineering, the processed data is used to give accurate results by applying multiple algorithms. A model is a set of algorithms that facilitate the process of finding relation between multiple dataset. An effective model can predict accurate results by finding exact insights of data.

5. PROJECT SYSTEM DESIGN

5. PROJECT SYSTEM DESIGN

Project Design is the strategic organization of ideas material and processes to set our project up for success once we launch. It includes dataflow diagram ,object models, architecture diagrams, UML diagrams, a detailed design and functionality of our project.

5.1 DFDS (DATA FLOW DIAGRAM)

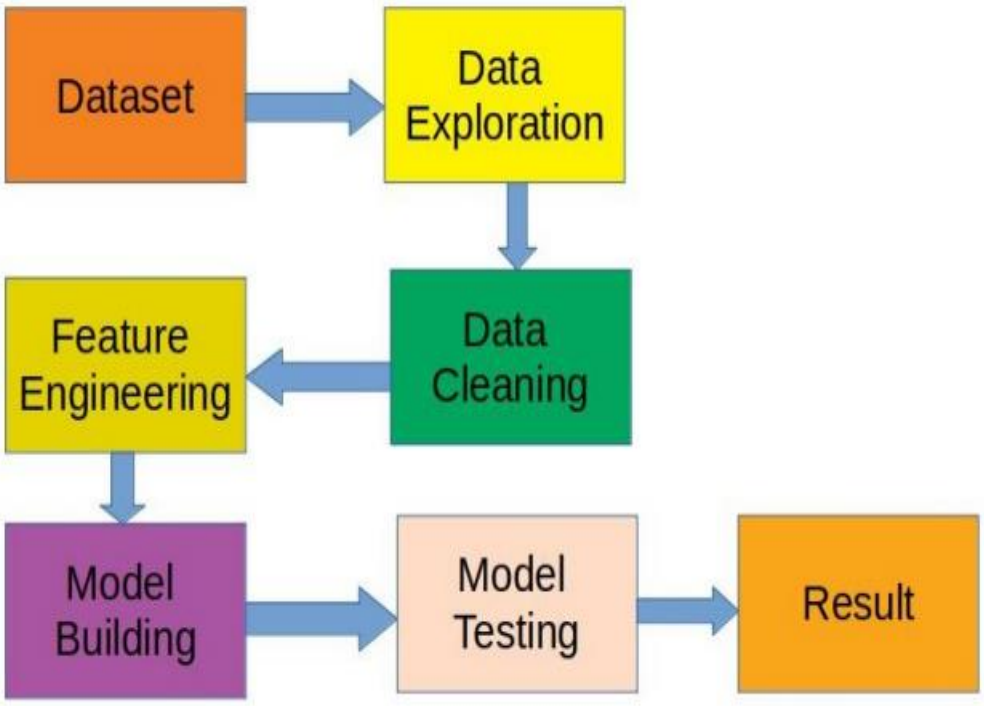


Fig 5.1.1 Data flow diagram

Data flow diagrams are used to graphically represent the flow of data in a business information system. DFD describes the processes that are involved in a system to transfer data from the input to the file storage and reports generation.

Data flow diagrams can be divided into logical and physical. The logical data flow diagram describes flow of data through a system to perform certain functionality of a business. The physical data flow diagram describes the implementation of the logical data flow.

5.2 E-R DIAGRAMS

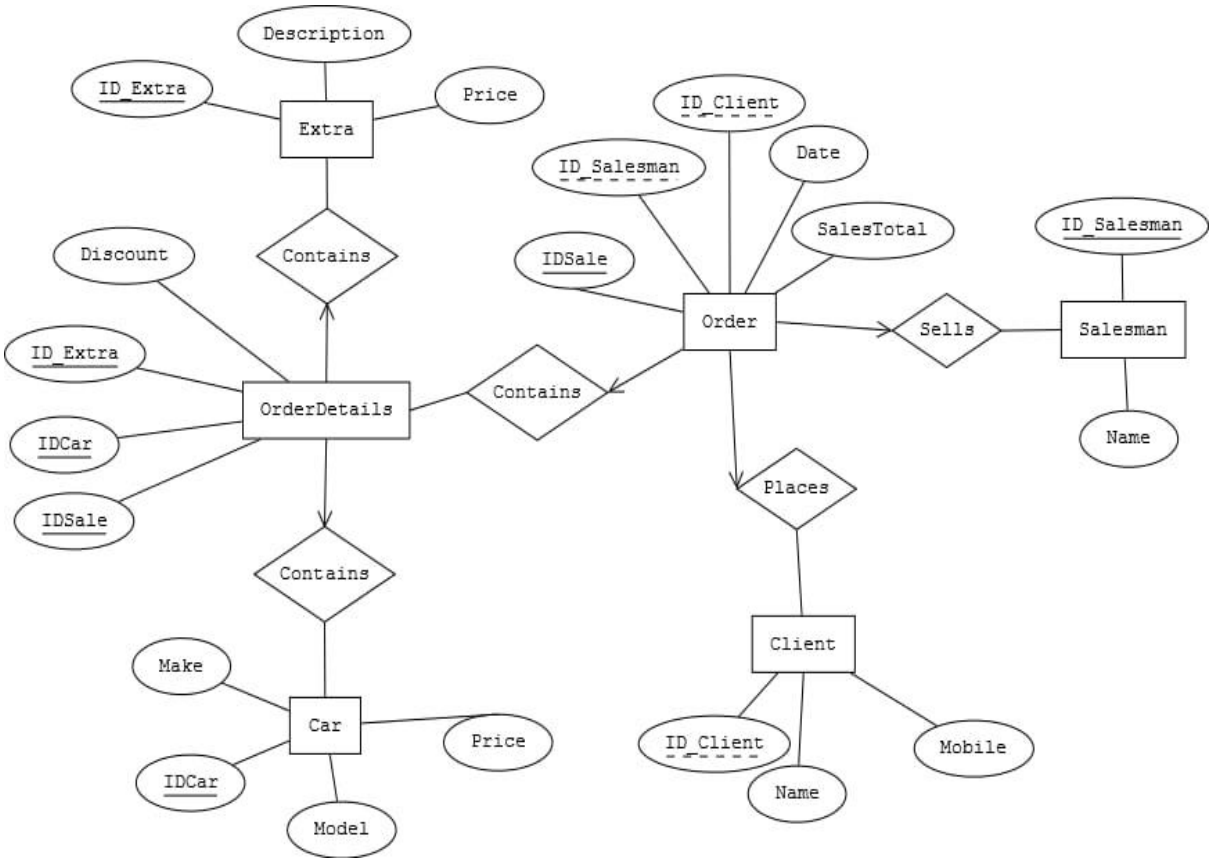


Fig 5.2 E-R Diagram

ER diagrams are related to data structure diagrams (DSDs), which focus on the relationships of elements within entities instead of relationships between entities themselves. ER diagrams also are often used in conjunction with data flow diagrams (DFDs), which map out the flow of information for processes or systems.

5.3 UML DIAGRAMS

USE CASE DIAGRAM

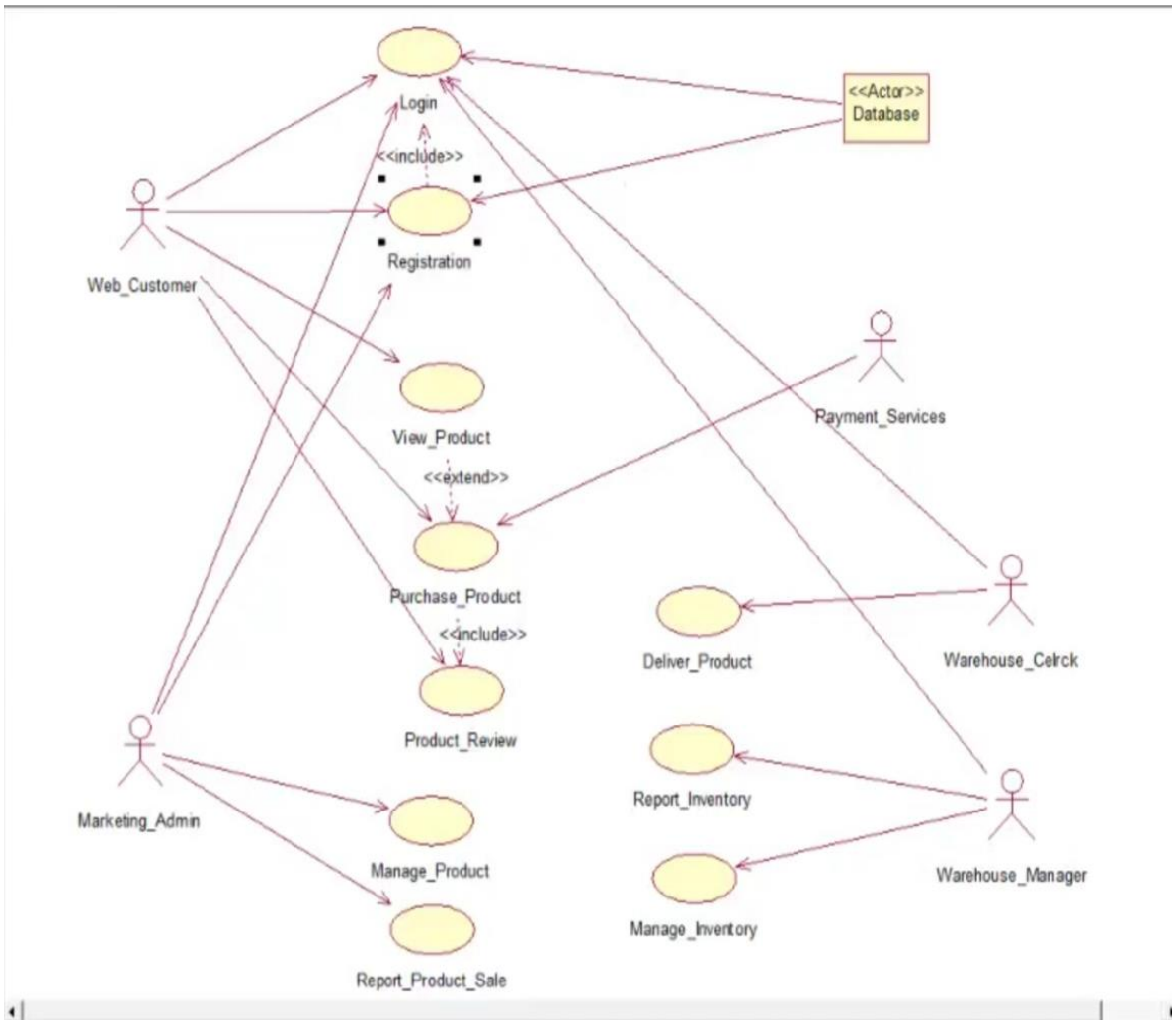


Fig 5.3.1 Use Case Diagram

CLASS DIAGRAM

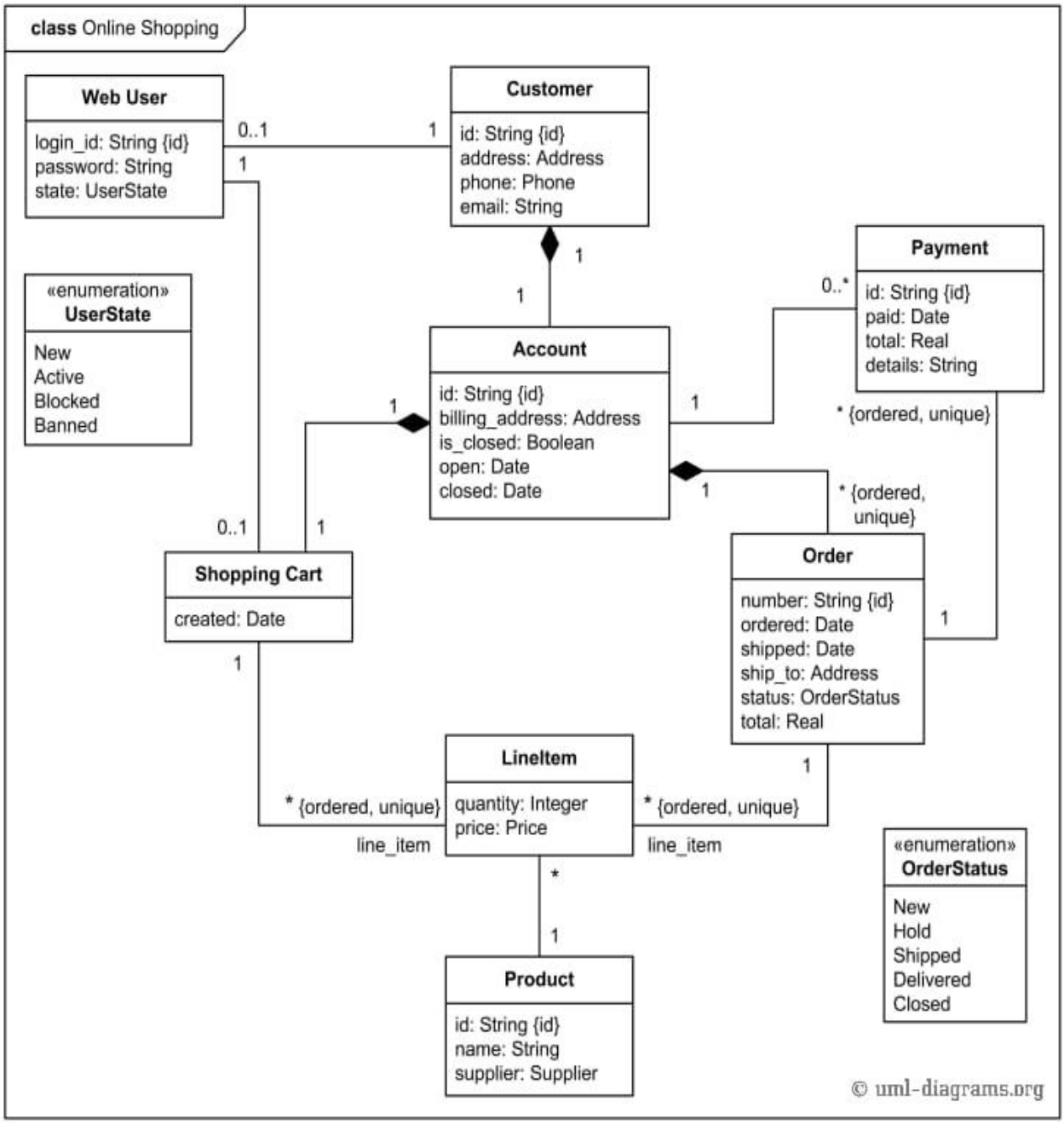


Fig 5.3.2 Class Diagram

SEQUENCE DIAGRAM

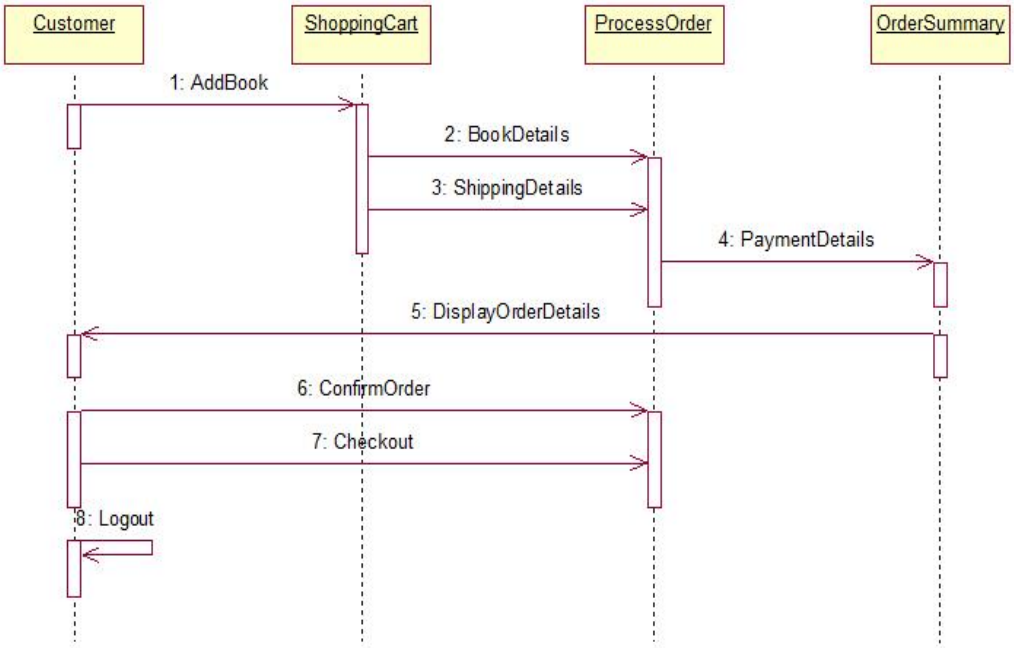


Fig 5.3.3 Sequence Diagram

6. PROJECT CODING

6. PROJECT CODING

6.1 CODE TEMPLATES

Randam forest Machine Learning Algorithm in python

```
import pandas as pd

from mlxtend.frequent_patterns import apriori

from mlxtend.frequent_patterns import association_rules

df = pd.read_excel(r'C:\Users\Shiva Reddy\Downloads\bhanu\Project17

mba\data.xlsx')

df.head()
```

```
df['Description'] = df['Description'].str.strip()

df.dropna(axis=0, subset=['InvoiceNo'], inplace=True)

df['InvoiceNo'] = df['InvoiceNo'].astype('str')

df = df[~df['InvoiceNo'].str.contains('C')]

basket = (df[df['Country'] =="France"]

          .groupby(['InvoiceNo', 'Description'])['Quantity']

          .sum().unstack().reset_index().fillna(0)

          .set_index('InvoiceNo'))



def encode_units(x):

    if x <= 0:

        return 0

    if x >= 1:

        return 1



basket_sets = basket.applymap(encode_units)

basket_sets.drop('POSTAGE', inplace=True, axis=1)



frequent_itemsets = apriori(basket_sets, min_support=0.07, use_colnames=True)
```

```
rules = association_rules(frequent_itemsets, metric="lift", min_threshold=1)

rules.head()
```

6.2 ATTRIBUTES USED IN DATASET

- **Item_Identifier**: Unique product ID
- **Item_Weight**: Weight of product
- **Item_Fat_Content**: Whether the product is low fat or not
- **Item_Visibility**: The % of total display area of all products in a store allocated to the particular product
- **Item_Type**: The category to which the product belongs
- **Item_MRP**: Maximum Retail Price (list price) of the product
- **Outlet_Identifier**: Unique store ID
- **Outlet_Establishment_Year**: The year in which store was established
- **Outlet_Size**: The size of the store in terms of ground area covered
- **Outlet_Location_Type**: The type of city in which the store is located
- **Outlet_Type**: Whether the outlet is just a grocery store or some sort of supermarket
- **Item_Outlet_Sales**: Sales of the product in the particular store. This is the outcome variable to be predicted.

6.3 LIBRARIES TO INSTALL IN PYTHON

```
pip install numpy==1.18.1

pip install pandas==0.25.3

pip install matplotlib==3.1.3

pip install keras==2.3.1

pip install tensorflow==1.14.0
```

pip install h5py==2.10.0

pip install opencv-python==4.2.0.32

pip install scikit-image

pip3 install package_name --user

6.54 Some observations:

- **Item_Visibility** has a min value of zero. This makes no practical sense because when a product is being sold in a store, the visibility cannot be 0.
- **Outlet_Establishment_Years** vary from 1985 to 2009. The values might not be apt in this form. Rather, if we can convert them to how old the particular store is, it should have a better impact on sales.
- The lower 'count' of **Item_Weight** and **Outlet_Size** confirms the findings from the missing value check.

7. PROJECT TESTING

7. PROJECT TESTING

7.1 VARIOUS TEST CASES

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, subassemblies, assemblies and/or a finished product. It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

TYPES OF TESTS

Unit testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

Integration testing

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

Functional test

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items:

Valid Input : identified classes of valid input must be accepted.

Invalid Input : identified classes of invalid input must be rejected.

Functions : identified functions must be exercised.

Output : identified classes of application outputs must be exercised.

Systems/Procedures : interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

System Test

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

Unit Testing

Unit testing is usually conducted as part of a combined code and unit test phase of the software lifecycle, although it is not uncommon for coding and unit testing to be conducted as two distinct phases.

Test strategy and approach

Field testing will be performed manually and functional tests will be written in detail.

Test objectives

- All field entries must work properly.
- Pages must be activated from the identified link.
- The entry screen, messages and responses must not be delayed.

Features to be tested

- Verify that the entries are of the correct format.
- No duplicate entries should be allowed.
- All links should take the user to the correct page.

Integration Testing

Software integration testing is the incremental integration testing of two or more integrated software components on a single platform to produce failures caused by interface defects. The task of the integration test is to check that components or software applications, e.g. components

in a software system or – one step up – software applications at the company level – interact without error.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

Acceptance Testing

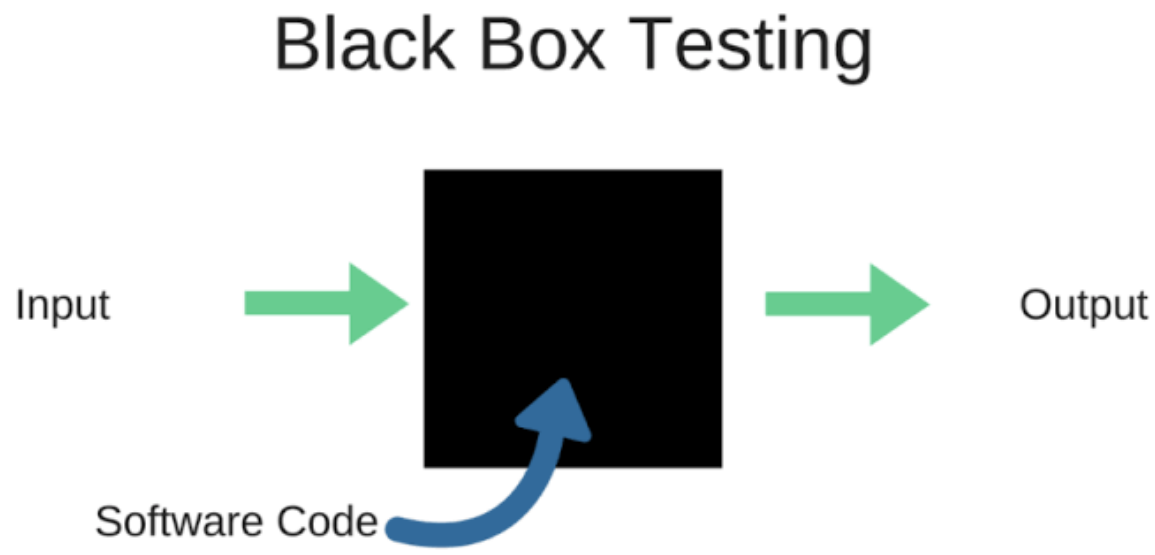
User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

7.2 BLACK BOX TESTING

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box .you cannot "see" into it. The test provides inputs and responds to outputs without considering how the software works.

The below Black-Box can be any software system you want to test. For Example, an operating system like Windows, a website like Google, a database like Oracle or even your own custom application. UnderBlack Box Testing, you can test these applications by just focusing on the inputs and outputs without knowing their internal code implementation.



Various approaches to black-box testing

There are a set of approaches for black-box testing.

Manual UI Testing: In this approach, a tester checks the system as a user. Check and verify the user data, error messages.

Automated UI Testing: In this approach, user interaction with the system is recorded to find errors and glitches. Testers can set record demand as per schedule.

Documentation Testing: In this approach, a tester purely checks the input and output of the software. Testers consider what system should perform rather than how. It is a manual approach to testing.

The tester doesn't need any technical knowledge to test the system. It is essential to understand the user's perspective.

Testing is performed after development, and both the activities are independent of each other. It works for a more extensive coverage which is usually missed out by testers as they fail to see the bigger picture of the software.

Test cases can be generated before development and right after specification. Black box testing methodology is close to agile.

7.3 WHITE BOX TESTING

The box testing approach of software testing consists of black box testing and white box testing. We are discussing here white box testing which also known as glass box is testing, structural testing, clear box testing, open box testing and transparent box testing. It tests internal coding and infrastructure of a software focus on checking of predefined inputs against expected and desired outputs. It is based on inner workings of an application and revolves around internal structure testing. In this type of testing programming skills are required to design test cases. The primary goal of white box testing is to focus on the flow of inputs and outputs through the software and strengthening the security of the software.

The term 'white box' is used because of the internal perspective of the system. The clear box or white box or transparent box name denote the ability to see through the software's outer shell into its inner workings.

Developers do white box testing. In this, the developer will test every line of the code of the program. The developers perform the White-box testing and then send the application or the software to the testing team, where they will perform the black box testing and verify the application along with the requirements and identify the bugs and sends it to the developer.

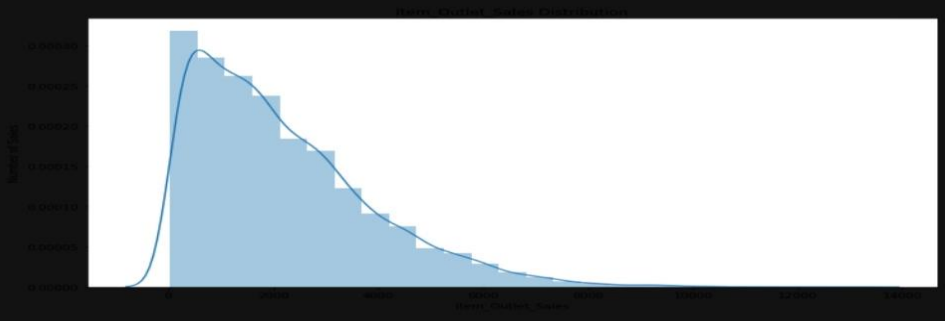
The developer fixes the bugs and does one round of white box testing and sends it to the testing team. Here, fixing the bugs implies that the bug is deleted, and the particular feature is working fine on the application.

8. OUTPUT SCREENS

8. OUTPUT SCREENS

8.1 USER INTERFACE

```
plt.figure(figsize=(12,7))
sns.distplot(train.Item_Outlet_S
ales, bins = 25)
plt.xlabel("Item_Outlet_Sales")
plt.ylabel("Number of Sales")
plt.title("Item_Outlet_Sales
Distribution")
```

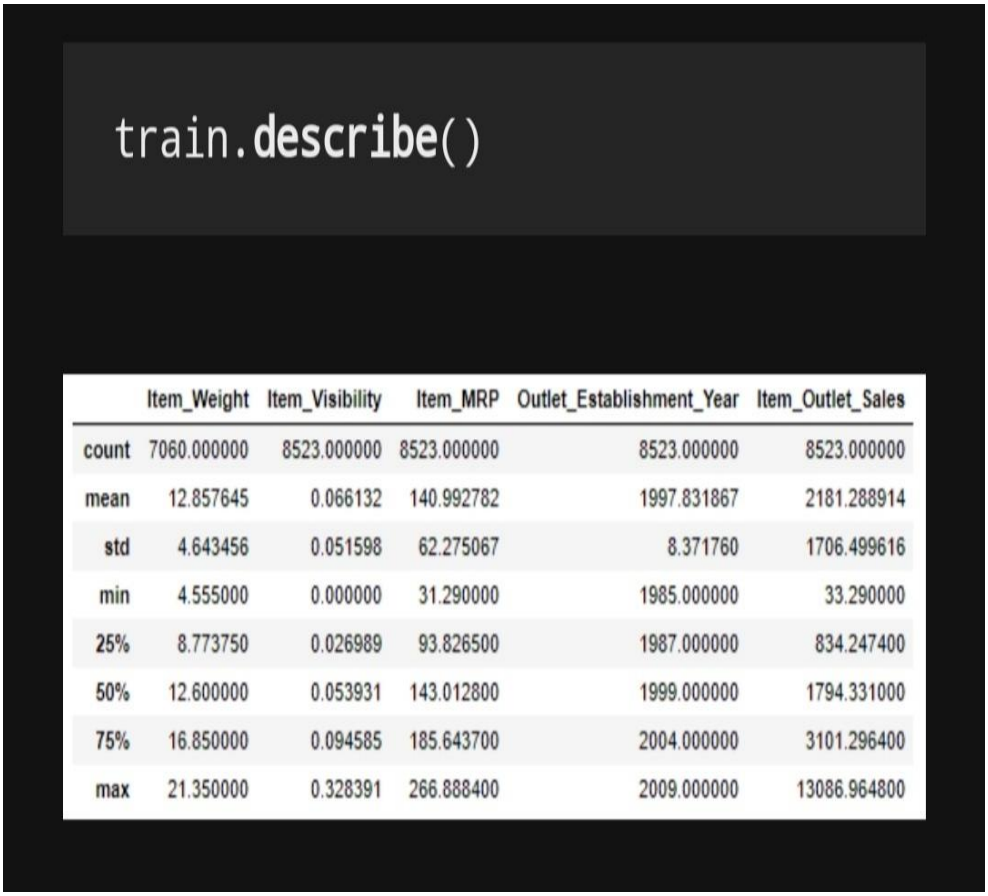


```
print ("Skew is:",
train.Item_Outlet_Sales.skew())
print("Kurtosis: %f" %
train.Item_Outlet_Sales.kurt())
```

```
train.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8523 entries, 0 to 8522
Data columns (total 12 columns):
Item_Identifier              8523 non-null object
Item_Weight                  7060 non-null float64
Item_Fat_Content             8523 non-null object
Item_Visibility              8523 non-null float64
Item_Type                    8523 non-null object
Item_MRP                     8523 non-null float64
Outlet_Identifier            8523 non-null object
Outlet_Establishment_Year    8523 non-null int64
Outlet_Size                  6113 non-null object
Outlet_Location_Type         8523 non-null object
Outlet_Type                  8523 non-null object
Item_Outlet_Sales            8523 non-null float64
dtypes: float64(4), int64(1), object(7)
memory usage: 799.1+ KB
```

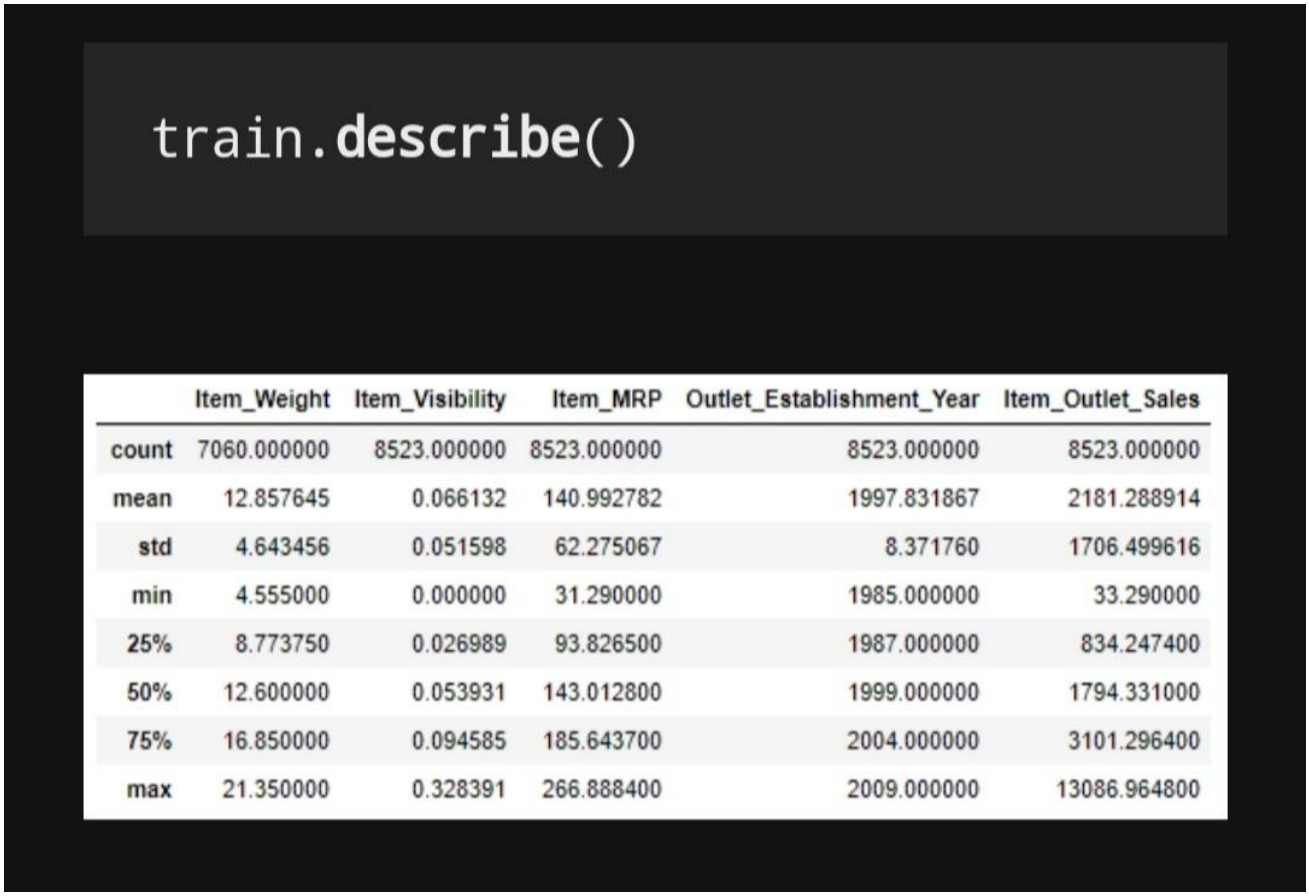
8.2 OUTPUT SCREENS



	Item_Weight	Item_Visibility	Item_MRP	Outlet_Establishment_Year	Item_Outlet_Sales
count	7060.000000	8523.000000	8523.000000	8523.000000	8523.000000
mean	12.857645	0.066132	140.992782	1997.831867	2181.288914
std	4.643456	0.051598	62.275067	8.371760	1706.499616
min	4.555000	0.000000	31.290000	1985.000000	33.290000
25%	8.773750	0.026989	93.826500	1987.000000	834.247400
50%	12.600000	0.053931	143.012800	1999.000000	1794.331000
75%	16.850000	0.094585	185.643700	2004.000000	3101.296400
max	21.350000	0.328391	266.888400	2009.000000	13086.964800

Fig 8.2: Generate Train and Test data

In above Screen,After generating model we can see total records available in dataset and then application using how many records for training and how many for testing. Now click on “Run Random Forest Algorithm” button to generate Random Forest model on train and test data.



	Item_Weight	Item_Visibility	Item_MRP	Outlet_Establishment_Year	Item_Outlet_Sales
count	7060.000000	8523.000000	8523.000000	8523.000000	8523.000000
mean	12.857645	0.066132	140.992782	1997.831867	2181.288914
std	4.643456	0.051598	62.275067	8.371760	1706.499616
min	4.555000	0.000000	31.290000	1985.000000	33.290000
25%	8.773750	0.026989	93.826500	1987.000000	834.247400
50%	12.600000	0.053931	143.012800	1999.000000	1794.331000
75%	16.850000	0.094585	185.643700	2004.000000	3101.296400
max	21.350000	0.328391	266.888400	2009.000000	13086.964800

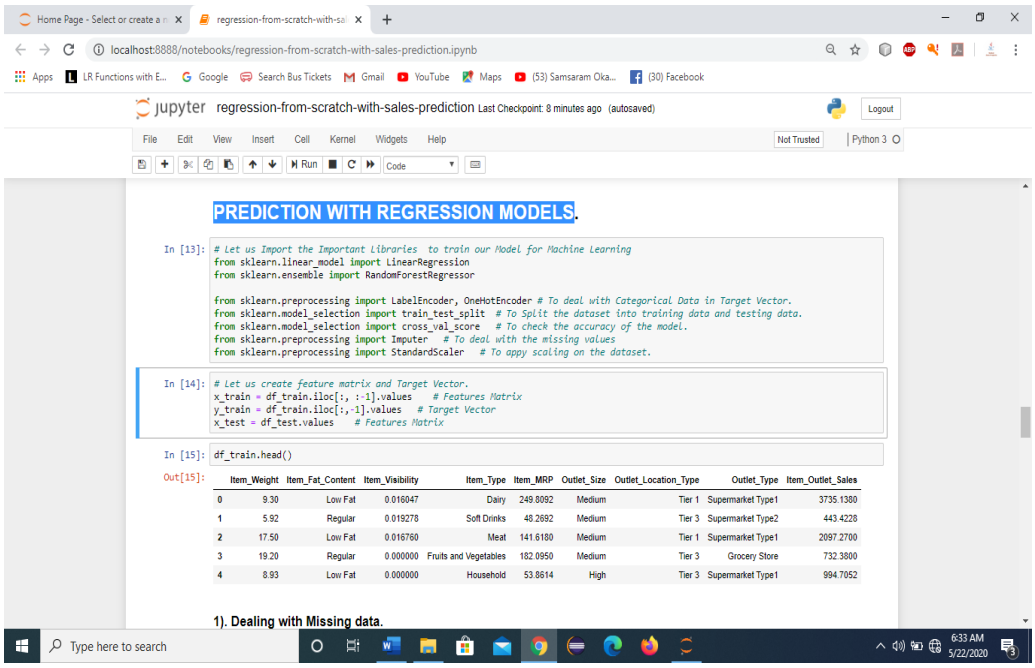
Fig 8.3: Run the Random Forest Algorithm

After Generating Random forest we get as In above screen Random Forest generate 99.78% percent accuracy while building model on train and test data. Now click on 'Detect Fraud From Test Data' button to upload test data and to predict whether test data contains normal or fraud transaction.

9. EXPERIMENTAL RESULTS

9. EXPERIMENTAL RESULTS

Considering the existing system and proposed system implementations, we come to a conclusion that the proposed system Blockchain ledger and AI technology embedded with cloud computing technology can advance the biomedical and health care domains in various novel ways, and we expect many new applications to emerge soon. As we implement the proposed project, we understand that the key aspects like the cost, time, size of the elements can be accessed to the at most level and enhance the privacy, security of data through exchange of data in the health care domains. This system is blockchain-based and provides data provenance, auditing, and control for shared medical data in cloud repositories among big data entities.



Fig

9.1: PREDICTION WITH REGRESSION MODELS

To run project double click on ‘run.bat’ file to get below screen

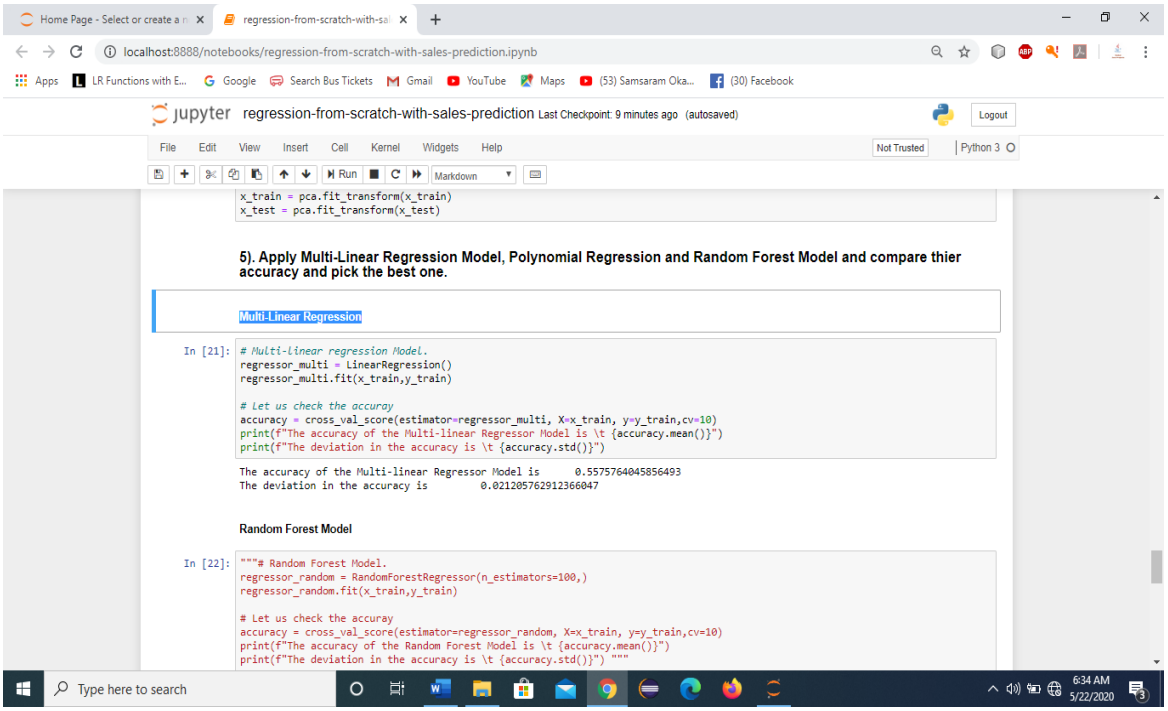


Fig 9.2: Train and Test models

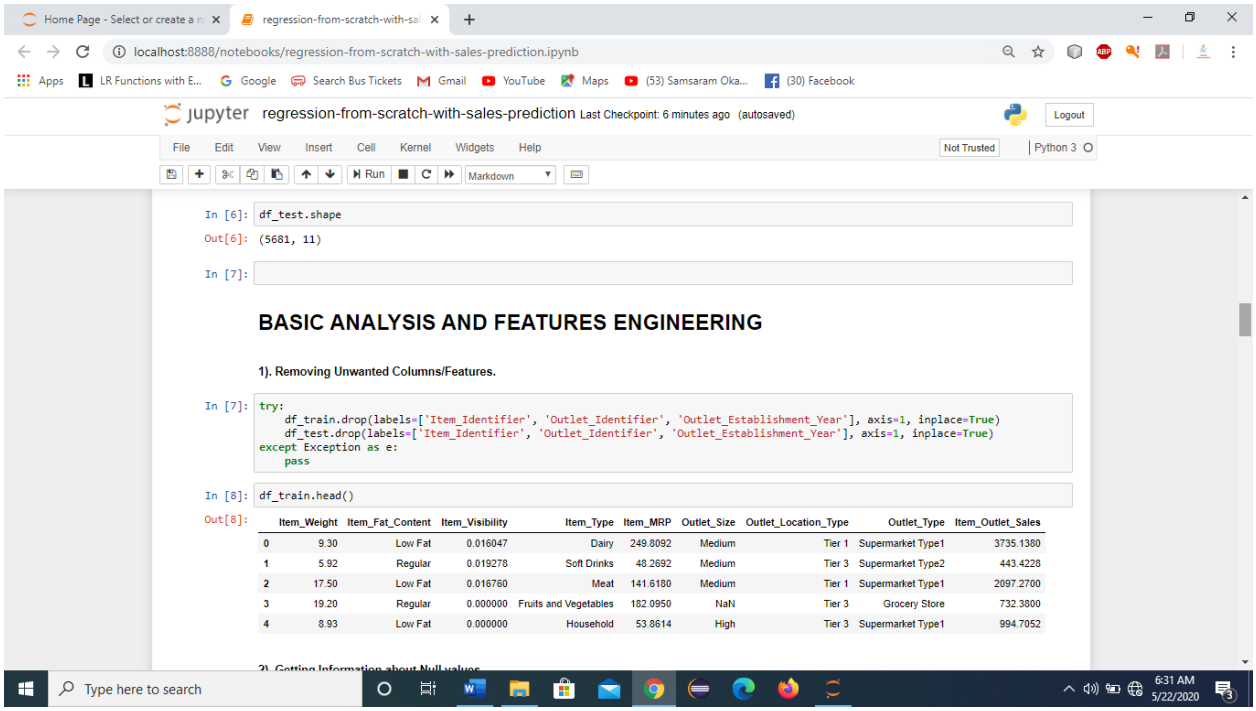
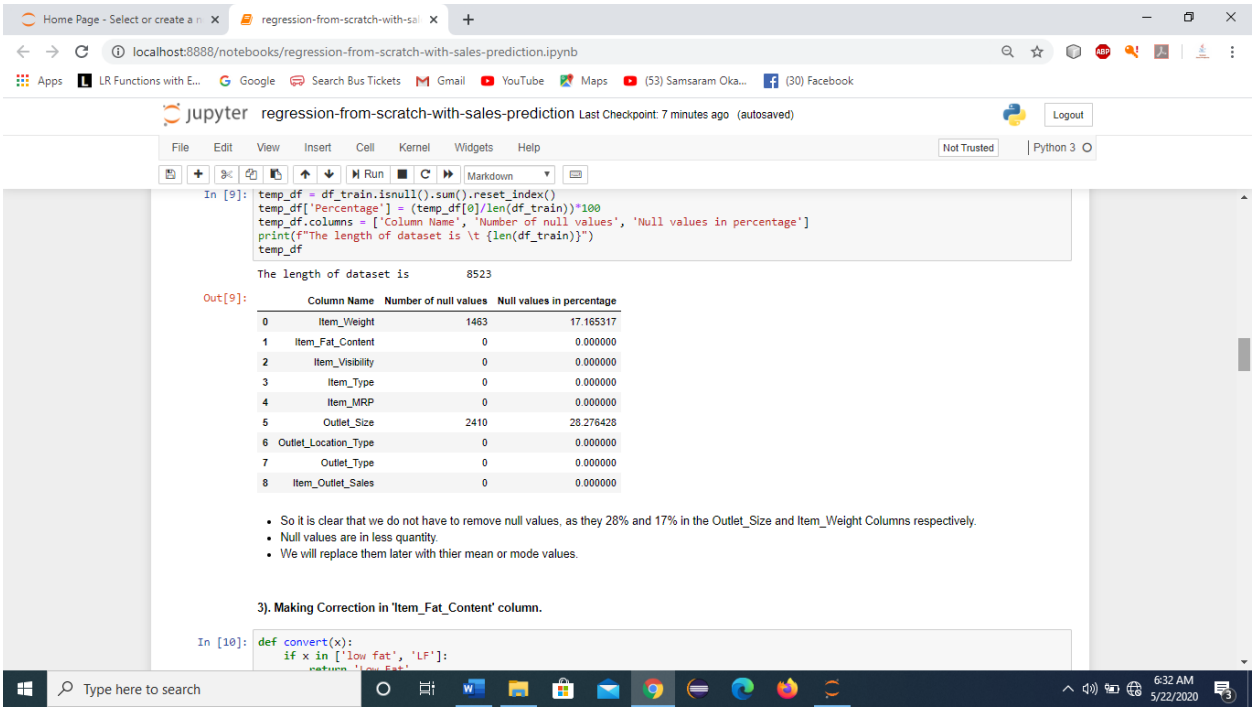


fig 9.3: Shape and Remove Null Values



In [9]:
```
temp_df = df_train.isnull().sum().reset_index()
temp_df['Percentage'] = (temp_df[0]/len(df_train))*100
temp_df.columns = ['Column Name', 'Number of null values', 'Null values in percentage']
print(f"The length of dataset is \t {len(df_train)}")
temp_df
```

```
The length of dataset is        8523
```

Out[9]:

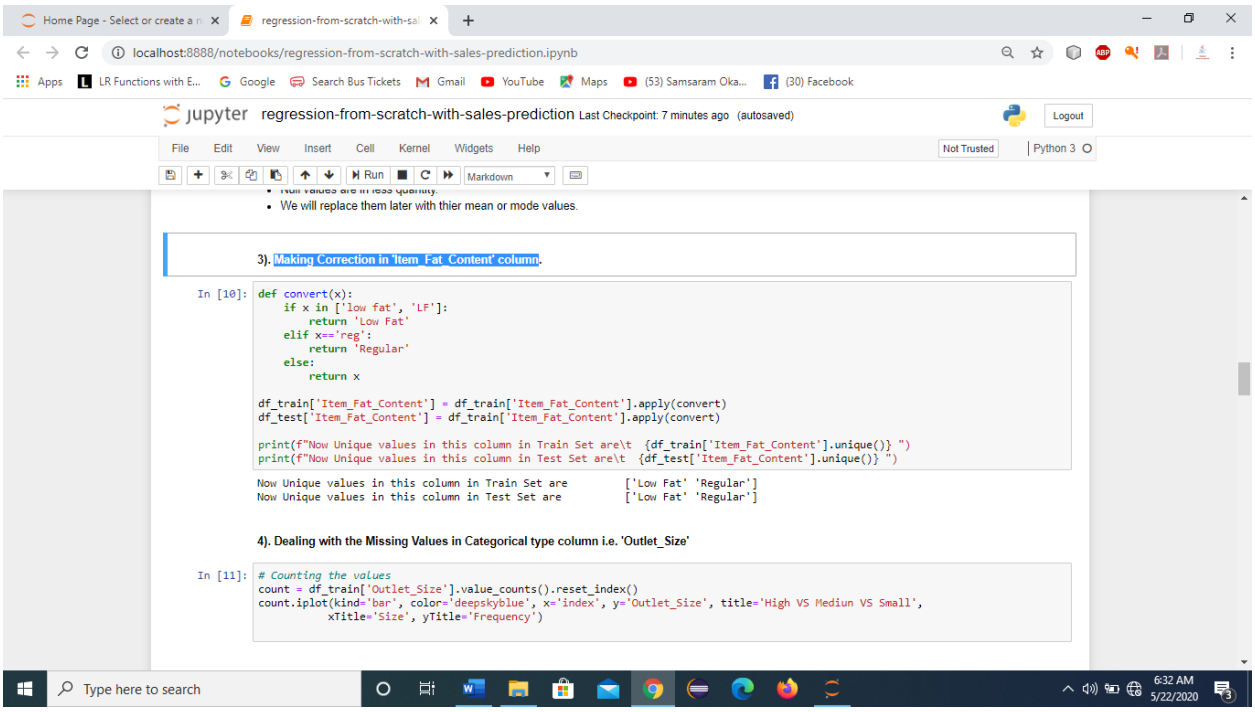
	Column Name	Number of null values	Null values in percentage
0	Item_Weight	1463	17.165317
1	Item_Fat_Content	0	0.000000
2	Item_Visibility	0	0.000000
3	Item_Type	0	0.000000
4	Item_MRP	0	0.000000
5	Outlet_Size	2410	28.276428
6	Outlet_Location_Type	0	0.000000
7	Outlet_Type	0	0.000000
8	Item_Outlet_Sales	0	0.000000

- So it is clear that we do not have to remove null values, as they 28% and 17% in the Outlet_Size and Item_Weight Columns respectively.
- Null values are in less quantity.
- We will replace them later with thier mean or mode values.

3). Making Correction in 'Item_Fat_Content' column.

In [10]:
```
def convert(x):
    if x in ['low fat', 'LF']:
```

fig 9.4:Complete description of dataset



- We will replace them later with thier mean or mode values.

3). Making Correction in 'Item_Fat_Content' column.

In [10]:
```
def convert(x):
    if x in ['low fat', 'LF']:
        return 'Low Fat'
    elif x=='reg':
        return 'Regular'
    else:
        return x

df_train['Item_Fat_Content'] = df_train['Item_Fat_Content'].apply(convert)
df_test['Item_Fat_Content'] = df_train['Item_Fat_Content'].apply(convert)

print(f"Now Unique values in this column in Train Set are\t  {df_train['Item_Fat_Content'].unique()} ")
print(f"Now Unique values in this column in Test Set are\t  {df_test['Item_Fat_Content'].unique()} ")
```

```
Now Unique values in this column in Train Set are         ['Low Fat' 'Regular']
Now Unique values in this column in Test Set are          ['Low Fat' 'Regular']
```

4). Dealing with the Missing Values in Categorical type column i.e. 'Outlet_Size'

In [11]:
```
# Counting the values
count = df_train['Outlet_Size'].value_counts().reset_index()
count.iplot(kind='bar', color='deepskyblue', x='index', y='Outlet_Size', title='High VS Mediun VS Small',
            xTitle='Size', yTitle='Frequency')
```

Fig 9.4: Making Correction in 'Item_Fat_Content' column

CONCLUSION AND FUTURE ENHANCEMENT

- This framework is to predict the future sales from given data of the previous year's using machine learning techniques. In this project, we have discussed how different machine learning models are build using different algorithms like Linear regression , Random forest model.
- These algorithms have been applied to predict the final result of sales. We have addresses in details about how the noisy data is been removed and the algorithms used to predict the result. Based on the sccuracy predicted by different models and we conclude that the random forest and multilinear regression approach is the best model. Our predictions help big mart to refine their methodologies and strategies which in turn helps them to increase their profit
- The ML algorithm that perform the best was XGBoost with RMSE = 1041 which got me in the first 25%. The next step will be looking at Hyperparameter Tuning and Ensembling.
- Hence, we propose a software tool for forecasting future sales volume based on the historical sales data. Using this tool, the accuracy of prediction for multiple linear regressions and random forests can be determined.

REFERENCES

1) Applied Linear Statistical Models", Fifth Edition by Kutner, Nachtsheim, Neter and L, Mc Graw Hill India, 2019.

2) Charles D. Kirkpatrick II and Julie R. Dahlquist, Technical Analysis: The Complete Resource for Financial Market Technicians , Pearson Education, Inc., 2020.

3) Demchenko, Yuri & de Laat, Cees &MembreyPeter,"Defining architecture components of the Big Data Ecosystem", 2018.

4) Blog: Big Sky, "The Data Analysis Process: 5 Steps To Better Decision Making", (URL: https://www.bigskyassociates.com/blog/bid/372186/The-Data- Analysis-Process-5-Steps-To-Better-DecisionMaking).

5) Blog: Dataaspirant, "HOW THE RANDOM FOREST ALGORITHM WORKS MACHINE LEARNING".

6) Mohit Gurnani, Yogesh Korke, Prachi Shah, Sandeep Udmale, "Forecasting of sales by using fusion of machine learning techniques",2017 International Conference on Data Management, Analytics and Innovation (ICDMAI), IEEE, October 2018.

7) Armstrong J,"Sales Forecasting", SSRN Electronic Journal, July 2018.

8) Samaneh Beheshti-Kashi, Hamid Reza, "A survey on retail sales forecasting and prediction in fashion markets", Systems Science & Control Engineering: An Open Access Journal.

9) Gopal Behera, Neeta Nain, "A Comparative Study of Big Mart Sales Prediction", 4th International Conference on Computer Vision and Image Processing, At MNIT Jaipur, September 2019.

10) Smola, A., &Vishwanathan, S. V. N. (2018). Introduction to machine learning. Cambridge University, UK, 32, 34.

11) U. N. Dulhare, "Prediction system for heart disease using Naive Bayes and particle swarm optimization,".

12) A. M. Rahat, A. Kahir, and A. K. M. Masum, "Comparison of Naive Bayes and SVM Algorithm based on Sentiment Analysis Using Review Dataset," in 2019 .

13) H. Pan, Y. Zhu, and L. Z. Xia, "Hierarchical PSO-adaboost based classifiers for fast and robust face detection," .

14) H. Kadam, R. Shevade, D. Ketkar, and S. Rajguru, "A forecast for big mart sales based on random forests and multiple linear regression,".

15) S. T. Zargar, J. Joshi, and D. Tipper, ''Prediction system for heart disease using Naive Bayes".

PUBLICATIONS

CONFERENCE

- International Conference on “Big mart sales using machine learning with data analysis” (ICICCI-21-0062)

- Paper ID: ICICCI-21-0062

ALL FOUR STUDENTS' ONE PAGE PROFILE

1.NAVYA RANI RAPETI(17K81A05H2)

NAVYA RANI RAPETI is currently pursuing her graduation from St Martin's Engineering College in the stream of Computer Science. She completed her intermediate from SR Junior College and 10th class from Presidency High School. She participated in various events, seminars and workshops during her graduation, some of them are

S.NO	EVENTS/SEMINARS/COURSES
1	Participated in Employability Skill development Program conducted by Zensar
2	Student of Smart Interviews
3	Participated in National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing"
4	Participated in Women online workshop on "Women in Cyber Security and Privacy in 2020"
5	Certification in Hacker rank (Python)
6	Certification in JavaScript By the Net Ninja in Cursa
7	Certification in Python Core in SoloLearn
8	Certification in MySQL database by the New Boston in Cursa
9	Certification in CyberSecurity by PacketHacks in Cursa
10	Participated in Anti-Drug Campaign conducted by Lush life Bistro

2.CHITLAPALLY BHAVANI(17K81A05D3)

CHITLAPALLY BHAVANI is currently pursuing her graduation from St Martin's Engineering College in the stream of Computer Science. She completed her intermediate from Sri Chaitanya Junior College and 10th class from Vivekananda high School. She participated in various events, seminars and workshops during her graduation, some of them are:

S.NO	EVENTS/SEMINARS/COURSES
1	Participated in Employability Skill development Program conducted by Zensar
2	Participated in National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing"
3	Participated in Women online workshop on "Women in Cyber Security and Privacy in 2020"
4	Certification in Hacker rank (Python)
5	Certification in Learn to code in Python3 in Udemy
6	Certification in Introduction to AI from ElementsofAI
7	Certification in The fundamentals of Digital Marketing in Google Digital Unlocked
8	Certification in MySQL database by The new Boston in Cursa
9	Certification in Programming with PHP for beginners in Cursa
10	Certification in 30 days to learn HTML and CSS in Cursa

3.GANJI KRISHNA SRI (17K81A05D9)

GANJI KRISHNA SRI is currently pursuing her graduation from St Martin's Engineering College in the stream of Computer Science. She completed her intermediate from Sri Chaitanya Junior College and 10th class from Sri Chaitanya Techno School. She participated in various events, seminars and workshops during her graduation, some of them are:

S.NO	EVENTS/SEMINARS/COURSES
1	Participated in Employability Skill development Program conducted by Zensar
2	Participated in National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing"
3	Participated in Women online workshop on "Women in Cyber Security and Privacy in 2020"
4	Certification in Management Managerial Accounting by nptelhrd in Cursa
5	Certification in Principles of Construction Management by nptelhrd in Cursa
6	Certification in Principles of Copy Writing by Appy Pie in Cursa
7	Certification in 30 days to learn HTML and CSS in Cursa
8	Certification in HTML by EJ Media in Cursa

4.KRISHNA TEJA (16K81A05E2)

TALAKOKKULA ASHWITHA is currently pursuing her graduation from St Martin's Engineering College in the stream of Computer Science. She completed her intermediate from Sri Chaitanya Junior College and 10th class from Sri Chaitanya Techno School. She participated in various events, seminars and workshops during her graduation, some of them are:

S.NO	EVENTS/SEMINARS/COURSES
1	Participated in Employability Skill development Program conducted by Zensar
2	Participated in National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing"
3	Participated in Women online workshop on "Women in Cyber Security and Privacy in 2020"
4	Certification in Programming with PHP for beginners by the Net Ninja in Cursa
5	Certification in Basic of AWS concepts by ExamPro in Cursa
6	Certification in Artificial Intelligence by Crash Course in Cursa
7	Certification in Python for Beginners in SoloLearn
8	Certification in JavaScript in SoloLearn

A

PROJECT REPORT

On

SEMI SUPERVISED MACHINE LEARNING APPROACH FOR DDOS DETECTION

Submitted by

1)Mr.RohanRaj(17K81A05G5) 2)Mr.RajeshKumar(17K81A05H0)

3)Mr.AjayKumar(17K81A05H5) 4)Ms.Lahari(17K81A05H9)

in partial fulfillment for the award of the

degree of

BACHELOR OF TECHNOLOGY

IN

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Under the Guidance of

Mr.EdigaLingappa

Assistant Professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ST.MARTIN'S ENGINEERING COLLEGE
An Autonomous Institute

Dhulapally, Secunderabad – 500 100

JUNE 2021

BONAFIDE CERTIFICATE

This is to certify that the project entitled SEMI SUPERVISED MACHINE LEARNING APPROACH FOR DDOS DETECTION, is being submitted by **Mr.ROHAN RAJ[17K81A05G5], Mr.RAJESHKUMAR[17K81A05H0], Mr.AJAYKUMAR [17K81A05H5], Ms.LAHARI[17K81A05H9]** in partial fulfillment of the requirement for the award of the degree of **BACHELOR OF TECHNOLOGY IN COMPUTER SCIENCE AND ENGINEERING** is recorded of bonafide work carried out by them. The result embodied in this report have been verified and found satisfactory.

Ediga Lingappa
Department of CSE

Head of the Department
Dr.M.NARAYANAN
Department of CSE

Internal Examiner

External Examiner

Place:

Date:

DECLARATION

We, the student of **Bachelor of Technology** in Department of Computer Science and Engineering, session: 2017 – 2021, St. Martin's Engineering College, Dhulapally, Kompally, Secunderabad, hereby declare that work presented in this Project Work entitled **SEMI SUPERVISED MACHINE LEARNING APPROACH FOR DDOS DETECTION** is the outcome of our own bonafide work and is correct to the best of our knowledge and this work has been undertaken taking care of Engineering Ethics. This result embodied in this project report has not been submitted in any university for award of any degree.

N. ROHAN RAJ 17K81A05G5

P. RAJESH KUMAR 17K81A05H0

S. AJAY KUMAR 17K81A05H5

V. LAHARI 17K81A05H9

ABSTRACT

The appearance of malicious apps is a serious threat to the Android platform. Most types of network interfaces based on the integrated functions, steal users' personal information and start the attack operations. In this paper, we propose an effective and automatic malware detection method using the text semantics of network traffic. In particular, we consider each HTTP flow generated by mobile apps as a text document, which can be processed by natural language processing to extract text-level features. Later, the use of network traffic is used to create a useful malware detection model. We examine the traffic flow header using N-gram method from the natural language processing (NLP). Then, we propose an automatic feature selection algorithm based on chi-square test to identify meaningful features. It is used to determine whether there is a significant association between the two variables. We propose a novel solution to perform malware detection using NLP methods by treating mobile traffic as documents. We apply an automatic feature selection algorithm based on N-gram sequence to obtain meaningful features from the semantics of traffic flows. Our methods reveal some malware that can prevent detection of antiviral scanners. In addition, we design a detection system to drive traffic to your own-institutional enterprise network, home network, and 3G / 4G mobile network. Integrating the system connected to the computer to find suspicious network behaviours.

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible and whose encouragements and guidance have crowded effects with success.

We extended our deep sense of gratitude to Principal**,** **Dr. P. SANTOSH KUMAR PATRA**, St. Martin's Engineering College, Dhulapally, for permitting us to undertake this project.

We are also thankful to **Dr. M.NARAYANAN**, Head of the Department, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

We would like to thank **Dr. T. POONGOTHAI,** Department of Computer Science and Engineering (AI & ML) for her encouragement and insightful comments and as well as our project coordinators **Dr. B.RAJALINGAM**, Associate Professor and **Mr. J.SUDHAKAR**, Assistant Professor, in Department of Computer Science and Engineering for their valuable support.

We would like to express our sincere gratitude and indebtedness to our project supervisor Mr.Ediga Lingappa, Associate Professor, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

Finally, we express thanks to all those who have helped us successfully to completing this project. Furthermore, we would like to thank our family and friends for their moral support and encouragement.

We express thanks to all those who have helped us in successfully completing the project.

N. Rohan Raj 17K81A05G5

P. Rajesh Kumar 17K81A05H0

S. Ajay Kumar 17K81A05H5

V. Lahari 17K81A05H9

TABLE OF CONTENTS

LIST OF OUTPUT SCREENS

LIST OF FIGURES

LIST OF ABBREVIATIONS

HTTP	HYPER TEXT TRANSFER PROTOCOL
TCP	TRANSMISSION CONTROL PROTOCOL
UDP	USER DATAGRAM PROTOCOL
DDOS	DISTRIBUTED DENIAL OF SERVICE
NLP	NATURAL LANGUAGE PROCESSING
SVM	SUPPORT VECTOR MACHINE

GLOSSARY OF TERMS

1. DDOS DETECTION
2. CO-CLUSTERING
3. ENTROPY ANALYSIS
4. INFORMATION GAIN RATIO
5. NSL-KDD
6. UNB ISCX 12
7. UNSW-NB 15
8. SUPERVISED
9. UNSUPERVISED
10. NATURAL PROCESSING LANGUAGE
11. N-GRAM
12. HYPERTEXT TRANSFER PROTOCAL
13. AUTHENTICATION

1. INTRODUCTION

1.1 PROJECT OVERVIEW

Distributed denial of service (DDoS) is one of the cyber-attack, which remains as a major attack on internet for past many years. DDoS detection based on Machine Learning techniques such as, Supervised and Unsupervised techniques has been already implemented which has some drawbacks like low detection accuracy and high false positive rates. In this paper, DDoS detection based on Semi-Supervised Machine learning technique is presented which is the combination of both supervised and unsupervised techniques that provides better results compared to the existing approaches. Unsupervised part consists of some estimation steps including clustering which reduces the false positive rates and increases the accuracy by reducing irrelevant data. In supervised part Random algorithm is used to accurately classify the DDoS attack data and it also reduces the false positive rate of unsupervised part.

In this approach, initially the network traffic data is read. Average entropy is estimated for the data and then by applying the clustering algorithm 3 clusters are formed. By making use of average entropy of data, the information gain ratio is calculated. Anomalous cluster is formed by combining the clusters which having highest gain ratio value. Random forest algorithm is applied for the obtained anomalous cluster to accurately classify the data and detect the DDoS attack. For better evaluation of performance of proposed approach, NSL-KDD network traffic dataset is used. The NSL-KDD dataset contains the different types of attack data.

1.2 PROJECT OBJECTIVE

The Objective of this project is to propose an effective and automatic malware detection method using the text semantics of network traffic. In particular, we consider each HTTP flow generated by mobile apps as a text document, which can be processed by natural language processing to extract text-level features. Later, the use of network traffic is used to create a useful malware detection model. We examine the traffic flow header using N-gram method from the natural language processing(NLP). Then, we propose an automatic feature selection algorithm based on chi-square test to identify meaningful features. It is used to determine whether there is a significant association between the two variables. We propose a novel solution to perform malware detection using NLP methods by treating mobile traffic as documents. The proposed approach consists of five major steps: Datasets preprocesing, estimation of network traffic Entropy, online co-clustering, information gain ratio computation and network traffic classification.

1.3 ORGANISATION OF CHAPTERS

Besides the introduction, the thesis is organized in other six chapters as follows:

Chapter 2, LITERATURE SURVEY: the review is made in the context of Semi Supervised ML Approaches for DDOS Detection with a particular attention on those implementations that assess the scalability and performances or their implementations. Most of the related work is on Machine Learning Approaches like Supervised, Unsupervised and Semi-Supervised, whereas a small part is on Network

traffic entropy estimation and Data Mining Techniques. It will be possible to notice that only a small subset of the literature actually focuses on the analysis of the systems in mass crises scenarios. Chapter 3, SOFTWARE AND HARDWARE REQUIREMENTS: this chapter discuss about the software and hardware required for the execution of the project. Chapter 4, SOFTWARE DEVELOPMENT ANALYSIS: this chapter explains the assumptions and technical specifications of the project. Chapter 5, PROJECT SYSTEM DESIGN: this chapter explains all the software development process with DFD and UML diagrams clearly. Chapter 6, PROJECT CODING: this chapter explains the design of the system, roles and responsibilities, as well as the requirements of a Semi-Supervised ML Approach using Co-Clustering Algorithm. Chapter 7, PROJECT TESTING: this chapter explains various test cases to test the project working. Chapter 8, OUTPUT SCREENS: explains a step by step process of the project execution. Chapter 9, EXPERIMENTAL RESULTS: tests and results are shown and explained in this chapter. The results are analyzed in the context of the thesis project and followed by discussion on systems throughput and resiliency, as well as the approaches to testing and analysis. Chapter 10, CONCLUSION AND FUTURE ENHANCEMENT: the chapter ends the project with a short summary of the main concepts mentioned in the thesis as well as the relevant results.

2. LITERATURE SURVEY

A literature survey or a literature review in a project report is that section which shows the various analyses and research made in the field of your interest and the results already published, taking into account the various parameters of the project and the extent of the project.

It is the most important part of your report as it gives you a direction in the area of your research. It helps you set a goal for your analysis - thus giving you your problem statement.

When you write a literature review in respect of your project, you have to write the researches made by various analysts - their methodology (which is basically their abstract) and the conclusions they have arrived at. You should also give an account of how this research has influenced your thesis.

2.1 SURVEY ON BACKGROUND

Various methodologies and techniques for reducing the effects of DDoS attacks in different network environments have been proposed and evaluated. The authors identified users' requests or demands to a specific resource and their communicative data. Then samples of such requests are sent to the detection systems to be judged for abnormalities. Also, Liu and Gu have used Learning Vector Quantisation (LVQ) neural networks to detect attacks. This is a supervised version of quantisation, which can be used for pattern recognition, multi-class classification and data compression tasks.

Akilandeswari and Shalinie have introduced a Probabilistic Neural Network Based Attack Traffic Classification to detect different DDoS attacks. However, the authors focus on separating Flash Crowd Event from Denial of Service Attacks. Gupta, Joshi and Misra have used a neural network to detect the number of zombies that have been involved in DDoS attacks. The process work- load is based on prediction using a feed-forward neural network. They listed the most popular DDoS attacks on cloud systems, classified and discussed intrusion detection systems along with their challenges.

Classification of the intrusion detection systems was knowledge-based and anomaly-based. Sub categorizations were done based on the scalability of the management system, user authentication and the response mechanism. Rashmi V. Deshmukh and Kailas K. Devadkar discussed DDoS attacks and provided a taxonomy of attacks in the cloud environment They classified defense mechanisms based on prevention, detection, and response to detection techniques Somani et al. presented insights into the JAC : A JOURNAL OF COMPOSITION THEORY Volume XIII, Issue V, MAY 2020 ISSN : 0731-6755 Page No:265 characterization, prevention, detection, and mitigation mechanisms of DDoS attacks in the cloud environment. A taxonomy of DDoS solutions was also presented and the solutions were categorized under prevention, detection, and mitigation. They concluded by discussing considerations to be made in selecting a defense solution.

Mahjabin et al. presented a review on different DDoS attacks. They discussed attack phases in a DDoS attack, variations and evolutions of attacks as well as attackers' targets and motivations. They classified and analyzed prevention and mitigation techniques based on their underlying principle of operation. Underlying principles for the prevention methods reviewed were filters, secure overlay service, load balancing, honey pots, and awareness-based prevention systems. Mitigation techniques were also categorized broadly based on detection, response, and tolerance-based systems. They concluded by listing the key features, advantages, and limitations of the prevention and detection mechanisms reviewe

Kalkan et al. presented DDoS attack scenarios in software defined networks (SDNs). Solutions to attacks were also broadly classified as intrinsic (having inherent properties) and extrinsic (depending on external factors). Solutions were also classified according to their defense function (detection, mitigation, and both detection and mitigation) and SDN switch intelligence (capable switch vs. dumb switch).

Zare et al. came up with a paper reviewing papers on DDoS attacks and countermeasures between the years 2000 and 2016. They discussed intrusion detection systems and analyzed countermeasures against DDoS attacks based on the location of the defense mechanism; source-end, core-end, victim-end, and distributed defense. There has been a great number of surveys on DDoS defenses in previous years. DDoS attacks are on the rise and there is a constant need to present the state-of-the-art defense mechanisms to aid researches in their attempt to combat such attacks. In previous surveys however, most defense categorizations were done based on their underlying principle of operation and not the function they perform in defending against attacks. Defense mechanisms were also not discussed into detail to give a greater understanding of their operation. Also, the individual defense mechanisms were not compared with each other, but rather, comparisons were mainly done based on the underlying principle of operation or location of deployment of the defense mechanism.

2.2 CONCLUSION ON SURVEY

The performances of network intrusion detection approaches, in general, rely on the distribution characteristics of the underlying network traffic data used for assessment. The DDoS detection approaches in the literature are under two main categories:

- Unsupervised approaches
- Supervised approaches.

Depending on the benchmark datasets used, unsupervised approaches often suffer from high false positive rate and supervised approach cannot handle large amount of network traffic data and their performances are often limited by noisy and irrelevant network data. Therefore, the need of combining both, supervised and unsupervised approaches arises to overcome DDoS detection issues.

3. SOFTWARE AND HARDWARE REQUIREMENTS

3.1 SOFTWARE REQUIREMENTS

Functional Requirements

Graphical User interface with the User.

Software Requirements

For developing the application, the following are the Software Requirements:

1. Python
2. Django
3. MySQL
4. MySQL client
5. Xampp Server

Operating Systems supported

Windows XP, 7, 8, 10

Technologies and Languages used to Develop

Python

Debugger and Emulator

Any Browser (Particularly Chrome)

3.2 HARDWARE REQUIREMENTS

For developing the application, the following are the Hardware Requirements:

- Processor: Pentium IV or higher
- RAM: 4GB
- Space on Hard Disk: 512GB

4. SOFWARE DEVELOPMENT ANALYSIS

4.1 OVERVIEW OF PROBLEM

Even though advanced Machine Learning(ML)techniques have been adopted for DDoS detection, the attack remains a major threat of the Internet. Most of the existing ML-based DDoS detection approaches are under two categories: supervised and unsupervised. Supervised ML approaches for DDoS detection rely on availability of labeled network traffic datasets. Whereas, unsupervised ML approaches detect attacks by analyzing the in coming network raffic. Both approaches are challenged by large amount of network traffic data, low detection accuracy and high false positive rates.

In this project, we present an online sequential semi-supervised ML approach for DDoS detection based on network Entropy estimation, Co-clustering, Information Gain Ratio and Exra-Trees algorithm.

4.2 DEFINE THE PROBLEM

The existing Machine Learning based DDoS detection approaches can be divided into three categories. Supervised ML approaches that use generated labeled network traffic datasets to build the detection model. Two major issues are facing the supervised approaches. First, the generation of labeled network traffic datasets is costly in terms of computation and time. Without a continuous update of their detection models, the supervised machine learning approaches are unable to predict the new legitimate and attack behaviors. Second, the presence of large amount of irrelevant normal data in the incoming network traffic is noisy and reduces the performances of supervised ML classifiers.

Unlike the first category, in the unsupervised approaches no labelled dataset is needed to build the detection model. The DDoS and the normal traffics are distinguished based on the analysis of their underlying distribution characteristics. However, the main drawback of the unsupervised approaches is the high false positive rates. In the high dimensional network traffic data the distance between points becomes meaningless and tends to homogenize. This problem, known as ‘the curse of dimensionality’, prevents unsupervised approaches to accurately detect attacks .

The semi-supervised ML approaches are taking advantages of both supervised and unsupervised approaches by the ability to work on labeled and unlabeled datasets. Also, the combination of supervised and unsupervised approaches allows to increase accuracy and decreases the false positive rates. However, semi-supervised approaches are also challenged by the drawbacks of both approaches. Hence, the semi-supervised approaches require a sophisticated implementation of its components in order to overcome the drawbacks of supervised and unsupervised approaches. In this paper we present an online sequential semisupervised ML approach for DDoS detection. A time based sliding window algorithm is used to estimate the entropy of the network header features of the incoming network traffic. When the entropy exceeds its normal range, the unsupervised co-clustering algorithm splits the incoming network traffic into three clusters. Then, an information gain ratio is computed based on the average entropy of the network header features between the network traffic subset of the current time window and each one of the obtained clusters. The network traffic data clusters that produce high information gain ratio are considered as anomalous and they are selected for preprocessing and classification using an ensemble classifiers based on the Extra-Trees algorithm.

To better evaluate the performance of the proposed approach three public network traffic datasets are used in the experiment, namely the NSL-KDD, the UNB ISCX IDS 2012 dataset and the UNSW-NB15. The experimental results are satisfactory when compared with the state-of-the-art DDoS detection methods.

4.3 MODULES OVERVIEW

There are four modules used in this project. They are listed as below :

- User Applications and URLs
- DDOS Attack Deduction
- Classifications of DDOS attack
- Graphical analysis

4.4 DEFINE THE MODULES

1. User Applications and URLs

This module defines that the urls are collected as the dataset in the system and also defines the types of devices used by the users and applications in it.

2. DDOS Attack Deduction

This module defines the deduction of DDOS attack by generating the network traffic and examines the network flow using Natural Language Processing.

3. Classification of DDOS Attack

This modules defines the classification of DDOS attacks based on the type of attack and labels using the algorithm.

4. Graphical Analysis

This module can be defined as the analyzing and the visualization of the no. of types of DDOS attacks by showing in different types of charts.

4.5 MODULE FUNCTIONALITY

1. User Applications and URLs

A user handles some various types of smart phones, desktops, laptops and tablets. If any kind of devices attacks for some unauthorized Malware softwares, those specific URLs are collected as the dataset for the system. In these threats are for user personal data including personal contacts, bank account numbers and any kind of personal documents with which hacking is possible.

2. DDOS Attack Deduction

A user may search the any link on the internet. Notably, not all network traffic data generated by malicious apps correspond to malicious traffic. Many malware take the form of repackaged benign apps. So, Malware can also contain the basic functions of a benign app. Subsequently, the network traffic they generate can be characterized by mixed benign and malicious network traffic. We examine the traffic flow header using Co-clustering algorithm from the natural language processing (NLP).

3. Classifications of DDOS Attack

Here, we compare the classification performance of Co-clustering algorithm with other popular machine learning algorithms. We have selected several popular classification algorithms. For all algorithms, we attempt to use multiple sets of parameters to maximize the performance of each algorithm. Using Co-clustering algorithm, algorithms classification for malware bag-of-words weightage.

4. Graphical Analysis

The graph analysis is done by the values taken from the result analysis part and it can be analyzed by the graphical representations, spline chart, bar chart and column chart in the system.

5. PROJECT SYSTEM DESIGN

5.1 DFDS IN CASE OF DATABASE PROJECTS

A data flow diagram shows the way information flows through a process or system. It includes data inputs and outputs, data stores, and the various sub processes the data moves through. DFDs are built using standardized symbols and notation to describe various entities and their relationships.

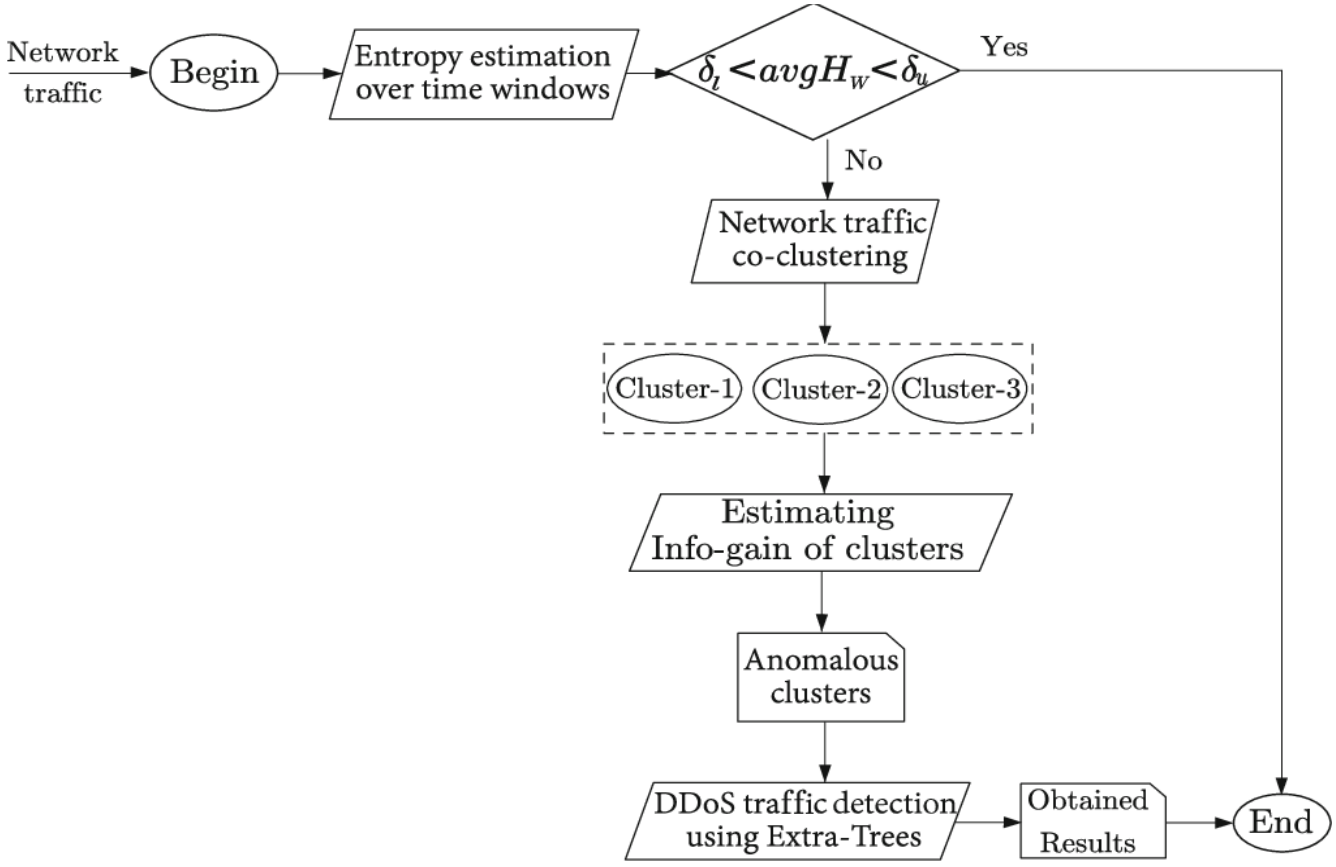


Fig.5.1.1: Data Flow Diagram

Data flow diagrams visually represent systems and processes that would be hard to describe in a chunk of text. You can use these diagrams to map out an existing system and make it better or to plan out a new system for implementation. Visualizing each element makes it easy to identify inefficiencies and produce the best possible system.

5.2 UML DIAGRAMS

UML stands for Unified Modeling Language. UML is a standardized general-purpose modeling language in the field of object-oriented software engineering. The standard is managed, and was created by, the Object Management Group.

The goal is for UML to become a common language for creating models of object oriented computer software. In its current form UML is comprised of two major components: a Meta-model and a notation. In the future, some form of method or process may also be added to; or associated with, UML.

The Unified Modelling Language is a standard language for specifying, Visualization, Constructing and documenting the artifacts of software system, as well as for business modelling and other non-software systems.

The UML represents a collection of best engineering practices that have proven successful in the modelling of large and complex systems.

The UML is a very important part of developing objects oriented software and the software development process. The UML uses mostly graphical notations to express the design of software projects.

GOALS:

The Primary goals in the design of the UML are as follows:

1. Provide users a ready-to-use, expressive visual modeling Language so that they can develop and exchange meaningful models.
2. Provide extendibility and specialization mechanisms to extend the core concepts.
3. Be independent of particular programming languages and development process.
4. Provide a formal basis for understanding the modeling language.
5. Encourage the growth of OO tools market.
6. Support higher level development concepts such as collaborations, frameworks, patterns and components.
7. Integrate best practices.

ARCHITECTURE DIAGRAM

An architecture diagram describes what you're building, how stakeholders interact with it, and where constraints lie. A design diagram explains how to build it.

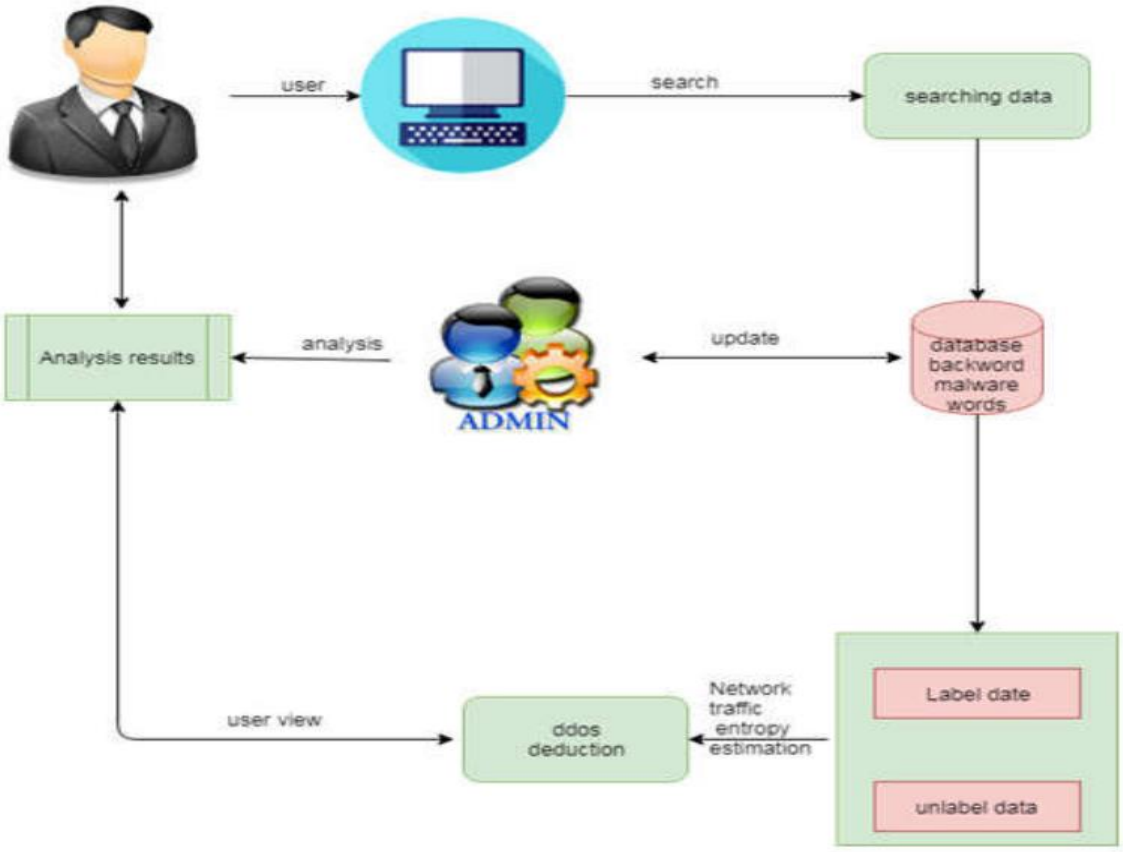


Fig.5.2.1: Architecture Diagram

Architectural diagrams show systems. Displaying information visually allows the viewer to see everything at a glance, including how things interact. This is especially useful when making changes: You'll be able to see the downstream effects of a given change more clearly.

USE CASE DIAGRAM

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.

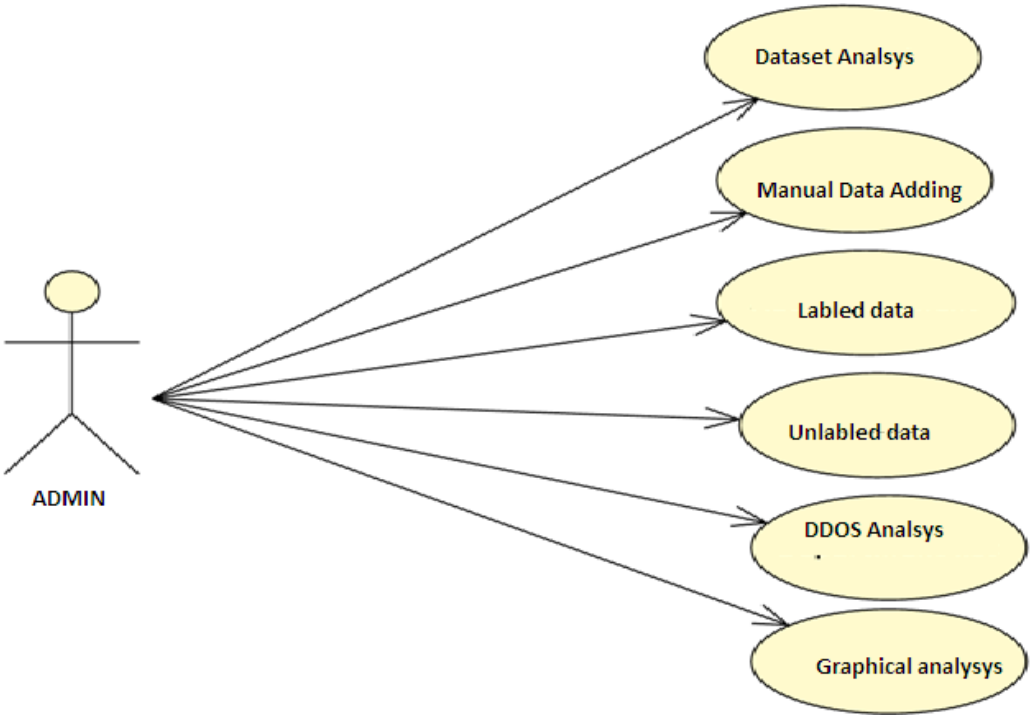


Fig.5.2.2: Use Case Diagram

CLASS DIAGRAM

In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.

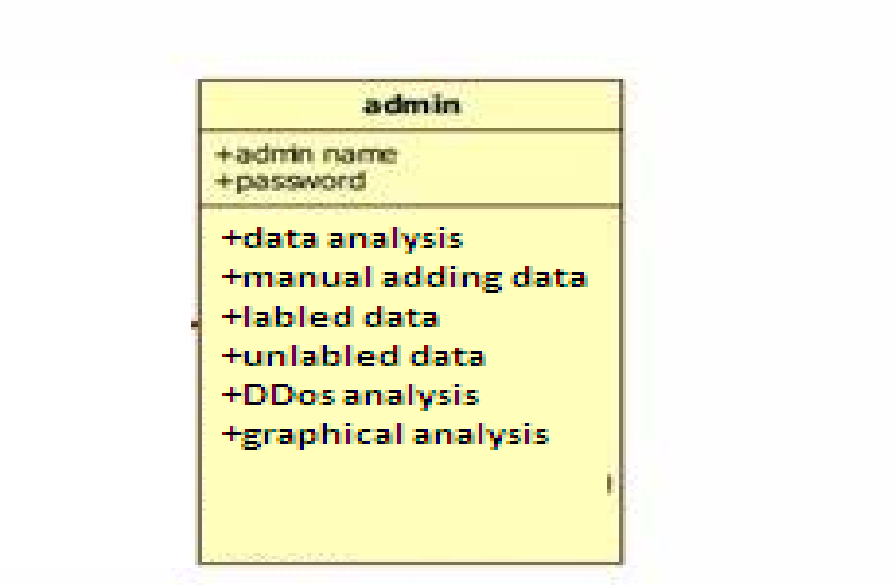


Fig.5.2.3 Class Diagram

SEQUENCE DIAGRAM

A sequence diagram in Unified Modelling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.

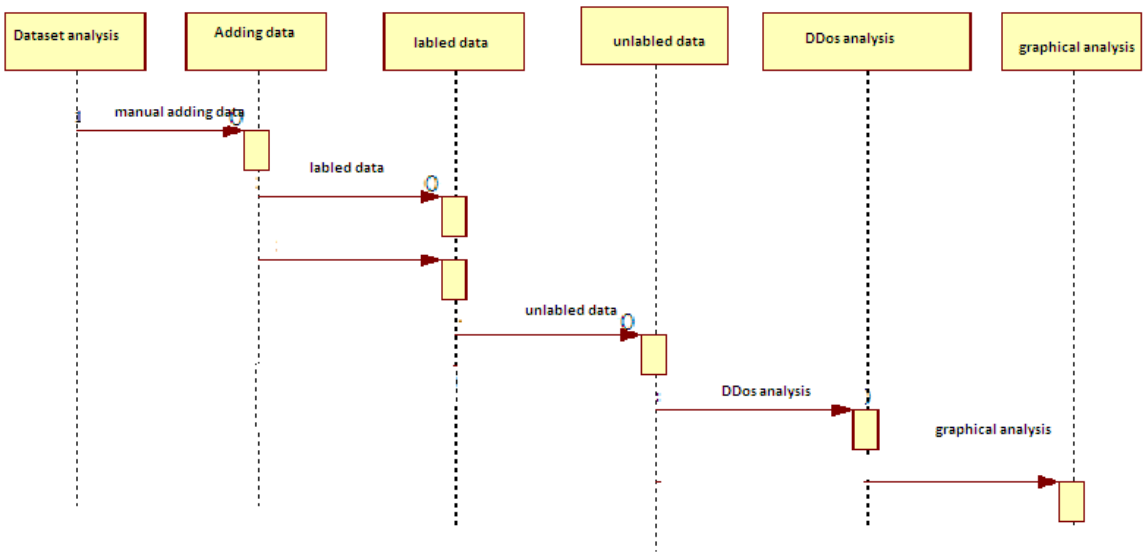


Fig.5.2.4: Sequence Diagram

ACTIVITY DIAGRAM

Activity diagram is basically a flowchart to represent the flow from one activity to another activity. The activity can be described as an operation of the system.

The control flow is drawn from one operation to another. This flow can be sequential, branched, or concurrent. Activity diagrams deal with all type of flow control by using different elements such as fork, join, etc.

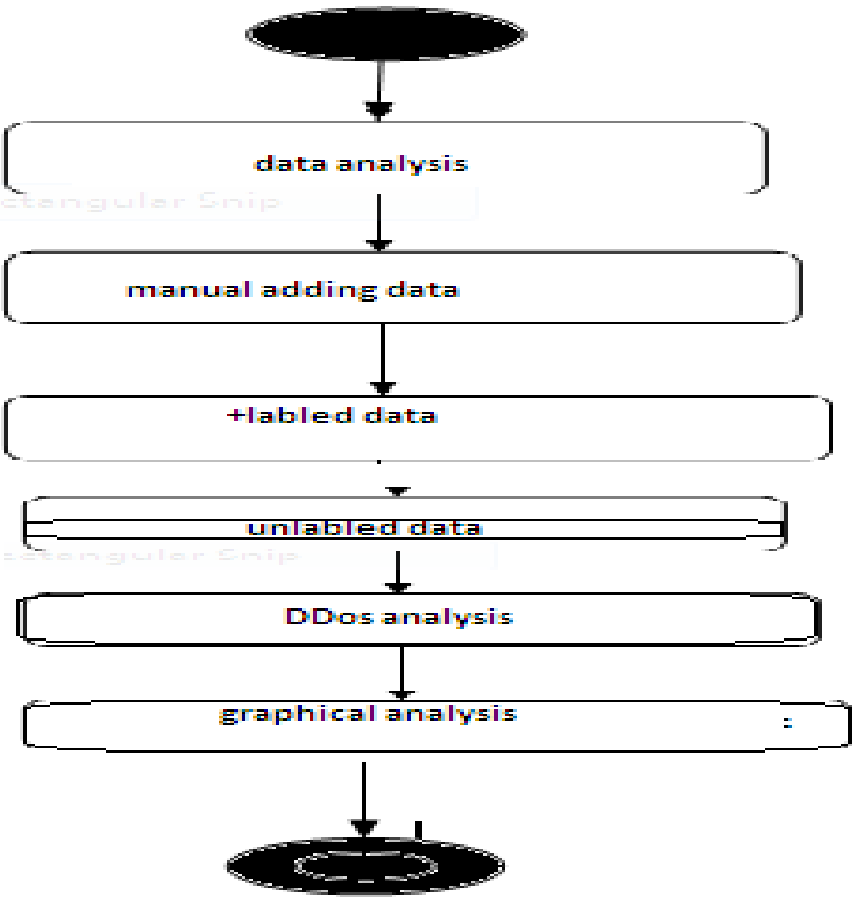


Fig.5.2.5: Activity Diagram

6. PROJECT CODING

6.1 CODE TEMPLATES

```
import re


from django.db.models import Q, Count
from django.shortcuts import render, redirect

# Create your views here.
from data_admins.models import ddos_dataset

def index(request):
    if request.method == "POST":
        if request.method == "POST":
            usid = request.POST.get('username')
            pswd = request.POST.get('password')
            if usid == 'admin' and pswd == 'admin':
                return redirect('userpage')
    return render(request,'index.html')

def register(request):
    return render(request,'register.html')

def userpage(request):
    obj = ddos_dataset.objects.all()
    return render(request,'userpage.html',{'object':obj})

def add_data(request):
    attack1 = []
    attack2, attack3, attack4, attack5, attack6, attack7, attack8, attack9 = [], [], [], [], [], [], [], []
    ans = ''
    txt = ''
    splt = ''
    if request.method == "POST":
        txt = request.POST.get("name")

        splt = (re.findall(r"[\w']+", str(txt)))

    for f in splt:
        if f in ('IPid','FDDI','x25','rangingdistance'):
            attack1.append(f)
        elif f in ('tcpchecksum','mtcp','controlflags','tcpoffset','tcpport'):
            attack2.append(f)
        elif f in ('ICMPID','udptraffic','udpunicorn','datagramid','NTP','RIP','TFTP'):
```

```
attack3.append(f)
elif f in ('GETID','POSTID','openBSD','appid','sessionid','transid','physicalid'):
            attack4.append(f)
        elif f in ('SYN','ACK','synpacket','sycookies'):
            attack5.append(f)
        elif f in ('serverattack','serverid','blockbankwidth'):
            attack6.append(f)
        elif f in ('monlist','getmonlist','NTPserver'):
            attack7.append(f)
        elif f in ('portid','FTPID','tryion','fragflag'):
            attack8.append(f)
        elif f in ('malwareid','gethttpid','httpid'):
            attack9.append(f)

    if len(attack1) > len(attack2) and len(attack1) > len(attack3) and len(attack1) > len(attack4) and len(
            attack1) > len(attack5) and len(attack1) > len(attack6) and len(attack1) > len(attack7) and len(
            attack1) > len(attack8) and len(attack1) > len(attack9):
        ans = "Ip Fragment Attack"
    elif len(attack2) > len(attack1) and len(attack2) > len(attack3) and len(attack2) > len(attack4) and len(
            attack2) > len(attack5) and len(attack2) > len(attack6) and len(attack2) > len(attack7) and len(
            attack2) > len(attack8) and len(attack2) > len(attack9):
        ans = "TCP Based Attack"
    elif len(attack3) > len(attack2) and len(attack3) > len(attack1) and len(attack3) > len(attack4) and len(
            attack1) > len(attack5) and len(attack1) > len(attack6) and len(attack1) > len(attack7) and len(
            attack1) > len(attack8) and len(attack1) > len(attack9):
        ans = "UDP Based Attack"
    elif len(attack4) > len(attack2) and len(attack4) > len(attack3) and len(attack4) > len(attack1) and len(
            attack4) > len(attack5) and len(attack4) > len(attack6) and len(attack4) > len(attack7) and len(
            attack4) > len(attack8) and len(attack4) > len(attack9):
        ans = "Layer Based Attack"
    elif len(attack5) > len(attack2) and len(attack5) > len(attack3) and len(attack5) > len(attack4) and len(
            attack5) > len(attack1) and len(attack5) > len(attack6) and len(attack5) > len(attack7) and len(
            attack5) > len(attack8) and len(attack5) > len(attack9):
        ans = "SYN Floods Attack"
    elif len(attack6) > len(attack2) and len(attack6) > len(attack3) and len(attack6) > len(attack4) and len(
            attack6) > len(attack5) and len(attack6) > len(attack1) and len(attack6) > len(attack7) and len(
            attack6) > len(attack8) and len(attack6) > len(attack9):
        ans = "Slowloris"
    elif len(attack7) > len(attack2) and len(attack7) > len(attack3) and len(attack7) > len(attack4) and len(
            attack7) > len(attack5) and len(attack7) > len(attack6) and len(attack7) > len(attack1) and len(
            attack7) > len(attack8) and len(attack7) > len(attack9):
        ans = "NTP Amplification"
    elif len(attack8) > len(attack2) and len(attack8) > len(attack3) and len(attack8) > len(attack4) and len(
            attack8) > len(attack5) and len(attack8) > len(attack6) and len(attack8) > len(attack7) and len(
            attack8) > len(attack1) and len(attack8) > len(attack9):
        ans = "Ping of Death Attack"
    elif len(attack9) > len(attack2) and len(attack9) > len(attack3) and len(attack9) > len(attack4) and len(
            attack9) > len(attack5) and len(attack9) > len(attack6) and len(attack9) > len(attack7) and
```

```
len(attack9) > len(attack8) and len(attack9) > len(attack1):
        ans = "HTTP Flood Attack"
    else:
        ans = "Unlabed Data"
    ddos_dataset.objects.create(ddos_data=txt,attack_result=ans)
    return render(request,'add_data.html')


def labeled_data(request):
    obj = ddos_dataset.objects.filter(Q(attack_result='Ip Fragment Attack')|Q ( attack_result='TCP Based
Attack') |Q(attack_result='UDP Based Attack') |Q (attack_result='NTP Amplification') |Q
(attack_result='HTTP Flood Attack')|Q (attack_result='Layer Based Attack')| Q(attack_result='Slowloris')
|Q (attack_result='Ping of Death Attack') |Q (attack_result='SYN Floods Attack'))
    return render(request,'labeled_data.html',{'object':obj})

def unlabeled_data(request):
    obj = ddos_dataset.objects.filter(attack_result='Unlabed Data')
    return render(request,'unlabeled_data.html',{'object':obj})

def ddos_analysis(request):
    chart = ddos_dataset.objects.values('attack_result').annotate(dcount=Count('attack_result'))
    return render(request,'ddos_analysis.html',{'objects':chart})

def chart_page(request,chart_type):
    chart = ddos_dataset.objects.values('attack_result').annotate(dcount=Count('attack_result'))
    return render(request,'chart_page.html',{'chart_type':chart_type,'objects':chart}
```

6.2 OUTLINE FOR VARIOUS FILES

We used Python programming to implement our project. Multiple Python files are used to implement our code. Various files used different types of modules in our project. Our project modules are – django.contrib.admin, django.apps.AppConfig, django.db.models, django.test.TestCase, django.shortcuts(render and redirect), django.conf.urls, etc. We also used various python modules like re, os, sys, decimal, MySQLdb, time, datetime, warnings, collections, etc.

6.3 CLASS WITH FUNCTIONALITY

In our project code, we have used various classes. They are:

1. DataAdminsConfig
2. ddos_dataset

Our first class configures the names of data_admins using the module AppConfig from django.apps and second class defines the dataset to be trained with ddos_data and attack-result as fields using models module from django.db.

6.4 METHODS INPUT AND OUTPUT PARAMETERS

In our project code, we implemented eight different methods. They are:

1. index(request)
2. register(request)
3. userpage(request)
4. add_data(request)
5. labeled_data(request)
6. unlabeled_data(request)
7. ddos_analysis(request)
8. chart_page(request, chart_type)

Our first method takes request as parameter and processes the request made by user and renders/loads the home page and also loads user page if the credentials given are correct. Second method takes request as parameter and processes the request and loads the registration page. Third method takes request as parameter and processes the request and returns/loads the user page. Fourth method takes request as parameter and processes the request made by the user and adds the data manually into the datasets based on the labels given to them, if no labels are given then it will be added to the unlabelled dataset. Fifth method takes the request as parameter, processes the request and filters the dataset into clusters i.e., different type of attacks and returns the results. Sixth method takes the request as parameter, processes the request and filters the unlabeled data into a dataset and returns the unlabeled dataset. Seventh method takes the request as parameter, processes the request made by the user and analyzes the objects' values, annotate them into a type of attack and returns the analysis of attacks. Eighth and the last method takes two parameters namely: request and chart_type and processes the data, analyzes and returns the data in different forms of charts like spline chart, bar chart and coloumn chart.

7. PROJECT TESTING

7.1 VARIOUS TEST CASES

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub-assemblies, assemblies and/or a finished product. It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

TYPES OF TESTS

Unit testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

Integration testing

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

Functional test

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items:

Valid Input : identified classes of valid input must be accepted.

Invalid Input : identified classes of invalid input must be rejected.

Functions : identified functions must be exercised.

Output : identified classes of application outputs must be exercised.

Systems/Procedures : interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

System Test

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

Unit Testing

Unit testing is usually conducted as part of a combined code and unit test phase of the software lifecycle, although it is not uncommon for coding and unit testing to be conducted as two distinct phases.

Test strategy and approach

Field testing will be performed manually and functional tests will be written in detail.

Test objectives

- All field entries must work properly.
- Pages must be activated from the identified link.
- The entry screen, messages and responses must not be delayed.

Features to be tested

- Verify that the entries are of the correct format
- No duplicate entries should be allowed
- All links should take the user to the correct page.

Integration Testing

Software integration testing is the incremental integration testing of two or more integrated

software components on a single platform to produce failures caused by interface defects.The task of the integration test is to check that components or software applications, e.g. components in a software system or – one step up – software applications at the company level – interact without error.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

Acceptance Testing

User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

7.2 BLACK BOX TESTING

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box .you cannot "see" into it. The test provides inputs and responds to outputs without considering how the software works. The below Black-Box can be any software system you want to test. For Example, an operating system like Windows, a website like Google, a database like Oracle or even your own custom application. Under Black Box Testing, you can test these applications by just focusing on the inputs and outputs without knowing their internal code implementation

Various approaches of Black Box testing

There are a set of approaches for black-box testing.

Manual UI Testing: In this approach, a tester checks the system as a user. Check and verify the user data, error messages.

Automated UI Testing: In this approach, user interaction with the system is recorded to find errors and glitches. Testers can set record demand as per schedule.

Documentation Testing: In this approach, a tester purely checks the input and output of the software. Testers consider what system should perform rather than how. It is a manual approach to testing.

The tester doesn't need any technical knowledge to test the system. It is essential to understand the user's perspective.

Testing is performed after development, and both the activities are independent of each other.

It works for a more extensive coverage which is usually missed out by testers as they fail to see the bigger picture of the software.

Test cases can be generated before development and right after specification.

Black box testing methodology is close to agile.

7.3WHITE BOX TESTING

The box testing approach of software testing consists of black box testing and white box testing. We are discussing here white box testing which is also known as glass box testing, structural testing, clear box testing, open box testing and transparent box testing**.**

It tests internal coding and infrastructure of a software focus on checking of predefined inputs against expected and desired outputs. It is based on inner workings of an application and revolves around internal structure testing. In this type of testing programming skills are required to design test cases. The primary goal of white box testing is to focus on the flow of inputs and outputs through the software and strengthening the security of the software.

The term 'white box' is used because of the internal perspective of the system. The clear box or white box or transparent box name denote the ability to see through the software's outer shell into its inner workings.

Developers do white box testing. In this, the developer will test every line of the code of the program. The developers perform the White-box testing and then send the application or the software to the testing

team, where they will perform the black box testing and verify the application along with the requirements and identify the bugs and sends it to the developer.

The developer fixes the bugs and does one round of white box testing and sends it to the testing team. Here, fixing the bugs implies that the bug is deleted, and the particular feature is working fine on the application.

Here, the test engineers will not include in fixing the defects for the following reasons:

- Fixing the bug might interrupt the other features. Therefore, the test engineer should always find the bugs, and developers should still be doing the bug fixes.
- If the test engineers spend most of the time fixing the defects, then they may be unable to find the other bugs in the application.

8. OUTPUT SCREENS

8.1 USER INTERFACES

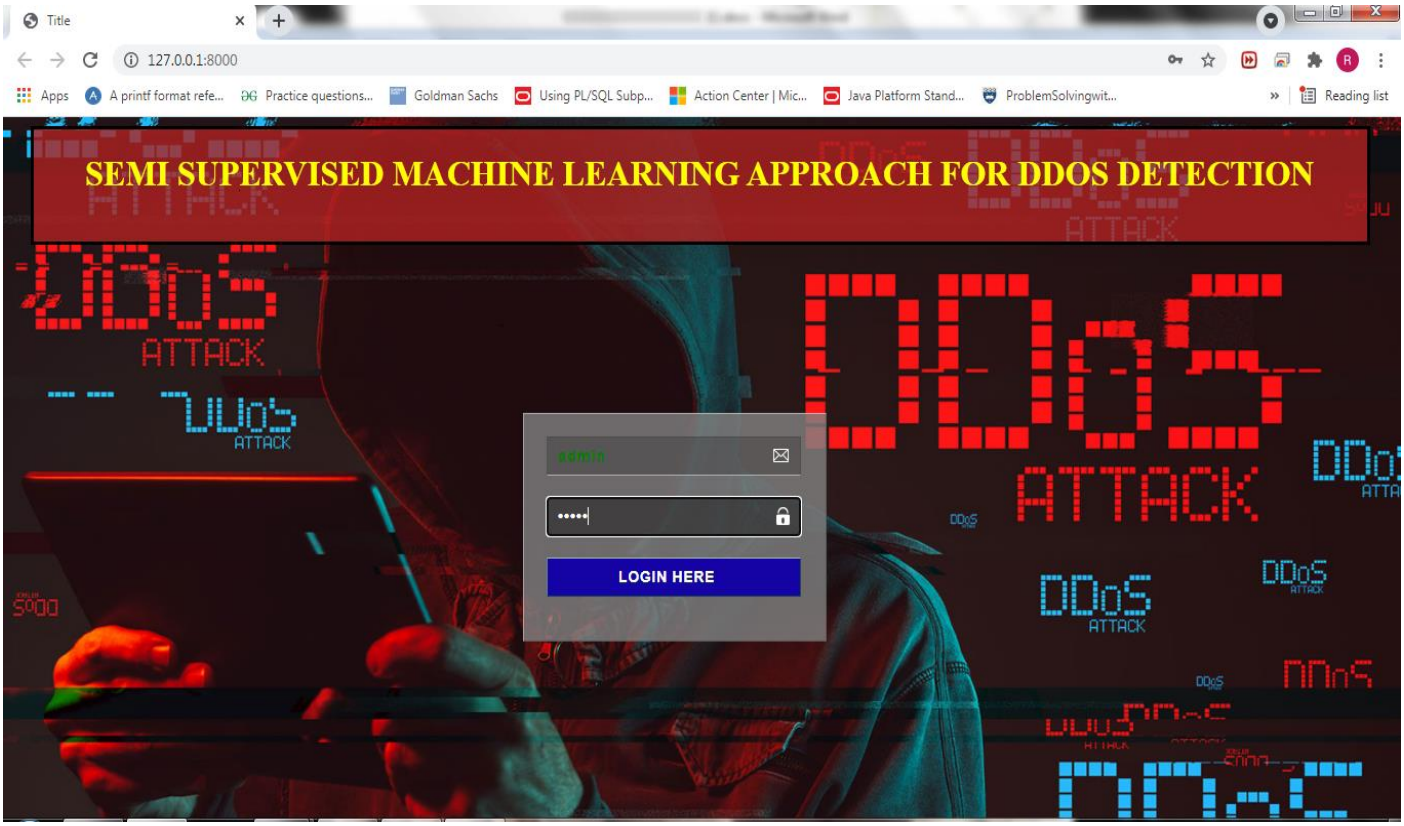


Fig.8.1.1: User Login Page

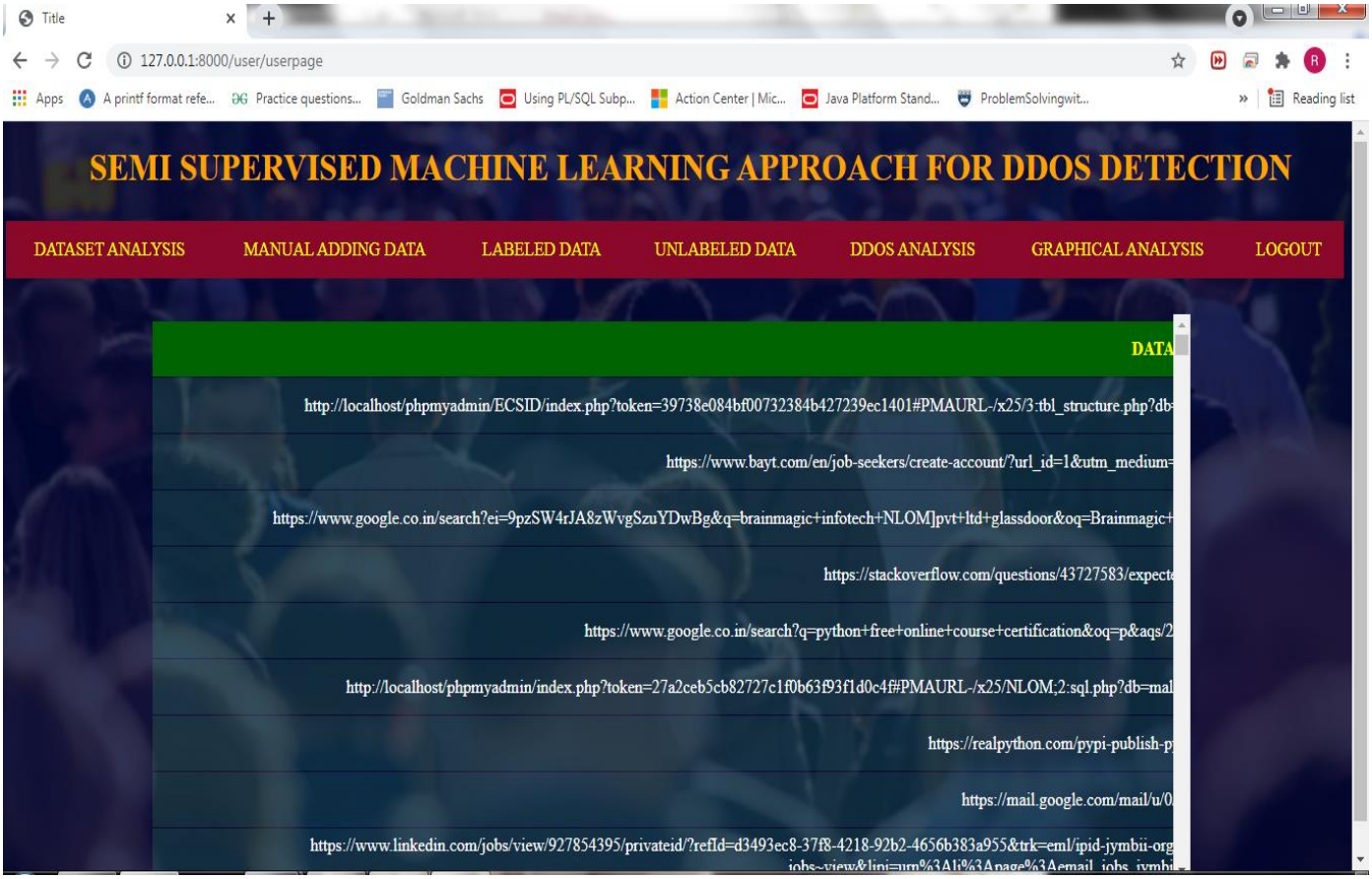


Fig.8.1.2: Home Page

8.2 OUTPUT SCREEENS

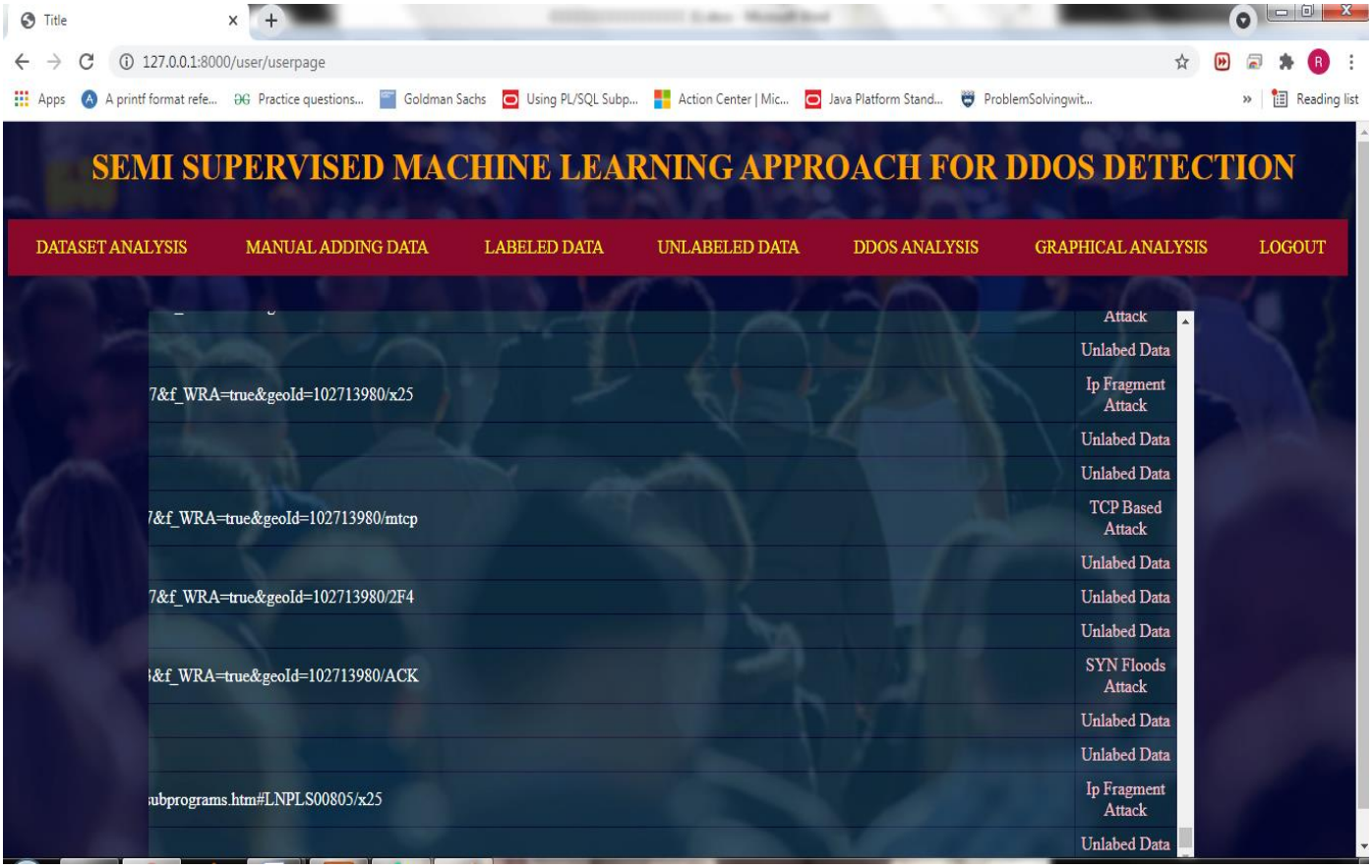


Fig.8.2.1: Home Page(DataSet Analysis)

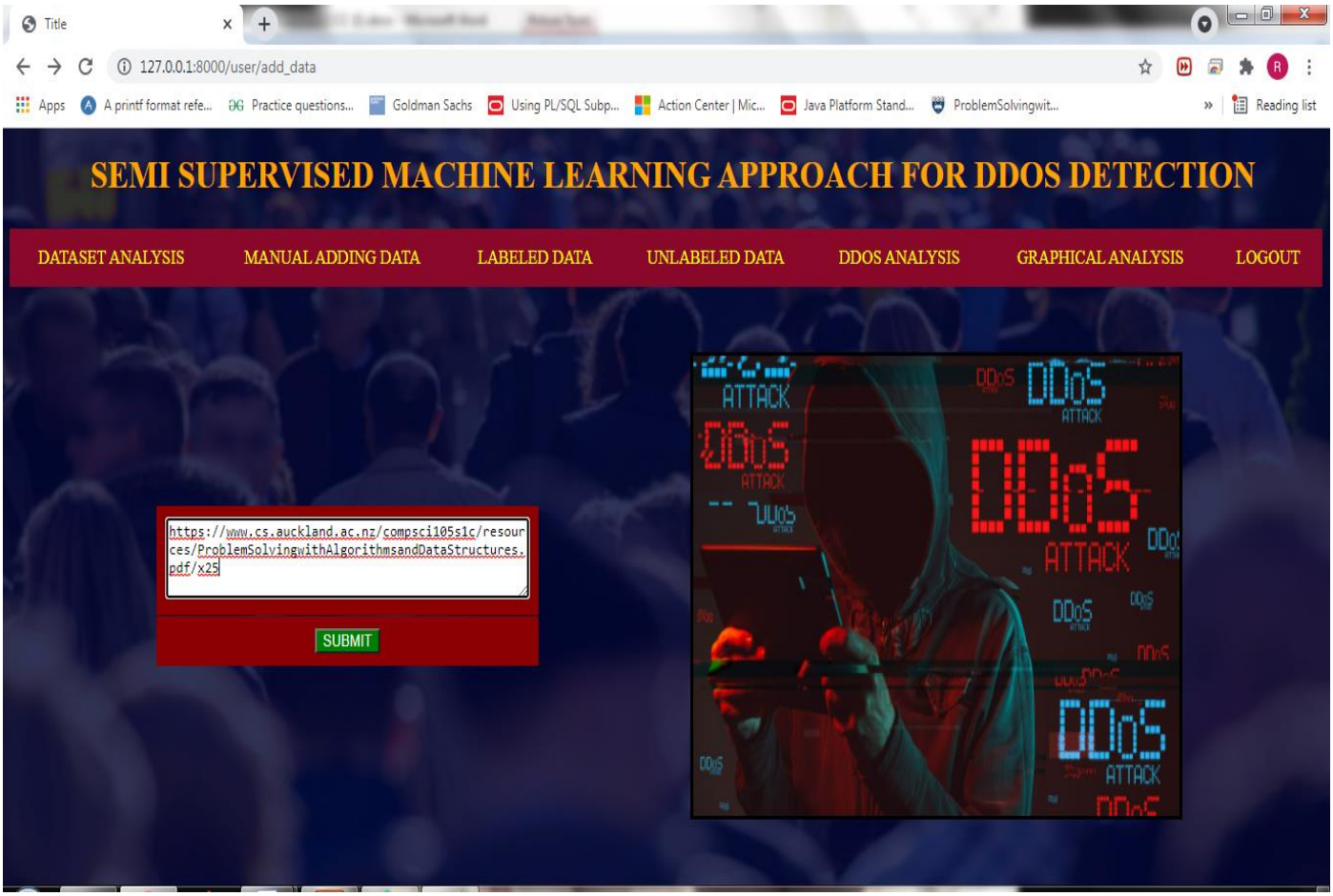


Fig.8.2.2: Manual Adding of Data

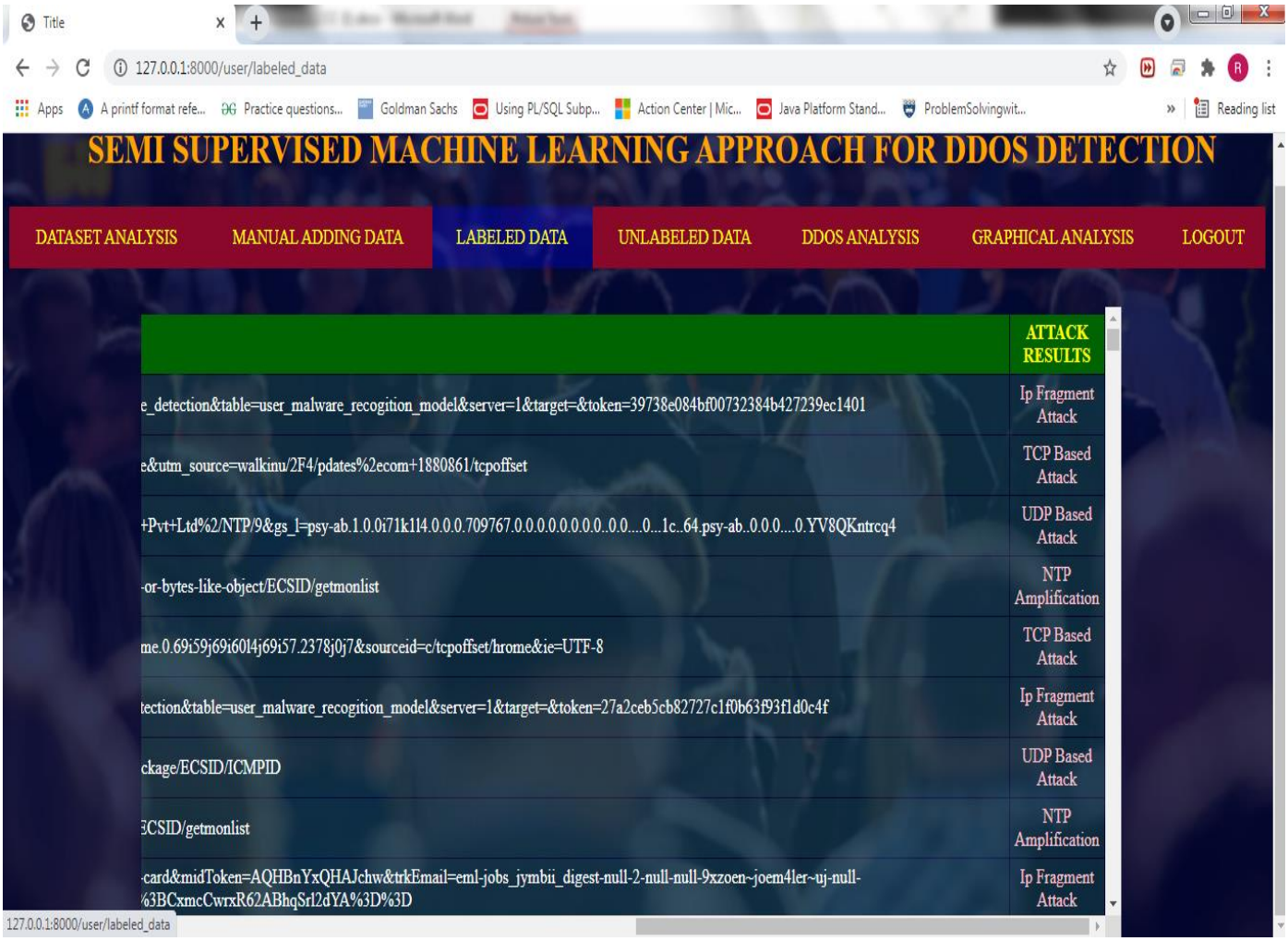


Fig.8.2.3: Labeled Data

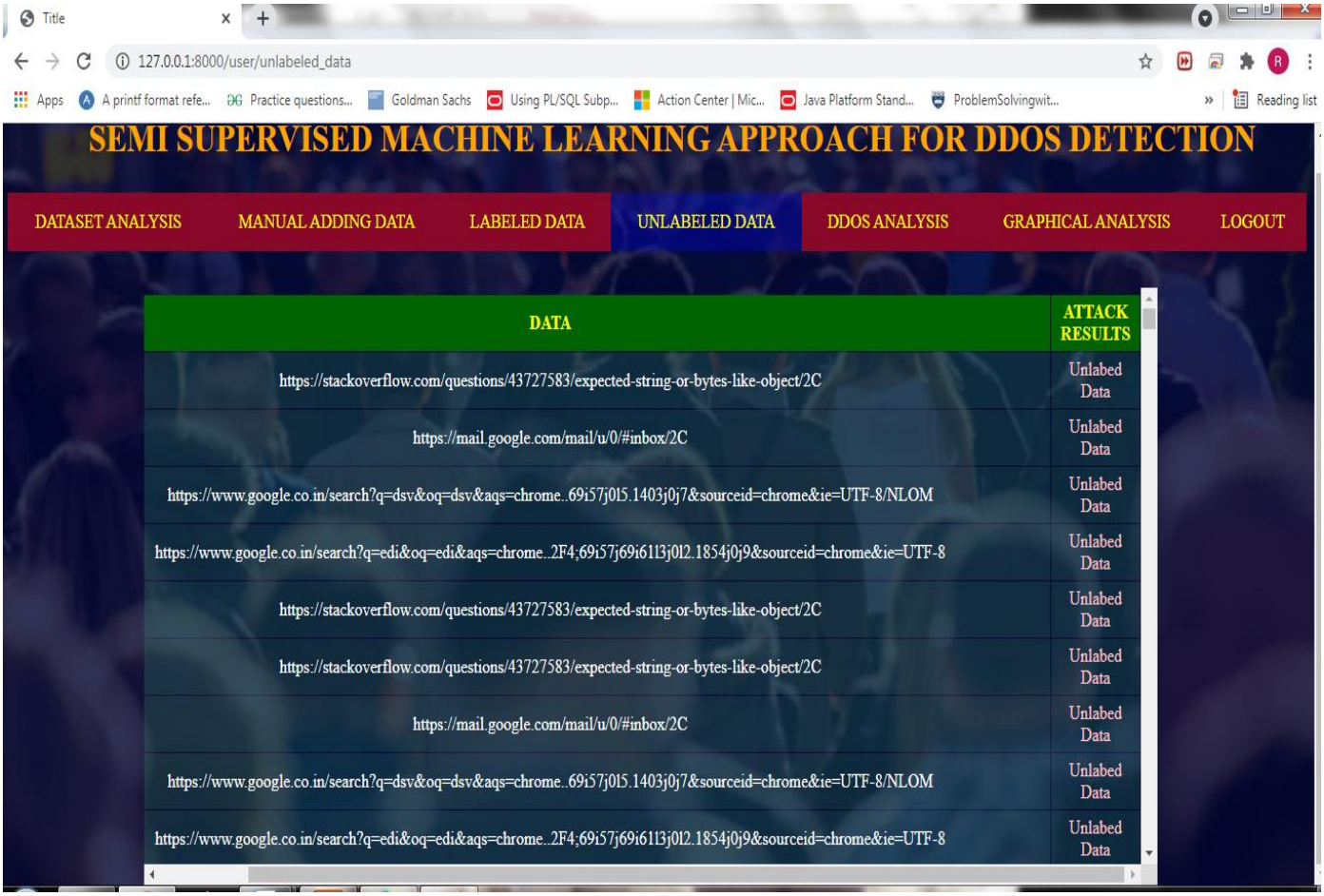


Fig.8.2.4: Unlabeled Data

9. EXPERIMENTAL RESULTS

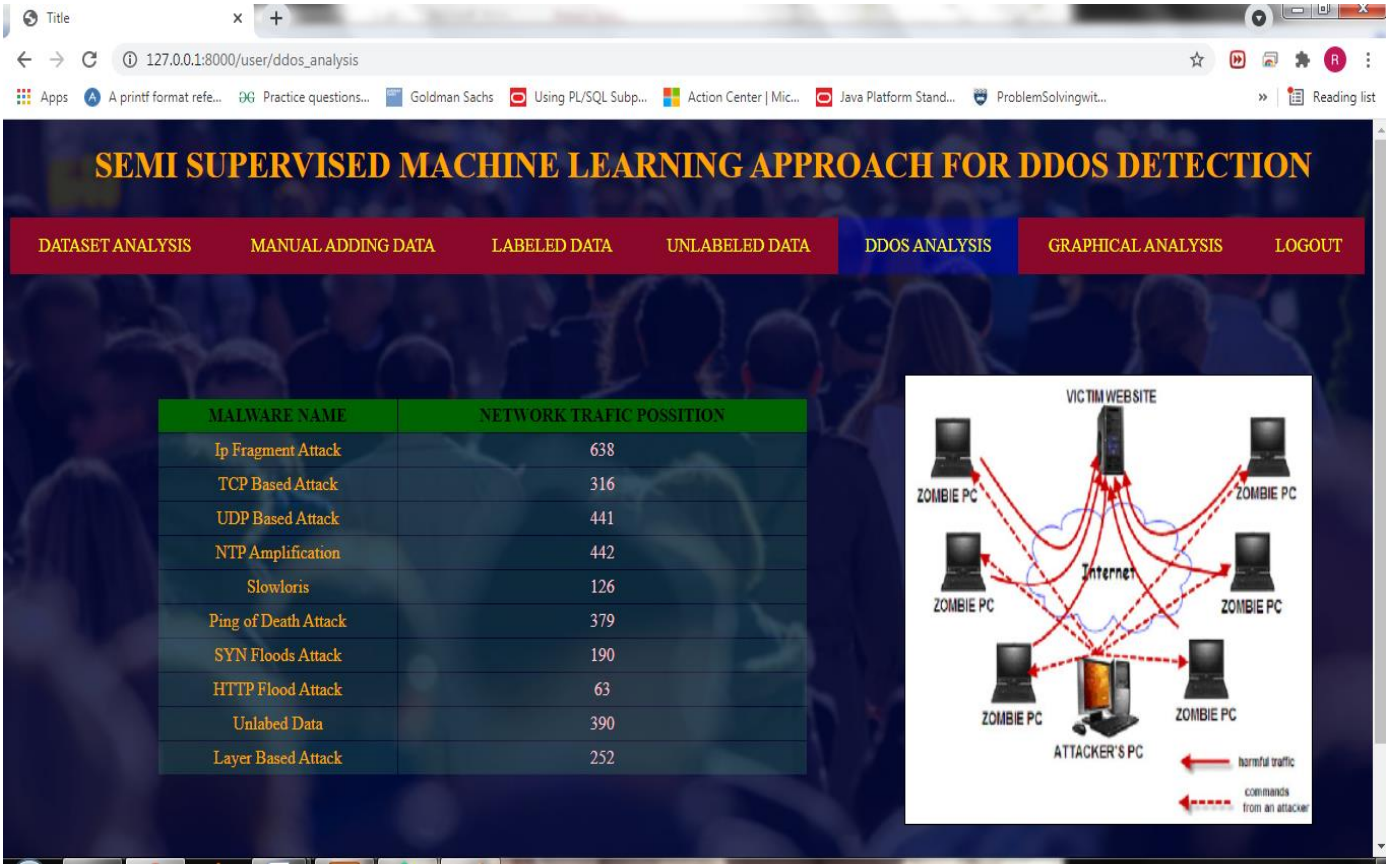


Fig.9.1: DDOS Analysis

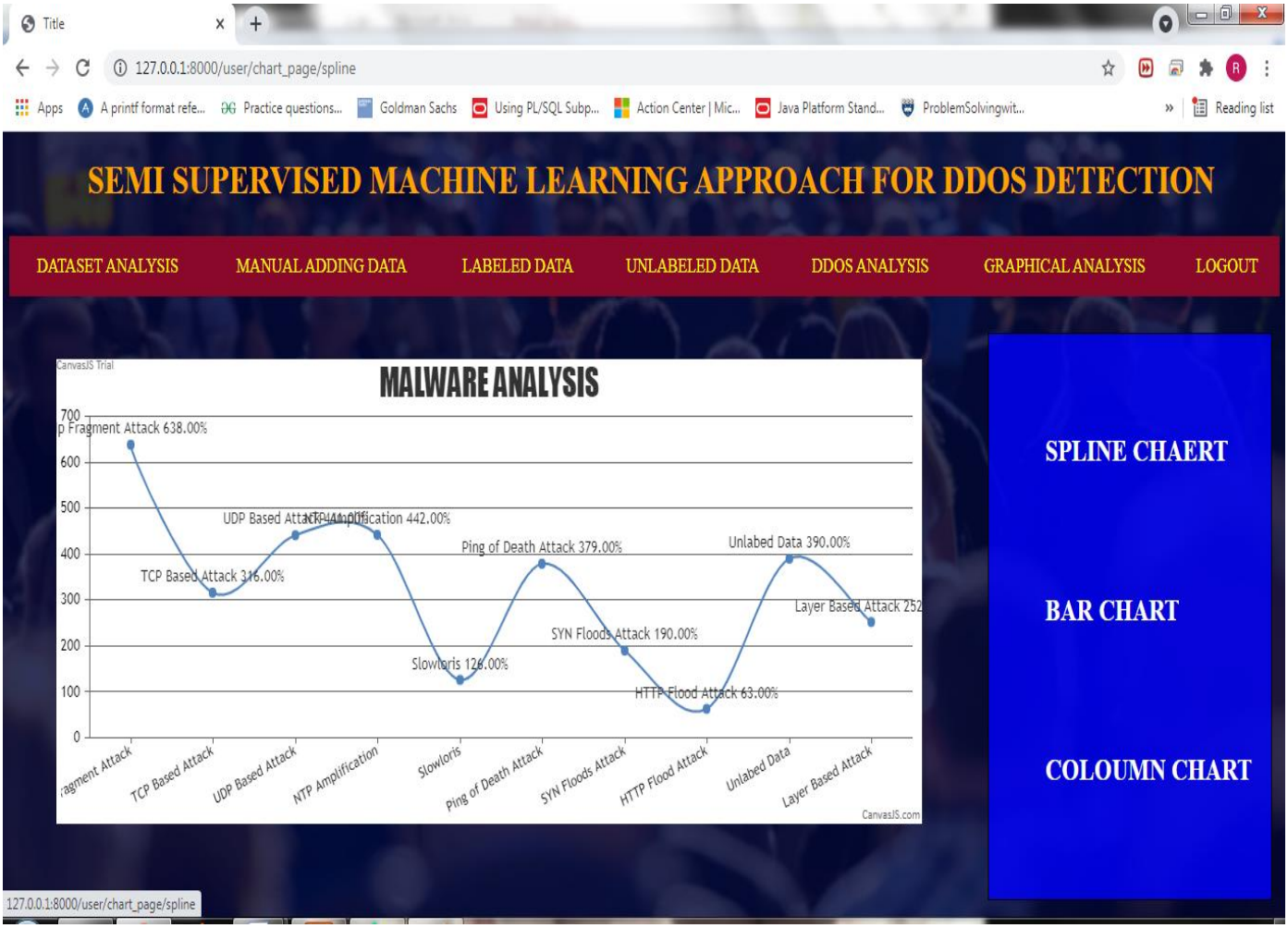


Fig.9.2: Graphical Analysis(Spline Chart)

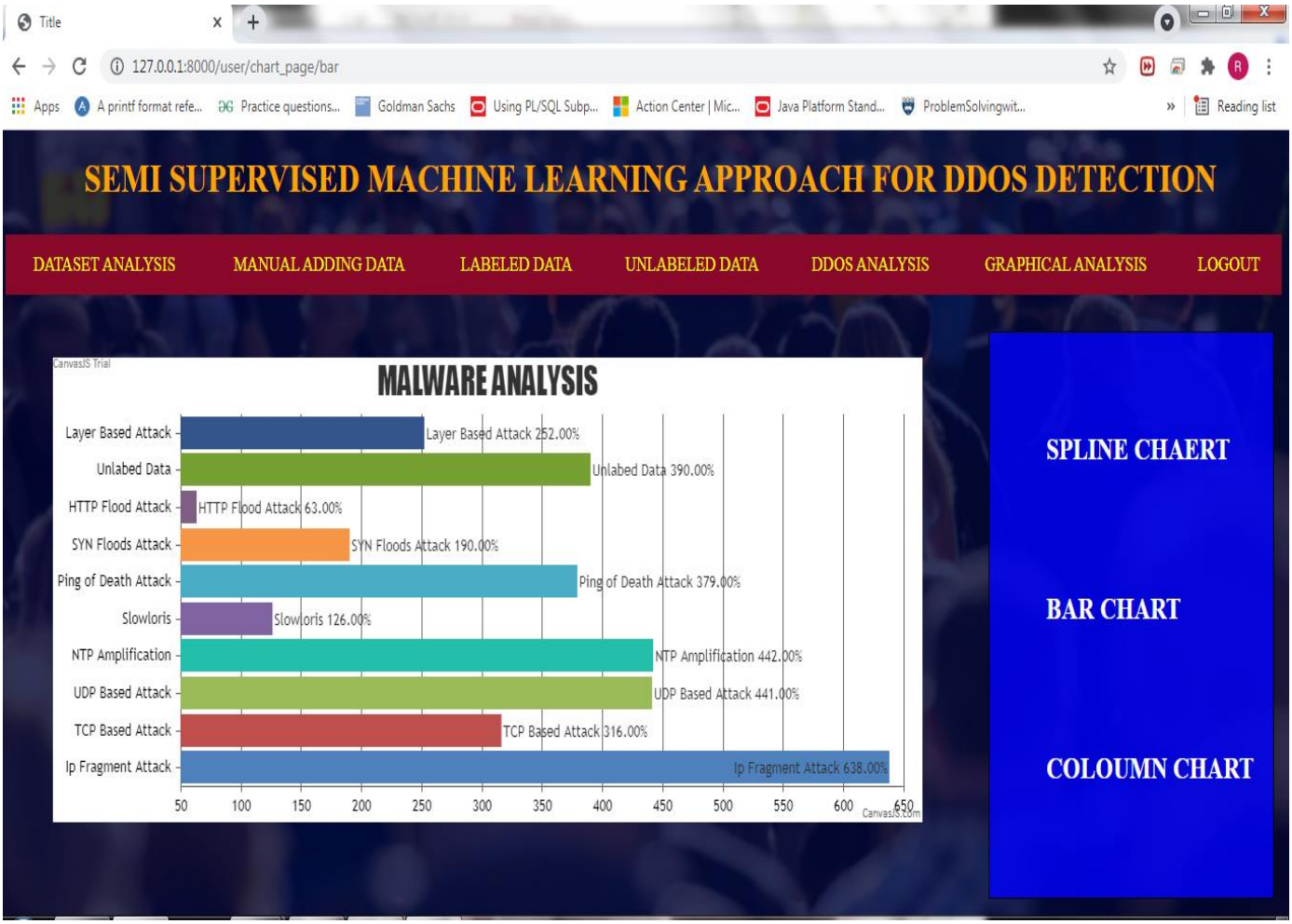


Fig.9.3: Graphical Analysis(Bar Chart)

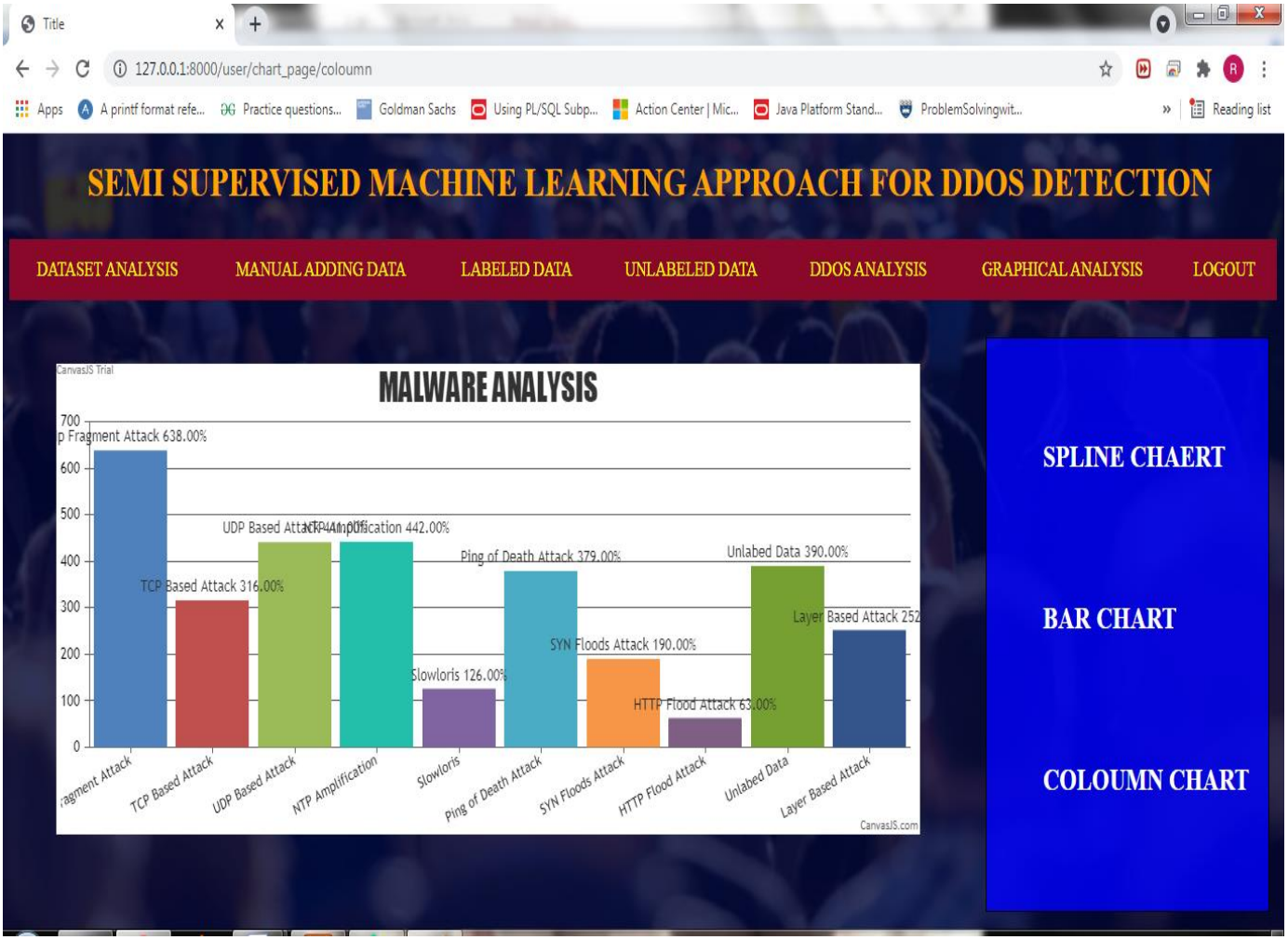


Fig.9.4: Graphical Analysis(Coloumn Chart)

10.CONCLUSION AND FUTURE ENHANCEMENT

In this project, we have proposed a semi-supervised DDoS detection approach based on entropy estimation, co-clustering, information gain ratio and the Extra-Trees ensemble classifiers. The entropy estimator estimates and analyzes the network traffic data entropy over a time-based sliding window. When the entropy exceeds its limits, the received network traffic during the current time window is split into three clusters using the co-clustering algorithm. Then, an information gain ratio is computed based on the average entropy of the network header features between the current time window subset and each one of the obtained clusters. The network traffic data clusters that produce high information gain ratio are considered as anomalous and selected for pre-processing and classification using an ensemble classifiers based on the Extra-Trees algorithm. Various experiments were conducted in order to assess the performance of the proposed method using three public benchmark datasets namely the NSL-KDD, the UNB ISCX 12 and the UNSW-NB15. The experiment results, in terms of accuracy and false positive rate, are satisfactory when compared with the state-of-the-art DDoS detection methods.

As a solution, we introduce a solution for mobile malware detection using network traffic flows, which assumes that each HTTP flow is a document and analyzes HTTP flow requests using NLP string analysis. The N-Gram line generation, feature selection algorithm, and SVM algorithm are used to create a useful malware detection model. Our evaluation demonstrates the efficiency of this solution, and our trained model greatly improves existing approaches and identifies malicious leaks with some false warnings. The harmful detection rate is 99.15%, but the wrong rate for harmful traffic is 0.45%. Using the newly discovered malware further verifies the performance of the proposed system. When used in real environments, the sample can detect 54.81% of harmful applications, which is better than other popular anti-virus scanners. As a result of the test, we show that malware models can detect our model, which does not prevent detecting other virus scanners.

Despite, that the proposed approach shows good performances with the public benchmark datasets, it is important to evaluate its performances in real world scenarios. For future work, we are planning to perform real world deployment of the proposed approach and evaluate it against several DDoS tools.

REFERENCES

1. DDOS detection using machine learning technique Sagar Pande, Aditya Khamparia, Deepak Gupta, Dang NH Thanh Recent Studies on Computational Intelligence, 59-68, 2021.

2. Semi-supervised k-means ddos detection method using hybrid feature selection algorithm Yonghao Gu, Kaiyue Li, Zhenyang Guo, Yongfei Wang IEEE Access 7, 64351-64365, 2019.

3. A novel distributed machine learning framework for semi-supervised detection of botnet attacks Gagan deep Kaur2018 Eleventh International Conference on Contemporary Computing (IC3), 1-7, 2018.

4. Leveraging machine learning approach to setup software-defined network (SDN) controller rules during DDoS attack Sajib Sen, Kishor Datta Gupta, Md Manjurul Ahsan Proceedings of International Joint Conference on Computational Intelligence, 49-60, 2020.

5. DDOS detection using machine learning technique Sagar Pande, Aditya Khamparia, Deepak Gupta, Dang NH Thanh Recent Studies on Computational Intelligence, 59-68, 2021.

6. A holistic approach for detecting ddos attacks by using ensemble unsupervised machine learning Saikat Das, Deepak Venugopal, Sajjan Shiva Future of Information and Communication Conference, 721-738, 2020.

7. Detecting mobile traffic anomalies through physical control channel fingerprinting: A deep semi-supervised approach Hoang Duy Trinh, Engin Zeydan, Lorenza Giupponi, Paolo Dini IEEE Access 7, 152187-152201, 2019.

8. Using semi-supervised learning for flow-based network intrusion detection Nandi O Leslie Cell 202, 528-0770, 2018.

9. Machine Learning in Software Defined Network Jiamei Liu, Qiaozhi Xu2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), 1114-1120, 2019.

10. IEEE Explor e.ieee.orgAn Overview of Machine Learning Based Approaches in DDoS Detectio Süreyya Atasever, İlker Özçelık, Şeref Sağiroğlu2020 28th Signal Processing and Communications Applications Conference (SIU), 1-4, 2020.

11. A reliable semi-supervised intrusion detection model: One year of network traffic anomaliesEduardo K Viegas, Altair O Santin, Vinicius V Cogo, Vilmar AbreuICC 2020-2020 IEEE International Conference on Communications (ICC), 1-6, 2020.

12. theguardian (2016) Ddos attack that disrupted internet was largest of its kind in history, experts say. https://www.theguardian.com/technology/2016/oct/26/ddos-attack-dyn-mirai-botnet.(Online; accessed 10 Apr 2017)

13. Idhammad M, Afdel K, Belouch M (2017) Dos detection method based on artificial neural networks. Int J Adv Comput Sci Appl (ijacsa) 8(4):465–471

14. Y. Gu, Y. Wang, Z. Yang, F. Xiong, and Y. Gao, ‘‘Multiple-features-based semi-supervised clustering DDoS detection method,’’ Math. Problems Eng., vol. 2017, Dec. 2017, Art. no. 5202836.

15. W. L. Al-Yaseen, Z. A. Othman, and M. Z. A. Nazri, ‘‘Multi-level hybrid support vector machine and extreme learning machine based on modified K-means for intrusion detection system,’’ Expert Syst. Appl., vol. 67, pp. 296–303, Jan. 2017.

PUBLICATIONS

JOURNAL (UGC approved Journal)

CONFERENCE (International Conference on “Innovations in Computers Networks, Computational Intelligence and IOT” [ICICCI-21]).

PAPER ID : ICICCI-21-0143

TOPIC : SEMI SUPERVISED MACHINE LEARNING APPROACH FOR DDOS DETECTION.

STUDENT'S PROFILE

Narra Rohan Raj is currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St.Martin's Engineering College. He completed his intermediate from Sri Chaitanya Junior College and 10th class from St. Marks High School. His Technical Skills include C, C++, JAVA, PYTHON. He is the student of Smart Interviews. Has certified in AI and ML course. Also participated in Leadership talk conducted by MHRD INNOVATION CELL. He has done few certification courses from online platforms such as Coursera and Udemy. He is an University level Basketball player and won fourteen college level tournaments conducted by different engineering colleges, and selected for JNTUH university basketball team. He has done an internship with a game developing company.

Rajesh Kumar Ponnala is a Software Engineer Intern at CustomFurnish.com and also pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin’s Engineering College. He completed his intermediate from Abhyaas Junior College and SSC from Vidya Niketan High School. His technical skills include C/C++, Python and MySQL. He also has a basic understanding of Core Java and frontend technologies like HTML, CSS, JavaScript. He took part in Employability Skill development Program conducted by Zensar. He is also a student of Smart Interviews. His participations include: 5-day International Hands-on Certification Training in Python Programming which was conducted by St.Martin’s Engineering College, Machine Learning with Python workshop conducted by TAM and IIC Online Sessions conducted by Institution's Innovation Council (IIC) of MHRD's Innovation Cell, New Delhi to promote Innovation, IPR, Entrepreneurship, and Start-ups among HEIs. His areas of interest are C++, Python, Artificial Intelligence, Machine Learning. Also, He has completed some courses on LinkedIn, SoloLearn and Progate.

S Ajay Kumar is currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. He completed his Intermediate from Narayana Junior college and 10th from Apex Central School. His Technical Skills include C, C++, JAVA, PYTHON. He also has basic understanding of SQL and HTML. He is a student of Smart Interviews. His participations include: National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020, the Global Webinar on Cloud & Big Data – Changing the way we work which was conducted Global Education & Careers Forum (GECF) on 12th August 2020.Also, participated in Leadership talk conducted by MHRD INNOVATION CELL and 3-days online workshop on Python Programming. He also completed few certification courses from online platforms like Coursera, CursaApp, Sololearn.

Lahari Vasabhakthula is currently pursuing her Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. She completed her Intermediate from Sri Chaitanya Junior college and 10th from Bhashyam High School. She completed her Internship from Electronics Corporation of India Limited (ECIL-ECIT), Hyderabad [June 2019 - July 2019] and developed a web Based application entitled "Cyber bulling detection". Her technical skills include C, C++, Python and SQL. She also has a basic understanding of Java (oops concepts). She is also a student of Smart Interviews. Her participations include: National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020, the Global Webinar on Cloud & Big Data – Changing the way we work which was conducted Global Education & Careers Forum (GECF) on 12th August 2020, Women online workshop on "Women in Cyber Security and Privacy in 2020" which was conducted from 6th to 10th July 2020.She completed few certification courses from online platforms like Coursera, CursaApp, Udemy, SoloLearn and PrepInsta.

APPENDICES

manage.py

```
import os
import sys

if __name__ == "__main__":
    os.environ.setdefault("DJANGO_SETTINGS_MODULE", "DDOS_ATTACK.settings")
    try:
        from django.core.management import execute_from_command_line
    except ImportError:
        # The above import may fail for some other reason. Ensure that the
        # issue is really that Django is missing to avoid masking other
        # exceptions on Python 2.
        try:
            import django
        except ImportError:
            raise ImportError(
                "Couldn't import Django. Are you sure it's installed and "
                "available on your PYTHONPATH environment variable? Did you "
                "forget to activate a virtual environment?"
            )
        raise
    execute_from_command_line(sys.argv)
```

views.py

```
import re


from django.db.models import Q, Count
from django.shortcuts import render, redirect

# Create your views here.
from data_admins.models import ddos_dataset

def index(request):
    if request.method == "POST":
        if request.method == "POST":
            usid = request.POST.get('username')
            pswd = request.POST.get('password')
            if usid == 'admin' and pswd == 'admin':
                return redirect('userpage')
    return render(request,'index.html')

def register(request):
    return render(request,'register.html')
```

```
def userpage(request):
    obj = ddos_dataset.objects.all()
    return render(request,'userpage.html',{'object':obj})

def add_data(request):
    attack1 = []
    attack2, attack3, attack4, attack5, attack6, attack7, attack8, attack9 = [], [], [], [], [], [], [], []
    ans = ''
    txt = ''
    splt = ''
    if request.method == "POST":
        txt = request.POST.get("name")

        splt = (re.findall(r"[\w']+", str(txt)))

    for f in splt:
        if f in ('IPid','FDDI','x25','rangingdistance'):
            attack1.append(f)
        elif f in ('tcpchecksum','mtcp','controlflags','tcpoffset','tcpport'):
            attack2.append(f)
        elif f in ('ICMPID','udptraffic','udpunicorn','datagramid','NTP','RIP','TFTP'):
            attack3.append(f)

        elif f in ('GETID','POSTID','openBSD','appid','sessionid','transid','physicalid'):
            attack4.append(f)
        elif f in ('SYN','ACK','synpacket','sycookies'):
            attack5.append(f)
        elif f in ('serverattack','serverid','blockbankwidth'):
            attack6.append(f)
        elif f in ('monlist','getmonlist','NTPserver'):
            attack7.append(f)
        elif f in ('portid','FTPID','tryion','fragflag'):
            attack8.append(f)
        elif f in ('malwareid','gethttpid','httpid'):
            attack9.append(f)

    if len(attack1) > len(attack2) and len(attack1) > len(attack3) and len(attack1) > len(attack4) and len(
            attack1) > len(attack5) and len(attack1) > len(attack6) and len(attack1) > len(attack7) and len(
            attack1) > len(attack8) and len(attack1) > len(attack9):
        ans = "Ip Fragment Attack"
    elif len(attack2) > len(attack1) and len(attack2) > len(attack3) and len(attack2) > len(attack4) and len(
            attack2) > len(attack5) and len(attack2) > len(attack6) and len(attack2) > len(attack7) and len(
            attack2) > len(attack8) and len(attack2) > len(attack9):
        ans = "TCP Based Attack"
    elif len(attack3) > len(attack2) and len(attack3) > len(attack1) and len(attack3) > len(attack4) and len(
            attack1) > len(attack5) and len(attack1) > len(attack6) and len(attack1) > len(attack7) and len(
            attack1) > len(attack8) and len(attack1) > len(attack9):
        ans = "UDP Based Attack"
    elif len(attack4) > len(attack2) and len(attack4) > len(attack3) and len(attack4) > len(attack1) and
```

```
len(attack4) > len(attack5) and len(attack4) > len(attack6) and len(attack4) > len(attack7) and len(
attack4) > len(attack8) and len(attack4) > len(attack9):
        ans = "Layer Based Attack"
    elif len(attack5) > len(attack2) and len(attack5) > len(attack3) and len(attack5) > len(attack4) and len(
            attack5) > len(attack1) and len(attack5) > len(attack6) and len(attack5) > len(attack7) and len(
            attack5) > len(attack8) and len(attack5) > len(attack9):
        ans = "SYN Floods Attack"
    elif len(attack6) > len(attack2) and len(attack6) > len(attack3) and len(attack6) > len(attack4) and len(
            attack6) > len(attack5) and len(attack6) > len(attack1) and len(attack6) > len(attack7) and len(
            attack6) > len(attack8) and len(attack6) > len(attack9):
        ans = "Slowloris"
    elif len(attack7) > len(attack2) and len(attack7) > len(attack3) and len(attack7) > len(attack4) and len(
            attack7) > len(attack5) and len(attack7) > len(attack6) and len(attack7) > len(attack1) and len(
            attack7) > len(attack8) and len(attack7) > len(attack9):
        ans = "NTP Amplification"
    elif len(attack8) > len(attack2) and len(attack8) > len(attack3) and len(attack8) > len(attack4) and len(
            attack8) > len(attack5) and len(attack8) > len(attack6) and len(attack8) > len(attack7) and len(
            attack8) > len(attack1) and len(attack8) > len(attack9):
        ans = "Ping of Death Attack"
    elif len(attack9) > len(attack2) and len(attack9) > len(attack3) and len(attack9) > len(attack4) and len(
            attack9) > len(attack5) and len(attack9) > len(attack6) and len(attack9) > len(attack7) and len(
            attack9) > len(attack8) and len(attack9) > len(attack1):
        ans = "HTTP Flood Attack"
    else:
        ans = "Unlabed Data"
    ddos_dataset.objects.create(ddos_data=txt,attack_result=ans)
    return render(request,'add_data.html')
def labeled_data(request):
    obj = ddos_dataset.objects.filter(Q(attack_result='Ip Fragment Attack')|Q ( attack_result='TCP Based
Attack') |Q(attack_result='UDP Based Attack') |Q (attack_result='NTP Amplification') |Q
(attack_result='HTTP Flood Attack')|Q (attack_result='Layer Based Attack')| Q(attack_result='Slowloris')
|Q (attack_result='Ping of Death Attack') |Q (attack_result='SYN Floods Attack'))
    return render(request,'labeled_data.html',{'object':obj})

def unlabeled_data(request):
    obj = ddos_dataset.objects.filter(attack_result='Unlabed Data')
    return render(request,'unlabeled_data.html',{'object':obj})

def ddos_analysis(request):
    chart = ddos_dataset.objects.values('attack_result').annotate(dcount=Count('attack_result'))
    return render(request,'ddos_analysis.html',{'objects':chart})

def chart_page(request,chart_type):
    chart = ddos_dataset.objects.values('attack_result').annotate(dcount=Count('attack_result'))
    return render(request,'chart_page.html',{'chart_type':chart_type,'objects':chart})
```

urls.py

```
from django.conf.urls import url
from django.contrib import admin
```

```
from data_admins import views as admins
urlpatterns = [
    url(r'^admin/', admin.site.urls),

    url('^$',admins.index,name="index"),
    url('user/register', admins.register, name="register"),
    url('user/add_data',admins.add_data,name="add_data"),
    url('user/userpage',admins.userpage,name="userpage"),
    url('user/labeled_data',admins.labeled_data,name="labeled_data"),
    url('user/unlabeled_data',admins.unlabeled_data,name="unlabeled_data"),
    url('user/ddos_analysis',admins.ddos_analysis,name="ddos_analysis"),
    url('user/chart_page/(?P<chart_type>\w+)',admins.chart_page,name="chart_page"),


]
```

A

PROJECT REPORT

On

STOCK MARKET TREND PREDICTION USING K-NEAREST NEIGHBORS(KNN) ALGORITHM

Submitted by

1) B. Sripriya(17K81A05D2) **2) S. Maneesha(17K81A05H7)**

3) G. Premika(17K81A05E3) **4) D. Sripriya(17K81A05D4)**

in partial fulfillment for the award of the

degree of

BACHELOR OF TECHNOLOGY

IN

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Under the Guidance of

Mr. D. Krishna, B.Tech, M.Tech,

Assistant Professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ST.MARTIN'S ENGINEERING COLLEGE
An Autonomous Institute

Dhulapally, Secunderabad – 500 100

JUNE 2021

BONAFIDE CERTIFICATE

This is to certify that the project entitled **STOCK MARKET TREND PREDICTION USING KNN ALGORITHM**, is being submitted by 1.**Ms. BYRAGONI SRIPRIYA** (**17K81A05D2), 2.Ms. SINGAM MANEESHA (17K81A05H7), 3. Ms. GOVINDARAJU PREMIKA (17K81A05E3), 4. Ms. DANDAMRAJU SRIPRIYA (17K81A05D4**) in partial fulfillment of the requirement for the award of the degree of **BACHELOR OF TECHNOLOGY IN COMPUTER SCIENCE ENGINEERING** is recorded of bonafied work carried out by them. The result embodied in this report have been verified and found satisfactory.

Assistant Professor

Mr. D. Krishna

Department of CSE

Head of the Department

Dr.M.NARAYANAN

Department of CSE

Internal Examiner

External Examine

Place:

Date

DECLARATION

We, the student of **Bachelor of Technology** in Department of Computer Science and Engineering', session: 2017 – 2021, St. Martin's Engineering College, Dhulapally, Kompally, Secunderabad, hereby declare that work presented in this Project Work entitled **STOCK MARKET TREND PREDICTION USING** K**NN ALGORITHM** is the outcome of our own bonafied work and is correct to the best of our knowledge and this work has been undertaken taking care of Engineering Ethics. This result embodied in this project report has not been submitted in any university for award of any degree.

Ms. B. Sripriya(17K81A05D2)

Ms. S. Maneesha(17K81A05H7)

Ms. G. Premika)17K81A05E3)

Ms. D. Sripriya(17K81A05D4)

ABSTRACT

This paper examines a hybrid model which combines a K-Nearest Neighbour (KNN) approach with a probabilistic method for the prediction of stock price trends. One of the main problems of KNN classification is the assumptions implied by distance functions. The assumptions focus on the nearest neighbour which are at the centroid of data points for test instances. This approach excludes the non-centric data points which can be statistically significant in the problem of predicting the stock price trends. For this it is necessary to construct an enhanced model that integrates KNN with a probabilistic method which utilizes both centric and non-centric data points in the computations of probabilities for the target instances. The embedded probabilistic method is derived from Bayes' theorem. The prediction outcome is based on a joint probability where the likelihood of the event of the nearest neighbour and the event of prior probability occurring together and at the same point in time where they are calculated. The proposed hybrid KNN Probabilistic model was compared with the standard classifiers that include KNN, Naive Bayes, One Rule (One R) and Zero Rule (Zero R). The test results showed that the proposed model outperformed the standard classifiers which were used for the comparisons. Keywords: Stock Price Prediction, K-Nearest Neighbour, Bayes' Theorem, Naive Bayes, Probabilistic Method.

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible and whose encouragements and guidance have crowded effects with success.

We extended our deep sense of gratitude to Principal**,Dr. P. SANTOSH KUMAR PATRA**, St. Martin's Engineering College, Dhulapally, for permitting us to undertake this project.

We are also thankful to **Dr. M.NARAYANAN**, Head of the Department, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

We would like to thank **Dr. T. POONGOTHAI,** Department of Computer Science and Engineering (AI & ML) for her encouragement and insightful comments and as well as our project coordinators **Dr. B.RAJALINGAM**, Assistant Professor and **Dr. N.SATHEESH**, Professor, in Department of Computer Science and Engineering for their valuable support.

We would like to express our sincere gratitude and indebtedness to our project supervisor Mr. D.Krishna, Assistant Professor,Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

Finally, we express thanks to all those who have helped us successfully to completing this project. Furthermore, we would like to thank our family and friends for their moral support and encouragement.

We express thanks to all those who have helped us in successfully completing the project.

Ms. B. SRIPRIYA(17K81A05D2)

Ms. S. MANEESHA(17K81A0H7)

Ms. G. PREMIKA(17K81A05E3)

Ms. D. SRIPRIYA(17K81A05D4)

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF OUTPUT SCREENS

LIST OF ABBREVATIONS

KNN	K-Nearest Neighbors
ANN	Artificial neural networks
GARCH	Generalized autoregressive conditional heteroscedasticity
SVR	Support vector regression
FLANN	Fast library for appropriate nearest neighbors

CHAPTER 1

INTRODUCTION

The stock market is an evolutionary, complex and a dynamic system. Market prediction is characterized by noise, data intensity, non-stationary, uncertainty and hidden relationships. The prediction of trend in stock market exchange has been a challenging and important research topic.

It is challenging because the data is noisy and not stationary. It is important because it can yield important results for decision makers.

Stock market is such a location where companies invest high capital and do their shares trading. Stock market prediction has disproved the Efficient Market Hypothesis which states that it is impossible to predict the market because it is efficient.

Researchers have proved that it is possible to predict the stock market. The ability of making future stock market prediction is an important factor for investors for making money.

It also helps investors to make selling or buying decisions to generate higher profits. The chief goal of this project is to add to the academic understanding of stock market prediction. The hope is that with a greater understanding of how the market moves, investors will be better equipped to prevent another financial crisis. The project will evaluate some existing strategies from a rigorous scientific perspective and provide a quantitative evaluation of new strategies.

There are several data mining algorithms that can be used for prediction purposes in the field of finance. Some examples would be the naive Bayes classifier, the k nearest neighbour (KNN) algorithm and the classification and the regression tree algorithm (Wu et al. 2007). All the mentioned algorithms could fill the purpose of the paper but it will center around the kNN algorithm as a method of predicting stock market movements as well as the MA formula. The movements will be detected by looking at a large amount of historical data and finding patterns to establish a well estimated forecast. This specific algorithm was chosen as it is a simple but a very effective algorithm to implement when looking at large amounts of data (Berson et al. 1999).The KNN algorithm simply states: "Objects that are ’near’ to each other will have similar prediction values as well. Thus if you know the prediction value of one of the objects you can predict it for its nearest neighbours" (Berson et al. 1999). As a comparison with the KNN algorithm, the MA formula was chosen. The MA formula has its simplicity as a common factor

with the KNN algorithm, but it is a statistical method used frequently by traders (Interactive Data Corp, 2014). In the finance world, stock trading is one of the most important activities. Stock market prediction is an act of trying to determine the future value of a stock. Everyday billions of dollars are traded on the exchange, and behind each dollar there is an investor hoping to profit in one way or another. Entire companies rise and fall daily based on the behaviour of the market. Investing in the stock market is risky, only if the investor is not aware of how market actually works. But if the investor is able to accurately predict market movements, it offers a tantalizing promises of wealth and influence. The chief goal of this project is to make use of machine learning approach for stock market prediction. The hope is that with a greater understanding of how the market moves, investors will be better equipped to prevent another financial crisis.

1.1 PROJECT OVERVIEW

Stock market prediction means determining the future value of a stock. The prediction of market value is of great importance to help in maximizing profits. If investor is able to accurately predict market movements, It may result in huge profits. So our project mainly deals in predicting stock values using KNN algorithm which is a machine learning approach that provides accurate prediction by looking at large amounts of historical data and patterns.

The project is used to evaluate stock market trend. Loss and gain plays vital role in competition market. So by predicting those loss and gain with the market may give profits gained by the company. Accurately predicting the market value is important. Our project will evaluate some existing strategies and also provide evaluation of new strategies.

1.2 PROJECT OBJECTIVES

- The objective of project is to get accurate values using KNN algorithm. This Machine learning algorithm assumes similar things that are near to each other.
- Similarity is known by calculating distance between points using several approaches like Euclid distance
- This algorithm is versatile, it can be used for classification, regression and search.

1.3 ORGANIZATION OF CHAPTERS

Besides the introduction, the thesis is organized in other six chapters as follows:

Chapter 2, LITERATURE SURVEY: the review is made in the context of EHR systems with a particular attention on those implementations that assess the scalability and performances or their implementations. Most of the related work is on blockchain solutions, whereas a small part is on cloud solutions.3 It will be possible to notice that only a small subset of the literature actually focuses on the analysis of the systems in mass crises scenarios.

Chapter 3, SOFTWARE AND HARDWARE REQUIREMENTS: this chapter discuss about the software and hardware required for the execution of the project

Chapter 4, SOFTWARE DEVELOPMENT ANALYASIS: this chapter explains the assumptions and technical specifications of the project.

Chapter 5, PROJECT SYSTEM DESIGN: this chapter explains all the software development process with dfd, E-R diagrams, and UML diagrams clearly.

Chapter 6, PROJECT CODING: this chapter explains the design of the system, roles and responsibilities, as well as the requirements of a EHRs management solution based on block chain.

Chapter 7, PROJECT TESTING: this chapter explains various test cases to test the project working.

Chapter 8, OUTPUT SCREENS: explains a step by step process of the project execution.

Chapter 9, EXPERIMENTAL RESULTS: tests and results are shown and explained in this chapter. The results are analyzed in the context of the thesis project and followed by discussion on systems throughput and resiliency, as well as the approaches to testing and analysis.

Chapter 10, CONCLUSION AND FUTURE ENHANCEMENT: the chapter ends the project with a short summary of the main concepts mentioned in the thesis as well as the relevant result

CHAPTER 2

LITERATURE SURVEY

A literature survey or a literature review in a project report is that section which shows the various analysis and research made in the field of your interest and the results already published, considering the various parameters of the project and the extent of the project. It is the most important part of our report as it gave us a direction in our research. It helped us set a goal for our analysis - thus giving us our problem statement.

2.1 SURVEY ON BACKROUND

Researchers have employed different machine learning classifiers for the prediction of stock market. This prediction is based on previous records called training data set. Normally 80% of the data set comprises the training data set. The data set that is tested on the trained classifier is called the testing dataset that comprises 20% of the data.

The proposed method begins with processing the data using data set 2, with each record contains a stock's financial features and the predicted outcomes in a structured categorical format. Using these records as inputs, stock price trends were predicted using the proposed hybrid KNN-Probabilistic model. Recently, prediction has been recognized as an important topic in Machine Learning. Python 3.7 was used as the tool to realize our research results in python programming language. A KNN machine learning algorithm and probabilistic method derived from Baye's theorem have been used for the prediction of stock market.

A nearest neighbour search (NNS) method produced an intended result by the use of KNN technique with technical analysis. This model applied technical analysis on stock market data which include historical price and trading volume. It applied technical indicators made up of stop loss, stop gain and RSI filters. The KNN algorithm part applied the distance function on the collected data. This model was compared with the buy-and-hold strategy by using the fundamental analysis approach.

Fast Library for Approximate Nearest Neighbours (FLANN) is used to perform the searches for choosing the best algorithm found to work best among a collection of algorithms in its library. Majhi etal. examined the FLANN model to predict the S&P 500 indices.

Artificial neural networks (ANN) exhibit high generalization power as compared to conventional statistical tools. ANN is able to infer from historical data to identify the characteristics of performing stocks. The information is reflected in technical and financial variables.

Neural network modelling can decode nonlinear regularities in asset price movements. Statistical inference and modifications to standard learning techniques prove useful in dealing with the salient features of economic data.

Shynkevich et al. studied how the performance of a financial forecasting model was improved by the use of a concurrent, and appropriately weighted news articles, having different degrees of relevance to the target stock.

In high dimensional data, not all features are relevant and have an influence on the outputs. An Enhanced Feature Representation Based on Linear Regression Model for Stock Market Prediction was evaluated to investigate the statistical metrics used in feature selection that extracts the most relevant features to reduce the high dimensionality of the data. The statistical metrics include Information Gain, Term Frequency-Invert Document Frequency and the Document Frequency.

Volatility indicates the risk of a security. The Generalized Autoregressive Conditional Heteroscedasticity (GARCH) process is an approach used to estimate volatility in financial markets. The Seemingly Unrelated Regressions (SUR) is a generalization of a linear regression model that comprises several indicator relationships that are linked by the fact that their volatilities are correlated. A GARCH-SUR model was evaluated and demonstrated that the existence of a significant relationship between the volatility of macroeconomic variables and the stock market volatility in the financial markets.

Maetal. proposed a hybrid financial time series model by combining Support Vector Regression (SVR), Trend model and Maximum Entropy (ME) based on ANN for forecasting trends in fund index. The study showed that the hybrid model extracted the financial features characteristics to formulate an improved predictive model.

A hybrid intelligent data mining methodology based on Genetic Algorithm - Support Vector Machine Model was reviewed to explore stock market tendency. This approach makes use of the genetic algorithm for variable selection in order to improve the speed of support vector machine by reducing the model complexity, and then the historical data is used to identify stock market trends. Hybrid techniques can be used to improve the existing forecasting models due to the limitation of ANN like black box technique. A combination of methods

such as fuzzy rule-based system, fuzzy neural network and Kalman filter with hybrid neuro-fuzzy architecture have been developed to predict financial time series data.

This research studies a hybrid approach through the use of KNN algorithm and a probabilistic method for predicting the stock price trends.

2.2 CONCLUSIONS ON SURVEY

.

These works provide basic background information that the existing system the in the finance world, stock trading is one of the most important activities. Stock market prediction is an act of trying to determine the future value of a stock. Everyday billions of dollars are traded on the exchange, and behind each dollar there is an investor hoping to profit in one way or another. Entire companies rise and fall daily based on the behaviour of the market. Investing in the stock market is risky, only if the investor is not aware of how market actually works. But if the investor is able to accurately predict market movements, it offers a tantalizing promises of wealth and influence. The chief goal of this project is to make use of machine learning approach for stock market prediction. The hope is that with a greater understanding of how the market moves, investors will be better equipped to prevent another financial crisis.

CHAPTER 3

SOFTWARE AND HARDWARE REQUIREMENTS

All computer software needs certain hardware components or other software resources to be present on a computer. These prerequisites are known as (computer) system requirements and are often used as a guideline as opposed to an absolute rule. Most software defines two sets of system requirements: minimum and recommended. With increasing demand for higher processing power and resources in newer versions of software, system requirements tend to increase over time. Industry analysts suggest that this trend plays a bigger part in driving upgrades to existing computer systems than technological advancements. Hardware requirements: The most common set of requirements defined by any operating system or software application is the physical computer resources, also known as hardware, a hardware requirements list is often accompanied by a hardware compatibility list (HCL), especially in case of operating systems. An HCL lists tested, compatible, and sometimes incompatible hardware devices for a particular operating system or application.

3.1 HARDWARE REQUIREMENTS :

- Processor - i3 processor or Above
- RAM - 4GB
- Hard Disk - 930GB
- Keyboard - 110 keys enhaced
- Mouse - Logitech

3.2 SOFTWARE REQUIREMENTS :

- Operating System - Windows10
- Programing Language - Python3.7

CHAPTER 4

SOFTWARE DEVELOPMENT ANALYSIS

Software development is a process of writing and maintaining the source code, but in a broader sense, it includes all that is involved between the conception of the desired software through to the final manifestation of the software, sometimes in a planned and structured process. Therefore, software development may include research, new development, prototyping, modification, reuse, reengineering, maintenance, or any other activities that result in software products

4.1 OVERVIEW OF PROBLEM

In the finance world, stock trading is one of the most important activities. Stock market prediction is an act of trying to determine the future value of a stock. Everyday billions of dollars are traded on the exchange, and behind each dollar there is an investor hoping to profit in one way or another. Entire companies rise and fall daily based on the behaviour of the market. Investing in the stock market is risky, only if the investor is not aware of how market actually works. But if the investor is able to accurately predict market movements, it offers a tantalizing promises of wealth and influence. The chief goal of this project is to make use of machine learning approach for stock market prediction. The hope is that with a greater understanding of how the market moves, investors will be better equipped to prevent another financial crisis.

4.2 DEFINE THE PROBLEM

In the existing system the entire companies rise and fail based on the behaviour of the market as the investor is not aware oh how it actually works so to overcome this we propose a Machine Learning (ML) approach that will be trained from the available stocks data and gain intelligence and then uses the acquired knowledge for an accurate prediction. In this context this study uses a machine learning technique called K-Nearest Neighbour

It is a simple and effective algorithm to implement when looking at large amounts of data. KNN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most

similar to the available categories. Similarity is known by calculating the distance between the data points on a graph.

4.3 MODULES OVERVIEW

- Download Data Set
- Corelation Data
- Data Preprocessing
- Run Knn with Uniform weights
- Run Knn With Distance Weights
- Predict Data
- Knn Accuracy

4.4 DEFINE THE MODULES

Download dataset :

In this module User/investor downloads the required stock dataset and their competitors stock dataset.

Correlation Data :

Correlation Data shows the strength of a relationship between two stocks and is expressed numerically by the correlation coefficient. The correlation coefficient's values range between -1.0 and 1.0.

Data preprocessing :

This module is responsible for encoding of data, dropping missing values, split labels and split dataset.

Run KNN with uniform weights :

This generate a KNN model with uniform weights and calculate accuracy of KNN model with uniform weights.

Run KNN with distance weights :

This generate a KNN model with distance weights and calculate accuracy of KNN model with distance weights.

Predict Data :

This module test how accurately the model work for future datasets.

KNN accuracy :

Accuracy comparision for different models using bar graph, in this x-axis represent algorithm names and y-axis represent accuracy rate.

4.5 MODULE FUNCTIONALITY

Download dataset :

In this module User/investor downloads the required stock dataset and their competitors stock dataset.

Correlation Data :

Correlation Data shows the strength of a relationship between two stocks and is expressed numerically by the correlation coefficient. The correlation coefficient's values range between -1.0 and 1.0.

Data preprocessing :

This module is responsible for encoding of data, dropping missing values, split labels and split dataset.

Run KNN with uniform weights :

This generate a KNN model with uniform weights and calculate accuracy of KNN model with uniform weights.

Run KNN with distance weights :

This generate a KNN model with distance weights and calculate accuracy of KNN model with distance weights.

Predict Data :

This module test how accurately the model work for future datasets.

KNN accuracy :

Accuracy comparision for different models using bar graph, in this x-axis represent algorithm names and y-axis represent accuracy rate.

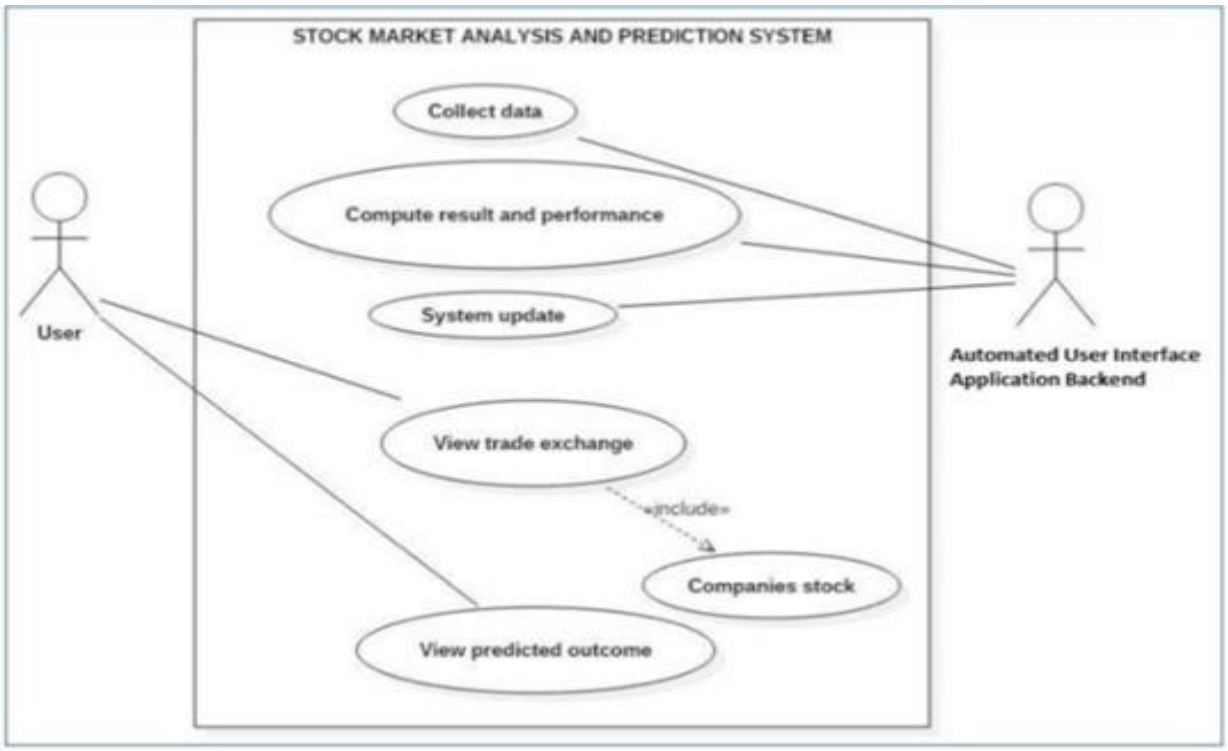
CHAPTER 5

PROJECT SYSTEM DESIGN

Systems design is the process of defining elements of a system like modules, architecture, components and their interfaces and data for a system based on the specified requirements. It is the process of defining, developing and designing systems which satisfies the specific needs and requirements of a business or organization.

A systemic approach is required for a coherent and well-running system. Bottom-Up or Top-Down approach is required to take into account all related variables of the system. A designer uses the modelling languages to express the information and knowledge in a structure of system that is defined by a consistent set of rules and definitions. The designs can be defined in graphical or textual modelling languages.

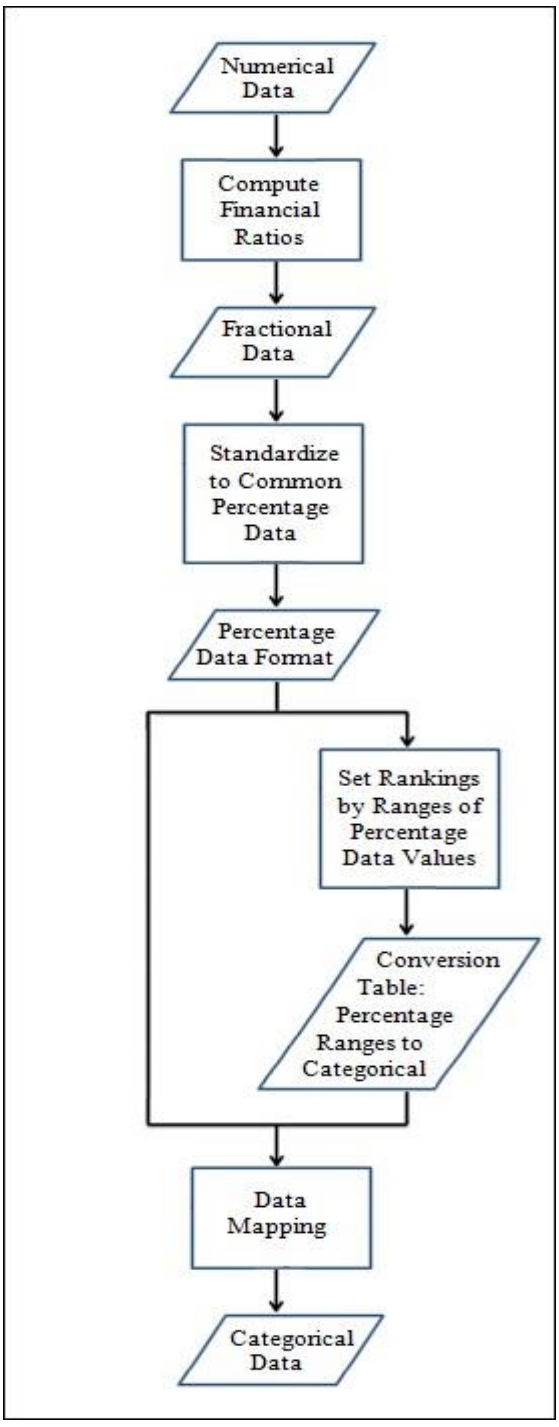
5.1 SYSTEM ARCHITECTURE



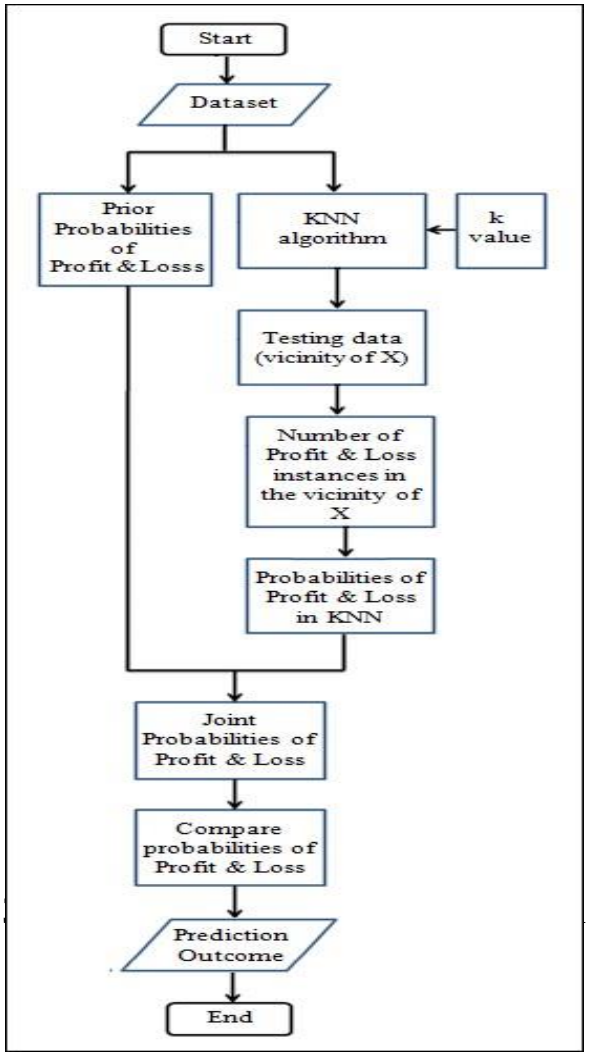
5.1 SYSTEM ARCHITECTURE

5.1 DFDS IN CASE OF DATABASE PROJECTS

A data flow diagram shows the way information flows through a process or system. It includes data inputs and outputs, data stores, and the various sub processes the data moves through. DFDs are built using standardized symbols and notation to describe various entities and their relationships



5.2 Flow chart of data transformation



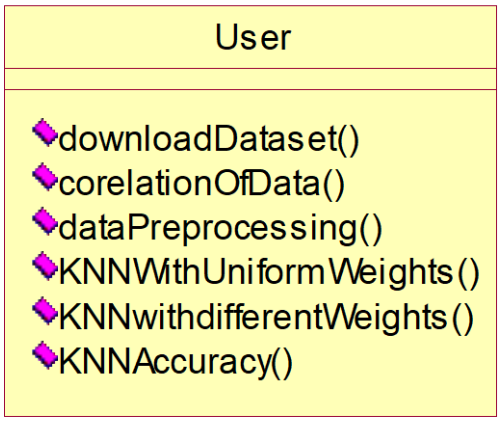
5.3 flowchart of knn probabilistic model

The flow chart in above Figures illustrates the process flow of the proposed model. Using the parallel approach, the model starts with computing the prior probabilities and the probabilities based on KNN approach simultaneously on both the Profit class and Loss class. KNN initialization process involves the use of the k value for the nearest neighbours of test instances. KNN then calculates the number of Profit class and Loss class instances based on the k number of nearest neighbours in the vicinity of each test instance. The outcome generated from KNN is then used by the probabilistic method for further classification.

5.2 UML DIAGRAMS

CLASS DIAGRAM :

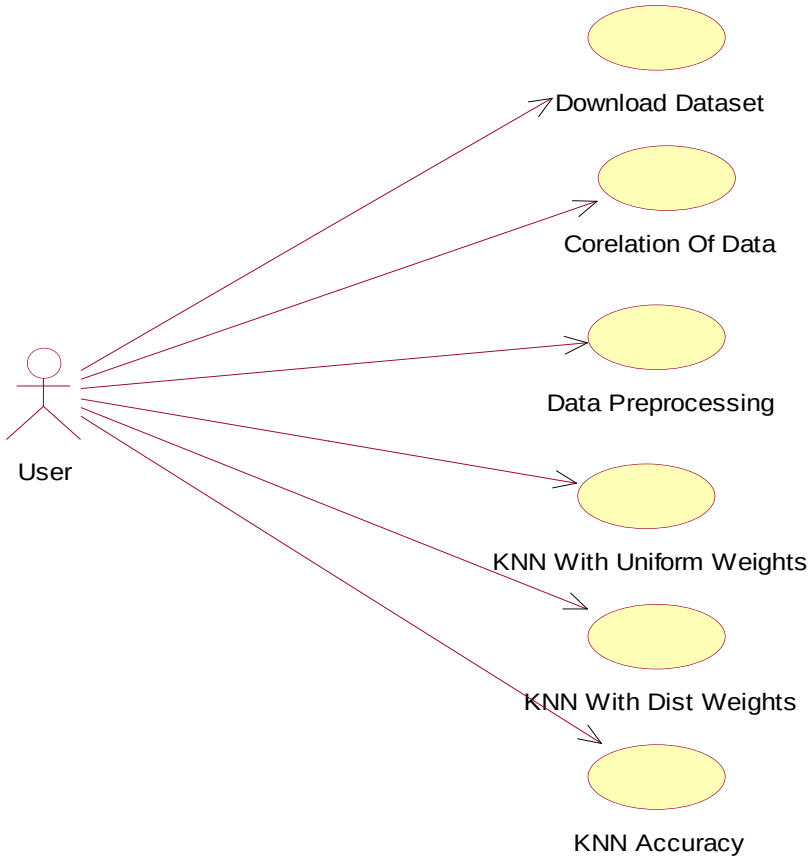
In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.



5.4 class diagram

USECASE DIAGRAM :

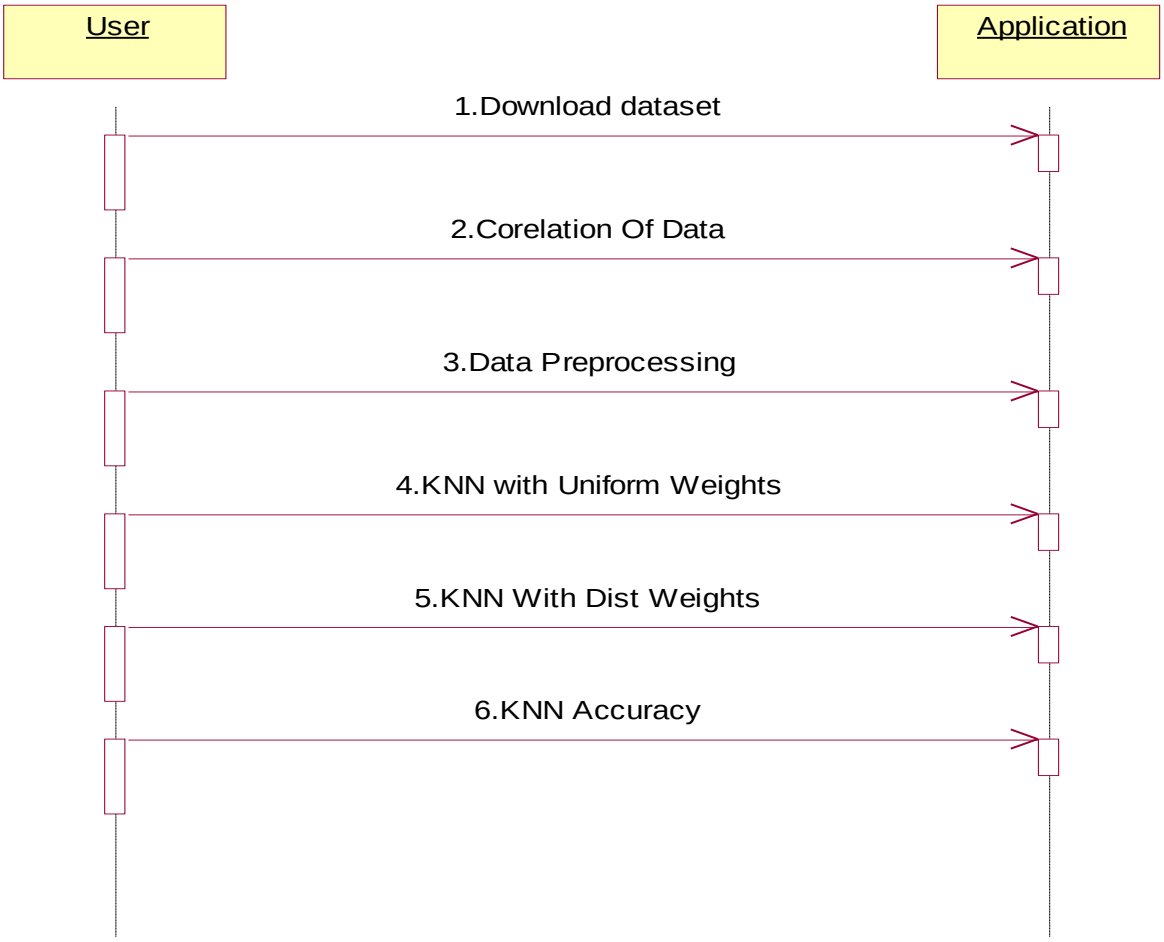
A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted



5.5 Usecase diagram

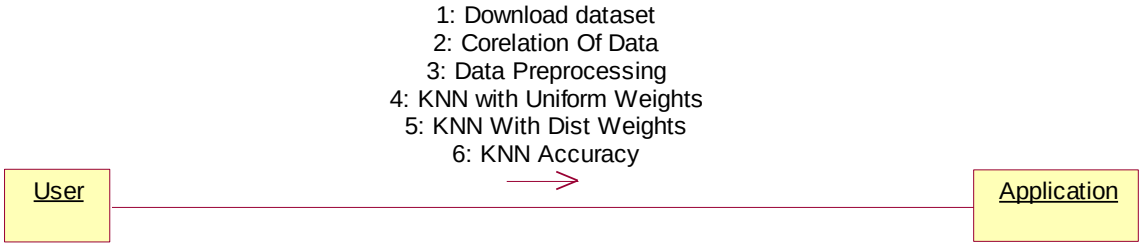
SEQUENCE DIAGRAM :

A sequence diagram in Unified Modeling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagram



5.6 Sequence diagram

COLLABORATION DIAGRAM :



CHAPTER 6
PROJECT CODING

6.1 CODE TEMPLATE

Stockpredictfinal.py

```
from tkinter import messagebox
from tkinter import *
from tkinter import simpledialog
import tkinter
from tkinter import filedialog
from imutils import paths
from tkinter.filedialog import askopenfilename

import pandas as pd
import datetime
import pandas_datareader.data as web
from pandas import Series, DataFrame
import matplotlib.pyplot as plt
from matplotlib import style
import matplotlib as mpl
from matplotlib import cm as cm
import math
import numpy as np
from sklearn import preprocessing
from sklearn.model_selection import train_test_split
from sklearn.neighbors import KNeighborsRegressor
```

```
import seaborn as sns
main = tkinter.Tk()
main.title("Stock Trend Using KNN")
main.geometry("1300x1200")

global dataFrame, dfreg
global moving_avg
global dfcomp
global clfknn
global clfknndist
global X, y, X_train, y_train, X_test, y_test,X_pred
global distknn, uniknn, knnunipred, knndistpred

def loadDataset():
    text.delete('1.0', END)
    global dataFrame
    global dfcomp
    start = datetime.datetime(2010, 1, 1)
    end = datetime.datetime(2017, 1, 11)

  dataFrame = web.DataReader("AAPL", 'yahoo', start, end)

    text.insert(END, "Shape of Apple Stock Dataset: "+str(dataFrame.shape)+"\n\n")
    text.insert(END, "Sample of Apple Stock Data: \n"+str(dataFrame.head(2))+"\n\n")
```

```
    dfcomp = web.DataReader(['AAPL', 'GE', 'GOOG', 'IBM', 'MSFT'], 'yahoo', start=start, end=end)['Adj Close']
    text.insert(END, "Shape of Apple Competitor Stock Dataset: " + str(dfcomp.shape) + "\n\n")
    text.insert(END, "Sample of Apple Competitor Stock Data: \n" + str(dfcomp.head(2)) + "\n\n")

    text.insert(END, "Dataset Downloaded from Yahoo Finance Dataset\n\n")

def dfcorr():
    text.delete('1.0', END)
    global dfcomp

    text.insert(END, "Correlation form Apple Competitor Stock\n\n")
    retscomp = dfcomp.pct_change()
    corr = retscomp.corr()
    text.insert(END, "correlation: \n"+str(corr)+"\n\n")

def dataPreProcess():
    text.delete('1.0', END)
    global dataFrame,dfreg
    global X, y, X_train, X_test, y_train, y_test,X_pred

    text.insert(END,"Data PreProcessing for Apple Stock Dataset\n\n")
    dfreg = dataFrame.loc[:,["Adj Close","Volume"]]
    dfreg["HL_PCT"] = (dataFrame["High"] - dataFrame["Low"]) / dataFrame["Close"] * 100.0
```

```
dfreg["PCT_change"] = (dataFrame["Close"] - dataFrame["Open"]) / dataFrame["Open"] * 100.0

# Drop missing value
dfreg.fillna(value=-99999, inplace=True)
# We want to separate 1 percent of the data to forecast
forecast_out = int(math.ceil(0.01 * len(dfreg)))

# Separating the label here, we want to predict the AdjClose
forecast_col = 'Adj Close'
dfreg['label'] = dfreg[forecast_col].shift(-forecast_out)
X = np.array(dfreg.drop(['label'], 1))

# Scale the X so that everyone can have the same distribution for linear regression
X = preprocessing.scale(X)
# Finally We want to find Data Series of late X and early X (train) for model generation and evaluation
X_pred = X[-forecast_out:]
X = X[:-forecast_out]
# Separate label and identify it as y
y = np.array(dfreg['label'])
y = y[:-forecast_out]

text.insert(END, "X lablels : \n"+str(X)+"\n\n")
text.insert(END, "Y lablels : \n"+str(y)+"\n\n")
text.insert(END, "Data spliting into Train and Test samples\n")
X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.3)
```

```
    text.insert(END, "number of Train Samples : " + str(len(X_train)) + "\n")
    text.insert(END, "number of Test Sample: " + str(len(X_test)) + "\n")

    text.insert(END, "Data Preprocessing Completed\n\n")

def uniformKNN():
    text.delete('1.0',END)
    global clfknn
    global uniknn

    # KNN Regression
    clfknn = KNeighborsRegressor(n_neighbors=5)
    clfknn.fit(X_train, y_train)

    uniknn = clfknn.score(X_train, y_train)

    text.insert(END, "Accuracy of KNN with Uniform weights : "+str(uniknn*100)+"\n\n")

def distKNN():
    text.delete('1.0', END)
    global clfknndist,knndistpred
    global distknn,knnunipred
    # KNN Regression
    clfknndist = KNeighborsRegressor(n_neighbors=5,weights='distance')
    clfknndist.fit(X_train, y_train)
    distknn = clfknndist.score(X_train, y_train)
```

```
    text.insert(END, "Accuracy of KNN with Distance weights : "+str(distknn*100)+"\n\n")
def predModel():
    text.delete('1.0', END)
    global clfknndist,clfknn,knnunipred,knndistpred
    global X, y, X_train, X_test, y_train, y_test


    filename = filedialog.askopenfilename(initialdir="Yahoo-Finance-Dataset")
    test = pd.read_csv(filename)
    text.insert(END, filename + " test file loaded\n"+str(test.columns)+"\n");
    x_pred = np.array(test.drop(['Unnamed: 0'],1))


    text.insert(END, "test Dataset: \n"+str(x_pred)+"\n\n");


    knndistpred = clfknndist.predict(x_pred)


    text.insert(END, "Predict values for KNN with Dist weights: \n" + str(knndistpred) + "\n\n");


    knnunipred = clfknn.predict(x_pred)


    text.insert(END, "Predict values for KNN with Uni Wights: \n" + str(knnunipred) + "\n\n");


def graph():
    text.delete('1.0', END)
```

```
global uniknn,distknn
global knnunipred,knndistpred
global dfreg

dfreg['Forecast'] = np.nan
last_date = dfreg.iloc[-1].name
last_unix = last_date
next_unix = last_unix + datetime.timedelta(days=1)

for i in knnunipred:
    next_date = next_unix
    next_unix += datetime.timedelta(days=1)
    dfreg.loc[next_date] = [np.nan for _ in range(len(dfreg.columns) - 1)] + [i]
dfreg['Adj Close'].tail(500).plot()
dfreg['Forecast'].tail(500).plot()
plt.legend(loc=4)
plt.xlabel('Date')
plt.ylabel('Price')
plt.savefig('knnUniformPredGraph.png')
plt.close()
for i in knndistpred:
    next_date = next_unix
    next_unix += datetime.timedelta(days=1)
    dfreg.loc[next_date] = [np.nan for _ in range(len(dfreg.columns) - 1)] + [i]
dfreg['Adj Close'].tail(500).plot()
dfreg['Forecast'].tail(500).plot()
```

```
    plt.legend(loc=4)
    plt.xlabel('Date')
    plt.ylabel('Price')
    plt.savefig('knnDistPredGraph.png')
    plt.close()
    height = [uniknn,distknn]
    bars = ('KNN with uniform weights Accuracy', 'KNN with distance weights Accuracy')
    y_pos = np.arange(len(bars))
    plt.bar(y_pos, height)
    plt.xticks(y_pos, bars)
    plt.show()
font = ('times', 16, 'bold')
title = Label(main, text='Stock Trend Prediction Using KNN')
title.config(bg='PaleGreen2', fg='Khaki4')
title.config(font=font)
title.config(height=3, width=120)
title.place(x=0,y=5)
font1 = ('times', 14, 'bold')
uploadButton = Button(main, text="Download Dataset", command=loadDataset)
uploadButton.place(x=700,y=100)
uploadButton.config(font=font1)
corrButton = Button(main, text="Correlation for Data", command=dfcorr)
corrButton.place(x=700,y=150)
corrButton.config(font=font1)
ppButton = Button(main, text="Data Preprocessing", command=dataPreProcess)
ppButton.place(x=700,y=200)
ppButton.config(font=font1)
```

```
uniformButton = Button(main, text="Run KNN with Uniform Weights", command=uniformKNN)

uniformButton.place(x=700,y=250)

uniformButton.config(font=font1)

distButton = Button(main, text="Run KNN with Distance Weights", command=distKNN)

distButton.place(x=700,y=300)

distButton.config(font=font1)

predButton = Button(main, text="Predict the Test Data ", command=predModel)

predButton.place(x=700,y=350)

predButton.config(font=font1)

graphButton = Button(main, text="KNN Accuracy", command=graph)

graphButton.place(x=700,y=400)

graphButton.config(font=font1)

font1 = ('times', 12, 'bold')

text=Text(main,height=30,width=80)

scroll=Scrollbar(text)

text.configure(yscrollcommand=scroll.set)

text.place(x=10,y=100)

text.config(font=font1)

main.config(bg='PeachPuff2')

main.mainloop()
```

6.2 OUTLINE FOR VARIOUS FILES

Stockpredictfinal.py

We used python programming to implement our project. We used single python file to implement or code. This file consists of various modules that we have used. Our project modules are – Download data, correlation data, data pre-processing,

KNN accuracy with uniform weights, KNN accuracy with distance weights, predict test data and KNN accuracy.

6.3 CLASS WITH FUNCTIONALITIES

In the project Home page, we have only one class i.e., USER and user navigate to seven functional modules, they are - Download data, correlation data, data pre-processing, KNN accuracy with uniform weights, KNN accuracy with distance weights, predict test data and KNN accuracy. Let us look into the brief overview of each functional module.

a) Download Dataset :

Here user download dataset from Yahoo Finance, which provides financial news, data and commentary including stock quotes, press release, financial reports etc.

b) Correlation Data :

Correlation data shows the strength of a relationship between two stocks which is expressed in numerically by the correlation coefficient. The correlation coefficient's value ranges from -1.0 to 1.0.

c) Data pre-processing :

This module is responsible for encoding of data, dropping missing values, split labels and split dataset.

d) KNN accuracy with Uniform weights :

Generates a KNN model with uniform weights and calculate accuracy of the values predicted using that model.

e) KNN accuracy with Distance weights :

Generates a KNN model with distance weights and calculate accuracy of the values predicted using that model.

f) Predict data :

Here we upload test data and predict future values using two models they are- KNN model with uniform weights and KNN model with distance weights.

g) KNN accuracy :

Accuracy comparison of models for different models using bar graph, in this x-axis represent algorithm names and y-axis represent accuracy rates.

6.4 METHODS INPUT AND OUTPUT PARAMETERS

We implemented various methods, which include :

loadDataset()

dfcorr()

dataPreProcess

uniformKNN()

distKNN()

preModel()

grapgh()

localDataset method is used to take stock dataset as input, dfcorr method is used to find correlation between various stocks, dataPreProcess takes the downloaded dataset and input process it and split data into training and training heads, unifromKNN using training and testing data generate a KNN Model and calculates accuracy, distKNN using training and testing data generate a KNN Model and calculates accuracy, preModel predicts future values and grapgh takes the predicted values as input and plot graph comparing accuracy of different models.

CHAPTER 7

PROJECT TESTING

7.1 VARIOUS TEST CASES

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

TYPES OF TESTS

Unit testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

Integration testing

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration

testing is specifically aimed at exposing the problems that arise from the combination of components.

Functional test

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals. Functional testing is centred on the following items:

- Valid Input : identified classes of valid input must be accepted.
- Invalid Input : identified classes of invalid input must be rejected.
- Functions : identified functions must be exercised.
- Output : identified classes of application outputs must be exercised.
- Systems/Procedures : interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

System Test

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of 49 system testing is the configuration oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

Unit Testing

Unit testing is usually conducted as part of a combined code and unit test phase of the software lifecycle, although it is not uncommon for coding and unit testing to be conducted as two distinct phases.

Test strategy and approach

Field testing will be performed manually and functional tests will be written in detail. Test objectives

➢ All field entries must work properly.

➢ Pages must be activated from the identified link.

➢ The entry screen, messages and responses must not be delayed. Features to be tested

➢ Verify that the entries are of the correct format

➢ No duplicate entries should be allowed

➢ All links should take the user to the correct page.

Integration Testing

Software integration testing is the incremental integration testing of two or more integrated software components on a single platform to produce failures caused by interface defects.

The task of the integration test is to check that components or software applications, e.g. components in a software system or – one step up – software applications at the company level – interact without error.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

Acceptance Testing

User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements.

Test Results: All the test cases mentioned above passed successfully. No defects encountered

7.2 Black Box Testing

Black box testing is a technique of software testing which examines the functionality of software without peering into its internal structure or coding. The primary source of black box testing is a specification of requirements that is stated by the customer. In this method, tester selects a function and gives input value to

examine its functionality, and checks whether the function is giving expected output or not. If the function produces correct output, then it is passed in testing, otherwise failed. The test team reports the result to the development team and then tests the next function. After completing testing of all functions if there are severe problems, then it is given back to the development team for correction. Generic steps of black box testing

➢ The black box test is based on the specification of requirements, so it is examined in the beginning.

➢ In the second step, the tester creates a positive test scenario and an adverse test scenario by selecting valid and invalid input values to check that the software is processing them correctly or incorrectly.

➢ In the third step, the tester develops various test cases such as decision table, all pairs test, equivalent division, error estimation, cause-effect graph, etc.

➢ The fourth phase includes the execution of all test cases.

➢ In the fifth step, the tester compares the expected output against the actual output.

➢ In the sixth and final step, if there is any flaw in the software, then it is cured and tested again.

7.3White Box Testing

The box testing approach of software testing consists of black box testing and white box testing. We are discussing here white box testing which also known as glass box is testing, structural testing, clear box testing, open box testing and transparent box testing. It tests internal coding and infrastructure of a software focus on checking of 42 predefined inputs against expected and desired outputs. It is based on inner workings of an application and revolves around internal structure testing. In this type of testing programming skills are required to design test cases. The primary goal of white box testing is to focus on the flow of inputs and outputs through the software and strengthening the security of the software.

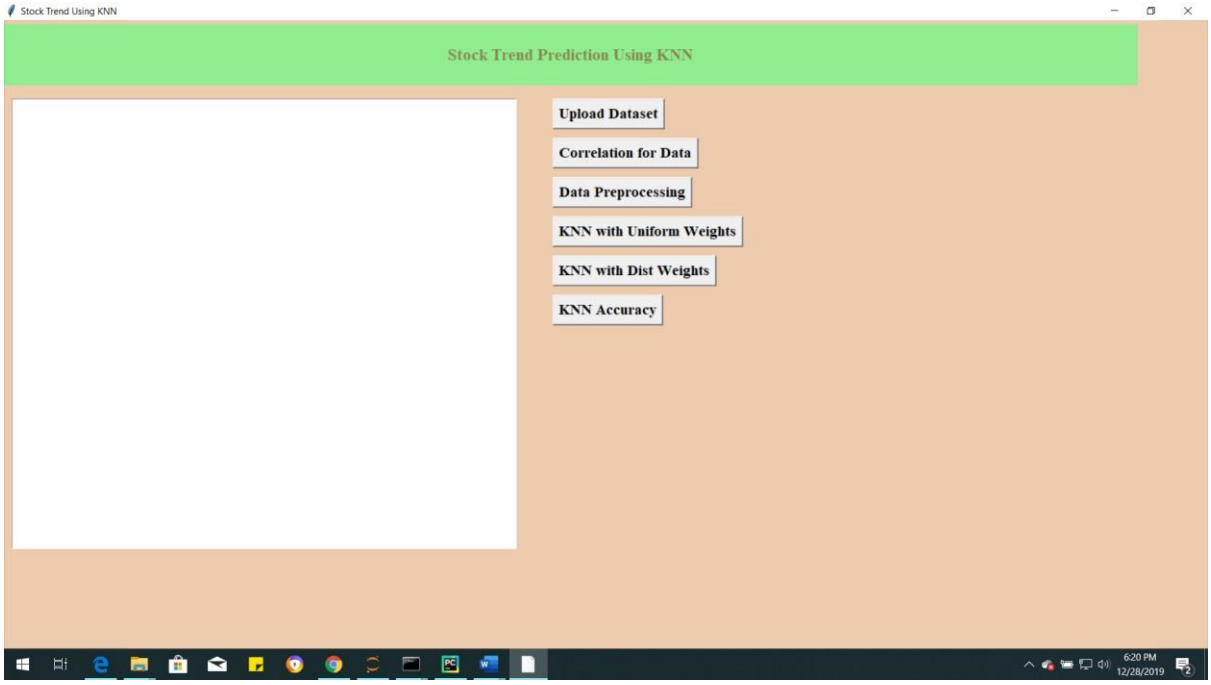
The term 'white box' is used because of the internal perspective of the system. The clear box or white box or transparent box name denote the ability to see through the software's outer shell into its inner workings.

Developers do white box testing. In this, the developer will test every line of the code of the program. The developers perform the White-box testing and then send the application or the software to the testing team. requirements and does one round of white box testing and sends it to the testing team.

CHAPTER 8

OUTPUT SCREENS

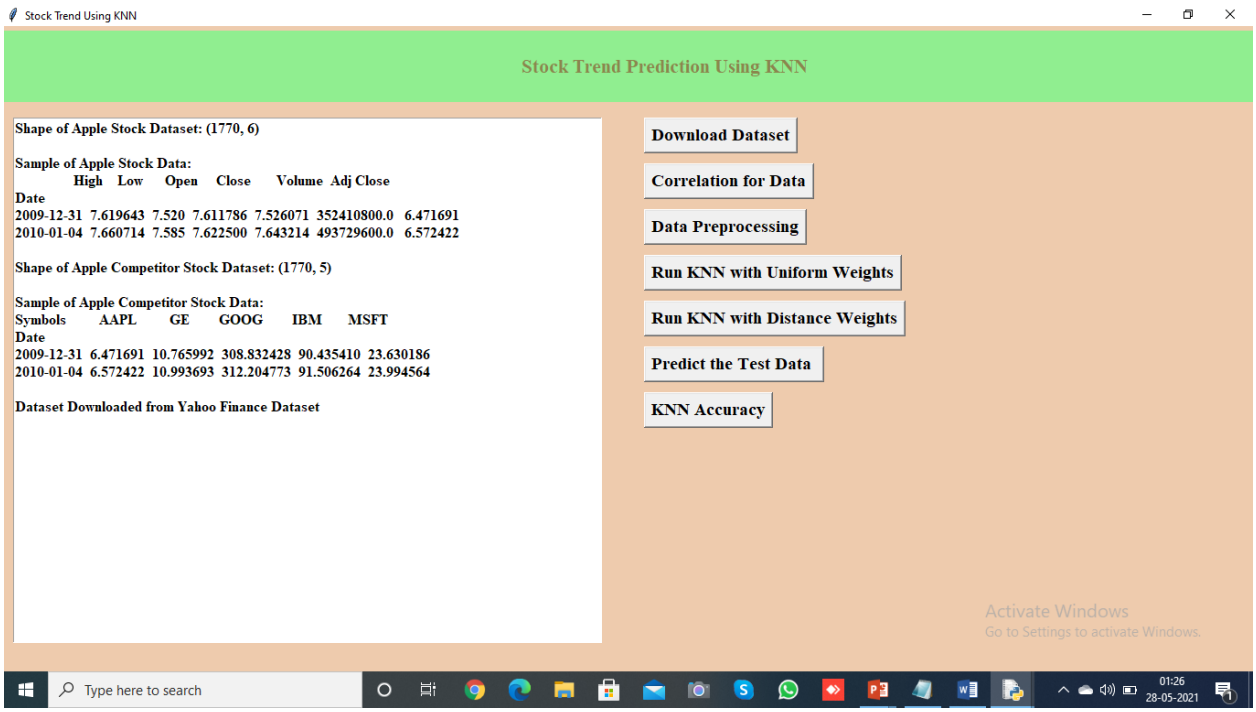
8.1 User Interfaces



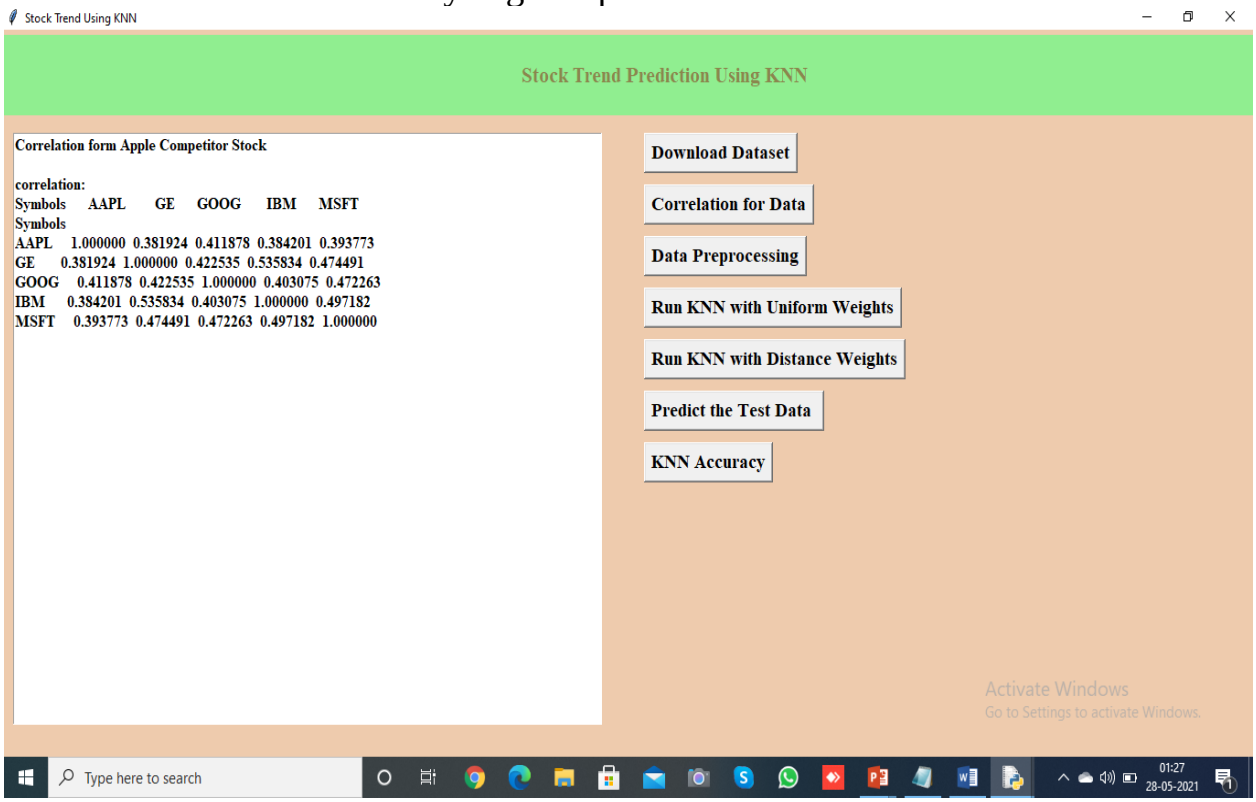
8.1 Home screen

8.2 Output Screens

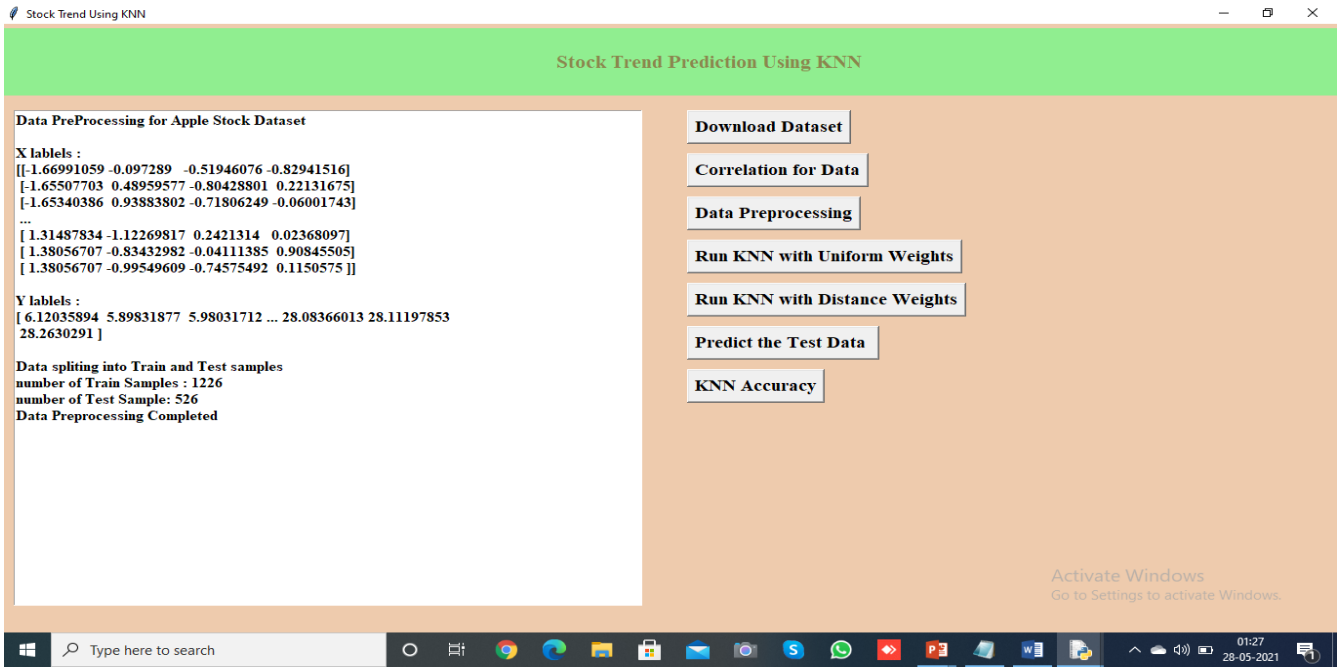
- Analysing data : loading yahoo finance dataset



- Correlation data : Analysing competitors stock data

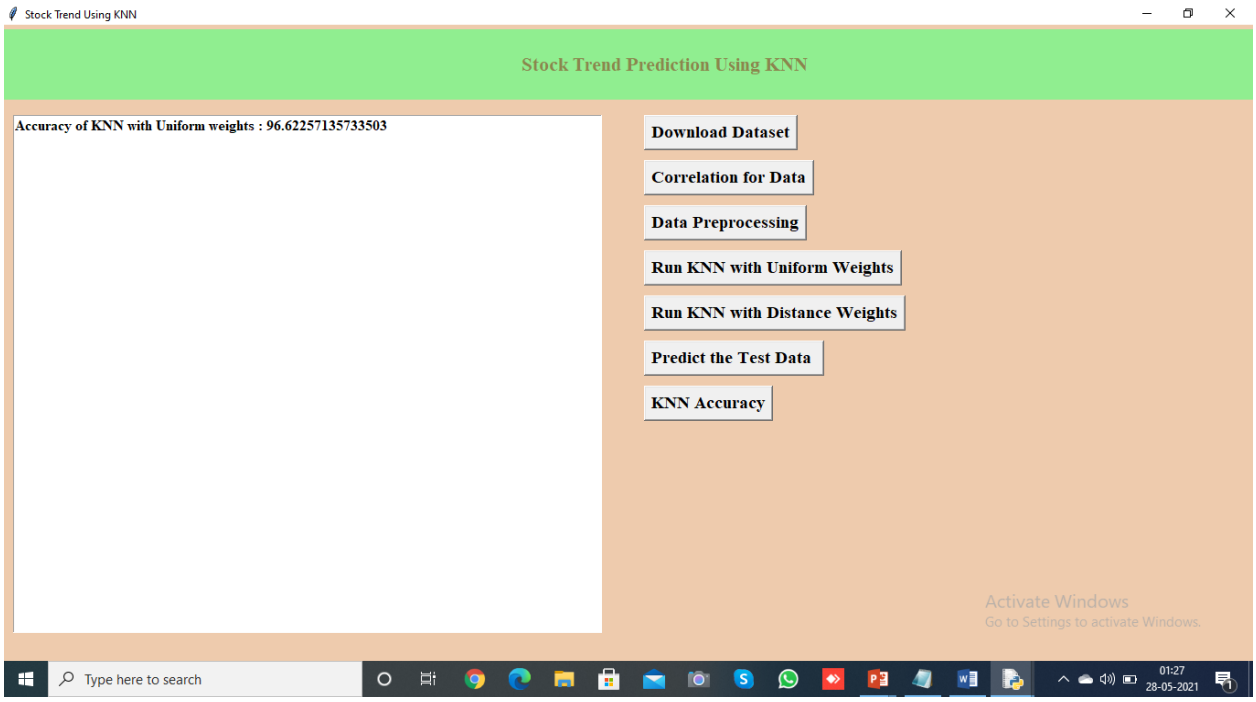


8.3 Corelation Data



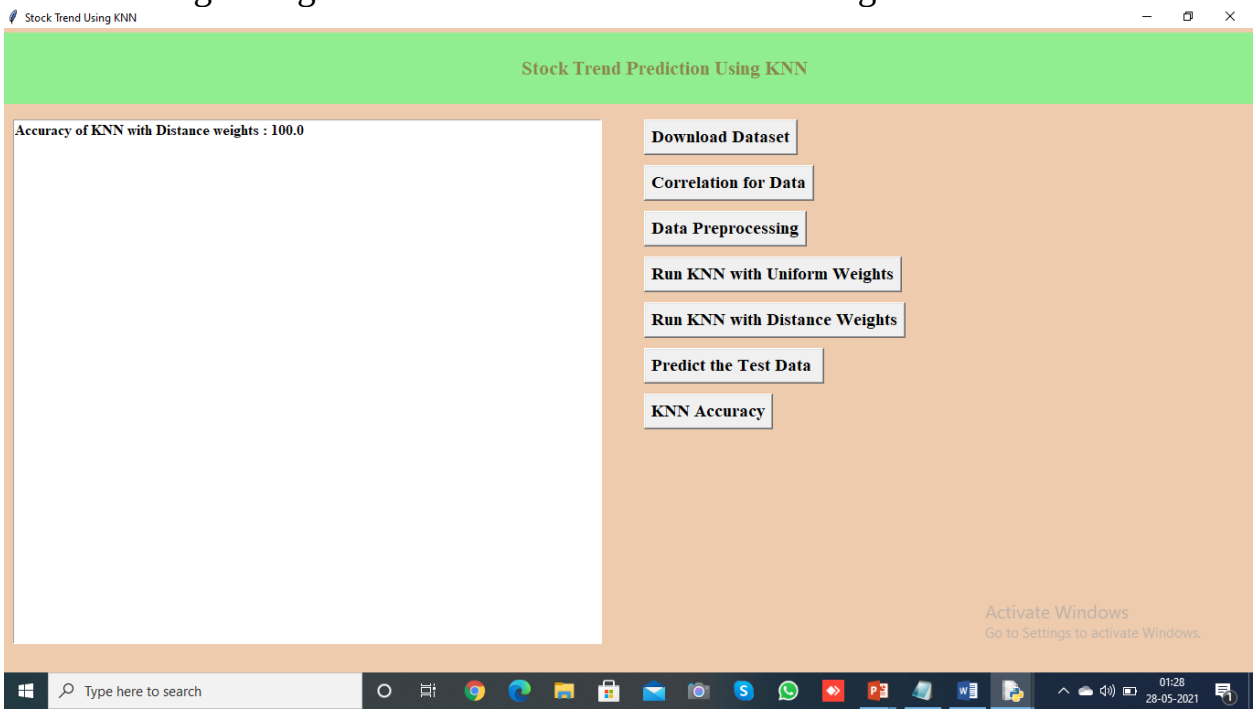
8.4 Data preprocessing and dataset split into training and testing heads

- Testing the knn generated model with uniform weights



8.5 Knn model with uniform weights accuracy

- Testing with generated knn model with distant weights

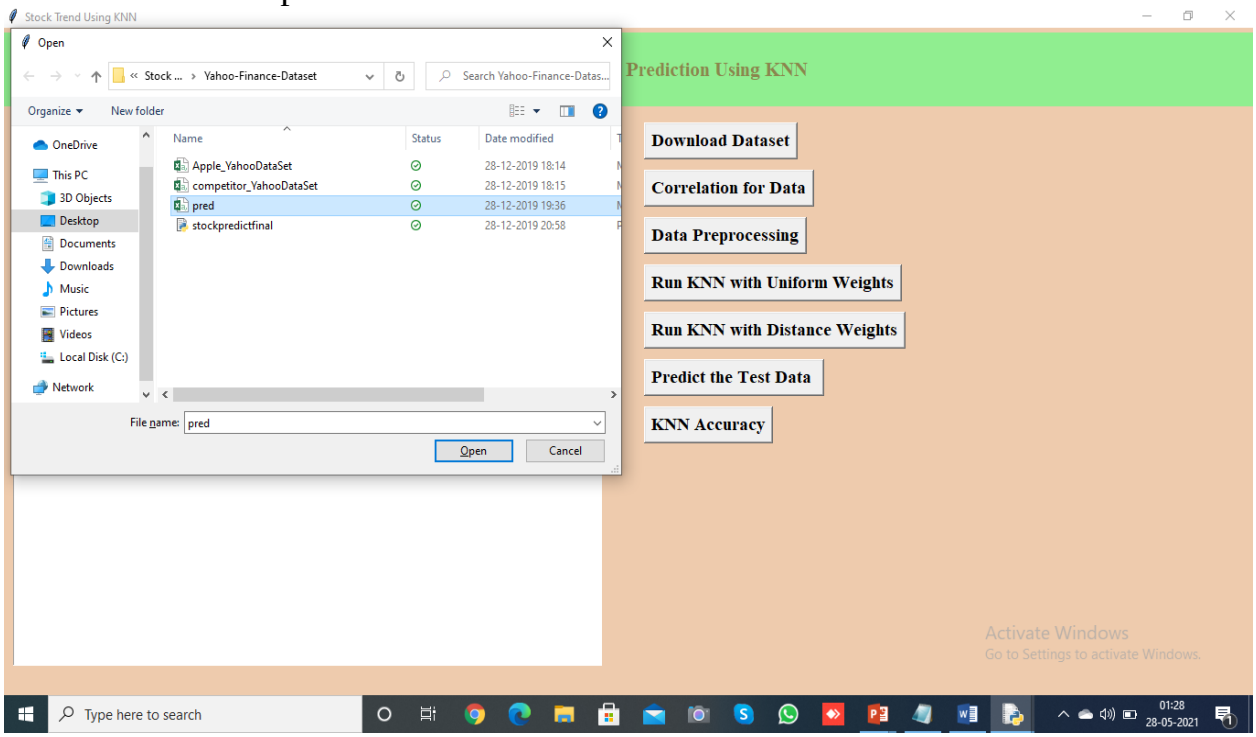


8.6 Knn model with distance weight accuracy

CHAPTER 9

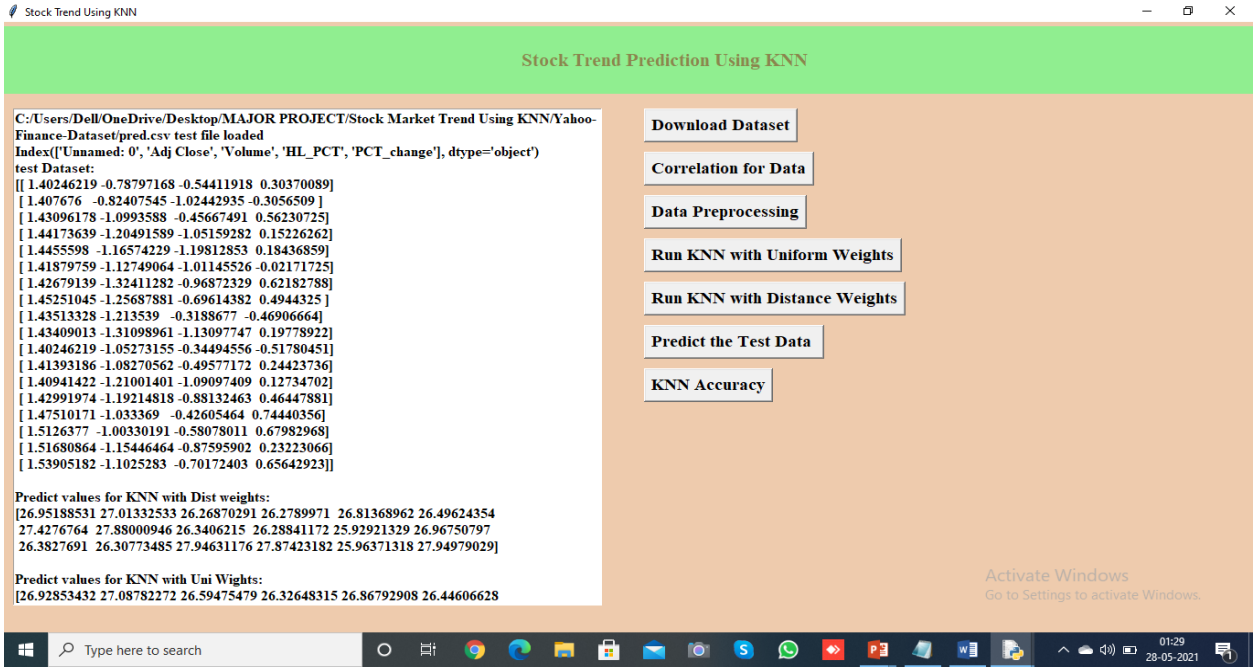
EXPERIMENTAL RESULT

- Test Data Upload



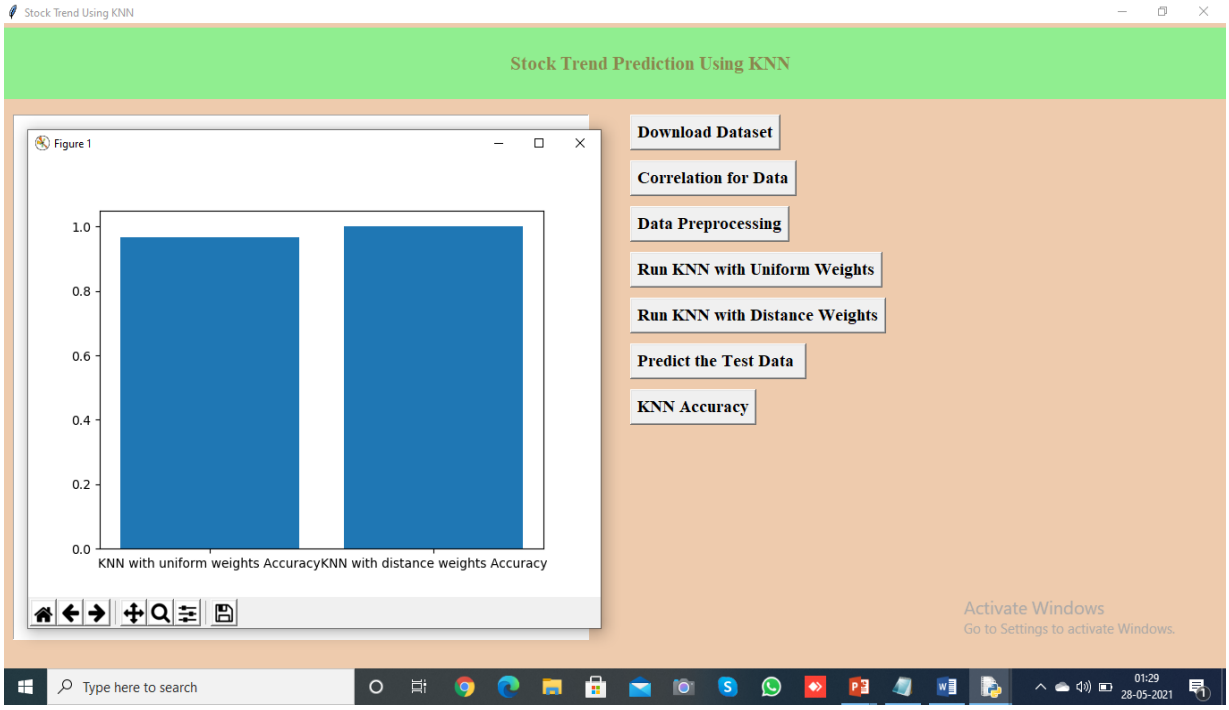
9.1 Test data upload

- Predicted future values



9.2 Predicted values using different models

- KNN Accuracy evaluation with different models



9.3 Knn models accuracy comparision

From the above graph we can see that the distance weights has little bit better accuracy compare to uniform weights, in above graph x-axis contains algorithm name and y-axis represents accuracy of that algorithm.

CHAPTER 10

CONCLUSION AND FUTURE ENHANCEMENT

- The aim of this research is to improve the statistical fitness of the proposed model to overcome a KNN problem due to its computation approach. The KNN classifier can compute the empirical distribution over the Profit and Loss class values in the k number of nearest neighbours. However, the outcome is less than adequate due to sparse data. The KNN classifier has under fitting issue as it does not cater to generalization of sparse data outside the range of nearest neighbourhood. We have compared a hybrid KNN-Probabilistic model with four standard algorithms on the problem of predicting the stock price trends. Our results showed that the proposed KNN-Probabilistic model leads to significantly better results compared to the standard KNN algorithm and the other classification algorithms. The limitation of the proposed model is that it applies a binary classification technique. The actual output of this binary classification model is a prediction score in twoclass. The score indicates the model's certainty that the given observation belongs to either the Profit class or Loss class.

- For future work, the knowledge component is to transform the binary classification into multiclass classification. The multiclass classification involves observation and analysis of more than the existing two statistical class values. Additional research will include the application of the probabilistic model to multiclass data in order to provide more specific information of each class value. The newly formed multiclass classification will contain five class labels named “Sell”, “Underperform”, “Hold”, “Outperform”, and “Buy”. In numerical values for mapping purpose, we will convert “Sell” to -2 which implies strongly unfavourable; “Underperform” to -1 which implies moderately unfavourable; “Hold” to 0 which implies neutral; “Outperform” to 1 which implies moderately favourable; and “Buy” to 2 which implies strongly favourable.

CHAPTER 11

REFERENCES

1)Charles D. Kirkpatrick II and Julie R. Dahlquist, Technical Analysis: The Complete Resource for Financial Market Technicians , Pearson Education, Inc., 2020.

2)Monica Tirea and Viorel Negru, Intelligent Stock Market Analysis System - A Fundamental and Macro-economical Analysis Approach, IEEE, 2020.

3)Yauheniya Shynkevicha, T.M. McGinnity, Sonya A. Coleman, and Ammar Belatreche, Forecasting movements of health-care stock prices Vol. 85, 2020.

4)Chi Ma, Junnan Liu, Hongyan Sun, and Haibin Jin, A hybrid financial time series model based on neural networks, IEEE, 2020.

5)Gérard Biau & Luc Devroye, Lectures on the Nearest Neighbour Method, Springer, 2019.

6)Saed Sayad, K Nearest Neighbors - Classification, [Online]. 2019.

7)Morris, Bayes' Theorem: A Visual Introduction for Beginners, Blue Windmill Media, 2019.

8)James V Stone, Bayes' Rule With R: A Tutorial Introduction to Bayesian Analysis, Sebtel Press, 2019.

9)Marco Scutari and Jean-Baptiste Denis, Bayesian Networks CRC Press, 2019.

10)Peter Bruce and Andrew Bruce, Practical Statistics for Data Scientists: O'Reilly Media, 2019.

11)Ciaran Walsh, Key Management Ratios, 4th Edition (Financial Times Series), Prentice Hall, 2019.

12)Seyed Enayatolah Alavi, Hasanali Sinaei, and Elham Afsharirad, Predict the Trend of Stock Prices Using Machine Learning Techniques, IAIEST, 2018.

13)Abdolhossein Zameni and Othman Yong, Share Price Performance of Malaysian IPOs Around Lock-Up Expirations, 2018, pp. 8094-8108.

14) Lida Nikmanesh and Abu Hassan Shaari Mohd Nor, Macroeconomic Determinants of Stock Market Volatility Vol 21(1), 2018, pp. 161- 180.

15) Luigi Troiano, Pravesh Kriplani, and Irene Díaz, Regression Driven F-Transform and Application to Smoothing of Financial Time Series, IEEE, 2018.

16)Othman Yong Ciaran Walsh, Key Management Ratios, 4th Edition (Financial Times Series), Prentice Hall, 2009.

17)Seyed Enayatolah Alavi, Hasanali Sinaei, and Elham Afsharirad, Predict the Trend of Stock Prices Using Machine Learning Techniques, IAIEST, 2015.

18)Lock Siew Han and Md Jan Nordin, Integrated Multiple Linear Regression-One Rule Classification Model for the Prediction of Stock Price Trend, Journal of Computer Sciences, Vol 13 (9), 2017, pp. 422-429.

19)Abdolhossein Zameni and Othman Yong, Share Price Performance of Malaysian IPOs Around Lock-Up Expirations, Advanced Science Letters, Vol 23(9), 2017, pp. 8094-8102

20)Hani A.K. Ihlayyel, Nurfadhlina Mohd Sharef, Mohd Zakree Ahmed Nazri, and Azuraliza Abu Bakar, An Enhanced Feature Representation Based On Linear Regression Model For Stock Market Prediction, Intelligent Data Analysis, Vol 22(1), 2018, pp. 45-76.

21)Lida Nikmanesh and Abu Hassan Shaari Mohd Nor, Macroeconomic Determinants of Stock Market Volatility: An Empirical Study of Malaysia and Indonesia, Asian Academy of Management Journal, Vol 21(1), 2016, pp. 161-180.

22)Luigi Troiano, Pravesh Kriplani, and Irene Díaz, Regression Driven F-Transform and Application to Smoothing of Financial Time Series, IEEE,2017.

23)Fu-Yuan Huang, Integration of an Improved Particle Swarm Algorithm and Fuzzy Neural Network for Shanghai Stock Market Prediction, 2008 Workshop on Power Electronics and Intelligent Transportation System.[Online].August 2008, IEEE, http://ieeexplore.ieee.org/document/4634852/, 2008, pp. 242–247.

24)Ching-hsue Cheng, Tai-liang Chen, and Hiajong Teoh, Multiple-Period Modified Fuzzy Time-Series for Forecasting TAIEX, Fourth International Conference on FuzzySystems and Knowledge Discovery (FSKD 2007), http://ieeexplore.ieee.org/document/4406190/, IEEE, 2007, pp. 2–6.

25)Pei-Chann Chang, Chin-Yuan Fan, and ShihHsin Chen, Financial Time Series Data Forecasting by Wavelet and TSK Fuzzy Rule Based System, Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007), http://ieeexplore.ieee.org/document/4406255/, IEEE, 2007, pp. 331–335.

CHAPTER 12

PUBLICATIONS

- **JOURNAL (UGC approved Journal)**
- **CONFERENCE (International Conference on "Innovations in computers Networks, Computational Intelligence and IoT" [ICICCI-21-0062],C13 batch).**
- **PAPER ID:ICICCI-21-0061**
- **PAPER TITLE:STOCK MARKET TREND PREDICTION USING KNN ALGORITHM**

CHAPTER 13

STUDENTS PROFILE

BYRAGONI SRIPRIYA is currently pursuing her Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. She completed her intermediate from Sri Gayatri Junior college and 10th class from Avanthi High School. She completed her 15 days of internship on "Machine learning through python" at code mania. Her technical skills include C, C++ and Python. She also has a basic understanding of Java. She took part in Employability Skill development Program conducted by Zensar. She is also a student of Smart Interviews. Her participations include: National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020, the Global Webinar on Cloud & Big Data – Changing the way we work which was conducted Global Education & Careers Forum (GECF) on 12th August 2020, Women online workshop on "Women in Cyber Security and Privacy in 2020" which was conducted from 6th to 10th July 2020. Her areas of interest are Python, Machine Learning and Deep Learning. She completed few certification courses from online platforms like Coursera, cursaApp, Udemy and Solo Learn.

SINGA MANEESHA is currently pursuing her Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. She completed her intermediate from Narayana Junior College and 10th class from N.S.K.K High School. She is one of the members of Coders Club in our college. Her technical skills include C, Python, Java and SQL.She took part in Employability Skill development Program conducted by Zensar. She is also a student of Smart Interviews. Her participations include: Women online workshop on "Women in Cyber Security and Privacy in 2020" which was conducted from 6th to 10th July 2020, Machine Learning Workshop conducted by TAM on 8th and 9th February 2018, Two - Days National level seminar on "Recent trends in cloud computing, Fog and Edge Computing " from 18th june to 19th june 2021 and International Conference "Innovations in Computer Networks, Computational Intelligence and IoT"(ICICCI-21) conducted on 25th and 26th june 2021. She completed few certification courses from online platforms which include –AI for everyone from Coursera, AWS fundamentals: Going Cloud native from Coursera, Data science math skills from Coursera, Leadership and Emotional intelligence from Coursera, Managing projects risks from Coursera, VR and 360 Video production from courser and Java script from Net Ninja. Her areas of interest are Artificial Intelligence, Machine Learning and Deep Learning. She is also an active sports player, participated in various intra-college sports which include Kabaddi, Cricket, Throw ball, Basket Ball.

GOVINDARAJU PREMIKA is currently pursuing her Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. She completed her intermediate from Narayana Junior College and 10th class from Bhashyam High School. She is one of the members of Coders Club in our college. Her technical skills include C, Python, Java and SQL.She took part in Employability Skill development Program conducted by Zensar. She is also a student of Smart Interviews. Her participations include: Workshop on windows Mobile app development at IARE college, Women online workshop on "Women in Cyber Security and Privacy in 2020" which was conducted from 6th to 10th July 2020, Two - Days National level seminar on "Recent trends in cloud computing, Fog and Edge Computing " from 18th june to 19th june 2021 and International Conference "Innovations in Computer Networks, Computational Intelligence and IoT"(ICICCI-21) conducted on 25th and 26th june 2021. She completed few certification courses from online platforms which include –AI for everyone from Coursera, AWS fundamentals: Going Cloud native from Coursera, Data science math skills from Coursera, Leadership and Emotional intelligence from Coursera, Managing projects risks from Coursera, VR and 360 Video production from courser and Java script from Net Ninja. Her areas of interest are Artificial Intelligence, Machine Learning and Deep Learning.

DANDAMRAJJU SRIPRIYA is currently pursuing her Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. She completed her intermediate from Sri Chaitanya Junior College and 10th class from Jeevadhan High School. She is one of the members of Coders Club in our college. Her technical skills include HTML, CSS, Bootstrap, Javascript, jQuery, C, Python, Java and SQL. She also has a basic understanding of Machine learning. She took part in Employability Skill development Program conducted by Zensar. She is also a student of Smart Interviews. Her participations include: Women online workshop on "Women in Cyber Security and Privacy in 2020" which was conducted from 6th to 10th July 2020, Machine Learning Workshop conducted by TAM on 8th and 9th February 2018, National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020, "Know More - Teach More ",Two - Days National level seminar on "Recent trends in cloud computing, Fog and Edge Computing " from 18th june to 19th june 2021 and International Conference "Innovations in Computer Networks, Computational Intelligence and IoT"(ICICCI-21) conducted on 25th and 26th june 2021. She completed few certification courses from online platforms which include – Front end development from Udemy, HTML, CSS, Javascript from courser, AI for Everyone from coursera and also participated in Hackathon with a team of 4 members to explain air pollution index in a particular area. Her areas of interest are Artificial Intelligence, Machine Learning and Deep Learning.

A

PROJECT REPORT

On

BLOCK CHAIN FOR SECURE EHRS SHARING OF MOBILE CLOUD BASED E-HEALTH SYSTEMS

Submitted by

1) A.Suraj Reddy(17K81A05C5) 2)E.Chakravarthi Reddy(17K81A05D5)

3) B.Niharika (17K81A05D1) 4) P.Nandini(17K81A05G9)

In partial fulfillment for the award of the degreeof

BACHELOR OF TECHNOLOGY

IN

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Under the Guidance of

Ms. S.Swetha

Assistant Professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ST.MARTIN'S ENGINEERING COLLEGE
An Autonomous Institute

Dhulapally, Secunderabad – 500 100

JUNE 2021

BONAFIDE CERTIFICATE

This is to certify that the project entitled **BLOCK CHAIN FOR SECURE EHRS SHARING OF MOBILE CLOUD BASED E-HEALTH SYSTEMS**, is being submitted by **1. Mr. A.Suraj Reddy (17K81A05C5), 2. Mr. E.Chakravarthi Reddy (17K81A05D5), 3. Ms.B.Niharika (17K81A05D1), 4. Ms. P. Nandini (17K81A05G9)** in partial fulfillment of the requirement for the award of the degree of **BACHELOR OF TECHNOLOGY IN COMPUTER SCIENCE AND ENGINEERING** is recorded of bonafide work carried out by them. The result embodied in this report have been verified and found satisfactory.

Assistant Professor
Ms. S.SWETHA
Department of CSE

Head of the Department
Dr.M.NARAYANAN
Department of CSE

Internal Examiner

External Examiner

Place:

Date:

DECLARATION

We, the student of **Bachelor of Technology** in Department of Computer Science and Engineering', session: 2017 – 2021, St. Martin's Engineering College, Dhulapally, Kompally, Secunderabad, hereby declare that work presented in this Project Work entitled **BLOCK CHAIN FOR SECURE EHRS SHARING OF MOBILE CLOUD BASED E-HEALTH SYSTEMS** is the outcome of our own bonafide work and is correct to the best of our knowledge and this work has been undertaken taking care of Engineering Ethics. This result embodied in this project report has not been submitted in any university for award of any degree.

Mr. A. Suraj Reddy (17K81A05C5)
Mr. E.Chakravarthi Reddy(17K81A05D5)
Ms B. Niharika (17K81A05D1)
Ms. P. Nandini (17K81A05G9)

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible and whose encouragements and guidance have crowded effects with success.

We extended our deep sense of gratitude to Principal**, Dr. P. SANTOSH KUMAR PATRA**, St. Martin's Engineering College, Dhulapally, for permitting us to undertake this project.

We are also thankful to **Dr. M.NARAYANAN**, Head of the Department, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

We would like to thank **Dr. T. POONGOTHAI,** Department of Computer Science and Engineering (AI & ML) for her encouragement and insightful comments and as well as our project coordinators **Dr. B.RAJALINGAM**, Associate Professor and **Dr. N. SATHEESH**, Professor, in Department of Computer Science and Engineering for their valuable support.

We would like to express our sincere gratitude and indebtedness to our project supervisor **Ms. S.SWETHA**, Assistant Professor, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

Finally, we express thanks to all those who have helped us successfully to completing this project. Furthermore, we would like to thank our family and friends for their moral support and encouragement.

We express thanks to all those who have helped us in successfully completing the project.

Mr. A. Suraj Reddy (17K81A05C5)

Mr. E.Chakravarthi Reddy (17K81A05D5)

Ms B. Niharika (17K81A05D1)

Ms. P. Nandini (17K81A05G9)

ABSTRACT

The main objective of this project is securely store and maintains the patient records in the healthcare. Healthcare is a data-intensive domain where a large amount of data is created, disseminated, stored, and accessed daily. The blockchain technology is used to protect the healthcare data hosted within the cloud.

The block that contain the medical data and the timestamp. Cloud computing will connect different healthcare providers. It allows healthcare provider to access the patient details more securely from anywhere. It preserve data from attackers. The data is encrypted prior to outsourcing to the cloud. The healthcare provider have to decrypt the data prior to download.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF OUTPUT SCREENS

LIST OF ABBREVIATIONS

APHL	Association of Public Health Laboratories
EHR	Electronic Health Record
EHR-S	Electronic Health Record System
HIE	Health Information Exchange

CHAPTER 1

INTRODUCTION

The recent advent of technology affects all elements of human life and is dynamic the means we tend to use and understand things antecedent. Rather like the changes technology has offered in numerous alternative sectors of life, it is conjointly finding new ways for improvement within the health care sector. The most advantage that advancement in technology is providing square measure to enhance security, user expertise, and alternative aspects of the health care sector. These advantages were offered by Electronic Health Record systems (EHRs). However, they still face some problems relating to the protection of medical records, user possession of information, knowledge integrity, etc. the answer to those problems may be the employment of unique technology, i.e., Blockchain. This technology offers to produce a secure, tamper-proof platform for storing medical records and alternative healthcare-related info.

Block chain may be a distributed system recording and storing group action records. A lot specifically, block chain could also be a shared, immutable record of peer-to-peer group actions engineered from coupled transaction blocks and keep throughout a digital ledger. Blockchain depends on established cryptology techniques to allow every participant throughout a network to move (e.g. store, exchange, and think about information), while not pre-existing trust between the parties. During a blockchain system, there is no central authority; instead, group action records are kept and distributed across all network participants. Interactions with the blockchain become notable to all or any or any participants and want verification by the network before data is additional, enabling trustless collaboration between network participants whereas recording associate immutable audit path of all interactions.

Given that one of the inherent properties of blockchain is its decentralized nature, during which information possession is placed within the hands of individual users, some have planned that blockchain is also a lot of fittingly applied to EHRs. Blockchain technology facilitates the secure transfer of patient medical records, manages the drugs offer chain, and helps attention researchers to diagnose patient health issues. The goal of EHR systems was to unravel the issues long-faced by paper-based healthcare records Associated to supply an economical system that may rework the state of the healthcare sector.

1.1 PROJECT OVERVIEW

We propose a methodology to overcome the security problems that are occurred in the existing system and effectively store the data over the cloud we introduce this system. The data user outsources the encrypted documents to the cloud.

The Data user get the each result, the proof and the public verification key, they itself or others can verify the freshness, authenticity, and completeness of the search result even without decrypting them.

1.2 PROJECT OBJECTIVES

- Secure, immutable and decentralized EHRs with patient owing his/her own health data
- Single version of the truth verified by the health care provider
- Full medical history of a patient at one single point
- Increased transparency

1.3 ORGANIZATION OF CHAPTERS

Besides the introduction, the thesis is organized in other six chapters as follows:

Chapter 2, LITERATURE SURVEY: the review is made in the context of EHR systems with a particular attention on those implementations that assess the scalability and performances or their implementations. Most of the related work is on blockchain solutions, whereas a small part is on cloud solutions.

It will be possible to notice that only a small subset of the literature actually focuses on the analysis of the systems in mass crises scenarios.

Chapter 3, SOFTWARE AND HARDWARE REQUIREMENTS: this chapter discuss about the software and hardware required for the execution of the project.

Chapter 4, SOFTWARE DEVELOPMENT ANALYASIS: this chapter explains the assumptions and technical specifications of the project.

Chapter 5, PROJECT SYSTEM DESIGN: this chapter explains all the software development process with dfd, E-R diagrams, and UML diagrams clearly.

Chapter 6, PROJECT CODING: this chapter explains the design of the system, roles and responsibilities, as well as the requirements of a EHRs management solution based on block chain.

Chapter 7, PROJECT TESTING: this chapter explains various test cases to test the project working.

Chapter 8, OUTPUT SCREENS: explains a step by step process of the project execution.

Chapter 9, EXPERIMENTAL RESULTS: tests and results are shown and explained in this chapter. The results are analyzed in the context of the thesis project and followed by discussion on systems throughput and resiliency, as well as the approaches to testing and analysis.

Chapter 10, CONCLUSION AND FUTURE ENHANCEMENT: the chapter ends the project with a short summary of the main concepts mentioned in the thesis as well as the relevant results.

CHAPTER 2

LITERATURE SURVEY

A literature survey or a literature review in a project report is that section which shows the various analysis and research made in the field of your interest and the results already published, considering the various parameters of the project and the extent of the project. It is the most important part of our report as it gave us a direction in our research. It helped us set a goal for our analysis - thus giving us our problem statement.

2.1 SURVEY ON BACKGROUND

V Ramani, T Kumar, A Bracken, M Liyanage and M Ylianttila., proposes a blockchain [1] based secure and efficient data accessibility mechanism for the patient and the doctor in a given healthcare system and able to protect the privacy of the patients as well. The security analysis of scheme shows that it can resist to well-known attacks along with maintaining the integrity of the system. An Ethereum based implementation has used to verify the feasibility of the proposed system. [2] Azaria et al., describes novel decentralized registry management system called MedRec to manage EMRs, using blockchain technology. The results obtained are demonstrated an innovative approach to the management of medical records, providing interoperability and accessibility through a complete registry. [3] Dubovitskaya., describes framework based on the management and exchange of EMR(Electronic Medical Record) data for care of cancer patients based on blockchain. They implement the framework in a prototype that guarantees privacy, security, availability and detailed access control over EMR data which results significantly reduce the response time to share EMR and improve decision making for medical care and reduce the overall cost. [4] Magyar., present a study of how blockchain technology can help solve the problem of secures data storage and guarantee its availability at the same time in an EHR system. The new technology solves an essential problem of access to data without endangering personal privacy. In this related work [5] Xia., proposed a secure and scalable access control system for confidential information using blockchain.

The results show that the system succeeds where traditional methods of passwords access control, firewalls and intrusion detection systems fail. Badr, S., Gomaa, I., and Abd-Elrahman., describes a protocol to achieve the preservation of patient [6] privacy which is called PBE-DA applying the Blockchain concept in an eHealth platform. [7] A.Ibrahim., describes Attribute-Based Access Control refers to defining access policies and authorizing access to resources and data based on certain attribute. Role Based Access Control which provides access based on predefined roles organizations sets a complex set of Boolean logic and conditions to ensure that access is provided if and only if all the conditions have been satisfied. Secure solution for Electronic Health Record System. [8] Z.Ying., proposed a process policy is attached to the ciphertext. Policy preserving EHR system on the basis of CP-ABE is a algorithm which can hide the entire access policy as well as recover the hidden attributes from the access matrix. Inspite of efficient results this demands external key authorities with a public key infrastructure which is complex and requires expensive resources to obtain EHRs. [9] R Wu,H.Hu., proposed a broker-based access control mechanism is used to generate composite EHR data schema. Based on the schema, distributed HER instances from various approach is only reached EHRs sharing on a computer stimulation with virtual machines and not applicable to smart phones. In related work.

N Rifi., proposed system permits users to request data after their identities and [10] cryptographic keys are verified which is considered as proof of work. In this, communication and authentication protocols and algorithms between entities were not fully investigated. This system effects the confidentiality and integrity of data. [11] Dagher ., framework based on blockchain for secure, interoperable the framework called Ancile uses smart contracts in blockchain based on Ethereum. The results show a blockchain system that achieves a high level of decentralization while recognizing that some nodes must be of a higher authority.

[12] Da Conceição., propose the implementation of large- scale information architecture to access EHR based on intelligent contracts as mediators of information. The main contribution is the framing of data privacy and accessibility problems in medical care as well as proposal of an integrated architecture based on blockchain. [13] Dias ., The author present an approach to solve the problem of managing access control

in eHealth. The results in general show that the approach is viable, which offers several advantages when compared with existing systems. [14] Kaur et al., The author propose a platform based on Blockchain that can be used to store and manage EMR in a cloud environment. This study offers a summary of the framework, the internal work and protocols for management of heterogeneous health data. [15] Mikula & Jacobsen., The authors proposed a system for the identities and accesses management using blockchain. They propose a prototype based on an open source blockchain framework called Hyperledger Fabric to demonstrate the viability of the system. The results confirm that identity and access management can be achieved in a decentralized, efficient and secure manner.

2.2 CONCLUSIONS ON SURVEY

These works provide basic background information that the existing system doesn't maintain and process the data securely. It doesn't provide the more accurate search result. Incorrect and misleading of data will produce the wrong analysis result and Low search Efficiency. The search delay of the scheme is proportional to the size of the database. It is not suitable for the large scale databases.

CHAPTER 3

SOFTWARE AND HARDWARE REQUIREMENTS

All computer software needs certain hardware components or other software resources to be present on a computer. These prerequisites are known as (computer) system requirements and are often used as a guideline as opposed to an absolute rule. Most software defines two sets of system requirements: minimum and recommended. With increasing demand for higher processing power and resources in newer versions of software, system requirements tend to increase over time. Industry analysts suggest that this trend plays a bigger part in driving upgrades to existing computer systems than technological advancements.

Hardware requirements: The most common set of requirements defined by any operating system or software application is the physical computer resources, also known as hardware, a hardware requirements list is often accompanied by a hardware compatibility list (HCL), especially in case of operating systems. An HCL lists tested, compatible, and sometimes incompatible hardware devices for a particular operating system or application

3.1 SOFTWARE REQUIREMENTS

- O/S : Windows 7/8/10.
- Language : Java.
- IDE : Net Beans 8.2
- Data Base : MySQL

3.2 HARDWARE REQUIREMENTS

- System : Pentium IV 2.4 GHz
- Hard Disk : 160 GB
- Ram : 4GB
- Monitor : 15 VGA color
- Mouse : Logitech.
- Keyboard : 110 keys enhanced

CHAPTER 4

SOFTWARE DEVELOPMENT ANALYASIS

Software development is a process of writing and maintaining the source code, but in a broader sense, it includes all that is involved between the conception of the desired software through to the final manifestation of the software, sometimes in a planned and structured process. Therefore, software development may include research, new development, prototyping, modification, reuse, reengineering, maintenance, or any other activities that result in software products

4.1 OVERVIEW OF PROBLEM

Over decades, medical facilities have evolved elegantly. Still most of us are the witness of the fact that whenever we see a doctor, we need to put forward our medical file in front of him/her. Our file contains our previous prescriptions, medical reports, X-Rays, MRIs etc. It is a tedious task to keep record of all these. To overcome the security problems that are occurred in the existing system and effectively store the data over the cloud we introduce this system.

4.2 DEFINE THE PROBLEM

The Existing system doesn't maintain and process the data securely. It doesn't provides the more accurate search result. Incorrect and misleading of data will produce the wrong analysis result. Low search Efficiency. The search delay of the scheme is proportional to the size of the database. It is not suitable for the large scale database.

By using the Block chain technology we intend to create an E-health system of EHRs that have high security of data where the Data user get the each result, the proof and the public verification key, they itself or others can verify the freshness, authenticity, and completeness of the search result even without decrypting them and user friendly.

4.3 MODULES OVERVIEW

Healthcare Provider

- Load patient Records
- Key Generation
- Encrypt patient Records
- Block Creation
- Upload and Download Patient Records

Cloud Service Provider

- View Patient Records
- Grant or Revoke Permission

4.4 DEFINE THE MODULES

Healthcare Provider

Data Selection and Loading

In this process, the health provider chooses patient healthcare records for uploading and maintaining the dataset in the cloud.

Key Generation

The secret key is generated using cryptographic algorithm. This key is used for encrypting the dataset.

Encrypt Patient Records

The data is encrypted for secure maintenance. So, that the unauthorized person can’t be able to access the data that are presented in the cloud.

Block Creation

- Each block contain patient record and it's timestamp.
- A block chain, originally block is a growing list of records called blocks.

Upload and Download Patient Records

After creating the block, the healthcare provider will upload the records into the cloud. Suppose, if they want to retrieve an record from cloud, first the healthcare provider search the record. Based on the search it will show the results. After getting an approval and key from the cloud service provider the healthcare provider can download the data.

Cloud Service Provider

View Patient Records

Any health care provider can view the patient records by searching the patients name and after getting permission from the cloud CSP, they can view records.

Grant or Revoke Permission

Cloud provider is responsible for grant and revokes the permissions.

4.5 MODULE FUNCTIONALITY

Healthcare Provider module – He / She has the access to upload and download the health records, in the process of uploading they have to encrypt the records and while downloading they have to decrypt by requesting permission from the cloud.

Registration

It is a process of enrolling or being enrolled into the cloud. To utilize the cloud documents, every healthcare provider should enroll. During this process your basic information like email, contacts etc., are collected and stored in the Cloud. The cloud id for a particular user will get automatically generated during the registration.

Cloud ID

Every user should create a Cloud ID and use it to identify something with near certainty that the identifier does not duplicate one that has already been, or will be, created to identify something else. Information labeled with Cloud ID by independent parties can therefore be later combined into a single database, or transmitted on the same channel, without needing to resolve conflicts between identifiers

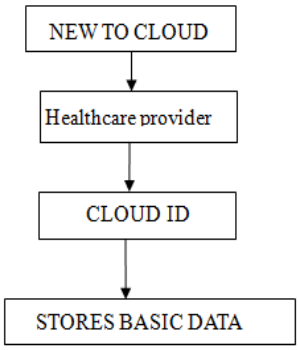


Fig 4.1 - flow of storing data in the cloud

Cloud Service Provider

The cloud service provider maintains all the patient records and also they can provide a permission to the user to access the data.

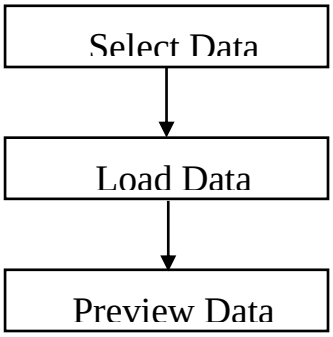


Fig 4.2 - flow of data in Cloud Service Provider

The Cloud Service Provider can view all the uploaded and downloaded documents in the Cloud. The CSP receives the document request from the Data User, verifies the authentication before granting permission. Then the CSP executes the query and returns the encrypted document according to the search token. And also returns an additional proof with the document, to verify the search result.

Public Verification Key

Public verification key is a security measure designed to make sure that your document outsourced in cloud doesn't get hacked. By verifying public key, the Data Owner and the Data User adding another layer of protection to the documents or files in the cloud by confirming each other's identities.

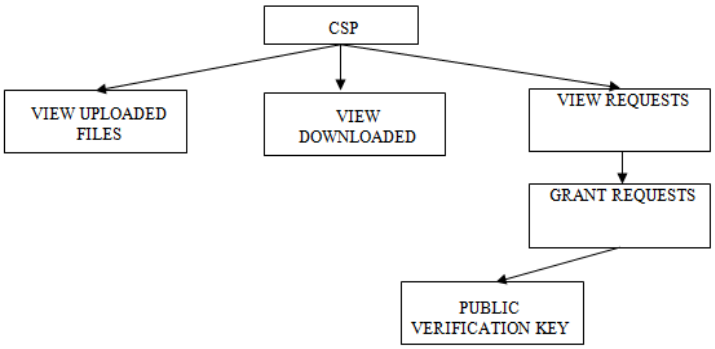


Fig 4.3 – Cloud Service Provider Flow

CHAPTER 5

PROJECT SYSTEM DESIGN

Systems design is the process of defining elements of a system like modules, architecture, components and their interfaces and data for a system based on the specified requirements. It is the process of defining, developing and designing systems which satisfies the specific needs and requirements of a business or organization.

A systemic approach is required for a coherent and well-running system. Bottom-Up or Top-Down approach is required to take into account all related variables of the system. A designer uses the modelling languages to express the information and knowledge in a structure of system that is defined by a consistent set of rules and definitions. The designs can be defined in graphical or textual modelling languages.

SYSTEM ARCHITECTURE

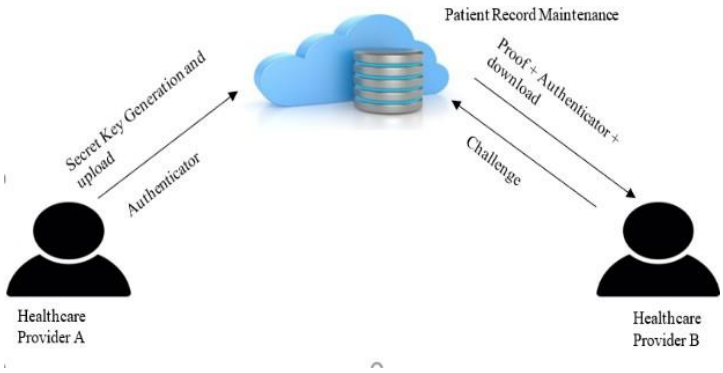


Fig 5.1 – System Architecture

5.1 DFDS IN CASE OF DATABASE PROJECTS

A data flow diagram shows the way information flows through a process or system. It includes data inputs and outputs, data stores, and the various sub processes the data moves through. DFDs are built using standardized symbols and notation to describe various entities and their relationships.

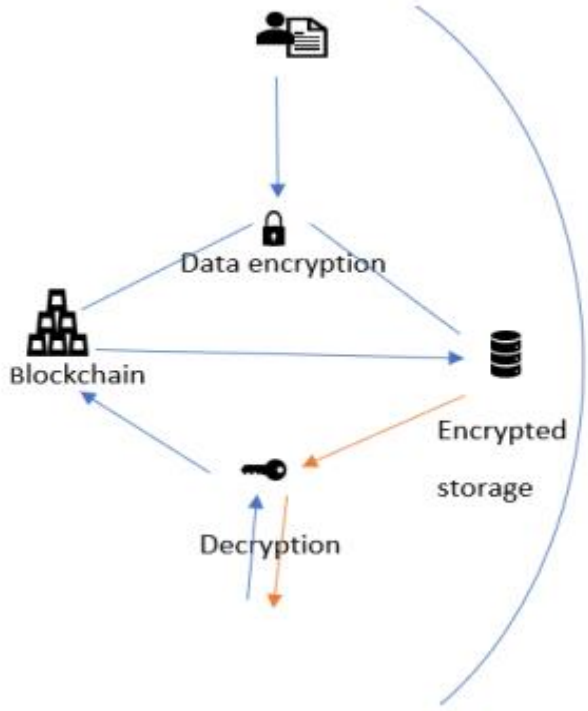


Fig 5.2 - Data Flow Diagram

Data flow diagrams visually represent systems and processes that would be hard to describe in a chunk of text. You can use these diagrams to map out an existing system and make it better or to plan out a new system for implementation. Visualizing each element makes it easy to identify inefficiencies and produce the best possible system.

5.2 E-R DIAGRAMS

ER diagrams are created based on three basic concepts: entities, attributes and relationships. **ER Diagrams** contain different symbols that use rectangles to represent entities, ovals to **define** attributes and diamond shapes to represent relationships. At first look, an **ER diagram** looks very similar to the flowchart.

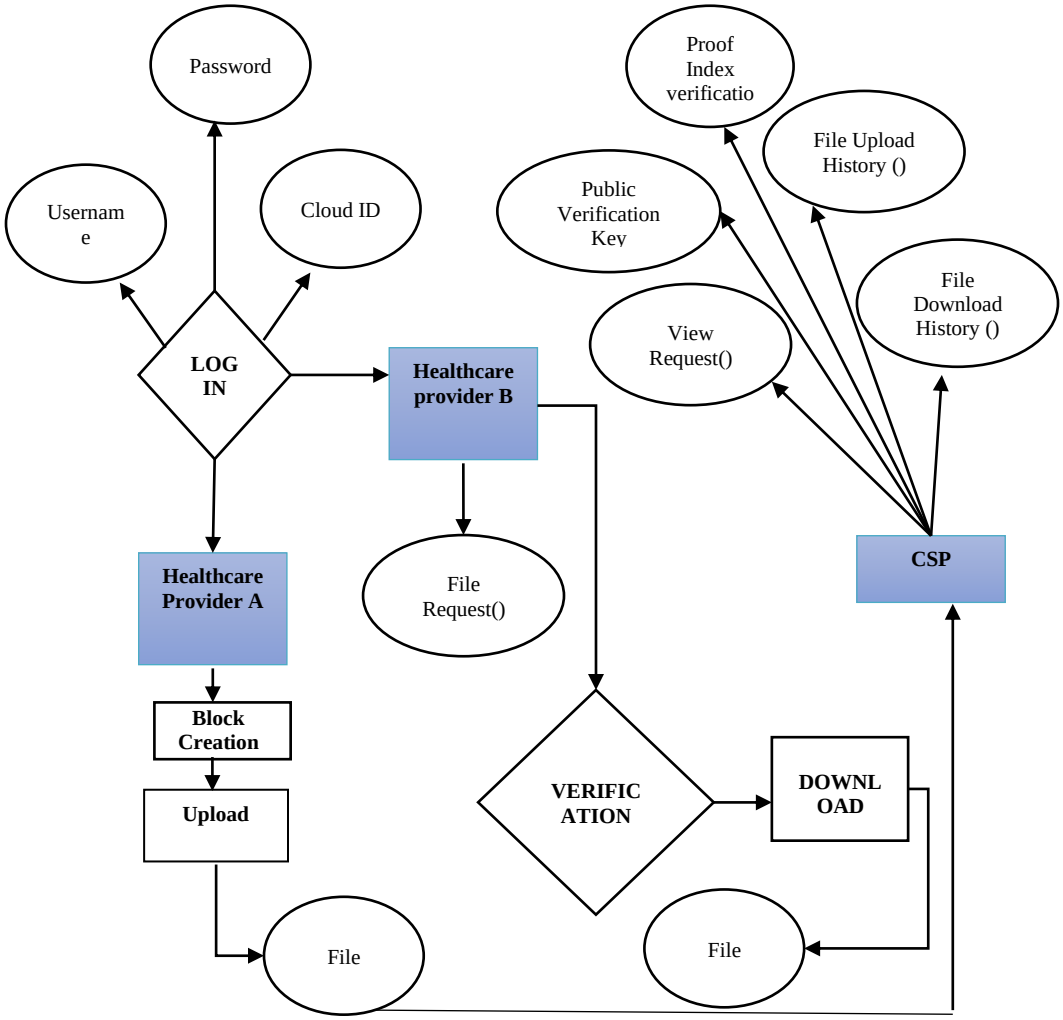


Fig 5.3 – E-R Diagram

5.3 UML DIAGRAMS

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.

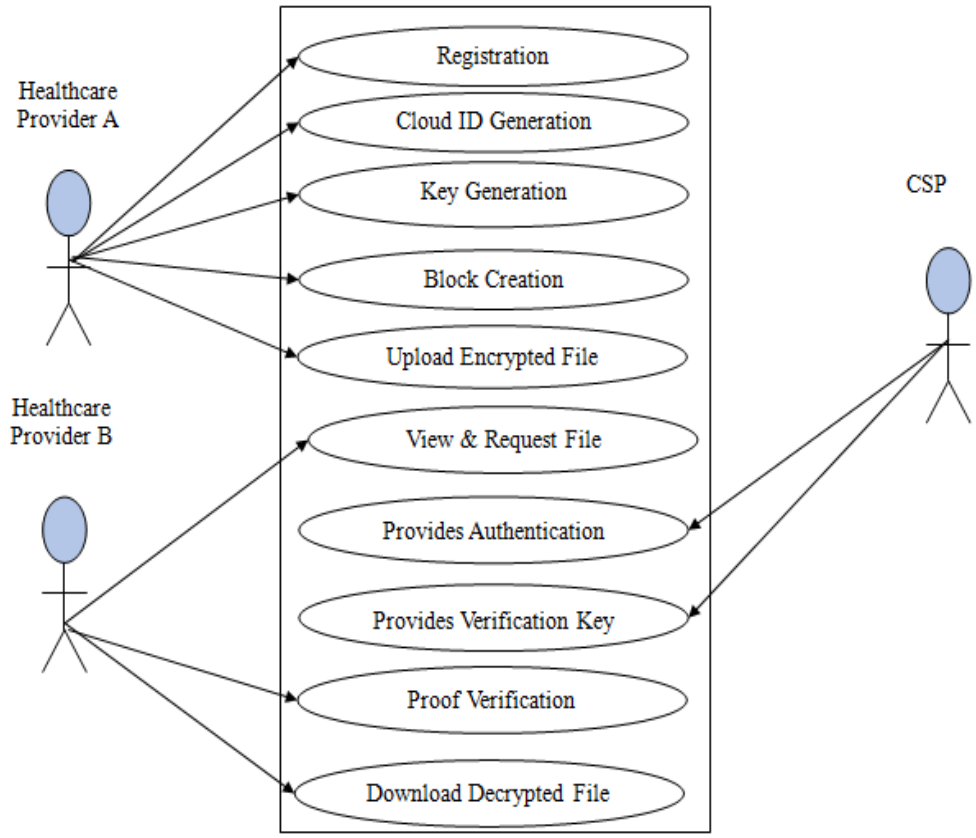


Fig 5.4 - Use Case Diagram

CLASS DIAGRAM

In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.

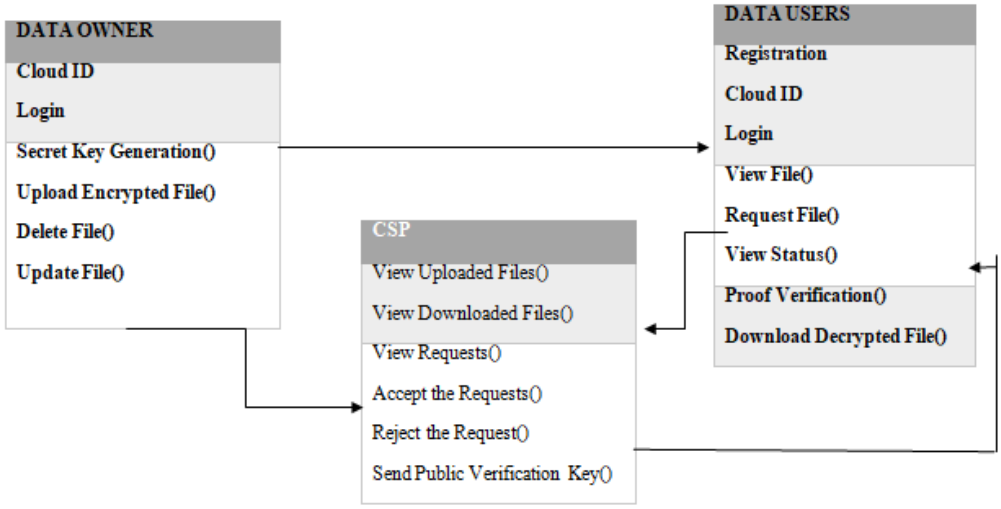


Fig 5.5 - Class Diagram

SEQUENCE DIAGRAM

A sequence diagram in Unified Modeling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.

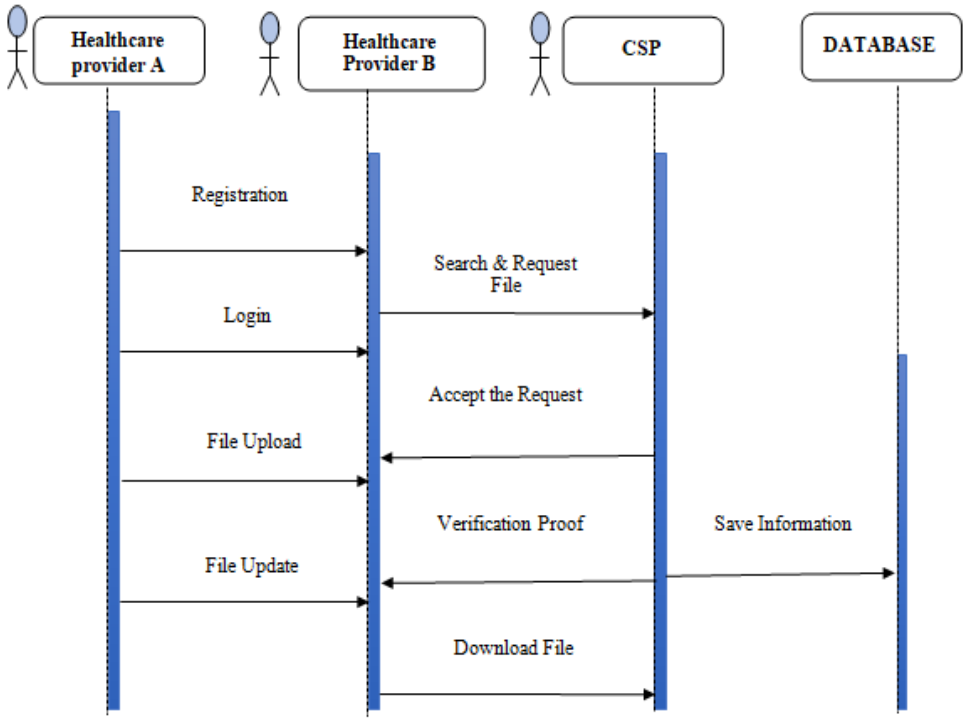


Fig 5.6 - Sequence Diagram

FLOW DIAGRAM

It is a collective term for a diagram representing a flow or set of dynamic relationships in a system. The term flow diagram is also used as a synonym for flowchart, and sometimes as a counterpart of the flowchart. Flow diagrams are used to structure and order a complex system, or to reveal the underlying structure of the elements and their interaction.

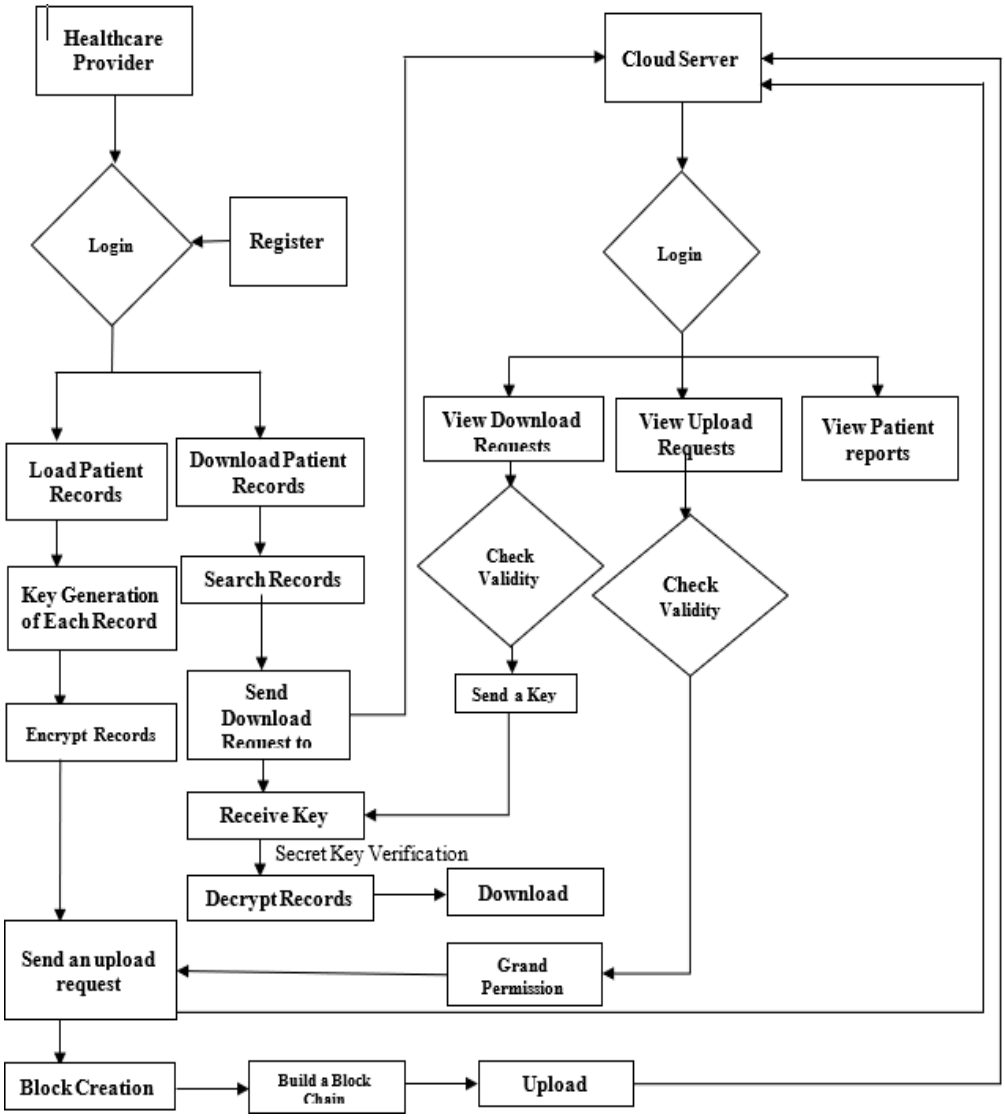


Fig 5.7 - Flow Diagram

CHAPTER 6

PROJECT CODING

6.1 CODE TEMPLATES

```
/* sample code */

package block_chain;

import static block_chain.upload.filepath;

import static block_chain.upload.nor;

import java.io.BufferedReader;

import java.io.File;

import java.io.FileReader;

import java.io.FileWriter;

import java.util.Date;

import java.sql.Timestamp;

import javax.swing.JOptionPane;

import java.awt.*;

import java.applet.*;

import java.io.IOException;

import java.util.logging.Level;

import java.util.logging.Logger;

import javax.swing.DefaultListModel;
```

```
public class block_Creation extends javax.swing.JFrame {

/* Creates new form block_Creation */

public block_Creation() {

initComponents();

this.setResizable(false);

this.setLocationRelativeTo(null);

}

/*  This method is called from within the constructor to initialize the form */

@SuppressWarnings("unchecked")

// <editor-fold defaultstate="collapsed" desc="Generated Code">

private void initComponents() {

jPanel13 = new javax.swing.JPanel();

jPanel14 = new javax.swing.JPanel();

jLabel10 = new javax.swing.JLabel();

jPanel15 = new javax.swing.JPanel();

jLabel11 = new javax.swing.JLabel();

jSeparator5 = new javax.swing.JSeparator();

jButton6 = new javax.swing.JButton();

jButton7 = new javax.swing.JButton();

jButton8 = new javax.swing.JButton();

jPanel1 = new javax.swing.JPanel();
```

```
jLabel13 = new javax.swing.JLabel();

jTextField1 = new javax.swing.JTextField();

jLabel14 = new javax.swing.JLabel();

jScrollPane1 = new javax.swing.JScrollPane();

jList1 = new javax.swing.JList<>();

jLabel12 = new javax.swing.JLabel();

setDefaultCloseOperation(javax.swing.WindowConstants.EXIT_ON_CLOSE);

jPanel13.setBackground(new java.awt.Color(153, 102, 255));

jPanel14.setBackground(new java.awt.Color(204, 204, 255));

jLabel10.setFont(new java.awt.Font("Cambria", 1, 24)); // NOI18N

jLabel10.setText("Blockchain: A Panacea for Healthcare Cloud-Based Data");

jPanel15.setBackground(new java.awt.Color(255, 255, 255));

jLabel11.setFont(new java.awt.Font("Cambria", 1, 18)); // NOI18N

jLabel11.setForeground(new java.awt.Color(102, 0, 255));

jLabel11.setText("Request Status and Upload");

jButton6.setBackground(new java.awt.Color(153, 102, 255));

jButton6.setFont(new java.awt.Font("Cambria", 1, 14)); // NOI18N

jButton6.setText("Upload");

jButton6.addActionListener(new java.awt.event.ActionListener() {

public void actionPerformed(java.awt.event.ActionEvent evt) {

jButton6ActionPerformed(evt);
```

```
} });
jButton7.setBackground(new java.awt.Color(153, 102, 255));
jButton7.setFont(new java.awt.Font("Cambria", 1, 14)); // NOI18N
jButton7.setText("Create and Bulid a Block Chain");
jButton7.addActionListener(new java.awt.event.ActionListener() {
public void actionPerformed(java.awt.event.ActionEvent evt) {
jButton7ActionPerformed(evt);
} });
jButton8.setBackground(new java.awt.Color(153, 102, 255));
jButton8.setFont(new java.awt.Font("Cambria", 1, 14)); // NOI18N
jButton8.setText("View Blocks");
jButton8.addActionListener(new java.awt.event.ActionListener() {
public void actionPerformed(java.awt.event.ActionEvent evt) {
jButton8ActionPerformed(evt);
} });
jPanel1.setBackground(new java.awt.Color(255, 255, 255));
jPanel1.setBorder(javax.swing.BorderFactory.createEtchedBorder());
jLabel13.setFont(new java.awt.Font("Cambria", 1, 14)); // NOI18N
jLabel13.setText("Block List");
jTextField1.setFont(new java.awt.Font("Cambria", 0, 14)); // NOI18N
jLabel14.setFont(new java.awt.Font("Cambria", 1, 14)); // NOI18N
```

```
jLabel14.setText("Number of Blocks");

jList1.setFont(new java.awt.Font("Cambria", 0, 14)); // NOI18N

jList1.addListSelectionListener(new javax.swing.event.ListSelectionListener() {

public void valueChanged(javax.swing.event.ListSelectionEvent evt) {

jList1ValueChanged(evt);

} });

jScrollPane1.setViewportView(jList1);

javax.swing.GroupLayout jPanel1Layout = new javax.swing.GroupLayout(jPanel1);

jPanel1.setLayout(jPanel1Layout);

jPanel1Layout.setHorizontalGroup(

jPanel1Layout.createParallelGroup(javax.swing.GroupLayout.Alignment.LEADING)

.addGroup(jPanel1Layout.createSequentialGroup()

.addGap(18, 18, 18)

.addGroup(jPanel1Layout.createParallelGroup(javax.swing.GroupLayout.Alignment.LE
ADING, false)

.addComponent(jLabel14)

.addComponent(jLabel13)

.addComponent(jTextField1)

.addComponent(jScrollPane1, javax.swing.GroupLayout.DEFAULT_SIZE, 504,
Short.MAX_VALUE))

.addContainerGap(javax.swing.GroupLayout.DEFAULT_SIZE,
Short.MAX_VALUE)));
```

```
jPanel1Layout.setVerticalGroup(

jPanel1Layout.createParallelGroup(javax.swing.GroupLayout.Alignment.LEADING)

.addGroup(jPanel1Layout.createSequentialGroup()

.addContainerGap()

.addComponent(jLabel14)

.addPreferredGap(javax.swing.LayoutStyle.ComponentPlacement.RELATED)

.addComponent(jTextField1, javax.swing.GroupLayout.PREFERRED_SIZE, 35,
javax.swing.GroupLayout.PREFERRED_SIZE)

.addGap(18, 18, 18)

.addComponent(jLabel13)

.addPreferredGap(javax.swing.LayoutStyle.ComponentPlacement.RELATED)

.addComponent(jScrollPane1, javax.swing.GroupLayout.PREFERRED_SIZE,
javax.swing.GroupLayout.DEFAULT_SIZE,
javax.swing.GroupLayout.PREFERRED_SIZE)

.addContainerGap(14, Short.MAX_VALUE)));

javax.swing.GroupLayout jPanel15Layout = new javax.swing.GroupLayout(jPanel15);

jPanel15.setLayout(jPanel15Layout);

jPanel15Layout.setHorizontalGroup(

jPanel15Layout.createParallelGroup(javax.swing.GroupLayout.Alignment.LEADING)

.addGroup(jPanel15Layout.createSequentialGroup()

.addGap(26, 26, 26)

.addGroup(jPanel15Layout.createParallelGroup(javax.swing.GroupLayout.Alignment.L
EADING, false)
```

.addComponent(jButton6, javax.swing.GroupLayout.DEFAULT_SIZE, 547, Short.MAX_VALUE)

.addComponent(jPanel1, javax.swing.GroupLayout.DEFAULT_SIZE, javax.swing.GroupLayout.DEFAULT_SIZE, Short.MAX_VALUE)

.addComponent(jLabel11)

.addComponent(jSeparator5, javax.swing.GroupLayout.DEFAULT_SIZE, 547, Short.MAX_VALUE)

.addComponent(jButton8, javax.swing.GroupLayout.DEFAULT_SIZE, javax.swing.GroupLayout.DEFAULT_SIZE, Short.MAX_VALUE)

.addComponent(jButton7, javax.swing.GroupLayout.DEFAULT_SIZE, 547, Short.MAX_VALUE))

.addContainerGap(30, Short.MAX_VALUE)));

jPanel15Layout.setVerticalGroup(

jPanel15Layout.createParallelGroup(javax.swing.GroupLayout.Alignment.LEADING)

.addGroup(jPanel15Layout.createSequentialGroup()

.addGap(25, 25, 25)

.addComponent(jLabel11)

.addPreferredGap(javax.swing.LayoutStyle.ComponentPlacement.RELATED)

.addComponent(jSeparator5, javax.swing.GroupLayout.PREFERRED_SIZE, 10, javax.swing.GroupLayout.PREFERRED_SIZE)

.addGap(18, 18, 18)

.addComponent(jButton7, javax.swing.GroupLayout.PREFERRED_SIZE, 36, javax.swing.GroupLayout.PREFERRED_SIZE)

.addGap(18, 18, 18)

.addComponent(jButton8, javax.swing.GroupLayout.PREFERRED_SIZE, 36, javax.swing.GroupLayout.PREFERRED_SIZE)

.addGap(18, 18, 18)

.addComponent(jPanel1, javax.swing.GroupLayout.PREFERRED_SIZE, javax.swing.GroupLayout.DEFAULT_SIZE, javax.swing.GroupLayout.PREFERRED_SIZE)

.addGap(18, 18, Short.MAX_VALUE)

.addComponent(jButton6, javax.swing.GroupLayout.PREFERRED_SIZE, 36, javax.swing.GroupLayout.PREFERRED_SIZE)

.addGap(21, 21, 21)));

jLabel12.setFont(new java.awt.Font("Cambria", 1, 24)); // NOI18N

jLabel12.setText(" Security and Privacy?");

javax.swing.GroupLayout jPanel14Layout = new javax.swing.GroupLayout(jPanel14);

jPanel14.setLayout(jPanel14Layout);

jPanel14Layout.setHorizontalGroup(

jPanel14Layout.createParallelGroup(javax.swing.GroupLayout.Alignment.LEADING)

.addGroup(jPanel14Layout.createSequentialGroup()

.addGroup(jPanel14Layout.createParallelGroup(javax.swing.GroupLayout.Alignment.TRAILING)

.addComponent(jPanel15, javax.swing.GroupLayout.PREFERRED_SIZE, javax.swing.GroupLayout.DEFAULT_SIZE, javax.swing.GroupLayout.PREFERRED_SIZE)

.addGroup(jPanel14Layout.createParallelGroup(javax.swing.GroupLayout.Alignment.TRAILING)

```
.addGroup(jPanel14Layout.createSequentialGroup()

.addContainerGap()

.addComponent(jLabel12, javax.swing.GroupLayout.PREFERRED_SIZE, 248,
javax.swing.GroupLayout.PREFERRED_SIZE))

.addGroup(javax.swing.GroupLayout.Alignment.LEADING,
jPanel14Layout.createSequentialGroup()

.addGap(43, 43, 43)

.addComponent(jLabel10))))

.addContainerGap(43, Short.MAX_VALUE)));

jPanel14Layout.setVerticalGroup(

jPanel14Layout.createParallelGroup(javax.swing.GroupLayout.Alignment.LEADING)

.addGroup(jPanel14Layout.createSequentialGroup()

.addGap(26, 26, 26)

.addComponent(jLabel10)

.addGap(21, 21, 21)

.addComponent(jLabel12)

.addPreferredGap(javax.swing.LayoutStyle.ComponentPlacement.RELATED, 20,
Short.MAX_VALUE)

.addComponent(jPanel15, javax.swing.GroupLayout.PREFERRED_SIZE,
javax.swing.GroupLayout.DEFAULT_SIZE,
javax.swing.GroupLayout.PREFERRED_SIZE)

.addGap(49, 49, 49)));

javax.swing.GroupLayout jPanel13Layout = new javax.swing.GroupLayout(jPanel13);
```

```
jPanel13.setLayout(jPanel13Layout);

jPanel13Layout.setHorizontalGroup(

jPanel13Layout.createParallelGroup(javax.swing.GroupLayout.Alignment.LEADING)

.addGroup(jPanel13Layout.createSequentialGroup()

.addGap(45, 45, 45)

.addComponent(jPanel14, javax.swing.GroupLayout.PREFERRED_SIZE,
javax.swing.GroupLayout.DEFAULT_SIZE,
javax.swing.GroupLayout.PREFERRED_SIZE)

.addContainerGap(43, Short.MAX_VALUE)));

jPanel13Layout.setVerticalGroup(

jPanel13Layout.createParallelGroup(javax.swing.GroupLayout.Alignment.LEADING)

.addGroup(jPanel13Layout.createSequentialGroup()

.addGap(49, 49, 49)

.addComponent(jPanel14, javax.swing.GroupLayout.PREFERRED_SIZE,
javax.swing.GroupLayout.DEFAULT_SIZE,
javax.swing.GroupLayout.PREFERRED_SIZE)

.addContainerGap(44, Short.MAX_VALUE)) );

javax.swing.GroupLayout layout = new javax.swing.GroupLayout(getContentPane());

getContentPane().setLayout(layout);

layout.setHorizontalGroup(

layout.createParallelGroup(javax.swing.GroupLayout.Alignment.LEADING)

.addComponent(jPanel13, javax.swing.GroupLayout.DEFAULT_SIZE,
javax.swing.GroupLayout.DEFAULT_SIZE, Short.MAX_VALUE));
```

```
layout.setVerticalGroup(
layout.createParallelGroup(javax.swing.GroupLayout.Alignment.LEADING)
.addComponent(jPanel13, javax.swing.GroupLayout.DEFAULT_SIZE,
javax.swing.GroupLayout.DEFAULT_SIZE, Short.MAX_VALUE));
pack();
}// </editor-fold>
private void jButton6ActionPerformed(java.awt.event.ActionEvent evt) {
// TODO add your handling code here:
JOptionPane.showMessageDialog(null, "Uploaded Successfully");
}
private void jButton7ActionPerformed(java.awt.event.ActionEvent evt) {
// TODO add your handling code here:
int i=1;
try{
BufferedReader br=new BufferedReader(new FileReader(filepath));
String s1="",s2="";
br.readLine();
while((s1=br.readLine())!=null){
System.out.println(s1+"\n");
s2+=s1+"\n\n";
File f=new File("./Cloud Me/Blocks/"+i);
f.mkdir();
```

```
File ff=new File("./Cloud Me/Blocks/"+i+"/"+i+".txt");

ff.delete();

FileWriter fw1=new FileWriter("./Cloud Me/Blocks/"+i+"/"+i+".txt",true);

fw1.write(s1);

fw1.write("\r\n");

fw1.close();

File ff1=new File("./Cloud Me/Blocks/"+i+"/"+"Time Stamp"+".txt");

ff1.delete();

FileWriter fw2=new FileWriter("./Cloud Me/Blocks/"+i+"/"+"Time Stamp"+".txt",true);

Date date= new Date();

long time = date.getTime();

System.out.println("Time in Milliseconds: " + time);

Timestamp ts = new Timestamp(time);

System.out.println("Current Time Stamp: " + ts);

fw2.write("Time in Milliseconds: " + time +"\n" + "Current Time Stamp: " + ts);

fw2.write("\r\n");

fw2.close();

i++; }

br.close();

JOptionPane.showMessageDialog(null, "Blocks Created Successfully");

}
```

```
catch (Exception ex)
{}
}
private void jButton8ActionPerformed(java.awt.event.ActionEvent evt) {
// TODO add your handling code here:
jTextField1.setText(""+nor);
DefaultListModel lmodel = new DefaultListModel();
for(int i=1;i<=nor;i++)
{
lmodel.addElement(""+i);
}
jList1.setModel(lmodel);
}
private void jList1ValueChanged(javax.swing.event.ListSelectionEvent evt) {
// TODO add your handling code here:
Desktop desktop = Desktop.getDesktop();
File dirToOpen = null;
try {
dirToOpen = new File("Cloud Me\\Blocks\\"+jList1.getSelectedValue());
desktop.open(dirToOpen);
} catch (IllegalArgumentException iae) {
```

```
System.out.println("File Not Found");

} catch (IOException ex) {

}}

/* @param args the command line arguments */

public static void main(String args[]) {

try {

for (javax.swing.UIManager.LookAndFeelInfo info :
javax.swing.UIManager.getInstalledLookAndFeels()) {

if ("Nimbus".equals(info.getName())) {

javax.swing.UIManager.setLookAndFeel(info.getClassName());

break;

} }

} catch (ClassNotFoundException ex) {

java.util.logging.Logger.getLogger(block_Creation.class.getName()).log(java.util.loggin
g.Level.SEVERE, null, ex);

} catch (InstantiationException ex) {

java.util.logging.Logger.getLogger(block_Creation.class.getName()).log(java.util.loggin
g.Level.SEVERE, null, ex);

} catch (IllegalAccessException ex) {

java.util.logging.Logger.getLogger(block_Creation.class.getName()).log(java.util.loggin
g.Level.SEVERE, null, ex);

} catch (javax.swing.UnsupportedLookAndFeelException ex) {
```

```
java.util.logging.Logger.getLogger(block_Creation.class.getName()).log(java.util.loggin
g.Level.SEVERE, null, ex);
}
/* Create and display the form */
java.awt.EventQueue.invokeLater(new Runnable() {
public void run() {
new block_Creation().setVisible(true);
}
});
}
// Variables declaration - do not modify
private javax.swing.JButton jButton6;
private javax.swing.JButton jButton7;
private javax.swing.JButton jButton8;
private javax.swing.JLabel jLabel10;
private javax.swing.JLabel jLabel11;
private javax.swing.JLabel jLabel12;
private javax.swing.JLabel jLabel13;
private javax.swing.JLabel jLabel14;
private javax.swing.JList<String> jList1;
private javax.swing.JPanel jPanel1;
private javax.swing.JPanel jPanel13;
```

```
private javax.swing.JPanel jPanel14;

private javax.swing.JPanel jPanel15;

private javax.swing.JScrollPane jScrollPane1;

private javax.swing.JSeparator jSeparator5;

private javax.swing.JTextField jTextField1;

// End of variables declaration

}
```

6.2 OUTLINE FOR VARIOUS FILES

We used Java programming to implement our project. We used Swing which is used to create window-based applications. It is built on the top of Abstract Windowing Toolkit API and entirely written in java. Our code consists of various modules that we have used. Our project modules are – health care provider and cloud provider. We also used various swing features and methods that provides to facilitate coding in an easier and quicker way

6.3 CLASS WITH FUNCTIONALITY

There are multiple classes in our code, some of which are:

Upload.java: This handles the uploading of dataset.

Log.java: This handles the backend linking of the entire code to the MySQL server in order to facilitate the final output.

Encryptrecord.java: This handles the encryption of records using AES algorithm.

Blockchain.java: This handles the creation of blocks and storing that in the cloud.

Finaldownload.java: This handles the decryption of records and downloading them.

6.4 METHODS INPUT AND OUTPUT PARAMETERS

We implemented multiple methods, few of which are:

UIManager.setLookAndFeel()

DriverManager.getConnection()

getSecretEncryptionKey()

encryptText()

decryptText()

First method UIManager.setLookAndFeel() is used to manage current look and feel, set of available to look and feel. Method DriverManager.getConnection() is used to set connection to the data base. Method getSecretEncryptionKey() is used to generate a key with the help of Cryptographic Algorithm. Method encryptText() and decryptText() are used to encrypt and decrypt the records with the help of AES algorithm. We also used some java swing API such as JButton, JTextField, JTextArea, JRadioButton, JCheckbox, JMenu, JColorChooser etc.

CHAPTER 7

PROJECT TESTING

7.1 VARIOUS TEST CASES

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

TYPES OF TESTS

Unit testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

Integration testing

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the

combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

Functional test

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items:

Valid Input : identified classes of valid input must be accepted.

Invalid Input : identified classes of invalid input must be rejected.

Functions : identified functions must be exercised.

Output : identified classes of application outputs must be exercised.

Systems/Procedures : interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

System Test

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

Unit Testing

Unit testing is usually conducted as part of a combined code and unit test phase of the software lifecycle, although it is not uncommon for coding and unit testing to be conducted as two distinct phases.

Test strategy and approach

Field testing will be performed manually and functional tests will be written in detail.

Test objectives

- All field entries must work properly.
- Pages must be activated from the identified link.
- The entry screen, messages and responses must not be delayed.

Features to be tested

- Verify that the entries are of the correct format
- No duplicate entries should be allowed
- All links should take the user to the correct page.

Integration Testing

Software integration testing is the incremental integration testing of two or more integrated software components on a single platform to produce failures caused by interface defects.

The task of the integration test is to check that components or software applications, e.g. components in a software system or – one step up – software applications at the company level – interact without error.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

Acceptance Testing

User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

7.2 BLACK BOX

Black box testing is a technique of software testing which examines the functionality of software without peering into its internal structure or coding. The primary source of black box testing is a specification of requirements that is stated by the customer.

In this method, tester selects a function and gives input value to examine its functionality, and checks whether the function is giving expected output or not. If the function produces correct output, then it is passed in testing, otherwise failed. The test team reports the result to the development team and then tests the next function. After completing testing of all functions if there are severe problems, then it is given back to the development team for correction.

Generic steps of black box testing

- The black box test is based on the specification of requirements, so it is examined in the beginning.
- In the second step, the tester creates a positive test scenario and an adverse test scenario by selecting valid and invalid input values to check that the software is processing them correctly or incorrectly.
- In the third step, the tester develops various test cases such as decision table, all pairs test, equivalent division, error estimation, cause-effect graph, etc.
- The fourth phase includes the execution of all test cases.
- In the fifth step, the tester compares the expected output against the actual output.
- In the sixth and final step, if there is any flaw in the software, then it is cured and tested again

7.3 WHITE BOX TESTING

The box testing approach of software testing consists of black box testing and white box testing. We are discussing here white box testing which also known as glass box is testing, structural testing, clear box testing, open box testing and transparent box testing. It tests internal coding and infrastructure of a software focus on checking of

predefined inputs against expected and desired outputs. It is based on inner workings of an application and revolves around internal structure testing. In this type of testing programming skills are required to design test cases. The primary goal of white box testing is to focus on the flow of inputs and outputs through the software and strengthening the security of the software.

The term 'white box' is used because of the internal perspective of the system. The clear box or white box or transparent box name denote the ability to see through the software's outer shell into its inner workings.

Developers do white box testing. In this, the developer will test every line of the code of the program. The developers perform the White-box testing and then send the application or the software to the testing team, where they will perform the black box testing and verify the application along with the requirements and identify the bugs and sends it to the developer. The developer fixes the bugs and does one round of white box testing and sends it to the testing team. Here, fixing the bugs implies that the bug is deleted, and the particular feature is working fine on the application.

CHAPTER 8

OUTPUT SCREENS

8.1 USER INTERFACES

Fig 8.1 – Welcome page

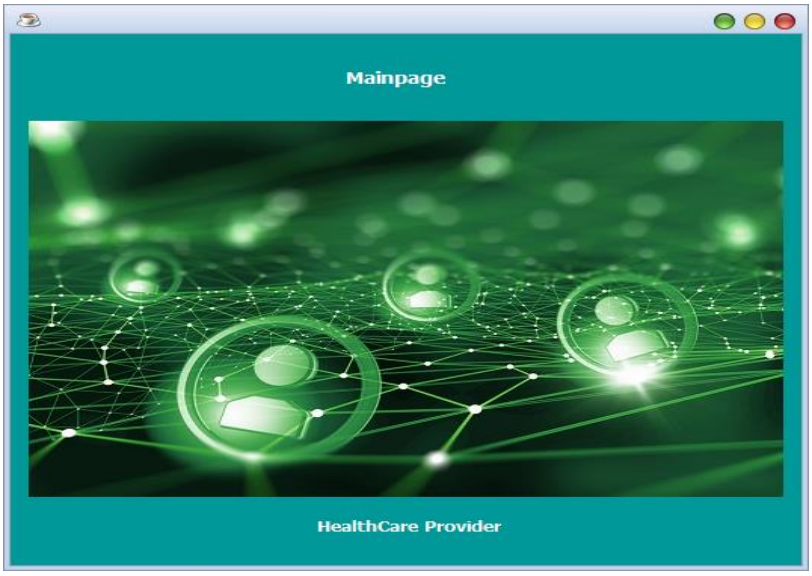


Fig 8.2 – Main Page

8.2 OUTPUT SCREENS

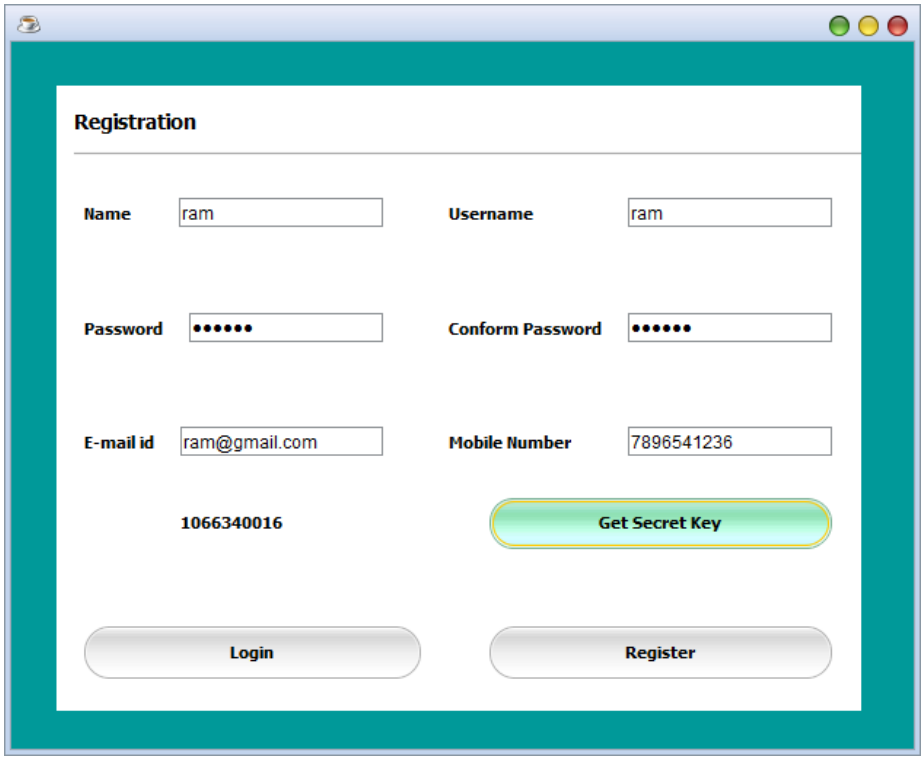


Fig 8.3 – Registration page

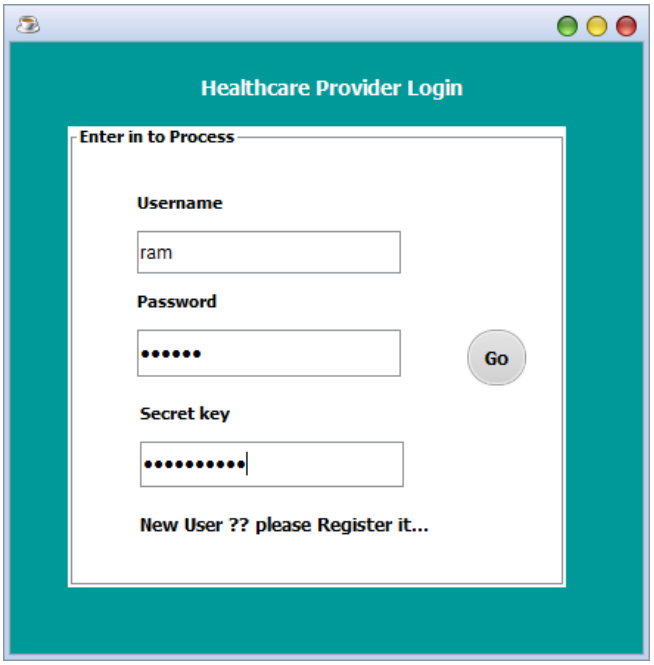


Fig 8.4 – HealthCare Provider Login Page

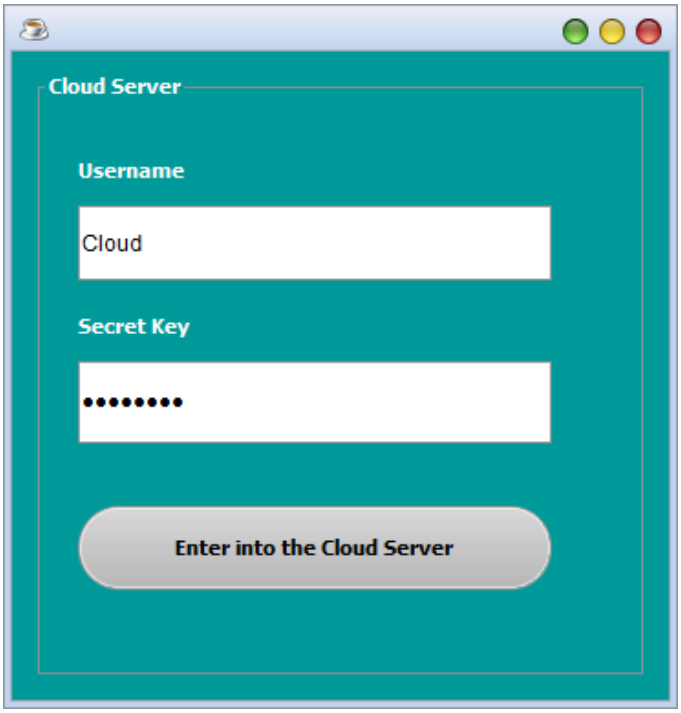


Fig 8.5 – Cloud Service Provider Login Page

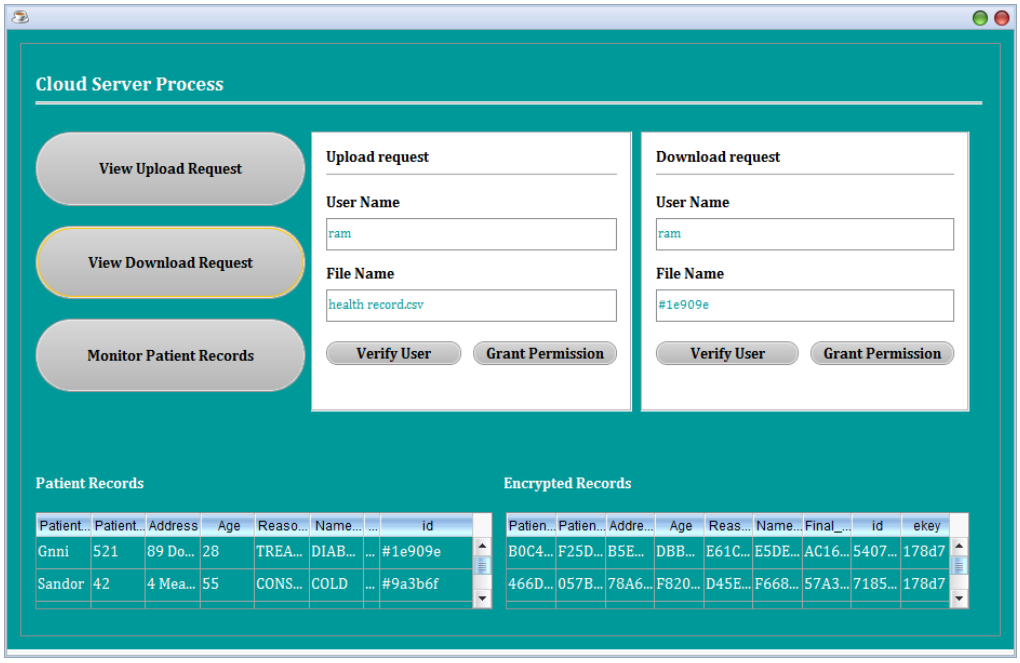


Fig 8.6 – CSP Home Page

CHAPTER 9

EXPERIMENTAL RESULTS

The health care provider first login's with his/her credentials

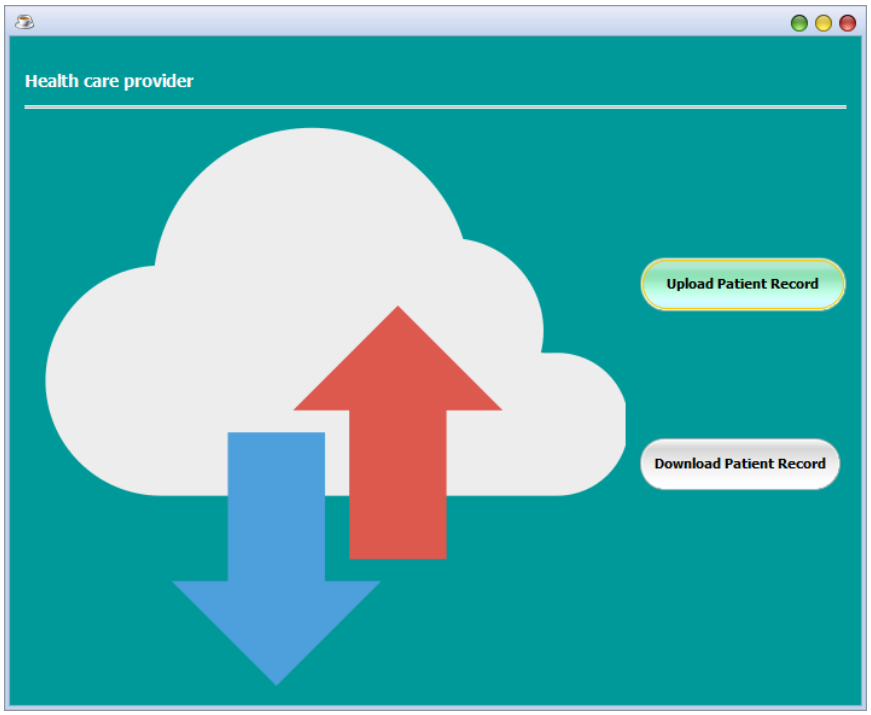


Fig 9.1 – Uploading and Download Request Page

Health Care Provider can now upload and download Records from this page.

Now let's see the uploading process

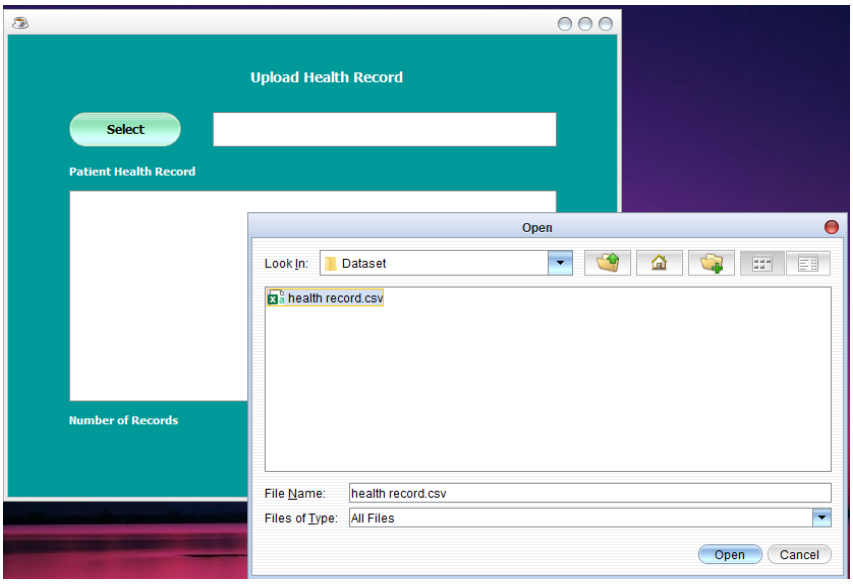


Fig 9.2 - Uploaded Data Set

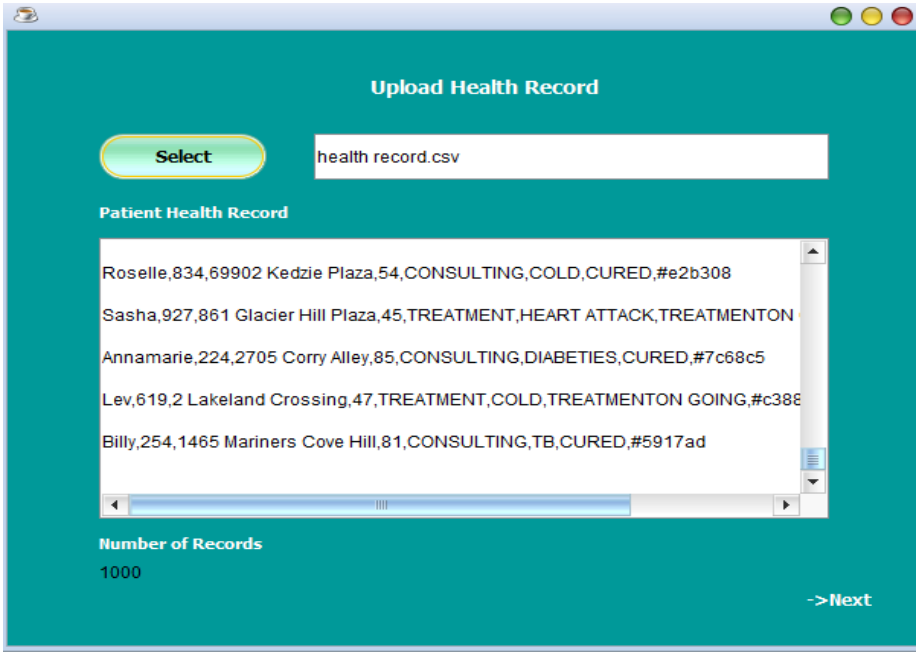


Fig 9.3 - Uploaded Data Set

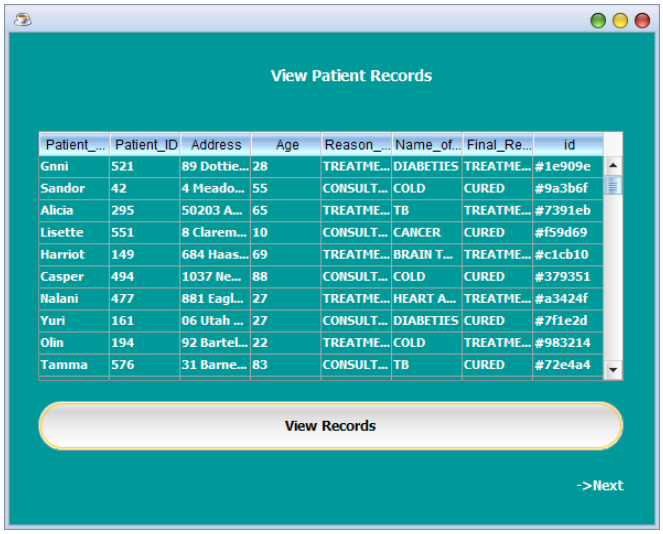


Fig 9.4 - View Patient Records.

Now the Health Care Provider can view the details of patient.

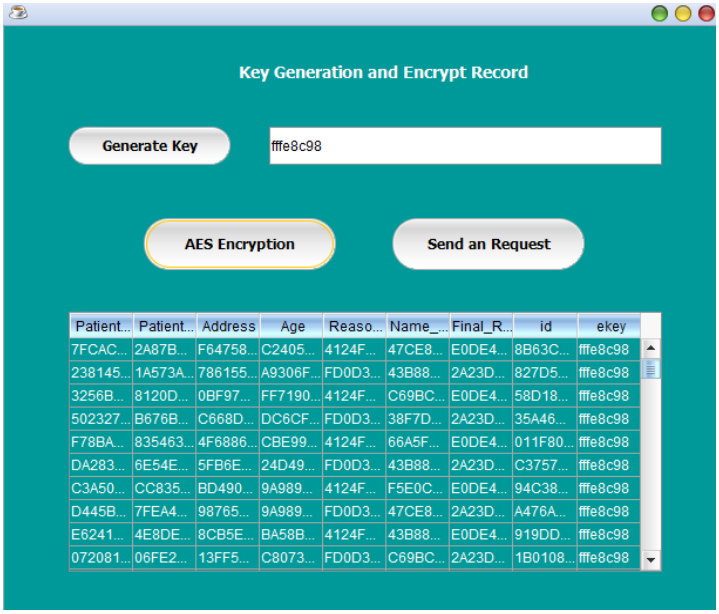


Fig 9.5 - Key Generation and Encrypt Records

Encrypting Records Using Key, which is generated by Cryptographic Algorithm, the Data is now Encrypted using AES Algorithm.

Now the CSP will Verify user and grant permission for uploading data

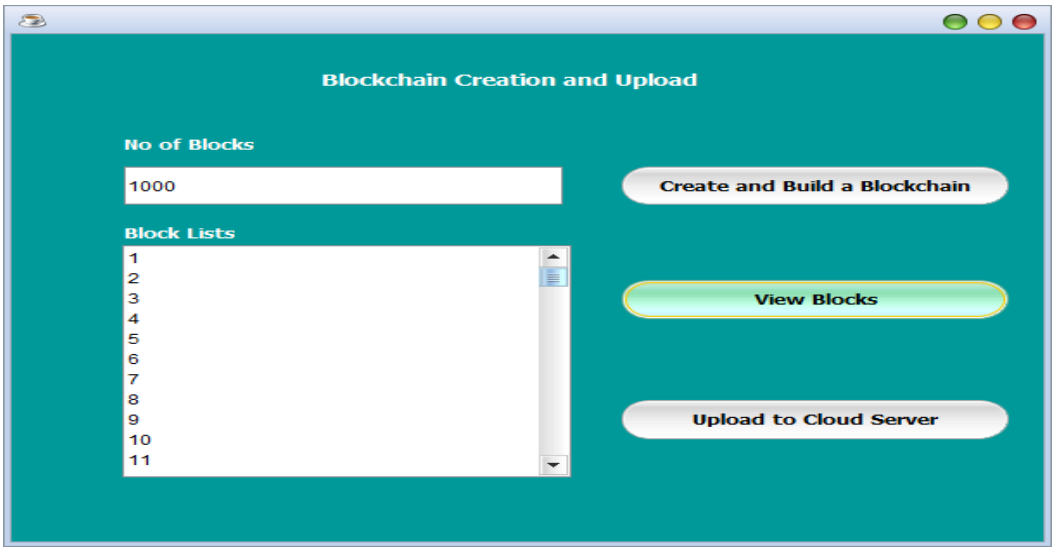


Fig 9.6 - Blockchain Creation and upload.

Now the Uploaded data is converted in to Blocks and upload it in Cloud.

When any other health care provider wants to view any persons records, then he/she has to enter there credentials' in the login page and search for the persons record.

Now the CSP will verify user and grant permission for downloading data.

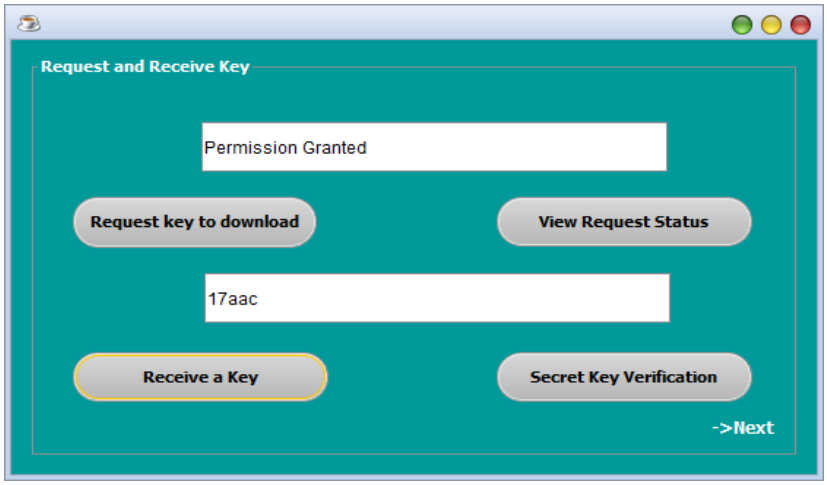


Fig 9.7 - Request and Receive Key.

After permission is granted a secret key will be received and Verified.

Now they can view and Download the record.

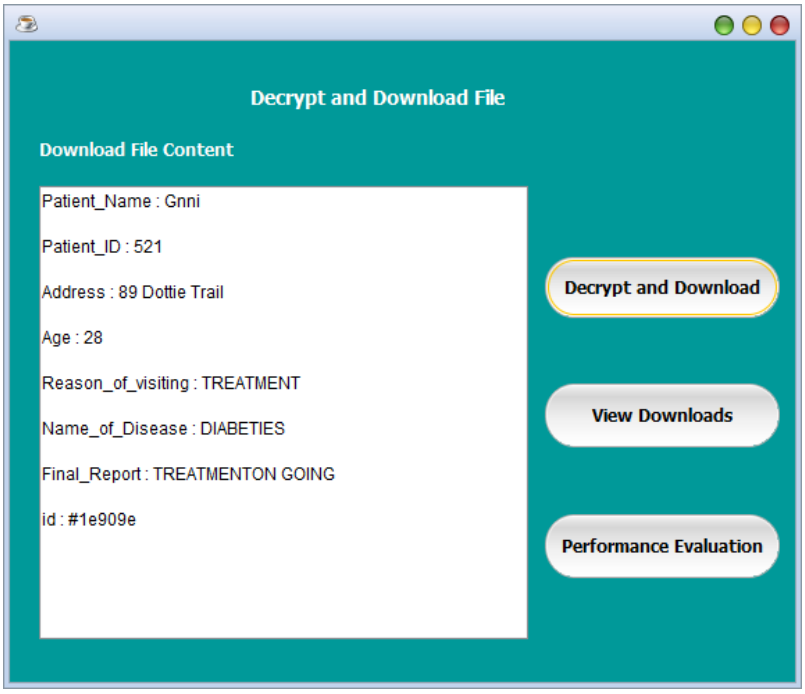


Fig 9.8 - Decrypt and Download File.

After Verification of Secret key, we can decrypt the record and download it.

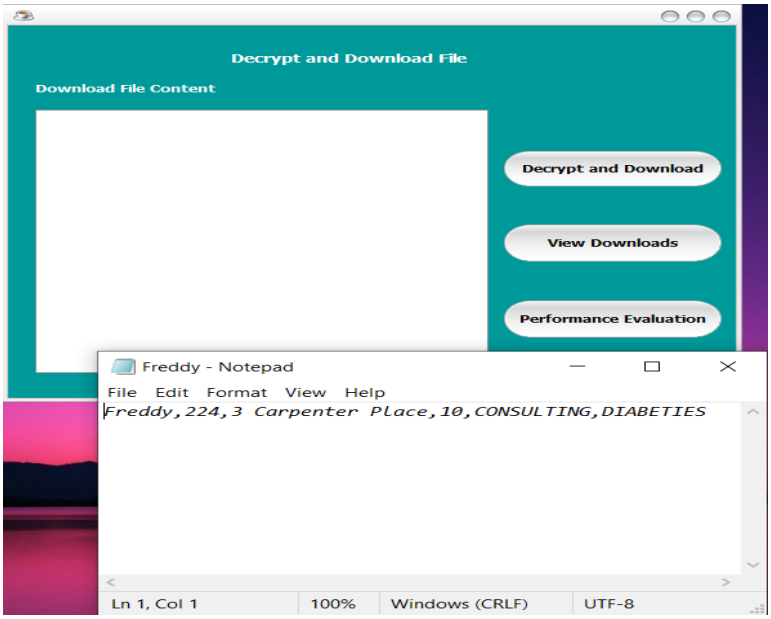


Fig 9.9 – Final output

This is the decrypted File which will be downloaded in our system.

CHAPTER 10

CONCLUSION AND FUTURE ENHANCEMENT

In the above project, the existing challenges of the EHR system are identified and the solution is provided to address these issues via a real prototype implementation. Patient records are maintained more securely and they'll be easily accessible by any health care providers. By building blockchain, it provides efficient search result verification, while preventing data freshness attacks and data integrity attacks in the cloud. The blockchain-enabled resolution may be a step towards the economic management of e-health records on clouds, which is promising in several health care applications

Data Classification based on Security: A cloud computing data center can store data from various users. To provide the level of security based on the importance of data, classification of data can be done. This classification scheme should consider various aspects like access frequency, update frequency and access by various entities etc. based on the type of data. Once the data is classified and tagged, then level of security associated with this specific tagged data element can be applied. Level of security includes confidentiality, encryption, integrity and storage etc. that are selected based on the type of data.

CHAPTER 11

REFERENCES

[1]V. Ramani, T. Kumar, A. Bracken, M. Liyanage, and M. Ylianttila, ‘‘Secure and effificient data accessibility in blockchain based healthcare systems,’’ in *Proc. GLOBECOM*, Dec. 2018

[2] Azaria, A., Ekblaw, A., Vieira, T., and Lippman, A., MedRec: Using blockchain for medical data access and permission management. Proc - 2016 2nd Int Conf Open Big Data, OBD. 25–30, 2016.

[3] Dubovitskaya, A., Xu, Z., Ryu, S., Schumacher, M., Wang, F., Secure and Trustable Electronic Medical Records Sharing using Blockchain. Proc AMIA Annu Symp. 650, 2017

[4] Magyar G. Blockchain: solving the privacy and research availability tradeoff for EHR data: A new disruptive technology in health data management. 2017 IEEE 30thJubil Neumann Colloq. 135– 140, 2017.

[5] Xia, Q., Sifah, E. B., Smahi, A., Amofa, S., and Zhang,X., BBDS: Blockchain-based data sharing for electronic medical records in cloud environments. Inform 8(2):44, 2017.

[6] Badr, S., Gomaa, I., and Abd-Elrahman, E., Multi-tier Blockchain Framework for IoT-EHRs Systems. Procedia Comput Sci. 141: 159–166, 2018.

[7]A. Ibrahim, B. Mahmood, and M. Singhal, ‘‘A secure framework for sharing electronic health records over clouds,’’ in *Proc. IEEE Serious Games Appl. Health*, May 2016.

[8].Z. Ying, L. Wei, Q. Li, X. Liu, and J. Cui, ‘‘A lightweight policy preserving EHR sharing scheme in the cloud,’’2018.

[9] R. Wu, G.-J. Ahn, and H. Hu, ‘‘Secure sharing of electronic health records in clouds,’’ in *Proc. 8th Int. Conf. Collaborative Comput., Netw., Appl. Worksharing (CollaborateCom)*, Oct. 2012,

[10]N. Rififi, E. Rachkidi, N. Agoulmine, and N. C. Taher, ‘‘Towards using blockchain technology for eHealth data access management,’’ in *Proc. IEEE 4th Int. Conf. Adv.Biomed. Eng.*, Oct. 2017

[11] Dagher, G. G., Mohler, J., Milojkovic, M., and Marella, P. B., Ancile: Privacypreserving framework for access control and inter-operability of electronic healthrecordsusingblockchain technology. Sustain Cities, 2018.

[12] da Conceição, A. F., da Silva, F. S. C., Rocha, V., Locoro, A., Barguil, J. M., Eletronic Health Records using Blockchain Technology.. 2018.

[13] Dias, J. P., Reis, L., Ferreira H. S., Martins, Â., Blockchain for Access Control in eHealth Scenarios. 2018.

[14] Kaur, H., Alam, M. A., Jameel, R., Mourya, A. K., and Proposed Solution, C.V.A.,Future Direction for Blockchain- Based Heterogeneous Medicare Data in Cloud Environment.
J. Med. Syst, 2018.

[15]Mikula, T., and Jacobsen, R. H., Identity and Access Management with Blockchain in Electronic Healthcare Records. 2018 21st Euromicro Conf Digit Syst Des. 699–706, 2018.

[16]J. Li, J. Wu, and L. Chen, “Block-secure: Blockchain based scheme for secure p2p cloud storage,” Information Sciences, vol. 465, pp. 219–231, Oct. 2018.

[17] Wu, A.; Zhang, Y.; Zheng, X.; Guo, R.; Zhao, Q.; Zheng,
D. Effificient and privacy-preserving traceable attribute- based encryption in blockchain. *Ann. Telecommun.* **2019**, *74*, 401–411.

[18]Huang, Q.; Ma, Z.; Yang, Y.; Niu, X.; Fu, J. Attribute based DRM scheme with dynamic usage control in cloud computing. *China Commun.* **2014**, *11*, 50–63.

CHAPTER 12

PUBLICATIONS

- JOURNAL (UGC APPROVED JOURNAL)

- CONFERENCE (Online Mega International Conference “Innovations in Computers Networks, Computational Intelligence and IoT ” ICICCI-21)

- PAPER ID:ICICCI-21-0067,

- PAPER TITLE:BLOCK CHAIN FOR SECURE EHRS SHARING OF CLOUD BASED E-HEALTH SYSTEMS

CHAPTER 13

STUDENTS PROFILE

Alluri Suraj Reddy is currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. He completed his intermediate from Sri Chaitanya Junior College and 10th class from St. Joseph's Convent High School. His technical skills include C, C++, Python, Java and MySQL. He also has a basic understanding of ML. He is one of the members of Street Cause SMEC NGO, He was a Volunteer(2018-2019) later got promoted as Associate president for Street Cause SMEC (2019-2020) and for his work, Street Cause Hyderabad NGO has awarded him as the "Exceptional EB"(2020). Now he is a member in Street Cause Gold. He is also a student of Smart Interviews. His participations include: Internship at SmartBridge which is collaborated with IBM on Machine Learning, Developed the project entitled "Predicting Life Expectancy using Machine Learning" (May 2020 – Jun 2020). Internship at Electronics Corporation of India Limited, Hyderabad (ECIL-ECIT). Developed the project entitled "Online Attendance Management system for Organization" (Jun 2019- Jul 2019). Attended the India First Leadership Talk webinar conducted by MHRD's Innovation Cell on 9th May 2020.Workshop on HTML/CSS conducted by TAM from 5th Jan 2018 to 3rd Feb 2018. Attended 2days Entrepreneurship Summit conducted at MLRIT, Hyd on 21st & 22nd Aug 2017. His areas of interest are Artificial Intelligent, Data Science. He completed few certification courses from online platforms like Coursera, udemy, Insidesherpa, CursaApp and SoloLearn.

Edara Chakravarthi Reddy is currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. He completed his Intermediate from Sri Chaitanya Junior College and 10th class from Viswabharathi English Medium High School. His technical skills include C, C++, Python and MySQL. He also has a Intermediate level of understanding in Cybersecurity. He is one of the members of Telangana Academy for Skill and Knowledge (**TASK**). His participations include: Internship at Verzeo which is collaborated with Microsoft on Cybersecurity, Developed the project entitled "SQL injection and Web-application Security" (May 2020 – Aug 2020). He took part in Employability Skill development Program conducted by Zensar. He is also a student of Smart Interviews. His participations include: National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020, "Know More - Teach More ", the Global Webinar on Cloud & Big Data – Changing the way we work which was conducted Global Education & Careers Forum (GECF) on 12th August 2020, "One Day Webinar on Internet of Things and Its Applications" conducted by Anand Institute of Technology on 21st May 2020 and IIC Online Sessions conducted by Institution's Innovation Council (IIC) of MHRD's Innovation Cell, New Delhi to promote Innovation, IPR, Entrepreneurship, and Start-ups among HEIs from 28th April to 22ndMay 2020. Attended the India First Leadership Talk webinar conducted by MHRD's Innovation Cell on 9th May 2020.Workshop on HTML/CSS conducted by TAM from 5th Jan 2018 to 3rd Feb 2018. His areas of interest are Artificial Intelligent, Data Science and Cybersecurity. He completed few certification courses from online platforms like Coursera, Udemy and CursaApp.

Niharika Bolla is currently pursuing her Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. She completed her intermediate from sri Chaitanya junior kalasala and 10th class from sri Chaitanya School. Her technical skills include C, Python and C++. Her participations include: National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020, "Know More - Teach More ", the Global Webinar on Cloud & Big Data – Changing the way we work which was conducted Global Education & Careers Forum (GECF) on 12th August 2020, Women online workshop on "Women in Cyber Security and Privacy in 2020" which was conducted from 6th to 10th July 2020, "Techskill E-Quiz" conducted by panimalar institute of technology on 9th may 2020.She completed certification courses from Coursera, and SoloLearn.

Peesaradolu Nandini is currently pursuing her Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. She completed her intermediate from Gauthami Junior College and 10th class from Vishal High School. Her technical skills include C, Python and C++. Her participations include: National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020, "Know More - Teach More ", the Global Webinar on Cloud & Big Data – Changing the way we work which was conducted Global Education & Careers Forum (GECF) on 12th August 2020, Women online workshop on "Women in Cyber Security and Privacy in 2020" which was conducted from 6th to 10th July 2020, "Techskill E-Quiz" conducted by Panimalar institute of technology on 9th may 2020. She completed certification courses from Coursera, and SoloLearn.

CHAPTER 14

APPENDICES

An appendix contains supplementary material that is not an essential part of the text itself but which may be helpful in providing a more comprehensive understanding of the research problem and/or is information which is too cumbersome to be included in the body of the paper. A separate appendix should be used for each distinct topic or set of data and always have a title descriptive of its contents.

APPENDIX A – DIGITAL HEALTH AGENCIES/ EHRS IN EACH PROVINCE/TERRITORY:

British Columbia

1. Digital Health Hub: http://www.canadadhh.com/

2. Digital Health Circle: https://www.digitalhealthcircle.ca/

3. BC Children's Hospital Digital Health Innovation Lab: https://www.bcchr.ca/dhil

4. BC Health Information Management Professionals Society: https://www.bchimps.org/

Alberta

1. Alberta Netcare: https://www.albertanetcare.ca/

2. MyHealth Records: https://myhealth.alberta.ca/mhr-features

Manitoba

1. LibreMD: https://www.libremd.com/

2. MBTelehealth: https://mbtelehealth.ca/

3. Shared Health Manitoba: https://sharedhealthmb.ca/

Saskatchewan

1. eHealth Saskatchewan: https://www.ehealthsask.ca/Pages/default.aspx

2. Sunrise Clinical Manager (SCM): https://www.saskatoonhealthregion.ca/locations_services/Services/Digital Health/Pages/Home.aspx

3. Lumeca: https://lumeca.com/

Ontario

1. eHealth Ontario: https://www.ehealthontario.on.ca/en/

2. Ontario MD: https://www.ontariomd.ca/

3. Ontario Telemedicine Network: https://otn.ca/

Quebec

1. Quebec Health Record: https://www.quebec.ca/en/health/your-health-information/quebec-health-record/

2. Cristal-Net: http://www.dcicristalnet.com/ BLOCKCHAIN IN HEALTH CARE 42 csagroup.org

New Brunswick

1. Accreon Health Cloud: https://accreon.com/interoperability/

2. eVisitNB: https://www.evisitnb.ca/

Nova Scotia

1. myHealthNS: https://www.myhealthns.ca/

2. NS Medical Devices: https://www.nsmedicaldevices.com/

Prince Edward Island

1. Health PEI: https://www.princeedwardisland.ca/en/information/health-pei/electronic-health-records-ehrs

Newfoundland and Labrador

The Newfoundland and Labrador Centre for Health Information (NLCHI): https://www.nlchi.nl.ca/

Yukon Territory

1. eHealth Yukon: http://www.hss.gov.yk.ca/ehealth.php

2. Yukon Hospitals (Telehealth): https://yukonhospitals.ca/yukon-hospital-corporation/telehealth

Northwest Territories

1. NWT HealthNet: https://www.hss.gov.nt.ca/en/services/nwt-healthnet

2. NWT Virtual Care: https://www.nthssa.ca/en/services/nwt-virtual-care

APPENDIX B – ADDITIONAL STANDARDS:

Standards, specifications or reports under development (except otherwise noted) by IEEE Blockchain Initiative:

- P2140.1 – Standard for General Requirements for Cryptocurrency Exchanges (Published)
- P2140.2 – Standard for Security Management for Customer Cryptographic Assets on Cryptocurrency Exchanges
- P2140.3 – Standard for User Identification and Anti-Money Laundering on Cryptocurrency Exchanges
- P2140.4 – Standard for Distributed/Decentralized Exchange Framework using DLT (Distributed Ledger Technology)
- 2140.5-2020 – IEEE Standard for a Custodian Framework of Cryptocurrency (Published)
- P2141.1 – Standard for the Use of Blockchain in Anti-Corruption Applications for Centralized Organizations
- P2141.2 – Standard for Transforming Enterprise Information Systems from Centralized Architecture into Blockchain-based Decentralized Architecture
- P2141.3 – Standard for Transforming Enterprise Information Systems from Distributed Architecture into Blockchain-based Decentralized Architecture BLOCKCHAIN IN HEALTH CARE 44 csagroup.org
- P2142.1 – Recommended Practice for E-Invoice Business Using Blockchain Technology
- 2143.1-2020 – IEEE Standard for General Process of Cryptocurrency Payment (Published)
- P2143.2 – Standard for Cryptocurrency Payment Performance Metrics
- P2143.3 – Standard for Risk Control Requirements for Cryptocurrency Payment
- P2145 – Standard for Framework and Definitions for Blockchain Governance
- P2146.1 – Standard for Entity-Based Risk Mutual Assistance Model through Blockchain Technology

- P2146.2 – Standard for External Data Retrieval of Blockchain for Risk Mutual Assistance Model
- 2418.2-2020 – IEEE Approved Draft Standard Data Format for Blockchain Systems
- P2418.3 – Standard for the Framework of Distributed Ledger Technology (DLT) Use in Agriculture
- P2418.4 – Standard for the Framework of Distributed Ledger Technology (DLT) Use in Connected and Autonomous Vehicles (CAVs)
- P2418.5 – Standard for Blockchain in Energy
- P2418.7 – Standard for the Use of Blockchain in Supply Chain Finance
- P2418.8 – Standard for Blockchain Applications in Governments
- P2418.9 – Standard for Cryptocurrency Based Security Tokens
- P2418.10 – Standard for Blockchain-based Digital Asset Management
- P2677.1 – Standard for Blockchain-based Omnidirectional Pandemic/epidemic Surveillance: Overarching Framework
- P2677.10 – Standard for Blockchain-based Omnidirectional Pandemic/epidemic Surveillance: Access to Personal Data
- P2677.11 – Standard for Blockchain-based Omnidirectional Pandemic/epidemic Surveillance: Access to Telecommunications Data
- P2677.12 – Standard for Blockchain-based Omnidirectional Pandemic/epidemic Surveillance: Access to Transportation Data
- P2677.20 – Standard for Blockchain-based Omnidirectional Pandemic/epidemic Surveillance: Requirements for Blockchain Infrastructure

APPENDIX C: SUPPLEMENTARY FILE:

Supplementary material related to this article can be found, in the online version, at:

https://ieeexplore.ieee.org/document/8717579

Blockchain: https://www.blockchain.com

A

PROJECTREPORT

On

PERSONAL VOICE ASSISSTANT

Submitted by

1)Ms.M.Aashritha (17K81AO5G0) 2)Mr.G.Suraj (17K81A05E0)

3)Ms.J.Srichandana(17K81AO5E5) 4)Mr.A.Sanjaykumar(17K81A05C4)

inpartialfulfillmentfortheawardof the

degreeof

BACHELOR OF TECHNOLOGY

IN

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Under the Guidance of

Ms.Laxmi Devi

Assisstant Professor(M.Tech)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ST.MARTIN'S ENGINEERING COLLEGE

An Autonomous Institute

Dhulapally, Secunderabad – 50010

JUNE 2021

BONAFIDE CERTIFICATE

This is to certify that the project entitled**PERSONAL VOICE ASSISTANT**, is being submitted by **1)Ms.M.Aashritha(17K81A05G0),2)Ms.J.SriChandana(17K81A05E5), 3)Mr.G.Suraj(17K81A05E0),4)Mr.A.SanjayKumar(17K81A05C4)**in partial fulfillment of the requirement forthe award of the degree of Bachelor of Technology in**COMPUTERSCIENCEENGINEERING**is recorded of bonafide work carried out by them.the result embodied in this report have been verified and found satisfactory.

Signature

Laxmi Devi

Department of CSE

Head of the Department

Dr.M.NARAYANAN

Department of CSE

Internal Examiner

External Examiner

Place:

Date:

DECLARATION

We, the student of **Bachelor of Technology** in Department of Computer Science and Engineering', session: 2017 – 2021, St. Martin's Engineering College, Dhulapally, Kompally, Secunderabad, hereby declare that work presented in this Project Work entitled Personal Voice Assisstant is the outcome of our own bonafide work and is correct to the best of our knowledge and this work has been undertaken taking care of Engineering Ethics. This result embodied in this project report has not been submitted in any university for award of any degree.

M.Aashritha 17K81A05G0

J.Srichandana 17K81A05E5

G.Suraj 17K81A05E0

A.Sanjaykumar 17K81A05C4

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible and whose encouragements and guidance have crowded effects with success. We extended our deep sense of gratitude to Principal**,Dr. P. SANTOSH KUMARPATRA**, St. Martin's Engineering College, Dhulapally, for permitting us to undertake this project. We are also thankful to **Dr. M.NARAYANAN**, Head of the Department, Computer Science and Engineering, St.Martin's Engineering College, Dhulapally, for his support and guidance throughout our project. We would like to thank **Dr. T. POONGOTHAI,** Department of Computer Science and Engineering (AI & ML) for her encouragement and insightful comments and as well as our project coordinators **Dr. B.RAJALINGAM**, Associate Professor and **Dr. N. SATHEESH**, Professor, in Department of Computer Science and Engineering for their valuable support. We would like to express our sincere gratitude and indebtedness to our project supervisor **Ms.Laxmi Devi, Assisstant professor**, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.Finally, we express thanks to all those who have helped us successfully to completing this project. Furthermore, we would like to thank our family and friends for their moral support and encouragement. We express thanks to all those who have helped us in successfully completing the project.

Ms.M.Ashrithaa 17K81A05G0
Ms.J.Sri Chandana 17K81A05E5
Mr.G.Suraj 17K81A05E0
Mr.A.Sanjay Kumar 17K81A05C4

ABSTRACT

In this project we introduce, personal voice assistant that use to take the user commands as input and perform tasks based on the user commands. It provides more efficient and natural interaction with support of multiple voice commands in the same utterance. This assistant has a unique face recognition technique through which only the authorized user can provide the command to the assistant and can perform their various tasks on system. Our paper reaches out to help our society by making their work easier as this system can tell the news, search what you want, send email by only your voice command, play game with you, set reminders, tell the location, forecast weather, can tell horoscope of you and endless number of tasks can be done by this. Thus our system can be used for the doing the multi-purpose tasks in robust and flexible approaches. Key Words: Speech Recognition, Face Recognition, TTS, Voice command, Voice assistant.

TABLEOFCONTENTS

LIST OF FIGURES

LIST OF OUTPUT FIGURES

LIST OF ACRONYMS

ACRONYM	DESCRIPTI
NLP	Natural Language Processing
SDLC	Software Development Life Cycle
GUI	Graphical User Interface
API	Application Programming Interface
CLI	Command Line Interface

1.INTRODUCTION

Today, we can ask voice assistants like Apple's Siri, Google Now to perform simple tasks like, "What's the weather", "Remind me to take pills in the morning", etc. in our own natural language. The next evolution of natural language interaction with voice assistants is in the form of task automation such as "turn on the air conditioner whenever the temperature rises above 30 degrees Celsius", or "if there is motion on the security camera after 10pm, call Bob". A voice assistant is a digital assistant that uses voice recognition, speech synthesis and natural language processing (NLP) to provide a service through a particular application. Now everyone wants to have an assistant who listen our call, anticipates our needs and can take necessary action when needed. This luxury life is now available with the help of Artificial Intelligence based on voice assistant. Voice assistants come's in small packages and can perform a variety of actions after hearing our commands. They can launch apps, open web browser, answer basic informational queries, tell horoscope, calculate your BMI, answer our queries, play music, send email, set reminders, make lists, and do basic math calculations, etc

1.1 PROJECT OVERVIEW

Describe this project or product and its intended audience, or provide a link or reference to the project charter.

1.2 PROJECT OBJECTIVES

- Voice Assistant is a virtual assistant that uses speech recognition, natural language processing and speech synthesis to take actions to help its users.
- This project mainly focuses on the voice which detects and gives the required output.

1.3 ORGANIZATION OF CHAPTERS

Besides the introduction, the thesis is organized in other six chapters as follows:

Chapter 2, LITERATURE SURVEY: the review is made in the context of hand gesture recognitionsystems with a particular attention on those implementations that assess the scalability and performances or their implementations. Most of the related work is on convolution neural network, whereas a small part is on cloud solutions. It will be possible to notice that only a small subset of the literature actually focuses on the analysis of the systems in mass crises scenarios. Chapter 3, SOFTWARE AND HARDWARE REQUIREMENTS: this chapter discuss about the software and hardware required for the execution of the project. Chapter 4, SOFTWARE DEVELOPMENT ANALYASIS: this chapter explains the

assumptions and technical specifications of the project. Chapter 5, PROJECT SYSTEM DESIGN: this chapter explains all the software development process with DFD and UML diagrams clearly. Chapter 6, PROJECT CODING: this chapter explains the design of the system, roles and responsibilities, as well as the requirements of a HGR management solution based on CNN. Chapter 7, PROJECT TESTING: this chapter explains various test cases to test the project working. Chapter 8, OUTPUT SCREENS: explains a step by step process of the project execution. Chapter 9, EXPERIMENTAL RESULTS: tests and results are shown and explained in this chapter. The results are analyzed in the context of the thesis project and followed by discussion on systems throughput and resiliency, as well as the approaches to testing and analysis. Chapter 10, CONCLUSION AND FUTURE ENHANCEMENT: the chapter ends the project with a short summary of the main concepts mentioned in the thesis as well as the relevant results.

2.LITERATURE SURVEY

2.1 SURVEY ON BACKGROUNDS

Yash Mittal el al. [1] proposed a multi-functional `Smart Home Automation System' (SHAS) that can be adapted to a user's voice and recognize the voice-commands, independent of the speaker's personal characteristics such as accent. An Arduino microcontroller board is used for processing and control which makes this system cost effective. Thus for converting existing homes into a smart home this prototype i.e. Smart Home Automation System (SHAS) can be used. Prerna Wadikar, Nidhi Sargar, Rahool Rangnekar, Prof. Pankaj Kunekar, [2] “Home Automation using Voice Commands in the Hindi Language”: The proposed of Home Automation in Hindi language Voice commands was to implemented the dedicated hardware i.e. Arduino Uno and using voice recognition module that makes the system more cost-efficient and robust. The system can work on various connected devices like light, fan, AC, etc. This system allows users to make decisions and to regulate the home appliances with the help of voice assistants. Steve Joseph, Chetan Jha, Dipesh Jain, Saurabh Gavali, Manish Salvi [7], “Voice based E-Mail for the Blind”: They design the system that was helpful for sending emails for the blind people without the need of visual interaction with the screen. Speech-to-Text Based Life Log System for Smartphones [8], the technique used was Microphone of Smartphone, STT (Speech-To-Text). From this the user are able to search life log sound files using Text. Aditi Bhalerao, Samira Bhilare, Anagha Bondade, Monal Shingade, Aradhana Deshmukh [9], “Smart Voice Assistant: a universal voice control solution for non-visual access to the Android operating system”, design the voice control solution for the mobile device through which user can do their task without accessing towards their mobile screen. Chen-Yen Peng et al. [10] designed and built a tailor-made function for users without their attempt. Commands are taken from Google Home’s voice recognition and Bluetooth signals are transferred to Raspberry Pi to control the connected devices. The proposed paper mainly focuses on researching combining characteristics of Google Home with Google Assistant Personal Voice Assistant using machine learning and thereby customizing this to meet the new needs of users. G. KALYAN KUMAR, K. PAVAM KUMAR REDD “CORTANA (Intelligent Assistant)” [11], describe general language and processing capabilities of the Cortana are derived from Tell me Networks and are combined with a Semantic search database called Satori which is very much used in searching the data

2.2 CONCLUSION ON SURVEY

This work is a command-based speech recognition system. Used speech recognition module to take the input using microphone and gtts module to convert voice to text .after converting speech to text it uses os module to open files and uses playsound module to play sound. We used tikinter module to create user interface.

3.SOFTWARE AND HARDWARE REQUIREMENTS

3.1 SOFTTWARE REQUIREMENTS

- Operating System : Windows XP.
- Platform : PYTHON TECHNOLOGY
- Tool : Spyder, Python 3.5
- Front End : Anaconda
- Back End : python anaconda script

3.2 HARDWARE REQUIREMENTS

- System: Pentium IV 2.4 GHz.
- Hard Disk : 40 GB.
- Monitor : 15 inch VGA Color.
- Mouse : Logitech Mouse.
- Ram : 512 MB
- Keyboard : Standard Keyboard

4.SOFTWARE DEVELOPMENT ANALYASIS

Waterfall Mode:

The Waterfall Model is a linear sequential flow. In which progress is seen as flowing steadily downwards (like a waterfall) through the phases of software implementation. This means that any phase in the development process begins only if the previous phase is complete. The waterfall approach does not define the process to go back to the previous phase to handle changes in requirement.

In this article, we will discuss the advantages and disadvantages of the waterfall, should we avoid it? when to use it? and the waterfall model pitfall, and why I see it as the father of the SDLC models.

Waterfall Model Phases

Waterfall Model contains the main phases similarly to other process models, you can read this article for more information about phases definitions.

When to use Waterfall Model?

Due to the nature of the waterfall model, it is hard to get back to the previous phase once completed. Although, this is can be very rigid in some software projects which need some flexibility, while, this model can be essential or the most suitable model for other software projects' contexts.

The usage of the waterfall model can fall under the projects which do not focus on changing the requirements, for example:

1. Projects initiated from a request for proposal (RFP), the customer has a very clear documented requirements
2. Mission Critical projects, for example, in a Space shuttle
3. Embedded systems.

We can notice some similarities of these types of projects that they cannot be delivered in iterative, incremental, or agile manner, for example, in embedded systems for the elevator, you cannot deliver an elevator who can go up only without going down, or handling only users requests from inside and ignore outside calls for the elevator.

Validation and Verification Model –V-Model

V-Model is mostly known as the validation and verification software development process model (The Vee Model), and It is one of the most know software development methodology. Although it is considered as an improvement to the waterfall model and it has some similarities as the process also based on sequential steps moving down in a linear way, it differs from the waterfall model as the steps move upwards after the coding phase to form the typical V shape. This V shape demonstrates the relationships between each phase of the development life cycle and its as sociated phase of testing.

This means that any phase in the development process begins only if the previous phase is complete and has a correspondence related testing phase which is performed against this phase completion. Similar to the Waterfall model, the V-Model does not define the process to go back to the previous phase to handle changes in requirement.

The technical aspect of the project cycle is considered as a V shape starting with the business needs on the upper left and ending with the user acceptance testing on the upper right.

V-Model Model Phases

The V-Model Model contains the main phases similarly to other process models, you can read this article for more information about SDLC phases definitions.

Moreover, it breaks down the testing phase into detailed steps to ensure the validation and verification process. So, it contains the below testing phases:

Unit Testing

The Unit testing is the testing at the code level and helps eliminate issues at an early stage, mainly the developer is responsible to perform the unit test for his code while not all the defects cannot be discovered at the unit testing.

Functional Testing

Functional testing is associated with the low-level design phase which ensures that collections of codes and units are working together probably to execute new function or service.

Integration Testing

Integration testing is associated with the high-level design phase. Integration testing ensures the integration between all system modules after adding any new functions or updates.

System Testing

System testing is associated with the system requirements and design phase. It combines the software, hardware, and the integration of this system with the other external systems.

User Acceptance Testing

UserAcceptance testing is associated with the business and operations analysis phase. The customer users are the main performers of this testing based on test cases and scenarios that cover the business requirements to ensure that they have delivered the right software as per the specifications.

4.1 OVERVIEW OF A PROBLEM

The voice assistant is design to make the work easier of the user. As user can give command to them without making visual access to the screen. Personal voice assistant has been arounded for a quiet period of time and is a common research subject.so we made an algorithm to take the command as input and actions are performed according to the commands

4.2 DEFINE THE PROBLEM

The voice assistant is design to make the work easier of the user. As user can give command to them without making visual access to the screen. The biggest disadvantage of this system is that confidential data can be accessed by unauthorized user so the privacy can be breached. Due to this, the confidentiality, integrity and availability (CIA) of user data is affected. Looking to this problem the security features of “Face Recognition” is designed so that it can detect the authorized user face and take user command as input and provide response via a synthesis voice. Facial recognition technology (FRT) is one of the most controversial new tools. It was first developed in the 1960s. It has recently become accessible to the mass market-to both law enforcement and private consumers

4.3 MODULES OVER VIEW

1)tikinter

2)speech recognition

3)gtts

4)os

4.4 DEFINE THE MODULE

1)tkinter: In Python, Tkinter is a standard GUI (graphical user interface) package. Tkinter is Python's default GUI module and also the most common way that is used for GUI programming in Python. Note that Tkinter is a set of wrappers that implement the Tk widgets as Python classes.

2)speech recognition: Voice Recognition Module is a compact easy-controlspeaking recognition board. It is a speaker-dependent module and supports up to 80 voice commands. Any sound could be trained as command.... Any 7 voice commands in the library could be imported into recognizer. It means 7 commands are effective at the same time.

3)gtts:(Google Text-to-Speech), a Python library and CLI tool to interface with Google Translate's text-to-speech API. Write spoken mp3 data to a file, a file-like object (bytestring) for further audio manipulation, or stdout . Or simply pre-generate Google Translate TTS request URLs to feed to an external program.

4)os: The OS module in Python provides functions for interacting with the operating system. OS comes under Python’s standard utility modules. This module provides a portable way of using operating system dependent functionality. The *os* and *os.path* modules include many functions to interact with the file system.

4.5 MODULE FUNCTIONALITY

A module is a separate unit of software or hardware. Typical characteristics of modular components include portability, which allows them to be used in a variety of systems, and interoperability, which allows them to function with the components of other systems.

5.PROJECT SYSTEM DESIGN

5.1 DFDS IN CASE OF DATABASE PROJECTS

DFD is a graphical representation which provides information flow between input and output data. It is also known as “Data Flow Chart or Bubble Chart”. A DFD is often used as a preliminary step to create an overview of the system, which can later be elaborated.

Level 0 DFD:

The user gives the input in the form of voice; this voice command is recognized by the application. Then it will check whether it is the authorized user, then action is performed as per the command given by the user. Command given is compared as a form of action and question and responsed with the dialog box or search through the knowledge base.

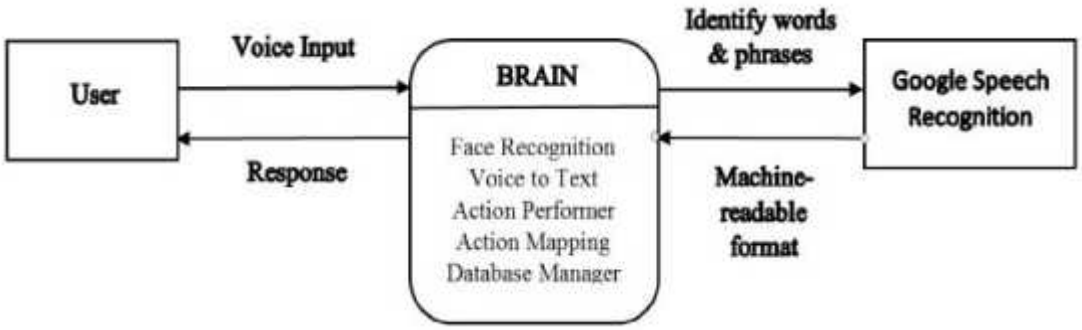


Fig 5.1 data flow diagram DFD0

Level 1 DFD:

Input is given by user in the form of voice. GoogleVoiceAPI will convert this voice data in text form and then the action is performed by the voice assistant according to the command given by the user by comparing with the dialog box and knowledge base.

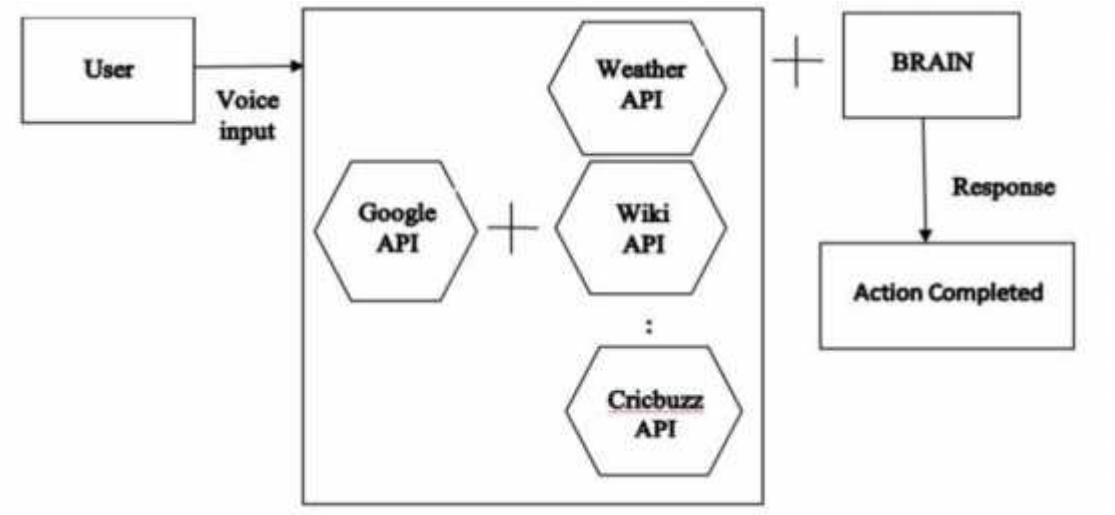


Fig 5.2 data flow diagram DFD1

5.2 UML DIAGRAMS

UML stands for Unified Modeling Language. UML is a standardized general-purpose modeling language in the field of object-oriented software engineering. The standard is managed, and was created by, the Object Management Group.

The goal is for UML to become a common language for creating models of object oriented computer software. In its current form UML is comprised of two major components: a Meta-model and a notation. In the future, some form of method or process may also be added to; or associated with, UML.

The Unified Modeling Language is a standard language for specifying, Visualization, Constructing and documenting the artifacts of software system, as well as for business modeling and other non-software systems.

The UML represents a collection of best engineering practices that have proven successful in the modeling of large and complex systems.

The UML is a very important part of developing objects oriented software and the software development process. The UML uses mostly graphical notations to express the design of software projects.

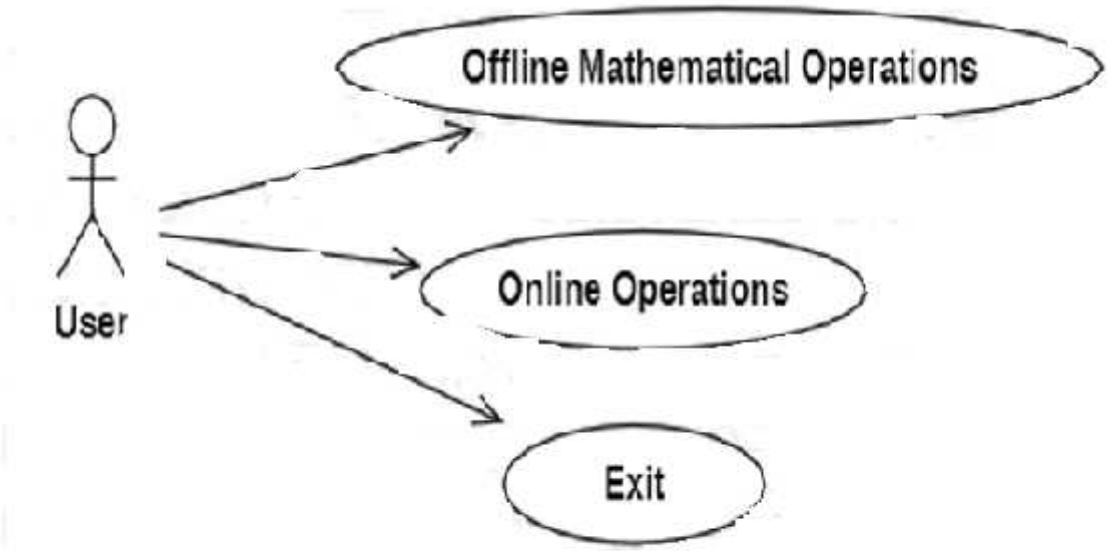


Fig 5.3 uml diagram

CLASS DIAGRAM:

In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.

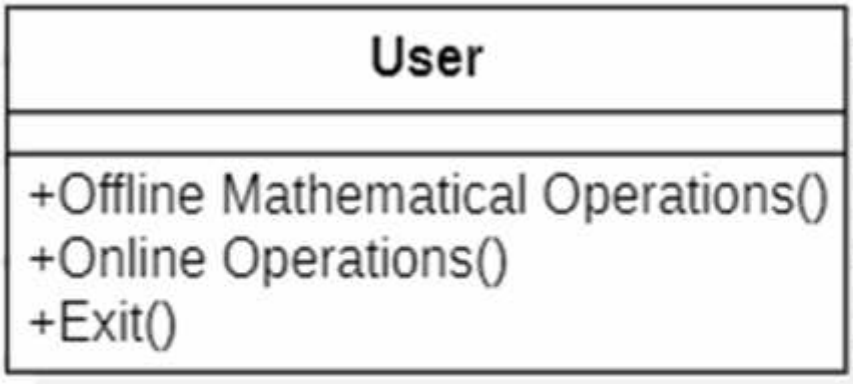


Fig 5.4 class diagram

SEQUENCE DIAGRAM:

A sequence diagram in Unified Modeling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.

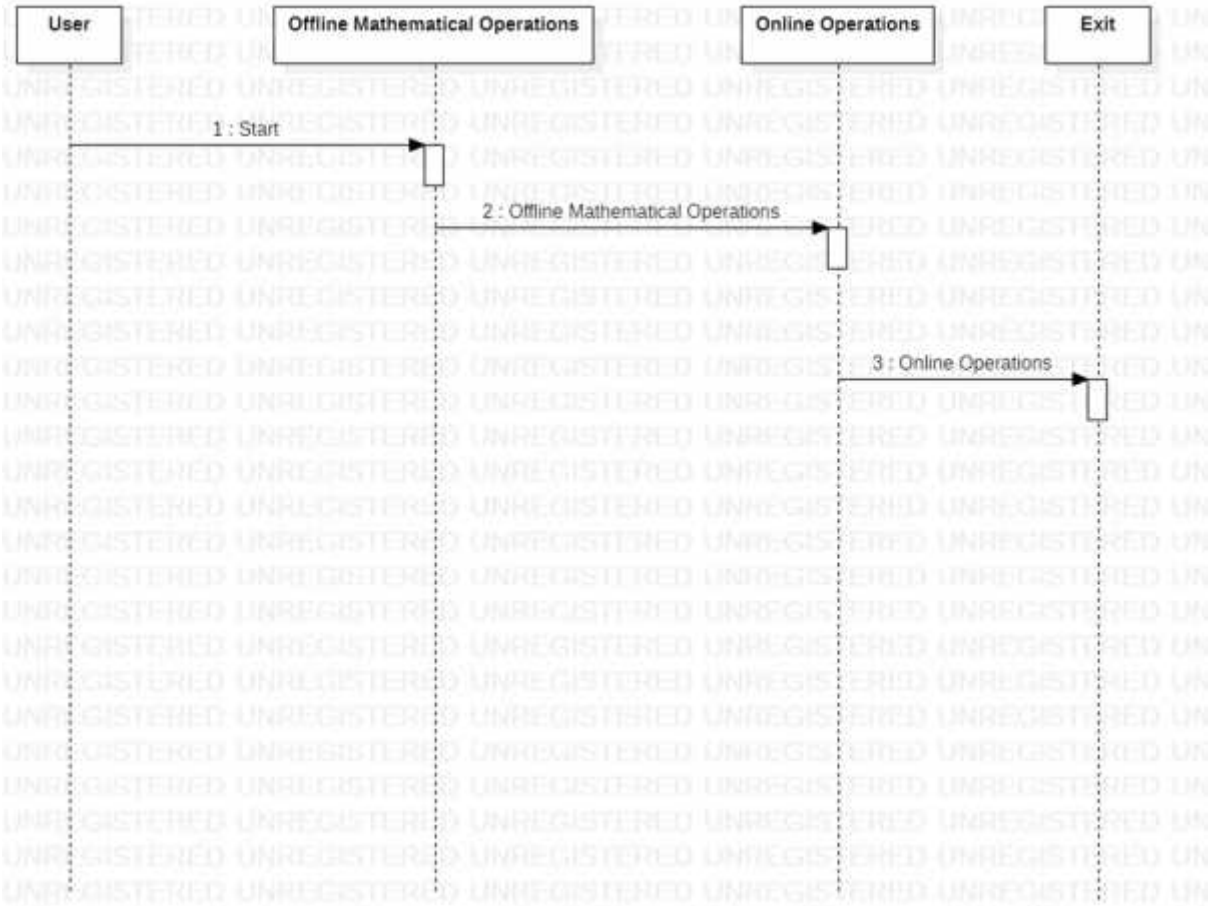


Fig 5.5 sequence diagram

ACTIVITY DIAGRAM:

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modeling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.

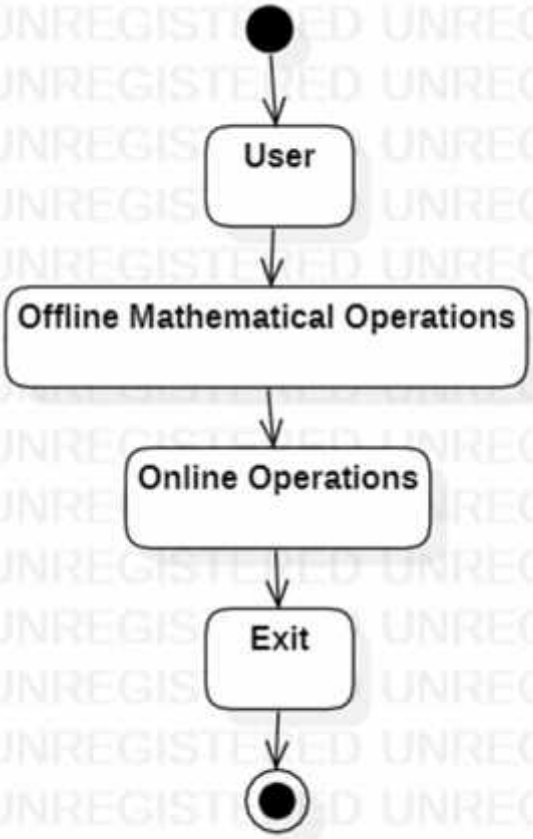


Fig 5.6 activity diagram

6.PROJECT CODING

6.1 CODE TEMPLATES

```
def play():

data = text.get("1.0",END)

print("test "+data)

t1 = gTTS(text=data, lang='en', slow=False)

t1.save("output.mp3")

playsound("output.mp3")
```

def runOffline(data):

```
data = data.lower()

current = 0

total = 0

count = 0

arr = data.split(" ")

i = 0

while i < len(arr):

print(str(arr[i])+" "+str(arr[i].isnumeric())+" "+str(i))

if arr[i].isnumeric():

current = float(arr[i])

count = count + 1
```

```
if total == 0:

total = current

#play()

if arr[i] == 'into' or arr[i] == '*':

i = i + 1

current = float(arr[i])

total = total * current

#play()

if arr[i] == 'plus' or arr[i] == '+':

i = i + 1

current = float(arr[i])

total = total + current

#play()

if arr[i] == 'minus' or arr[i] == '-':

i = i + 1

current = float(arr[i])

total = total - current

#play()

if arr[i] == 'divide':

total = total / current
```

```
#play()

if arr[i] == 'percentage':

i = i + 1

current = float(arr[i])

total = total / count

#play()

if arr[i] == 'power':

i = i + 1

current = float(arr[i])

total = pow(total,current)

#play()

if arr[i] == 'root':

i = i + 1

current = float(arr[i])

total = total ** count

#play()

if arr[i] == 'increase':

i = i + 1

current = float(arr[i])

total = total + 1
```

```
#play()

if arr[i] == 'decrease':

i = i + 1

current = float(arr[i])

total = total - 1

#play()

if arr[i] == 'table':

i = i + 2

current = int(arr[i])

for k in range(1,11):

text.insert(END,str(current)+" * "+str(k)+" = "+str(current * k)+"\n")

#play()

i = i + 1

return total

def offline():

text.delete('1.0', END)

r = sr.Recognizer()

mic = sr.Microphone()

with mic as source:

text.insert(END,"speak\n")
```

```
text.update_idletasks()

print('speak')

r.adjust_for_ambient_noise(source)

audio = r.record(source,duration=5)

data = r.recognize_google(audio)

text.insert(END,"Received Command : "+data+"\n")

text.update_idletasks()

total = runOffline(data+"\n\n")

text.insert(END,"Computed Result Below\n\n"+str(total)+"\n")

text.update_idletasks()

play()

def runOnline(command):

data = command.lower()

data = data.strip("\n")

data = data.strip()

arr = data.split(" ")

i = 0

while i < len(arr):

print(str(i)+" "+arr[i])

if arr[i].strip("\n").strip() == 'search':
```

```
i = i + 1

url = "https://www.google.com.tr/search?q={}".format(arr[i])

webbrowser.open_new_tab(url)

if arr[i].strip("\n").strip() == 'settings':

subprocess.Popen([r"C:/Windows/System32/DpiScaling.exe"])

if arr[i].strip("\n").strip() == 'wi-fi':

i = i + 1

if arr[i].strip("\n").strip() == 'on':

os.system("netsh interface set interface 'Wifi' enabled")

if arr[i].strip("\n").strip() == 'off':

os.system("netsh interface set interface 'Wifi' disabled")

if arr[i].strip("\n").strip() == 'open':

i = i + 1

name = arr[i]

name = name.strip("\n")

name = name.strip()

print(str(os.path.exists('E:/'+name+".txt"))+" "+str(name))

if os.path.exists('E:/'+name+".txt"):

os.system("start " + 'E:/'+name+".txt")

if os.path.exists('E:/'+name+".doc"):
```

```
os.system("start " + 'E:/'+name+".doc")

if os.path.exists('E:/'+name+".docx"):

os.system("start " + 'E:/'+name+".docx")

if os.path.exists('E:/'+name+".pdf"):

os.system("start " + 'E:/'+name+".pdf")

if os.path.exists('E:/'+name+".jpg"):

os.system("start " + 'E:/'+name+".jpg")

if os.path.exists('E:/'+name+".png"):

os.system("start " + 'E:/'+name+".png")

if os.path.exists('E:/'+name+".gif"):

os.system("start " + 'E:/'+name+".gif")

i = i + 1

def online():

text.delete('1.0', END)

r = sr.Recognizer()

mic = sr.Microphone()

with mic as source:

text.insert(END,"speak\n")

text.update_idletasks()

print('speak')
```

```
r.adjust_for_ambient_noise(source)

audio = r.record(source,duration=5)

data = r.recognize_google(audio)

text.insert(END,"Received Command : "+data+"\n")

text.update_idletasks()

runOnline(data+"\n\n")

#play()

def close():

main.destroy()

font = ('times', 15, 'bold')

title = Label(main, text='PERSONAL VOICE ASSISTANT')

#title.config(bg='powder blue', fg='olive drab')

title.config(font=font)

title.config(height=3, width=120)

title.place(x=0,y=5)

font1 = ('times', 13, 'bold')

ff = ('times', 12, 'bold')

offlineButton = Button(main, text="Offline Mathematical Operations", command=offline)

offlineButton.place(x=300,y=100)

offlineButton.config(font=ff)
```

```
onlineButton = Button(main, text="Online Operations", command=online)

onlineButton.place(x=300,y=150)

onlineButton.config(font=ff)

exitButton = Button(main, text="Exit", command=close)

exitButton.place(x=300,y=200)

exitButton.config(font=ff)

font1 = ('times', 12, 'bold')

text=Text(main,height=25,width=130)

scroll=Scrollbar(text)

text.configure(yscrollcommand=scroll.set)

text.place(x=10,y=250)

text.config(font=font1)

main.config()

main.mainloop()
```

6.2 OUT LINE FOR VARIOUS FILES

We used Python programming to implement our project. A single python file is used to implement our code. This file consists of various modules that we have used. Our projectmodules are -gtts,speech_recognition,subprocess. We also used various python modules like tkinter,os, subprocess,playsound, webbrowser.

6.3METHODS INPUT AND OUTPUT PARAMETERS:

In our project code, we implemented six different methods. They are:

1. play()

2. runoffline()

3. offline()

4. runonline()

5. online()

6.close()

Our first method play() doesn’t take any input parameters but after successful execution, it speaks out the actions done. Second method runoffline() it takes the input parameter data that is command given by us and outputs the result.Third Method offline() it doesn’t have input parameter it takes the input from user and transfers command to runoffline method. runonline() it takes the input parameter data that is the command we spoke, and performs action according to command.online() methid doesn’t take any input but takes command from user and transfer data to runonline method .close() don’t have any parameters but upon clicking this button, it will close the project window.

7. PROJECT TESTING

7.1 VARIOUS TEST CASES

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

7.2 BLACK BOX TESTING

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box .you cannot “see” into it. The test provides inputs and responds to outputs without considering how the software works.

7.3 WHITE BOX TESTING

White Box Testing is a testing in which in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is purpose. It is used to test areas that cannot be reached from a black box level.

Unit Testing

Unit testing is usually conducted as part of a combined code and unit test phase of the software lifecycle, although it is not uncommon for coding and unit testing to be conducted as two distinct phases.

Test strategy and approach

Field testing will be performed manually and functional tests will be written in detail.

Test objectives

- All field entries must work properly.
- Pages must be activated from the identified link.

- The entry screen, messages and responses must not be delayed.

Features to be tested

- Verify that the entries are of the correct format
- No duplicate entries should be allowed
- All links should take the user to the correct page.

Integration Testing

Software integration testing is the incremental integration testing of two or more integrated software components on a single platform to produce failures caused by interface defects.

The task of the integration test is to check that components or software applications, e.g. components in a software system or – one step up – software applications at the company level – interact without error.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

Acceptance Testing

User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements.

Test Results: All the test cases mentioned above passed successfully. No defects

8.OUTPUT SCREENS

8.1 USER INTERFACES

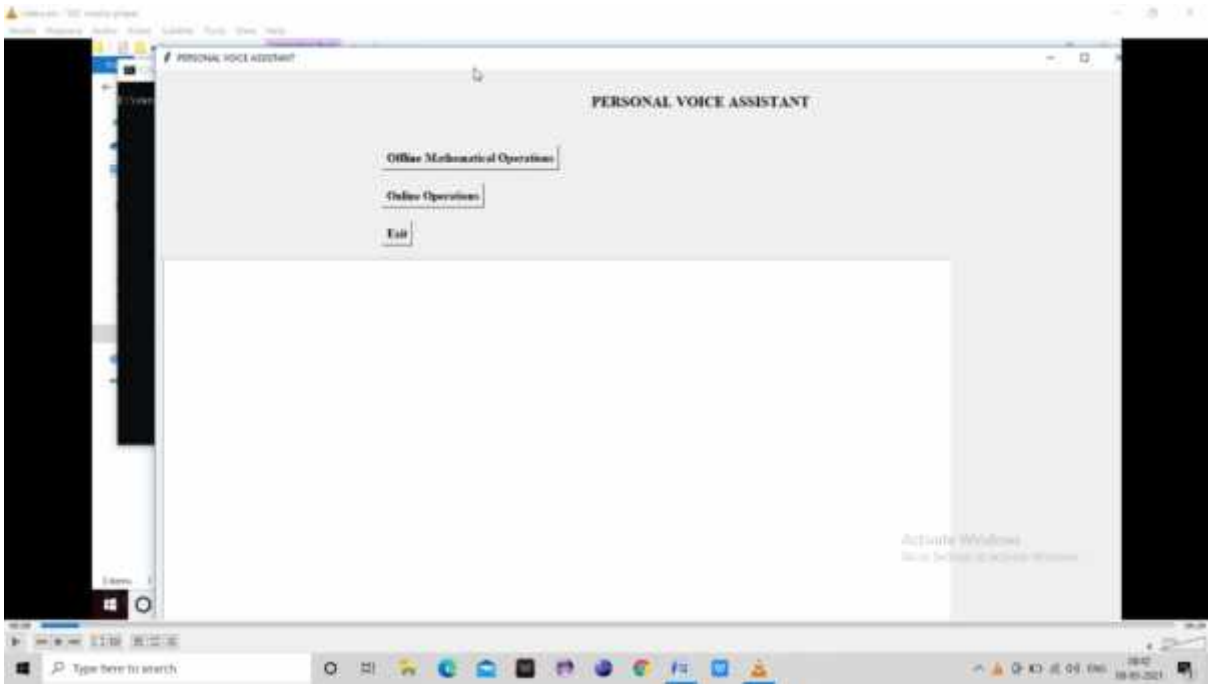
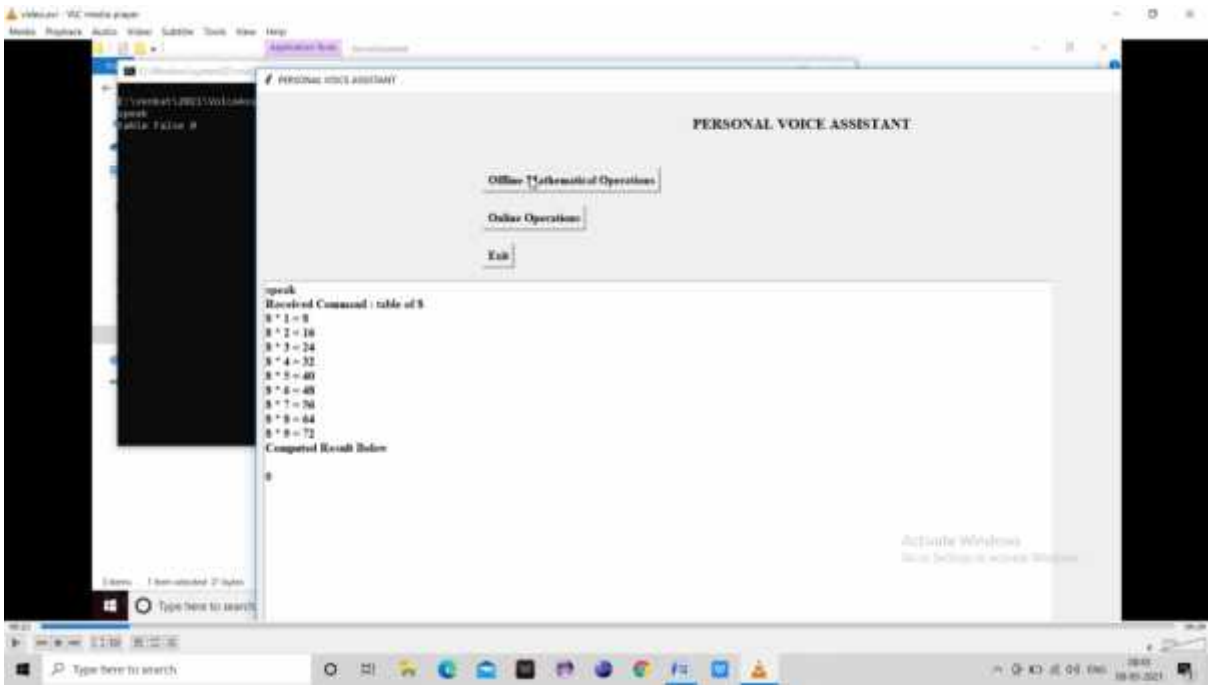


Fig 8.1.1 user interface

8.2 OUTPUT SCREENS



8.2.1 Result of command table of 8

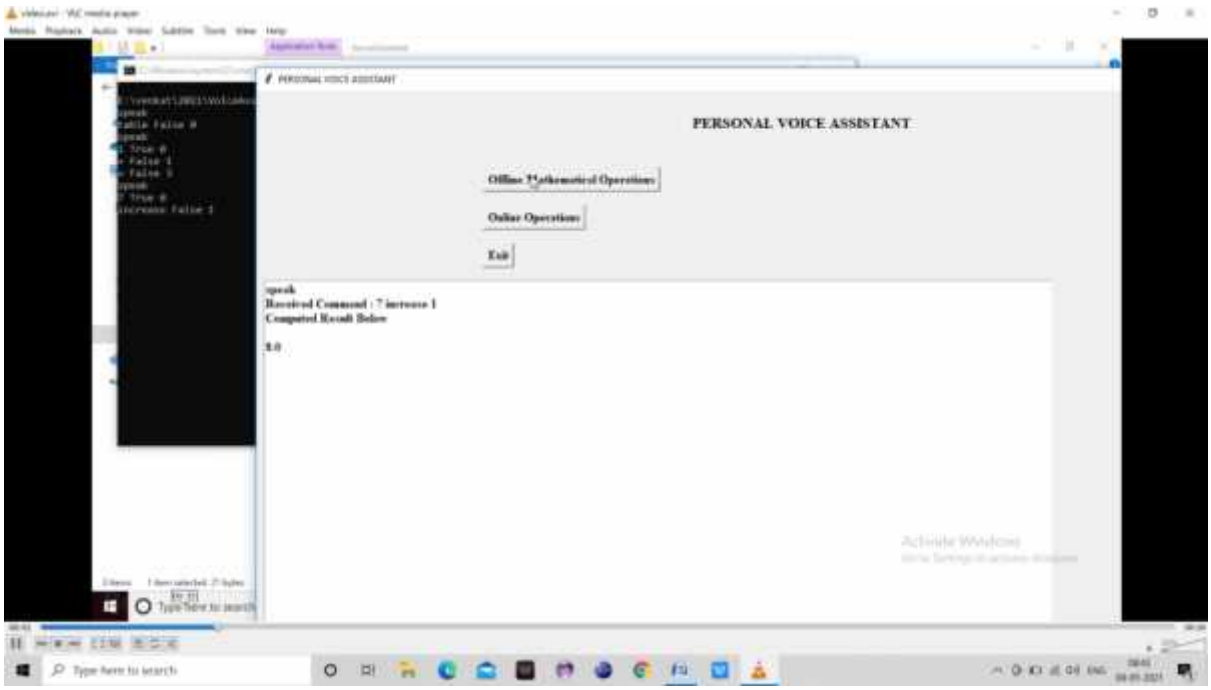


Fig 8.2.2 result of command 7 increase 1

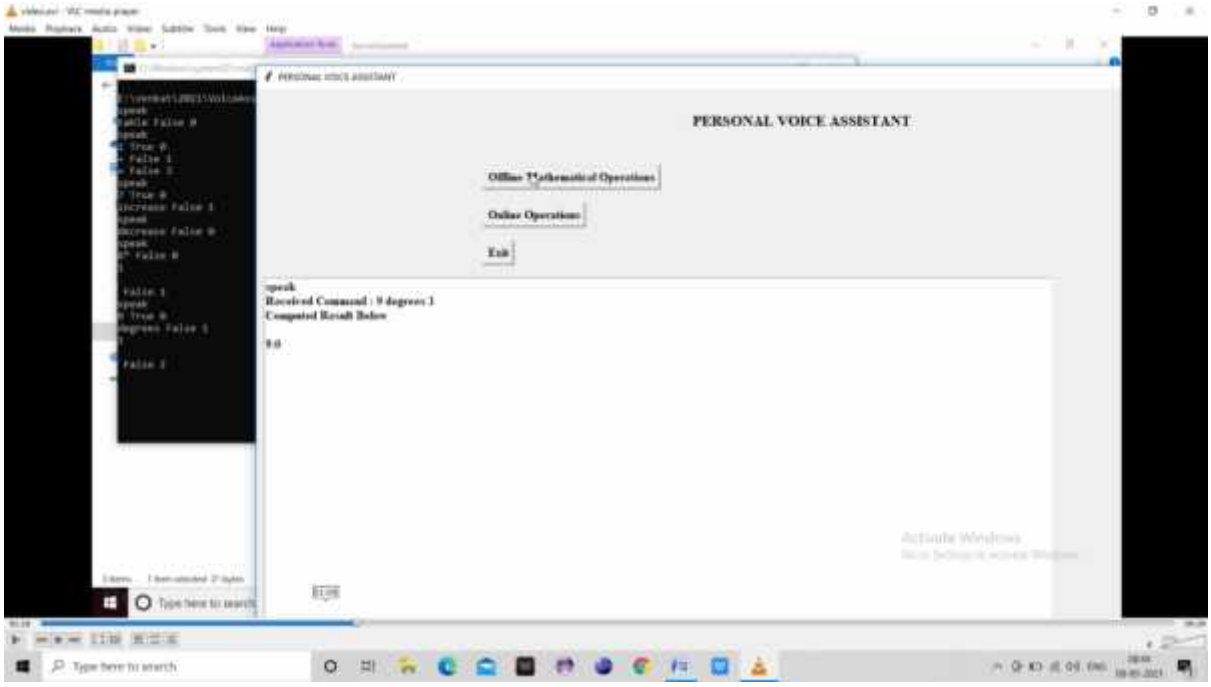


Fig 8.2.3 result of command 9 degrees 1

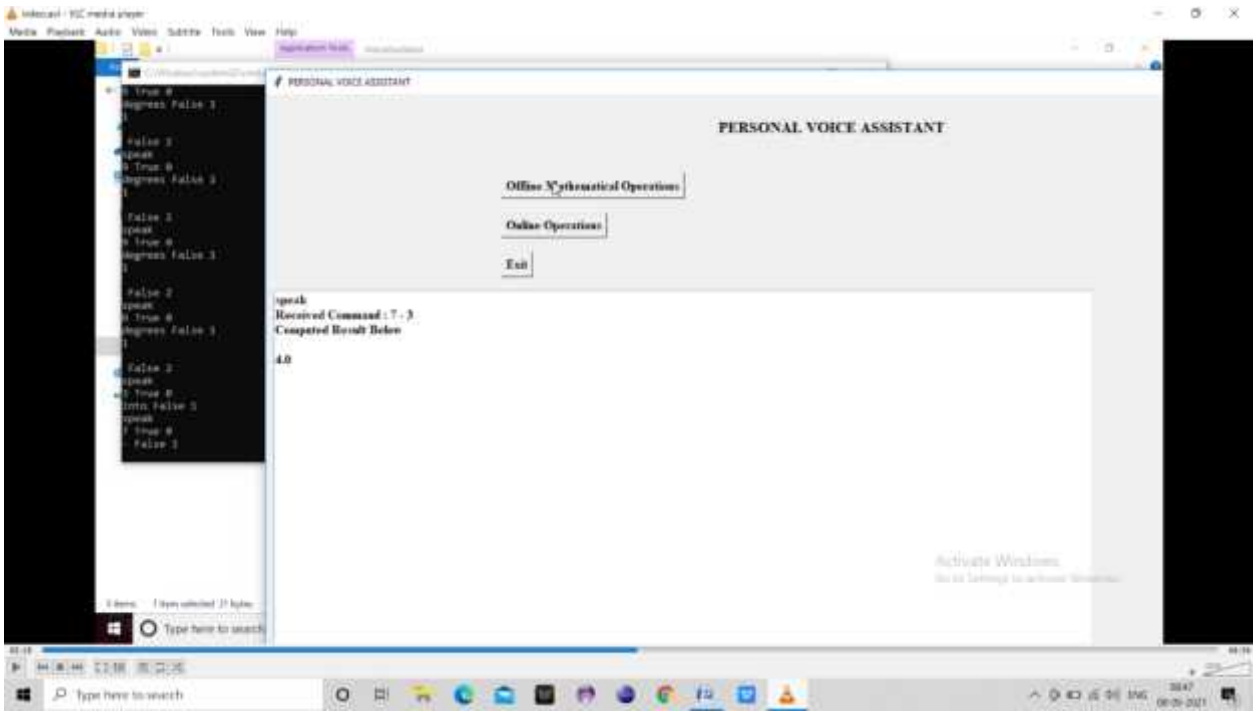


Fig 8.2.4 result of command 7-3

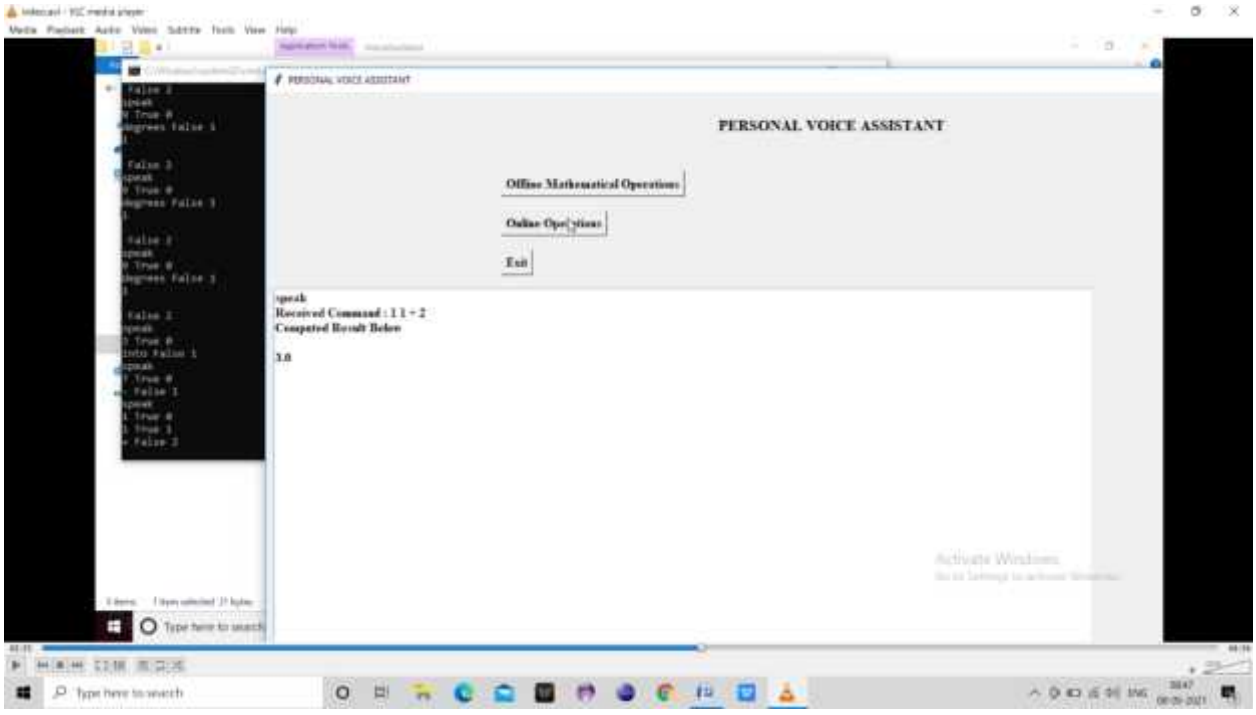


Fig 8.2.5 result of command 11+2

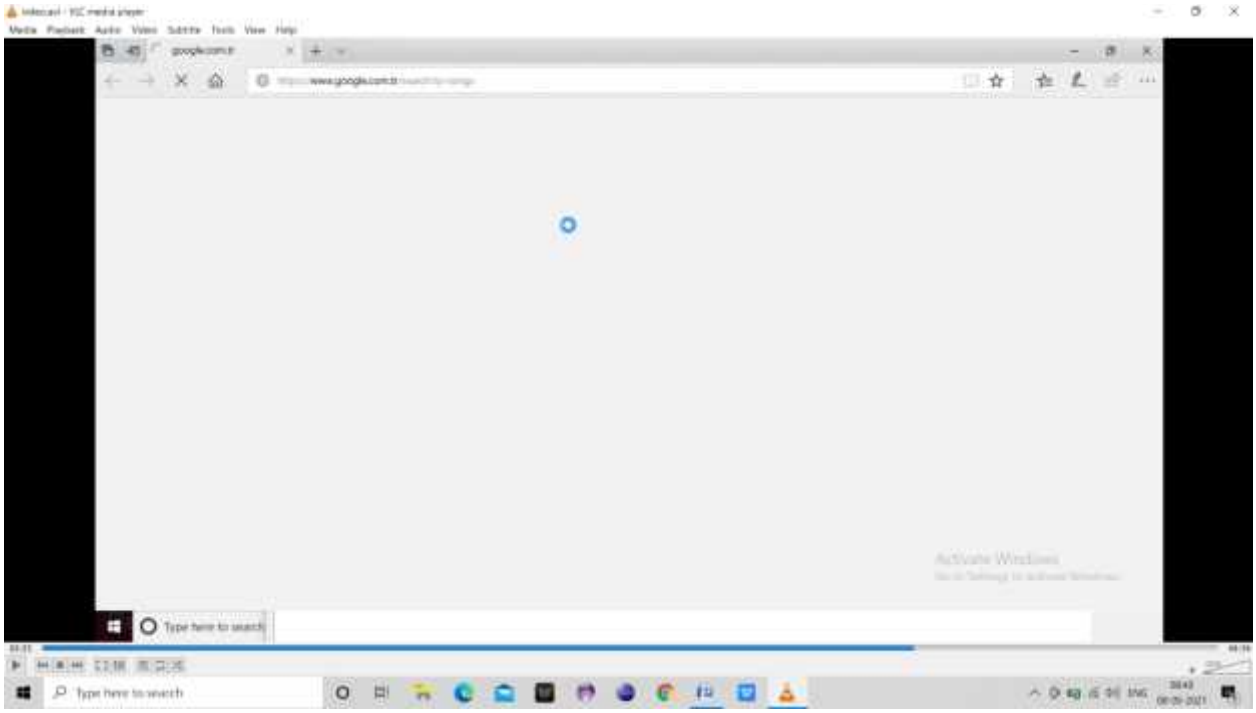


Fig 8.2.6 result of command search

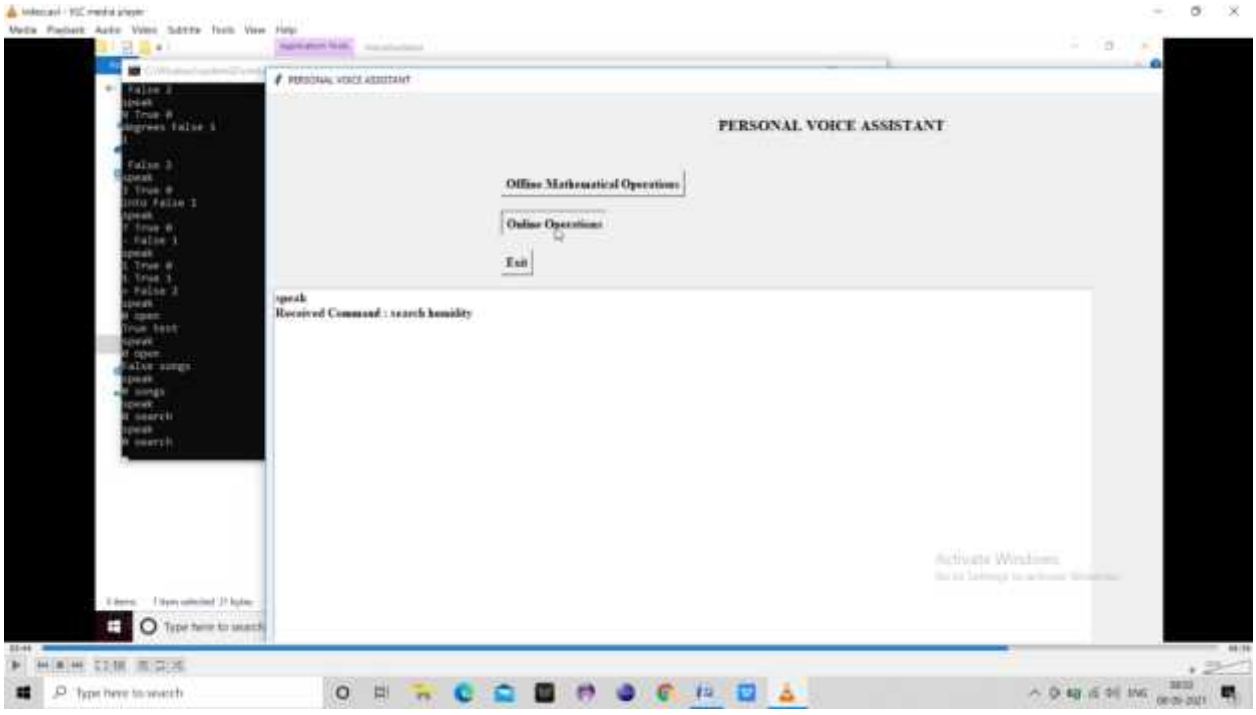


Fig 8.2.7 gave the command of search humidity

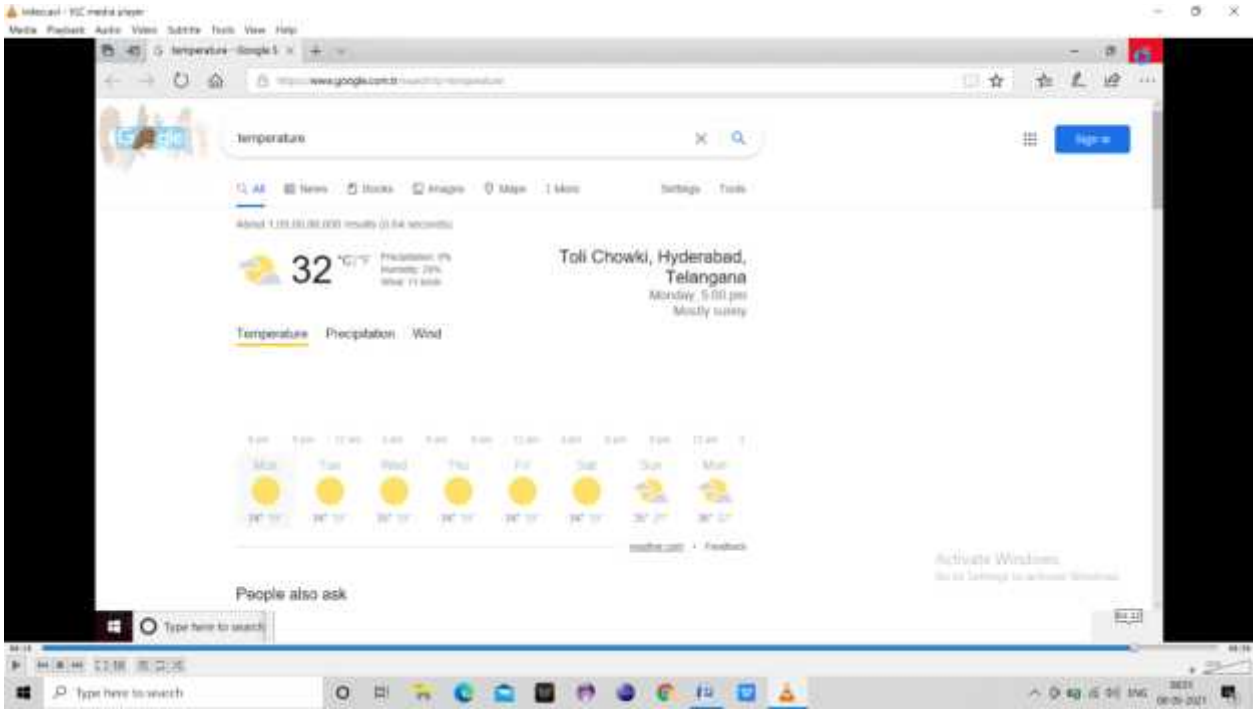


Fig 8.2.8 result of command search humidity

9.EXPERIMENTALS RESULTS

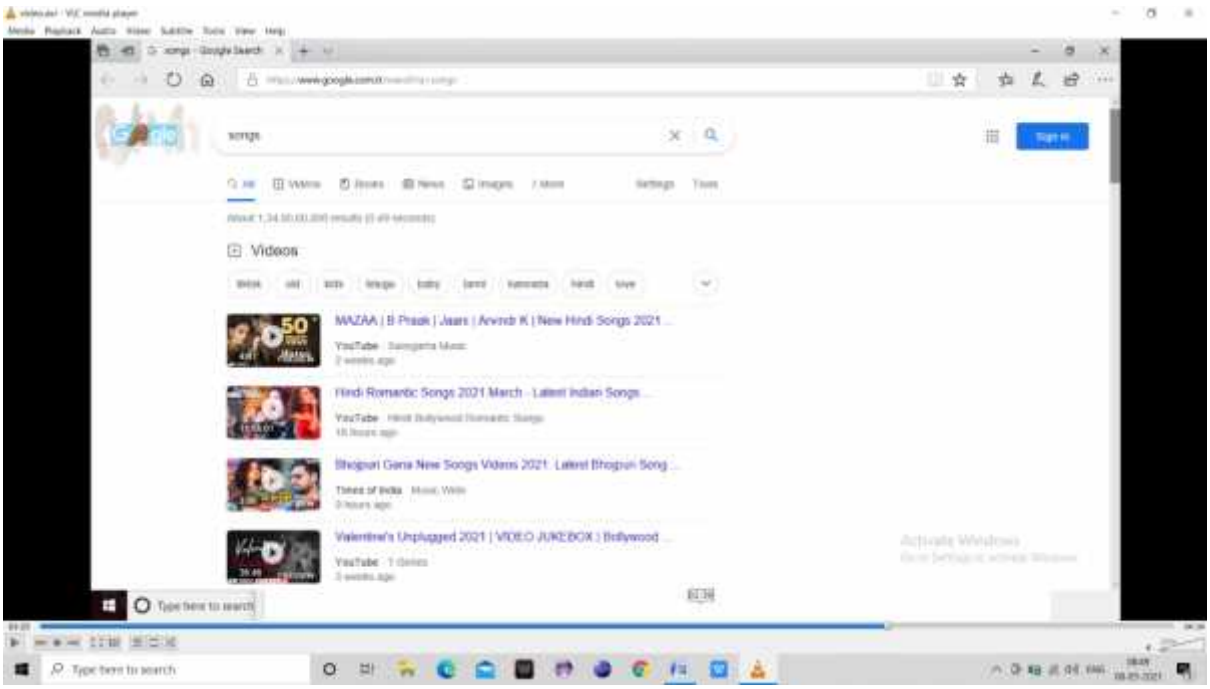


Fig 9.1 result of the command search songs

10.CONCLUSIONAND FUTURE ENHANCEMENT

CONCLUSION

This system is designed in such a method wherein the user can accommodate to it effortlessly. Our proposed system personal voice assistant can be implemented using speech recognition module thus makes the system more secure and robust. The contributions of Smart Voice Assistant are twofold. It is the voice control application that provides enhancements to all applications running on a system by synthesizing commands set from onscreen context. Personal voice assistant can benefit large number of users with universal eyes free and hands free voice control of their system.The advantage of voice commands over multi-touch when interacting with a screen non-visually is that it does not require targets to be located and thus avoids the problems with pointing, it saves time and reading of News can be possible by the blind people. This can do variety of tasks like, do calculation, updates about the stock and the endless tasks for the user. Thus making one's life comfortable and at the same time remotely accessible via voice commands.

FUTURE ENHANCEMENT

1.In future we will add various other languages.

2. Focus on security.

3. Acknowledging with devices or electronics.

4. Adding of face recognition feature.

5.

REFERENCES

[1] Yash Mittal, Pradhi Toshniwal "A voice-controlled multifunctional Smart Home Automation System", 2015 India Conference (INDICON), IEEE.

[2] Prerna Wadikar, Nidhi Sargar, Rahool Rangnekar, Prof.Pankaj Kunekar "Home Automation using Voice Commands in the Hindi Language", 2020, IRJET.

[3] Nagesh Singh Chauhan,"Build Your First Voice Assistant",March,2019 https://towardsdatascience.com/build-your-first-voiceassistant.

[4] Manohar Swamynathan, “Mastering Machine Learning with Python”, Karnataka: Apress, 2017.

[5] Yu Zhong, T. V. Raman, Casey Burkhardt, FadiBiadsy and Jerey P. Bigham, "JustSpeak: Enabling Universal Voice Control on Android", April 2014.

[6] Sukhada Chokkadi, Sannidhan MS, Sudeepa K B, Abhir Bhandary “A Study on various state of the art of the Art Face Recognition System using Deep Learning Techniques”, 2019, IJATCSE.

[7] Steve Joseph, Chetan Jha, Dipesh Jain, Saurabh Gavali, Manish Salvi “Voice based E-Mail for the Blind”, 2020, IRJET.

[8] Dongmahn SEO, Suhyun KIM, Gyuwon SONG, Seung-gil, "Speech-to-Text-based Life Log System for Smartphones”, IEEE International Conference on Consumer Electronics (ICCE), 2014.

[9] Aditi Bhalerao, Samira Bhilare, Anagha Bondade, Monal Shingade, Aradhana Deshmukh, “Smart Voice Assistant: a universal voice control solution for non-visual access to the Android operating system”, Volume: 04, Issue: 01, Jan-2017.

[10] Chen-Yen Peng and Rung-Chin Chen “Voice Recognition by Google Home and Raspberry Pi for smart Socket Control” 2018 Tenth International Conference on Advanced Computational Intelligence (ICACI) March 29– 31, 2018, Xiamen, China.

PUBLICATIONS

JOURNAL (UGC approved Journal)

CONFERENCE (International Conference on “Innovations in Computers Networks, Computational Intelligence and IOT” [ICICCI-21]).

PAPER ID : ICICCI-21-0136

TOPIC : HAND GESTURE RECOGNITION USING CONVOLUTION NEURAL NETWORK.

STUDENTS PROFILE:

Marupalli Aashritha Goud is currently pursuing her Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. She completed her intermediate from Narayana Junior College and 10th class from St. Andrew's High School. She is one of the members of Coders Club in our college. Her responsibilities in that group include mentoring and motivating students to take coding as a serious hobby. Her technical skills include C, Python. She also has a basic understanding of C++ and Java. She is also a student of Smart Interviews. Her participations include: National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing"which was conducted from 10th to 12th December 2020, "Know More - Teach More ", the Global Webinar on Cloud & Big Data – Changing the way we work which was conducted Global Education & Careers Forum(GECF) on 12th August 2020, Women online workshop on "Women in Cyber Security and Privacy in 2020" which was conducted from 6th to 10th July 2020, "Know More - Teach More ", the Global Webinar on Cyber Threats and Defense Techniques conducted by GECF on 22nd July 2020, "One Day Webinar on Internet of Things and Its Applications" conducted by Anand Institute of Higher Technology on 21st May 2020 and IIC Online Sessions conducted by Institution's Innovation Council (IIC) of MHRD's Innovation Cell, New Delhi to promote Innovation, IPR, Entrepreneurship, and Start-ups among HEIs from 28thApril to 22nd May 2020. Her areas of interest are Python,Machine Learning and building serverless applications. She completed few certification courses from online platforms like Coursera and currently working on Full Stack Developer in Udemy.

Gariboyina Suraj is currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. He completed his intermediate in narayana junior college, and 10th class from Sindhu Vidyalayam High School. His technical skills include C, Python and Java. He also has a basic understanding of C++. He took part in Employability Skill development Program conducted by Zensar. He is also a student of Smart Interviews. His participations include: National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020. His areas of interest are Python, Artificial Intelligence, Machine Learning and Deep Learning. He completed few certification courses from online platforms like Coursera, CursaApp and SoloLearn.He also completed the internship in python at hexnbit .completed frontend fundamentals in pirple.com.

Akkapathri sanjay kumar is currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin’s Engineering College. He completed his intermediate in sr.junior college and 10th class from narendra high school nizamabad. His technical skills include C, Python and Java. He also has a basic understanding of C++. He took part in Employability Skill development Program conducted by Zensar. He is also a student of Smart Interviews. His participations include: National Level Three Day Online Workshop on “AI & ML in Speech and Audio Processing” which was conducted from 10th to 12th December 2020. His areas of interest are Python, Artificial Intelligence, Machine Learning and Deep Learning. He completed few certification courses from online platforms like Coursera, CursaApp and SoloLearn.

Srichandana Julakanti is currently pursuing her Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin’s Engineering College. She completed her intermediate from Narayana Junior College and 10th class from Kakatiya High School. Her technical skills include C, Java,C++. She also has a basic understanding of Spring framework and also front end technologies like React. She is also a student of Smart Interviews. Her participations include: National Level Three Day Online Workshop on “AI & ML in Speech and Audio Processing” which was conducted from 10th to 12th December 2020, “Know More - Teach More “, the Global Webinar on Cloud & Big Data – Changing the way we work which was conducted Global Education & Careers Forum (GECF) on 12th August 2020, Women online workshop on “Women in Cyber Security and Privacy in 2020” which was conducted from 6th to 10th July 2020, “Know More - Teach More “, the Global Webinar on Cyber Threats and Defense Techniques conducted by GECF on 22nd July 2020, “One Day Webinar on Internet of Things and Its Applications” conducted by Anand Institute of Higher Technology on 21st May 2020 and IIC Online Sessions conducted by Institution's Innovation Council (IIC) of MHRD's Innovation Cell, New Delhi to promote Innovation, IPR, Entrepreneurship, and Start-ups among HEIs from 28th April to 22nd May 2020. Her areas of interest are Java , developing Web applications using Spring Framework and React as Front End Technology. She completed few certification courses from online platforms like Coursera and currently working on React js. in Udemy.

APPENDICES

```
import os
import subprocess
import webbrowser
from tkinter import *

import speech_recognition as sr
from gtts import gTTS
from playsound import playsound

main = Tk()
main.title("PERSONAL VOICE ASSISTANT")
main.geometry("1300x1200")

def play():
data = text.get("1.0",END)
print("test "+data)
t1 = gTTS(text=data, lang='en', slow=False)
t1.save("output.mp3")
playsound("output.mp3")

def runOffline(data):
data = data.lower()
current = 0
total = 0
count = 0
arr = data.split(" ")
i = 0
while i < len(arr):
print(str(arr[i])+" "+str(arr[i].isnumeric())+" "+str(i))
if arr[i].isnumeric():
current = float(arr[i])
count = count + 1
if total == 0:
total = current
#play()
if arr[i] == 'into' or arr[i] == '*':
i = i + 1
current = float(arr[i])
total = total * current
#play()
if arr[i] == 'plus' or arr[i] == '+':
i = i + 1
```

```
current = float(arr[i])
total = total + current
#play()
if arr[i] == 'minus' or arr[i] == '-':
i = i + 1
current = float(arr[i])
total = total - current
#play()
if arr[i] == 'divide':
total = total / current
#play()
if arr[i] == 'percentage':
i = i + 1
current = float(arr[i])
total = total / count
#play()
if arr[i] == 'power':
i = i + 1
current = float(arr[i])
total = pow(total,current)
#play()
if arr[i] == 'root':
i = i + 1
current = float(arr[i])
total = total ** count
#play()
if arr[i] == 'increase':
i = i + 1
current = float(arr[i])
total = total + 1
#play()
if arr[i] == 'decrease':
i = i + 1
current = float(arr[i])
total = total - 1
#play()
if arr[i] == 'table':
i = i + 2
current = int(arr[i])
        for k in range(1,11):
text.insert(END,str(current)+" * "+str(k)+" = "+str(current * k)+"\n")
#play()
i = i + 1
return total
```

```
def offline():
text.delete('1.0', END)
r = sr.Recognizer()
mic = sr.Microphone()
with mic as source:
text.insert(END,"speak\n")
text.update_idletasks()
print('speak')
r.adjust_for_ambient_noise(source)
audio = r.record(source,duration=5)
data = r.recognize_google(audio)
text.insert(END,"Received Command : "+data+"\n")
text.update_idletasks()
total = runOffline(data+"\n\n")
text.insert(END,"Computed Result Below\n\n"+str(total)+"\n")
text.update_idletasks()
play()



def runOnline(command):
data = command.lower()
data = data.strip("\n")
data = data.strip()
arr = data.split(" ")
i = 0
while i < len(arr):
print(str(i)+" "+arr[i])
if arr[i].strip("\n").strip() == 'search':
i = i + 1
url = "https://www.google.com.tr/search?q={}".format(arr[i])
webbrowser.open_new_tab(url)
if arr[i].strip("\n").strip() == 'settings':
subprocess.Popen([r"C:/Windows/System32/DpiScaling.exe"])
if arr[i].strip("\n").strip() == 'wi-fi':
i = i + 1
if arr[i].strip("\n").strip() == 'on':
os.system("netsh interface set interface 'Wifi' enabled")
if arr[i].strip("\n").strip() == 'off':
os.system("netsh interface set interface 'Wifi' disabled")
if arr[i].strip("\n").strip() == 'open':
i = i + 1
```

```
name = arr[i]
name = name.strip("\n")
name = name.strip()
print(str(os.path.exists('E:/'+name+".txt"))+" "+str(name))
if os.path.exists('E:/'+name+".txt"):
os.system("start " + 'E:/'+name+".txt")
if os.path.exists('E:/'+name+".doc"):
os.system("start " + 'E:/'+name+".doc")
if os.path.exists('E:/'+name+".docx"):
os.system("start " + 'E:/'+name+".docx")
if os.path.exists('E:/'+name+".pdf"):
os.system("start " + 'E:/'+name+".pdf")
if os.path.exists('E:/'+name+".jpg"):
os.system("start " + 'E:/'+name+".jpg")
if os.path.exists('E:/'+name+".png"):
os.system("start " + 'E:/'+name+".png")
if os.path.exists('E:/'+name+".gif"):
os.system("start " + 'E:/'+name+".gif")
i = i + 1


def online():
text.delete('1.0', END)
r = sr.Recognizer()
mic = sr.Microphone()
with mic as source:
text.insert(END,"speak\n")
text.update_idletasks()
print('speak')
r.adjust_for_ambient_noise(source)
audio = r.record(source,duration=5)
data = r.recognize_google(audio)
text.insert(END,"Received Command : "+data+"\n")
text.update_idletasks()
runOnline(data+"\n\n")
#play()



def close():
 main.destroy()


font = ('times', 15, 'bold')
```

```
title = Label(main, text='PERSONAL VOICE ASSISTANT')
#title.config(bg='powder blue', fg='olive drab')
title.config(font=font)
title.config(height=3, width=120)
title.place(x=0,y=5)

font1 = ('times', 13, 'bold')
ff = ('times', 12, 'bold')

offlineButton = Button(main, text="Offline Mathematical Operations", command=offline)
offlineButton.place(x=300,y=100)
offlineButton.config(font=ff)


onlineButton = Button(main, text="Online Operations", command=online)
onlineButton.place(x=300,y=150)
onlineButton.config(font=ff)

exitButton = Button(main, text="Exit", command=close)
exitButton.place(x=300,y=200)
exitButton.config(font=ff)


font1 = ('times', 12, 'bold')
text=Text(main,height=25,width=130)
scroll=Scrollbar(text)
text.configure(yscrollcommand=scroll.set)
text.place(x=10,y=250)
text.config(font=font1)

main.config()
main.mainloop()
```

A

PROJECT REPORT

On

MISSING CHILD IDENTIFICATION SYSTEM USING DEEP LEARNING AND MULTICLASS SVM

Submitted by

1) R.Sreeharipriya(17K81A05M9) 2) Nishika Lewis (17K81A05M1)
3) G.Rahul (17K81A05K8) 4) S.Sreeja Reddy (17K81A05N1)

in partial fulfillment for the award of the

degree of

BACHELOR OF TECHNOLOGY

IN

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Under the Guidance of

Ms.S.Swetha
Assistant Professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ST.MARTIN'S ENGINEERING COLLEGE
An Autonomous Institute

Dhulapally, Secunderabad – 500100

JUNE 2021

BONAFIDE CERTIFICATE

This is to certify that the project entitled "Missing Child Identification System Using Deep Learning and Multiclass SVM", is being submitted by **R.Sreeharipriya (17K81A05M9), Nishika Divya Lewis (17K81A05M1), G.Rahul (17K81A05K8) and S.Sreeja Reddy (17K81A05N1)** in partial fulfillment of the requirement for the award of the degree of **BACHELOR OF TECHNOLOGY IN COMPUTER SCIENCE AND ENGINEERING** is recorded of bonafide work carried out by them. The results embodied in this report have been verified and found satisfactory.

Assistant Professor

Ms.S.SWETHA

Department of CSE

Head of the Department

Dr.M.NARAYANAN

Department of CSE

Internal Examiner

External Examiner

Place:

Date:

DECLARATION

We, the students of **Bachelor of Technology** in the Department of **'Computer Science and Engineering'**, session: 2017 – 2021, St. Martin's Engineering College, Dhulapally, Kompally, Secunderabad, hereby declare that work presented in this Project Work entitled Missing Child Identification System Using Deep Learning and Multiclass SVM is the outcome of our own bonafide work and is correct to the best of our knowledge and this work has been undertaken taking care of Engineering Ethics. This result embodied in this project report has not been submitted in any university for award of any degree.

R.Sreehariprіya (17K81A05M9)
Nishika Divya Lewis (17K81A05M1)
G.Rahul (17K81A05K8)
S.Sreeja Reddy (17K81A05N1)

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible and whose encouragements and guidance have crowded effects with success.

We extended our deep sense of gratitude to Principal**, Dr. P. SANTOSH KUMAR PATRA**, St. Martin's Engineering College, Dhulapally, for permitting us to undertake this project.

We are also thankful to **Dr. M.NARAYANAN**, Head of the Department, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

We would like to thank **Dr. T. POONGOTHAI,** Department of Computer Science and Engineering (AI & ML) for her encouragement and insightful comments and as well as our project coordinators **Dr. B.RAJALINGAM**, Associate Professor and **Mr. J.SUDHAKAR**, Assistant Professor, in Department of Computer Science and Engineering for their valuable support.

We would like to express our sincere gratitude and indebtedness to our project supervisor Ms.S.Swetha, Assistant Professor, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for her support and guidance throughout our project.

Finally, we express thanks to all those who have helped us successfully to completing this project. Furthermore, we would like to thank our family and friends for their moral support and encouragement.

We express thanks to all those who have helped us in successfully completing the project.

R.Sreehariprіya (17K81A05M9)
Nishika Divya Lewis (17K81A05M1)
G.Rahul (17K81A05K8)
S.Sreeja Reddy (17K81A05N1)

ABSTRACT

Every year, many children are reported missing in India. A large number of children remain unidentified in missing child cases. This paper describes a novel application of deep learning methodology for identifying the reported missing child from photos of a large number of children available, using face recognition. The public can upload photographs of suspicious children to a common portal, along with landmarks and comments. The photo will be automatically compared to the missing child's registered photos in the repository. The input child image is classified, and the photo with the best match is chosen from a database of missing children. Using the facial image uploaded by the public, a deep learning model is trained to correctly identify the missing child from the missing child image database. For face recognition, the Convolutional Neural Network (CNN), a highly effective deep learning technique for image-based applications, is used. A pre-trained CNN model VGG-Face deep architecture is used to extract face descriptors from images. In contrast to traditional deep learning applications, our algorithm only employs a convolution network as a high-level feature extractor, with child recognition handled by a trained SVM classifier. Using the best performing CNN model for face recognition, VGG-Face, and properly training it results in a deep learning model that is insensitive to noise, illumination, contrast, occlusion, image pose, and child age, and it outperforms previous methods in face recognition-based missing child identification. The classification performance of the child identification system is 99.41%. It was tested on 43 children.

TABLE OF CONTENTS

LIST OF TABLES

Table 1. List of Tables

LIST OF FIGURES

Table 2. List of Figures

LIST OF OUTPUT SCREENS

Table 3. List of output screens

LIST OF ABBREVIATIONS

CNN	Convolutional Neural Networks
SVM	Support Vetor Machine
UML	Unified Modelling Language
GUI	Graphical User Interface
FIR	First Information Report
CLF	Common Loudspeaker Format
COTS	Commercial Off-The-Shelf
HTML	Hyper Text Markup Language
CSS	Cascading Style Sheets
VGG	Visual Geometry Group
SVC	Support Vector Classifier
DCNN	Deep Convolutional Neural Networks
MHA	Ministry of Home Affairs
RAM	Random Access Memory
CPU	Central Processing Unit
ER	Entity-Relationship
DFD	Data Flow Diagram

Table 4. List of abbreviations

GLOSSARY OF TERMS

TERM	MEANING
Deep Learning	Deep learning is part of a broader family of machine learning methods based on artificial neural networks with representation learning.
Convolutional Neural Network	In deep learning, a convolutional neural network is a class of deep neural network, most commonly applied to analyse visual imagery.
Support Vector Machine	Support vector machines (SVMs) are particular linear classifiers which are based on the margin maximization principle.
Visual Geometry Group - 16	It usually refers to a deep convolutional network for object recognition developed and trained by Oxford's renowned Visual Geometry Group (VGG).
Phenotype	The term "phenotype" refers to the observable physical properties of an organism.
Prototype	First or preliminary version of a device or vehicle from which other forms are developed.
Occlusion	Blockage or closing.
Repository	A central location in which data is stored and managed.

Table 5. Glossary of Terms

1. INTRODUCTION

Children are a country's most valuable asset. The future of any country is dependent on the proper education of its children. India is the world's second most populous country, with children constituting a sizable proportion of the total population. However, many children go missing in India each year for a variety of reasons, including abduction or kidnapping, runaway children, trafficked children, and lost children. A deeply disturbing fact about India's missing children is that, while 174 children go missing on average every day, half of them go untraced. Children who go missing are susceptible to being exploited and abused for a range of reasons. According to a National Crime Records Bureau (NCRB) report cited by the Ministry of Home Affairs (MHA) in Parliament (LS Q no. 3928, 20-03-2018), more than one lakh children (1,11,569 in actual numbers) were reported missing until 2016, with 55,625 of them remaining untraced at the end of the year. Many non-governmental organizations claim that the number of missing children is much higher than reported. The majority of missing child cases are reported to police. For various reasons, a child missing in one region may be found in another region or state. Even if a child is found, he or she may be difficult to identify among the reported missing cases. This paper describes a framework and methodology for developing an assistive tool for locating a missing child. A concept for maintaining a virtual space is proposed, in which recent photographs of children provided by parents when reporting missing cases are saved in a repository. The public is encouraged to take photographs of children in suspicious situations and upload them to the portal. The application will include an automatic search for this photo among the missing child case images. This assists police officers in locating the child anywhere in India. When a child is found, the photograph taken at the time is compared to the images uploaded by the police/guardian at the time the child went missing. In some cases, the child has been missing for an extended period of time. This age difference is reflected in the images because ageing changes the shape of the face and the texture of the skin. It is necessary to develop a feature discriminator that is resistant to ageing effects. When compared to other face recognition systems, this is the most difficult challenge in missing child identification. Also, a child's facial appearance can change due to changes in pose, orientation, illumination, occlusions, background noise, and so on. The image taken by the public may be of poor quality, as some of them may have been taken from a distance without the child's knowledge.

1.1 PROJECT OVERVIEW

We propose a methodology for missing child identification which combines facial feature extraction based on deep learning and matching based on CNN and SVM. The proposed system utilizes face recognition for missing child identification.

It consists of a national portal for storing details of missing child along with the photo. Whenever a child missing is reported, the photo of the missing child is put into the portal. Public can search for any matching child in the database for the images with them. The system will prompt the most matching cases. Once the matching is found, the officer can get the details of the child.

1.2 PROJECT OBJECTIVE

The objective of this project is to use facial feature extraction to identify reported missing children from photographs of many youngsters. This project aims to make it easier for anyone to report the details of a missing child without having to go through the lengthy process of filing a FIR with the local police station.

The central focus of this project is to match the facial features of a child who has been reported missing for a long time with the picture present in the repository.

1.3 ORGANIZATION OF CHAPTERS

This documentation consists of 10 different chapter and they are:

1. Introduction – This chapter covers the overview of our project and its objectives.
2. Literature Survey – This includes the details of our survey.
3. Software and Hardware Requirements – We specify our software and hardware requirements here.
4. Software Development Analysis – This section includes the problem definition and details of the modules we used in our project.
5. Project System Design – This chapter includes the design part of our project which includes UML diagrams.
6. Project Coding – This section contains the details of our project code.
7. Project Testing – The details of test cases and testing are included in this chapter.
8. Output Screens – This contains the screenshots of how our project looks like when executed.
9. Experimental Results – This chapter contains the screenshots of our results.
10. Conclusion and Future Enhancements – This covers the conclusion of our project and the possible future developments.

2. LITERATURE SURVEY

A literature survey or a literature review in a project report is that section which shows the various analysis and research made in the field of your interest and the results already published, considering the various parameters of the project and the extent of the project.

It is the most important part of our report as it gave us a direction in our research. It helped us set a goal for our analysis - thus giving us our problem statement.

2.1 SURVEY ON BACKGROUND

1) Missing People Detection System

Authors : Aryan Patel, Dhru Prajapati, Dimple Jadhav, Mudra Doshi

Their proposed framework utilizes Face Recognition for missing individuals' recognizable proof. Here the general public or police who finds a suspicious person (child, mentally challenged person, etc.) on the road uploads an image of that person into the portal. Their system extracts the face encodings of the image and compare with that of the face encodings of the previously existing images in the database. If a match is found, an alert message/notification will be sent to the user. If a match is not found, then the person will be provided with the option of registering that face as a new entry to their database. Whenever public upload a picture, the face encodings of the image are extracted then compared to the face encodings of the pictures stored within the database. The user is notified that a match is found alongside the image from the database that matched with the uploaded picture.

2) The Lore of speculation and analysis using machine learning and image matching

Authors : K. Bharath, Paithankar Sumit, S. Amudha

Face patterns are generated using Histogram of Oriented Gradients (HOG) algorithm. The images are made black and white. Here, the part of the images that looks more like the original HOG face pattern is found. Finally, the detected face is bounded by a bounding box. Sixty-eight specific points (landmarks) that are existing on every face are figured out by using the face landmark estimation algorithm. From the landmarks found, image transformations like scaling, shearing and rotation are used by the OpenCV’s affine transformation to make the lips and eyes appear in the same location on every image. The face images are then passed through deep convolutional neural network. By doing

this, we obtain 128 measurements which are 128- dimension hyper sphere. And no one knows which parts of the face the 128 measurements representing. All we know is that the network outputs the same 128 numbers for two different images of the same person. A linear SVM classifier is used to recognize the face. The classifier has been trained in such a way that it can take the measurements from a test image and gives the closest match as output.

3) LBPH-based Enhanced Real-Time Face Recognition

Authors : Farah Deeba, Aftab Ahmed

They developed a facial recognition system based on the Local Binary Pattern Histogram (LBPH) method to treat the real-time recognition of the human face in the low and high-level images. We aspire to maximize the variation that is relevant to facial expression and open edges so to sort of encode edges in a very cheap way. These highly successful features are called the Local Binary Pattern Histogram (LBPH).

4) Face Recognition System Based on LBPH Algorithm

Authors : Abhishek Pratap Singh, SunilKumar S Manvi, Pratik Nimbal, Gopal Krishna Shyam

In the proposed face recognition system, they used Local Binary Patterns Histogram algorithm for recognizing faces. The whole procedure is divided into three major components, i.e. detection of faces, facial feature extraction, and classification of the image. The Face detection process describes the face of a person in input image. In feature extraction, facial landmarks are extracted and to make an LBPH histogram that gives the completely unique result and then in recognition process the histogram of the input image is compared with database histogram using the classifier. The result shows that the system can recognize a known and unknown person.

5) Additive angular margin loss for deep face recognition.

Authors : Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou.

Arcface

The proposed ArcFace has a clear geometric interpretation due to the exact correspondence to the geodesic distance on the hypersphere. They present arguably the most extensive experimental evaluation of all the recent state-of-the-art face recognition methods on over 10 face recognition

benchmarks including a new large-scale image database with trillion level of pairs and a large-scale video dataset. They show that ArcFace consistently outperforms the state-of-the-art and can be easily implemented with negligible computational overhead. They release all refined training data, training codes, pre-trained models and training logs, which helped reproduce the results.

6) Efficient Face Recognition System for Identifying Lost People

Authors : Bharath Darshan Balar, D S Kavya, Chandana M, Anush E, Vishwanath R Hulipalled

A Flask API interface accompanies their model to give the user a better user experience. When the user opens their application they will be asked to upload a image of the missing person. If a match is found, they will be provided with the image and details about the match. If match is not found, they will be asked if they want to register that image as a new entry into their database. If they wish to register, they will be asked to enter the details about the image.

7) Face recognition performance under aging.

Authors : Debayan Deb, Lacey Best-Rowden, and Anil K Jain

They fit multilevel statistical models to genuine comparison scores (similarity between images of the same face) from the two COTS face matchers. This allowed them to analyze the degradation in recognition performance due to elapsed time between a probe (query) and its enro llment (gallery) image. They account for face image quality to obtain a better estimate of trends due to aging, and analyze whether longitudinal trends in genuine scores differ by subject gender and race. Based on the results of their statistical model, they infer that the state-of-the-art COTS matchers can verify 99% of the subjects at a false accept rate (FAR) of 0.01% for up to 10.5 and 8.5 years of elapsed time. Beyond this time lapse of 8.5 years, there is a significant loss in face recognition accuracy. This study extends and confirms the findings of earlier longitudinal studies on face recognition.

8) Ongoing Face Recognition Vendor Test

Authors : Patrick J. Grother, Mei Ngan, and Kayee Hanaoka

The algorithms implement one-to-many identification of faces appearing in two- dimensional images. Three datasets were used - the primary dataset is comprised of 26.6 million reasonably well-controlled

live portrait photos of 12.3 million individuals. Three smaller datasets containing more unconstrained photos are also used: 200 thousand side-view images; 3.2 million webcam images; 2.5 million photojournalism and amateur photographer photos. The report will be useful for comparison of face recognition algorithms and assessment of absolute capability.

9) Longitudinal study of child face recognition

Authors : Debayan Deb, Neeta Nain, and Anil K Jain

Multilevel statistical models are fit to genuine comparison scores from the CLF dataset to determine the decrease in face recognition accuracy over time. Additionally, they analyze both the verification and open-set identification accuracies in order to evaluate state-of-the-art face recognition technology for tracing and identifying children lost at a young age as victims of child trafficking or abduction.

10) Age estimation guided convolutional neural network for age-invariant face recognition

Authors : Tianyue Zheng, Weihong Deng, and Jiani Hu

They propose a novel deep face recognition network called age estimation guided convolutional neural network (AE-CNN) to separate the variations caused by aging from the person specific features which are stable. The carefully designed CNN model can learn age-invariant features for face recognition.

2.2 CONCLUSIONS ON SURVEY

These works provide basic background information about various techniques and algorithms in deep learning and machine learning in order to improve face recognition, face features detection, feature comparison and feature matching. They use multiple algorithms like ArcFace, LBPH, CNN, COTS, etc under various contexts. Deep learning and Machine Learning algorithms are the major part of image recognition involved projects since they are the fundamental aspect to these projects. Without effective and necessary knowledge about these concepts in detail, it would've been extremely difficult for us to choose and apply the required algorithm to make our project's output reliable. Every reference gave an example to how to use deep learning and machine learning to work on our project and produce the existing project.

3. SOFTWARE AND HARDWARE REQUIREMENTS

Requirement is a condition or capability possessed by the software or system component in order to solve a real world problem. The problems can be to automate a part of a system, to correct shortcomings of an existing system, to control a device, and so on.

Requirements describe how a system should act, appear or perform. For this, when users request for software, they provide an approximation of what the new system should be capable of doing. Requirements differ from one user to another and from one business process to another.

The purpose of the requirements document is to provide a basis for the mutual understanding between the users and the designers of the initial definition of the software development life cycle (SDLC) including the requirements, operating environment and development plan.

Requirements help to understand the behavior of a system, which is described by various tasks of the system. For example, some of the tasks of a system are to provide a response to input values, determine the state of data objects, and so on. Note that requirements are considered prior to the development of the software. The requirements, which are commonly considered, are classified into three categories, namely, functional requirements, non-functional requirements, and domain requirements.

The functional requirements should be complete and consistent. Completeness implies that all the user requirements are defined. Consistency implies that all requirements are specified clearly without any contradictory definition. Generally, it is observed that completeness and consistency cannot be achieved in large software or in a complex system due to the problems that arise while defining the functional requirements of these systems. The different needs of stakeholders also prevent the achievement of completeness and consistency. Due to these reasons, requirements may not be obvious when they are,'first specified and may further lead to inconsistencies in the requirements specification.

The non-functional requirements (also known as **quality requirements)** are related to system attributes such as reliability and response time. Non-functional requirements arise due to user requirements, budget constraints, organizational policies, and so on. These requirements are not related directly to any particular function provided by the system.

Non-functional requirements should be accomplished in software to make it perform efficiently. For example, if an aeroplane is unable to fulfill reliability requirements, it is not approved for safe operation. Similarly, if a real time control system is ineffective in accomplishing non-functional requirements, the control functions cannot operate correctly.

System requirements are the configuration that a system must have in order for a hardware or software application to run smoothly and efficiently. Failure to meet these requirements can result in installation problems or performance problems. The former may prevent a device or application from getting installed, whereas the latter may cause a product to malfunction or perform below expectation or even to hang or crash.

System requirements are also known as minimum system requirements.

Hardware system requirements often specify the operating system version, processor type, memory size, available disk space and additional peripherals, if any, needed. Software system requirements, in addition to the requirements, may also specify additional software dependencies (e.g., libraries, driver version, framework version). Some hardware/software manufacturers provide an upgrade assistant program that users can download and run to determine whether their system meets a product's requirements.

Some products include both minimum and recommended system requirements. A video game, for instance, may function with the minimum required CPU and GPU, but it will perform better with the recommended hardware. A more powerful processor and graphics card may produce improved graphics and faster frame rates (FPS).
Some system requirements are not flexible, such as the operating system(s) and disk space required for software installation. Others, such as CPU, GPU, and RAM requirements may vary significantly between the minimum and recommended requirements. When buying or upgrading a software program, it is often wise to make sure your system has close to the recommended requirements to ensure a good user experience.

3.1 SOFTWARE REQUIREMENTS

- Operating System: Windows 10.
- Platform: PYTHON TECHNOLOGY - PyCharm 2020.1.1.
- Front End: Python, HTML, CSS.
- Back End: MYSQL.

3.2 HARDWARE REQUIREMENTS

- System: Intel(R) Core (TM) i7-8550U CPU @ 1.80GHz   1.99 GHz.
- Hard Disk: 1 TB.
- Monitor: Intel(R) UHD Graphics 620.
- Ram: 8 GB.

4. SOFTWARE DEVELOPMENT ANALYSIS

Software development is a process of writing and maintaining the source code, but in a broader sense, it includes all that is involved between the conception of the desired software through to the final manifestation of the software, sometimes in a planned and structured process. Therefore, software development may include research, new development, prototyping, modification, reuse, re-engineering, maintenance, or any other activities that result in software products.

4.1 OVERVIEW OF THE PROBLEM

There is lack of coordination in departments dealing with issues concerning children and timeliness for the police department in handling this process. The fact that a child can be housed in a children's home for quite a while without the care givers' knowledge even from searching from different police stations shows lack of convergence among the agencies involved in child protection. Based on these two issues, we can conclude that there is no effective system in place to report missing children and informing the parents and/or relatives whenever the child is found.

There is a need for proactive support from various sectors to address the issue of missing children in a systematic way. Effective mechanisms should be established to ensure coordination and information sharing. Tracing the missing children becomes nearly impossible when there is no database to capture anddisseminate comprehensive information for the children. Therefore, a solution that can help the child welfare agencies to efficiently trace missing children and promptly communicating to their care givers is needed.

4.2 DEFINING THE PROBLEM

Mostly missing child cases are reported to the police. The child missing from one region may be found in another region or another state, for various reasons. So even if a child is found, it is difficult to identify him/her from the reported missing cases.

A framework and methodology for developing an assistive tool for tracing missing child is described further.

An idea for maintaining a virtual space is proposed, such that the recent photographs of children given by parents at the time of reporting missing cases is saved in a repository.

The public is given provision to voluntarily take photographs of children in suspected situations and uploaded in that portal.

4.3 MODULES OVERVIEW

1) Data processing

Pre-processing input raw image in the context of face recognition involves acquiring the face region and standardizing images in a format compatible with the CNN architecture employed. Each CNN has a different input size requirement. The photographs of missing child acquired by a digital camera or mobile phone are taken and categorized into separate cases for creating the database of face recognition system. The face region in each image is identified and cropped for getting the input face images.

2) Data Upload

It consists of a national portal for storing details of missing child along with the photo. Whenever a child missing is reported, along with the FIR, the admin uploads the photo of the missing child into the portal. The public can upload photo of any suspicious child at any time into the portal with details like place where the child is found, name of the child, their name and phone number. The photo uploaded by the users will be automatically compared with photos of the registered missing children and if a matching photo with sufficient score is found, then the result will be displayed whether the child is found in the missing database or not.

3) Searching Data

Whenever users upload photo of a suspected child, the system generates template vector of the facial features from the uploaded image, and it checks whether the uploaded photo is present in the missing child database or not.

4.4 DEFINING THE MODULES

The project mainly consists of 5 modules :

1)Public module

2)Image processing module

3)Search module

4)Result module

5)Admin module

4.5 MODULE FUNCTIONALITY

1)Public Module -The public can upload the image of suspected child in the portal along with details like place where the child is found, name of the child, their name and phone number.

2)Image Processing Module -The uploaded image is processed, and the face region is cropped. The image is processed to the required pixels and the facial features are extracted.

3)Search Module -The system checks the whether the extracted facial features match with the facial features of images present in the missing repository or not.

4)Result Module -This displays to public whether the child is present in the missing repository or not.

5)Admin Module -The admin can check the database and report about the missing children

5. PROJECT SYSTEM DESIGN

Systems design is the process of defining elements of a system like modules, architecture, components and their interfaces and data for a system based on the specified requirements. It is the process of defining, developing and designing systems which satisfies the specific needs and requirements of a business or organization. A systemic approach is required for a coherent and well-running system. Bottom-Up or Top-Down approach is required to take into account all related variables of the system. A designer uses the modelling languages to express the information and knowledge in a structure of system that is defined by a consistent set of rules and definitions. The designs can be defined in graphical or textual modelling languages.

Unified Modelling Language has been used by us to describe software both structurally and behaviourally with notations.

SYSTEM ARCHITECTURE

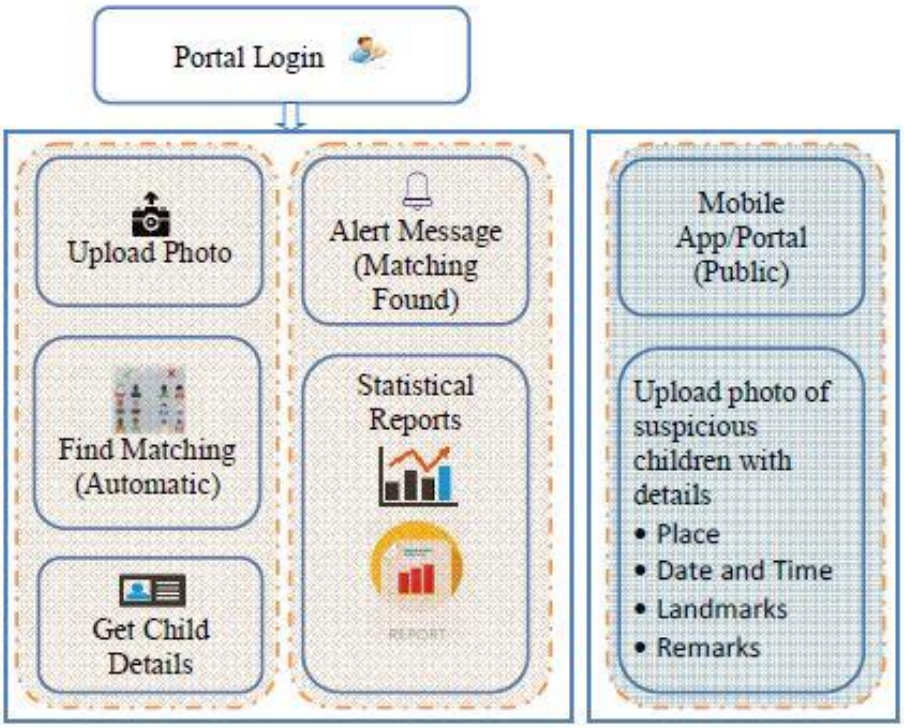


Fig 5.1 System Architecture

It consists of a national portal for storing details of missing child along with the photo. Whenever a child missing is reported, along with the FIR, the concerned officer uploads the photo of the missing child into the portal.

Public can search for any matching child in the database for the images with them. The system will prompt the most matching cases. Once the matching is found, the officer can get the details of the child.

5.1 DATA FLOW DIAGRAMS

CONTEXT-LEVEL DIAGRAM :

This process is handled by four categories of users namely: the searcher, authorized parties, managers and system administrator. The main process is to trace a lost child.

The system administrator registers and views the authorized parties and managers. This creates credentials for logging in. The authorized parties then input the details for the lost and found persons after a successful login.

The searchers then register themselves and conduct a thorough system guided search in the system. An option of adding a new missing person is also provided in case the searched person is not found.

The management then views reports generated by the system.

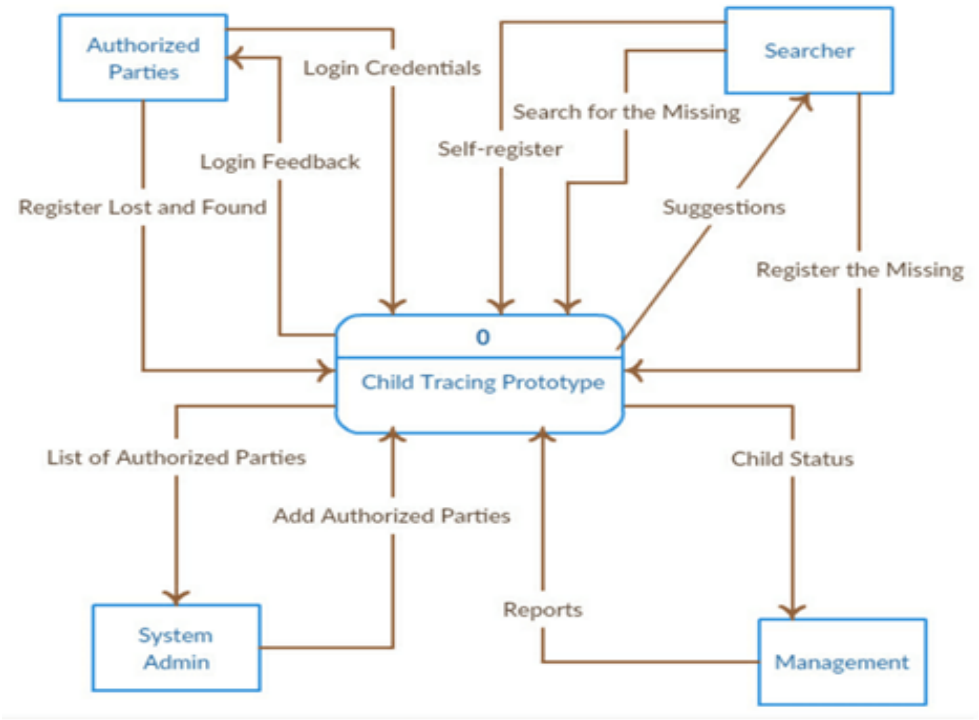


Fig 5.2 Context Level Diagram

LEVEL-0 DIAGRAM

As illustrated in the context diagram, the major processes involved in tracing a missing person/child are:

Managing users:

The managed users include both the authorized parties and the managers. This process involves creating new users and their roles; modifying and deleting the existing users.

Registration of lost and found persons:

This process provides an interface to input the details of the lost but found persons. This data includes but not limited to the phenotypes; current location; the contacts of the current care giver; name, if known; and the child's photo.

Self-registration:

For anyone to log into this system, authentication is required. Setting up registration centers for the

people with legitimate interest on a missing person is quite expensive. This meansthat the searchers have to register themselves. As a result, an interface for self-registration is required. Here, the searchers input their personal details plus the contact for receiving notifications.

Searching for missing persons:

Here is an interactive process for searching the missing person. The searchers have to log in or register, in case it is their first time before proceeding to conduct the search. The system computes the information gain for all the attributes and prompts the searcher to select among the available options of the root attribute. Information gain for the remaining attributes, keeping in mind the previous attribute, is computed. The searcher is again and again prompted to select an option among the ones presented for the attribute, among the ones remaining, with the highest information gain. This step is repeated until searched person is identified. This person is then marked as found and the search ends. Otherwise, the searcher registers the person, if not found,as explained in the registration process.

Registration of missing persons:

The searcher enters the details of the missing person into the system. These include the phenotypes, name and the photo of the person. The searcher is also prompted to select the preferred channel of notification (through SMS or email).

Synchronizing lost and found and missing persons:

This is a background process which tries to find a match between the persons reported as missing and the ones in the lost and found list. This is triggered by an update in the list of lostand found persons by an ‘authorized party’. If similarities reach a certain threshold, a notificationis sent to the searcher, requesting a search process. This then follows the search process explained.

Preparing reports:

The system administrator and/or the managers enter the desired report dates. This is followed a selection of the report category and sub-category, if necessary, and clicking the ‘view report’ button. This process extracts reports from lost and found persons; users; missing persons; the found persons; and searchers data stores. All these processes have been summarized and diagrammatically represented in the level-0 diagram below:

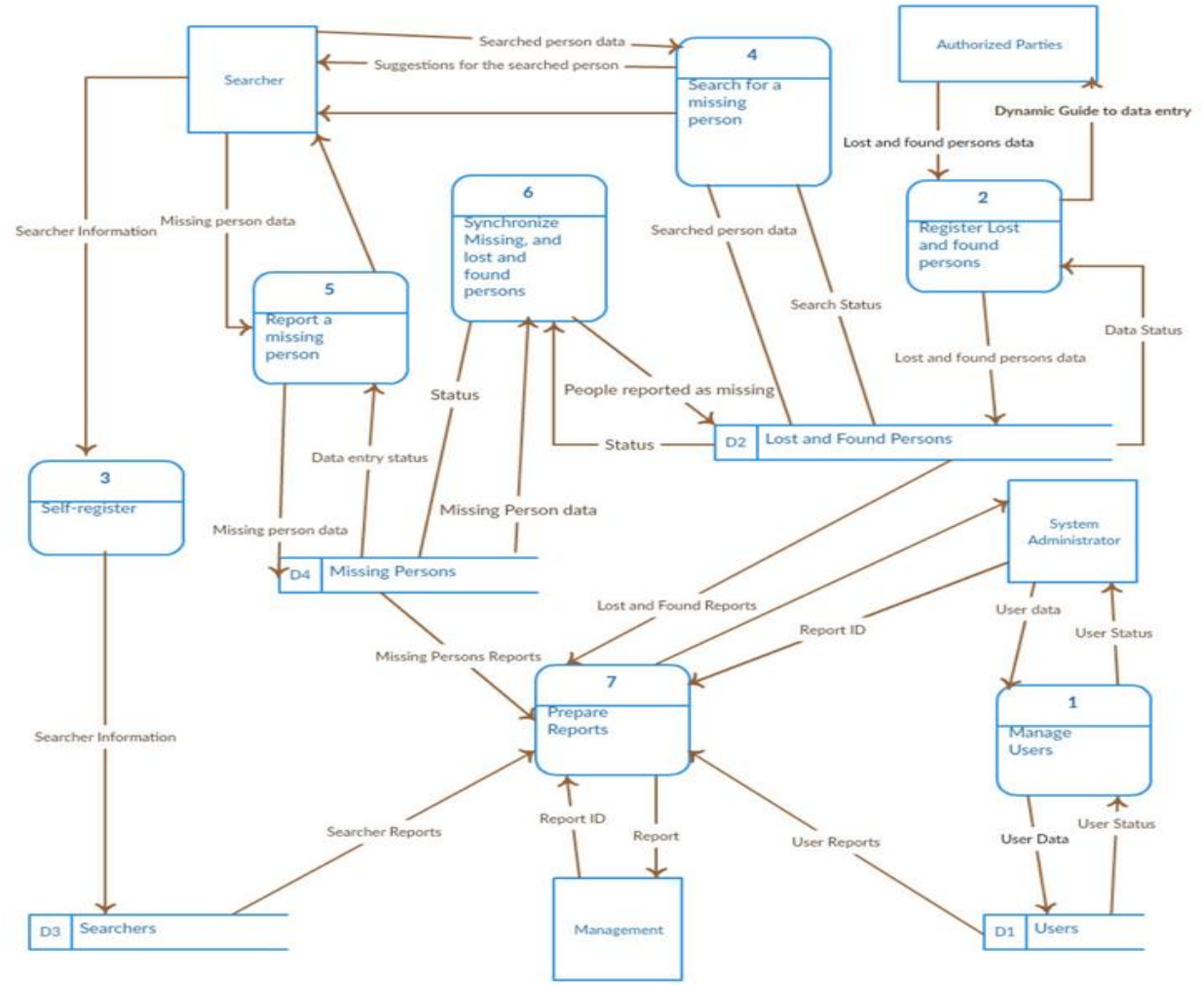


Fig 5.3 Level-0 DFD

5.2 ENTITY-RELATIONSHIP DIAGRAM

An entity relationship model is a high-level conceptual model which describes data in terms of entities, their attributes and their relationships (Riccardi, 2002). The entity relationship diagram shows how is represented and organized in the database schema without specifying the actual data (Pagh, 2006).

The system administrator has the user id attribute as the primary key. The relationship between the system administrator and the user is one to many. This shows that one system administrator can manage more than one entity user.

The user entity has username attribute as the primary key. The entities manager and authorized party borrow attributes from user. These borrowed attributes include name, username and password. The

two have “ISA” relationship with the entity user. In addition to these attributes, authorized party entity has location and organization attributes.

Lost and found persons have several phenotype attributes which describe them. The relationship between the authorized party entity and the lost and found entity is a many to many relationship. This means that one or more authorized parties can register one or more lost and found persons. Just like the lost and found entity, the missing person entity has phenotype attribute which describe it. The searcher entity has username, password, phone number and nameas the attributes. Username is the key attribute. The name attribute is multivalued as it can take several values. There exists a many to many relationship between the searcher and the lost and found entity. This means that one or more searchers can search for the same or different lost and found persons. A similar relationship exists between the searcher and the missing person entity which means that one or more searchers can register one or more missing persons.

Entity report has date, details and report id as its attributes. The report id is the primary key. Considering the relationship between entities system administrator and report, one system administrator can view one or more reports. One or more managers can also view one or more reports. As shown in the one to many relationship.

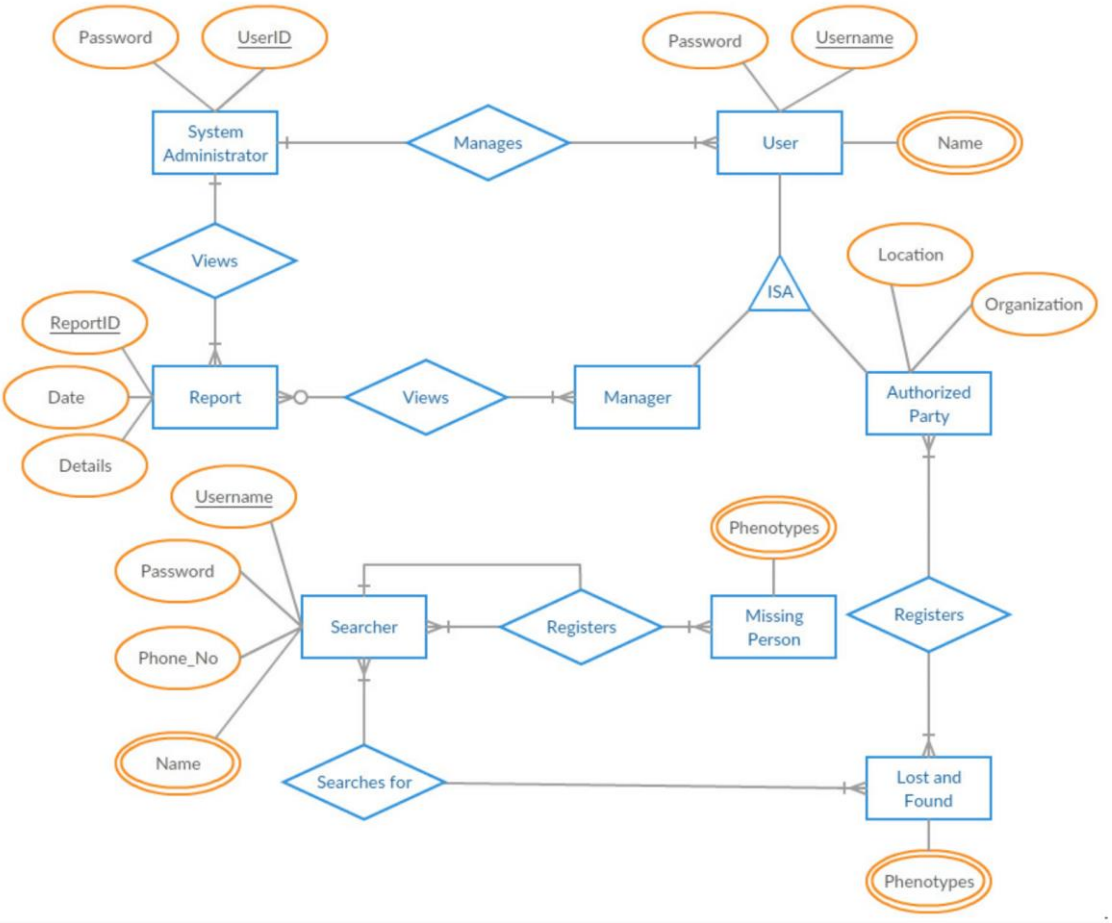


Fig 5.4 E-R Diagram

5.3 UML DIAGRAMS

CLASS DIAGRAM

A class diagram provides a pictorial representation of all the classes in an object oriented system; their attributes and methods; their connections; their interactions and inheritances if any. In simpler terms, classes represent objects whose roles are similar and to what extend the objects of the classes "know" about each other (Felici, 2011).

The system administrator can login, add one to many new users, modify the existing and log out. The system administrator can also view one or more reports. Managers and authorized parties are both users. They inherit attributes; and login and log out functions from the superclass"user". In addition to the inherited functions, one or many managers can view zero or more reports while one or many authorized parties can register zero or more missing persons. A searcher on the other hand can login; search one or more lost and found persons; register zero or more missing persons and log out.

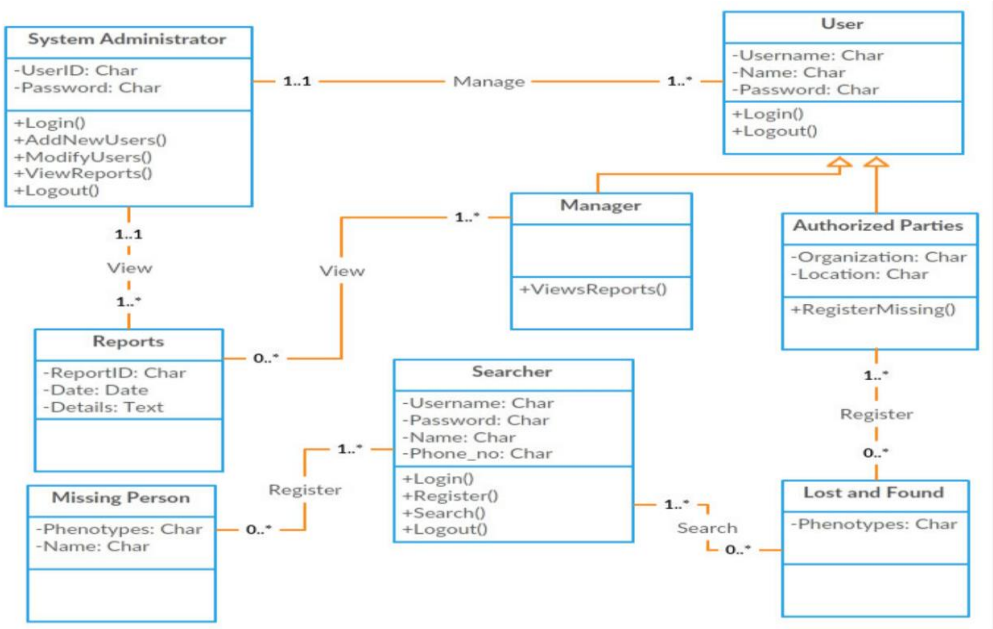


Fig 5.5 Class Diagram

USE CASE DIAGRAM

A use case is simply a list of actions which typically define the interactions between an actor and the system with an aim of achieving a certain goal. Each interaction is a single unit of work and captures a “contract” for the behavior of the system under discussion to deliver a singlegoal (Kettenis, 2007). Most of the functional requirements are captured by the use case. The diagram is displayed below:

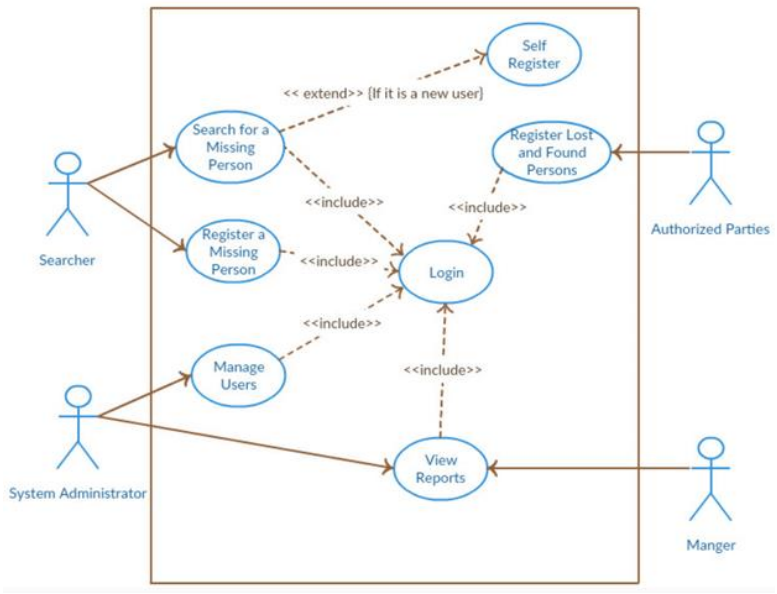


Fig 5.6 Use-case Diagram

SEQUENCE DIAGRAM

The sequence diagram in this case provides a visual representation the object interactions during the searching process. This includes the actor and the objects the actors interact with throughout the execution of the search.

The searcher initiates the process by executing the ‘trace person’ action. This is then followed by a loop where the searcher selects from displayed phenotypes which narrow down from a wider to a focused view depending on their entropies.

This process repeats until when the searched person is found. The searcher then selects the person and endsthe search process. This marks the person as found and is followed by an automatic update of thelist of found persons.

If the searched person was not found, the searcher enters the details of the searched person. This registration process prompts the searcher to provide the contact through which, ifthe person is found, a notification will be send through.

After this, any subsequent registration of lost and found persons trigger synchronization of the list of the missing persons and those reported as lost and found. In case there is a close relationship between the searched person and the one entered, a notification is automatically send to the searcher, prompting him/her to conduct another search.

The following diagram depicts the entire scenario described above in further detail :

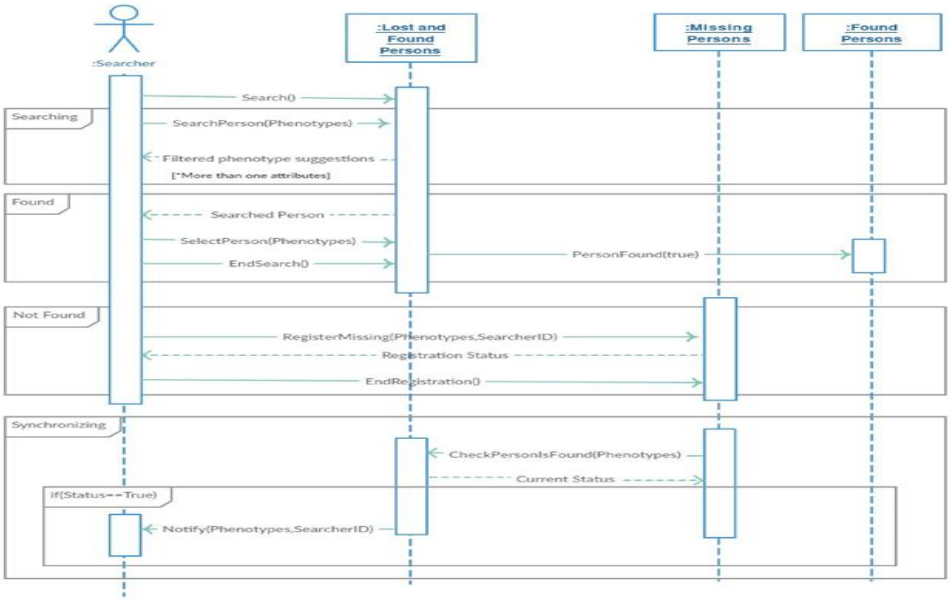


Fig 5.7 Sequence Diagram

6. PROJECT CODING

Project Coding is the process of designing and building an executable computer program to accomplish a specific computing result or to churn out a particular prototype or product. Programming involves tasks such as: analysis, generating algorithms, profiling algorithms' accuracy and resource consumption, and the implementation of algorithms in a chosen programming language (commonly referred to as coding). The source code of a program is written in one or more languages that are intelligible to programmers, rather than machine code, which is directly executed by the central processing unit. The purpose of programming is to find a sequence of instructions that will automate the performance of a task (which can be as complex as an operating system) on a computer, often for solving a given problem. Proficient programming thus often requires expertise in several different subjects, including knowledge of the application domain, specialized algorithms, and formal logic.

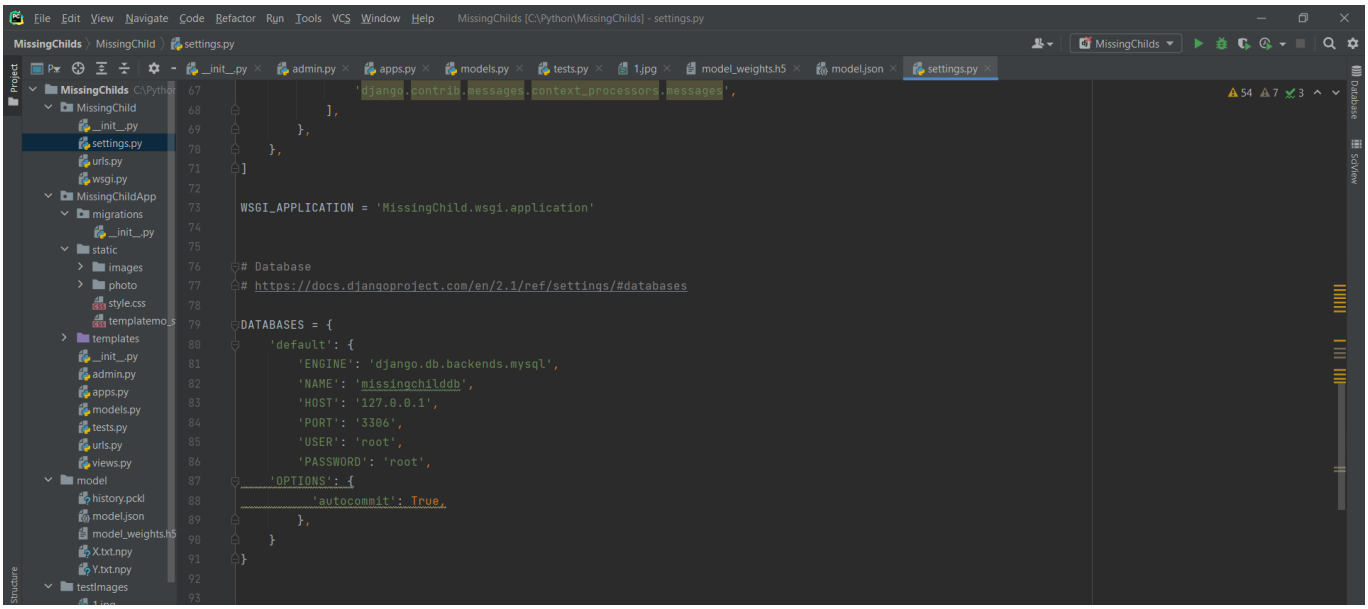
6.1 CODE TEMPLATES

1) This template contains the overview of multiple Django Server tools that we have imported for authentication, content type verification, message display, session checking, security and context processors for debugging, requests etc.

```
INSTALLED_APPS = [
    'django.contrib.admin',
    'django.contrib.auth',
    'django.contrib.contenttypes',
    'django.contrib.sessions',
    'django.contrib.messages',
    'django.contrib.staticfiles',
    'MissingChildApp'
]

MIDDLEWARE = [
    'django.middleware.security.SecurityMiddleware',
    'django.contrib.sessions.middleware.SessionMiddleware',
    'django.middleware.common.CommonMiddleware',
    'django.middleware.csrf.CsrfViewMiddleware',
    'django.contrib.auth.middleware.AuthenticationMiddleware',
    'django.contrib.messages.middleware.MessageMiddleware',
    'django.middleware.clickjacking.XFrameOptionsMiddleware',
]

ROOT_URLCONF = 'MissingChild.urls'

TEMPLATES = [
    {
        'BACKEND': 'django.template.backends.django.DjangoTemplates',
        'DIRS': [
      os.path.join('C:/Python/MissingChild/MissingChildApp', 'templates'),
   ],
        'APP_DIRS': True,
        'OPTIONS': {
            'context_processors': [
                'django.template.context_processors.debug',
                'django.template.context_processors.request',
                'django.contrib.auth.context_processors.auth',
                'django.contrib.messages.context_processors.messages',
            ],
        },
    },
]
```

Fig 6.1 Code Template (1)

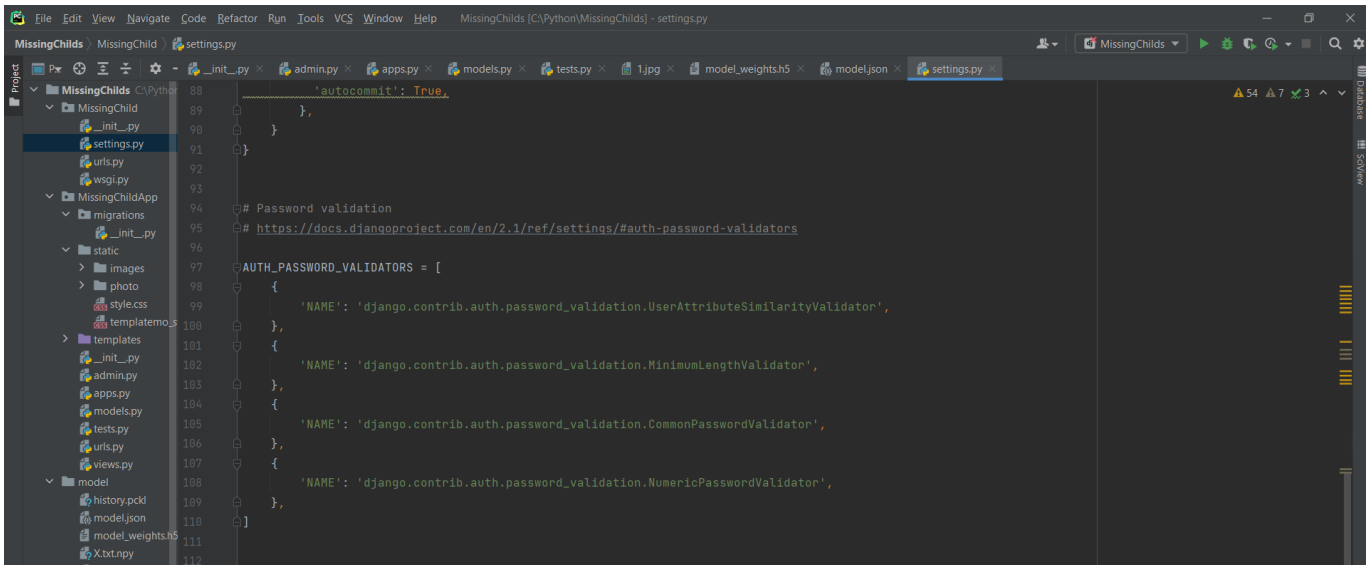
2) This template displays the details to access the database that is linked to the python code.



```
                'django.contrib.messages.context_processors.messages',
            ],
        },
    },
]

WSGI_APPLICATION = 'MissingChild.wsgi.application'


# Database
# https://docs.djangoproject.com/en/2.1/ref/settings/#databases

DATABASES = {
    'default': {
        'ENGINE': 'django.db.backends.mysql',
        'NAME': 'missingchilddb',
        'HOST': '127.0.0.1',
        'PORT': '3306',
        'USER': 'root',
        'PASSWORD': 'root',
    'OPTIONS': {
            'autocommit': True,
        },
    }
}
```

Fig 6.2 Code Template(2)

3) This template displays the user attribute validation statements from the source code.



```
        'autocommit': True,
        },
    }
}


# Password validation
# https://docs.djangoproject.com/en/2.1/ref/settings/#auth-password-validators

AUTH_PASSWORD_VALIDATORS = [
    {
        'NAME': 'django.contrib.auth.password_validation.UserAttributeSimilarityValidator',
    },
    {
        'NAME': 'django.contrib.auth.password_validation.MinimumLengthValidator',
    },
    {
        'NAME': 'django.contrib.auth.password_validation.CommonPasswordValidator',
    },
    {
        'NAME': 'django.contrib.auth.password_validation.NumericPasswordValidator',
    },
]
```

Fig 6.3 Code Template(3)

6.2 OUTLINE FOR VARIOUS FILES

We used Python programming to implement our project. We also used HTML and CSS to develop our webpage. Our code consists of various modules that we have used. Our project modules are – Public Upload, Images, Search, Result Display and Admin. We also used various Django Server and various AI and deep learning methods that Django's embedded server provides to facilitate coding in an easier and quicker way.

6.3 CLASS WITH FUNCTIONALITY

There are multiple classes in our code, some of which are :

1) Templates: They handle the backend context processors for requests, debugging, authentication and messaging.
2) Middleware: The middleware class handles the view, sessions, security and frame options.
3) Django Installed Apps: These handle the content types and static files through the code.
4) Validator: They handle the validation of multiple users in and out of the website based on the credentials that the user would provide.
5) Database Linking: This handles the backend linking of the entire code to the MySQL server in order to facilitate the final output.

6.4 METHODS INPUT AND OUTPUT PARAMETERS

We implemented multiple methods, few of which are :

1. get_image()

2. cnnModel()

3. StandardScaler()

4. svm=SVC() , etc.

Our first method get_image() takes in the images and for processing recognition. cnnModel() doesn't have any input parameters. StandardScaler() is used for image scaling while feature extraction.

The svm=SVC() takes in the Kernel, probability and random state to further classify and give out the accurate results later using and print() with the parameterized statement of whether or not the child has been found.

7. PROJECT TESTING

Project Testing is a method to check whether the actual software product matches expected requirements and to ensure that software product is Defect free. It involves execution of software/system components using manual or automated tools to evaluate one or more properties of interest. The purpose of software testing is to identify errors, gaps or missing requirements in contrast to actual requirements.

Some prefer saying Software testing as a White Box and Black Box Testing. In simple terms, Software Testing means the Verification of Application Under Test (AUT). This tutorial introduces testing software to the audience and justifies its importance.

Project testing is important because, if there are any bugs or errors in the software, it can be identified early and can be solved before delivery of the software product. Properly tested software product ensures reliability, security and high performance which further results in time saving, cost effectiveness and customer satisfaction.

Typically Testing is classified into three categories.

- Functional Testing
- Non-Functional Testing or Performance Testing
- Maintenance (Regression and Maintenance)

7.1 VARIOUS TEST CASES

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub-assemblies, assemblies and/or a finished product. It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

We have performed multiple tests under the broad categorisation of white-box and black-box testing which further include unit, integration, boundary value and statement covering etc.

ID	Case	Expected Outcomes	Comments
1.0	**Login**		
1.1	Password or username left out	Error Dialog Box	Pass
1.2	Wrong password or username entered	Error Dialog Box	Pass
2.0	**User Registration**		
2.1	Leaving out a required field	Error Dialog Box	Pass
3.0	**Lost and Found Person Registration**		
3.1	Leaving out a required field	Error Dialog Box	Pass
4.0	**Searcher Registration**		
4.1	Leaving out a required field	Error Dialog Box	Pass
5.0	**Missing Person registration**		
5.1	Leaving out a required field	Error Dialog Box	Pass
6.0	**Role Access**		
6.1	Accessing unauthorized Page	Session destroyed. User redirected to the login page	Pass

Table 6. Test Cases Tabulation

7.2 BLACK-BOX TESTING

Black Box Testing is a software testing method in which the functionalities of software applications are tested without having knowledge of internal code structure, implementation details and internal paths. Black Box Testing mainly focuses on input and output of software applications and it is entirely based on software requirements and specifications. It is also known as Behavioral Testing.

Here are the generic steps followed to carry out any type of Black Box Testing.

- Initially, the requirements and specifications of the system are examined.
- Tester chooses valid inputs (positive test scenario) to check whether SUT processes them correctly. Also, some invalid inputs (negative test scenario) are chosen to verify that the SUT is able to detect them.

- Tester determines expected outputs for all those inputs.
- Software tester constructs test cases with the selected inputs.
- The test cases are executed.
- Software tester compares the actual outputs with the expected outputs.
- Defects if any are fixed and re-tested.

Types of Black Box Testing

There are many types of Black Box Testing but the following are the prominent ones -

- Functional testing - This black box testing type is related to the functional requirements of a system; it is done by software testers.
- Non-functional testing - This type of black box testing is not related to testing of specific functionality, but non-functional requirements such as performance, scalability, usability.
- Regression testing - Regression Testing is done after code fixes, upgrades or any other system maintenance to check the new code has not affected the existing code.

Black Box Testing Techniques

Following are the prominent Test Strategy amongst the many used in Black box Testing

- Equivalence Class Testing: It is used to minimize the number of possible test cases to an optimum level while maintains reasonable test coverage.
- Boundary Value Testing: Boundary value testing is focused on the values at boundaries. This technique determines whether a certain range of values are acceptable by the system or not. It is very useful in reducing the number of test cases. It is most suitable for the systems where an input is within certain ranges.
- Decision Table Testing: A decision table puts causes and their effects in a matrix. There is a unique combination in each column.

7.3 WHITE-BOX TESTING

White Box Testing is software testing technique in which internal structure, design and coding of software are tested to verify flow of input-output and to improve design, usability and security.

It is one of two parts of the Box Testing approach to software testing. Its counterpart, Blackbox testing, involves testing from an external or end-user type perspective. On the other hand, White box testing in software engineering is based on the inner workings of an application and revolves around internal testing.

The term "WhiteBox" was used because of the see-through box concept. The clear box or WhiteBox name symbolizes the ability to see through the software's outer shell (or "box") into its inner workings. Likewise, the "black box" in "Black Box Testing" symbolizes not being able to see the inner workings of the software so that only the end-user experience can be tested.

White box testing involves the testing of the software code for the following:

- Internal security holes
- Broken or poorly structured paths in the coding processes
- The flow of specific inputs through the code
- Expected output
- The functionality of conditional loops
- Testing of each statement, object, and function on an individual basis

The testing can be done at system, integration and unit levels of software development. One of the basic goals of whitebox testing is to verify a working flow for an application. It involves testing a series of predefined inputs against expected or desired outputs so that when a specific input does not result in the expected output, you have encountered a bug.

To give you a simplified explanation of white box testing, we have divided it into two basic steps. This is what we do when testing an application using the white box testing technique:

STEP 1) UNDERSTAND THE SOURCE CODE

The first thing a tester will often do is learn and understand the source code of the application. Since white box testing involves the testing of the inner workings of an application, the tester must be very knowledgeable in the programming languages used in the applications they are testing. Also, the testing person must be highly aware of secure coding practices. Security is often one of the primary objectives of testing software. The tester should be able to find security issues and prevent

attacks from hackers and naive users who might inject malicious code into the application either knowingly or unknowingly.

Step 2) CREATE TEST CASES AND EXECUTE

The second basic step to white box testing involves testing the application's source code for proper flow and structure. One way is by writing more code to test the application's source code. The tester will develop little tests for each process or series of processes in the application. This method requires that the tester must have intimate knowledge of the code and is often done by the developer.

The goal of WhiteBox testing in software engineering is to verify all the decision branches, loops, statements in the code.

A major White box testing technique is Code Coverage analysis. Code Coverage analysis eliminates gaps in a Test Case suite. It identifies areas of a program that are not exercised by a set of test cases. Once gaps are identified, you create test cases to verify untested parts of the code, thereby increasing the quality of the software product

There are automated tools available to perform Code coverage analysis. Below are a few coverage analysis techniques a box tester can use:

Statement Coverage:- This technique requires every possible statement in the code to be tested at least once during the testing process of software engineering.

Branch Coverage - This technique checks every possible path (if-else and other conditional loops) of a software application.

Apart from above, there are numerous coverage types such as Condition Coverage, Multiple Condition Coverage, Path Coverage, Function Coverage etc. Each technique has its own merits and attempts to test (cover) all parts of software code. Using Statement and Branch coverage you generally attain 80-90% code coverage which is sufficient.

Following are important WhiteBox Testing Techniques:

- Statement Coverage

- Decision Coverage
- Branch Coverage
- Condition Coverage
- Multiple Condition Coverage
- Finite State Machine Coverage
- Path Coverage
- Control flow testing
- Data flow testing

8. OUTPUT SCREENS

An output screen is a device used to display output. An output screen could be a separate monitor or another display device used only to display the output being received from the computer or other devices.

Here, in the screen prints given below, we can see that the user interface screens consist of the home page which describes the purpose of the portal, and the public access screen which allows users to upload the credentials of the child that they found in order to check whether the child exists in the repository or not.

8.1 USER INTERFACE

1) HOME PAGE

This allows the user to look at the various sections of the website like the home page, official login, public upload screen etc, so that they can navigate through the web page in an easier way.

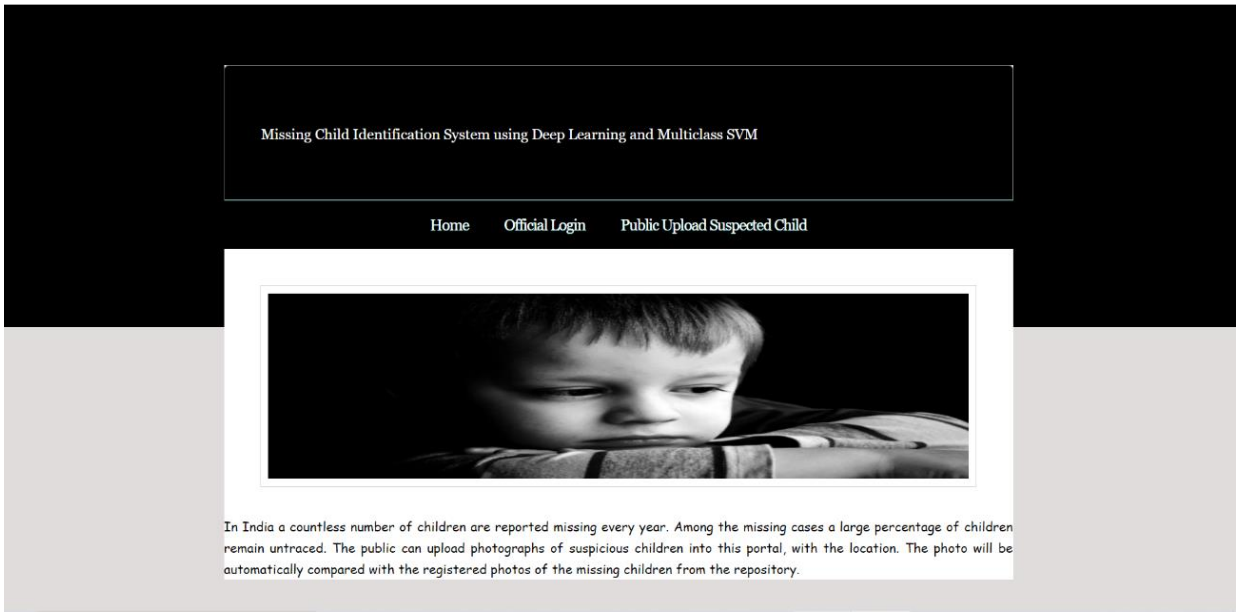


Fig 8.1 User Interface

8.2 OUTPUT SCREENS

1) The official login screen helps the user (administrator, here) to check the activity going on in the web page.

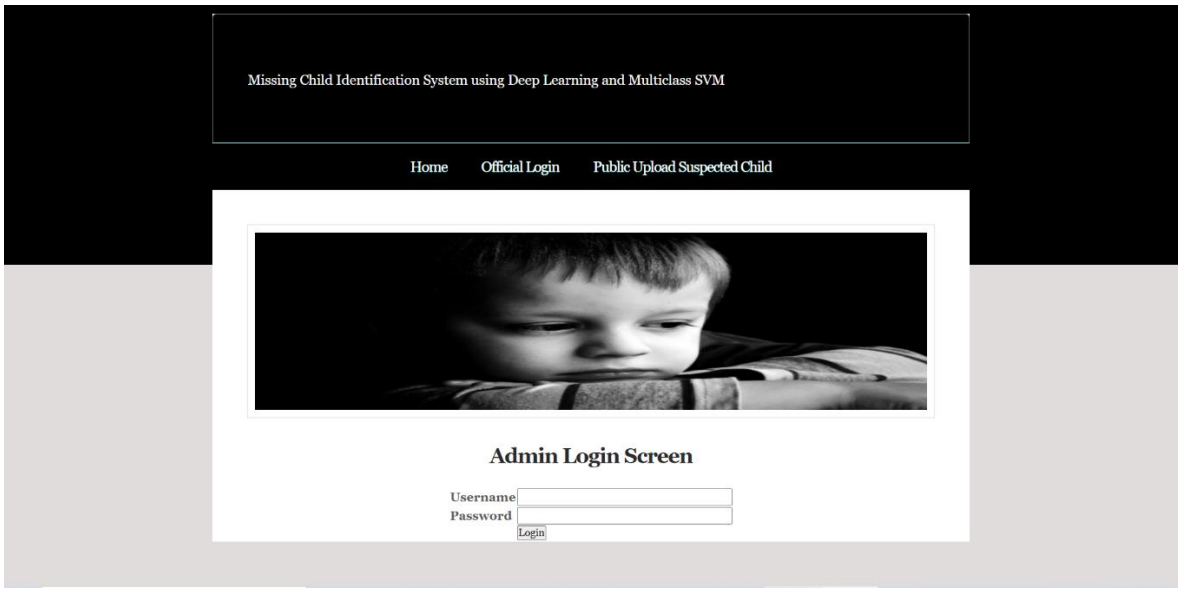


Fig 8.2 Admin Login Screen

2) The public upload suspected child screen helps users to upload the details of the child that they want to look for. The admin would later contact and fetch extra details for them, if they need them.

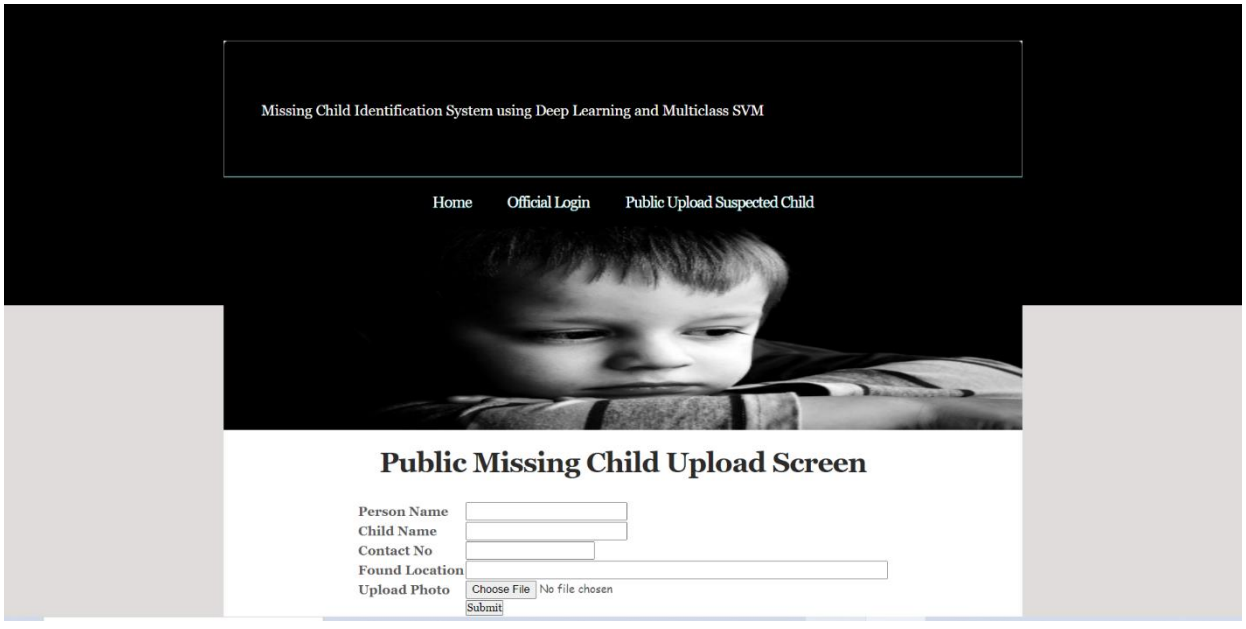


Fig 8.3 Upload Screen

9. EXPERIMENTAL RESULT

1) The details like the name of the uploader, name of the child (temporary, if not known), contact number, location and photo are uploaded by the searching party

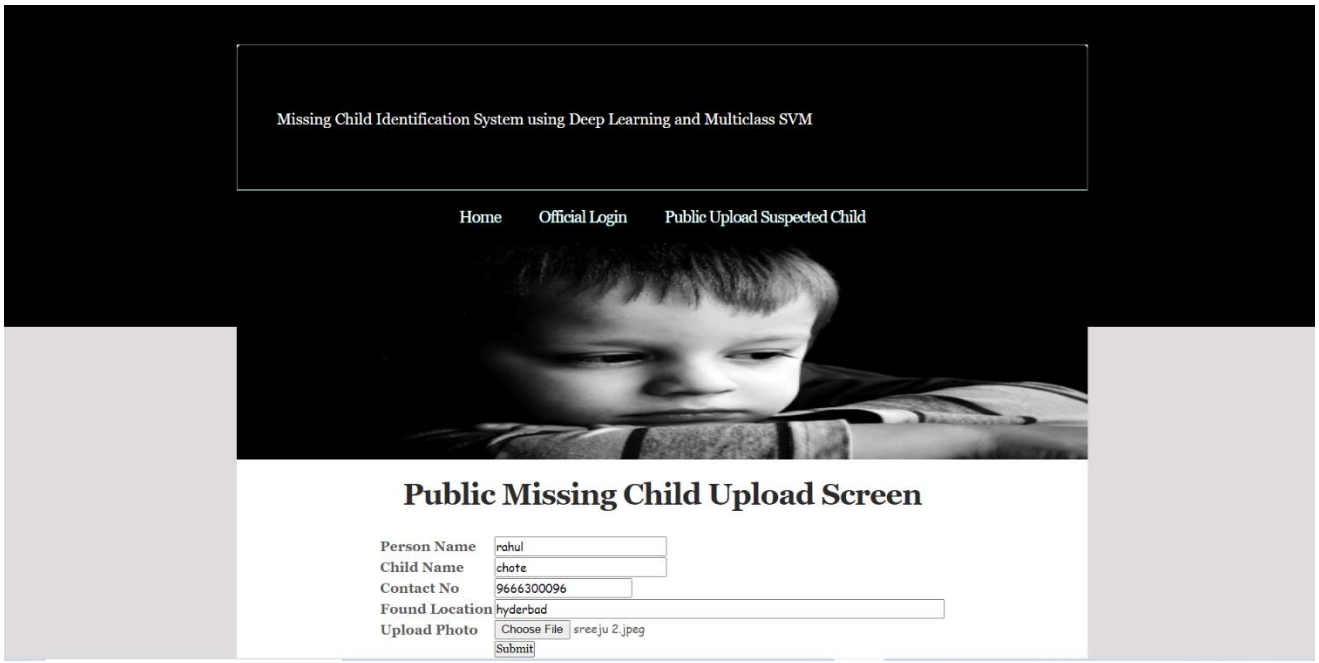


Fig 9.1 Details Entered

2) The result as to whether or not the child exists is displayed after the user enters the details he wishes to look for.

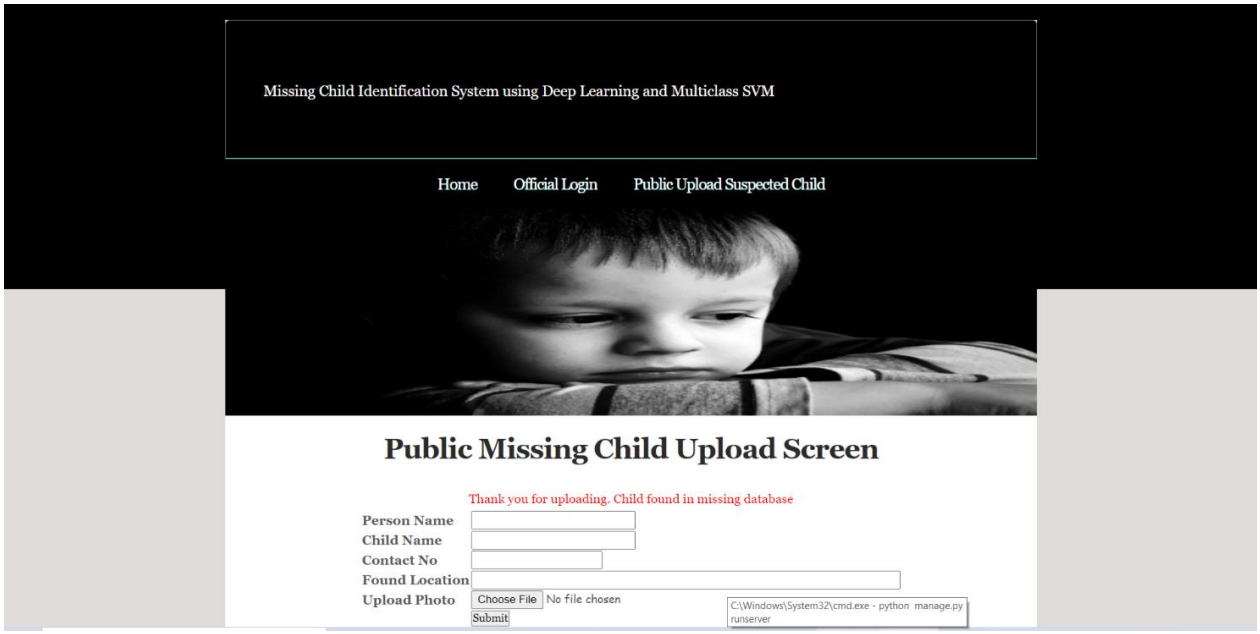


Fig 9.2 Result Displayed

3) Admin output includes the tracking of activity that goes on in the web-page. The status of various children in the database since the first search up until the latest search result can be found.

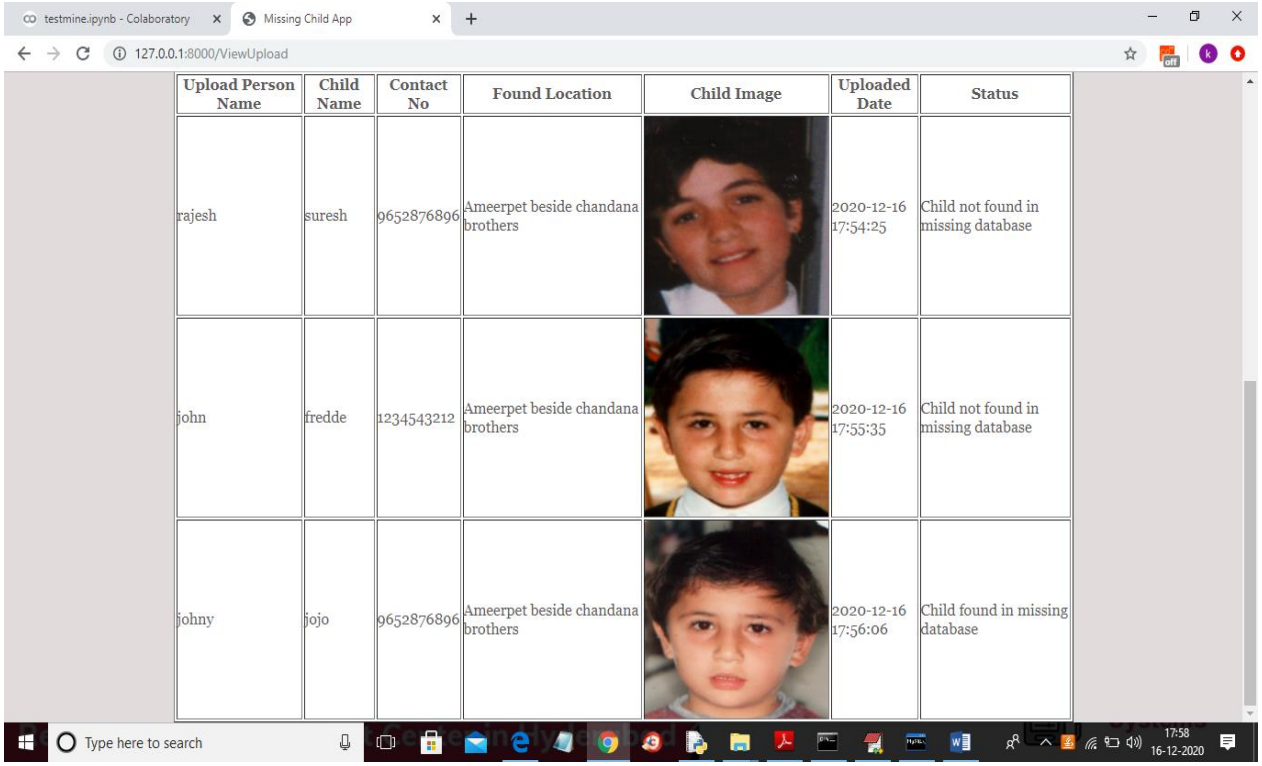


Fig 9.3 Activity Tracking

10.CONCLUSION AND FUTURE ENHANCEMENT

A missing child identification system is proposed, which combines the powerful CNN based deep learning approach for feature extraction and support vector machine classifier for classification of different child categories.

This system is evaluated with the deep learning model which is trained with feature representations of children faces. By discarding the softmax of the VGG-Face model and extracting CNN image features to train a multi class SVM, it was possible to achieve superior performance.

Performance of the proposed system is tested using the photographs of children with different lighting conditions, noises and also images at different ages of children.

The classification achieved a higher accuracy of 99.6% which shows that the proposed methodology of face recognition could be used for reliable missing children identification.

Some of the limitations of our prototype include the restriction on the no.of images loaded- a large no.of images would result in reduced performance of selected CNN architecture (VGG-16), also, our system can recognize up to only the young adult stage of the child, and extremely low resolutions cannot be handled as of now.

In the future, we can extend the existing CNN architecture – VGG-16 to its advanced version VGG-19 (which is better layered, thus can handle increased image load). Also, we could develop it to be more efficient in recognizing children irrespective of their grown age, and the uploaded image resolution/ other conditions (lighting, slight blurs etc). Another feature that we look forward to add is the “current location extractor”.

11.REFERENCES

[1] MISSING PEOPLE DETECTION SYSTEM Aryan Patel1, Dhru Prajapati2, International Research Journal of Engineering and Technology (IRJET) Volume: 08 Issue: 05 | May 2021

[2] K. Bharath, Paithankar Sumit, S. Amudha, (2020). The Lore of speculation and analysis using machine learning and image matching, in International Journal of Trendy Research in Engineering and Technology Volume 4 Issue 4 August 2020.

[3] Farah Deeba, Aftab Ahmed, "LBPH-based Enhanced Real-Time Face Recognition", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 5, 2019

[4] Abhishek Pratap Singh, SunilKumar S Manvi, Pratik Nimbal, Gopal Krishna Shyam, "Face Recognition System Based on LBPH Algorithm", International Journal of Engineering and Advanced Technology (IJEAT) Volume-8, Issue-5S, May, 2019.

[5] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In CVPR, 2019.

[6] BharathDarshanBalar, D S Kavya, Chandana M, Anush E, Vishwanath R Hulipalled, "Efficient Face Recognition System for Identifying Lost People", International Journal of Engineering and AdvancedTechnology (IJEAT) Volume-8, Issue-5S, May 2019

[7] Debayan Deb, Lacey Best-Rowden, and Anil K Jain. Face recognition performance under aging. In CVPRW, 2017.

[8] Patrick J. Grother, Mei Ngan, and Kayee Hanaoka. Ongoing Face Recognition Vendor Test (FRVT), Part 2: Identification. NIST Interagency Report, 2018.

[9] Debayan Deb, Neeta Nain, and Anil K Jain. Longitudinal study of child face recognition. In IEEE ICB, 2018.

[10] Tianyue Zheng, Weihong Deng, and Jiani Hu. Age estimation guided convolutional neural network for age-invariant face recognition. In CVPRW, 2017

12.PUBLICATIONS

CONFERENCE :

International Conference on “Innovations in Computers Networks, Computational Intelligence and IoT” (ICICCI – 21)

Paper ID : ICICCI – 21 – 0049

13. STUDENT PROFILES

R.Sreeharipriya is a Bachelor of Technology student at St. Martin's Engineering College studying Computer Science and Engineering. She finished her schooling in Kendriya Vidyalaya A.F.S, Begumpet, which included both her 10th and 12th grades. Python and Java are among her technical skills. She also knows the fundamentals of C++. Her participations include: a National Level Seminar on "Recent Trends in Cloud Computing, Fog, and Edge Computing" on the 18th and 19th of June 2021, a National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" from the 10th to the 12th of December 2020, and a Women online workshop on "Women in Cyber Security and Privacy in 2020" from the 6th to the 10th of July 2020. In June 2019, she worked as a summer intern at Aspirevision Technologies PvtLtd on a machine learning project to develop a "Text Summarizer" using NLP (Natural Language Processing). Python, Artificial Intelligence, Machine Learning, and Deep Learning are among her interests. She spends her free time taking online certification courses related to her field of study as well as personal interests from platforms such as Coursera and CursaApp. Quite apart from her academic field of study, she has a keen interest in Creative and Content Writing, which she pursued and developed stories individually and as part of multiple assignments while working as a Campus Reporter in StuMagz (July to October 2019) and also doing freelance writing for some clients. She has a large passion in fields such as public relations, consulting, and management, which she hopes to pursue in the coming years after completing her undergraduate degree.

S.Sreeja Reddy is currently pursuing her Bachelor of Technology in the stream of Computer Science And Engineering at St.Martin's Engineering College. She completed her intermediate from Narayana Junior College and 10th class from Sadhu Vaswani International School. Her technical skills include C, Java and Python. She also has a basic understanding of C++. She took part in E-Summit program conducted at Marri Laxman Reddy Institute Of Technology in 2018 and completed few certification courses from online platforms like Coursera and CursaApp.

G. Rahul is currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. He has completed his Secondary Education from St. Joseph's Public School and Higher Secondary Education from Narayana Junior College. His technical skills include Java, C, Python. He attended E-summit, Entrepreneurship carnival, which was hosted by EDC – MLRIT in association with Nucleus Tech and SUMVN in the year 2017. He has completed online courses (AWS Fundamentals: Going Cloud Native, Data Science Math Skill, Leadership and Emotional Intelligence, Managing Project Risks and Fundamentals, Python) from Coursera and CursaApp.

Nishika Divya Lewis is a student at St. Martin's Engineering College, pursuingBachelor of Technology in Computer Science and Engineering, as well as an Associate Curriculum Developer for coding at Cuemath Pvt.Ltd which is a leading global after-school math and coding platform for K-12 classes. She completed her Intermediate Education from TriVidyaa Junior College and schooling from Narayana Olympiad High School. She was recruited from a list ofapplicants to be a member of AIESEC which is a non-governmental not-for-profit organisation recognised by UNESCO which provided young people with leadership development, cross-cultural internships, and global volunteer exchange experiences from January 2018- December 2018. In January 2019, she was selected to lead the marketing department for TechnologyAwareness Month (TAM), a significant event at St. Martin's Engineering College, and was promoted to Vice President the following year (2020) based on her abilities and commitment. In June 2019 she worked as a summer intern at Smart Bridge Pvt.Ltd on a machine learning project ' Survival Analysis on Diabetes' using IBM Watson Studio and Node Red. She is fond of graphic designing & creating content. Her technical skills include C, C++, Java and Python. She is passionate towards learning newthings and aspires to become a Full Stack Web Developer in the near future.

14. APPENDICES

Appendix A: User Requirements Questionnaire

User Requirements

Questionnaire

This research will be used for academic purpose only. Its main objective is to collect the user requirements to create a child tracing prototype. Kindly provide your honest answers inthe following questions. Please note that your responses will be treated as private and confidential.

1. Missing children are easy to trace, provided that they are in safe hands (with the police or transferred to a children's home).

- ᛉ Strongly Agree
- ᛉ Agree
- ᛉ Neutral
- ᛉ Disagree
- ᛉ Strongly Disagree

2. The current process of tracing a missing child/person is efficient.

- ᛉ Strongly Agree
- ᛉ Agree
- ᛉ Neutral
- ᛉ Disagree
- ᛉ Strongly Disagree

3. A searcher is promptly alerted when a missing relative is found but not yet re-united with his/her family.

- ᛉ Strongly Agree
- ᛉ Agree
- ᛉ Neutral
- ᛉ Disagree
- ᛉ Strongly Disagree

4. The current process of tracing a missing child/person is user friendly.

ᵞ Strongly Agree
ᵞ Agree
ᵞ Neutral
ᵞ Disagree
ᵞ Strongly Disagree

5. I believe that the current processes and systems, if any, for tracing a missing child/person are secure and the records are safely kept.

ᵞ Strongly Agree
ᵞ Agree
ᵞ Neutral
ᵞ Disagree
ᵞ Strongly Disagree

6. If a proper computer system is implemented, I believe that tracing missing children (persons) would be made easier.

ᵞ Strongly Agree
ᵞ Agree
ᵞ Neutral
ᵞ Disagree
ᵞ Strongly Disagree

7. How sensitive are you while giving out personal information of your loved ones?

ᵞ Very Sensitive
ᵞ Not Very Sensitive
ᵞ Neutral

Appendix B: System Usability Questionnaire

This research will be used for academic purpose only. Its main objective is to find out users' experience in using the child tracing prototype. Kindly provide your honest opinion onthe same. Please note that your responses will be treated as private and confidential.

System Usability Scale

Kindly rate the child tracing prototype about the following:

1. The user interface is very user friendly.

ɣ Strongly Agree

ɣ Agree

ɣ Neutral

ɣ Disagree

ɣ Strongly Disagree

2. I can use this prototype with the minimum training.

ɣ Strongly Agree

ɣ Agree

ɣ Neutral

ɣ Disagree

ɣ Strongly Disagree

3. Searching for a missing person using this system will take a shorter duration as compared to the current methods.

ɣ Strongly Agree

ɣ Agree

ɣ Neutral

ɣ Disagree

ɣ Strongly Disagree

4. This question is practical and aims at testing the accuracy of the prototype. Kindly provide a list of five lost and found children to the researcher. After the researcher inputs them into the system, try searching them and note down your findings (Among them, how many were correctly identified?)

……..

5. The system provides a convenient way of tracing the missing persons

ɾ Strongly Agree
ɾ Agree
ɾ Neutral
ɾ Disagree
ɾ Strongly Disagree

6. I will use this system in case any of my loved ones goes missing.

ɾ Strongly Agree
ɾ Agree
ɾ Neutral
ɾ Disagree
ɾ Strongly Disagree

7. How likely are you to recommend this system to the other users?

ɾ Very Likely
ɾ Likely
ɾ Neutral
ɾ Not Likely
ɾ Not Likely At All

8. Any Comments

……
……
……
……
……
……
……
……
……
……
……
……

Appendix C: Interview Questions

Interview Questions

This research is will be used for academic purpose only. Its main objective is to find out users' experience in using the child tracing prototype. Kindly provide your honest opinion onthe same. Please note that your responses will be treated as private and confidential.

Interviewee: ……. **Location:** …….

Medium: …….. **Date:** ……

1. If a child/person is missing, what should an ordinary citizen do?

   ……
   ……
   ……
   ……
   ……
   ……
   ……
   ……
   ……

2. What is the current procedure of tracing a missing child/person?

   ……
   ……
   ……
   ……
   ……
   ……
   ……
   ……
   ……
   ……

3. What are the challenges in tracing a missing child/person?

……
……
……
……
……
……
……
……

4. How do you handle children who have been brought to you by “Good Samaritans”?

……
……
……
……
……
……
……
……

5. What is the current process of re-uniting a lost and found child/person to his/her family?

……
……
……
……
……
……
……
……

6. What are the challenges faced while re-uniting the lost and found child/person to his/her relatives?

……
……
……
……
……
……
……
……

7. In your opinion, what do you think needs to be improved in the process of:

    a. Tracing a missing child/person?

……
……
……
……
……
……
……
……
……
……
……

    b. Re-uniting a lost and found child/person to his/her family?

……
……
……
……
……
……
……
……
……
……

A

PROJECT REPORT

On

CARTOON OF AN IMAGE

Submitted by

1) Ch.Saipriya (17K81A05K2) **2) S.Alekhya(17K81A05N0)**
3) Terala Ramya(17K81A05N8) **4) K.Bhavyasri (17K81A05P6)**

in partial fulfillment for the award of the

degree of

BACHELOR OF TECHNOLOGY

IN

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Under the Guidance of

Ravi Krishna Ayyappa
Assistant Professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ST.MARTIN'S ENGINEERING COLLEGE
An Autonomous Institute

Dhulapally, Secunderabad – 500100

JUNE 2021

BONAFIDE CERTIFICATE

This is to certify that the project entitled "Cartoon of an Image", is being submitted by **Ch.Saipriya(17K81A05K2), S.Alekhya (17K81A05N0), Terala Ramya(17K81A05N8) and K.Bhavyasri(17K81A05P6)** in partial fulfillment of the requirement for the award of the degree of **BACHELOR OF TECHNOLOGY IN COMPUTER SCIENCE AND ENGINEERING** is recorded of bonafide work carried out by them. The results embodied in this report have been verified and found satisfactory.

Assistant Professor
Mr. Ravi Krishna Ayyappa
Department of CSE

Head of the Department
Dr. M. NARAYANAN
Department of CSE

Internal Examiner

External Examiner

Place:

Date:

DECLARATION

We, the students of **Bachelor of Technology** in the Department of **'Computer Science and Engineering'**, session: 2017 – 2021, St. Martin's Engineering College, Dhulapally, Kompally, Secunderabad, hereby declare that work presented in this Project Work entitled Cartoon of an Image is the outcome of our own bonafide work and is correct to the best of our knowledge and this work has been undertaken taking care of Engineering Ethics. This result embodied in this project report has not been submitted in any university for award of any degree.

Ch.Saipriya	(17K81A05K2)
S.Alekhya	(17K81A05N0)
TeralaRamya	(17K81A05N8)
K.Bhavyasri	(17K81A05P6)

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible and whose encouragements and guidance have crowded effects with success.

We extended our deep sense of gratitude to Principal**, Dr. P. SANTOSH KUMAR PATRA**, St. Martin's Engineering College, Dhulapally, for permitting us to undertake this project. We are also thankful to **Dr. M.NARAYANAN**, Head of the Department, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

We would like to thank **Dr. T. POONGOTHAI,** Department of Computer Science and Engineering (AI & ML) for her encouragement and insightful comments and as well as our project coordinators **Dr. B.RAJALINGAM**, Associate Professor and **Dr. SANTHOSH KUMAR**, Associate Professor, in Department of Computer Science and Engineering for their valuable support.

We would like to express our sincere gratitude and indebtedness to our project supervisor **MR.P.R.K.AYYAPPA**, Assistant Professor, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for her support and guidance throughout our project.

Finally, we express thanks to all those who have helped us successfully to completing this project.

Furthermore, we would like to thank our family and friends for their moral support and encouragement.

We express thanks to all those who have helped us in successfully completing the project.

Ch.Saipriya (17K81A05K2)

S.Alekhya (17K81A05N0)

TeralaRamya (17K81A05N8)

K.Bhavyasri (17K81A05P6)

ABSTRACT

Cartoons are humorous, satirical, and at times opinionated. Drawing cartoons is however, not easy. Only those well-trained artists who possess this great skill can do it well. Recently, many technologies have been developed to make it possible to create cartoons entirely on the computer. This can be recreating and helps one to have a cartoonic view of everything. To cartoonize images and different objects and blend them accordingly as we require. Our aim is to create a cartoon which doesn't look like a filter applied on an image but, is actually a cartoonic view of an input image. In order to get the basic cartoon effect, we just need the bilateral filter and some edge detection mechanism. We can access this cartoon images through an application where you can also save them and make changes. This project proposes the method that makes an input target image into exaggerated cartoon-like images by using reference images. To deform a target image, we extract feature points from a target image and define the feature point model on reference images. And then, we apply feature based warping method to this deformation. For our result be felt more cartoonish, we additionally apply the luminance quantization method and the edge enhancement method to the deformed target image. At this time, we control intensities of the target image deformation, the luminance quantization and the edge enhancement for the capability that is able to create various results.

TABLE OF CONTENTS

LIST OF TABLES

Table 1. List of Tables

LIST OF FIGURES

Table 2. List of Figures

LIST OF ACRONYMS

<AAM>	Active Appearance Model
<SDLC>	Software Development Life Cycle
<UML>	Unified Modeling Language
<BUA>	Bottom Up Approach
<TDA>	Top Down Approach
<OpenCV>	Open Computer Vision
<FPS>	Frames per second
<CPU>	Control Processing Unit
<GPU>	Graphics Processing Unit
<CG>	Computer Graphics
<RAM>	Random Access Memory
<DFD>	Data Flow Diagram

Table 4. List of Acronyms

1. INTRODUCTION

Cartoons are very popular. The cartoons are not limited to children but adults also provide their liking to them. There is lot of work in recent time in improving sharpness of cartoon images. The real world is portrayed in the cartoon animation. Cartoon whether series or movies are too very popular. There has been research everyday regarding the cartooning improvement and image development. These have been a lot of cartoon image generation method. The cartoons were generally created by artists but due to lot of cartoon works different applications that can create the images into scenes came into act.

Cartoon image plays essential role in our daily lives especially in entertainment, education, advertisement. Cartooning of an image is an interesting project under image processing where it takes an input image, processes it and produces an output as a cartoon. The method of image processing is used to do some processes on a picture like an image enhancement or to remove some functional data from the image.Cartooning of an image is an interesting project under image processing where it takes an input image, processes it and produces an output as a cartoon**.** Cartoon exaggerates target. This fact is a special feature of cartoon and it makes cartoons as the cartoon. But, the exaggeration on the cartoon is very hard to express to every users. So, only cartoon specialists make cartoons. To help novice to easily create a cartoon, studies those create cartoon like images using a computer were progressed, called cartoon rendering. But, most of cartoon rendering method couldn't express their results variously, because their results made by a fixed algorithm. And some other cartoon rendering methods provide various results by textures or user interactions. But, their methods were not intuitive methods. So this project proposes an cartooning method which every users can easily create cartoon-like result were difficult to use to novices. To enhance the shortcoming of previous cartoon rendering method images. we deform the input image using the reference image and apply the cartooning to it. At this time, the user can control the deforming / cartooning intensities of the target image. The main contribution of this project is as follows. It proposes the cartooning method using reference images for every users to generate the result easily. By the controlling of deforming / cartooning intensities, we can create various results from an input image.

1.1 PROJECT OVERVIEW

We propose a methodology for cartooning a image which makes use of AAM searching method by exaggerating reference images. The proposed system utilizes a real image for cartooning the image. It extracts feature points from a target image and define the feature point model on reference images. And then, we apply feature based warping method to this deformation. To make the image more cartoonish, it additionally apply the luminance quantization method and the edge enhancement method to the deformed target image. The system will give the cartoonic image of the given image.

1.2 PROJECT OBJECTIVE

The objective of this project is to use AAM for extracting facial features and warping for giving it a definition then edge exaggeration for making it black and white then the luminance method for adding a color pop. This project aims to make it easier for anyone who is a novice for cartoon image theme and mainly for an artist to make the work easier to create cartoons of a sample image instead of thinking imaginarily. The central focus of this project is to convert the real image to a cartoon-like- image.

1.3 ORGANIZATION OF CHAPTERS

This documentation consists of 10 different chapter and they are:

1. Introduction – This chapter covers the overview of our project and its objectives.
2. Literature Survey – This includes the details of our survey.
3. Software and Hardware Requirements – We specify our software and hardware requirements here.
4. Software Development Analysis – This section includes the problem definition and details of the modules we used in our project.
5. Project System Design – This chapter includes the design part of our project which includes UML diagrams.
6. Project Coding – This section contains the details of our project code.
7. Project Testing – The details of test cases and testing are included in this chapter.
8. Output Screens – This contains the screenshots of how our project looks like when executed.
9. Experimental Results – This chapter contains the screenshots of our results.
10. Conclusion and Future Enhancements – This covers the conclusion of our project and the possible future developments.

2. LITERATURE SURVEY

A literature survey or a literature review in a project report is that section which shows the various analysis and research made in the field of your interest and the results already published, considering the various parameters of the project and the extent of the project.

It is the most important part of our report as it gave us a direction in our research. It helped us set a goal for our analysis - thus giving us our problem statement.

2.1 SURVEY ON BACKGROUND

Library Cartoons: A Literature Review of Library perspective on Library-themed Car y-themed Cartoons, Caricatures, and Comics Julia B. Chambers is a MLIS upand-comer at San Jose State University's School of Library and Information Science. To comprehend contrasting perspectives on past occasions, antiquarians, political theory researchers, and sociologists have examined political and publication kid's shows with topics going from decisions to financial approach to human rights. However sparse examination has been devoted to kid's shows with library topics. The creator of this paper inspects peer-explored writing regarding the matter of library kid's shows, including verifiable foundation, examination of ongoing subjects, and contentions for advancing library-themed kid's shows, exaggerations, and funnies. The creator finds a huge hole in the writing on this theme and presumes that data experts would profit by an extensive substance investigation of library-themed kid's shows to improve comprehension of the essentialness of libraries during noteworthy occasions, survey public view of libraries, and distinguish patterns after some time. Researchers have examined and dissected the impact and estimation of publication kid's shows in the United States since the beginning of the twentieth century, not long after kid's shows turned into a standard element in East Coast papers. In a 1933 article, American craftsmanship and scholarly pundit Elizabeth Luther Cary contended that American exaggeration gave understanding into history, uncovering perspectives or elective mentalities that papers and history books have in any case neglected to record. Twenty years afterward, Stephen Becker (1959), creator of Comic Art in America, agreed that early instances of exaggeration served to make up for editorial shortfalls, in some cases going about as the solitary satisfactory source for editorial excessively indecent or touchy to show up in composed publications. Richard Felton Outcault's Yellow Kid publication kid's shows, distributed in 1896 in the New York World, are one model: "[Yellow Kid] brought something new and disturbing into American homes: the ghettos, and ghetto children, and

customary savagery, and slang, and the arrogance of destitution" (Becker, 1959, p. 13). Contemporary publication kid's shows keep on filling in as an adequate arrangement for circulating disputable perspectives (Kuipers, 2011), frequently with the purpose of influencing public assessment. In an investigation of political kid's shows with official political decision subjects, Edwards and Ware (2005) analyzed the effect of publication kid's shows on open assessment and presumed that negative personifications of electors added to public indifference toward the discretionary cycle. Comparative decisions about the intensity of comic craftsmanship to impact general assessment were accounted for in an examination by Josh Greenberg (2002), whose exploration recommended that kid's shows may assist individuals with interpreting life occasions. Conversely, different researchers have inspected political kid's shows as a reflection of general assessment instead of a provocateur of thought. Anyway the writing, here, presents opposing ends. Edward Holley and Norman Stevens (1969), for example, contend that kid's shows are an exact depiction of public assessment, while others highlight proof showing that kid's shows don't really mirror the overall view nor fill in as opportune pictures of notable occasions (Gilmartin and Brunn, 1998; Meyer, Seidler, Curry, and Aveni, 1980). Concentrated as artistic expressions (Robb, 2009), Zeitgeist ephemera (Holley and Stevens, 1969), essential sources (Thomas, 2004), and even problem solvers (Edwards and Product, 2005; see additionally Marin-Arrese, 2008; Neuberger and Kremar, 2008), article kid's shows have been the subject of investigation in an assortment of scholarly trains. Nonetheless, insufficient exploration has been devoted to the subject of kid's shows or cartoons containing library topics. Indeed, the creator of this writing survey discovered just one investigation, directed by Alireza Isfandyari-Moghaddam and Vahideh Kashi-Nahanji (2010), dedicated to the substance examination of topics in a little choice of library kid's shows, and that review neglected to portray its determination cycle or on the other hand the strategy for content examination utilized. However library-themed kid's shows exist in wealth and go back to the late 1800s. Library kid's shows not just offer a wide scope of critique on custodians, library financing, and the digitization of data, however they likewise give exceptional understanding into the historical backdrop of libraries in the U.S. For instance, an unmistakable reading material utilized in initial library science classes, Foundations of Library and Data Science by Richard E. Rubin (2010), presents a commendatory perspective on Andrew Carnegie's $56 million commitment toward the development of thousands of libraries across America (p. 60). While Rubin takes note of that a few people scrutinized Carnegie's gifts as a type of social control, there is no notice of the public's shock over the taxation rate they made. Nor is there notice of the see held by some that the development of these libraries was simply about Carnegie's personality as opposed to about the public great. However various article kid's shows, for example, the two models underneath, mock Carnegie's

generosity, disparage his self-image, and issue critique on the taxation rate at last delivered by his endowment of public libraries to urban communities around the nation. Notwithstanding the rich history of library kid's shows, many examination inquiries concerning kid's shows containing library subjects have never been tended to in the writing. For example, what were probably the most punctual library kid's shows in this nation? What were normal topics? Have the subjects changed throughout the long term – particularly since the Internet turned into a broad examination device? Most significantly, does the investigation of library kid's shows matter?

2.2 CONCLUSIONS ON SURVEY

This survey of academic writing on the subject of library kid's shows distinguishes past regions of study, features some topical patterns, and contends that a far reaching content examination of library-themed kid's shows would add to the field of library science similarly that researchers indifferent orders have utilized publication kid's shows to enhance their comprehension of notable occasions, investigate public discernment, and recognize patterns.

3. SOFTWARE AND HARDWARE REQUIREMENTS

Requirement is a condition or capability possessed by the software or system component in order to solve a real world problem. The problems can be to automate a part of a system, to correct shortcomings of an existing system, to control a device, and so on.

Requirements describe how a system should act, appear or perform. For this, when users request for software, they provide an approximation of what the new system should be capable of doing. Requirements differ from one user to another and from one business process to another.

The purpose of the requirements document is to provide a basis for the mutual understanding between the users and the designers of the initial definition of the software development life cycle (SDLC) including the requirements, operating environment and development plan.

Requirements help to understand the behavior of a system, which is described by various tasks of the system. For example, some of the tasks of a system are to provide a response to input values, determine the state of data objects, and so on. Note that requirements are considered prior to the development of the software. The requirements, which are commonly considered, are classified into three categories, namely, functional requirements, non-functional requirements, and domain requirements.

The functional requirements should be complete and consistent. Completeness implies that all the user requirements are defined. Consistency implies that all requirements are specified clearly without any contradictory definition. Generally, it is observed that completeness and consistency cannot be achieved in large software or in a complex system due to the problems that arise while defining the functional requirements of these systems. The different needs of stakeholders also prevent the achievement of completeness and consistency. Due to these reasons, requirements may not be obvious when they are first specified and may further lead to inconsistencies in the requirements specification.

The non-functional requirements (also known as **quality requirements)** are related to system attributes such as reliability and response time. Non-functional requirements arise due to user requirements, budget constraints, organizational policies, and so on. These requirements are not related directly to any particular function provided by the system.

Non-functional requirements should be accomplished in software to make it perform efficiently. For example, if an aero plane is unable to fulfill reliability requirements, it is not approved for safe operation. Similarly, if a real time control system is ineffective in accomplishing non-functional requirements, the control functions cannot operate correctly.

System requirements are the configuration that a system must have in order for a hardware or software application to run smoothly and efficiently. Failure to meet these requirements can result in installation problems or performance problems. The former may prevent a device or application from getting installed, whereas the latter may cause a product to malfunction or perform below expectation or even to hang or crash.

System requirements are also known as minimum system requirements.

Hardware system requirements often specify the operating system version, processor type, memory size, available disk space and additional peripherals, if any, needed. Software system requirements, in addition to the requirements, may also specify additional software dependencies (e.g., libraries, driver version, frame work version). Some hardware/software manufacturers provide an upgrade assistant program that users can download and run to determine whether their system meets a product's requirements.

Some products include both minimum and recommended system requirements. A video game, for instance, may function with the minimum required CPU and GPU, but it will perform better with the recommended hardware. A more powerful processor and graphics card may produce improved graphics and faster frame rates (FPS).
Some system requirements are not flexible, such as the operating system(s) and disk space required for software installation. Others, such as CPU, GPU, and RAM requirements may vary significantly between the minimum and recommended requirements. When buying or upgrading a software program, it is often wise to make sure your system has close to the recommended requirements to ensure a good user experience.

3.1 SOFTWARE REQUIREMENTS

- Operating System : Windows XP.
- Platform : PYTHON TECHNOLOGY
- Tool : Python 3.9
- Back End : Python IDLE 3.9

3.2 HARDWARE REQUIREMENTS

- System : Pentium IV 2.4 GHz.
- Hard Disk : 40 GB.
- Monitor : 15 inch VGA Color.
- Mouse : Logitech Mouse.
- Ram : 512 MB
- Keyboard :Standard Keyboard

4. SOFTWARE DEVELOPMENT ANALYSIS

Software development is a process of writing and maintaining the source code, but in a broader sense, it includes all that is involved between the conception of the desired software through to the final manifestation of the software, sometimes in a planned and structured process. Therefore, software development may include research, new development, prototyping, modification, reuse, re-engineering, maintenance, or any other activities that result in software products.

4.1 OVERVIEW OF THE PROBLEM

Moreover, as evident from various researches in the field of image classification, unlike traditional methods like multilayer perceptron, Convolutional Neural Networks extracts successively larger features in a hierarchical set of layers and thus, is best suited for our problem consisting of images with a large number of variations. Cartoon exaggerates target. This fact is a special feature of cartoon and it makes cartoons as the cartoon. But, the exaggeration on the cartoon is very hard to express to every users. So, only cartoon specialists make cartoons. To help novice to easily create a cartoon, studies those create cartoon like images using a computer were progressed, called cartoon rendering. But, most of cartoon rendering method couldn't express their results variously, because their results made by a fixed algorithm. And some other cartoon rendering methods provide various results by textures or user interactions. But, their methods were not intuitive methods. So they were difficult to use to novices. There are many systems proposed earlier for cartoon image in similar ways but each and every has their limitations.

4.2 DEFINING THE PROBLEM

The cartoon of an image takes a real image as input and then converts the loaded image into cartoon-like images by using target images. To deform into a cartoon image, we extract feature points from real image apply some techniques to get cartoonish effect on real image. The cartoon of an image – converts the given sample image to a cartoonish filter image initially by using warping method. AAM searching algorithm can be used to extract feature points from the target image. Edge exaggeration and luminous quantization are methods used to convert coloured image with blur effect. Finally, we can use adaptive threshold technique to achieve a pure cartoonic image. The initial goal is to display a window to upload a picture which is saved in user's device. The major goal is then to display a converted cartoonish image along with the real image.

For this, we first convert the image to gray – scale and then we apply the media blur filter. Next step is to identify the edges in the image using Edge Detection. Adaptive threshold technique is used there-after to suppress non maximum points and to highlight edge points.

4.3 MODULES OVERVIEW

1) AAM Searching

AAM is an efficient fitting algorithm for extracting features. The goal of the AAM search is to find the model parameters that generate a synthetic image as close as possible to a given input image and to use the resulting AAM parameters for interpretation. Fitting the model and the target image is treated as a nonlinear optimization problem. Therefore, the fitting task requires a huge amount of computation when the standard nonlinear optimization techniques such as the gradient descent method are used.

2) Warping

Image warping is the process of digitally manipulating an image such that any shapes portrayed in the image have been significantly distorted. Warping may be used for correcting image distortion as well as for creative purposes. The same techniques are equally applicable to video. This technique make use of pre-defined reference images to form a deformed image which is used for further process. These pre-defined reference images are fixed sample images.

3) Edge Exaggeration

Edge enhancement is an image processing filter that enhances the edge contrast of an image or video in an attempt to improve its apparent sharpness. The filter works by identifying sharp edge boundaries in the image, such as the edge between a subject and a background of a contrasting colour, and increasing the image contrast in the area immediately around the edge. This has the effect of creating subtle bright and dark highlights on either side of any edges in the image, called overshoot and undershoot, leading the edge to look more defined when viewed from a typical viewing distance. It is also widely used in computer printers especially for font or/and graphics to get a better printing quality. Edge enhancement can be either an analog or a digital process. Analog edge enhancement may be used, for example, in all-analog video equipment such as modern CRT televisions.

4) Bilateral Filtering

A bilateral filter is a non-linear, edge-preserving, and noise-reducing smoothing filter for images. It replaces the intensity of each pixel with a weighted average of intensity values from nearby pixels. This weight can be based on a Gaussian distribution. Crucially, the weights depend not only on Euclidean distance of pixels, but also on the radiometric differences (e.g., range differences, such as color intensity, depth distance, etc.). This preserves sharp edges. This filter is the key element in the color image processing chain, as it homogenizes color regions while preserving edges, even over Getting a blurred version of the original image. For this, we first convert the image to gray – scale and then we apply the media blur filter.

4.3 DEFINING THE MODULES

The project mainly consists of 4 modules :

1) Numpy module

2) OpenCV module

3) Upload Image module

4) Result module

4.4 MODULE FUNCTIONALITY

1) Numpy Module – NumPy (Numerical Python) is an open-source library for the Python programming language. It is used for scientific computing and working with arrays. We have to import Numpy in python in order to make use of opencv.

2) OpenCV Module - OpenCV is a huge open-source library for computer vision it can process images and videos to identify objects, faces, or even the handwriting of a human. It is integrated with Numpy library for numerical operations, whatever operations one can do in Numpy can be combined with OpenCV.

3)Upload Image Module -The system pop ups a window on screen to upload a image for process.

4)Result Module -This displays the output as a cartoon like image side by the original image.

5. PROJECT SYSTEM DESIGN

System design is the process of defining elements of a system like modules, architecture, components and their interfaces and data for a system based on the specified requirements. It is the process of defining, developing and designing systems which satisfies the specific needs and requirements of a business or organization. A systemic approach is required for a coherent and well-running system. Bottom-Up or Top-Down approach is required to take into account all related variables of the system. A designer uses the modeling languages to express the information and knowledge in a structure of system that is defined by a consistent set of rules and definitions. The designs can be defined in graphical or textual modeling languages.

Unified Modeling Language has been used by us to describe software structurally with notations.

SYSTEM ARCHITECTURE

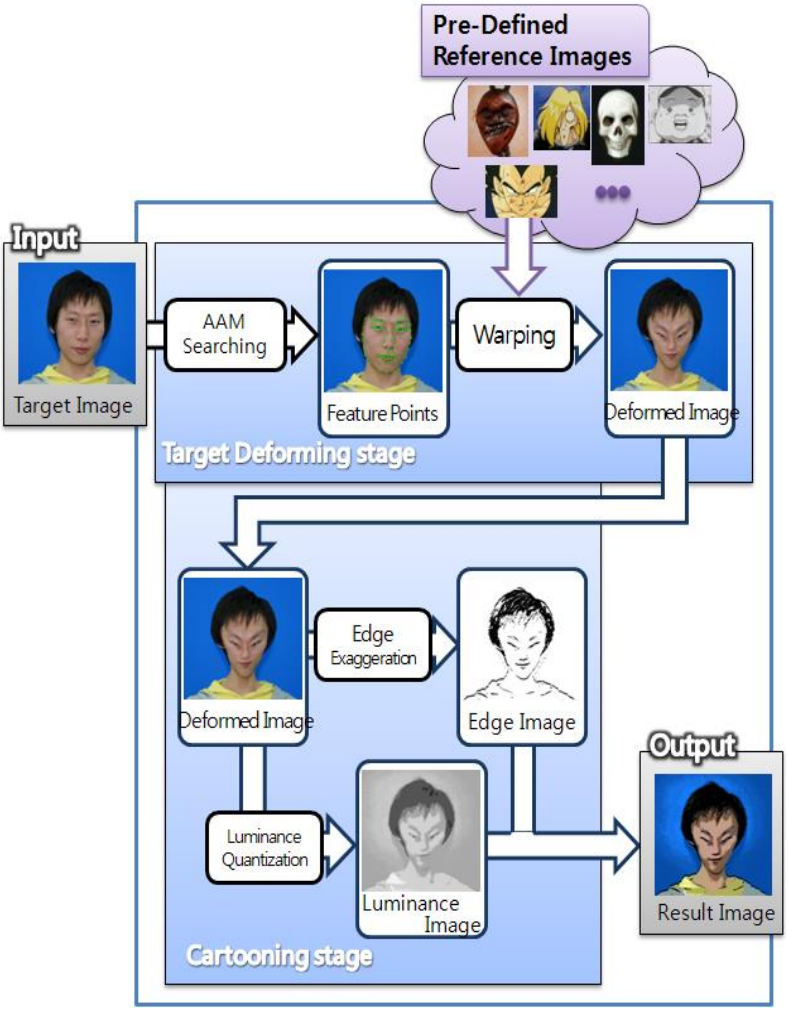


Fig 5.1 System Architectur

Downscale the image and then apply bilateral filter to get a cartoon flavour. Then again we upscale the image. Bilateral Filter: This filter is the key element in the colour image processing chain, as it homogenizes colour regions while preserving edges, even over getting a blurred version of the original image. For this, we first convert the image to gray – scale and then we apply the media blur filter. Next step is to identify the edges in the image using Edge Detection.

Edge Detection- Process of identifying edges in an image to be used as a fundamental asset in image analysis and locating areas with strong intensity contrasts.

The Open Computer Vision Library (OpenCV) provides a standard toolkit for performing basic and complex image processing algorithms like these.

5.1 DATA FLOW DIAGRAMS

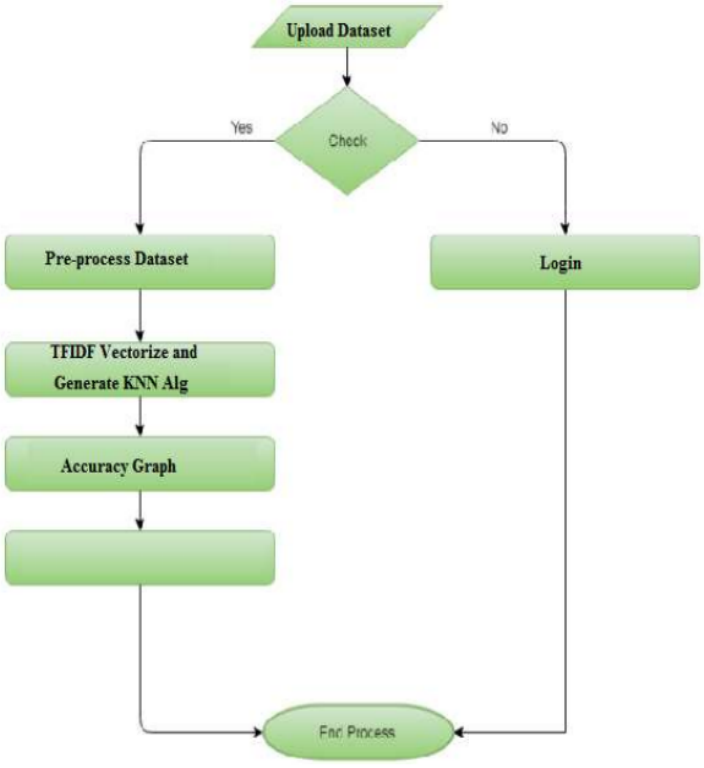


Fig 5.2 Data Flow Diagram

Data flow diagrams are used to graphically represent the flow of data in a business information system. DFD describes the processes that are involved in a system to transfer data from the input to the file storage and reports generation. Data flow diagrams can be divided into logical and physical. The logical data flow diagram describes flow of data through a system to perform certain

functionality of a business. The physical data flow diagram describes the implementation of the logical data flow.

DFD graphically representing the functions, or processes, which capture, manipulate, store, and distribute data between a system and its environment and between components of a system. The visual representation makes it a good communication tool between User and System designer. Structure of DFD allows starting from a broad overview and expand it to a hierarchy of detailed diagrams. DFD has often been used due to the following reasons.

5.2 UML DIAGRAMS

UML is an acronym that stands for Unified Modeling Language. Simply put, UML is a modern approach to modeling and documenting software. In fact, it‘s one of the most popular business process modeling techniques.

It is based on diagrammatic representations of software components. As the old proverb says: ―a picture is worth a thousand words‖. By using visual representations, we are able to better understand possible flaws or errors in software or business processes.

UML was created as a result of the chaos revolving around software development and documentation. In the 1990s, there were several different ways to represent and document software systems. The need arose for a more unified way to visually represent those systems and as a result, in 1994-1996, the UML was developed by three software engineers working at Rational Software. It was later adopted as the standard in 1997 and has remained the standard ever since, receiving only a few updates.

GOALS:

The Primary goals in the design of the UML are as follows:

1. Provide users a ready-to-use, expressive visual modeling Language so that they can develop and exchange meaningful models.

2. Provide extendibility and specialization mechanisms to extend the core concepts.

3. Be independent of particular programming languages and development process.

4. Provide a formal basis for understanding the modeling language.

5. Encourage the growth of OO tools market.

CLASS DIAGRAM

In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.

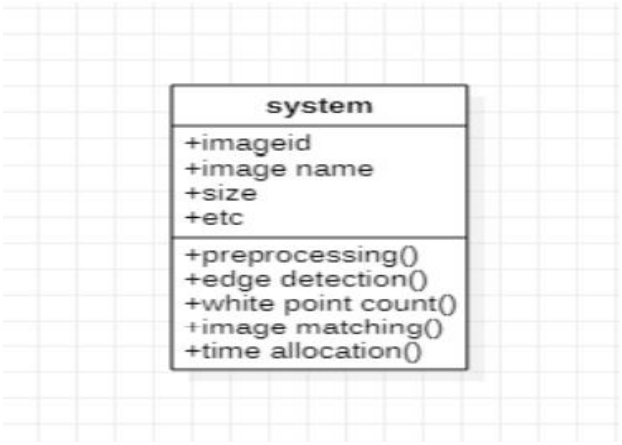


Fig 5.3 Class Diagram

USE CASE DIAGRAM

The use case diagram of the project cartooning of an image is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.

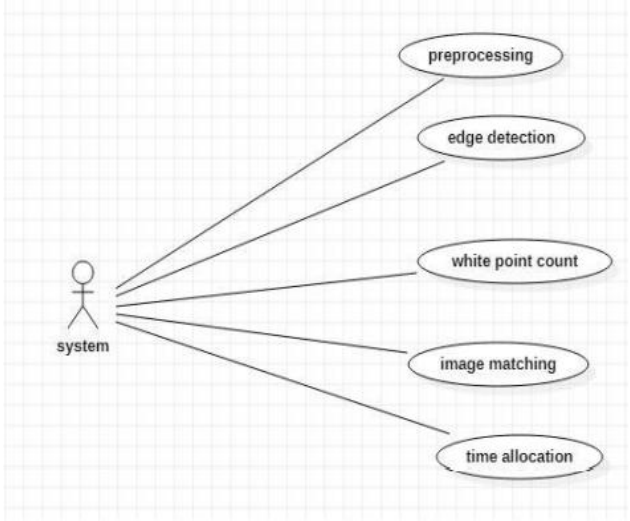


Fig 5.4 Use-case Diagram

SEQUENCE DIAGRAM

The sequence diagram of the project cartooning of an image in Unified Modeling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. It simply depicts the interaction between the objects in a sequential order, so in our project this sequence diagram shows how the sequence of objects functions.

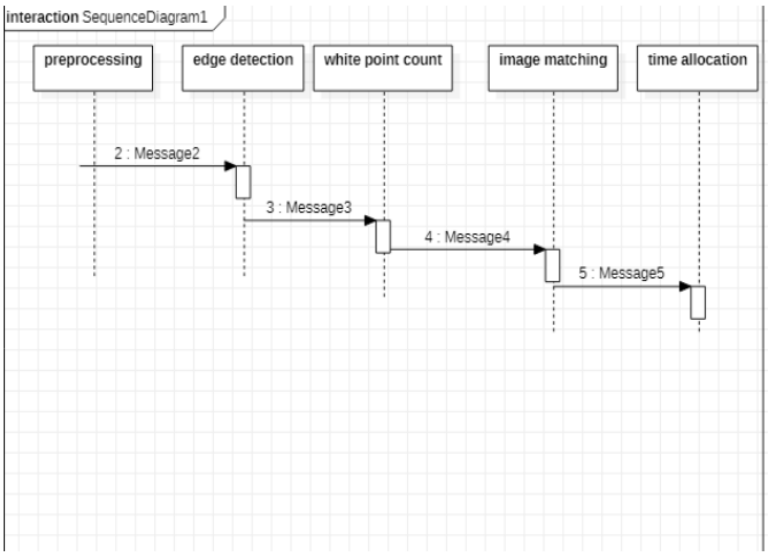


Fig 5.5 Sequence Diagram

COMPONENT DIAGRAM

Component diagram is a special kind of diagram in UML. The purpose is also different from all other diagrams discussed so far. It does not describe the functionality of the system but it describes the components used to make those functionalities. Thus from that point of view, component diagrams are used to visualize the physical components in a system. These components are libraries, packages, files, etc. Component diagrams can also be described as a static implementation view of a system. Static implementation represents the organization of the components at a particular moment. A single component diagram cannot represent the entire system but a collection of diagrams is used to represent the whole. UML Component diagrams are used in modeling the physical aspects of object-oriented systems that are used for visualizing, specifying, and documenting component-based systems and also for constructing executable systems through forward and reverse engineering. Component diagrams are essentially class diagrams that focus on a system's components that often used to model the static implementation view of a system.

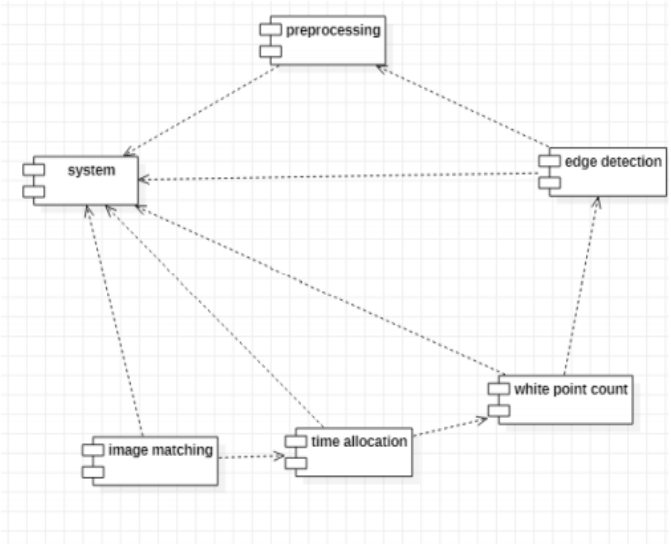


Fig 5.6 Component Diagram

ACTIVITY DIAGRAM

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modeling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.

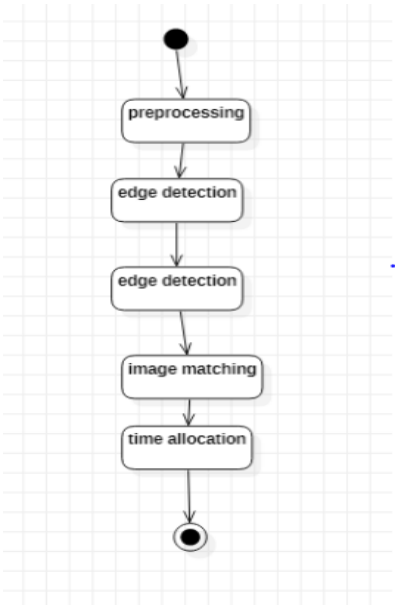


Fig 5.7 Activity Diagram

6. PROJECT CODING

Project Coding is the process of designing and building an executable computer program to accomplish a specific computing result or to churn out a particular prototype or product. Programming involves tasks such as: analysis, generating algorithms, profiling algorithms' accuracy and resource consumption, and the implementation of algorithms in a chosen programming language (commonly referred to as coding). The source code of a program is written in one or more languages that are intelligible to programmers, rather than machine code, which is directly executed by the central processing unit. The purpose of programming is to find a sequence of instructions that will automate the performance of a task (which can be as complex as an operating system) on a computer, often for solving a given problem. Proficient programming thus often requires expertise in several different subjects, including knowledge of the application domain, specialized algorithms, and formal logic.

6.1 CODE TEMPLATES

1. Open up the Command Prompt and run the python interpreter as shown in the screenshot below.

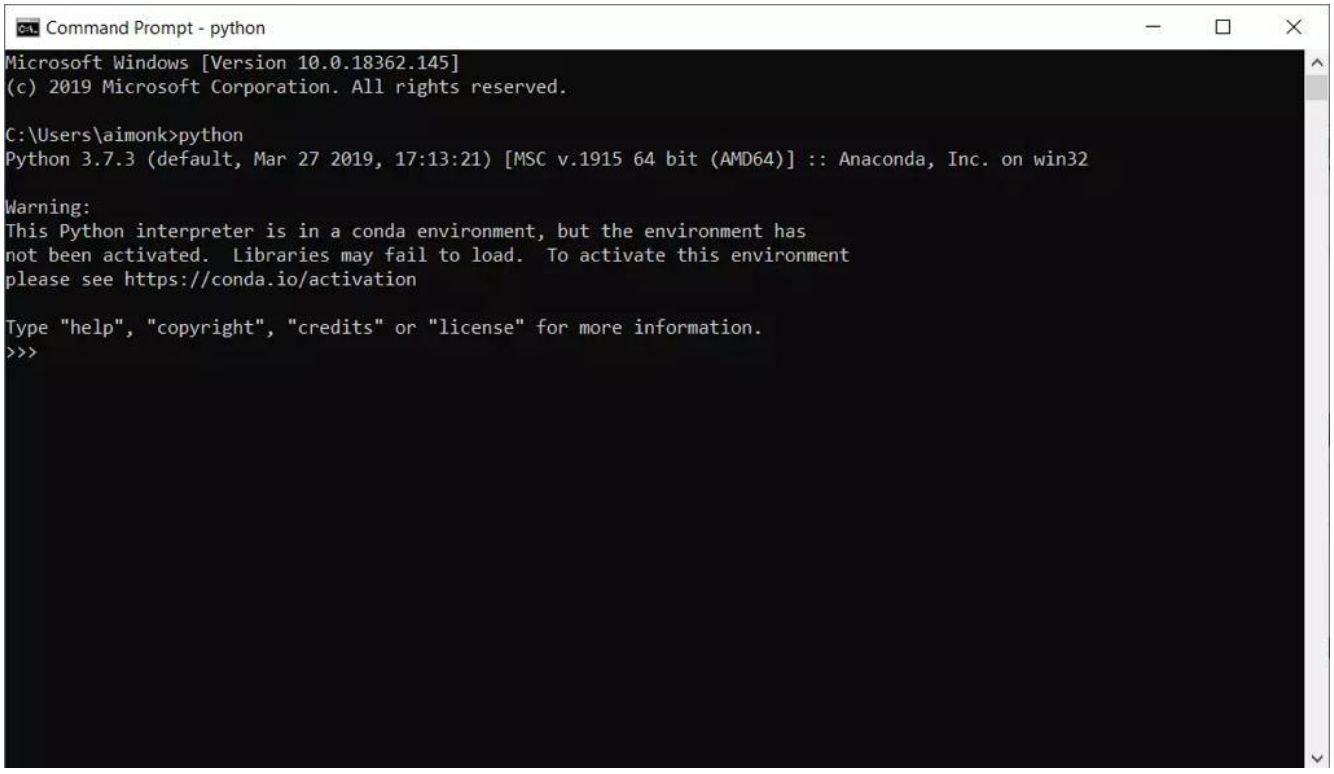


Fig 6.1 Code Template

2. Import the opencv library and print the version of opencv as shown in the next screenshot.

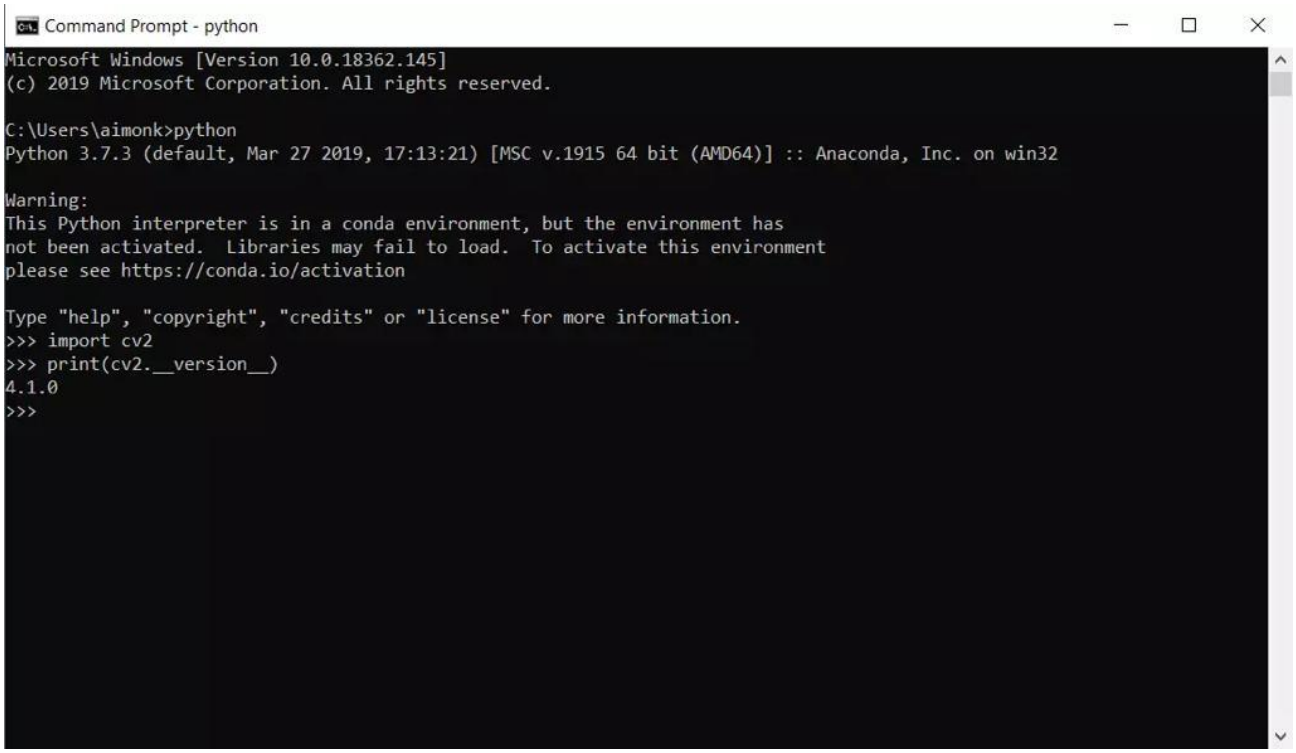


Fig 6.1.1 Code Template

3. Code snippet of the project code.

```
import cv2
import numpy as np
from tkinter.filedialog import *

photo = askopenfilename()
img = cv2.imread(photo)

grey = cv2.cvtColor(img, cv2.COLOR_BGR2GRAY)
grey = cv2.medianBlur(grey, 5)
edges = cv2.adaptiveThreshold(
    grey, 255, cv2.ADAPTIVE_THRESH_MEAN_C, cv2.THRESH_BINARY, 5, 4)
# cartoonize
color = cv2.bilateralFilter(img, 9,200,200)
cartoon = cv2.bitwise_and(color, color, mask=edges)

cv2.imshow("Image", img)
cv2.imshow("Cartoon", cartoon)
# save
x=cv2.imwrite("cartoon.jpg", cartoon)
y=cv2.resize(x,(100,100))
cv2.waitKey(0)
cv2.destroyAllWindows()
```

Fig 6.1.2 Code Template

6.2 OUTLINE FOR VARIOUS FILES

We used Python programming to implement our project which an open source software. We imported numpy module and cv2 library from python IDLE in latest version. Our code consists of various modules that we have used. Our project module is Upload a image for getting the result.

6.3 METHODS INPUT AND OUTPUT PARAMETERS

We implemented multiple methods, few of which are :

1. askopenfilename()

2. cv2.adaptiveThreshold()

3. cv2.bilateralFilter()

4. cv2.imshow()

5. cv2.medianBlur(), etc

Our first method askopenfilename() takes in the images and for processing recognition. Cv2.adaptiveThreshold() is used to control the black point feature on the given image. Cv2.bilateralFilter() is used for image scaling while feature extraction.

The cv2.medianBlur() adds a blur effect for the image using a grey scale point. The next method cv2.imshow() is used to save the input image and display the output image.

7. PROJECT TESTING

Project Testing is a method to check whether the actual software product matches expected requirements and to ensure that software product is Defect free. It involves execution of software/system components using manual or automated tools to evaluate one or more properties of interest. The purpose of software testing is to identify errors, gaps or missing requirements in contrast to actual requirements.

Some prefer saying Software testing as a White Box and Black Box Testing. In simple terms, Software Testing means the Verification of Application Under Test (AUT). This tutorial introduces testing software to the audience and justifies its importance.

Project testing is important because, if there are any bugs or errors in the software, it can be identified early and can be solved before delivery of the software product. Properly tested software product ensures reliability, security and high performance which further results in time saving, cost effectiveness and customer satisfaction.

Typically Testing is classified into three categories.

- Functional Testing
- Non-Functional Testing or Performance Testing
- Maintenance (Regression and Maintenance)

7.1 VARIOUS TEST CASES

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub-assemblies, assemblies and/or a finished product. It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

We have performed multiple tests under the broad categorisation of white-box and black-box testing which further include unit, integration, boundary value and statement covering etc.

TYPES OF TESTS

Unit testing:

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

Integration testing:

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

Functional test:

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items:

Valid Input : identified classes of valid input must be accepted.

Invalid Input : identified classes of invalid input must be rejected.

Functions : identified functions must be exercised.

Output : identified classes of application outputs must be exercised.

Systems/Procedures : interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

System Test:

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

7.2 BLACK BOX TESTING:

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box .you cannot ―see‖ into it. The test provides inputs and responds to outputs without considering how the software works.

7.3 WHITE BOX TESTING:

White Box Testing is a testing in which in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is purpose. It is used to test areas that cannot be reached from a black box level.

Test strategy and approach

Field testing will be performed manually and functional tests will be written in detail.

Test objectives

- All field entries must work properly.
- Pages must be activated from the identified link.
- The entry screen, messages and responses must not be delayed.

Features to be tested

- Verify that the entries are of the correct format
- No duplicate entries should be allowed
- All links should take the user to the correct page.

Integration Testing:

Software integration testing is the incremental integration testing of two or more integrated software components on a single platform to produce failures caused by interface defects. The task of the integration test is to check that components or software applications, e.g. components in a software system or – one step up – software applications at the company level – interact without error.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

Acceptance Testing:

User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements. Test Results: All the test cases mentioned above passed successfully. No defects encountered.

8. OUTPUT SCREENS

An output screen is a device used to display output. An output screen could be a separate monitor or another display device used only to display the output being received from the computer or other devices.

Here, in the screen prints given below, we can see that the user interface screens consist of the home page which describes the purpose of the portal, and the public access screen which allows users to upload the credentials of the child that they found in order to check whether the child exists in the repository or not.

8.1 USER INTERFACE

1. The below window is of user interface where the user will be compile and running the program**.**

Fig 8.1 User Interface

8.2 OUTPUT SCREEN

2. After running the program this window will pop up on the screen to take an input.

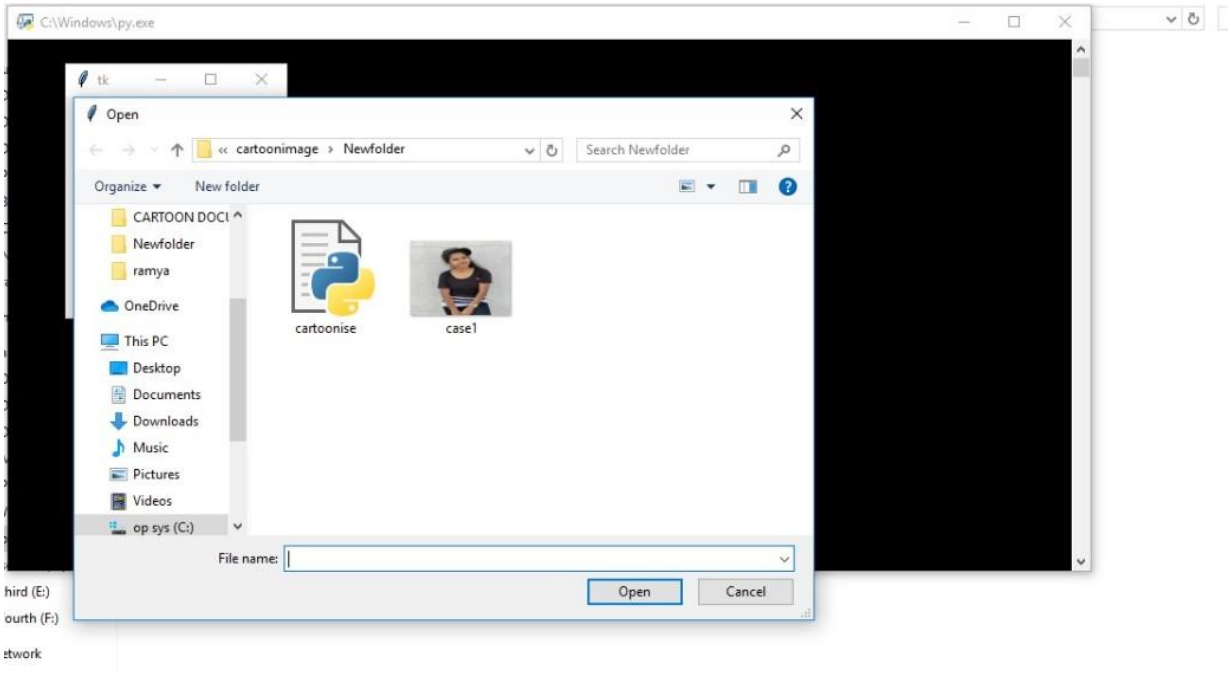


Fig 8.2 Output Screen

9. EXPERIMENTAL RESULT

1) The displayed image will be the output of the project which will look like a cartoon image of a basic image.

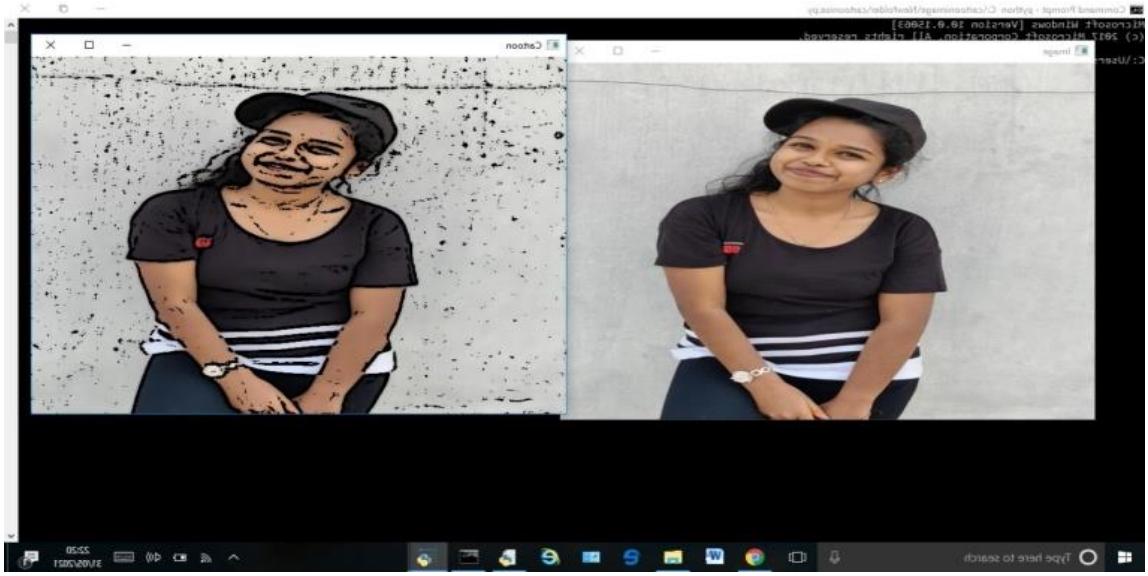


Fig 9.1 Result Displayed

2) The result is like cartoon image for a given image of different grey scale and black point.

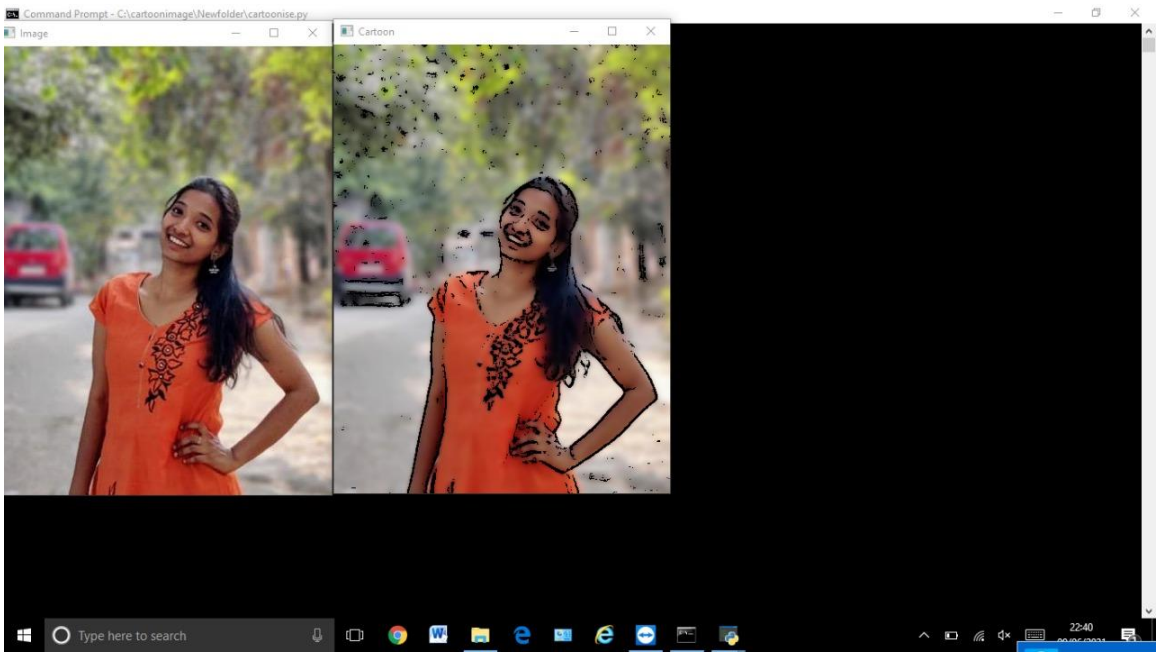


Fig 9.2 Result Displayed

10. CONCLUSION AND FUTURE ENHANCEMENT

Cartooning of an image is an efficient method for extracted cartoon objects. The test results show that the developed method could extract meaningful objects well in different characters and backgrounds. The extracted cartoon objects expect to be effectively used in cartoon image retrieval because they can represent the color characteristics of cartoon objects well. The future enhancement of this paper could be adding an extra feature of live capturing of images and them use them for cartooning.

This project proposed an cartooning method using reference image. It generates the result image by deforming the target to the form of the reference image and cartoons on it. It is very easy to use to every user. By adjusting the deformation intensity of the target and the cartooning intensity, users can generate various result images.

The proposal of this paper has a restriction of algorithm that the feature point model of the reference image has to be predefined and provided. And it also has the limit that the deformation for the exaggeration of the target image can be applied only to the frontal face. The improving method of these problems is being researched now.

If the system of this paper adopts the model extracting feature points from the face of diverse angles through the expansion of AAM in the future, it will be able to deform the specific target on the video input. And the research to deform a target image by using several reference images is also in the plan.

11.REFERENCES

[1] Mohapatra, H.; Rath, A. Advancing generation Z employability through new forms of learning: quality assurance and recognition of alternative credentials; ResearchGate, 2020.

[2] Mohapatra, H.; Rath, A. K. Fundamentals of software engineering: Designed to provide an insight into the software engineering concepts; BPB, 2020.

[3] The Comic Culture, Anime is NOT a cartoon - The Comic Culture – Medium, May.1, 2016.Accessed on: Dec.29 2020 [online]. Available:https://medium.com/@RichardSeghers/anime-is-not-acartoon-95bcfb416431

[4] Haoxiang Li, Zhe Lin, Xiaohui Shen, Jonathan Brandt, Gang Hua, “A Convolutional Neural Network for Cascade Face Dectection,” The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5325-5334.

[5] Brachmann Anselm, Erhardt Barth, Christoph Redies, Brachmann, “Using CNN Features to Better Understand What Makes Visual Artworks Special,” Front. Psychol. (2017).

[6] Karayev, Sergey, Matthew Trentacoste, Helen Han, Aseem Agarwala, Trevor Darrell, Aaron Hertzmann, and Holger Winnemoeller, "Recognizing image style," arXiv preprint arXiv:1311.3715 (2018).

[7] Gatys, Leon A., Alexander S. Ecker, and Matthias Bethge, "Image style transfer using convolutional neural networks," IEEE conference on computer vision and pattern recognition(CVPR), pp. 2414-2423. 2016,.in press.

[8] Martinsson, Hans, and Tor Sandstrom, "Gray scaling in high performance mask making," In Photomask and Next-Generation Lithography Mask Technology XII, vol. 5853, pp. 1031-1042. International Society for Optics and Photonics, 2017.

[9] Khotanzad, Alireza, and C. Chung, "Application of multi-layer perceptron neural networks to vision problems," Neural Computing & Applications 7, no. 3(1998): 249-259.

[10] Wang, Fei, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. "Residual attention network for image classification." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3156-3164.

12.PUBLICATIONS

CONFERENCE:

International Conference on “Innovations in Computers Networks, Computational Intelligence and IoT” (ICICCI – 21)

Paper ID : ICICCI – 21 – 0044

13.STUDENT PROFILES

S.Alekhya is a Bachelor of Technology student at St. Martin's Engineering College studying Computer Science and Engineering. She finished her schooling from Bhashyam Public School , Intermediate From Sri Chaitanya Junior College. Python and C are among her technical skills. She also knows the fundamentals of C++, SQL. Her participations include: Women online workshop on “Women in Cyber Security and Privacy in 2020” from the 6th to the 10th of July 2020, HTML & CSS workshop conducted by TAMv5 in 2018. She spends her free time taking online certification courses related to her field of study as well as personal interests from platforms such as Coursera and Great Learning she took part in Employability Skill Development Program conducted by Zensar. She is interested towards learning new things and aspires to build a career path in Software Testing in future.

Kandula Bhavyasri is currently pursuing her Bachelor of Technology in the stream of Computer Science And Engineering at St.Martin's Engineering College. She completed her 11^{th} and 12^{th} grades from National Institute of Open schooling and 10th class from St Peter's Grammar School. Her technical skills include C, Java and Python. She also has a basic understanding of C++. She took part in E-Summit program conducted at Marri Laxman Reddy Institute Of Technology in 2018 and completed few certification courses from online platforms like Coursera and Free learning.

Saipriya challagulla is currently pursuing her Bachelor of Technology in the stream of Computer Science And Engineering at St.Martin's Engineering College. She completed her intermediate from sri chaitanya college and 10th class from The creek planet School. Her technical skills include C, Python. She also has a basic understanding of C++.She participated in TECHGYAN workshop conducted in IIT-H, in the field of IoT. She is passionate towards network essentials and cloud computing. She has completed few certification courses from online platform like coursera.

Terala Ramya is a student at St. Martin's Engineering College, pursuing Bachelor of Technology in Computer Science and Engineering, as well as an Software Developer at HCL Technologies which is a leading multinational company. She completed her Intermediate Education from Sri Chaitanya Junior College and schooling from TVR Model High School. Her technical skills include C, C++, Java, SQL and Python. . Her participations include: Women online workshop on “Women in Cyber Security and Privacy in 2020” from the 6th to the 10th of July 2020. She is passionate towards learning new things and aspires to become a Full Stack Web Developer in the near future. She spends her free time taking online certification courses related to her field of study as well as personal interests from platforms such as Coursera and CursaApp.

14. APPENDICES

Appendix A: User Requirements Questionnaire

This research will be used for academic purpose only. Its main objective is to collect the user requirements to create a cartoon images. Kindly provide your honest answers in the following questions. Please note that your responses will be treated as private and confidential.

1. Cartoon like images are fun to create.
   - ϒ Strongly Agree
   - ϒ Agree
   - ϒ Neutral
   - ϒ Disagree
   - ϒ Strongly Disagree
2. The current process of getting a cartoon like image for normal image is efficient.
   - ϒ Strongly Agree
   - ϒ Agree
   - ϒ Neutral
   - ϒ Disagree
   - ϒ Strongly Disagree
3. The final output is similar to the expected cartoon image.
   - ϒ Strongly Agree
   - ϒ Agree
   - ϒ Neutral
   - ϒ Disagree
   - ϒ Strongly Disagree

4. The current process of converting image is user friendly.

ϒ Strongly Agree
ϒ Agree
ϒ Neutral
ϒ Disagree
ϒ Strongly Disagree

5. I believe that the current processes and systems, for converting a cartoon image are very simple yet stores by default.

ϒ Strongly Agree
ϒ Agree
ϒ Neutral
ϒ Disagree
ϒ Strongly Disagree

6. If a proper computer system is implemented, I believe that converting (persons) would be made easier.

ϒ Strongly Agree
ϒ Agree
ϒ Neutral
ϒ Disagree
ϒ Strongly Disagree

7. How sensitive are you try new weird things for fun on yourself?

ϒ Very Sensitive
ϒ Not Very Sensitive
ϒ Neutral

Appendix B: System Usability Questionnaire

This research will be used for academic purpose only. Its main objective is to find out users' experience in Cartoon Images . Kindly provide your honest opinion on the same. Please note that your responses will be treated as private and confidential.

System Usability Scale:

Kindly rate the child tracing prototype about the following:

1. The user interface is very user friendly.

ϒ Strongly Agree

ϒ Agree

ϒ Neutral

ϒ Disagree

ϒ Strongly Disagree

2. I can use this prototype with the minimum training.

ϒ Strongly Agree

ϒ Agree

ϒ Neutral

ϒ Disagree

ϒ Strongly Disagree

3. By using this system converting an image will take a shorter duration as compared to the current methods.

ϒ Strongly Agree

ϒ Agree

ϒ Neutral

ϒ Disagree

ϒ Strongly Disagree

4. This question is practical and aims at testing the accuracy of the prototype. Kindly upload an image of a person/child to the system. After the researcher inputs them into the system, try changing the values of the grey-scale and black points of the conversion and note down what changes did you observe(Among them, how many were similarly looking like a cartoon ?)

……..

5. The system provides a convenient way of converting an image into cartoon like image.

ϒ Strongly Agree
ϒ Agree
ϒ Neutral
ϒ Disagree
ϒ Strongly Disagree

6. I will use this system to make fun of my friends.

ϒ Strongly Agree
ϒ Agree
ϒ Neutral
ϒ Disagree
ϒ Strongly Disagree

7. How likely are you to recommend this system to the other users?

ϒ Very Likely
ϒ Likely
ϒ Neutral
ϒ Not Likely
ϒ Not Likely At All

8. Any Comments

……
……
……
……
……
……
……
……
……
……
……
……

Appendix C: Interview Questions

Interview Questions

This research is will be used for academic purpose only. Its main objective is to find out users' experience in using the child tracing prototype. Kindly provide your honest opinion on the same. Please note that your responses will be treated as private and confidential.

Interviewee: ……. **Location:** …….

Medium: …….. **Date:** ……

1. If a child/person want a cartoon image of normal image, what should an ordinary citizen do?

   ……
   ……
   ……
   ……
   ……
   ……
   ……
   ……
   ……

2. What is the current conversion process?

   ……
   ……
   ……
   ……
   ……
   ……
   ……
   ……
   ……
   ……

3. What are the challenges did you face through the process?

……
……
……
……
……
……
……
……

4. What is the difference between cartoon and an anime? Are they alike?

……
……
……
……
……
……
……
……

5. What is the process running in the backend of conversion?

……
……
……
……
……
……
……
……

6. What are the challenges in the backend of conversion?

……
……
……
……
……
……
……
……

7. In your opinion, what do you think needs to be improved in the process of:
    a. May be add more features?

……
……
……
……
……
……
……
……
……
……
……

    b. Does converted look like cartoon?

……
……
……
……
……
……
……
……
……
……

A

PROJECT REPORT

On

BEHAVIOR ANALYSIS FOR MENTALLY AFFECTED PEOPLE

Submitted by

1)B. Jessica Dolly(17K81A05K0) **2)P. Malla Reddy(17K81A05M4)**

3)P. Nikhil Reddy(17K81A05M6) **4)T. Gowri(17K81A05N7)**

in partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Under the Guidance of

Mrs. M. Naga Triveni

Assistant Professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ST. MARTIN'S ENGINEERING COLLEGE
An Autonomous Institute

Dhulapally, Secunderabad – 500 100

JUNE 2021

BONAFIDE CERTIFICATE

This is to certify that the project entitled "Behavior Analysis for Mentally Affected People", is being submitted by **B. Jessica Dolly (17K81A05K0), P. Malla Reddy (17K81A05M4), P. Nikhil Reddy (17K81A05M6) and T. Gowri (17K81A05N7)** in partial fulfillment of the requirement for the award of the degree of **BACHELOR OF TECHNOLOGY IN COMPUTER SCIENCE AND ENGINEERING** is recorded of bonafide work carried out by them. The result embodied in this report have been verified and found satisfactory.

Assistant Professor
Mrs. M. Naga Triveni
Department of CSE

Head of the Department
Dr.M.NARAYANAN
Department of CSE

Internal Examiner

External Examiner

Place:

Date:

DECLARATION

We, the student of **Bachelor of Technology** in Department of '**Computer Science and Engineering**', session: 2017 – 2021, St. Martin's Engineering College, Dhulapally, Kompally, Secunderabad, hereby declare that work presented in this Project Work entitled "Behavior Analysis for Mentally Affected People" is the outcome of our own bonafide work and is correct to the best of our knowledge and this work has been undertaken taking care of Engineering Ethics. This result embodied in this project report has not been submitted in any university for award of any degree.

B. Jessica Dolly (17K81A05K0)
P. Malla Reddy (17K81A05M4)
P. Nikhil Reddy (17K81A05M6)
T. Gowri (17K81A05N7)

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible and whose encouragements and guidance have crowded effects with success.

We extended our deep sense of gratitude to Principal**,** **Dr. P. SANTOSH KUMAR PATRA**, St. Martin's Engineering College, Dhulapally, for permitting us to undertake this project.

We are also thankful to **Dr. M.NARAYANAN**, Head of the Department, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

We would like to thank **Dr. T. POONGOTHAI,** Department of Computer Science and Engineering (AI & ML) for her encouragement and insightful comments and as well as our project coordinators **Dr. B.RAJALINGAM**, Associate Professor and **Dr. R.SANTHOSHKUMAR**, Associate Professor, in Department of Computer Science and Engineering for their valuable support.

We would like to express our sincere gratitude and indebtedness to our project supervisor <Guide Name, Designation>, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

Finally, we express thanks to all those who have helped us successfully to completing this project. Furthermore, we would like to thank our family and friends for their moral support and encouragement.

We express thanks to all those who have helped us in successfully completing the project.

B. Jessica Dolly (17K81A05K0)
P. Malla Reddy (17K81A05M4)
P. Nikhil Reddy (17K81A05M6)
T. Gowri (17K81A05N7)

ABSTRACT

Unlike most other health conditions, the treatment of mental illness relies on subjective measurement. In addition, the criteria for diagnosing mental illnesses are based on broad categories of symptoms that do not account for individual deviations from these criteria. The increasing availability of personal digital devices, such as smartphones that are equipped with sensors, offers a new opportunity to continuously and passively measure human behavior in situ. This promises to lead to more precise assessment of human behavior and ultimately individual mental health. More refined modeling of individual mental health and a consideration of individual context, assessed through continuous monitoring, opens the way for more precise and personalized digital interventions that may help increase the number of positive clinical outcomes in mental healthcare. In this paper, we provide a conceptual review of such techniques for measuring, modeling, and treating mental illness and maintaining mental health.

TABLE OF CONTENTS

LIST OF TABLES

Table 1. List of Tables

LIST OF FIGURES

Table 2. List of Figures

LIST OF OUTPUT SCREENS

Table 3. List of output screens

LIST OF ABBREVIATIONS

CNN	Convolutional Neural Networks
UML	Unified Modelling Language
GUI	Graphical User Interface
DCNN	Deep Convolutional Neural Networks
RAM	Random Access Memory
CPU	Central Processing Unit
ER	Entity-Relationship
DFD	Data Flow Diagram

Table 4. List of abbreviations

GLOSSARY OF TERMS

TERM	MEANING
Deep Learning	Deep learning is part of a broader family of machine learning methods based on artificial neural networks with representation learning.
Convolutional Neural Network	In deep learning, a convolutional neural network is a class of deep neural network, most commonly applied to analyse visual imagery.
Prototype	First or preliminary version of a device or vehicle from which other forms are developed.
Occlusion	Blockage or closing.
Repository	A central location in which data is stored and managed.

Table 5. Glossary of Terms

1. INTRODUCTION

1.1 PROJECT OVERVIEW

- Over the last 10 years, technology has become more proximal to human activity. As more and more people adopt today's technology, healthcare involving technology in some respect is becoming increasingly acceptable. Ownership of smartphones is especially prevalent among underserved minority groups: 47% of black non-Hispanics and 49% of Hispanics own smart phones, compared to 42% of non-Hispanic whites .From the perspective of mental healthcare, mobile technology appears to be a feasible medium for delivering care; for example, a recent community-based survey of over 1,500 people with serious mental illnesses found that over 80% of patients with bipolar disorder (BD) owned and used mobile phones regularly for calling, texting, and the internet (Ben-Zeev, Davis, Kaiser, Krzsos, & Drake, 2013).

- Using Facial Landmarks we are detecting emotions, more robust and powerful than the earlier used fisherface classifier, but also requiring some more code and modules. Nothing insurmountable though. We need to do a few things:

1. Get images from a webcam
2. Detect Facial Landmarks
3. Train a machine learning algorithm
4. Predict emotions

- The first thing to do is find ways to transform these nice dots overlaid on your face into features to feed the classifier. Features are little bits of information that describe the object or object state that we are trying to divide into categories. Is this description a bit abstract? Imagine you are in a room without windows with only a speaker and a microphone. I am outside this room and I need to make you guess whether there is a cat, dog or a horse in front of me. The rule is that I can only use visual characteristics of the animal, no names or comparisons. What do I tell you? Probably if the animal is big or small, that it has fur, that the fur is long or short, that it has claws or hooves, whether it has a tail made of flesh or just from hair, etcetera

- Each bit of information I pass you can be considered a feature, and based the same feature set for each animal, you would be pretty accurate if I chose the features well. How you extract features from your source data is actually where a lot of research is, it's not just

about creating better classifying algorithms but also about finding better ways to collect and *describe* data. The same classifying algorithm might function tremendously well or not at all depending on how well the information we feed it is able to discriminate between different objects or object states. If, for example, we would extract eye colour and number of freckles on each face, feed it to the classifier, and then expect it to be able to predict what emotion is expressed, we would not get far. However, the facial landmarks from the same image material describe the position of all the "moving parts" of the depicted face, the things you use to express an emotion. This is certainly useful information!

1.2 SCOPE AND OBJECTIVE

- Significant gender differences in help seeking have also been found, with 11% of female students looking for help in comparison to 6% of males. A global survey found that while males made up 43.8% of the student body, they only comprised of 33.9% of clients who presented to college counselling centres, suggesting that males tend not to seek help for mental health problems. While females generally have higher rates of mood and anxiety disorders this only partially accounts for the gender difference found in help seeking.

- Studies corroborate that mental health problems can impact severely on a student's life. Indeed, mental health problems considerably disrupt learning ability, with psychopathology, particularly anxiety and depression, being associated with lower grades. In addition, students who had lifetime suicide plans and attempts when entering university obtained significantly lower grades, as did those who engaged in non-suicidal self-injury. Issues with attention and concentration can also impact on grades in addition to mental wellbeing. For example, ADHD is often co-morbid with a range of mental health disorders. Moreover, research has found that of those with DSM IV/CIDI mental health disorders in the previous 12 months, 83.1% of disorders commenced before students started college and that pre-matriculation onset was associated with higher attrition rates and lower university entry rates. It is important therefore to establish baseline prevalence rates of disorders and to understand the socio-demographic predictors of mental health and behavioral problems when students first enter universities.

- This point of entry information may be very beneficial for universities, helping them to provide adequate support for students and addressing problems early, minimizing risk and improving grades and retention rates. A report examining the mental health of students in higher education recommends the use of longitudinal studies to gain greater insight into psychopathology in the

student body. Research such as that carried out by the WHO World Mental Health Surveys International College Student Project (WMH-ICS) will gather important information about the wellbeing of the student population. Conducted as part of this initiative, the Ulster University Student Wellbeing Study aims to examine and monitor student health and wellbeing during their time at university.

1.3 ORGANIZATION OF CHAPTERS

This documentation consists of 10 different chapter and they are:

1. Introduction – This chapter covers the overview of our project and its objectives.
2. Literature Survey – This includes the details of our survey.
3. Software and Hardware Requirements – We specify our software and hardware requirements here.
4. Software Development Analysis – This section includes the problem definition and details of the modules we used in our project.
5. Project System Design – This chapter includes the design part of our project which includes UML diagrams.
6. Project Coding – This section contains the details of our project code.
7. Project Testing – The details of test cases and testing are included in this chapter.
8. Output Screens – This contains the screenshots of how our project looks like when executed.
9. Experimental Results – This chapter contains the screenshots of our results.
10. Conclusion and Future Enhancements – This covers the conclusion of our project and the possible future developments.

2. LITERATURE SURVEY

A literature survey or a literature review in a project report is that section which shows the various analysis and research made in the field of your interest and the results already published, considering the various parameters of the project and the extent of the project. It is the most important part of our report as it gave us a direction in our research. It helped us set a goal for our analysis - thus giving us our problem statement.

2.1 SURVEY ON BACKGROUND

- **Persistence of mental health problems and needs in a college student population**

  AUTHOR: Kara Zivin 1, Daniel Eisenberg, Sarah E Gollust, Ezra Golberstein

  We conducted a baseline web-based survey of students attending a large public university in fall 2005 and a two-year follow-up survey in fall 2007. We used brief screening instruments to measure symptoms of mental disorders (anxiety, depression, eating disorders), as well as self-injury and suicidal ideation. We estimated the persistence of these mental health problems between the two time points, and determined to what extent students with mental health problems perceived a need for or used mental health services (medication or therapy). We conducted logistic regression analyses examining how baseline predictors were associated with mental health and help-seeking two years later. Over half of students suffered from at least one mental health problem at baseline or follow-up. Among students with at least one mental health problem at baseline, 60% had at least one mental health problem two years later. Among students with a mental health problem at both time points, fewer than half received treatment between those time points.

- **The impact of lifetime suicidality on academic performance in college freshmen**

  AUTHOR : **P Mortier , K Demyttenaere, R P Auerbach , J G Green , R C Kessler , G Kiekens , M K Nock , R Bruffaerts**

  As part of the World Mental Health Surveys International College Student project, web-based self-reported STB of KU Leuven (Leuven, Belgium) incoming freshmen (N=4921; response rate=65.4%) was collected, as well as academic year percentage (AYP), and the departments to which students belong. Single- and multilevel multivariate analyses were conducted, adjusted for gender, age, parental educational level, and comorbid lifetime emotional problems. Lifetime suicide plan and attempt upon college entrance were associated with significant decreases in AYP

(3.6% and 7.9%, respectively). A significant interaction was found with average departmental AYP, with STB more strongly associated with reduced AYP in departments with lower than higher average AYP Limited sample size precluded further investigation of interactions between department-level and student-level variables. No information was available on freshman secondary school academic performance.

- **A Mental disorder among college students in the World Health Organization World Mental Health Surveys**

  AUTHOR: R P Auerbach 1, J Alonso 2, W G Axinn 3, P Cuijpers 4, D D Ebert5, J G Green 6, I Hwang

  Although mental disorders are significant predictors of educational attainment throughout the entire educational career, most research on mental disorders among students has focused on the primary and secondary school years. The World Health Organization World Mental Health Surveys were used to examine the associations of mental disorders with college entry and attrition by comparing college students (n = 1572) and non-students in the same age range (18-22 years; n = 4178), including non-students who recently left college without graduating (n = 702) based on surveys in 21 countries (four low/lower-middle income, five upper-middle-income, one lower-middle or upper-middle at the times of two different surveys, and 11 high income). Lifetime and 12-month prevalence and age-of-onset of DSM-IV anxiety, mood, behavioral and substance disorders were assessed with the Composite International Diagnostic Interview (CIDI).

- **Prevalence and correlates of depression, anxiety, and suicidality among university students**

  AUTHOR: Daniel Eisenberg 1, Sarah E Gollust, Ezra Golberstein, Jennifer L Hefner

  Mental health among university students represents an important and growing public health concern for which epidemiological data are needed. A Web-based survey was administered to a random sample at a large public university with a demographic profile similar to the national student population. Depressive and anxiety disorders were assessed with the Patient Health Questionnaire (R. L. Spitzer, K. Kroenke, J. B. W. Williams, & the Patient Health Questionnaire Primary Care Study Group, 1999). Nonresponse weights were constructed with administrative data and a brief non-respondent survey. The response rate was 56.6% (N = 2,843). The estimated prevalence of any depressive or anxiety disorder was 15.6% for undergraduates and 13.0% for graduate students.

Suicidal ideation in the past 4 weeks was reported by 2% of students. Students reporting financial struggles were at higher risk for mental health problems (odds ratios = 1.6-9.0). These findings highlight the need to address mental health in young adult populations, particularly among those of lower socioeconomic status. Campus communities reach over half of young adults and thus represent unique opportunities to address mental health issues in this important age group.

2.2 CONCLUSIONS ON SURVEY

Previous studies suggest that many students have mental health and behavioral problems while at university which can impact on their wellbeing and may result in elevated attrition rates. The current study extends on these findings, providing important information on baseline rates of mental health and behavioral problems, along with help seeking in a representative sample of students commencing university in NI. Such findings mean that those in need of help are identified early and provided with information on available services. This may lead to improved retention rates and academic success, as well as maintaining or improving psychological health and wellbeing beyond the university years. A review of the evidence suggests that it is essential to increase awareness among students about the services and support that is available, as well as providing guidance for university staff to assist students with mental health difficulties.

3. SOFTWARE AND HARDWARE REQUIREMENTS

Requirement is a condition or capability possessed by the software or system component in order to solve a real world problem. The problems can be to automate a part of a system, to correct shortcomings of an existing system, to control a device, and so on.

Requirements describe how a system should act, appear or perform. For this, when users request for software, they provide an approximation of what the new system should be capable of doing. Requirements differ from one user to another and from one business process to another.

The purpose of the requirements document is to provide a basis for the mutual understanding between the users and the designers of the initial definition of the software development life cycle (SDLC) including the requirements, operating environment and development plan.

Requirements help to understand the behavior of a system, which is described by various tasks of the system. For example, some of the tasks of a system are to provide a response to input values, determine the state of data objects, and so on. Note that requirements are considered prior to the development of the software. The requirements, which are commonly considered, are classified into three categories, namely, functional requirements, non-functional requirements, and domain requirements.

The functional requirements should be complete and consistent. Completeness implies that all the user requirements are defined. Consistency implies that all requirements are specified clearly without any contradictory definition. Generally, it is observed that completeness and consistency cannot be achieved in large software or in a complex system due to the problems that arise while defining the functional requirements of these systems. The different needs of stakeholders also prevent the achievement of completeness and consistency. Due to these reasons, requirements may not be obvious when they are first specified and may further lead to inconsistencies in the requirements specification.

The non-functional requirements (also known as quality requirements) are related to system attributes such as reliability and response time. Non-functional requirements arise due to user requirements, budget constraints, organizational policies, and so on. These requirements are not related directly to any particular function provided by the system.

Non-functional requirements should be accomplished in software to make it perform efficiently. For example, if an aeroplane is unable to fulfill reliability requirements, it is not approved for safe operation. Similarly, if a real time control system is ineffective in accomplishing non-functional requirements, the control functions cannot operate correctly.

System requirements are the configuration that a system must have in order for a hardware or software application to run smoothly and efficiently. Failure to meet these requirements can result in installation problems or performance problems. The former may prevent a device or application from getting installed, whereas the latter may cause a product to malfunction or perform below expectation or even to hang or crash.

System requirements are also known as minimum system requirements. Hardware system requirements often specify the operating system version, processor type, memory size, available disk space and additional peripherals, if any, needed.

Software system requirements, in addition to the requirements, may also specify additional software dependencies (e.g., libraries, driver version, framework version). Some hardware/software manufacturers provide an upgrade assistant program that users can download and run to determine whether their system meets a product's requirements.

Some products include both minimum and recommended system requirements. A video game, for instance, may function with the minimum required CPU and GPU, but it will perform better with the recommended hardware. A more powerful processor and graphics card may produce improved graphics and faster frame rates (FPS). Some system requirements are not flexible, such as the operating system(s) and disk space required for software installation. Others, such as CPU, GPU, and RAM requirements may vary significantly between the minimum and recommended requirements. When buying or upgrading a software program, it is often wise to make sure your system has close to the recommended requirements to ensure a good user experience.

3.1 SOFTWARE REQUIREMENTS

Operating System	Windows Family.
Framework	DJANGO
Coding Langauge	Python

3.2 HARDWARE REQUIREMENTS

Processor	Pentium I3 or Higher
Speed	3.4 GHz
RAM	4 GB Min
Hard Disk	40 GB
Key Board	Standard Windows Keyboard
Mouse	Optical Mouse
Monitor	14’ Colour Monitor

4. SOFTWARE DEVELOPMENT ANALYSIS

Software development is a process of writing and maintaining the source code, but in a broader sense, it includes all that is involved between the conception of the desired software through to the final manifestation of the software, sometimes in a planned and structured process. Therefore, software development may include research, new development, prototyping, modification, reuse, re-engineering, maintenance, or any other activities that result in software products

4.1 OVERVIEW OF PROBLEM

In real life, people express their emotion on their face to show their psychological activities and attitudes in the interaction with other people. The primary focus of this project is to determine which emotion an input image that contains one facial emotion belongs to. Because human face is complex to interpret, emotion recognition can be specifically divided into classification of basic emotion and classification of compound emotion. For the goals of our project, the essential problem is to focus on the classification of 7 basic emotions (shown at below): Figure4.1.1: The examples of 7 basic emotions: Happy, Sad, Neutral, Angry, Fear, Surprise, and Disgust.

Fig 4.1.1: The examples of 7 basic emotions: Happy, Sad, Neutral, Angry, Fear, Surprise, and Disgust

In conclusion, we want to construct a system by which an input image that contains one expression belonging to one of the 7 basic emotions can generate an output that correctly labels the input image.

4.2 DEFINE THE PROBLEM

Most of the people undergo mental illness, any kind of illness which impacts in their behaviors when they are with group of people. Few people are open to show their vulnerability to others and few people are not that easy to go people.

So, a methodology for detecting an assistive tool for predicting behaviors is described further.

An idea for maintaining a virtual space is proposed, such that the images of people given at the time of case studies and surveys are saved in a repository.

The public is given provision to voluntarily take photographs of themselves and upload in that application, through that they can predict their emotions and behavior.

4.3 MODULES OVERVIEW

- **Data processing:** Pre-processing input raw image in the context of face recognition involves acquiring the face region and standardizing images in a format compatible with the CNN architecture employed. Each CNN has a different input size requirement. The photographs of people acquired by a digital camera or mobile phone are taken and categorized into separate cases for creating the database of face recognition system. The face region in each image is identified and cropped for getting the input face images.

- **CNN:** In general, CNN contains 2 blocks of layers: convolutional layers and fully-connected Layers (also known as Dense Layers). The convolutional layers are used to perform the task of feature extraction as it detects and produces signals of feature by doing dot product between convolutional kernels and the feature map or images. On the other hand , the fully-connected layers consists of units as neuron cells that mathematically represents the linear operation and the operation of activation function. The last layer, a softmax classifier will generate the output in the format of the probability of each class in multinomial distribution, and the correct label will be the one with the maximum probability.

- For the Convolution layers, the structure includes convolution filters, activation function, and maxpooling filters. The convolution filters are kernels with certain size that perform dot product between the images or feature maps and filters. In such process, one filter represents one specific feature the model intends to detect, and this feature can be located anywhere on the image or

feature-map since the kernel is connected with every region with size of kernel on the image or feature maps.

- In addition, the kernel is sensitive enough to detect the feature since the feature is the region with high cross-corelation. Besides convolutional layers, the activation function will serve as a threshold that only allows the acceptable signal from previous convolutional layers to pass through, thereby making the system more complex so that the feature can be easily distinguished.
- The last type of convolutional block is the Maxpooling layer. The maxpooling layer performs the task to downsample the image or feature map, thereby reducing the computation and directing the next layers to focus on more detailed features. In conclusion, these 3 types of layers construct the uniqueness of convolution block.

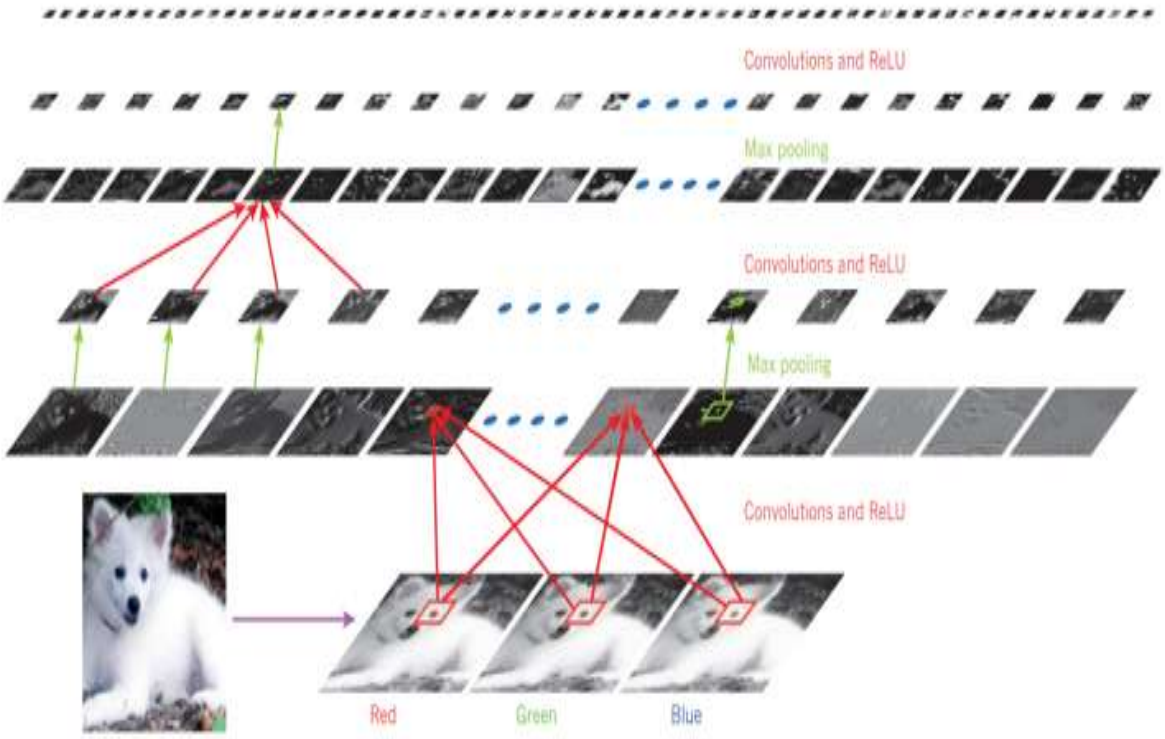


Fig 4.3.1: Example of how convolutional filters work. In this example, the image is in RGB color space, and the filters are detecting the their corresponding features

In addition to convolutional block, the fully connected layers will serve as classifier. Each unit of the layer contains the weight matrix, and through the linear transformation and activation function the output becomes the input of next layers of units. In contrast to traditional linear transformation,

The activation function ReLU will act in the same way as the convolutional layers to make the system more easily distinguish the feature. In addition, the last layer is normally a softmax classifier, and the result is typical the one with maximum probability in multinomial distribution.

- **The Advantages and Disadvantages of CNN:**

The major advantages of this algorithm are:

A. The feature can be captured regardless of its location.

B. The users do not have to design the filters to extract feature.

C. The negative effects of variance of lights can be reduced because the model is trained to learn the effect of noise.

However, it is apparent that this algorithm has several major issues:

A. The model requires a very large dataset to train because it has to cover the as many situations as possible. However, it is difficult to collect the dataset of emotion.

B. The model takes a very long to train from beginning (2 to 3 weeks)

C. The training requires machine with very good handware, and they are expensive and consumes a large amount of energy.

4.4 DEFINE THE MODULES

The project mainly consists of 5 modules :

- Load and Preprocess Dataset
- Train CNN Algorithm
- Capturing Image
- Detect Emotion
- Detect Emotion from images

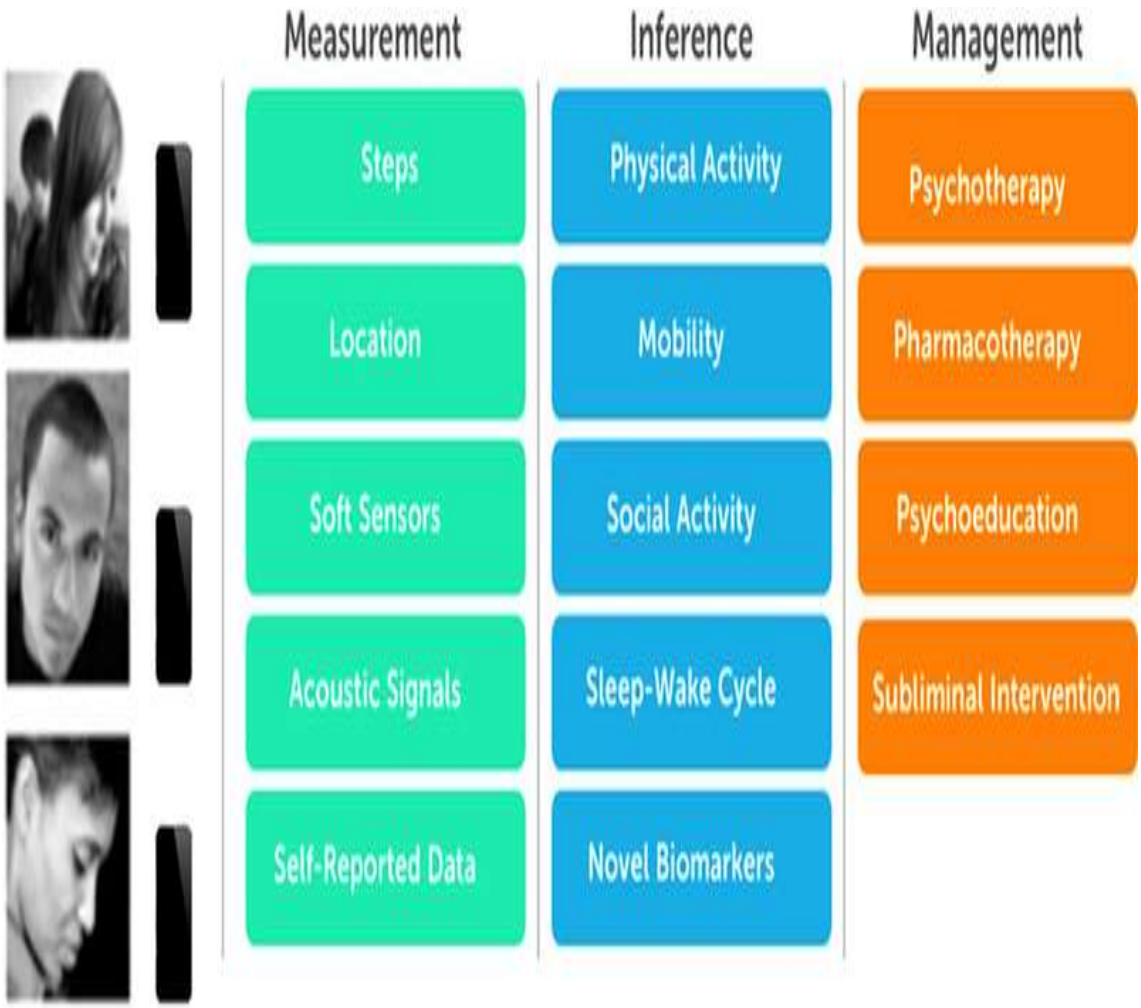
4.5 MODULE FUNCTIONALITY

- Load and Preprocess Dataset – It first loads and preprocesses the dataset and gives the total number of images which are present.
- Train CNN Algorithm – It trains the algorithm accordingly to achieve the emotion from the images.
- Capturing Image – It captures the image of the person, if the image is of small size it gets padded and gets resized. The captured image which is of the form RGB (Red Green Blue) turns into gray color to avoid noise to get accurate outcome.
- Detect Emotion – It detects the emotion the given image or the captured image.
- Detect Emotion from images – We can not only capture the live image but also we can upload the downloaded image from the repository to detect the emotion.

5. PROJECT SYSTEM DESIGN

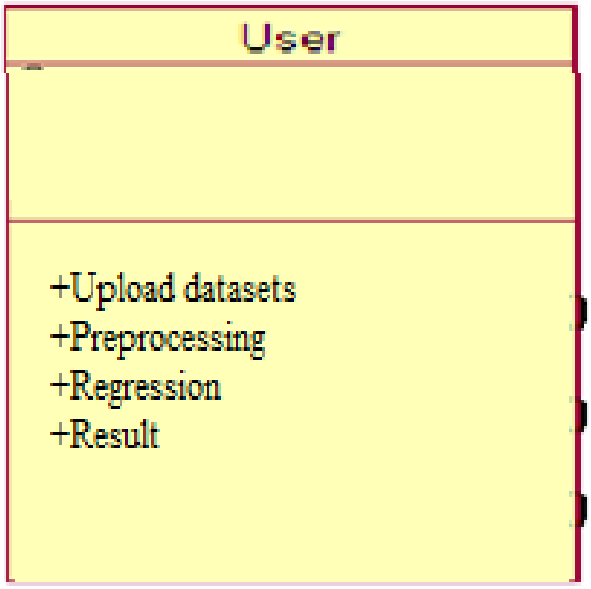
Systems design is the process of defining elements of a system like modules, architecture, components and their interfaces and data for a system based on the specified requirements. It is the process of defining, developing and designing systems which satisfies the specific needs and requirements of a business or organization. A systemic approach is required for a coherent and wellrunning system. Bottom-Up or Top-Down approach is required to take into account all related variables of the system. A designer uses the modelling languages to express the information and knowledge in a structure of system that is defined by a consistent set of rules and definitions. The designs can be defined in graphical or textual modelling languages. Unified Modelling Language has been used by us to describe software both structurally and behaviourally with notations.

5.1 ARCHITECTURAL DIAGRAM

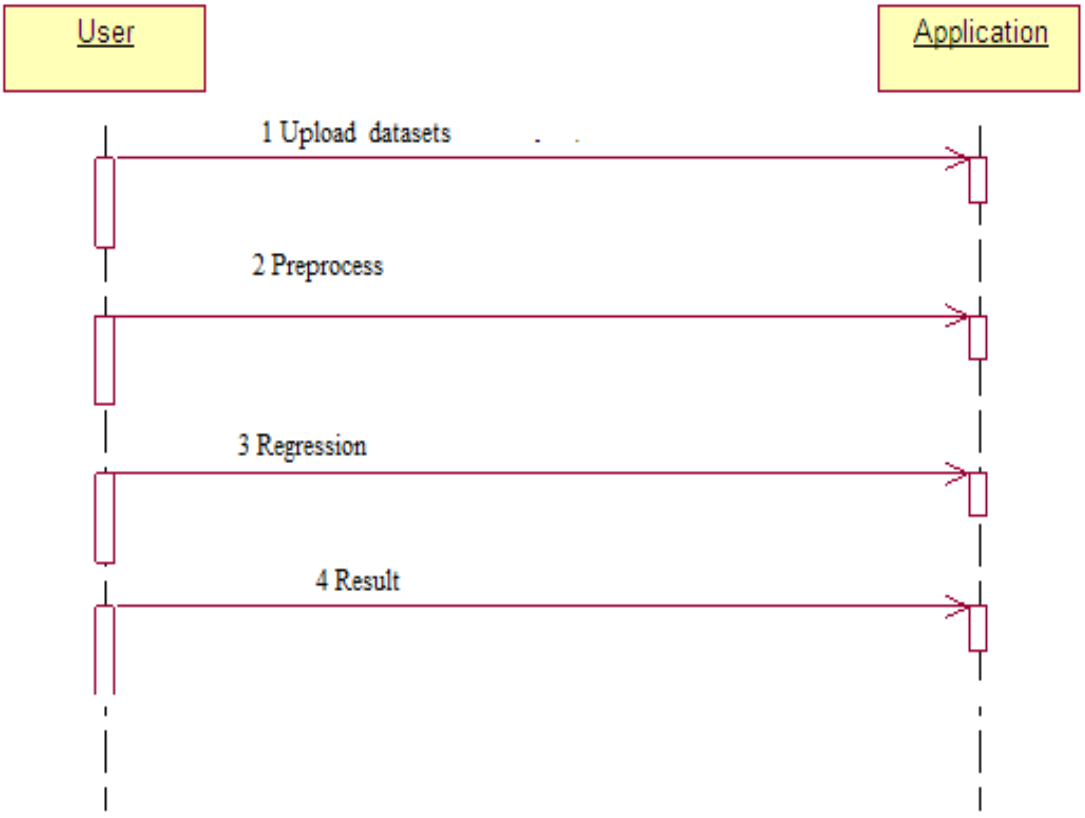


5.2 UML DIAGRAMS

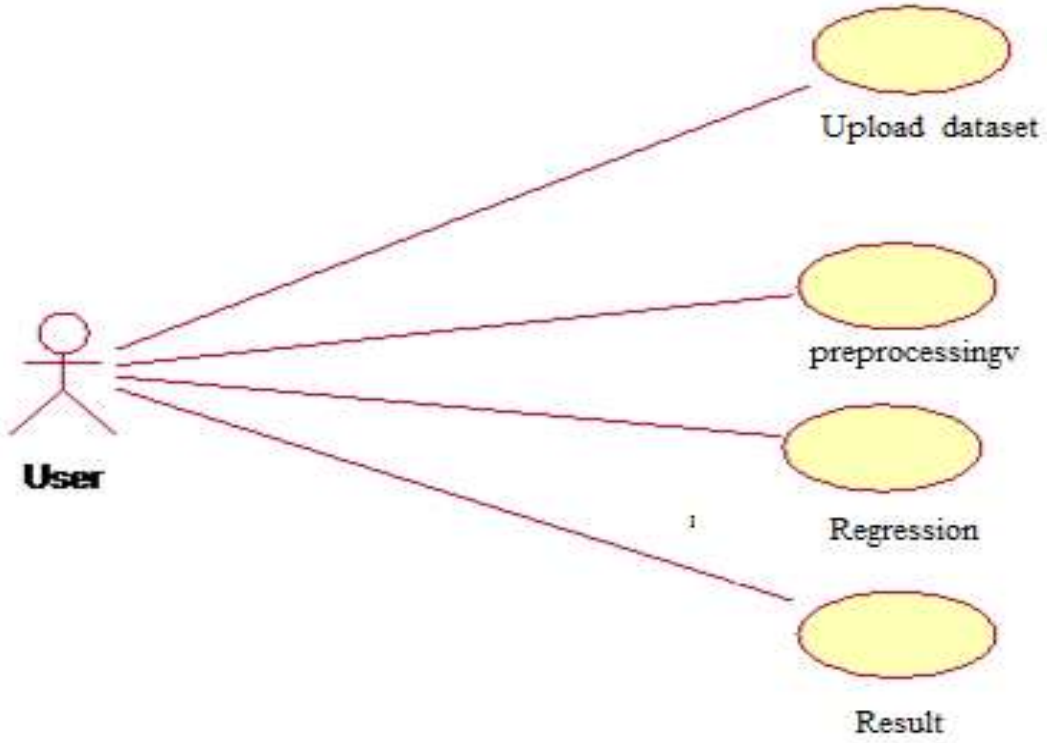
5.2.1 CLASS DIAGRAM



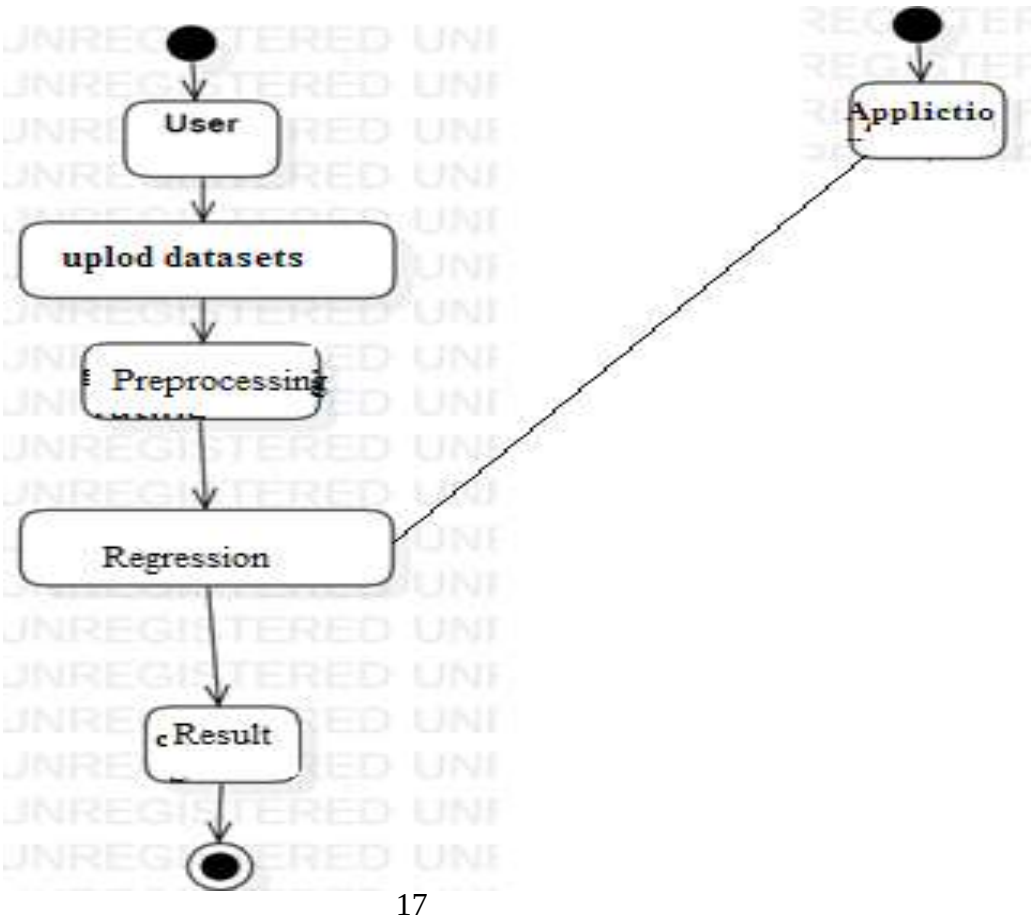
5.2.2 SEQUENCE DIAGRAM



5.2.3 USE CASE DIAGRAM



5.2.4 ACTIVITY DIAGRAM



6. PROJECT CODING

Project Coding is the process of designing and building an executable computer program to accomplish a specific computing result or to churn out a particular prototype or product. Programming involves tasks such as: analysis, generating algorithms, profiling algorithms' accuracy and resource consumption, and the implementation of algorithms in a chosen programming language (commonly referred to as coding). The source code of a program is written in one or more languages that are intelligible to programmers, rather than machine code, which is directly executed by the central processing unit. The purpose of programming is to find a sequence of instructions that will automate the performance of a task (which can be as complex as an operating system) on a computer, often for solving a given problem. Proficient programming thus often requires expertise in several different subjects, including knowledge of the application domain, specialized algorithms, and formal logic.

6.1 CODE TEMPLATES

File Edit Format Run Options Window Help

```
from tkinter import messagebox
from tkinter import *
from tkinter import simpledialog
import tkinter
import sys
sys.path.append(r"c:\users\jessica\appdata\roaming\python\python38\site-packages")
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
from tkinter import simpledialog
from tkinter import filedialog
import os
import cv2
import numpy as np
from sklearn.decomposition import PCA
from sklearn import svm
from sklearn.metrics import accuracy_score
from sklearn.model_selection import train_test_split
import pickle
import PIL.Image, PIL.ImageTk
from tkinter.filedialog import askopenfilename
```

Fig 6.1.1: Importing all the libraries which are useful

File Edit Format Run Options Window Help

```
def __init__(self, window, window_title, video_source=0):
    global cart
    global text
    cart = []
    self.window = window
    self.window.title(window_title)
    self.window.geometry("1300x1200")
    self.video_source = video_source
    self.vid = MyVideoCapture(self.video_source)
    self.canvas = tkinter.Canvas(window, width = self.vid.width, height = self.vid.height)
    self.canvas.pack()
    self.font1 = ('times', 13, 'bold')
    self.btn_snapshot=tkinter.Button(window, text="Load & Preprocess Dataset", command=self.processDataset)
    self.btn_snapshot.place(x=10,y=50)
    self.btn_snapshot.config(font=self.font1)
    self.btn_train=tkinter.Button(window, text="Train CNN Algorithm", command=self.trainmodel)
    self.btn_train.place(x=10,y=100)
    self.btn_train.config(font=self.font1)
    self.btn_predict=tkinter.Button(window, text="Capture Person", command=self.capturePerson)
    self.btn_predict.place(x=10,y=150)
    self.btn_predict.config(font=self.font1)

    self.btn_person=tkinter.Button(window, text="Detect Emotion", command=self.predict)
    self.btn_person.place(x=10,y=200)
    self.btn_person.config(font=self.font1)

    self.btn1_person=tkinter.Button(window, text="Detect Emotion From Images", command=self.predictImage)
    self.btn1_person.place(x=10,y=250)
    self.btn1_person.config(font=self.font1)

    self.img_canvas = tkinter.Canvas(window, width = 200, height = 200)
    self.img_canvas.place(x=10,y=300)

    self.text=Text(window,height=35,width=45)
    scroll=Scrollbar(self.text)
    self.text.configure(yscrollcommand=scroll.set)
    self.text.place(x=1000,y=50)
    self.text.config(font=self.font1)

    self.cascPath = "haarcascade_frontalface_default.xml"
    self.faceCascade = cv2.CascadeClassifier(self.cascPath)
```

Fig 6.1.2: Creating a window for capturing image

```
File Edit Format Run Options Window Help
    def capturePerson(self):
        option = 0
        ret, frame = self.vid.get_frame()
        img = frame
        gray = cv2.cvtColor(frame, cv2.COLOR_BGR2GRAY)
        faces = self.faceCascade.detectMultiScale(gray,1.3,5)
        print("Found {0} faces!".format(len(faces)))
        for (x, y, w, h) in faces:
            cv2.rectangle(frame, (x, y), (x+w, y+h), (0, 255, 0), 2)
            img = frame[y:y + h, x:x + w]
            img = cv2.resize(img,(48,48))
            option = 1
        if option == 1:
            cv2.imwrite("test.jpg",img);
            cv2.imshow("Image Captured",frame)
            cv2.waitKey(0)
        else:
            messagebox.showinfo("Face or person not detected","Face or person not detected")

    def processDataset(self):
        global X, Y
        global X_train, X_test, y_train, y_test
        global pca
        X = np.load('model/X.txt.npy')
        Y = np.load('model/Y.txt.npy')
        print(X.shape)
        X = np.reshape(X, (X.shape[0],(X.shape[1]*X.shape[2]*X.shape[3])))
        pca = PCA(n_components = 100)
        X = pca.fit_transform(X)
        print(X.shape)
        X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=0.2)
        messagebox.showinfo("Total number of images found in dataset is : "+str(len(X)),"Total number of images found in dataset is : "+str(len(X)))
```

Fig 6.1.3: Turning the image into Gray and resizing its original size

```
File Edit Format Run Options Window Help
    def predictImage(self):
        names = ['angry','disgusted','fearful','happy','neutral','sad','surprised']
        global pca
        filename = filedialog.askopenfilename(initialdir="testImages")
        img = cv2.imread(filename)
        img = cv2.resize(img, (32,32))
        im2arr = np.array(img)
        im2arr = im2arr.reshape(1,32,32,3)
        im2arr = im2arr.astype('float32')
        im2arr = im2arr/255
        test = im2arr
        test = np.reshape(test, (test.shape[0],(test.shape[1]*test.shape[2]*test.shape[3])))
        test = pca.transform(test)
        predict = classifier.predict(test)[0]

        img = cv2.imread(filename)
        img = cv2.resize(img, (500,400))
        cv2.putText(img, 'Behaviour Detected as : '+names[predict], (10, 25),  cv2.FONT_HERSHEY_SIMPLEX,0.7, (255, 0, 0), 2)
        cv2.imshow('Behaviour Detected as : '+names[predict], img)
        cv2.waitKey(0)


    def update(self):
        ret, frame = self.vid.get_frame()
        if ret:
            self.photo = PIL.ImageTk.PhotoImage(image = PIL.Image.fromarray(frame))
            self.canvas.create_image(0, 0, image = self.photo, anchor = tkinter.NW)
            self.window.after(self.delay, self.update)
```

Fig 6.1.4: Predicting the behavior from the image

6.2 OUTLINE FOR VARIOUS FILES

- **Tkinter:** This framework provides Python users with a simple way to create GUI elements using the widgets found in the Tk toolkit. Tk widgets can be used to construct buttons, menus, data fields, etc.

- **Sys:** The sys module in Python provides various functions and variables that are used to manipulate different parts of the Python runtime environment. It allows operating on the interpreter as it provides access to the variables and functions that interact strongly with the interpreter.

- **Matplotlib:** It is an amazing visualization library in Python for 2D plots of arrays. Matplotlib is a multi-platform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack. matplotlib. pyplot is a collection of functions that make matplotlib work like MATLAB. Each pyplot function makes some change to a figure: e.g., creates a figure, creates a plotting area in a figure, plots some lines in a plotting area, decorates the plot with labels, etc.

- **Numpy:** NumPy is the fundamental package for scientific computing in Python. NumPy arrays facilitate advanced mathematical and other types of operations on large numbers of data. NumPy aims to provide an array object that is up to 50x faster than traditional Python lists. The array object in NumPy is called ndarray, it provides a lot of supporting functions that make working with ndarray very easy.

- **Pandas:** Pandas DataFrame is two-dimensional size-mutable, potentially heterogeneous tabular data structure with labeled axes (rows and columns). Indexing and Selecting Data. Working with Missing Data. Iterating over rows and columns.

- **Simpledialog:** It’s a module contains convenience classes and functions for creating simple modal dialogs to get a value from the user. The above three functions provide dialogs that prompt the user to enter a value of the desired type.

- **Filedialog:** It helps you open, save files or directories. This is the type of dialog you get when you click file, open. This dialog comes out of the module, there's no need to write all the code manually.

- **OS:** The OS module in Python provides functions for interacting with the operating system. OS comes under Python's standard utility modules. This module provides a portable way of using operating system dependent functionality path* modules include many functions to interact with the file system.

- **OpenCV:** CV2 is a cross-platform library using which we can develop real-time computer vision applications. It mainly focuses on image processing, video capture and analysis including features like face detection and object detection.

6.3 CLASS WITH FUNCTIONALITY

In systems engineering and requirements engineering, a non-functional requirement is a requirement that specifies criteria that can be used to judge the operation of a system, rather than specific behaviors. They are contrasted with functional requirements that define specific behavior or functions. Non-functional requirements add tremendous value to business analysis. It is commonly misunderstood by a lot of people. It is important for business stakeholders, and Clients to clearly explain the requirements and their expectations in measurable terms. If the non-functional requirements are not measurable then they should be revised or rewritten to gain better clarity. For example, User stories help in mitigating the gap between developers and the user community in Agile Methodology.

- **Usability:** Prioritize the important functions of the system based on usage patterns. Frequently used functions should be tested for usability, as should complex and critical functions. Be sure to create a requirement for this.

- **Reliability:** Reliability defines the trust in the system that is developed after using it for a period of time. It defines the likeability of the software to work without failure for a given time period. The number of bugs in the code, hardware failures, and problems can reduce the reliability of the software.

Your goal should be a long MTBF (mean time between failures). It is defined as the average period of time the system runs before failing. Create a requirement that data created in the system will be retained for a number of years without the data being changed by the system. It’s a good idea to also include requirements that make it easier to monitor system performance.

- **Performance:** What should system response times be, as measured from any point, under what circumstances? Are there specific peak times when the load on the system will be unusually high? Think of stress periods, for example, at the end of the month or in conjunction with payroll disbursement.

- **Supportability:** The system needs to be cost-effective to maintain. Maintainability requirements may cover diverse levels of documentation, such as system documentation, as well as test documentation, e.g. which test cases and test plans will accompany the system.

6.4 METHODS INPUT AND OUTPUT PARAMETERS

We implemented multiple methods, few of which are :

- predict_image(self) - predicts the image where the face of a person is recognized.
- resize(img) - resizes the image to fit into the frame.
- reshape(img) - reshapes the image if it is inappropriate.
- transform( ) - transforms the color image into a gray image.
- get_frame( ) - gives the output frame.
- processDataset( ) - loads and processes the dataset.
- trainModel( ) - trains the model and predicts the emotion from the image.

7. PROJECT TESTING

Project Testing is a method to check whether the actual software product matches expected requirements and to ensure that software product is Defect free. It involves execution of software/system components using manual or automated tools to evaluate one or more properties of interest.

The purpose of software testing is to identify errors, gaps or missing requirements in contrast to actual requirements. Some prefer saying Software testing as a White Box and Black Box Testing. In simple terms, Software Testing means the Verification of Application Under Test (AUT).

This tutorial introduces testing software to the audience and justifies its importance. Project testing is important because, if there are any bugs or errors in the software, it can be identified early and can be solved before delivery of the software product. Properly tested software product ensures reliability, security and high performance which further results in time saving, cost effectiveness and customer satisfaction.

Typically Testing is classified into three categories.

- Functional Testing
- Non-Functional Testing or Performance Testing
- Maintenance (Regression and Maintenance)

Functional Testing

- Functional tests provide systematic protests that functions tested are accessible as stated by the business and technical necessities, system certification and user guides.

  Functional difficult is centered on the subsequent items:

- **Valid Input** is used to identified classes of valid input must be accepted.
- **Invalid Input** is used to identified classes of illegal input must be disallowed.
- **Functions** is used to identified purposes must be exercised.
- **Output** is used to classify modules of request outputs.

- **Systems/Procedures** is used to interfacing systems or events must be appealed. Organization and grounding of functional tests is focused on supplies, key functions, or special test belongings. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and succeeding processes must be well-thought-out for testing. Before functional testing is complete, supplementary tests are identified and the operative value of recent test is resulted.

7.1 VARIOUS TEST CASES

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub-assemblies, assemblies and/or a finished product. It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement. We have performed multiple tests under the broad categorisation of white-box and black-box testing which further include unit, integration, boundary value and statement covering etc.

ID	**CASE**	**EXPECTED OUTCOMES**	**COMMENTS**
1.0	Load and Process Dataset		
	Number of images in the dataset	Total number of images	Pass
2.0	Train CNN Algorithm	Trains the algorithm	Pass
3.0	Capture Person on Cam		
3.1	Captures Person	Detects face	Pass
3.2	Captures Things	Doesn’t detect face	Pass
4.0	Detect Emotion	Detects 1 of the emotion Happy / Sad / Neutral / Angry / Frustrated / Disgusted / Fearful	Pass

Table 7.1. Test Cases Tabulation

- **Unit Testing**

Unit testing includes the design of test belongings that validate that the internal program logic is operative properly, and that program input produces valid outputs. All choice branches and internal code flow should be authorized. It is the testing of separate software units of the request. It is done after the close of an individual unit before integration. This is a structural testing, that relies on data of its structure and is invasive. Unit tests achieve basic tests at factor level and test a specific commercial process, application, and/or system formation. Unit tests ensure that each single path of business process completes accurately to the documented provisions and contains clearly defined inputs and probable results.

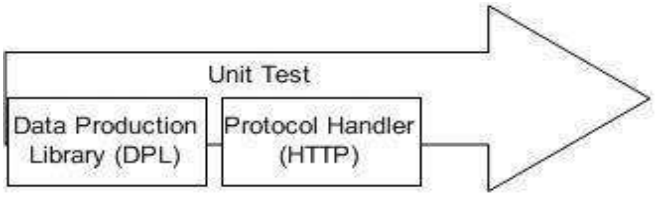


FIG 7.1 - UNIT TESTING

- **Integration Testing**

Integration tests are calculated to test integrated software components to regulate if they actually run as one program. Testing is occasion driven and is more concerned with the basic result of screens or fields. Integration tests validate that although the workings were individually approval, as shown by positively unit testing, the grouping of components is correct and dependable. Integration testing is specifically aimed at revealing the problems that arise from the grouping of components.

- **System Testing**

System testing confirms that the entire combined software system meets supplies. It tests a configuration to ensure known and predictable outcomes. An example of system testing is the arrangement oriented system mixing test. System testing is based on procedure similes and flows, emphasizing pre-driven process links and addition points.

7.2 BLACK-BOX TESTING

Black Box Testing is testing the software short of any knowledge of the inner mechanisms, building or language of the module actuality tested. Black box tests, as most other kinds of tests, must be printed from a final source document, such as requirement or necessities file, such as specification or requirement file. It is a testing in which the software below test is treated, as a black box you cannot “see” into it. The test provides inputs and respond to outputs without seeing how the software works.

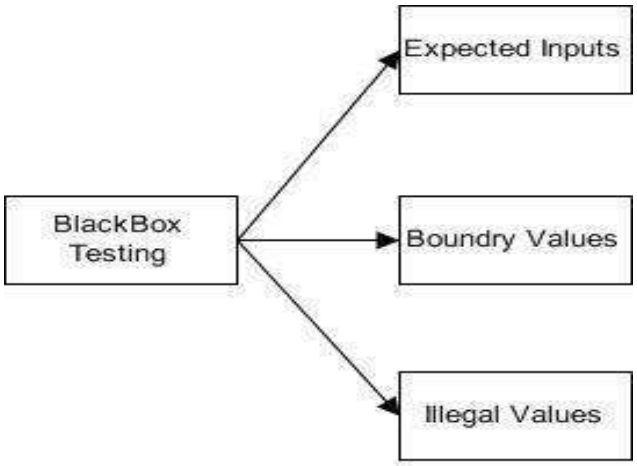


FIG 7.2 BLACK BOX TESTING

Here are the generic steps followed to carry out any type of Black Box Testing.

• Initially, the requirements and specifications of the system are examined.

• Tester chooses valid inputs (positive test scenario) to check whether SUT processes them correctly. Also, some invalid inputs (negative test scenario) are chosen to verify that the SUT is able to detect them.

• Tester determines expected outputs for all those inputs.

• Software tester constructs test cases with the selected inputs.

• The test cases are executed.

• Software tester compares the actual outputs with the expected outputs.

• Defects if any are fixed and re-tested.

Types of Black Box Testing

There are many types of Black Box Testing but the following are the prominent ones –

• Functional testing - This black box testing type is related to the functional requirements of a system; it is done by software testers.

• Non-functional testing - This type of black box testing is not related to testing of specific functionality, but non-functional requirements such as performance, scalability, usability.

• Regression testing - Regression Testing is done after code fixes, upgrades or any other system maintenance to check the new code has not affected the existing code.

Black Box Testing Techniques

Following are the prominent Test Strategy amongst the many used in Black box Testing

• Equivalence Class Testing: It is used to minimize the number of possible test cases to an optimum level while maintains reasonable test coverage.

• Boundary Value Testing: Boundary value testing is focused on the values at boundaries. This technique determines whether a certain range of values are acceptable by the system or not. It is very useful in reducing the number of test cases. It is most suitable for the systems where an input is within certain ranges.

• Decision Table Testing: A decision table puts causes and their effects in a matrix. There is a unique combination in each column.

7.3 WHITE-BOX TESTING

White Box Testing is a challenging in which the software tester has information of the inner workings, construction and language of the software, or at least its drive. It is used to test areas that cannot be stretched from a black box level.

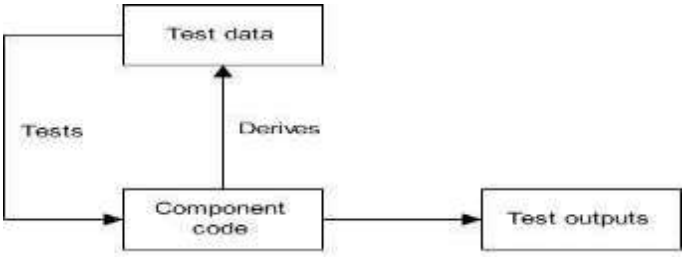


FIG 7.3 - WHITE BOX TESTING

In white box testing, code is visible to testers so it is also called Clear box testing, Open box testing, Transparent box testing, Code-based testing and Glass box testing.29 It is one of two parts of the Box Testing approach to software testing. Its counterpart, Blackbox testing, involves testing from an external or end-user type perspective. On the other hand, White box testing in software engineering is based on the inner workings of an application and revolves around internal testing. The term "WhiteBox" was used because of the see-through box concept. The clear box or WhiteBox name symbolizes the ability to see through the software's outer shell (or "box") into its inner workings. Likewise, the "black box" in "Black Box Testing" symbolizes not being able to see the inner workings of the software so that only the end-user experience can be tested.

White box testing involves the testing of the software code for the following:

• Internal security holes

• Broken or poorly structured paths in the coding processes

• The flow of specific inputs through the code

• Expected output

• The functionality of conditional loops

• Testing of each statement, object, and function on an individual basis

The testing can be done at system, integration and unit levels of software development. One of the basic goals of whitebox testing is to verify a working flow for an application. It involves testing a series of predefined inputs against expected or desired outputs so that when a specific input does not result in the expected output, you have encountered a bug.

To give you a simplified explanation of white box testing, we have divided it into two basic steps. This is what we do when testing an application using the white box testing technique:

STEP 1) UNDERSTAND THE SOURCE CODE

The first thing a tester will often do is learn and understand the source code of the application. Since white box testing involves the testing of the inner workings of an application, the tester must be very knowledgeable in the programming languages used in the applications they are testing. Also, the testing person must be highly aware of secure coding practices. Security is often one of the primary objectives of testing software. The tester should be able to find security issues and prevent 30 attacks from hackers and naive users who might inject malicious code into the application either knowingly or unknowingly.

Step 2) CREATE TEST CASES AND EXECUTE

The second basic step to white box testing involves testing the application's source code for proper flow and structure. One way is by writing more code to test the application's source code. The tester will develop little tests for each process or series of processes in the application. This method requires that the tester must have intimate knowledge of the code and is often done by the developer.

The goal of WhiteBox testing in software engineering is to verify all the decision branches, loops, statements in the code.

A major White box testing technique is Code Coverage analysis. Code Coverage analysis eliminates gaps in a Test Case suite. It identifies areas of a program that are not exercised by a set of test cases. Once gaps are identified, you create test cases to verify untested parts of the code, thereby increasing the quality of the software product.

There are automated tools available to perform Code coverage analysis. Below are a few coverage analysis techniques a box tester can use:

Statement Coverage:- This technique requires every possible statement in the code to be tested at least once during the testing process of software engineering

Branch Coverage - This technique checks every possible path (if-else and other conditional loops) of a software application.

Apart from above, there are numerous coverage types such as Condition Coverage, Multiple Condition Coverage, Path Coverage, Function Coverage etc. Each technique has its own merits and attempts to test (cover) all parts of software code. Using Statement and Branch coverage you generally attain 80-90% code coverage which is sufficient. Following are important WhiteBox Testing Techniques:

• Statement Coverage

• Decision Coverage

• Branch Coverage

• Condition Coverage

• Multiple Condition Coverage

• Finite State Machine Coverage

• Path Coverage

8. OUTPUT SCREENS

An output screen is a device used to display output. An output screen could be a separate monitor or another display device used only to display the output being received from the computer or other devices.

Here, in the screen prints given below, we can see that the user interface screens consist of a application in which there are few buttons like Capture Image, Train Algorithm, Detect Emotion etc. where a person can load and process the dataset, train the algorithm, capture the image and get the predicted emotion from the image.

8.1 USER INTERFACE

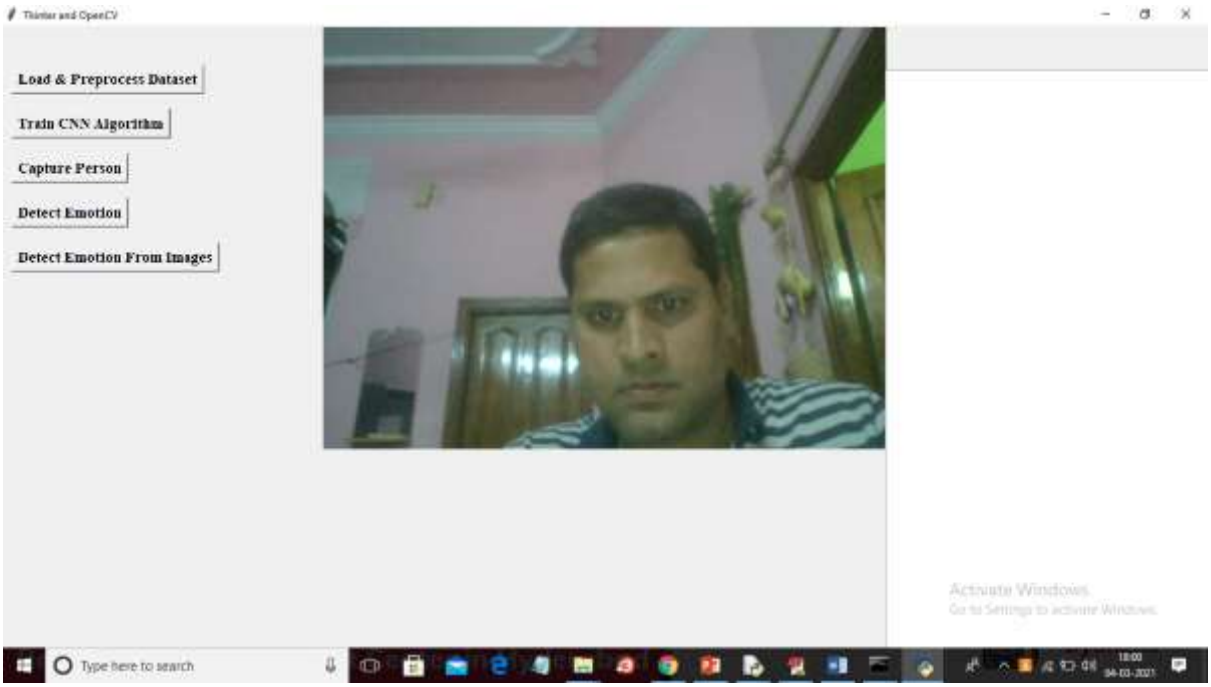


Fig 8.1: In above screen web cam started and now click on 'Load & Preprocess Dataset' button to read images and process them. This process may take 2 to 3 minutes of time.

8.2 OUTPUT SCREENS

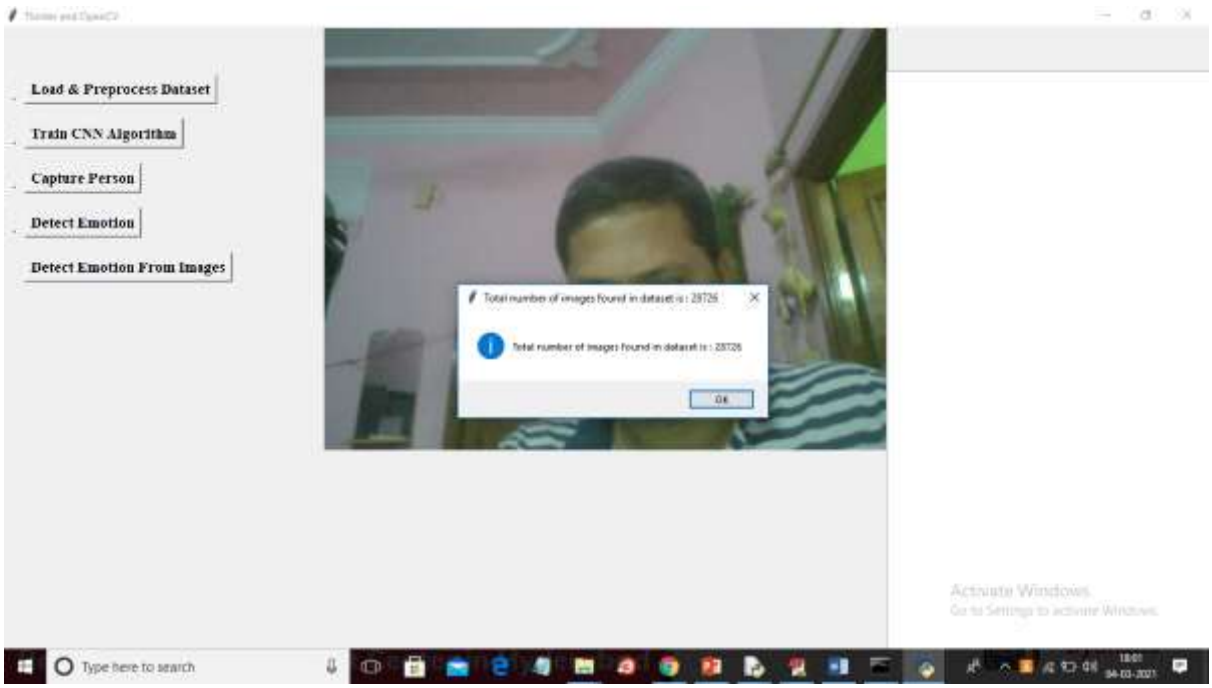


Fig. 8.2.1: In above screen we can see application process 28726 images and now click on 'Train CNN Algorithm' button to train CNN with all those images

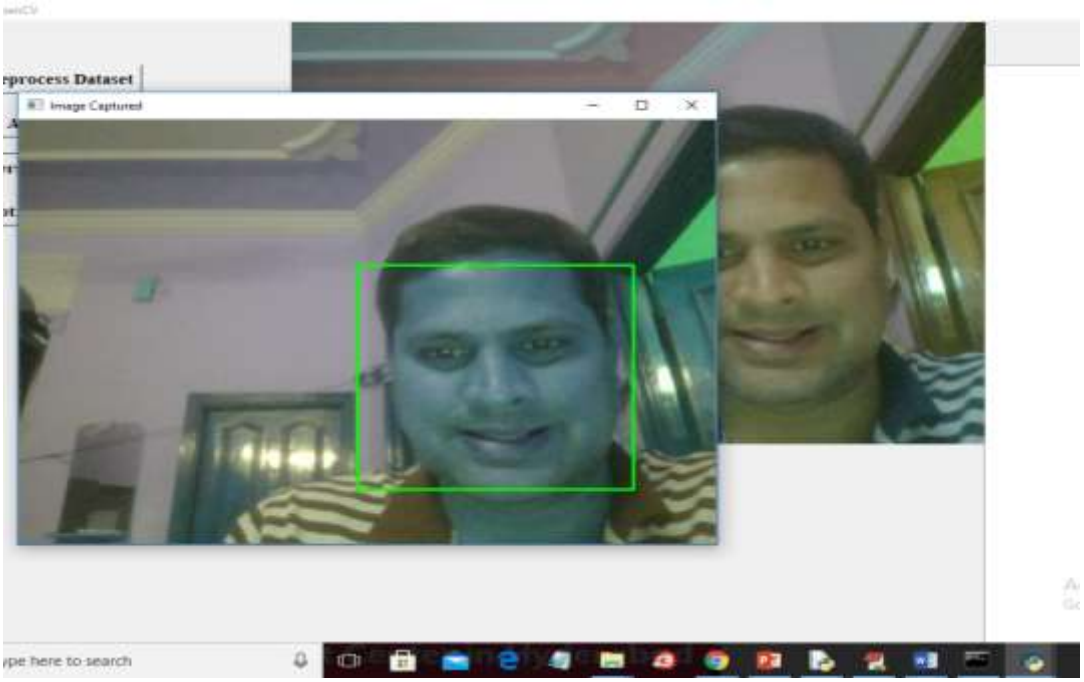


Fig. 8.2.2: In the above screen we see that the image is captured, it turns the colour picture into gray

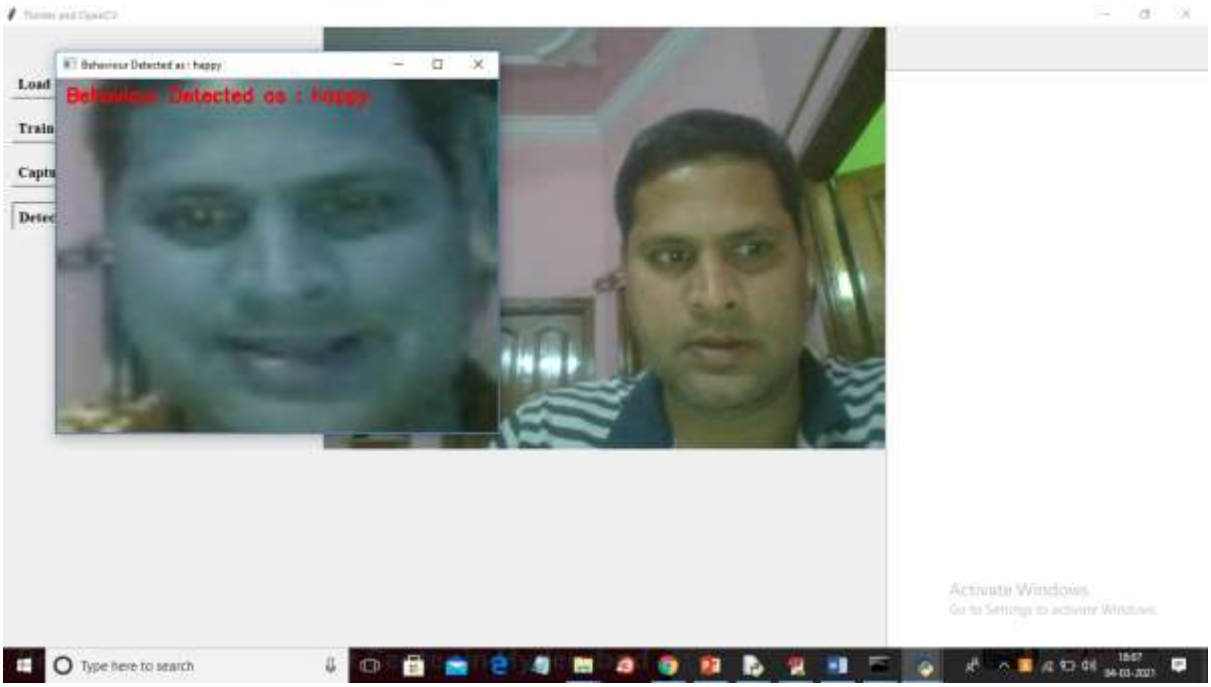


Fig.8.2.3: The emotion is detected as Happy

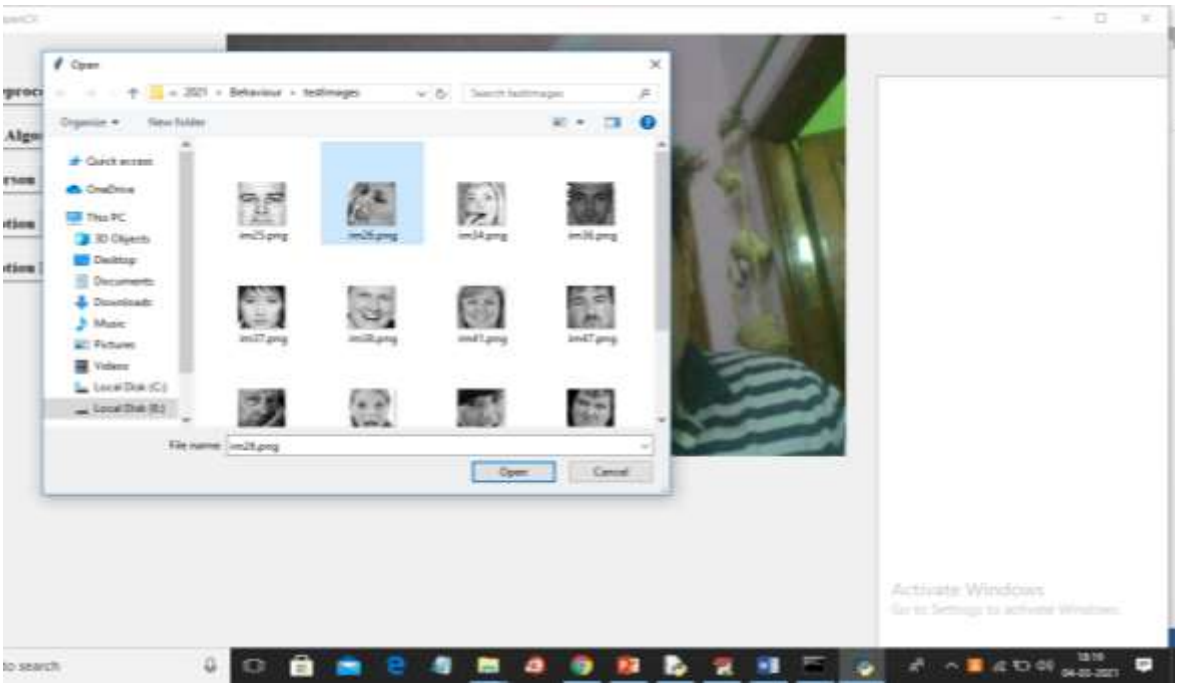


Fig.8.2.4: An image can also be selected to detect emotion

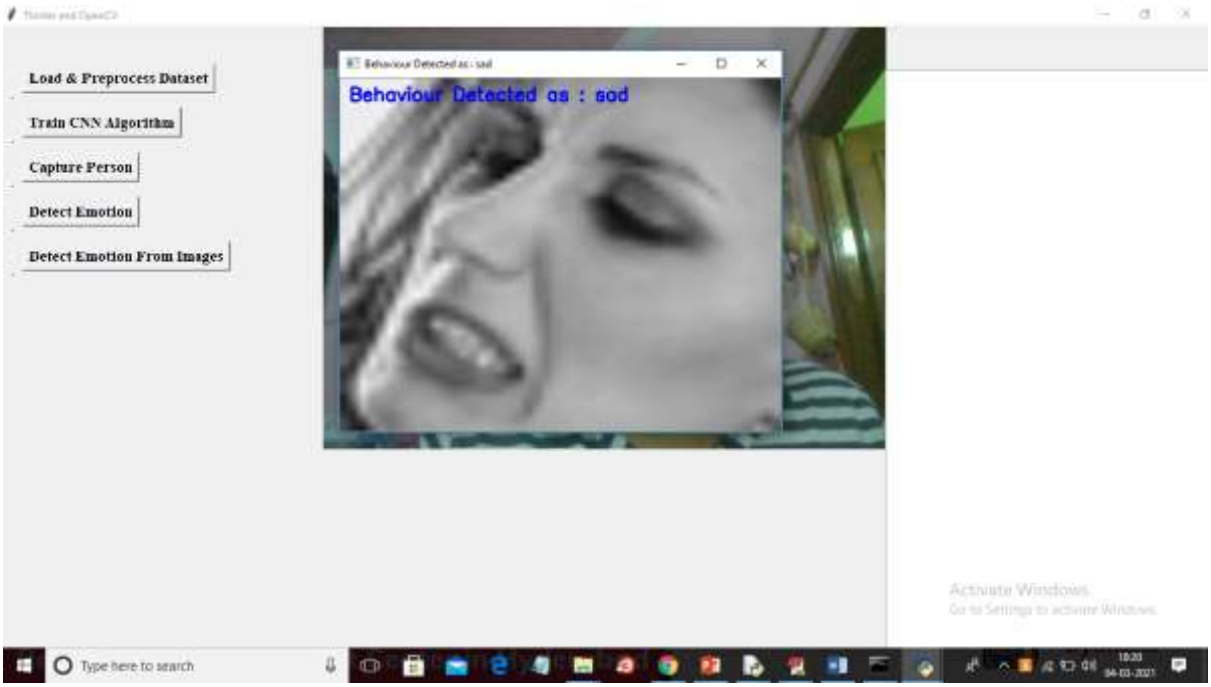


Fig.8.2.5: From the image the emotion is detected as Sad

1) Webcam connection: using this module application will be connected to live webcam
2) Load & Preprocess Dataset: using this module application read all dataset images from numpy array and then normalize and extract features from images.
3) Train CNN Algorithm: Extracted features will be used to train CNN algorithm
4) Capture Person: using this module we will capture person image and then detect face from that image
5) Detect Emotion: This module will take detected face as input and then by using CNN algorithm will predict person mental behaviour as SAD, HAPPY, NEUTRAL, ANGRY etc.

9. EXPERIMENTAL RESULT

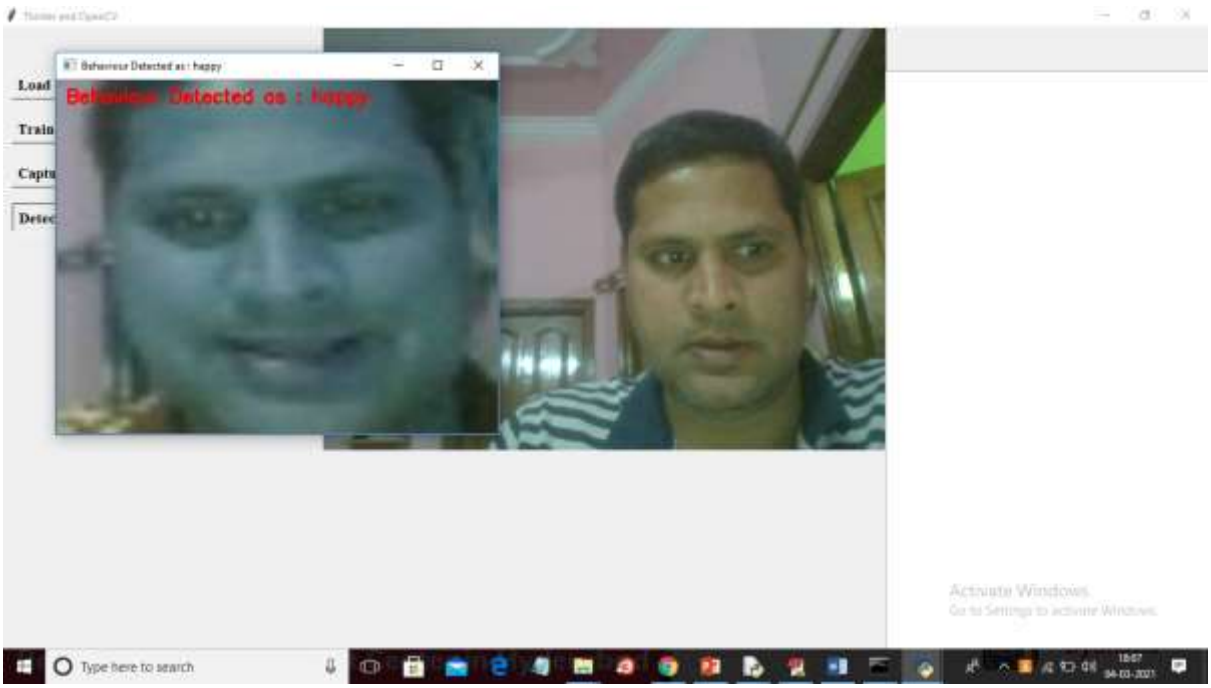


Fig. 9.1: The Behavior is detected as Happy

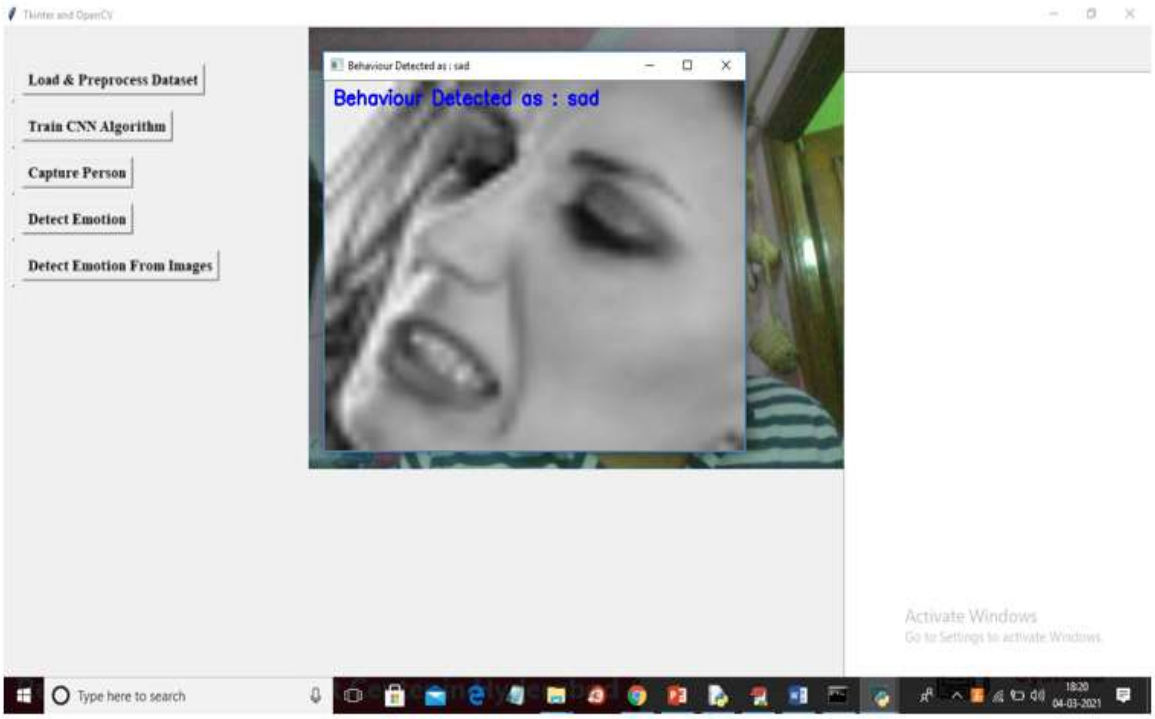


Fig. 9.2: The Behavior from a selected image is detected as Sad

10. CONCLUSION AND FUTURE ENHANCEMENT

The aim was to classify facial expressions into one of seven emotions by using various models. A conventional method and smaller convolutional networks were used and tested before arriving at the proposed model.

The ability of the model to make predictions in effectively real-time, indicates that real world uses of facial emotion recognition is barred only by the relative inaccuracies of the model itself.

In the future, an in depth analysis of the top 2 predicted emotions may lead to a much more accurate and reliable system. Further training samples for the more difficult to predict emotion of disgust will definitely be required in order to perfect such a system.

The real-time capacity of the model in addition to its quick training time and near-state-of-the-art accuracy allows the model to be adapted and used in nearly any use-case.

This also implies that with some work, the model could very well be deployed into real-life applications for effective utilization in domains such as in healthcare, marketing and the video game industry.

11. REFERENCES

1. Kiekens G, Claes L, Demyttenaere K, Auerbach RP, Green JG, Kessler RC, et al. Lifetime and 12- Month Nonsuicidal Self-Injury and Academic Performance in College Freshmen. Suicide & life-threatening behavior. 2016.
2. Larcombe W, Finch S, Sore R, Murray CM, Kentish S, Mulder RA, et al. Prevalence and socio-demographic correlates of psychological distress among students at an Australian university. Studies in Higher Education. 2016.
3. Higgins A, Doyle L, Downes C, Murphy R, Sharek D, DeVries J, et al. The LGBTIreland Report: national study of the mental health and wellbeing of lesbian, gay, bisexual, transgender and intersex people in Ireland. Dublin; 2016.
4. Association ACH. American College Health Association-National College Health Assessment II: Fall 2015 Reference Group Undergraduates Executive Summary. Hanover, MD; 2016.
5. Auerbach RP, Alonso J, Axinn WG, Cuijpers P, Ebert DD, Green JG, et al. Mental disorders among college students in the World Health Organization World Mental Health Surveys. Psychological medicine. 2016.
6. Mortier P, Demyttenaere K, Auerbach RP, Green JG, Kessler RC, Kiekens G, et al. The impact of lifetime suicidality on academic performance in college freshmen. Journal of affective disorders. 2015.
7. Prince JP. University student counseling and mental health in the United States: Trends and challenges. Mental Health & Prevention. 2015.
8. McLafferty M, Armour C, McKenna A, O'Neill S, Murphy S, Bunting B. Childhood adversity profiles and adult psychopathology in a representative Northern Ireland study. Journal of Anxiety Disorders. 2015.
9. Reetz DR, Krylowicz B, Mistler B. The association for university and college counseling center directors annual survey. 2014.

10. McIntyre D, Rowland M, Choi K, Sarkin A. Gender differences in the relationships between mental health symptoms, impairment, and treatment-related behaviors among college students. Mental Health & Prevention. 2014.

11. Steel Z, Marnane C, Iranpour C, Chey T, Jackson JW, Patel V, et al. The global prevalence of common mental disorders: a systematic review and meta-analysis 1980–2013. International journal of epidemiology. 2014.

12. PUBLICATIONS

CONFERENCE : International Conference on “Innovations in Computers Networks, Computational Intelligence and IoT” (ICICCI – 21)

Paper ID : ICICCI – 21 – 0108

13. STUDENT PROFILES

Buddha Jessica Dolly is a student at St. Martin's Engineering College, pursing Bachelor of Technology in Computer Science and Engineering, as well as a Programmer Analyst Trainee at Cognizant which is a leading Multinational Company. She completed her Intermediate Education from Narayana Junior College and Schooling from St. Peter's Grammar School. She took part in the employability development program conducted by ZENSAR. She participated in the events like “know more -teach more”, the global webinar on cloud and bigdata -changing the way we work which was conducted by Global education and careers forum (GECF) on 12th august 2020 , Women online workshop on “women in cyber security and privacy in 2020” which was conducted from 6th to 10th July 2020 In June 2019 she worked as a summer intern at Ridhan Technologies on web development project "Law Firm Website" using Html, Css, JavaScript. She is fond of graphic designing and creating content. Her technical skills include C, C++, Java, Python and MySql. She is passionate towards learning new things and aspires to become a greater version of herself in the technical field.

P. MALLA REDDY is currently pursuing his Bachelor of Technology (B.Tech) in the stream of Computer Science and Engineering (CSE) at St. Martin's Engineering College. He completed his intermediate from Sri Chaitanya junior college with 95% and class 10 from Shantiniketan School with 9.0 CGPA. His technical skills are C, C++, Python. He took part in the E-summit program conducted at Marri Lakshman Reddy Institute of Technology in 2018. He is very active in sports and a good Cricket player. He is also a very talented cricket player in the college sports team and participated in a few tournaments. He completed a few certification courses from online platforms like Coursera and SoloLearn.

GOWRI TEKMAL is currently pursuing her Bachelor of technology (B tech) in the stream of Computer Science and Engineering (CSE) at St. Martins engineering College . She completed her intermediate from Sri gayatri junior college with 90% and class 10 from Sri Sai Ram High School with 9.2cgpa. Her technical skills are C, C++, Python. She has a basic knowledge in Java. She is the representative (CR) of her class . She took part in the employability development program conducted by ZENSAR. She participated in the events like “know more -teach more”, the global webinar on cloud and bigdata -changing the way we work which was conducted by Global education and careers forum (GECF) on 12th august 2020 , Women online workshop on “women in cyber security and privacy in 2020” which was conducted from 6th to 10th July 2020 , “one day webinar on internet of things and its applications” conducted by Anand institute of higher technology on 21st may 2020. Her areas of interest are artificial intelligence(AI) , machine learning(ML) , deep learning , Python. She has completed a few certification courses from online platforms like Coursera , cursa and SoloLearn.

Perudi Nikhil Reddy is currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. He completed his intermediate from Narayana Junior College and 10th class from Nagarjuna Talent School. His responsibilities in that group include mentoring and motivating students to take coding as a serious hobby. His technical skills include C, Python and Java. He also has a basic understanding of C++. He took part in Employability Skill development Program conducted by Zensar. He also has a basic understanding of C++. "One Day Webinar on Internet of Things and Its Applications" conducted by Anand Institute of Higher Technology on 21st May 2020. April to 22nd May 2020. His areas of interest are Python, Artificial Intelligence, Machine Learning and Deep Learning. He completed few certification courses from online platforms like Coursera, Cursa App and Solo Learn.

14. APPENDICES

14.1 Conventional Methods and Their Issues:

Like every other classification problems, the emotion recognition problem requires an algorithm to complete feature extraction and categorical classification. In order to classify an emotion, we need to extract certain feature from data and build an model that can classify the input based on the feature. The procedure can be outlined as following:

1. Data Pre-processing: The data pre-processing is to standardize the data. The typical way is to set the mean of the data to 0 and to also divide the data by the standard deviation .
2. Feature Extraction: The typical conventional method is to detect the face and extract the Action Units (AU)(shown in figure) from the face, and certain emotion contain the combination of AUs code as feature.
3. Model Construction: The conventional classifier can be either supervised or unsupervised algorithm. A typical example of supervised algorithm is Support Vector Machine, and the examples of unsupervised algorithm include Principle Component Analysis (PCA) and Linear Discriminant Analysis (LDA).
4. Label or Result Generation The typical way to generate label or result is to find which decision boundary has the minimum Euclidean distance from the data.

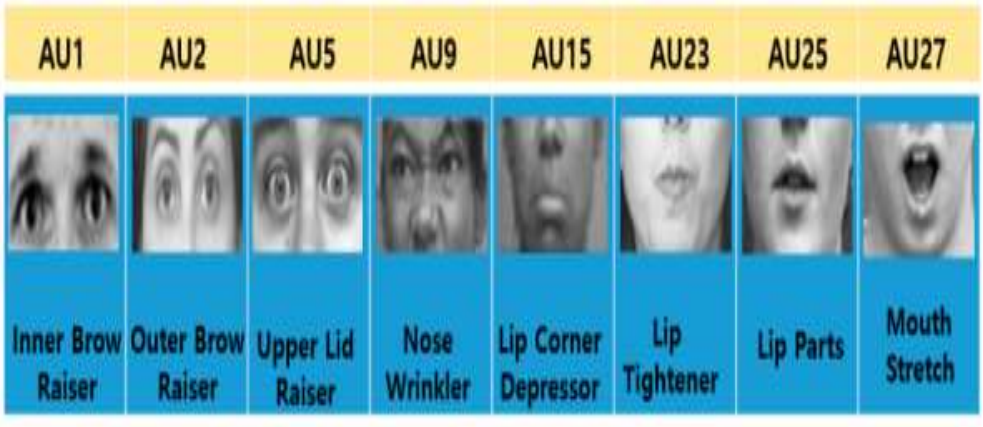


Fig 14.1: Examples of some Action Units

14.2 The Issue with Conventional Method:

The issues of Conventional method are:

A. Variance of Lights Since each image is taken in the completely different background and lighting conditions, the intra-class noise of lights will distort the model to classify the emotion. As results, the same type of emotions may be classified differently because of the effect of lighting noise.

B. Variance of Location Since the feature is typically extracted by filters such application of Local Binary Pattern in, the location of the feature, therefore, may affect the functionality of the feature extraction. As results, the AU may be extracted incorrectly if the face has is rotated or is in different part of the image.

These two issues are major problems faced by conventional algorithms.

In practice of other researches, each class typically consists of at least 700 to 1000 images to form the training dataset as mentioned in section 5 of the report. However, we do not have such large amount publicly available dataset for the project to conduct training. We, therefore, apply data augmentation to generate more data from original dataset to artificially create the variance of lights and the shift of locations.

1. Random noise: We added three kinds of noises for augmentation (Gaussian, Salt, and Poisson) – Gaussian noise: The noise is in a Gaussian distribution and randomly added into images – Salt: The noise is to add 1 randomly into images – Poisson: The noise is in a Poisson distribution and randomly added into images.
2. Horizontal flip: Adding an effect of flipping the images horizontally.
3. Rotation: We rotate the image from 0◦ to 50◦by increment of 10◦

Fig 14.2: This is an example of augmented data.

A

PROJECT REPORT

On

WOMEN SAFETY IN INDIAN CITIES USING MACHINE LEARNING ON TWEETS

Submitted by

Mr.ASai Anirudh(17K81A05J3)

Ms.CHSonanjali Devi(17K81A05K3)

Ms.KDhamini(17K81A05L5)

Ms.Shubhanshi Pandey(17K81A05N2)

in partial fulfilment for the award of the

degree of

BACHELOR OF TECHNOLOGY

IN

DEPARTMENT OF COMPTER SCIENCE AND ENGINEERING

Under the Guidance of

DR.G.JAWAHERLALNEHRU

ASSOCIATE PROFESSOR

ST.MARTIN'S ENGINEERING COLLEGE
An Autonomous Institute
Dhulapally, Secunderabad – 500 100
JUNE 2021

BONAFIDE CERTIFICATE

This is to certify that the project entitled**WOMEN SAFETY IN INDIAN CITIES USING MACHINE LEARNING ON TWEETS**, is being submitted by **Mr.A SAI ANIRUDH17K81A05J3, Ms.CH SONANJALI DEVI 17K81A05K3, Ms.KDHAMINI 17K81A05L5,Ms. SHUBHANSHI PANDEY17K81A05N2** in partial fulfillment of the requirement for the award of the degree of **BACHELOR OFTECHNOLOGY INCOMPUTER SCIENCE AND ENGINEERING**is recorded of bonafide work carried out by them. The result embodied in this report have been verified and found satisfactory.

<Signature>

DR.G.JAWAHERLALNEHRU

Department of CSE

Head of the Department

Dr.M.NARAYANAN

Department of CSE

Internal Examiner External Examiner

Place:

Date:

DECLARATION

We, the students of **Bachelor of Technology** in Department of Computer Science and Engineering', session: 2017 – 2021, St. Martin's Engineering College, Dhulapally, Kompally, Secunderabad, hereby declare that work presented in this Project Work entitled **WOMEN SAFETY IN INDIAN CITIES USING MACHINE LEARNING ON TWEETS** is the outcome of our own bonafide work and is correct to the best of our knowledge and this work has been undertaken taking care of Engineering Ethics. This result embodied in this project report has not been submitted in any university for award of any degree.

A SAI ANIRUDH 17K81A05J3

CH SONANJALI DEVI 17K81A05K3

K DHAMINI 17K81A05L5

SHUBHANSHI PANDEY 17K81A05N2

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible and whose encouragements and guidance have crowded effects with success.

We extended our deep sense of gratitude to Principal**, Dr. P. SANTOSH KUMARPATRA**, St. Martin's Engineering College, Dhulapally, for permitting us to undertake this project.

We are also thankful to Dr.**DR.G.JAWAHERLALNEHRU**, Head of the Department, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

We would like to thank **Dr. T. POONGOTHAI,** Department of Computer Science and Engineering (AI & ML) for her encouragement and insightful comments and as well as our project coordinators **Dr. B.RAJALINGAM**, Associate Professor and **Mr. J.SUDHAKAR**, Assistant Professor, in Department of Computer Science and Engineering for their valuable support.

We would like to express our sincere gratitude and indebtedness to our project supervisor Dr.GJawaherlalNehru, Assistant Professor, Computer Science and Engineering,St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

Finally, we express thanks to all those who have helped us successfully to completing this project. Furthermore, we would like to thank our family and friends for their moral support and encouragement.

We express thanks to all those who have helped us in successfully completing the project.

A SAI ANIRUDH 17K81A05J3

CH SONANJALI DEVI 17K81A05K3

K DHAMINI 17K81A05L5

SHUBHANSHI PANDEY 17K81A05N2

ABSTRACT

Women and girls have been experiencing a lot of violence and harassment in public places in various cities starting from stalking and leading to abuse harassment or abuse assault. This research paper basically focuses on the role of social media in promoting the safety of women in Indian cities with special reference to the role of social media websites and applications including Twitter platform Facebook and Instagram. This paper also focuses on how a sense of responsibility on part of Indian society can be developed the common Indian people so that we should focus on the safety of women surrounding them. Tweets on Twitter which usually contains images and text and also written messages and quotes which focus on the safety of women in Indian cities can be used to read a message amongst the Indian Youth Culture and educate people to take strict action and punish those who harass the women. Twitter and other Twitter handles which include hash tag messages that are widely spread across the whole globe sir as a platform for women to express their views about how they feel while we go out for work or travel in a public transport and what is the state of their mind when they are surrounded by unknown men and whether these women feel safe or not?

TABLEOFCONTENTS

LISTOFFIGURES

.

LIST OF OUTPUT SCREENS

LISTOFABBREVIATIONS

BMP	Bitmap
CPU	Cental Processing Unit
GB	Giga Byte
GUI	Graphical User Interface
SVM	Support Vector Machine
CNN	Convolutional Neural Networks
TP	Tweepy Packages

1.INTRODUCTION

1.1PROJECT OVERVIEW

Twitter in this modern era has emerged as an ultimate microblogging social network consisting over hundred million users and generate over five hundred million messages known as ‘Tweets’ every day. Twitter with such a massive audience has magnetized users to emit their perspective and judgmental about every existing issue and topic of internet, therefore twitter is an informative source for all the zones like institutions, companies and organizations. On the twitter, users will share their opinions and perspective in the tweets section. This tweet can only contain 140 characters, thus making the users to compact their messages with the help of abbreviations, slang, shot forms, emoticons, etc. In addition to this, many people express their opinions by using polysemy and sarcasm also. Hence twitter language can be termed as the unstructured. From the tweet, the sentiment behind the message is extracted. This extraction is done by using the sentimental analysis procedure. Results of the sentimental analysis can be used in many areas like sentiments regarding a particular brand or release of a product, analyzing public opinions on the government policies, people thoughts on women, etc. In order to perform classification of tweets and analyze the outcome, a lot of study has been done on the data obtained by the twitter. We also review some studies on machine learning in this paper and research on how to perform sentimental analysis using that domain on twitter data. The paper scope is restricted to machine learning algorithm and models. Staring at women and passing comments can be certain types of violence and harassments and these practices, which are unacceptable, are usually normal especially on the part of urban life. Many researches that have been conducted in India shows that women have reported sexual harassment and other practices as stated above. Such studies have also shown that in popular metropolitan cities like Delhi, Pune, Chennai and Mumbai, most women feel they are unsafe when surrounded by unknown people. On social media, people can freely express what they feel about the Indian politics, society and many other thoughts. Similarly, women can also share their experiences if they have faced any violence or sexual harassment and this brings innocent people together in order to stand up against such incidents. From the analysis of tweets text collection obtained by the twitter, it includes names of people who has harassed the women and also names of

women or innocent people who have stood against such violent acts or unethical behavior of men and thus making them uncomfortable to walk freely in public.

1.2 PROJECT OBJECTIVES

This project is to analyse women safety using social networking messages and by applying machine learning algorithms on it. Now-a-days almost all peoples are using social networking sites to express their feelings and if any women feel unsafe in any area, then she will express negative words in her post/tweets/messages and by analysing those messages we can detect which area is more unsafe for women's his project is to analyse women safety using social networking messages and by applying machine learning algorithms on it. Now-a-days almost all peoples are using social networking sites to express their feelings and if any women feel unsafe in any area, then she will express negative words in her post/tweets/messages and by analysing those messages we can detect which area is more unsafe for women's.

1.3.ORGANIZATION OF CHAPTERS

1.INTRODUCTION

In this chapter,We described the Overview of our project which summarizes the existing and proposed part and Objective of our project which summarizes the goal of our project .

2.LITERATURE SURVEY

In this chapter,Webroadly specified the Survey on Background which included the references that we surveyed and Conclusions on Survey which included the concepts that we have taken from the reference papers.

3.SOFTWARE AND HARDWARE REQUIREMENTS

In this chapter,We specified the requirements of Software and Hardware which are included in our project.

4.SOFTWARE DEVELOPMENT ANALYSIS

In this chapter,We described the Overview of our problem,Defined the problem.We also specified the Overview of Modules,Defined the Modules and showed the Functionality of the Modules.

5.PROJECT SYSTEM DESIGN

In this chapter,We showed the Unified Modeling Language(UML) diagrams of the system design.The UML Diagram include Usecase,Sequence,State Chart,Component and Deployment.

6.PROJECT CODING

In this chapter,We included Code templates,Outline for various files which includes all libraries,Class with Functionality and input and output parameters of methods.

7.PROJECT TESTING

In this chapter,We explained various test cases such as Unit,Integration,Functional and System testing.It also includes Black Box and White Box testings.

8.OUTPUT SCREENS

In this chapter,We showed the User Interface and Output Screens of our project.

9.EXPERIMENTAL RESULTS

In this chapter,We measured the performance of our algorithm based on accuracy metric.

10.CONCLUSION AND FUTURE ENHANCEMENTS

In this chapter, it covers the conclusion of our project and the possible future developments.

2.LITERATURE SURVEY

2.1 SURVEY ON BACKGROUND

1. Luciano Barbosa and Junlan Feng. "Robust sentiment detection on twitter from biased and noisy data." Proceedings of the 23rd international conference on computational linguistics: posters. Association for Computational Linguistics, 2010.

In this paper, we propose an approach to automatically detect sentiments on Twitter messages (tweets) that explores some characteristics of how tweets are written and meta-information of the words that compose these messages. Moreover, we leverage sources of noisy labels as our training data. These noisy labels were provided by a few sentiment detection websites over twitter data. In our experiments, we show that since our features are able to capture a more abstract representation of tweets, our solution is more effective than previous ones and also more robust regarding biased and noisy data, which is the kind of data provided by these sources.

2. Michael Gamon. "Sentiment classification on customer feedback data: noisy data, large feature vectors, and the roleof linguistic analysis." Proceedings of the 20th international conference on Computational Linguistics. Association for Computational Linguistics, 2004.

We demonstrate that it is possible to perform automatic sentiment classification in the very noisy domain of customer feedback data. We show that by using large feature vectors in combination with feature reduction, we can train linear support vector machines that achieve high classification accuracy on data that present classification challenges even for a human annotator. We also show that, surprisingly, the addition of deep linguistic analysis features to a set of surface level word n-gram features contributes consistently to classification accuracy in this domain.

3. Gupta B, Negi M, Vishwakarma K, Rawat G &Badhani P (2017). "Study of Twitter sentiment analysis using machine learning algorithms on Python." International Journal of Computer Applications, 165(9) 0975-8887.

Twitter is a platform widely used by people to express their opinions and display sentiments on different occasions. Sentiment analysis is an approach to analyze data and retrieve sentiment that it embodies. Twitter sentiment analysis is an application of sentiment analysis on data from Twitter (tweets), in order to extract sentiments conveyed by the user. In the past decades, the research in this field has consistently grown. The reason behind this is the challenging format of the tweets which makes the processing difficult. The tweet format is

very small which generates a whole new dimension of problems like use of slang, abbreviations etc. In this paper, we aim to review some papers regarding research in sentiment analysis on Twitter, describing the methodologies adopted and models applied, along with describing a generalized Python based approach.

4. Sahayak V, Shete V & Pathan A (2015). "Sentiment analysis on twitter data." International Journal of Innovative Research in Advanced Engineering (IJIRAE), 2(1), 178-183.

With the advancement of web technology and its growth, there is a huge volume of data present in the web for internet users and a lot of data is generated too. Internet has become a platform for online learning, exchanging ideas and sharing opinions. Social networking sites like Twitter, Facebook, Google+ are rapidly gaining popularity as they allow people to share and express their views about topics, have discussion with different communities, or post messages across the world. There has been lot of work in the field of sentiment analysis of twitter data. This survey focuses mainly on sentiment analysis of twitter data which is helpful to analyze the information in the tweets where opinions are highly unstructured, heterogeneous and are either positive or negative, or neutral in some cases. In this paper, we provide a survey and a comparative analyses of existing techniques for opinion mining like machine learning and lexicon-based approaches, together with evaluation metrics. Using various machine learning algorithms like Naive Bayes, Max Entropy, and Support Vector Machine, we provide research on twitter data streams.We have also discussed general challenges and applications of Sentiment Analysis on Twitter

2.2.CONCLUSIONS ON SURVEY

Throughout the research paper we have discussed about various machine learning algorithms that can help us to organize and analyze the huge amount of Twitter data obtained including millions of tweets and text messages shared every day. These machine learning algorithms are very effective and useful when it comes to analyzing of large amount of data including the SPC algorithm and linear algebraic Factor Model approaches which help to further categorize the data into meaningful groups. Support vector machines is yet another form of machine learning algorithm that is very popular in extracting Useful information from the Twitter and get an idea about the status of women safety in Indian cities.

3.SOFTWARE AND HARDWARE REQUIREMENTS

3.1.SOFTWARE REQUIREMENTS

Operating system : Windows 7 Ultimate.

Coding Language : Python.

Front-End : Python.

Designing : Html,css,javascript.

Data Base : MySQL.

3.2.HARDWARE REQUIREMENTS

Operating system : Windows 7 Ultimate.

Coding Language : Python.

Front-End : Python.

Designing : Html,css,javascript.

Data Base : MySQL.

PYTHON

Python is a **high-level, interpreted**, **interactive** and **object-oriented scriptinglanguage**. Python is designed to be highly readable. It uses English keywords frequently where as other languages use punctuation, and it has fewer syntactical constructions than other languages.

- **Python is Interpreted:** Python is processed at runtime by the interpreter. You do not need to compile your program before executing it. This is similar to PERL and PHP.
- **Python is Interactive:** You can actually sit at a Python prompt and interact with the interpreter directly to write your programs.
- **Python is Object-Oriented:** Python supports Object-Oriented style or technique of programming that encapsulates code within objects.
- **Python is a Beginner's Language:** Python is a great language for the beginner-level programmers and supports the development of a wide range of applications from simple text processing to WWW browsers to games.

History of Python

Python was developed by Guido van Rossum in the late eighties and early nineties at the National Research Institute for Mathematics and Computer Science in the Netherlands.

Python is derived from many other languages, including ABC, Modula-3, C, C++, Algol-68, SmallTalk, and Unix shell and other scripting languages.

Python is copyrighted. Like Perl, Python source code is now available under the GNU General Public License (GPL).

Python is now maintained by a core development team at the institute, although Guido van Rossum still holds a vital role in directing its progress.

Python Features

Python's features include:

- **Easy-to-learn:** Python has few keywords, simple structure, and a clearly defined syntax. This allows the student to pick up the language quickly.
- **Easy-to-read:** Python code is more clearly defined and visible to the eyes.
- **Easy-to-maintain:** Python's source code is fairly easy-to-maintain.

- **A broad standard library:** Python's bulk of the library is very portable and cross-platform compatible on UNIX, Windows, and Macintosh.
- **Interactive Mode:** Python has support for an interactive mode which allows interactive testing and debugging of snippets of code.
- **Portable:** Python can run on a wide variety of hardware platforms and has the same interface on all platforms.
- **Extendable:** You can add low-level modules to the Python interpreter. These modules enable programmers to add to or customize their tools to be more efficient.
- **Databases:** Python provides interfaces to all major commercial databases.
- **GUI Programming:** Python supports GUI applications that can be created and ported to many system calls, libraries and windows systems, such as Windows MFC, Macintosh, and the X Window system of Unix.
- **Scalable:** Python provides a better structure and support for large programs than shell scripting.

Python has a big list of good features:

- It supports functional and structured programming methods as well as OOP.
- It can be used as a scripting language or can be compiled to byte-code for building large applications.
- It provides very high-level dynamic data types and supports dynamic type checking.
- IT supports automatic garbage collection.

MySQL

MySQL is an open-source relational database management system (RDBMS).[5][6] Its name is a combination of "My", the name of co-founder Michael Widenius's daughter,[7] and "SQL", the abbreviation for Structured Query Language. A relational database organizes data into one or more data tables in which data types may be related to each other; these relations

help structure the data. SQL is a language programmers use to create, modify and extract data from the relational database, as well as control user access to the database. In addition to relational databases and SQL, an RDBMS like MySQL works with an operating system to implement a relational database in a computer's storage system, manages users, allows for network access and facilitates testing database integrity and creation of backups.

MySQL was created by a Swedish company, MySQL AB, founded by Swedes David Axmark, Allan Larsson and Finland Swede Michael "Monty" Widenius. Original development of MySQL by Widenius and Axmark began in 1994.[22] The first version of MySQL appeared on 23 May 1995. It was initially created for personal usage from mSQL based on the low-level language ISAM, which the creators considered too slow and inflexible. They created a new SQL interface, while keeping the same API as mSQL. By keeping the API consistent with the mSQL system, many developers were able to use MySQL instead of the (proprietarily licensed) mSQL antecedent.[23]

Features

- A broad subset of ANSI SQL 99, as well as extensions
- Cross-platform support
- Stored procedures, using a procedural language that closely adheres to SQL/PSM[79]
- Triggers
- Cursors
- Updatable views
- Online Data Definition Language (DDL) when using the InnoDB Storage Engine.
- Information schema
- Performance Schema that collects and aggregates statistics about server execution and query performance for monitoring purposes.[80]
- A set of SQL Mode options to control runtime behavior, including a strict mode to better adhere to SQL standards.
- X/Open XA distributed transaction processing (DTP) support; two phase commit as part of this, using the default InnoDB storage engine
- Transactions with savepoints when using the default InnoDB Storage Engine. The NDB Cluster Storage Engine also supports transactions.
- ACID compliance when using InnoDB and NDB Cluster Storage Engines[81]
- SSL support
- Query caching
- Sub-SELECTs (i.e. nested SELECTs)
- Built-in replication support
    - Asynchronous replication: master-slave from one master to many slaves[82][83] or many masters to one slave[84]
    - Semi synchronous replication: Master to slave replication where the master waits on replication[85][86]
    - Synchronous replication: Multi-master replication is provided in MySQL Cluster.[87]

- Virtual Synchronous: Self managed groups of MySQL servers with multi master support can be done using: Galera Cluster[88] or the built in Group Replication plugin[89]

MySQL as a service

Some cloud platforms offer MySQL "as a service". In this configuration, application owners do not have to install and maintain the MySQL database on their own. Instead, the database service provider takes responsibility for installing and maintaining the database, and application owners pay according to their usage.[101] Notable cloud-based MySQL services are the Amazon Relational Database Service; Oracle MySQL Cloud Service, Azure Database for MySQL, Rackspace; HP Converged Cloud; Heroku and Jelastic. In this model the database service provider takes responsibility for maintaining the host and database.

Advantages of MySQL

1. **Data Security**

MySQL is globally renowned for being the most secure and reliable database management system used in popular web applications including WordPress, Drupal, Joomla, Facebook and Twitter. The data security and support for transactional processing that accompany the recent version of MySQL can greatly benefit any business, especially if it is an eCommerce business that involves frequent money transfers.

2. **On-Demand Scalability**

MySQL offers unmatched scalability to facilitate the management of deeply embedded apps using a smaller footprint, even in massive warehouses that stack terabytes of data. On-demand flexibility is the star feature of MySQL. This open-source solution allows complete customization to eCommerce businesses with unique database server requirements.

3. **High Performance**

MySQL features a distinct storage-engine framework that facilitates system administrators to configure the MySQL database server for a flawless performance. Whether it is an eCommerce website that receives a million queries every single day or a high-speed transactional processing system, MySQL is designed to meet even the most demanding applications while ensuring optimum speed, full-text indexes and unique memory caches for enhanced performance.

4. **Round-the-Clock Uptime**

MySQL comes with the assurance of 24×7 uptime and offers a wide range of high-availability solutions, including specialized cluster servers and master/slave replication configurations.

5. **Comprehensive Transactional Support**

MySQL tops the list of robust transactional database engines available on the market. With features such as complete atomic, consistent, isolated, durable transaction support; multi-version transaction support; and unrestricted row-level locking, it is the go-to solution for full data integrity. It guarantees instant deadlock identification through server-enforced referential integrity.

6. **Complete Workflow Control**

With an average download and installation time of less than 30 minutes, MySQL means usability from day one. Whether your platform is Linux, Microsoft, Macintosh or UNIX, MySQL is a comprehensive solution with self-management features that automate everything from space expansion and configuration to data design and database administration.

7. **Reduced Total Cost of Ownership**

By migrating current database apps to MySQL, enterprises enjoy significant cost savings on new projects. The dependability and ease of management can save troubleshooting time that is otherwise wasted in fixing downtime issues and performance problems.

8. **The Flexibility of Open Source**

All the fears and worries that arise in an open-source solution can be brought to an end with MySQL's round-the-clock support and enterprise indemnification. The secure processing and trusted software of MySQL combine to provide effective transactions for large-volume projects. It makes maintenance, debugging and upgrades fast and easy while enhancing the end-user experience

4.SOFTWARE DEVELOPMENT ANALYSIS

4.1.OVERVIEW OF PROBLEM

Nowadays women are experiencing lots of violence such as harassment in places in several cities. This starts from stalking which then leads to abusive harassment or also called abuse assault. In this paper we mainly focus on the role of social media which can be used to promote the safety of women in India, given the special reference to the participation of many social media websites or applications such as Twitter, Facebook and Instagram platforms.

4.2.DEFINE THE PROBLEM

Sentiment Analysis (SA) is a subject of study that investigates people's sentiments, views, assessments, appraisals, attitudes and emotions in the direction of entities such as individuals, services, organizations, issues, products, topics and their characteristics. It is also known as opinion mining, sentiment mining, subjectivity analysis, review mining, opinion extraction, emotion analysis, etc. Furthermost of the prevailing approaches have used the terms sentiment analysis and opinion mining interchangeably. According to the research Mathematics, this mining is defined as a quintuple. Sentiment mining = (t, s, h, T) Where't' is the target opinion, 's' is the sentiment about 't', 'h' is the holder opinion and 'T' is the time. There are 3Approaches in Sentiment Analysis.

They are: 1. Machine-learning approach, 2. Lexicon-based approach, 3. Hybrid approach.

Sentiment methods Classification Pros and Cons Lexicon based Dictionary based. Corpus based. Ensemble approaches. Pros: Best for domain reliant on, larger-term coverage. Cons: Only Finite number of words in the lexicons Machine learning Based Support vector machines. Bayesian networks. Naïve Bayes. Random forest. Pros: Capacity to adjust and make prepared models for explicit purposes and settings. Cons: Low relevance for new information, since it is important of marked information. Hybrid based Lexicon and machine learning based. Pros: High exactness of new information. Slant vocabulary developed utilizing public assets for assumption discovery. Notion words as highlights in the AI techn4.

4.3.MODULES OVERVEIW

Modules are a way to organize your course by weeks, units, chapters, topics or whatever organizational structure works for your course. With modules, you create a one-directional linear flow of what you would like your students to do. Modules can be accessed by clicking the Modules button in the Course Tools Menu along the left side of any course. You may also choose to have the Modules page display as your course Home page.You are already familiar with modules, as this orientation course is set up using Modules. Each module can contain files, discussions, assignments, quizzes, and any other learning materials that you would like to use. You can easily add items to a module that you have already created in the course. You can also create new items on the fly within the module. This allows you to create the structure of the course while developing new learning materials.Modules can easily be reordered to fit the flow of the course by simply dragging and dropping. Elements within the modules can also be reorganized by dragging and dropping.

Modules can be released on specific dates. You can also create release conditions (e.g. a module cannot be accessed until a previous module has been completed). All modules appear on the Module page. At this time, you cannot hide modules. You can collapse modules so that only module headings display; however, any user can expand a module to display contents in his/her account. If a module is set to release on a specific date, students will be able to see a list of module contents, but the list will be grayed out and items will remain inaccessible until the release date.

1. TensorFlow
2. Pandas
3. Scikit – learn
4. Matplotlib
5. NumPy

4.4.DEFINE THE MODULES

TensorFlow:

TensorFlow is a free and open-source software library for dataflow and differentiable programming across a range of tasks.

It is a symbolic math library, and is also used for machine learning applications such as neural networks.

It is used for both research and production at Google.

TensorFlow was developed by the Google Brain team for internal Google use.

It was released under the Apache 2.0 open-source license on November 9, 2015.

Its particular focus is on training & interference of deep neural network.

It is written in Python, C++.

It works on Linux, macOS, Windows, Android, JavaScript.

Pandas is an open-source Python Library providing high-performance data manipulation and analysis tool using its powerful data structures.

Python was majorly used for data munging and preparation.

It had very little contribution towards data analysis. Pandas solved this problem.

Using Pandas, we can accomplish five typical steps in the processing and analysis of data, regardless of the origin of data load, prepare, manipulate, model, and analyze.

Python with Pandas is used in a wide range of fields including academic and commercial domains including

finance, economics, Statistics, analytics, etc

In particular, it offers data structures and operations for manipulating numeric tables and time series

The name is derived from the term “panel data”, an econometrics term for data sets. It’s name is play on the phase “python data analysis” itself.

It is written in Python, C.

It is initially released on 11th January 2008.

Scikit – learn:

Scikit-learn provides a range of supervised and unsupervised learning algorithms via a consistent interface in Python.

It is licensed under a permissive simplified BSD license and is distributed under many Linux distributions, encouraging academic and commercial use.

It is free software machine learning library for the python programming language.

It is initially released in June 2007.

Matplotlib:-

It is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms.

Matplotlib can be used in Python scripts, the Python and I Python shells, the Jupyter Notebook, web application servers, and four graphical user interface toolkits.

Matplotlib tries to make easy things easy and hard things possible.

You can generate plots, histograms, power spectra, bar charts, error charts, scatter plots, etc., with just a few lines of code. For examples, see the sample plots and thumbnail gallery.

For simple plotting the pyplot module provides a MATLAB-like interface, particularly when combined with I Python.

For the power user, you have full control of line styles, font properties, axes properties, etc, via an object oriented interface or via a set of functions familiar to MATLAB users

It is written in Python

It is released in 2003

It was originally Written by John D.Hunder

NumPy:

NumPy is a general-purpose array-processing package.

It provides a high performance multidimensional array object, and tools for working with these arrays.

It is the fundamental package for scientific computing with Python. It contains various features including these important ones:

A powerful N-dimensional array object

Sophisticated (broadcasting) functions

Tools for integrating C/C++ and Fortran code

Useful linear algebra, Fourier transform, and random number capabilities

Besides its obvious scientific uses, NumPy can also be used as an efficient multi-dimensional container of generic data.

It is Written in Python & C.

4.5.MODULE FUNCTIONALITY

Modules present:

1.Upload Image

2.Train Dataset

3.Upload Test & Classify.

4.user

5.Exit

1.Upload Image:

we apply each component on all the preparation pictures. For each component, it finds the best limit which will characterize the countenances to positive and negative. Be that as it may, clearly, there will be blunders or misclassifications. We select the elements with least mistake rate, which implies they are the elements that best orders the auto and non-auto pictures.

• So now you take a picture. Take each 24x24 window. Apply 6000 elements to it. Check on the off chance that it is auto or not.

2.Train Dataset:

Now every single conceivable size and areas of every part is utilized to ascertain a lot of components. (Simply envision what amount of calculation it needs? Indeed, even a 24x24 window comes about more than 160000

3.Upload Test & Classify:

This velocity and the distance of the camera in feet from the car (i.e. the height of camera above the car) is printed on the output screen. For this use multiple object detection algorithms could have been used but the algorithm of developing the Haar cascade and its implementation proves to be the best since it is the least time consuming, most efficient and highly reliable.

4. User:

User will login through the login with the credentials and will do the whole process.

5. Exit:

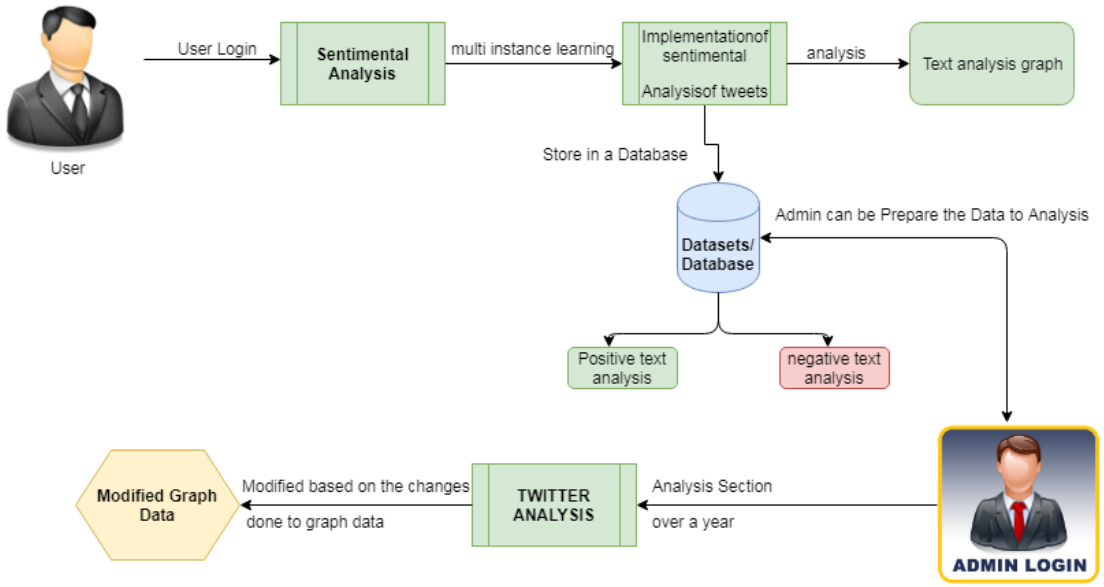
Once after the whole process is done with the vehicle detection the user will logout.

SUPPORT VECTOR MACHINE

"Support Vector Machine" (SVM) is a supervised machine learning algorithm which can be used for both classification and regression challenges. However, it is mostly used in classification problems. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiate the two classes very well (look at the below snapshot). Support Vectors are simply the co-ordinates of individual observation. Support Vector Machine is a frontier which best segregates the two classes (hyper-plane/ line). More formally, a support vector machine constructs a hyper plane or set of hyper planes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks like outliers detection. Intuitively, a good separation is achieved by the hyper plane that has the largest distance to the nearest training-data point of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier. Whereas the original problem may be stated in a finite dimensional space, it often happens that the sets to discriminate are not linearly separable in that space. For this reason, it was proposed that the

original finite-dimensional space be mapped into a much higher-dimensional space, presumably making the separation easier in that space.

5.PROJECT SYSTEM DESIGN



5.1.UML DIAGRAMS

UML stands for Unified Modeling Language. UML is a standardized general-purpose modeling language in the field of object-oriented software engineering. The standard is managed, and was created by, the Object Management Group.

The goal is for UML to become a common language for creating models of object oriented computer software. In its current form UML is comprised of two major components: a Meta-model and a notation. In the future, some form of method or process may also be added to; or associated with, UML.

The Unified Modeling Language is a standard language for specifying, Visualization, Constructing and documenting the artifacts of software system, as well as for business modeling and other non-software systems.

The UML represents a collection of best engineering practices that have proven successful in the modeling of large and complex systems.

The UML is a very important part of developing objects oriented software and the software development process. The UML uses mostly graphical notations to express the design of software projects. **GOALS:**

The Primary goals in the design of the UML are as follows:

1. Provide users a ready-to-use, expressive visual modeling Language so that they can develop and exchange meaningful models.
2. Provide extendibility and specialization mechanisms to extend the core concepts.
3. Be independent of particular programming languages and development process.
4. Provide a formal basis for understanding the modeling language.
5. Encourage the growth of OO tools market.
6. Support higher level development concepts such as collaborations, frameworks, patterns and components.
7. Integrate best practices.

Characteristics of UML

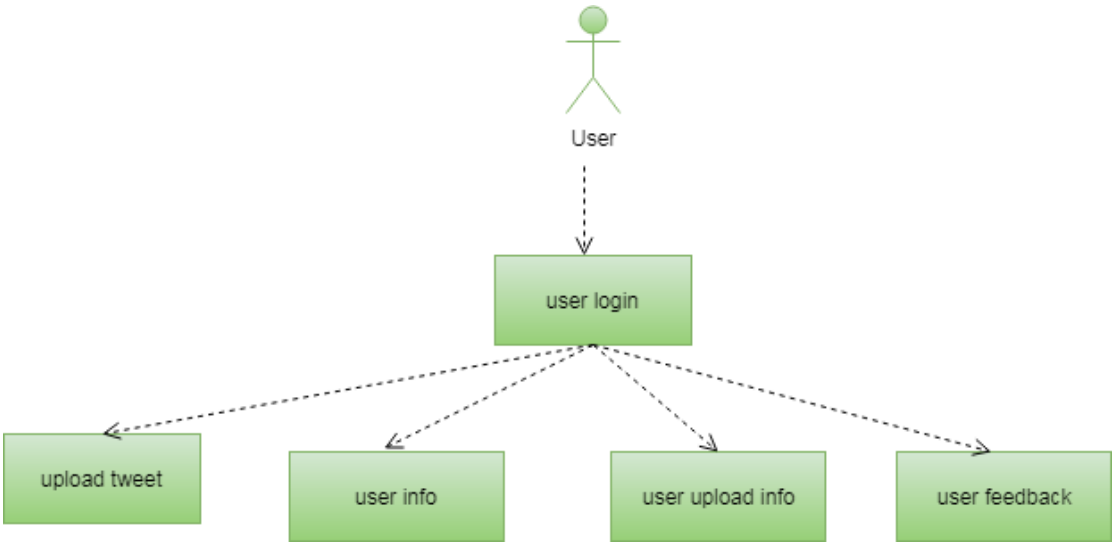
The UML has the following features:

- It is a generalized modeling language.
- It is distinct from other programming languages like C++, Python, etc.
- It is interrelated to object-oriented analysis and design.
- It is used to visualize the workflow of the system.
- It is a pictorial language, used to generate powerful modeling artifacts.

USE CASE DIAGRAM

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.

A.User



B.Admin

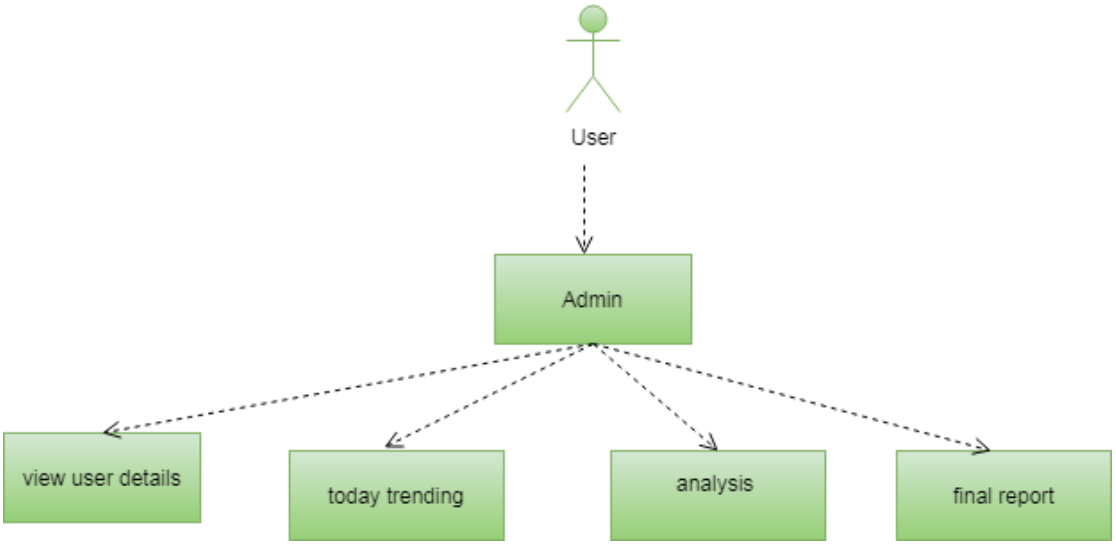
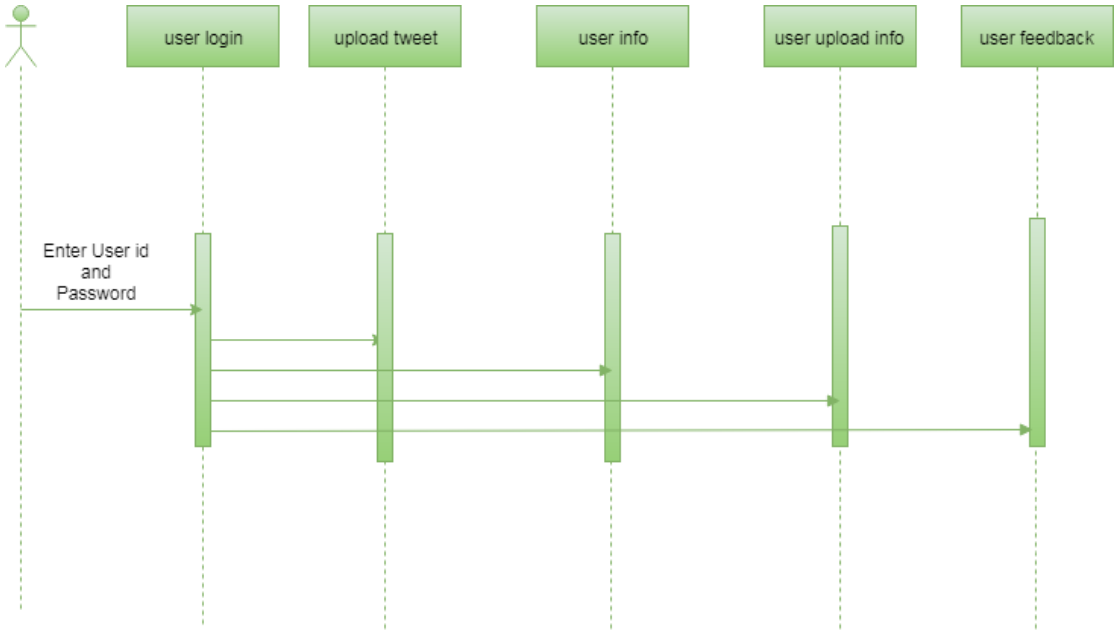


Fig 2.1 Use Case Diagram

SEQUENCE DIAGRAM

A sequence diagram in Unified Modeling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.

A.User



B.Admin

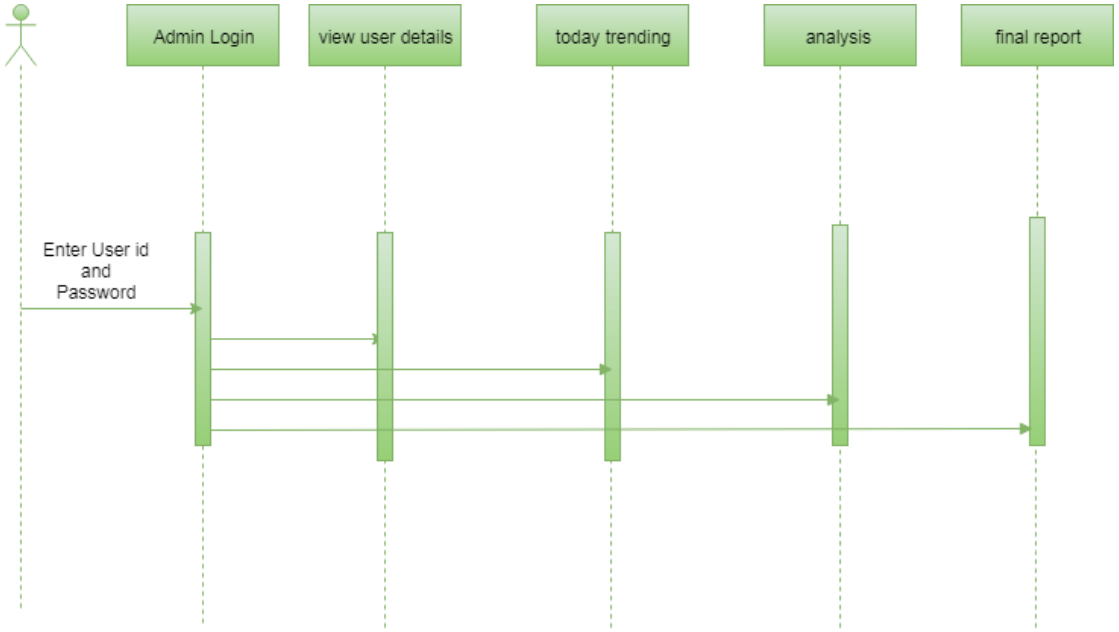
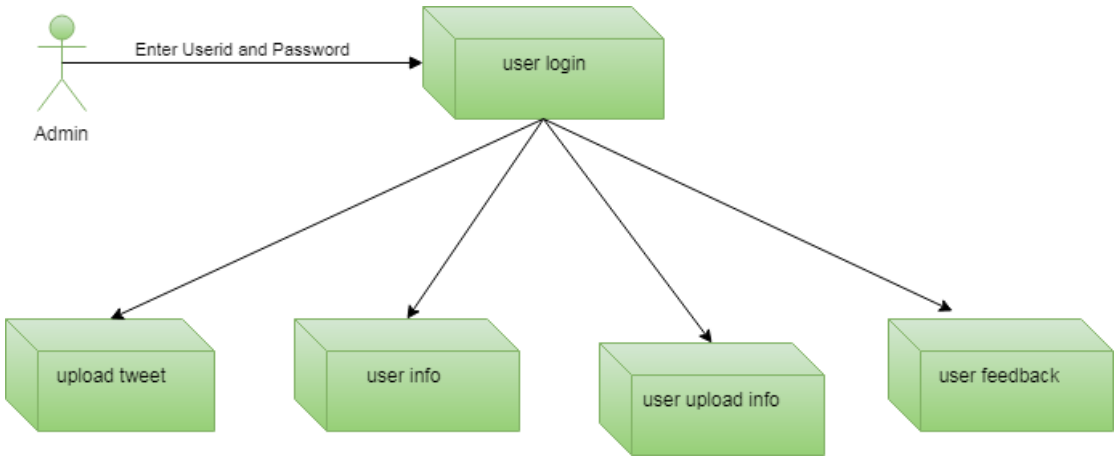


Fig 2.2 Sequence Diagram

COMPONENT DIAGRAM

a.User



b.Admin

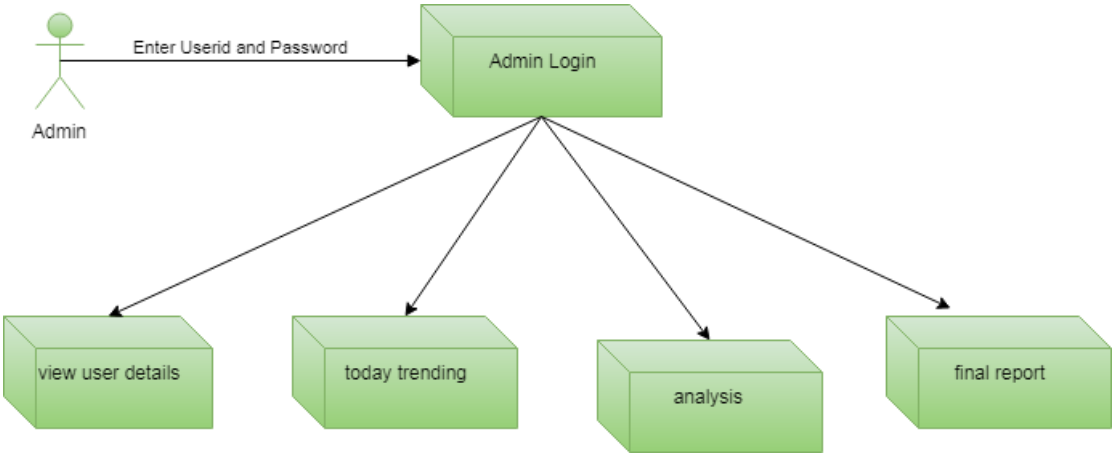


Fig 2.3 Component Diagram

CLASS DIAGRAM

In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.

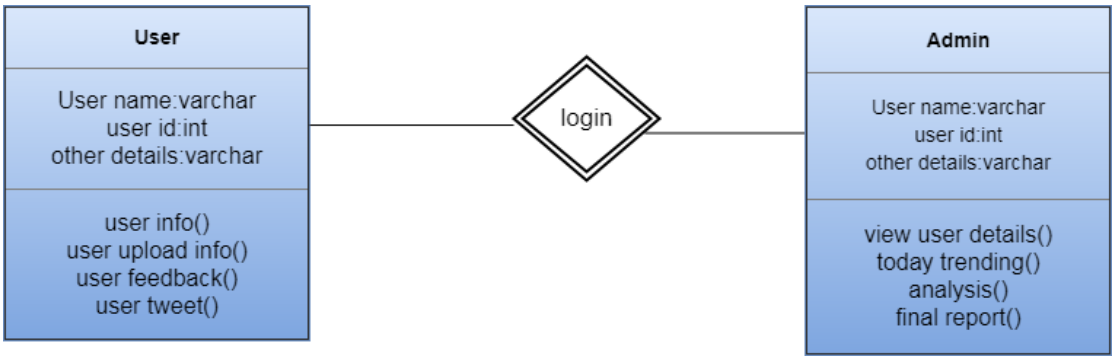


Fig 2.4 Class Diagram

6.PROJECT CODING

6.1.CODE TEMPLATES

```
import tkinter

from textblob import TextBlob

from tkinter import *

import matplotlib.pyplot as plt

import numpy as np

import pandas as pd

from string import punctuation

from nltk.corpus import stopwords

main = tkinter.Tk()

main.title("Analysis of Women Safety in Indian Cities Using Twitter data")

main.geometry("1200x1200")

global filename

tweets_list = []

clean_list = []

global pos, neu, neg

def tweetCleaning(doc):

tokens = doc.split()

table = str.maketrans('', '', punctuation)

tokens = [w.translate(table) for w in tokens]

tokens = [word for word in tokens if word.isalpha()]

stop_words = set(stopwords.words('english'))

tokens = [w for w in tokens if not w instop_words]
```

```
tokens = [word for word in tokens if len(word) > 1]

tokens = ' '.join(tokens)

return tokens

def upload():

global filename

filename = filedialog.askopenfilename(initialdir="dataset")

text.delete('1.0', END)

text.insert(END, filename + " loaded\n");

def read():

tweets_list.clear()

train = pd.read_csv(filename, encoding='iso-8859-1')

for i in range(len(train)):

tweet = train.get_value(i, 'Text')

tweets_list.append(tweet)

text.insert(END, tweet + "\n")

def clean():

text.delete('1.0', END)

clean_list.clear()

for i in range(len(tweets_list)):

tweet = tweets_list[i]

tweet = tweet.strip("\n")

tweet = tweet.strip()

tweet = tweetCleaning(tweet.lower())
```

```
clean_list.append(tweet)

text.insert(END, tweet + "\n")

def machineLearning():

text.delete('1.0', END)

global pos, neu, neg

pos = 0

neu = 0

neg = 0

for i in range(len(clean_list)):

tweet = clean_list[i]

blob = TextBlob(tweet)

if blob.polarity<= 0.2:

neg = neg + 1

text.insert(END, tweet + "\n")

text.insert(END, "Predicted Sentiment : NEGATIVE\n")

text.insert(END, "Polarity Score : " + str(blob.polarity) + "\n")

text.insert(END,'============================================================
==========================\n')

if blob.polarity> 0.2 and blob.polarity<= 0.5:

 neu = neu + 1

text.insert(END, tweet + "\n")

text.insert(END, "Predicted Sentiment : NEUTRAL\n")

text.insert(END, "Polarity Score : " + str(blob.polarity) + "\n")

text.insert(END,
```

```
'============================================================
==========================\n')
if blob.polarity> 0.5:
pos = pos + 1
text.insert(END, tweet + "\n")
text.insert(END, "Predicted Sentiment : POSITIVE\n")
text.insert(END, "Polarity Score : " + str(blob.polarity) + "\n")
text.insert(END,
'============================================================
==========================\n')
def bar():
label_X = []
category_X = []
text.delete('1.0', END)
text.insert(END,"Saftey Factor\n\n")
text.insert(END,'Positive : '+str(pos)+"\n")
text.insert(END,'Negative : '+str(neg)+"\n")
text.insert(END,'Neutral : '+str(neu)+"\n\n")
text.insert(END,'Length of tweets : '+str(len(clean_list))+"\n")
text.insert(END,'Positive : '+str(pos)+' / '+ str(len(clean_list))+' = '+str(pos/len(clean_list))+'%\n')
text.insert(END,'Negative : '+str(neg)+' / '+ str(len(clean_list))+' = '+str(neg/len(clean_list))+'%\n')
text.insert(END,'Neutral : '+str(neu)+' / '+ str(len(clean_list))+' = '+str(neu/len(clean_list))+'%\n')
label_X.append('Positive')
label_X.append('Negative')
```

```
label_X.append('Neutral')

category_X.append(pos)

category_X.append(neg)

category_X.append(neu)

plt.graph(category_X,labels=label_X,autopct='%1.1f%%')

plt.title('Women Saftey& Sentiment Graph')

plt.axis('equal') plt.show()

font = ('times', 16, 'bold')

title = Label(main, text='Analysis of Women Safety in Indian Cities Using Twitter data')

title.config(bg='purple', fg='white')

title.config(font=font)

title.config(height=3, width=120)

title.place(x=0, y=5)

font1 = ('times', 14, 'bold')

uploadButton = Button(main, text="Upload Tweets Dataset", command=upload)

uploadButton.place(x=50, y=100)

uploadButton.config(font=font1)

readButton = Button(main, text="Read Tweets", command=read)

readButton.place(x=50, y=150)

readButton.config(font=font1)

text = Text(main, height=25, width=150)

scroll = Scrollbar(text)

text.configure(yscrollcommand=scroll.set)
```

```
cleanButton = Button(main, text="Tweets Cleaning", command=clean)

cleanButton.place(x=210, y=150)

cleanButton.config(font=font1)

text = Text(main, height=25, width=150) scroll = Scrollbar(text)

text.configure(yscrollcommand=scroll.set)

mlButton = Button(main, text="Run Machine Learning Algorithm", command=machineLearning)

mlButton.place(x=400, y=150)

mlButton.config(font=font1)

graphButton = Button(main, text="Women Saftey Graph", command=graph)

graphButton.place(x=730,y=150)

graphButton.config(font=font1)

font1 = ('times', 12, 'bold')

text = Text(main, height=25, width=150)

scroll = Scrollbar(text)

text.configure(yscrollcommand=scroll.set)

text.place(x=10, y=200)

text.config(font=font1)

main.config(bg='purple')

main.mainloop()
```

6.2.OUTLINE FOR VARIOUS FILES:

We used Python programming to implement our project. A single python file is used to implement our code. This file consists of various modules that we have used. Our project

modules are - Upload Image, Train Dataset, Upload Test & Classify, User and exit. We also used various python modules like pandas, matplotlib, numpy, tensorflow, sklearn.

Python too supports file handling and allows users to handle files i.e., to read and write files, along with many other file handling options, to operate on files. The concept of file handling has stretched over various other languages, but the implementation is either complicated or lengthy, but alike other concepts of Python, this concept here is also easy and short. Python treats file differently as text or binary and this is important. Each line of code includes a sequence of characters and they form text file. Each line of a file is terminated with a special character, called the EOL or End of Line characters like comma {,} or newline character. It ends the current line and tells the interpreter a new one has begun. Let's start with Reading and Writing files.

Reading and writing data to files using Python is pretty straightforward. To do this, you must first open files in the appropriate mode. Here's an example of how to use Python's "with open(…) as …" pattern to open a text file and read its contents:

```
withopen('data.txt','r')asf:
data=f.read()
```

open() takes a filename and a mode as its arguments. r opens the file in read only mode.

NumPy introduces a simple file format for ndarray objects. This .npy file stores data, shape, dtype and other information required to reconstruct the ndarray in a disk file such that the array is correctly retrieved even if the file is on another machine with different architecture.

6.3.CLASS WITH FUNCTIONALITY:

In our project code, we implemented six different methods. They are:

1.UploadImage()

2.TrainDataset()

3.UploadTest& Classify ()

4.user()

5.Exit()

Our first method upload Image() doesn't take any input parameters but after successful execution, it displays a message "Image loaded". Our second method TrainDataset () will take 1 parameters and it will train the whole dataset being considered.Our third .UploadTest& Classify () study all the images and will predict the car model based on the pattern.the last two are the important methods that are mainly used for the logging and logout purpose of the project. Once after the prediction is using the image user will exit with the help of the exit () method.

6.4.METHODS INPUT AND OUTPUT PARAMETERS

INPUT DESIGN

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the following things:

- What data should be given as input?
- How the data should be arranged or coded?
- The dialog to guide the operating personnel in providing input.
- Methods for preparing input validations and steps to follow when error occur.

OBJECTIVES

1.Input Design is the process of converting a user-oriented description of the input into a computer-based system. This design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerized system.

2.It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The goal of designing input is to make data entry easier and to be free from

errors. The data entry screen is designed in such a way that all the data manipulates can be performed. It also provides record viewing facilities.

3.When the data is entered it will check for its validity. Data can be entered with the help of screens. Appropriate messages are provided as when needed so that the user will not be in maize of instant. Thus the objective of input design is to create an input layout that is easy to follow

OUTPUT DESIGN

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making.

1. Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is designed so that people will find the system can use easily and effectively. When analysis design computer output, they should Identify the specific output that is needed to meet the requirements.

2.Select methods for presenting information.

3.Create document, report, or other formats that contain information produced by the system.

The output form of an information system should accomplish one or more of the following objectives.

- Convey information about past activities, current status or projections of the
- Future.
- Signal important events, opportunities, problems, or warnings.
- Trigger an action.
- Confirm an action.

7.PROJECT TESTING

7.1.VARIOUS TEST CASES

UNIT TESTING

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

INTEGRATION TESTING

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

Functional test

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items:

Valid Input : identified classes of valid input must be accepted.

Invalid Input : identified classes of invalid input must be rejected.

Functions : identified functions must be exercised.

Output : identified classes of application outputs must be exercised.

Systems/Procedures : interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

System Test

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

7.2.BLACK BOX TESTING

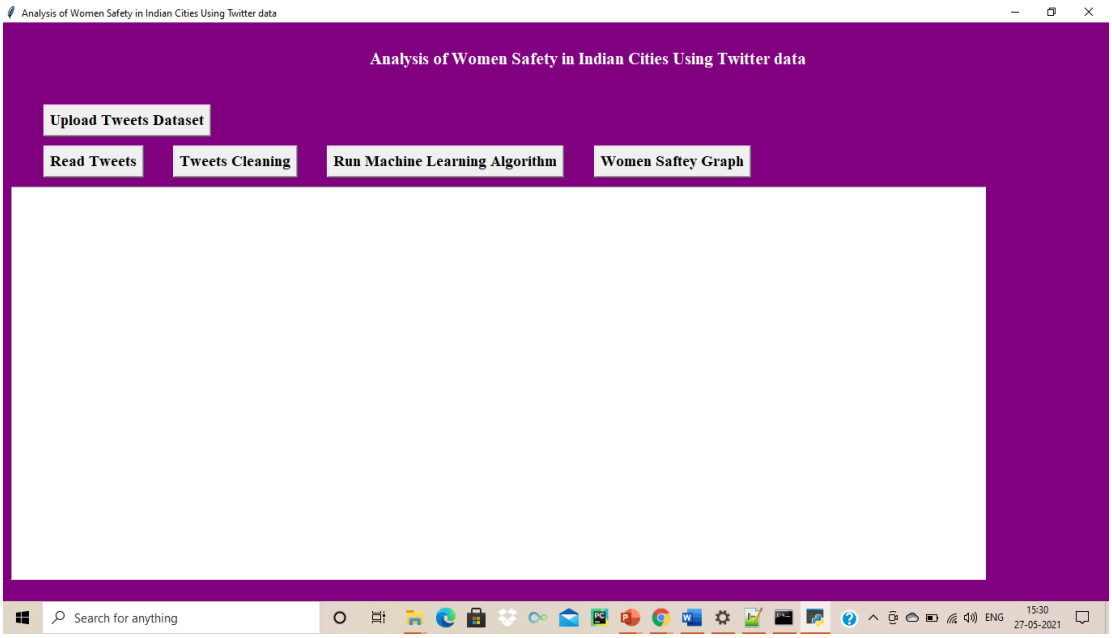
Black Box Testing is testing the software without any knowledge of the inner workings, structure or
language of the module being tested. Black box tests, as most other kinds of tests, must be written
from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated,
as a black box .you cannot “see” into it. The test provides inputs and responds to outputs without
considering how the software works.

7.3.WHITE BOX TESTING

White Box Testing is a testing in which in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is purpose. It is used to test areas that cannot be reached from a black box level.

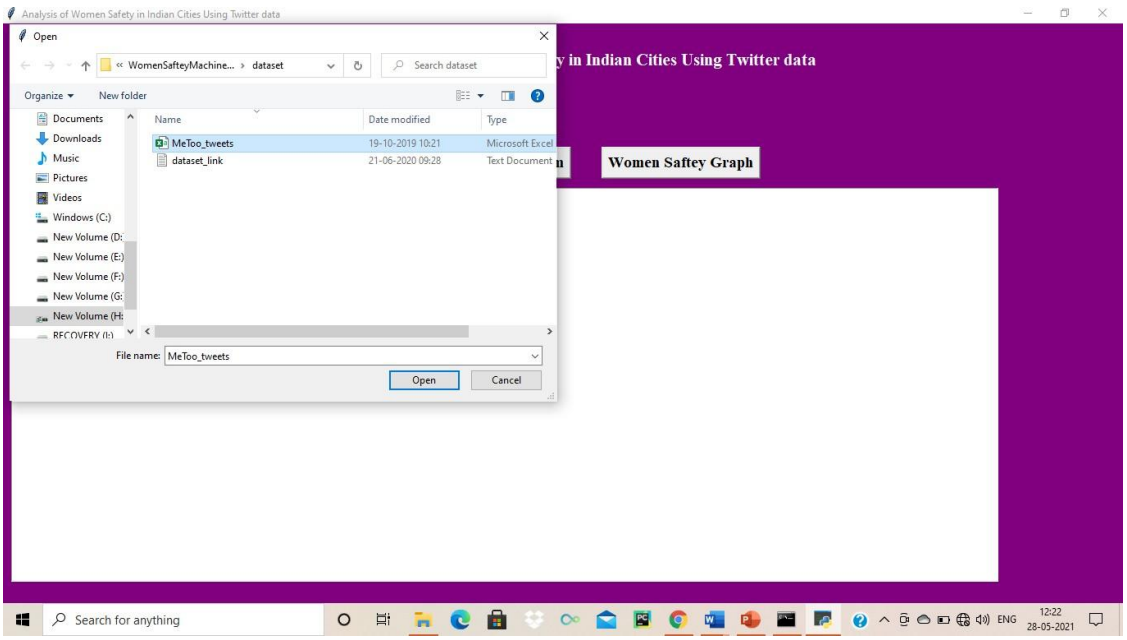
8.OUTPUT SCREENS

8.1.USER INTERFACES

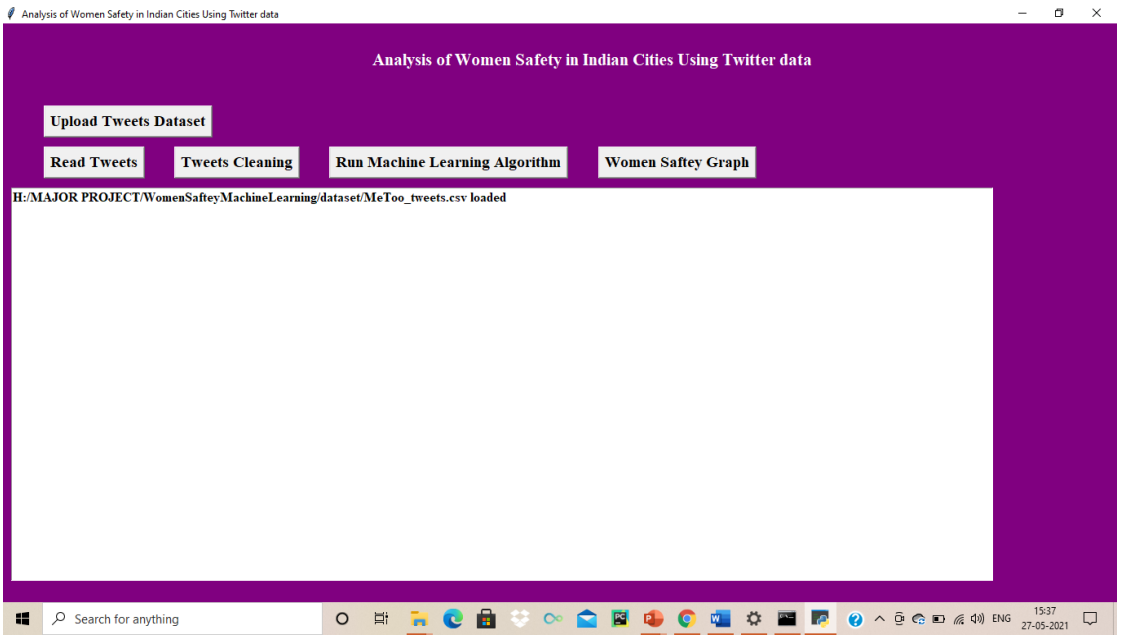


In abovescreenclickon‘UploadTweetsDataset’buttonanduploadtweets

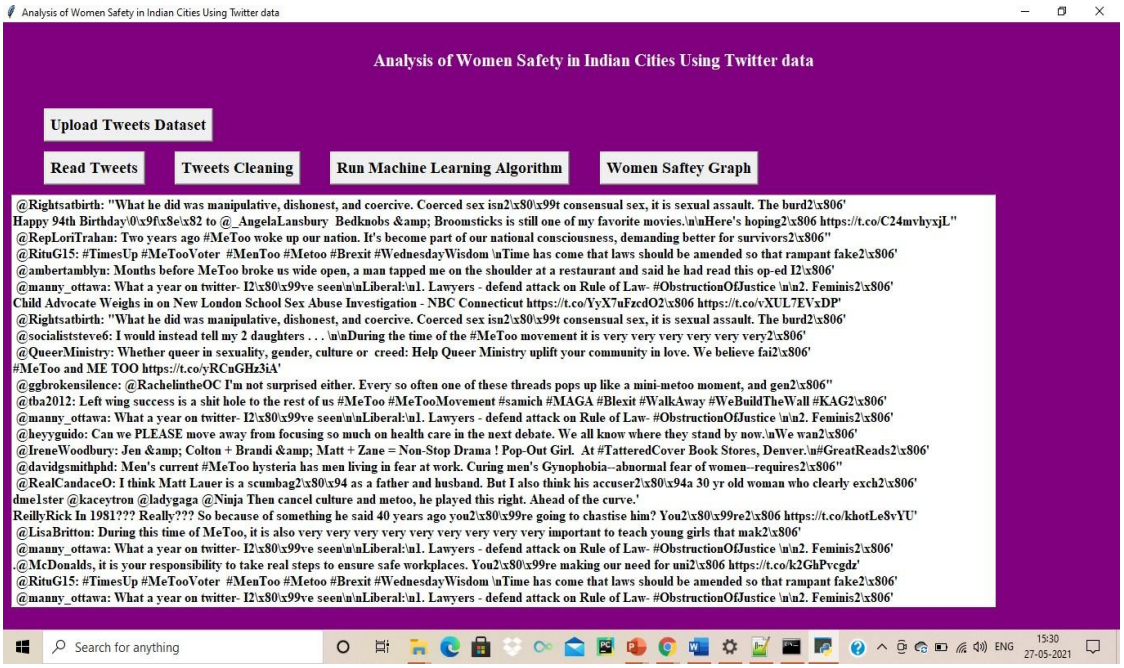
8.2.OUTPUT SCREENS



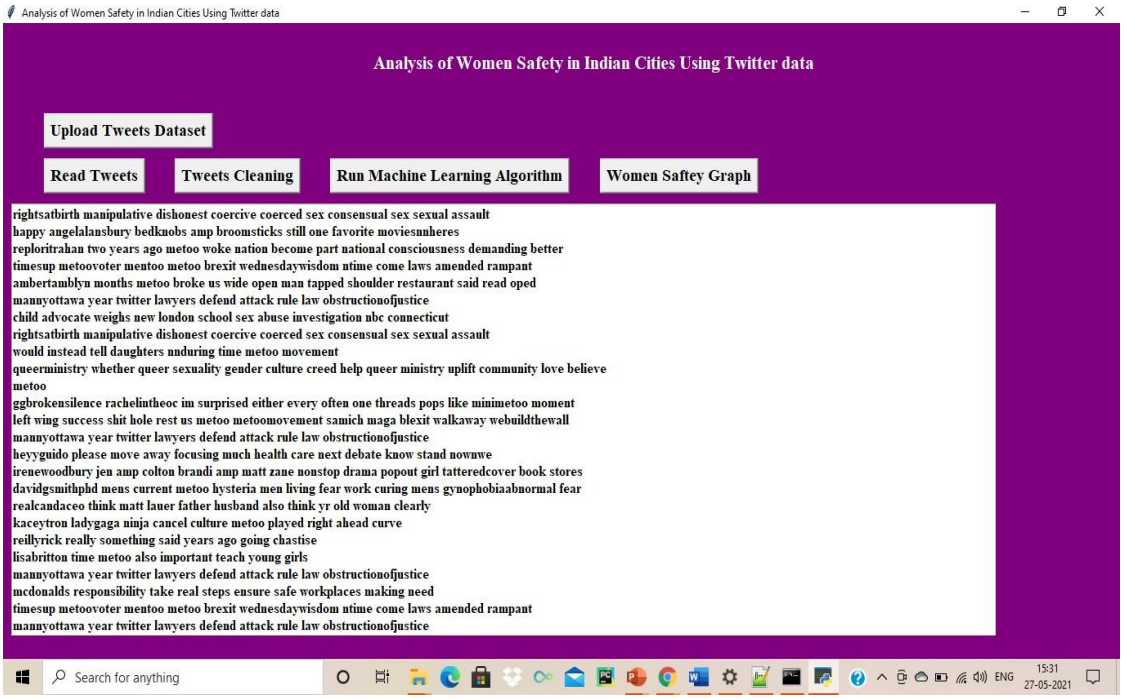
Uploading Dataset



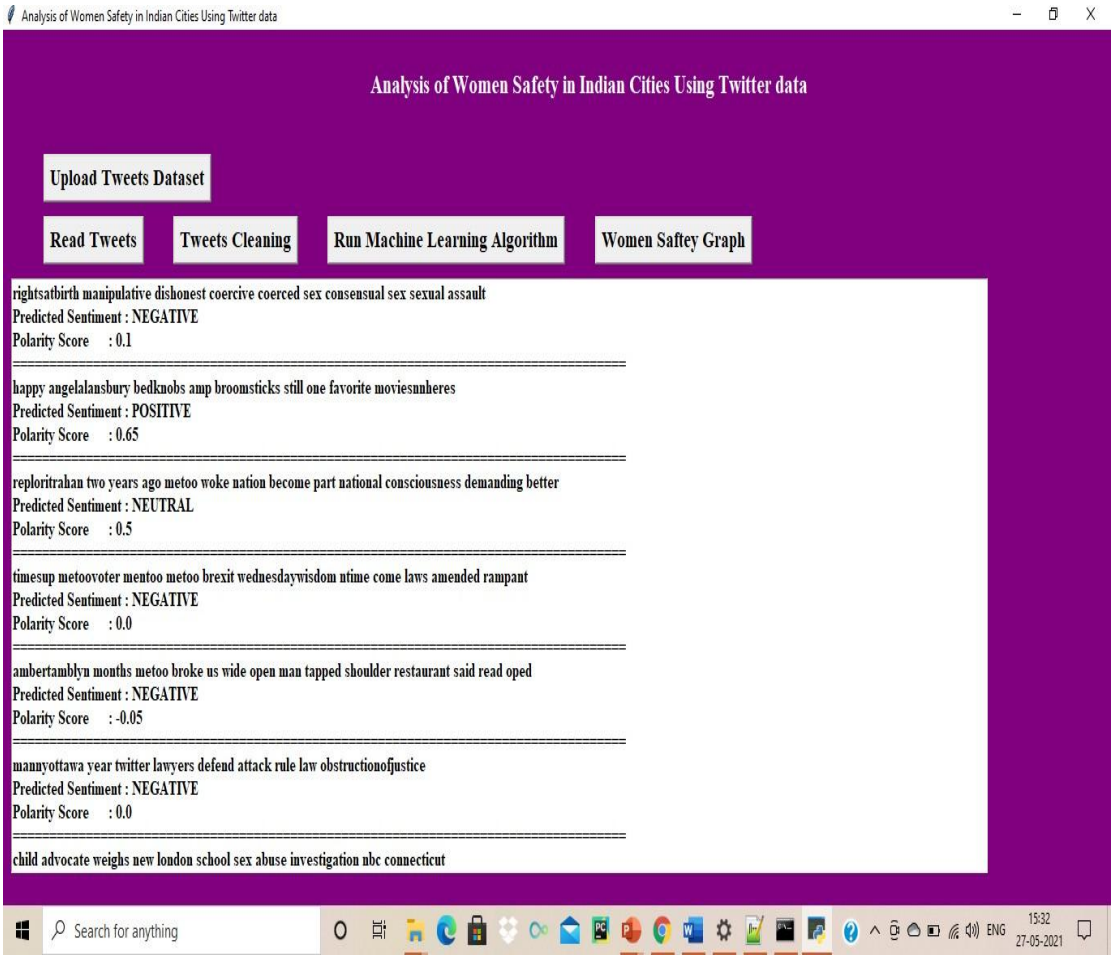
Dataset Uploaded



ReadingTweets

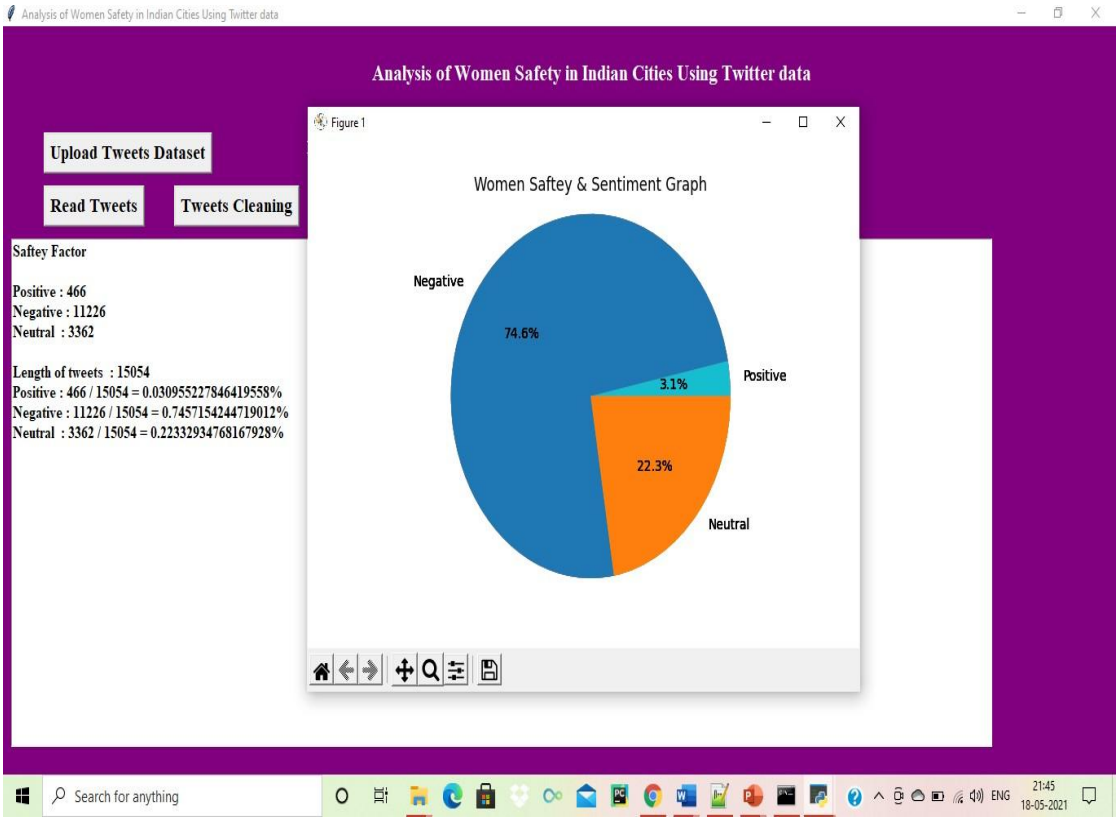


In above screen we can see all special symbols and stop words remove from tweetsandonlycleanwordsarethereandnowclickon‘RunMachineLearningAlgorithm’butt onto predict sentiments from tweets.



In above screen each tweet having tweet text and then displaying tweets sentimentswithpolarity score. Scroll downabovetext areatoseeall tweets.

9.EXPERIMENTAL RESULTS



Now click on 'Women Safety Graph' button to get below results and by seeing thatresultusercaneasilyunderstandwhetherareaissafeornot.Ifareaissafethenmorepeopl es will express either positive or neutral tweets and if not safe then morepeopleswill discuss negativetweets.

10. CONCLUSION AND FUTURE ENHANCEMENT

Throughout the research paper we have discussed about various machine learning algorithms that can help us to organize and analyze the huge amount of Twitter data obtained including

millions of tweets and text messages shared every day. These machine learning algorithms are very effective and useful when it comes to analyzing of large amount of data including the SPC algorithm and linear algebraic Factor Model approaches which help to further categorize the data into meaningful groups. Support vector machines is yet another form of machine learning algorithm that is very popular in extracting Useful information from the Twitter and get an idea about the status of women safety in Indian cities.

REFERENCES

1 VIKRAMCHANDRA&RAMPURSRINATH(2020).“AnalysisofWomenSafetyusingMachineLearningonTweets”.InternationalResearch Journal of Engineering and Technology (IRJET) p-ISSN: 2395-0072

2 Edukondalu, B., and P. Neelima. "Sentiment Analysis on Social medianetwork." *InternationalJournalonFutureRevolutioninComputerScience &Communication Engineering*6.2(2020):01-08.

3 Kumar, Deepak, and Shivani Aggarwal. "Analysis of women safety inIndian cities using machine learning on tweets." *2019 Amity InternationalConferenceon Artificial Intelligence(AICAI)*.IEEE,2019.

4 Gupta B, Negi M, Vishwakarma K, Rawat G &Badhani P (2017). “StudyofTwittersentimentanalysisusingmachinelearningalgorithmsonPython.” International Journal of Computer Applications, 165(9) 0975-8887.

5 Mamgain N, Mehta E, Mittal A & Bhatt G (2016, March). “Sentimentanalysis of top colleges in India using Twitter data.” In ComputationalTechniques,inInformationandCommunicationTechnologies(ICCT ICT),2016 International Conferenceon (pp.525-530).IEEE

6 Shah, S., Kumar, K., &Sarvananguru, R. K. (2016). Sentimental analysisoftwitterdatausingclassifieralgorithms. *InternationalJournalofElectrical andComputer Engineering*,*6*(1),357.

7 Sahayak V, Shete V & Pathan A (2015). “Sentiment analysis on twitterdata.”InternationalJournalofInnovativeResearchinAdvancedEngineering (IJIRAE),2(1),178-183.

8 Agarwal, A., Xie, B., Vovsha, I., Rambow, O., &Passonneau, R. J.(2011, June). Sentiment analysis of twitter data. In *Proceedings of theworkshoponlanguageinsocial media (LSM2011)*(pp.30-38).

9 Jiang, L., Yu, M., Zhou, M., Liu, X., & Zhao, T. (2011, June). Target-dependenttwittersentimentclassification.In*Proceedingsofthe49th*

annualmeetingoftheassociationforcomputationallinguistics:humanlanguagetechnologies(pp.151-160).

10 Kouloumpis, E., Wilson, T., & Moore, J. (2011, July). Twitter sentimentanalysis:Thegoodthebadandtheomg!.In *ProceedingsoftheInternationalAAAIConferenceonWebandSocialMedia*(Vol.5,No.1).

11 Pak, A., &Paroubek, P. (2010, May). Twitter as a corpus for sentimentanalysisandopinionmining.In *LREc*(Vol.10,No.2010,pp.1320-1326).

12 Bifet,A.,&Frank,E.(2010,October).Sentimentknowledgediscoveryintwitterstreamingdata.In *Internationalconferenceondiscoveryscience*(pp.1-15).Springer,Berlin,Heidelberg.

13 Barbosa,Luciano,andJunlanFeng."Robustsentimentdetectionontwitter from biased and noisy data." Proceedings of the 23rd internationalconferenceoncomputationallinguistics:posters.AssociationforComputational Linguistics,2010.

14 Bermingham, Adam, and Alan F. Smeaton. "Classifying sentiment inmicroblogs: is brevity an advantage?." Proceedings of the 19th ACMinternationalconferenceonInformationandknowledgemanagement.ACM,2010.

15 Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classificationusingdistantsupervision.*CS224Nprojectreport,Stanford*,*1*(12),2009.

16 Agarwal, Apoorv, FadiBiadsy, and Kathleen R. Mckeown. "Contextualphrase-level polarity analysis using lexical affect scoring and syntactic n-grams." Proceedings of the 12th Conference of the European Chapter oftheAssociationforComputationalLinguistics.AssociationforComputational Linguistics,2009.

17 Eugene Charniak and Mark Johnson. "Coarse-to-fine nbest parsing andMaxEntdiscriminativereranking."Proceedingsofthe43rdannualmeetingonassociationforcomputationallinguistics.AssociationforComputational Linguistics,2005.

18 Michael Gamon. "Sentiment classification on customer feedback data:noisydata,largefeaturevectors,andtheroleoflinguisticanalysis."Proceedingsofthe20thinternationalconferenceonComputational Linguistics.Association forComputationalLinguistics,2004.

19 Soo-MinKimandEduardHovy."Determiningthesentimentof opinions."Proceedingsofthe20thinternationalconference on Computational Linguistics. Association for Computational Linguistics, 2004

20 DanKleinandChristopherD.Manning."Accurateunlexicalizedparsing." Proceedings of the 41st Annual Meeting on Association for ComputationalLinguisticsVolume1.AssociationforComputational Linguistics,2003.

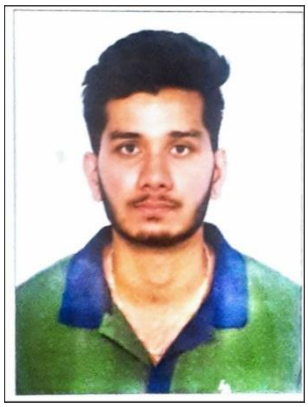
PUBLICATIONS

JOURNAL (UGC APPROVED JOURNAL)

CONFERENCE(INTERNATIONAL CONFERENCE ON "INNOVATIONS IN COMPUTERS NETWORKS,COMPUTATIONAL INTELLIGENCE AND IOT"[ICICCI-21,D4 BATCH].

TOPIC:WOMEN SAFETY IN INDIAN CITIES USING MACHINE LEARNING ON TWEETS.

STUDENT PROFILE



Anirudh Aiysolais currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. He completed his intermediate from Narayana Junior College and 10^{th} class from Narayana High School. His Technical Skills include C and Python .He also has a basic understanding of C++. He had also done Participation in Enterpreneurship ESUMMIT certified at MLRIT College and Machine Learning Workshop. Some External skills are: He is a college ambassador in youth marketing company called Grapevine . He also completed some certifications in Artificial Inteligence, MYSQL Database and Java Script by the Net Ninja.

CH.Sonanjali Devi is currently pursuing her Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. She completed her intermediate from Sri Chaitanya Junior College and 10th class from Bhashyam High School. Her Technical Skills include C and Java .She also has a basic understanding of C++. She had also done Participation in Enterpreneurship ESUMMIT certified at Sri Nidhi Engineering College and Machine Learning Workshop. Some External skills are: She is a college ambassador in youth marketing company called Grapevine and also she participated in district wise Athletics and Basketball competetions. She also completed some certifications in Artificial Inteligence and Java Script by the Net Ninja.

KantipamuDhaminiis currently pursuing her Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. She completed her intermediate from Sri Chaitanya Junior Kalasala and 10th class from Pratibha Vidyaniketan High School. Her Technical Skills include C ,MYSQL and Python . She had also done Participation in ESUMMIT certified at MLRIT College and Machine Learning Workshop. Some External skills are: She is a college ambassador in youth marketing company called Grapevine and also she participated in district wise Athletics and Basketball competetions. She also completed some certifications in Artificial Inteligence, Cybersecurity and Java Script by the Net Ninja.

Shubhanshi Pandey is currently pursuing her Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. She completed her intermediate from Narayana Junior College and 10^{th} grade from JMK International School. Her Technical Skills include C and Python .She also has a basic understanding of C++. She had also done Participation in Enterpreneurship ESUMMIT certified at MLRIT College and Machine Learning Workshop. Some External skills are: She is a college ambassador in youth marketing company called Viral Vision . She has also participated in district wise Badminton competitions. She has also completed some certifications in Artificial Inteligence, MYSQL Database and Amazon Web Services by coursera.

APPENDICES

[1] A.Pak and P. Paroubek. „Twitter as a Corpus for Sentiment Analysis and Opinion Mining". In Proceedings of the Seventh Conference on International Language Resources and Evaluation, 2010, pp.1320-1326

[2] R. Parikh and M. Movassate, “Sentiment Analysis of User- GeneratedTwitter Updates using Various Classi_cation Techniques",CS224N Final Report, 2009

[3] Go, R. Bhayani, L.Huang. “Twitter Sentiment ClassificationUsing Distant Supervision". Stanford University, Technical Paper,2009

[4] L. Barbosa, J. Feng. “Robust Sentiment Detection on Twitterfrom Biased and Noisy Data". COLING 2010: Poster Volume,pp. 36-44.

[5] Bifet and E. Frank, "Sentiment Knowledge Discovery inTwitter Streaming Data", In Proceedings of the 13th InternationalConference on Discovery Science, Berlin, Germany: Springer,2010, pp. 1-15.

[6] Agarwal, B. Xie, I. Vovsha, O. Rambow, R. Passonneau, “Sentiment Analysis of Twitter Data", In Proceedings of the ACL 2011Workshop on Languages in Social Media,2011 , pp. 30-38

[7] Dmitry Davidov, Ari Rappoport." Enhanced Sentiment Learning Using Twitter Hashtags and Smileys". Coling 2010: Poster Volumepages 241{249, Beijing, August 2010

[8] Po-Wei Liang, Bi-Ru Dai, “Opinion Mining on Social MediaData", IEEE 14th International Conference on Mobile Data Management,Milan, Italy, June 3 - 6, 2013, pp 91-96, ISBN: 978-1-494673-6068-5, http://doi.ieeecomputersociety.org/10.1109/MDM.2013.

[9] Pablo Gamallo, Marcos Garcia, “Citius: A Naive-Bayes Strategyfor Sentiment Analysis on English Tweets", 8th InternationalWorkshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland,Aug 23-24 2014, pp 171-175.

[10] Neethu M,S and Rajashree R,” Sentiment Analysis in Twitter using Machine Learning Techniques” 4th ICCCNT 2013,at Tiruchengode, India. IEEE – 31661

[11] P. D. Turney, “Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews,” in Proceedings of the 40th annual meeting on association for computational linguistics, pp. 417–424, Association for Computational Linguistics, 2002.

[12] J. Kamps, M. Marx, R. J. Mokken, and M. De Rijke, “Using wordnet to measure semantic orientations of adjectives,” 2004.

[13] R. Xia, C. Zong, and S. Li, “Ensemble of feature sets and classification algorithms for sentiment classification,” Information Sciences: an International Journal, vol. 181, no. 6, pp. 1138–1152, 2011.

[14] ZhunchenLuo, Miles Osborne, TingWang, An effective approachto tweets opinion retrieval", Springer Journal onWorldWideWeb,Dec 2013, DOI: 10.1007/s11280-013- 0268-7.

[15] Liu, S., Li, F., Li, F., Cheng, X., &Shen, H.. Adaptive cotraining SVM for sentiment classification on tweets. In Proceedings of the 22nd ACMinternational conference on Conference on information &knowledgemanagement (pp. 2079-2088). ACM,2013.

[16] Pan S J, Ni X, Sun J T, et al. “Cross-domain sentiment classification viaspectral feature alignment”. Proceedings of the 19th internationalconference on World wide web. ACM, 2010: 751-760.

[17] Wan, X..“A Comparative Study of Cross-Lingual SentimentClassification”. In Proceedings of the The 2012 IEEE/WIC/ACMInternational Joint Conferences on Web Intelligence and IntelligentAgent Technology-Volume 01 (pp. 24-31).IEEE Computer Society.2012

[18] Socher, Richard, et al. "Recursive deep models for semanticcompositionality over a sentiment Treebank." Proceedings of theConference on Empirical Methods in Natural Language Processing(EMNLP). 2013.

[19] Meng, Xinfan, et al. "Cross-lingual mixture model for sentimentclassification." Proceedings of the 50th Annual Meeting of theAssociation for Computational Linguistics Volume 1,2012

[20] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., &Stede, M..“Lexiconbasedmethods for sentiment analysis”. Computational linguistics, 2011:37(2), 267-307.

[21] Li, S., Xue, Y., Wang, Z., & Zhou, G..“Active learning for cross-domainsentiment classification”. In Proceedings of the Twenty-Thirdinternational joint conference on Artificial Intelligence (pp. 2127-2133).AAAI Press,2013

[22] Bollegala, D., Weir, D., & Carroll, J.. Cross-Domain SentimentClassification using a Sentiment Sensitive Thesaurus. Knowledge andData Engineering, IEEE Transactions on, 25(8), 1719-1731,2013

[23] Pang, B.and Lee, L. “A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts”. 42nd Meeting of the Association for Computational Linguistics[C] (ACL-04). 2004, 271-278. [24] V. M. K. Peddinti and P. Chintalapoodi, “Domain adaptation in sentiment analysis of twitter,” in Analyzing Microtext Workshop, AAAI, 2011.

A

PROJECT REPORT

On

ACCIDENT DETECTION SYSTEM

Submitted by

1) Mr.Suyash Singh (17K81A05N5)

2) Mr.AzeemPasha (17K81A05L8)

3) Mr.B.Vikas (17K81A05J8)

4) Mr.T.Saiteja (17K81A05P1)

in partial fulfilment for the award of the

degree of

BACHELOR OF TECHNOLOGY

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Under the Guidance of

Mr. C.Yosepu B.Tech,M.Tech,(Ph.D)

Associate Professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ST.MARTIN'S ENGINEERING COLLEGE

An Autonomous Institute
Dhulapally, Secunderabad – 500 100
JUNE 2021

BONAFIDE CERTIFICATE

This is to certify that the project entitled **ACCIDENT DETECTION SYSTEM, is** being submitted by **Mr. THOTA SAI TEJA 17K81A05P1, Mr. BEGARI VIKAS 17K81A05J8, Mr. SUYASH SINGH 17K81A05N5, Mr. AZEEM PASHA 17K81A05L8,** in partial fulfilment of the requirement for the award of the degree of **BACHELOR OF TECHNOLOGY in COMPUTER SCIENCE AND ENGINEERING is** recorded of bonafide work carried out by them. The result embodied in this report have been verified and found satisfactory.

Project Guide

Mr.C.Yosepu

Department of CSE

Head of the Department

Dr.M.NARAYANAN

Department of CSE

Internal Examiner

External Examiner

Place:

Date:

DECLARATION

We, the student of Bachelor of Technology in Department of Computer Science and Engineering', session: 2017 – 2021, St. Martin's Engineering College, Dhulapally, Kompally, Secunderabad, hereby declare that work presented in this Project Work entitled Accident Detection System is the outcome of our own bonafide work and is correct to the best of our knowledge and this work has been undertaken taking care of Engineering Ethics. This result embodied in this project report has not been submitted in any university for award of any degree.

Suyash Singh (17K81A05N5)

B.Vikas (17K81A05J8)

T.Saiteja (17K81A05P1)

Azeem Pasha (17K81A05L8)

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible and whose encouragements and guidance have crowded effects with success.

We extended our deep sense of gratitude to Principal, **Dr. P. SANTOSH KUMAR PATRA**, St. Martin's Engineering College, Dhulapally, for permitting us to undertake this project.

We are also thankful to **Dr.M.NARAYANAN**, Head of the Department, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

We would like to thank **Dr. T. POONGOTHAI**, Department of Computer Science and Engineering (AI & ML) for her encouragement and insightful comments and as well as our project coordinators **Dr.B.RAJALINGAM**, Associate Professor and **Mr. J.SUDHAKAR**, Assistant Professor, in Department of Computer Science and Engineering for their valuable support.

We would like to express our sincere gratitude and indebtedness to our project supervisor **Mr.C.Yosepu,** Associate Professor Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

Finally, we express thanks to all those who have helped us successfully to completing this project. Furthermore, we would like to thank our family and friends for their moral support and encouragement

We express thanks to all those who have helped us in successfully completing the project.

Suyash Singh(17K81A05N5)

B.Vikas (17K81A05J8)

T.Saiteja (17K81A05P1)

Azeem Pasha (17K81A05L8)

ABSTRACT

Accidents have been a major cause of deaths in India. More than 80% of accident-related deaths occur not due to the accident itself but the lack of timely help reaching the accident victims. In highways where the traffic is really light and fast-paced an accident victim could be left unattended for a long time. The intent is to create a system which would detect an accident based on the live feed of video from a CCTV camera installed on a highway. The idea is to take each frame of a video and run it through a deep learning convolution neural network model which has been trained to classify frames of a video into accident or non-accident. Convolutional Neural Networks has proven to be a fast and accurate approach to classify images. CNN based image classifiers have given accuracy's of more than 95% for comparatively smaller datasets and require less preprocessing as compared to other image classifying algorithms.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF OUTPUT SCREENS

LIST OF ACRONYMS

ROI	Region of interest
BMP	Bitmap
UML	Unified modelling language
CNN	Convolution neural network
GUI	Graphical User Interface

CHAPTER 1

INTRODUCTION

1. INTRODUCTION

Recent inter vehicular studies are acquiring commercial interest via the DSRC/WAVE standard in Vehicular Ad Hoc Networks (VANETs). Possible future services among vehicles are topic of many studies .In VANETs, vehicles are able to communicate with each other in vehicle-to vehicle (V2V) or with roadside network infrastructure in vehicle-to-Roadside Communication (V2R) manner.

Some of the envisioned applications for vehicular networks are : vehicle collision warning, security distance warning, driver assistance, cooperative driving, cooperative cruise control, dissemination of road information, internet access, map location, automatic parking, driverless vehicles(Boukerche et al., 2008)

Most of applications need traffic speed and travel time measurements. These measurements can be used to help roadway users to decide which route to use or when to depart etc. Also these measurement can be saved to analyze traffic speed and travel time patterns for different time intervals. Currently local detectors at specific points along the road are used to measure the speed. New approach is to equip vehicles with communication and location devices to measure their speed and travel time. Some studies have shown that cellular networks can be used to identify vehicle's location using cellular phone base station communication records(Bar-Gera, 2007)

Safe navigation support has also become one of the main research topic with the help of DSRC/WAVE standardization(Jiang et al., 2006). For instance, collision or road condition warning messages can be forwarded to following vehicles. Beside DSRC/WAVE standards, 2/3G cellular networks can be used to enable message exchange among vehicles(Boukerche et al., 2008; Lee and Gerla, 2010)

In this study, we will use machine learning methods to analyze collected information from

vehicles to detect forward collisions. Drivers will be alerted about collision and they will have time to take precaution to avoid piled-up collision.

PROJECT OVERVIEW

- The high demand of automobiles has also increased the traffic hazards and the road accidents. Life of the people is under high risk. This is because of the lack of best emergency facilities available in our country
- This application provides the optimum solution to poor emergency facilities provided to the roads accidents in the most feasible waye of the lack of best emergency facilities available in our country
- The usage of auto mobiles has improved linearly over the past decade, which increased in the risk of human life. This is because due to the insufficient emergency facilities

PROJECT OBJECTIVES

- This project mainly focuses on the detection algorithm based on Accident detection apparent feature information, that is, detects and classifies the vehicle target in the actual traffic picture.

- Its main difficulty lies in the picture of the vehicle target will change due to lighting, angle of view and the interior of the vehicle.

ORGANIZATION OF THE CHAPTERS

This thesis is organized in the following chapters:

Chapter 1: Introduction

Accidents have been a major cause of deaths in India. More than 80% of accident-related deaths occur not due to the accident itself but the lack of timely help reaching the accident victims. In highways where the traffic is really light and fast-paced an accident victim could be left unattended for a long time. The intent is to create a system which would detect an accident based

on the live feed of video from a CCTV camera installed on a highway. The idea is to take each frame of a video and run it through a deep learning convolution neural network model which has been trained to classify frames of a video into accident or non-accident. Convolutional Neural Networks has proven to be a fast and accurate approach to classify images. CNN based image classifiers have given accuracy's of more than 95% for comparatively smaller datasets and require less preprocessing as compared to other image classifying algorithms.

Chapter 2: Literature Survey

This section discusses about various works related to the development of the automotive industry, the intelligent vehicle technology is being developed. As the development of the automotive technology, the proportion of electrical and electronic system will be expanded to more than 50%. Particularly, LDWS (Lane Departure Warning System) technique is actively developing to generate a warning signal if the lane is separated. In this paper, we apply the Hough transform with optimized the accumulator cells in the four ROI in parallel and detects lanes with highly efficient. The result of verification of an algorithm showed a recognition rate of 94.3% on average. And Hough transform recognized only linear but the detection was successful in the curve.

Chapter 3: Software and Hardware Requirements

To be used efficiently, all computer software needs certain hardware components or other software resources to be present on a computer. These prerequisites are known as (computer) system requirements and are often used as a guideline as opposed to an absolute rule. Most software defines two sets of system requirements: minimum and recommended. This section outlines minimum software and hardware requirements for deploying the project. Requirements may vary based on utilization and observing performance of pilot projects is recommended prior to scale out

Chapter 4: Software Development Analysis

This section contains development and implementation details of the design parameters. Developer's code based on the system specifications and requirements. Following company procedures and guidelines, front-end developers build interfaces and back-ends while database administrators create relevant data in the database. The programmers also test and review each other's code.

Chapter 5: Project System Design

Design is the stage of the software development process. Here, architects and developers draw up advanced technical specifications they need to create the software to requirements. Stakeholders will discuss factors such as risk levels, team composition, applicable technologies, time, budget, project limitations, method and architectural design.

Chapter 6: Project Coding

A programming project produces a well-designed executing system that solves a specified distributed programming problem. A project code is used to represent a one- time, or intermittent departmental event or activity. Any person can use a project code on a transaction, regardless of the project manager or home organization. This section describes some of the coding templates, outline of various files, class with functionalities, the various methods of input and output parameter.

Chapter 7: Project Testing

The testing phase checks the software for bugs and verifies its performance before delivery to users. In this stage, expert testers verify the product's functions to make sure it performs according to the requirements analysis document. Testers use exploratory testing if they have experience with that software or a test script to validate the performance of individual components of the software. They notify developers of defects in the code. If developers confirm the flaws are valid, they improve the program, and the testers repeat the process until the software is free of bugs and behaves according to requirements.

Chapter 8: Output screens

The output of the programmed project is being screened with the screenshots. This section will contain the screenshots of the execution at intermediate stages of the execution. In a nutshell it will contain all the interfaces and the final output screens of the project.

Chapter 9: Experimental results

This section will contain about the experimental results of our project.

CHAPTER 2
LITERATURE SURVEY

2.LITERATURE SURVEY

2.1.Lane recognition algorithm using the Hough transform with applied accumulator cells in multichannel ROI

AUTHORS: Cho, Jae-Hyun, Young-Min Jang, and Sang-Bock Cho.

With the development of the automotive industry, the intelligent vehicle technology is being developed. As the development of the automotive technology, the proportion of electrical and electronic system will be expanded to more than 50%. Particularly, LDWS (Lane Departure Warning System) technique is actively developing to generate a warning signal if the lane is separated. In this paper, we apply the Hough transform with optimized the accumulator cells in the four ROI in parallel and detects lanes with highly efficient. The result of verification of an algorithm showed a recognition rate of 94.3% on average. And Hough transform recognized only linear but the detection was successful in the curve.

Simple Robust Road Lane Detection Algorithm AUTHORS:Low,

Chan Yee, HairiZamzuri, and SaifulAmriMazlan

Lane detection plays an important role in intelligent vehicle systems. Therefore, this paper presents a robust road lane marker detection algorithm to detect the left and right lane markers. The algorithm consists of optimization of Canny edge detection and Hough Transform. The system captures images from a front viewing vision sensor placed facing the road behind the windscreen as input. Then a series of image processing is applied to generate the road model. Canny edge detection performs features recognition then followed by Hough Transform lane generation. The algorithm detects visible left and right lane markers on the road based on realtime video processing.

An adaptive road ROI determination algorithm for lane detection

AUTHORS: Ding, Dajun, Chanho Lee, and Kwang-yeob Lee

Road conditions can provide important information for driving safety in driving assistance system. The input images usually include unnecessary information and road conditions need to

be analyzed only in a region of interest (ROI) to reduce the amount of computation. In this paper, a vision-based road ROI determination algorithm is proposed to detect the road region using the positional information of a vanishing point and line segments. The line segments are detected using Hough Transform. The road ROI can be determined automatically and adaptively in every frame. The proposed method is applied to various video images from black boxes, and is verified to be robust.

Edge Detection of Color Road Image Based on Lab ModelAUTHORS:

Fani, Hongli, and Weihua Wang.

Color image edge detection algorithm has gradually become the hot spot. Due to the shortage of previous road image detection algorithms, by analyzing the characteristics of Lab model and the road image, a new algorithm for color road image edge detection was presented. The original color data in RGB color model were converted to Lab color model, and the difference information between the gray image from L channel and the red-green image was obtained with difference image method, and the threshold was got through difference information with the optimal threshold value algorithm, then the edge detection was carried out. Through experiment analysis and comparison, we can see that the algorithm in this paper has higher resistance to noise and retains better edges for color road image edge detection than the traditional algorithms.

ACCIDENT AVOIDING SYSTEM USING LANE DETECTION

AUTHORS: Nalla, Phaneendra, GCL AbhiramanGoud, and V. Padmaja

This Model based on machine vision is a human decision-make like solution to avoid lane departure fatalities with high reliability. The goal of this model titled â€œAccident avoider using edge detectionâ€ is to implement an image processing algorithm to detect lanes on the road and give a textual warning on departure from the lane. In this paper, the model of vision-based lane departure warning system and the realization is described at first. Then the method of lane detection is illustrated, which is composed of five steps: image preprocessing, binary processing and dynamical threshold choosing, lane detection and departure detection. After that, the solution of how to perform the departure decision-making is proposed and demonstrated. Simulink was used and tested for implementation of this model under graphical user interface unit.

CHAPTER 3

SOFTWARE AND HARDWARE REQUIREMENTS

3. SOFTWARE AND HARDWARE REQUIREMENTS

REQUIREMENT ANALYSIS

The project involved analyzing the design of few applications so as to make the application more users friendly. To do so, it was really important to keep the navigations from one screen to the other well ordered and at the same time reducing the amount of typing the user needs to do. In order to make the application more accessible, the browser version had to be chosen so that it is compatible with most of the Browsers.

REQUIREMENT SPECIFICATION

Functional Requirements

- Graphical User interface with the User.

Software Requirements

For developing the application the following are the Software Requirements:

1. Python
2. Django

Operating Systems supported

1. Windows 7
2. Windows XP
3. Windows 8

Technologies and Languages used to Develop

1.Python

Debugger and Emulator

▪Any Browser (Particularly Chrome)

Hardware Requirements

For developing the application the following are the Hardware Requirements:

- Processor: Pentium IV or higher
- RAM: 256 MB
- Space on Hard Disk: minimum 512MB

CHAPTER 4

SOFTWARE DEVELOPMENT ANALYSIS

4. SOFTWARE DEVELOPMENT ANALYASIS

FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are,

- **ECONOMICAL FEASIBILITY**
- **TECHNICAL FEASIBILITY**
- **SOCIAL FEASIBILITY**

ECONOMICAL FEASIBILITY

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

TECHNICAL FEASIBILITY

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

SOCIAL FEASIBILITY

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

CHAPTER 5

PROJECT SYSTEM DESIGN

5. PROJECT SYSTEM DESIGN

INPUT DESIGN

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the following things:

- What data should be given as input?
- How the data should be arranged or coded?
- The dialog to guide the operating personnel in providing input.
- Methods for preparing input validations and steps to follow when error occur.

OBJECTIVES

1. Input Design is the process of converting a user-oriented description of the input into a computer-based system. This design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerized system.

2. It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The goal of designing input is to make data entry easier and to be free from errors. The data entry screen is designed in such a way that all the data manipulates can be performed. It also provides record viewing facilities.

3. When the data is entered it will check for its validity. Data can be entered with the help of screens. Appropriate messages are provided as when needed so that the user will not be in maize of instant. Thus the objective of input design is to create an input layout that is easy to follow

OUTPUT DESIGN

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making.

1. Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is designed so that people will find the system can use easily and effectively. When analysis design computer output, they should Identify the specific output that is needed to meet the requirements.

2.Select methods for presenting information.

3.Create document, report, or other formats that contain information produced by the system.

The output form of an information system should accomplish one or more of the following objectives.

- Convey information about past activities, current status or projections of the
- Future.
- Signal important events, opportunities, problems, or warnings.
- Trigger an action.
- Confirm an action.

UML DIAGRAMS

UML is a modern approach to modelling and documenting software. In fact, it's one of the most popular business process modelling techniques. It is based on diagrammatic representations of software components. As the old proverb says: "a picture is worth a thousand words". By using visual representations, we are able to better understand possible flaws or errors in software or business processes. Mainly, UML has been used as a general-purpose modelling language in the

field of software engineering. However, it has now found its way into the documentation of several business processes or workflows. For example, activity diagrams, a type of UML diagram, can be used as a replacement for flowcharts. They provide both a more standardized way of modelling workflows as well as a wider range of features to improve readability and efficacy

There are two broad categories of diagrams and they are again divided into subcategories −

- Behavioral Diagrams
- Structural Diagrams

BEHAVIORAL DIAGRAMS

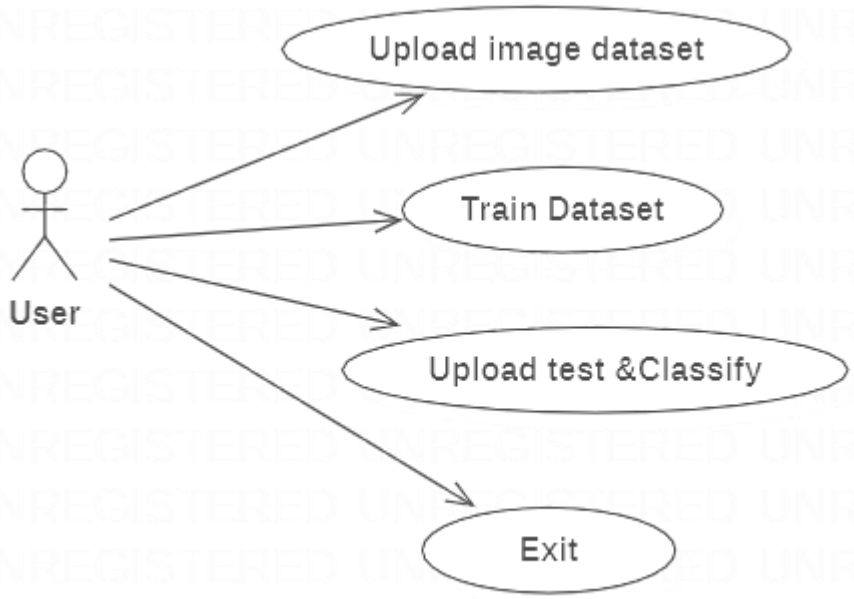
Any system can have two aspects, static and dynamic. So, a model is considered as complete when both the aspects are fully covered. Behavioural diagrams basically capture the dynamic aspect of a system. Dynamic aspect can be further described as the changing/moving parts of a system.

UML has the following five types of behavioral diagrams −

- Use case diagram
- Sequence diagram
- Collaboration diagram
- Statechart diagram
- Activity diagram

USE CASE DIAGRAM

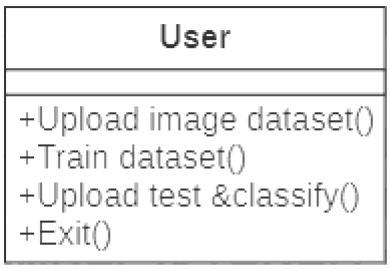
A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.



-

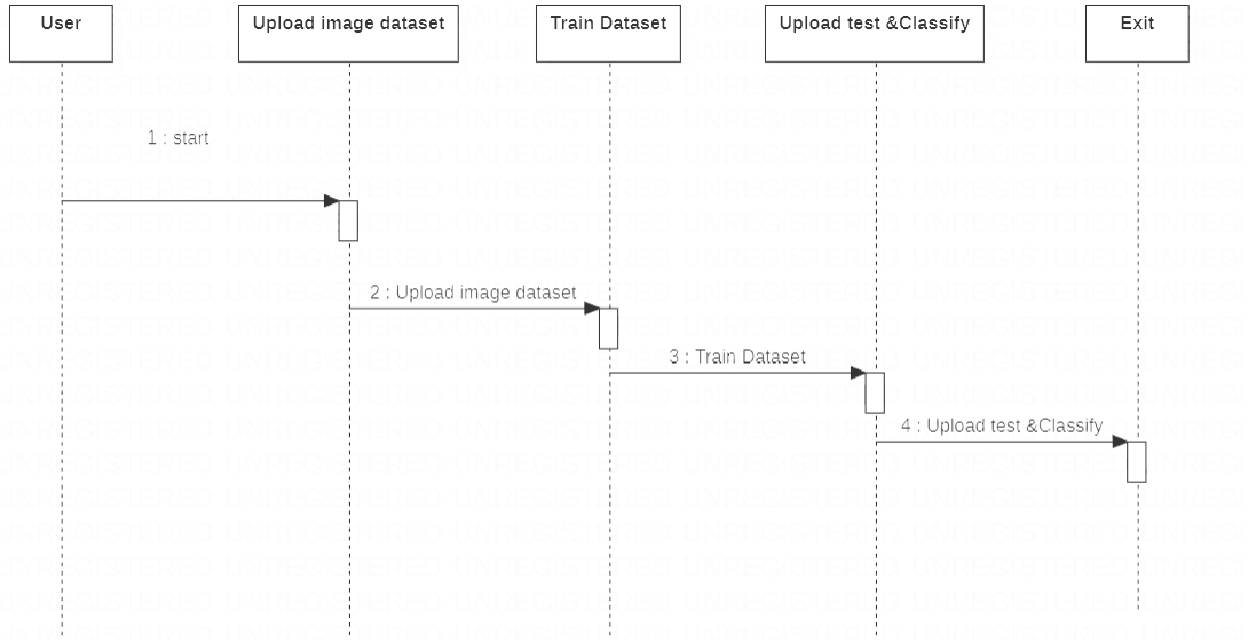
CLASS DIAGRAM

In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.



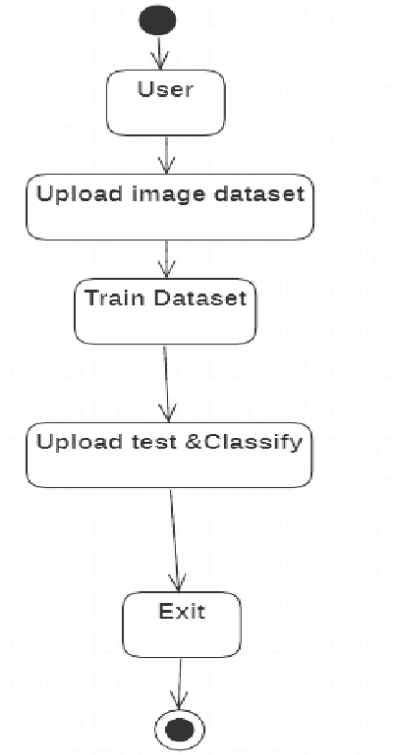
SEQUENCE DIAGRAM

A sequence diagram in Unified Modeling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.



ACTIVITY DIAGRAM:

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modeling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control



CHAPTER 6
PROJECT CODING

6. PROJECT CODING

CODING TEMPLATES

For implementing the system various libraries, modules and functions contained bythese libraries and modules have been imported as shown in the coding template below:

```
from tkinter import messagebox from tkinter import *
from tkinter import simpledialog import tkinter from
tkinter import filedialog from tkinter.filedialog import
askopenfilename import time import cv2 import
tensorflow.compat.v1 as tf from collections import
namedtuple from collections import defaultdict from
io import StringIO from PIL import Image import
numpy as np import winsound


tf.disable_v2_behavior() main =
tkinter.Tk() main.title("Accident
Detection")
main.geometry("1300x1200")


global filename global
detectionGraph global
msg


def loadModel():
    global detectionGraph
detectionGraph = tf.Graph()     with
detectionGraph.as_default():
od_graphDef = tf.GraphDef()         with
tf.gfile.GFile('model/frozen_inference_graph.pb', 'rb') as file:
```

```
serializedGraph = file.read()
od_graphDef.ParseFromString(serializedGraph)
tf.import_graph_def(od_graphDef, name='')

messagebox.showinfo("Training model loaded","Training model loaded")

def beep():
    frequency = 2500  # Set Frequency To 2500 Hertz     duration = 1000
# Set Duration To 1000 ms == 1 second winsound.Beep(frequency,
duration)

def uploadVideo():     global filename     filename =
filedialog.askopenfilename(initialdir="videos") pathlabel.config(text=filename) text.delete('1.0',
END) text.insert(END,filename+" loaded\n");

def calculateCollision(boxes,classes,scores,image_np):
    global msg
      #cv2.putText(image_np,                    "NORMAL!",                    (230,                    50),
cv2.FONT_HERSHEY_SIMPLEX, 1.0, (255, 255, 255), 2, cv2.LINE_AA)     for i, b in
enumerate(boxes[0]):         if classes[0][i] == 3 or classes[0][i] == 6 or classes[0][i] ==
8:             if scores[0][i] > 0.5:                 for j, c in enumerate(boxes[0]):
                  if (i != j) and (classes[0][j] == 3 or classes[0][j] == 6 or classes[0][j] == 8) and
scores[0][j]> 0.5:
                      Rectangle = namedtuple('Rectangle', 'xminyminxmaxymax') ra =
Rectangle(boxes[0][i][3], boxes[0][i][2], boxes[0][i][1], boxes[0][i][3]) rb =
Rectangle(boxes[0][j][3], boxes[0][j][2], boxes[0][j][1], boxes[0][j][3]) ar =
rectArea(boxes[0][i][3], boxes[0][i][1],boxes[0][i][2],boxes[0][i][3]) col_threshold =
0.6*np.sqrt(ar) area(ra, rb)                     if (area(ra,rb)<col_threshold) :
                          print('accident')
msg = 'ACCIDENT!' beep()
                          return True
else:
```

```
                return False
def rectArea(xmax, ymax, xmin, ymin):
    x = np.abs(xmax-xmin)     y =
np.abs(ymax-ymin) return x*y


def load_image_into_numpy_array(image):
  (im_width, im_height) = image.size
  return                np.array(image.getdata()).reshape((im_height,im_width,
3)).astype(np.uint8)

def area(a, b):  # returns None if rectangles don't intersect      dx =
min(a.xmax, b.xmax) - max(a.xmin, b.xmin) dy = min(a.ymax,
b.ymax) - max(a.ymin, b.ymin)      return dx*dy

def detector():     global msgmsg = "      cap
= cv2.VideoCapture(filename)     with
detectionGraph.as_default():
        with tf.Session(graph=detectionGraph) as sess:
            while True:
                ret, image_np = cap.read()
image_np_expanded = np.expand_dims(image_np, axis=0)
image_tensor = detectionGraph.get_tensor_by_name('image_tensor:0')
boxes = detectionGraph.get_tensor_by_name('detection_boxes:0')
scores = detectionGraph.get_tensor_by_name('detection_scores:0')
classes = detectionGraph.get_tensor_by_name('detection_classes:0')
try:
num_detections = detectionGraph.get_tensor_by_name('num_detections:0')
                (boxes, scores, classes, num_detections) = sess.run([boxes, scores, classes,
num_detections], feed_dict={image_tensor: image_np_expanded}) calculateCollision(boxes,
classes, scores, image_np)
                cv2.putText(image_np,   msg,    (230,    50), cv2.FONT_HERSHEY_SIMPLEX, 1.0,
(255, 0, 0), 2, cv2.LINE_AA)
```

```
                cv2.imshow('Accident Detection', image_np)
 except:
                cv2.destroyAllWindows()
                break
 if cv2.waitKey(25) & 0xFF == ord('q'):
                cv2.destroyAllWindows()
break

def exit():
main.destroy()
font = ('times', 16, 'bold') title = Label(main,
text='Accident Detection') title.config(bg='light cyan',
fg='pale violet red')  title.config(font=font)
title.config(height=3,  width=120)
title.place(x=0,y=5)
font1 = ('times', 13, 'bold')
uploadButton = Button(main, text="Load & Generate CNN Model",
command=loadModel) uploadButton.place(x=50,y=100) uploadButton.config(font=font1)

pathlabel = Label(main) pathlabel.config(bg='light cyan',
fg='pale violet red')  pathlabel.config(font=font1)
pathlabel.place(x=460,y=100)

webcamButton= Button(main, text="Browse System Videos",
command=uploadVideo) webcamButton.place(x=50,y=150) webcamButton.config(font=font1)

webcamButton= Button(main, text="Start Accident Detector",
command=detector) webcamButton.place(x=50,y=200) webcamButton.config(font=font1)

exitButton = Button(main, text="Exit", command=exit)
exitButton.place(x=330,y=250) exitButton.config(font=font1)
```

```
font1 = ('times', 12, 'bold') text=Text(main,height=20,width=150)
scroll=Scrollbar(text) text.configure(yscrollcommand=scroll.set)
text.place(x=10,y=250) text.config(font=font1)
main.config(bg='snow3') main.mainloop()
```

CHAPTER 7

PROJECT TESTING

7.PROJECT TESTING

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

TYPES OF TESTS

Unit testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

Integration testing

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

Functional test

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items:

Valid Input : identified classes of valid input must be accepted.

Invalid Input : identified classes of invalid input must be rejected.

Functions : identified functions must be exercised.

Output : identified classes of application outputs must be exercised.

Systems/Procedures : interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

System Test

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

White Box Testing

White Box Testing is a testing in which in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is purpose. It is used to test areas that cannot be reached from a black box level.

Black Box Testing

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box .you cannot “see” into it. The test provides inputs and responds to outputs without considering how the software works.

Unit Testing

Unit testing is usually conducted as part of a combined code and unit test phase of the software lifecycle, although it is not uncommon for coding and unit testing to be conducted as two distinct phases.

Test strategy and approach

Field testing will be performed manually and functional tests will be written in detail.

Test objectives

- All field entries must work properly.
- Pages must be activated from the identified link.
- The entry screen, messages and responses must not be delayed.

Features to be tested

- Verify that the entries are of the correct format
- No duplicate entries should be allowed
- All links should take the user to the correct page.

Integration Testing

Software integration testing is the incremental integration testing of two or more integrated software components on a single platform to produce failures caused by interface defects.

The task of the integration test is to check that components or software applications, e.g. components in a software system or – one step up – software applications at the company level – interact without error.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

Acceptance Testing

User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

CHAPTER 8

OUTPUT SCREENS

8. OUTPUT SCREENS

8.1 USER INTERFACES AND OUTPUT INTERFACE.

Analysing Dataset: This project is trained with images where vehicles collided and accident occurred and in test video if anything such collision happens between vehicles then application detect as accident. Training is done with tensor flow and CNN Algorithm. To run project double click on run.bat file to get below screen

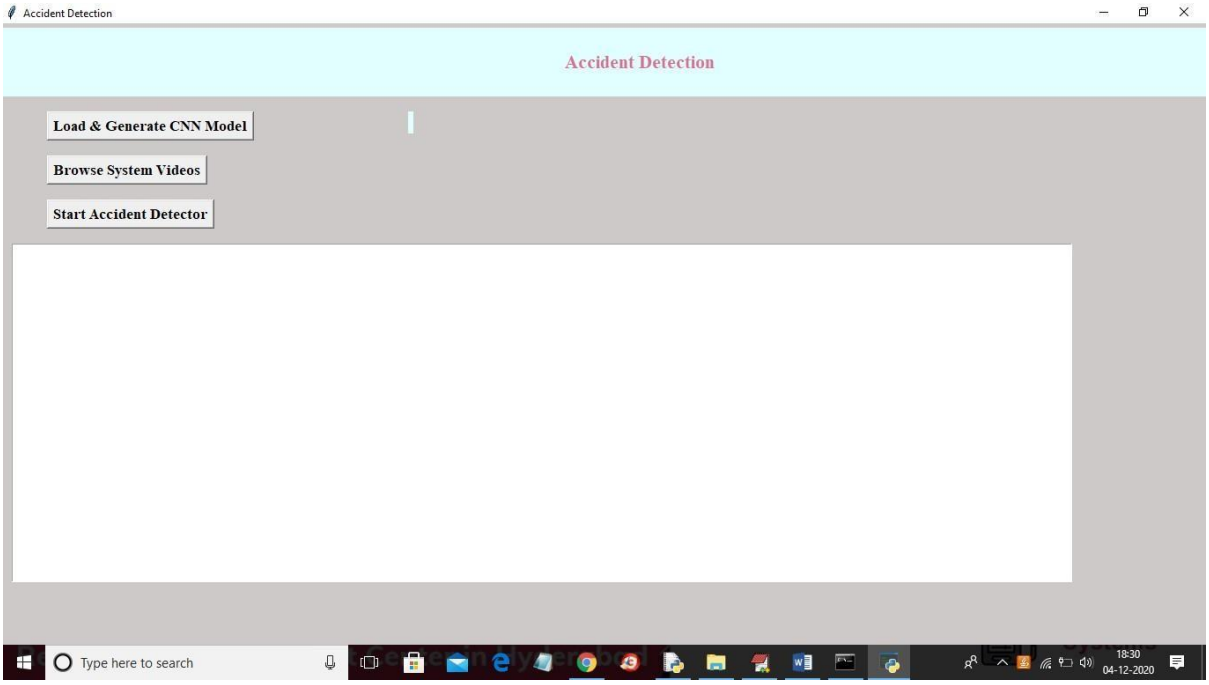


Figure-1 : Main Interface

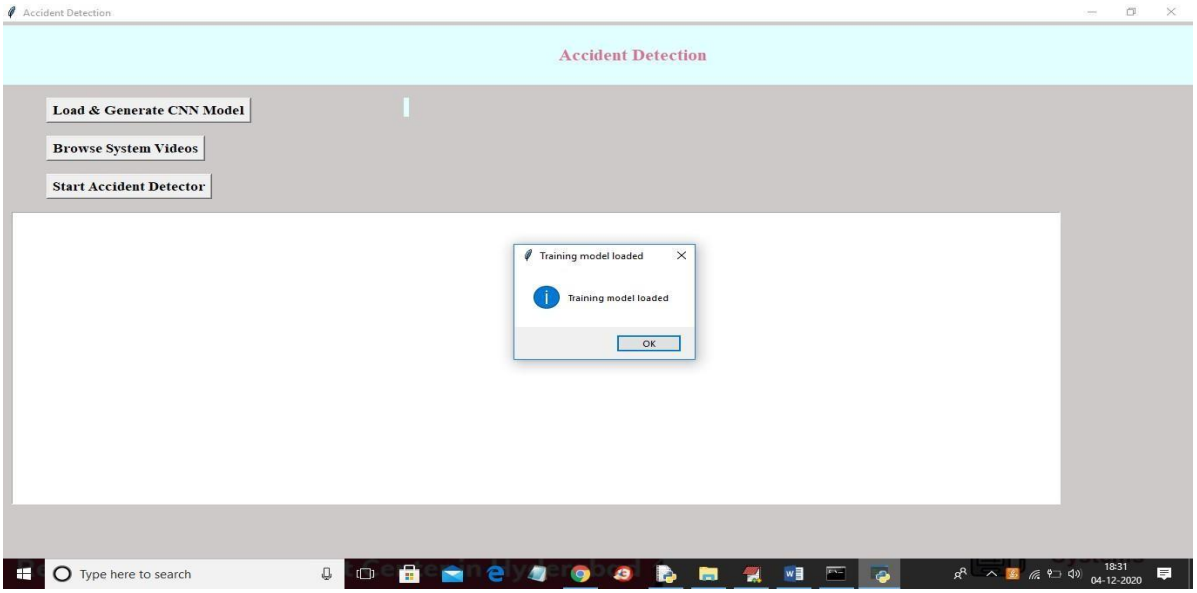


Figure-2: Uploading of Data set

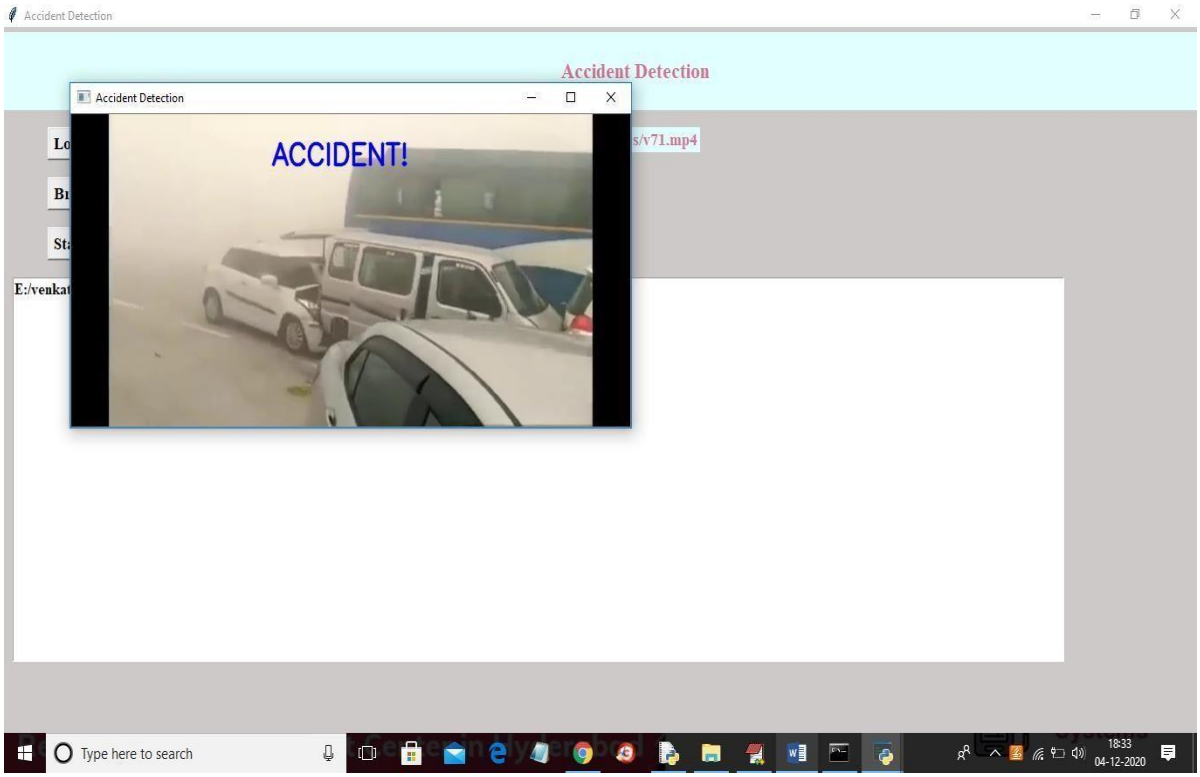


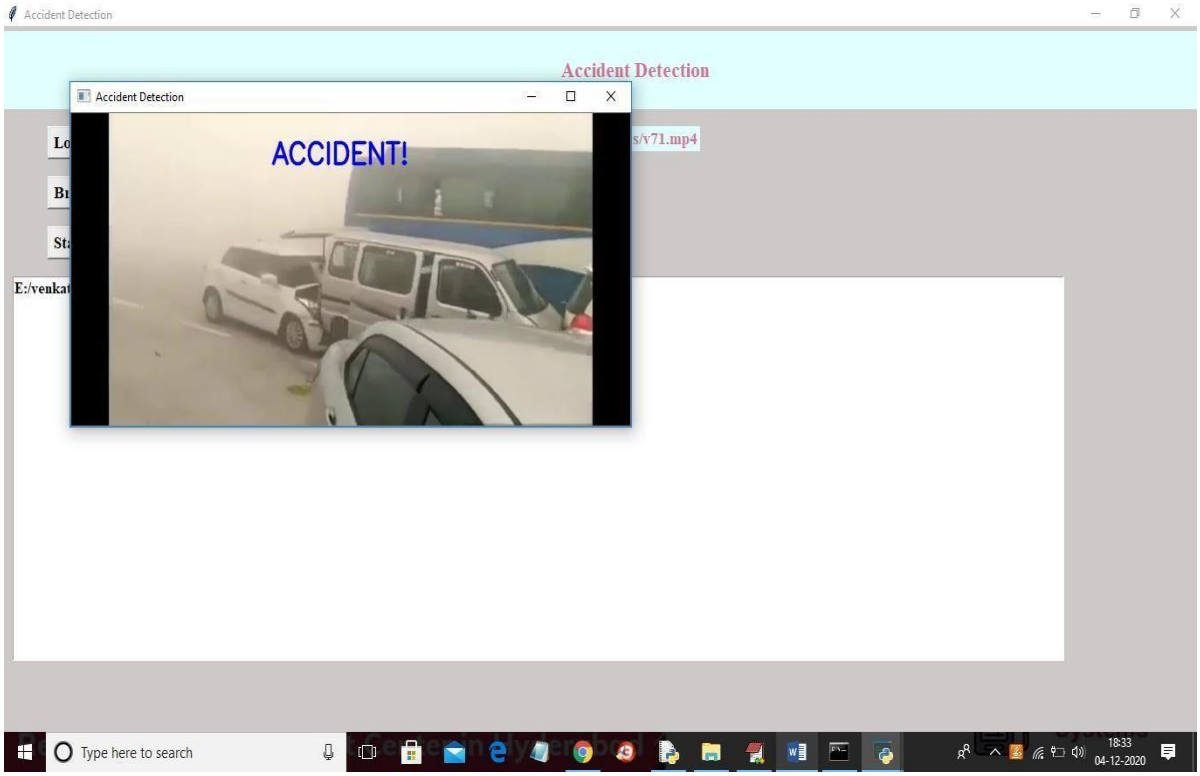
Figure-3: Accident Detection

In above screen video start playing and upon accident detection will get the screen with beep sound

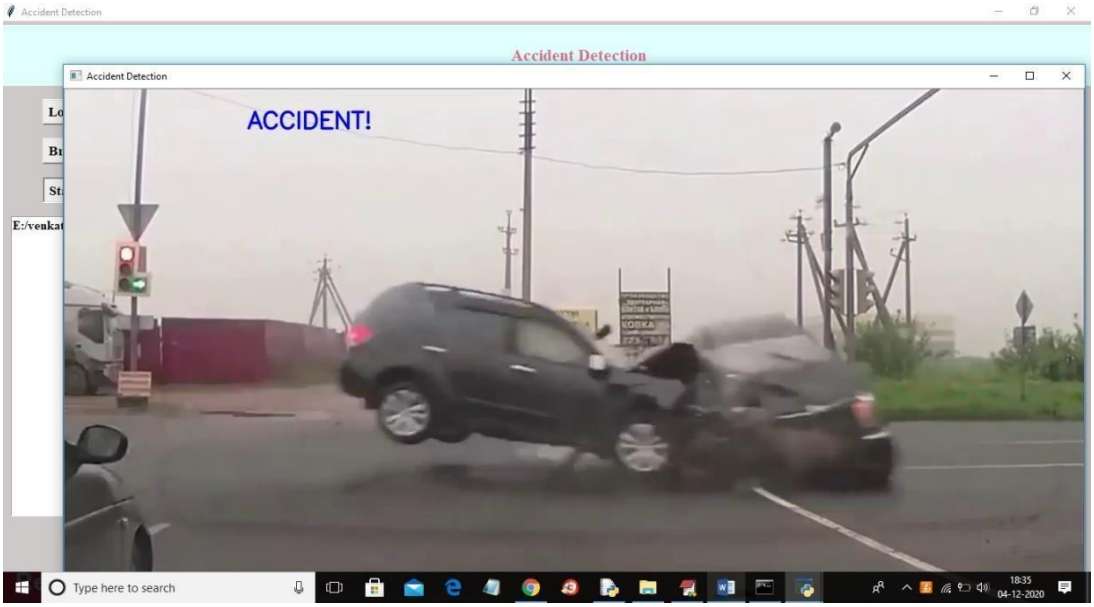
CHAPTER 9

EXPERIMENTAL RESULTS

9.EXPERIMENTAL RESULTS



In above screen video start playing and upon accident detection will get the screen with beep sound



In above screen upon collision then accident display message will appear with beep sound

CONCLUSION AND FUTURE ENHANCEMENT

10. CONCLUSION AND FUTURE ENHANCEMENT

The system can detect the accident and confirms the seriousness of the accident and then alert medical assist center to provide emergency medical aid to accident victim.

1. Accelerometer and heartbeat sensor are used to determine whether an accident had occurred.
2. The smart phone with the android app will send message to the nearest medical center
3. The system will also inform the friends and family of the victim through message
4. A buzzer is also provided to alert the fellow passengers on the road that an accident has occurred to invite their help
5. Accident detection and alert system are highly relevant in these days and this project aims at developing a low cost solution for the same for the benefit of the society.

REFERENCES

[1] Angshuman, G., 2004. An Incident Detection Algorithm Based On a Discrete State Propagation Model of Traffic Flow.

[2] Bar-Gera, H., 2007. Evaluation of a cellular phone-based system for measurements of traffic speeds and travel times: A case study from Israel. Transportation Research Part C: Emerging Technologies 15, 380 - 391.

[3] Boukerche, A., Oliveira, H., Nakamura, E., Loureiro, A., 2008. Vehicular Ad Hoc Networks: A New Challenge for Localization-Based Systems. Computer Communications 31, 2838– 2849

[4] Caffery, J.J., Stuber, G.L., 1998. Overview of radiolocation in CDMA cellular systems. IEEE Communications Magazine 36, 38-45.

[5] Chen, M.Y., Sohn, T., Chmelev, D., Haehnel, D., Hightower, J., Hughes, J., LaMarca, A., Potter, F., Smith, I., Varshavsky, A., 2006. Practical Metropolitan-Scale Positioning for GSM Phones, in: Dourish, P., Friday, A. (Eds.), UbiComp 2006: Ubiquitous Computing. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 225-242.

[6] Edmond Chin-Ping Chang, 1992. A Neural Network Approach to Freeway Incident Detection. IEEE, pp. 641-647

[7] Edmond Chin-Ping Chang, KunhuangHuarng, 1993. Fuzzy Set Applications For Freeway Incident Detection. IEEE, pp. 439-443.

[8] Faouzi, N.E.E., Leung, H., Kurian, A., 2011. Data fusion in intelligent transportation systems: Progress and challenges - A survey. Inf. Fusion 12, 4–10.

[9] Han-Lee Song, 1994. Automatic vehicle location in cellular communications systems. IEEE Transactions on Vehicular Technology 43, 902-908.

[10] Krakiwsky, E.J., Harris, C.B., Wong, R.V.., 1988. A Kalman filter for integrating dead reckoning, map matching and GPS positioning, in: , IEEE Position Location and Navigation Symposium, 1988. Record. Navigation into the 21st Century. IEEE PLANS '88. Presented at the , IEEE Position Location and Navigation Symposium, 1988. Record. Navigation into the 21st Century. IEEE PLANS '88, IEEE, pp. 39-46

PUBLICATION

- “Innovation in computer networks, Computational Intelligence and IoT”[ICICCI-21].

- Paper id:0139

Suyash Singh is currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. He completed his intermediate from Krishna Public School and 10th class from Krishna Public School. His Technical Skills include C,Python and Java. he also has a basic understanding of C++. He had also doneParticipation in National Level Three Day Online Workshop on "AI &ML External skills are: He is a. He also completed some certifications in Coursera and Cursa.

1	Certificate in Programming in PHP
2	Certificate in HTML
3	Certificate in JavaScript
4	Certificate in Web Development
5	Certificate in CSS
6	Certificate in 5-Day Online International Hands-On Certification Training in " Python Programming"

Azeem Pasha is currently pursuing her Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. He completed his intermediate from Sri Chaitanya Junior college and 10th class from Bright concept School. His technical skills include C, Python and Java. he also has a basic understanding of C++. he took part in Employability Skill development Program conducted by Zensar.His participations include: National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020, "Know More - Teach More ", the Global Webinar on Cloud & Big Data – Changing the way we work which was conducted Global Education& Careers Forum (GECF) on 12th August , "One Day Webinar on Internet of Things and Its Applications" conducted by Anand Institute of Higher Technology on 21st May 2020 . April to 22nd May 2020. His areas of interest are Python, Artificial Intelligence, Machine Learning and Deep Learning. he completed few certification courses from online platforms like Coursera, CursaApp and SoloLearn.

1	Certificate in MySQL database by ThenewBoston
2	Certificate in Web Development for beginners by learn code.academy
3	Certificate in Word by GCFLearnFree
4	Certificate in Managing Project Risks and Changes
5	Certificate in AWS Fundamentals: Going Cloud Native
6	Certificate in Data Science Math Skills

Begari Vikas is currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. He completed his intermediate from VignanVidyalayam Junior College and 10th class from Vignan Global Gen School. His Technical Skills include C,Python and Java. he also has a basic understanding of C++. He had also doneParticipation in Enterpreneurship ESUMMIT certified at MLRIT College and Machine Learning Workshop.Some External skills are: He isa college ambassador in youth marketing company called Grapevine and also he participated in district wise Athletics and Volley ballcompetetions. He also completed some certifications in Coursera and Cursa.

1	Certificate in MySQL database by ThenewBoston
2	Certificate in Web Development for beginners by learn code.academy
3	Certificate in Word by GCFLearnFree
4	Certificate in Managing Project Risks and Changes
5	Certificate in AWS Fundamentals: Going Cloud Native
6	Certificate in Data Science Math Skills

ThotaSaiTejais currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. He completed his intermediate from Sri Chaitanya Junior College and 10^{th} class from Brilliant Grammer High School. His technical skills include C,Python and Java. he also has a basic understanding of C++. He is also a Professional Footballer. His participations include: All India Football Sub Junior Nationals 2015 represented behalf of Telangana, All India Football Junior Nationals 2017 represened behalf of Telangana, I-League U18 represented behalf of Fateh Hyderabad AFC, All India University South Zone Football Tournament 2019 represented behalf of Jawaharlal Technical University Hyderabad(JNTUH), AWS Fundamentals: Going Cloud-Native an online non-credit course authorized by Amazon Web Services and offered through Courser on 07/02/2020, Leadership And Emotional Intelligence an online non-credit course authorzed by Indian School of Business and offered through Coursera, Java Script By Net Ninja on 05/05/2021, Programming With PHP for beginners by Net Ninja, Managing Project Risks and Changes an online non-credit course authorized bu university of California, Irivine and offered through Coursera. His areas of interest are Python, Artificial Intelligence, Machine Learning and Deep Learning. He completed few certification courses from online platforms like Coursera, CursaApp and Net Ninja..

A

PROJECT REPORT

On

CONTAMINENT ZONE ALERTING APPLICATION

Submitted by

1)Mr. D.Sampath Rao (Regd.No 17K81A05K4)

2)Ms. P.Anil Kumar (Regd.No17K81A05M2)

3)Ms. JL.Neha (Regd.No 17K81A05L3)

4)Mr. G.Vijay Chandu (Regd.No 17K81A05L1)

in partial fulfillment for the of award

the degree of

BACHELOR OF TECHNOLOGY

IN

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Under the Guidance of

Dr. T. Poongothai

B.E., M.E., Ph.D.,

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ST. MARTIN'S ENGINEERING COLLEGE
An Autonomous Institute

Dhulapally, Secunderabad – 500 100

JUNE 2021

BONAFIDE CERTIFICATE

This is to certify that the project entitled **CONTAMINENT ZONE ALERTING APPLICATION**, is being submitted by 1.**Mr. D.Sampath Rao (Regd.No 17K81A05K4), 2.Mr. P.Anil Kumar Reddy (Regd.No, 17K81A05M2), 3.Ms. JL.Neha (Regd.No 17K81A05L3), 4.Mr. G.Vijay Chandu (Regd.No 17K81A05L1**) in partial fulfillment of the requirement for the award of the degree of **BACHELOR OF TECHNOLOGY IN COMPUTER SCIENCE AND ENGINEERING** is recorded of Bonafide work carried out by them. The result embodied in this report have been verified and found satisfactory.

Dr.T.POONGOTHAI
Department of CSE

Head of the Department
Dr.M.NARAYANAN
Department of CSE

Internal Examiner

External Examiner

Place:

Date:

DECLARATION

We, the student of **Bachelor of Technology** in Department of Computer Science and Engineering', session: <2017 – 2021>, St. Martin's Engineering College, Dhulapally, Kompally, Secunderabad, hereby declare that work presented in this Project Work entitled **Contaminent Zone Alerting Application** is the outcome of our own Bonafide work and is correct to the best of our knowledge and this work has been undertaken taking care of Engineering Ethics. This result embodied in this project report has not been submitted in any university for award of any degree.

Mr. D. Sampath Rao (17K81A05K4)
Mr. P. Anil Kumar Reddy (17K81A05M2)
Ms. JL. Neha (17K81A05L3)
Mr. G. Vijay Chandu (17K81A05L1)

ABSTRACT

The World Health Organization has declared the outbreak of the novel coronavirus, Covid-19 as pandemic across the world. With its alarming surge of affected cases throughout the world, lockdown, and awareness (social distancing, use of masks etc.) among people are found to be the only means for restricting the community transmission. In a densely populated country like India, it is very difficult to prevent the community transmission even during lockdown without social awareness and precautionary measures taken by the people. Recently, several containment zones had been identified throughout the country and divided into red, orange and green zones, respectively. The red zones indicate the infection hotspots, orange zones denote some infection and green zones indicate an area with no infection. This project mainly focuses on development of an Android application which can inform people of the Covid-19 containment zones and prevent trespassing into these zones. This Android application updates the locations of the areas in a Google map which are identified to be the containment zones. The application also notifies the users if they have entered a containment zone and uploads the user's IMEI number to the online database. To achieve all these functionalities, many tools, and APIs from Google like Geofencing API are used in this application. Therefore, this application can be used as a tool for creating further social awareness about the arising need of precautionary measures to be taken by the people of India.

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible and whose encouragements and guidance have crowded effects with success.

We extended our deep sense of gratitude to Principal**,** **Dr. P. SANTOSH KUMAR PATRA**, St. Martin's Engineering College, Dhulapally, for permitting us to undertake this project.

We are also thankful to **Dr. M. NARAYANAN**, Head of the Department, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

We would like to thank **Dr. T. POONGOTHAI,** Department of Computer Science and Engineering (AI & ML) for her encouragement and insightful comments and as well as our project coordinators **Dr. B. RAJALINGAM**, Associate Professor and **Dr. R. SANTOSH KUMAR**, Associate Professor, in Department of Computer Science and Engineering for their valuable support.

We would like to express our sincere gratitude and indebtedness to our project supervisor **Dr. T. POONGOTHAI**, Professor, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

Finally, we express thanks to all those who have helped us successfully to completing this project. Furthermore, we would like to thank our family and friends for their moral support and encouragement.

We express thanks to all those who have helped us in successfully completing the project.

Mr. D. Sampath Rao (17K81A05K4)

Mr. P. Anil Kumar Reddy (17K81A05M2)

Ms. JL. Neha (17K81A05L3)

Mr. G. Vijay Chandu (17K81A05L1)

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

API	Application User Interface
GPS	Global Positioning System
SOAP	Simple Object Access Protocol
IMEI	International Mobile Equipment Identity
FTP	File Transfer Protocol
HTTP	Hyper Text Transfer Protocol
WHO	World Health Organization
JSON	JavaScript Object Notation

1. INTRODUCTION

Currently there are several research works undergoing in the country to prevent Covid-19 cases from rising. Previously our country was importing medical kits like PPE (Personal Protection Kits), mask from outside, but now it has been successful in developing these kits. Along with taking initiatives to Figure this disease, our country has also taken steps to make people aware of the disease. The news and media have a great part in creating this awareness by informing the public about the preventive measures that can keep them away from infection. Awareness among the people to carry out all the preventive measures can immensely help to reduce spread of the virus. The country has created containment zones throughout the cities wherever Covid-19 cases have been reported to prevent further spread of the virus. These containment zones have been kept isolated from the outside public to ensure no contamination occurs outside. After more than 2 months of the lockdown, the government has relaxed some of the lockdown rules and has permitted reopening of government offices, bus and other road transportation facilities and shopping markets. People can move inside the city for work and other purposes. But the containment zones are still being kept isolated, and new containment zones are being formed wherever Covid-19 cases have been reported. These zones are highly contagious as droplets with virus coughed out from an unscreened asymptomatic patient can travel up to 8 m. Though these containment zones are guarded by policemen, still there remains a chance that people might unknowingly step into them. In this situation where people can move in the city, these containment zones pose a risk of infection to these city dwellers.

1.1 PROJECT OVERVIEW

We focus on developing a mobile based application to provide information regarding the Covid-19 containment zones in West Bengal. The application further tracks the user's location and provides notification alert if the user has entered a containment zone. The application also provides daily Covid-19 case statistics to the users to keep them updated. The application is developed on Android SDK and uses to store the location data. Android's geofencing client is used to create geofences around the containment zones and notification manager is used to provide notifications. The application also uses RESTful web services to show the Covid-19 cases in West Bengal. We have tested our application with different users in different locations across West Bengal and it works efficiently and is able to attain our target. These containment zones have been kept isolated from the outside public to ensure no contamination occurs outside. After more than 2 months of the lockdown, the government has relaxed some of the lockdown rules and has permitted reopening of government offices, bus and other road transportation facilities and shopping markets. People can move inside the city for work and other purposes. But the containment zones are still being kept isolated, and new containment zones are being formed wherever Covid-19 cases have been reported. These zones are highly contagious as droplets with virus coughed out from an unscreened asymptomatic patient can travel up to 8 m.

1.2 PROJECT OBJECTIVES

The main objective of the project is to deliver an android alerting application. The application works in such a way to help user alerting before entering a containment zone. The system is developed to meet the necessity of urgency travel in pandemic situation. The object is aimed at safety of the people from covid-19 or other out breaks. The Android application shows the location of the containment zones to the users. It also notifies the user when he or she trespasses the boundary of a containment zone or stays in the containment zones. The application also notifies the users if they have entered a containment zone and uploads the user's IMEI number to the online database. To achieve all these functionalities, many tools, and APIs from Google like Geo-fencing API are used in this application. Tests have been carried out in various containment zones across Hyderabad for the validation of the Android application. The chosen containment zones for the testing of the application were visited one by one. It is highlighted that the application sends notification alerts within 5–8 seconds on entering. The application can be further used for many purposes like locating all the containment zones in the city.

1.3 ORGANIZATION OF CHAPTERS

1.INTRODUCTION: Currently there are several research works undergoing in the country to prevent Covid-19 cases from rising. Previously our country was importing medical kits like PPE (Personal Protection Kits), mask from outside, but now it has been successful in developing these kits. Along with taking initiatives to Figure this disease, our country has also taken steps to make people aware of the disease.

2.LITERATURE SURVEY: The massive outbreak of the COVID-19 has prompted various scientists, researchers, laboratories, and organizations around the world to conduct large scale research to help develop vaccines and other treatment strategies.

3.SOFTWARE AND HARDWARE REQUIREMENTS: System requirements are the required specifications a device must have in order to use certain hardware or software. For example, a computer may require a specific I/O port to work with a peripheral_device. A smartphone may need a specific operating system to run a particular app.

4.SOFTWARE DEVELOPMENT ANALYASIS: Geofencing is a location-based service in which an app or other software uses GPS, RFID, Wi-Fi or cellular data to trigger a pre-programmed action when a mobile device or RFID tag enters or exits a virtual boundary set up around a geographical location, known as a geofence.

5.PROJECT SYSTEM DESIGN: Systems design is the process of defining the architecture, product design, modules, interfaces, and data for a system to satisfy specified requirements. Systems design could be seen as the application of systems theory to product development.

6.PROJECT CODING: Coding projects are one of the many ways you can learn the coding skills you need. Some people think that only experienced programmers can complete a project. In fact, there are projects available for all skill levels across many different programming genres.

7.PROJECT TESTING: Project Testing Phase means a group of activities designated for investigating and examining progress of a given project to provide stakeholders with information about actual levels of performance and quality of the project. ... Planning – the team makes a plan of key procedures and steps of testing.

8.OUTPUT SCREENS: Project Output is the final measurable result received upon successful completion of a project when all planned tasks and activities are accomplished and project deliverables are produced. ... Output of a project is received through a complex of activities that define the project lifecycle.

9.EXPERIMENTAL RESULTS: The results are simply the end of the scientific experiment: What you found in your study.

10.CONCLUSION AND FUTURE ENHANCEMENT: The Conclusions section sums up the key points of your discussion, the essential features of your design, or the significant outcomes of your investigation. As its function is to round off the story of your project, it should: be written to relate directly to the aims of the project as stated in the Introduction.

CHAPTER 2

2. LITERATURE SURVEY

2.1 SURVEY ON BACKGROUND

Introduction

The COVID-19, an acronym for ‘‘Coronavirus Disease2019’’, is a respiratory illness caused by the severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2), a contagious virus belonging to a family of single-stranded, positive-sense RNA viruses known as Coronaviridae. Much like the influenza virus, SARS-CoV-2 attacks the respiratory system and causes ailments such as cough, fever, fatigue, and breathlessness. While the exact source of the virus is unknown, scientists have mapped the genome sequence of theSARS-CoV-2 and determined it to be a member of the β-CoV genera of the coronavirus family, which typically derives its gene sources from bats and rodents. The COVID-19 was first reported to affect human life in Wuhan City, in the Hubei province of China in December 2019. Since then, the COVID-19 has spread like wildfire throughout the rest of the world, marking its presence in 213 countries and independent territories. COVID-19 statistics for the worst affected countries and regions of the world have been presented. According to the WHO, the current global tally1 of confirmed coronavirus cases stands at 3,090,445 while the death tolahs reached 217,769. The rapid rise in the number of COVID-19 incidents worldwide has prompted the need for immediate countermeasures to curb the catastrophic effects of the COVID-19 outbreak. To this end, this paper evaluates the use of varied technologies such as IoT, UAVs, AI, blockchain, and 5G, which could help mitigate the adverse effects of this pandemic and expedite the recovery process. However, before exploring the potential technological solutions for COVID-19 pandemic impact management, we provide a comprehensive review of the COVID-19, including its clinical features, diagnosis, treatment, and the impact of its outbreak on the global economy.

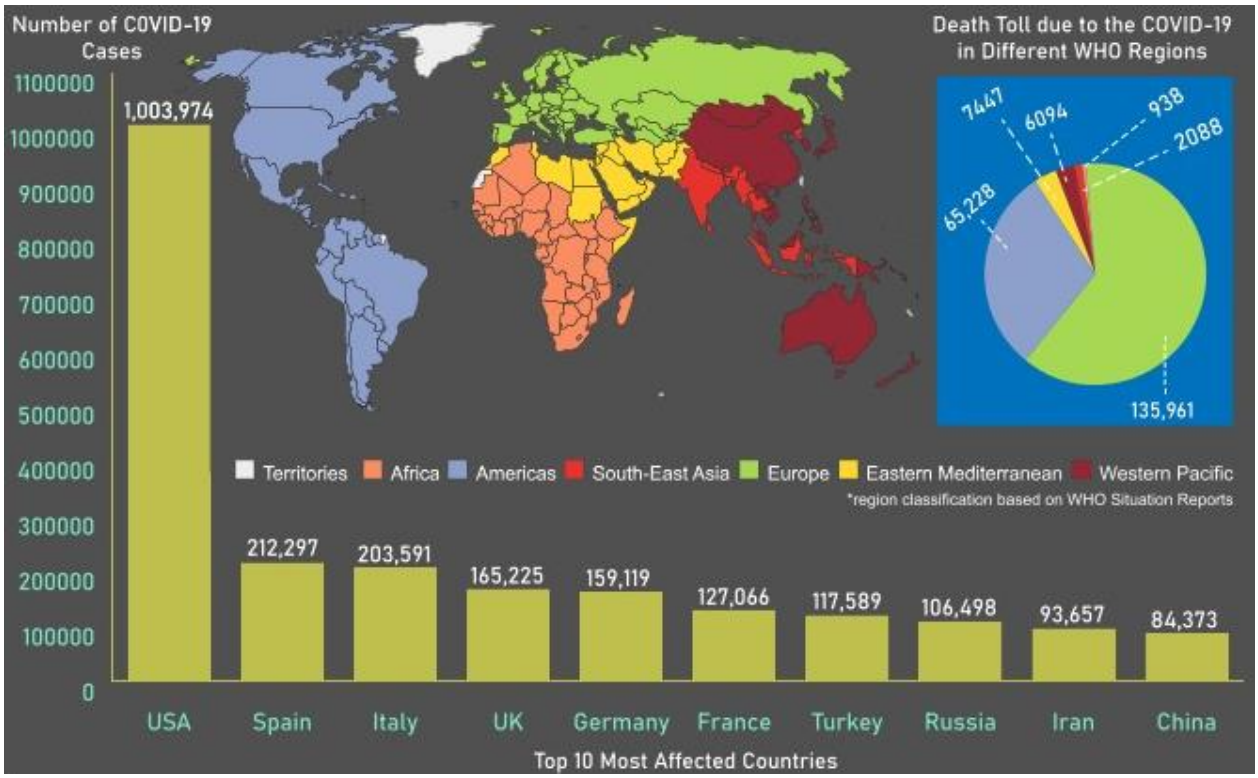


Figure 2.1: Top 10 Affected Countries

Background

According to the WHO, viral infections, particularly the ones caused by different coronaviruses, continue to emerge and pose a severe public health problem. Coronaviruses are spherical positive-sense RNA viruses ranging from 600Å - 1400Å in diameter, with proteins known as spikes protruding from its surface, which impart a crown-like structure to them under the electron microscope. The past two decades has witnessed the emergence of several viral outbreaks with different forms of coronavirus at the helm, such as the 2002-2004 SARS-CoV outbreak, and the more recent middle east respiratory syndrome coronavirus (MERS-CoV) infection of 2012. The SARS-CoV outbreak originated in the Guangdong province of China and later spread to more than 37 countries worldwide, causing over 8000 infections and around 774 deaths. The first case of MERS-CoV infection was detected in Saudi Arabia, which initiated a large-scale outbreak in the middle eastern countries that ultimately led to 871 fatalities. The COVID-19 outbreak came to light on 31 December 2019 when 27 cases of pneumonia of unknown Ethology were reported at the WHO's country office in China. The epicentre of the outbreak was linked to Wuhan's wholesale market for seafood and other exotic animals, including snakes, bats, and marmots.

A new strain of a highly contagious β-coronavirus, SARS-CoV-2, has been deemed responsible for the rapid outbreak of COVID-19. Distinguishing characteristics of the virus include its extremely contagious nature and relatively long (1-14 days) incubation period. During this period, a person can be infected by the virus and not show any symptoms at all. Therefore, people infected with the disease may unknowingly serve as silent carriers of the virus, contributing to a high basic reproductive number2 for the COVID-19 virus. While some studies indicate that SARS-CoV-2 could be susceptible to heat and ultraviolet (UV) light, there is no specific treatment or vaccine for the infection to date, and the management protocols for the disease are evolving as of this writing.

Figure 2.2: Incidents Reported Over a Year

Clinical Features

COVID-19 manifests with clinical features ranging from the asymptomatic state (no symptoms) to acute respiratory distress syndrome (ARDS) and multiple organ dysfunction syndrome (MODS). According to the results of a recent study conducted by the WHO in collaboration with China, of the 55,924 laboratory-confirmed COVID-19 cases that were examined, a majority exhibited clinical characteristics such as fever, dry cough, fatigue, and sputum production. At the same time, only a handful of patients showcased symptoms such as sore throat, headache, myalgia, and breathlessness, while symptoms such as nausea, nasal congestion, haemoptysis, diarrhoea, and conjunctival congestion were found to be very rare. While most of the COVID-19 patients developed a mild to moderate disease, a few patients were diagnosed with a severe (13.8%) and a critical (6.1%) form of the same. Patients with a severe or a critical form of the disease often develop bluish lips/face and are prone to a

variety of complications, including ARDS, acute heart injury, and secondary infection. According to the US Centres for Disease Control and Prevention (CDC), the individuals at the highest risk for severe illness from the COVID-19 include older adults (people above the age of 60) and people with existing medical conditions, such as diabetes, hypertension, asthma, and cardiovascular disease.

Related Works

The massive outbreak of the COVID-19 has prompted various scientists, researchers, laboratories, and organizations around the world to conduct large scale research to help develop vaccines and other treatment strategies. In the months following the COVID-19 outbreak, several papers examining different aspects of the COVID-19 have been published. To determine the clinical characteristics of the COVID-19, Dawe Wang et al. have studied 138 infected patients in Wuhan, China. The authors have taken into account specifics such as demographics, signs & symptoms, and medical history of all the patients to assess their cases carefully. The authors have also presented the laboratory findings of these patients to demonstrate the effects of the SARS-CoV-2 virus on different vital organs of the body. Nanshan Chen et al. studied 99 patients with the COVID-19, 49 of whom had a direct link to the Huanan seafood market in Wuhan, known to be the COVID-19 epicentre. Their findings of the epidemiological, clinical, and radiological characteristics of the disease have been published. In their findings, they report that among all the patients that were studied, 17% developed acute respiratory distress syndrome (ARDS), and among them, 11% died of multiple organ dysfunction syndrome (MODS).

Disease	Causative Agent	Year(s)	Death Toll	Classificiation
Spanish Flu	H1N1	1918-1919	~50 million	Pandemic
Asian Flu	H2N2	1957-1958	~1.1 million	Pandemic
Hong Kong Flu	H3N2	1968-1969	~1 million	Pandemic
SARS	SARS-CoV	2002-2004	774	Outbreak
Swine Flu	H1N1 (new strain)	2009-2010	~151,700 to 575,400	Pandemic
MERS	MERS-CoV	2012-present	871*	Outbreak
Asian Lineage Avian Influenza	H7N9	2013-2017	~605	Epidemic
Ebola Virus Disease (EVD)	Zaire ebolavirus	2014-2016	11,325	Epidemic
COVID-19	SARS-CoV-2	2019-present	217,769*	Pandemic

Table: 2.1 Spread of Various Diseases over 100 Years

Fang Jiang et al. have reviewed six published studies recognizing the clinical characteristics of the COVID-19. In their work, they have summarized these studies and, in doing so, provided a brief overview of clinical features and treatments of the COVID-19. The authors have reviewed the existing literature on computed tomography (CT) characteristics of

COVID-19 available on platforms such as PubMed, Google Scholar, and Elsevier, among others. The primary issue with both these works is that they review a small subset of a much broader subject. To this end, the authors provide a brief overview of the COVID-19 outbreak in terms of its clinical features, prevention, diagnosis, and treatment. Although these surveys shed some light on the current scenario of COVID-19 outbreak, they give a very brief and limited idea about the exact situation.

Different Stages of Covid-19 Outbreak

According to the WHO, the COVID-19 pandemic is regarded as having four main classes of transmission that remain consistent throughout the world to facilitate better communication and understanding amongst the countries. Such a categorization makes it simpler for other countries to enforce policies which they think would assist in preventing the outbreak, for example, imposing travel bans, shutting down schools & colleges, and enforcing partial or complete lockdown. For better understanding, we have portrayed the WHO transmission classes as different stages of the COVID19 outbreak keeping in line with several media reports. The onset of different stages of the COVID-19 outbreak in four countries, namely, China, Spain, Italy, and the USA.

A. STAGE I - IMPORTED CASES ONLY: The first stage of the COVID-19 outbreak in a particular nation is characterized by its first reported incident of the disease, in this case, COVID-19. In this stage, the disease does not spread locally, and the infection is usually limited to the people with travel history to an already affected region.

B. STAGE II - SPORADIC CASES/LOCAL TRANSMISSION The second stage of the COVID-19 outbreak occurs when there are a few sporadic cases of the disease in the country. It happens when people who are already infected with the disease spread it to people with whom they come into contact, usually immediate family members, friends, and colleagues. At this stage, it is possible to perform contact tracing and limit the spread of the disease by quarantining the infected people.

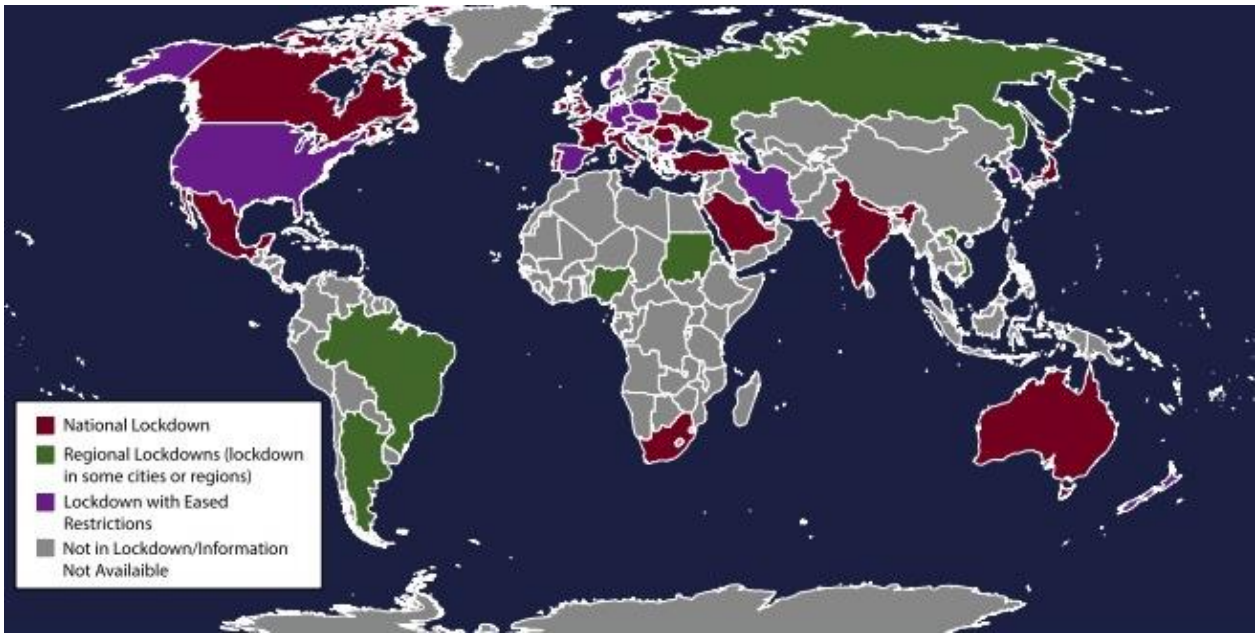


Figure 2.3 Countries in lockdown as of 19 April 2020 (data source: media reports)

C. STAGE III - CLUSTERS OF CASES The third stage of the COVID-19 outbreak in a country is marked by the presence of several clusters of COVID-19 cases, i.e., when the disease-causing virus starts circulating within a geographic location and infects individuals who have neither a history of travel nor contacts with someone who does. At this stage, it becomes hard to trace the source of the virus transmission, and geographical lockdown becomes highly necessary to prevent the outbreak from reaching stage IV.

D. STAGE IV - COMMUNITY TRANSMISSION The fourth stage of the COVID-19 pandemic in a country is associated with community transmission, i.e., larger outbreaks of local transmission in a country, leading to an extremely high number of reported incidents and deaths. At this stage, the outbreak gets out of control, and finding a cure or vaccine is the only way to mitigate the impact of the disease. Countries like Iran, Turkey, Canada, and the USA are currently in the fourth stage of the COVID-19 pandemic.

Methods

Reviewing the existing apps developed for COVID-19 was conducted by a systematic app review. The study was carried out following the steps discussed in the following subsections.

A. Systematic Search Criteria and Selection

A systematic review of mobile applications across three major mobile app stores: Android Google Play, Apple Store, and Microsoft Store was conducted on 30th April 2020. The search keywords used to find out the related applications were: ‘‘coronavirus,’’ ‘‘COVID 19,’’ ‘‘corona,’’ ‘‘corona outbreak,’’ ‘‘corona pandemic,’’ ‘‘corona and symptom monitoring,’’ ‘‘corona and self-care,’’ ‘‘COVID 19 and symptom monitoring,’’ ‘‘corona and COVID 19,’’ ‘‘corona and self-isolation,’’ ‘‘corona and quarantine,’’ and ‘‘self-care and COVID 19.’’ Each term was searched in each of the three app stores listed. App titles, screenshots of the apps, and app descriptions were considered for the preliminary screening. The second round of exclusion criteria focused on removing the duplicate apps that were found in multiple stores or from multiple search terms or the apps that were no longer available (some of the apps were removed from the app stores). Two members of this research team meticulously reviewed the apps in each round of exclusion. Finally, a total of 25 apps were selected, downloaded, reviewed, and evaluated.

B. Extracting the Study Data

To attain the research goal, we extracted data related to the objective of the apps, functionalities provided by the apps, the platform (operating system) of the selected apps, country context, language of the apps, user ratings, first release, and user comments. The objectives and features of the apps were extracted both from the app description and by the experimental use of the apps. For extracting the features, we first went through the app description and kept notes of the stated features. In some cases, when the app language was not English, we used Google translator. After that, each app was installed and explored to re-check (verify) the extracted features and update the list of extracted features as observed from the experimental use of the app. We ensured that the features extracted against each app from the description must be present in the actual app through the experimental use of the app. Our approach helped us to explore the stated features as well as other features (if any) from our experimental use of the apps.

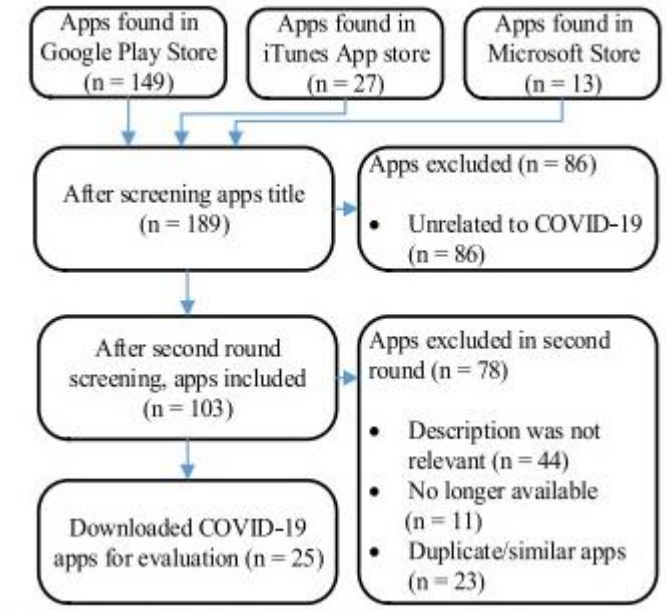


Figure 2.4: App Inclusion and Exclusion Process Flowchart

C. Data Analysis

An affinity diagram approach was used to analyse the data to find out the objectives of the apps. A noticing collecting-thinking approach was used to analyse the user comments (n = 1574) of the selected applications to identify the design characteristics of the apps. The other data including the review scores, development platform, application languages, and the country-context were analysed using descriptive statistics and data synthesizing. We note that our approach of data collection from online sources is widely used in different disciplines, and known as ethnography. Each of the selected apps was investigated to extract their functionalities or features. Each author of this article separately participated in exploring the app functionalities and then in grouping (mapping or clustering) process through an affinity diagram. Firstly, each author extracts the functionalities and assigned names to the functionalities. After that they met together to compare and verify the extracted functionalities for each application. The naming of the extracted functionalities, conflicting functionality names, and disagreements were resolved by discussion through consensus to finalize the functionalities of the selected apps. Secondly, each researcher separately categorized the extracted functionalities into different groups based on their similarities. Next, an intuitive name was given to each group (which are considered as the objective of the application) and then drew the affinity diagram. Finally, the four affinity diagrams (by

four authors) were analysed together to review the grouping or clustering. Thereafter, we came up with one affinity diagram with the final clustering.

The review comments were independently analysed by two researchers. The comments were meticulously read to notice, and code as we discussed above. The assigned keywords or codes were then collected and categorized to represent common themes. After that, both researchers thought about their coding to refine, verify and update the code names, categorizations, relations among the categorizations. Finally, all authors met to compare the coding and classifications. The inter-coder agreement was 0.91. The disagreements were resolved by discussion through consensus. Using an iterative process, all researchers read, sorted, reread, and recombined the data until consensus was achieved.

Discussion

The review found that only a limited number of mobile applications have been developed. However, we note that more application development efforts will eventually come in the future. Our review revealed the main purposes of developing the mobile applications and the functionalities to achieve these objectives for the prevention, mitigation, and containment of COVID-19. The review study also explored the key factors or concerns that affect the end-user experience. Our results indicate that users prefer applications with higher reliability, performance, responsiveness, supportive, ease-of-use, usefulness, security, privacy, and flexibility. We also observed that culturally sensitive applications are needed. COVID-19 pandemic creates a crucial situation in the affected countries.

Implications

This research has implications for future research and practice. The outcomes of this review will greatly contribute to the health institutes, health workers, practitioners, and the governments of the COVID-19 affected countries. Our paper summarizes the currently developed apps to raise awareness about the existing mobile applications, their functionalities, and design characteristics. The findings can help to take necessary initiatives in developing new applications or updating the existing applications to receive maximum benefits out of these applications during the pandemic. The review study also can be considered as a requirement elicitation study. The app developers may consider the revealed objectives and functionalities as the user requirements. For example, if the professionals or governments want to develop a mobile application to ensure population awareness about the COVID-19 pandemic, they may incorporate the services associated with the ’raise

awareness' objective. In contrast, if a government would like to develop a new application to provide all possible services related to COVID-19, it may consider the nine objectives and the entire set of functionalities revealed from this study as shown in Figure 2.2. This review also helps to understand the key information system design characteristics that need to be addressed to develop such applications. The revealed design characteristics can be considered as the design recommendations to the practitioners for developing and evaluating such applications. The design recommendations based on the revealed design characteristics are as follows. The applications should provide quick and accurate responses. The apps need to be usable and useful. The applications should provide authentic information in order to be reliable. The applications should also perform well with the required functionalities. This will also ensure that the applications are useful and supportive with respect to the user requirements. The security and privacy issues should be addressed properly since these aspects are critical to many users during the vulnerable time. Responsive and flexible app design and development need to be ensured. The application should be easy to use in order to ensure that different types of users can use it. Finally, the application should be culturally sensitive, when possible; since users may prefer to have a contextual (local) app in their own language. In sum, the findings of our study would be a great source of inspiration for governments to take necessary initiatives to develop new applications and promote the adoption of existing/new mobile application(s). Finally, for app development companies, these outcomes provide an indication for developing new and innovative mobile apps targeting the affected countries.

Limitations

The study presented in this paper has a few limitations that are important to acknowledge. Firstly, the criteria (search strings/keywords) chosen for selecting the relevant apps for this study may not cover all the available applications, especially the applications that are named in a local language. Secondly, this study does not include the relevant apps that are developed or became available in the app stores after April 30, 2020. Thirdly, the review data was analysed through the qualitative approach using the affinity diagram and noticing-collecting-thinking. Qualitative analysis is subjective and partly depends on analyser's skills, expertise, and knowledge. Therefore, there might be some flaws in the data extraction, coding, and clustering process. However, in order to alleviate these limitations, researchers meticulously conducted checks for apps selection and data analysis, and made adjustments where necessary through discussions. Fourth, in this study, we did not investigate which features

are essential and which are nice-to-have. Thus, a future study may explore or classify the essential and nice-to-have features to achieve the identified objectives. Finally, the online review data collected from the app stores may contain biases. Therefore, future research can use multiple methods to collect and analyse data to investigate the validity of our findings.

2.2 CONCLUSIONS ON SURVEY

We have conducted a brief survey on the existing apps published in Google play store which are related to Covid-19. Efforts have been made to include most of the apps in the survey. The summary of the survey is given in Table 1 which includes the name of the apps, the description of the apps given in Google Play store by their developers and our comments on the apps after using them. The survey shows that there are several apps developed in the country to Figure and contain COVID-19. Most of the states of our country have their own apps with specific features and functionality to help their citizens to stop COVID19 spread, get medical assistance during a crisis, create awareness, and understand safety precautions. The study also shows that there are a limited number of apps which show the COVID-19 containment zones in the country or state and out of these none has the functionality of notifying and alerting the user when they have entered a containment zone. Therefore, no app in the Google Play store is comparable with our proposed application because the idea behind the development of the proposed app is different. This highlights the novelty of the proposed app.

CHAPTER 3

3. SOFTWARE AND HARDWAREREQUIREMENTS

3.1 SOFTWARE REQUIREMENTS

- Browser with V8 Engine
- HTML AND CSS
- JAVA SCRIPT
- Node JS
- PostgreSQL
- JAVA/Kotlin
- Bootstrap
- Android studio

3.2 HARDWARE REQUIREMENTS

Windows

- 64-bit Microsoft® Windows® 8/10
- x86_64 CPU architecture; 2nd generation Intel Core or newer, or AMD CPU with support for a Windows Hypervisor
- 8 GB RAM or more
- 8 GB of available disk space minimum (IDE + Android SDK + Android Emulator)
- 1280 x 800 minimum screen resolution

Mac

- MacOS® 10.14 (Mojave) or higher
- ARM-based chips, or 2nd generation Intel Core or newer with support for Hypervisor Framework
- 8 GB RAM or more
- 8 GB of available disk space minimum (IDE + Android SDK + Android Emulator)
- 1280 x 800 minimum screen resolution

Linux

- Any 64-bit Linux distribution that supports Gnome, KDE, or Unity DE; GNU C Library (glib) 2.31 or later.
- x86_64 CPU architecture; 2nd generation Intel Core or newer, or AMD processor with support for AMD Virtualization (AMD-V) and SSSE3
- 8 GB RAM or more
- 8 GB of available disk space minimum (IDE + Android SDK + Android Emulator)
- 1280 x 800 minimum screen resolution

CHAPTER 4

4. SOFTWARE DEVELOPMENT ANALYSIS

4.1 OVERVIEW OF PROBLEM

The news and media have a great part in creating this awareness by informing the public about the preventive measures that can keep them away from infection. Awareness among the people to carry out all the preventive measures can immensely help to reduce spread of the virus. The country has created containment zones throughout the cities wherever Covid-19 cases have been reported to prevent further spread of the virus. These containment zones have been kept isolated from the outside public to ensure no contamination occurs outside.

1.COVID-19 Tracker:

This website can show if you are in or near Containment Zone.

With the help of the new COVID-19 Hotspot's tracker, users will be able to see what locations are identified as containment zones across India. This is also important because of lockdown restrictions many people will begin to travel for work, it is important to know the coronavirus hotspots that are in your path or near you

2.AAROGYA SETU:

•Aarogya Setu is a mobile application developed by the Government of India to connect essential health services with the people of India in our combined Figure against COVID-19.

•The app is aimed at augmenting the initiatives of the Government of India, particularly the Department of Health, in proactively reaching out to and informing the users of the app regarding risks, best practices and relevant advisories pertaining to the containment of COVID – 19.

3.COVID CARE

•COVID CARE-Quarantine and Contact Health Tracing for Covid Suspects in Arunachal Pradesh (COVID CARE 2020)

The application provides means of self-updating body temperature and Covid-19 symptoms by people of Arunachal Pradesh, thereby helping to monitor them remotely.

4.2 PROBLEM DEFINITION

Currently there are several research works undergoing in the country to prevent Covid-19 cases from rising. Previously our country was importing medical kits like PPE (Personal Protection Kits), mask from outside, but now it has been successful in developing these kits. Along with taking initiatives to Figure this disease, our country has also taken steps to make people aware of the disease. The news and media have a great part in creating this awareness by informing the public about the preventive measures that can keep them away from infection. Awareness among the people to carry out all the preventive measures can immensely help to reduce spread of the virus. The country has created containment zones throughout the cities wherever Covid-19 cases have been reported to prevent further spread of the virus. These containment zones have been kept isolated from the outside public to ensure no contamination occurs outside.

After more than 2 months of the lockdown, the government has relaxed some of the lockdown rules and has permitted reopening of government offices, bus and other road transportation facilities and shopping markets. People can move inside the city for work and other purposes. But the containment zones are still being kept isolated, and new containment zones are being formed wherever Covid-19 cases have been reported. These zones are highly contagious as droplets with virus coughed out from an unscreened asymptomatic patient can travel up to 8 m.

Though these containment zones are guarded by policemen, still there remains a chance that people might unknowingly step into them. In this situation where people can move in the city, these containment zones pose a risk of infection to these city dwellers. Therefore, informing people about the location of the containment zones can help them bypass and avoid these zones and thereby reduce the chance of community transmission.

we focus on developing a mobile based application to provide information regarding the Covid-19 containment zones in West Bengal. The application further tracks the user's location and provides notification alert if the user has entered a containment zone. The application also provides daily Covid-19 case statistics to the users to keep them updated. The application is developed on Android SDK and uses to store the location data. Android's geofencing client is used to create geofences around the containment zones and notification Manager is used to provide notifications.

4.3 MODULES OVERVIEW

Geofencing is a location-based service in which an app or other software uses GPS, RFID, Wi-Fi or cellular data to trigger a pre-programmed action when a mobile device or RFID tag enters or exits a virtual boundary set up around a geographical location, known as a geofence.

Depending on how a geofence is configured it can prompt mobile push notifications, trigger text messages or alerts, send targeted advertisements on social media, allow tracking on vehicle fleets, disable certain technology or deliver location-based marketing data.

Some geofences are set up to monitor activity in secure areas, allowing management to see alerts when anyone enters or leaves a specific area. Businesses can also use geofencing to monitor employees in the field, automate time cards and keep track of company property.

To make use of geofencing, an administrator or developer must first establish a virtual boundary around a specified location in GPS- or RFID-enabled software. This can be as simple as a circle drawn 100 feet around a location on Google Maps, as specified using APIs when developing a mobile app. This virtual geofence will then trigger a response when an authorized device enters or exits that area, as specified by the administrator or developer.

A geofence is most commonly defined within the code of a mobile application, especially since users need to opt-in to location services for the geofence to work. If you go to a concert venue, they might have an app you can download that will deliver information about the event. Or, a retailer might draw a geofence around its outlets to trigger mobile alerts for customers who have downloaded the retailer's mobile app. In these cases, a geofence that is managed by the retailer is programmed into the app, and users can opt to decline location access for the app.

A geofence can also be set up by end-users using geofencing capabilities in their mobile apps. These apps, such as iOS Reminders, allow you to choose an address or location where you want to trigger a specific alert or push notification. This is called an "if this, then that" command, where an app is programmed to trigger an action based off another action. For example, "If I'm five feet from my front door, turn on my lights." Or you might ask a reminder app to send you an alert once you reach a specific location.

Geofencing isn't just for mobile apps – it's used to control and track vehicles in the shipping industry, cattle in agriculture industry and – you'll see this topic pop up in drone discussions. Nearly every drone is pre-programmed to accommodate geofencing, which are usually set

up around airports, open-air venues and even the White House. The FAA can set up these drone-resistant geofences upon request – some barriers will stop a drone in mid-air, while others will trigger a warning message to the user. Some drone geofences will ask for a users' authorization – a process that ties the user's identity to their drone – so that law enforcement can keep track on unmanned drones.

With the rising popularity of mobile devices, geofencing has become a standard practice for plenty of businesses. Once a geographic area has been defined, the opportunities are seemingly endless for what companies can do, and it has become especially popular in marketing and social media.

4.4 DEFINE THE MODUES

Geofencing API from Android is used to create virtual boundaries or fences around geographical locations (Create and monitor geofences 2020). The developers can add geofences at different locations by providing the latitudes and longitudes along with radius to define the virtual boundary at that location. Geofencing technology senses the user's current location and checks whether the location is inside any of the geofences created. A broadcast receiver receives intent contained in a pending intent (an android API) sent by the location services when the user has entered, dwelt, or exited a geofence as shown in Figure. 6 and can initiate a background work or send a notification. The geofence transitions events include enter, exit, and dwell and multiple transition events can be set for the geofences. In this application, the dwell transition is set for the containment zones with a loitering delay of 5 seconds and an expiration duration set to never expire. The broadcast receiver is set to initiate a notification by the notification manager upon receiving an intent. Once the geofences are set, the user would receive notification on entering and dwelling inside a containment zone.

User Registration

The Alert User Registration form contains all the information maintained by the server in earlier releases, as well as new fields to support sending alerts through the web services plug-in or any other alert plug-in.

Location Tracking

As the number of mobile devices and time spent on them increases, mobile technology is reaching new levels of sophistication and advancements. One such technology is the Global Positioning System (GPS) which has evolved to be more precise since its inception. GPS has powered up two significant features in today's mobile devices: Geofencing and Location Based Tracking (or Geolocation).

Red-zone Identification

Geofencing API from Android is used to create virtual boundaries or fences around geographical locations. The developers can add geofences at different locations by providing the latitudes and longitudes along with radius to define the virtual boundary at that location. Geofencing technology senses the user's current location and checks whether the location is

inside any of the geofences created. A broadcast receiver receives intent contained in a pending intent (an android API) sent by the location services when the user has entered, dwelt, or exited a geofence and can initiate a background work or send a notification

RESTful API

Representational State Transfer (REST) API or RESTful web services are architectural styles for communications often used in web services development (RESTful API 2020). These APIs use less bandwidth than the Simple Object Access Protocol (SOAP) and hence they are useful for cloud applications. The RESTful API uses the HTTP methodologies which are defined by the RFC 2616 protocol. The information stored in a RESTful API are resources which can be read, updated, or deleted using resource methods like GET, POST or DELETE. The resources are accessed using Uniform Resource Identifiers (URIs). In this application, we have used a RESTful API from COVID19 India API (COVID19 India API 2020) and the resource that we have used is the Country and State wise data. We have used the GET method to receive the data of Telangana as a JSON object in the application

Alert Notification

Geofencing combines awareness of the user's current location with awareness of the user's proximity to locations that may be of interest. To mark a location of interest, you specify its latitude and longitude. To adjust the proximity for the location, you add a radius. The latitude, longitude, and radius define a geofence, creating a circular area, or fence, around the location of interest.

4.5 MODULES FUNCTIONALITY

User Registration

The Alert User Registration form contains all the information maintained by the server in earlier releases, as well as new fields to support sending alerts through the web services plug-in or any other alert plug-in.

The alert method is determined by the value in the Plugin Name field. When the Web Services Plugin Name is specified, you provide the end point URL for the Web service to which the alert will be sent in the Plugin Values field. You must define the Web service using the WSDL installed with BMC Remedy AR System. In this case, BMC Remedy AR System sends alerts to the plug-in server for processing by a specified plug-in. For information about sending alerts by Web services, see Using Web services with alerts.

You can configure other applications to register and deregister alert users by using C or Java API calls, by using a Web service, or by creating and deleting entries manually. (To deregister users through a Web service, you must use one of the deregister operations described in Registering and deregistering users by web service.)

Location Tracking

As the number of mobile devices and time spent on them increases, mobile technology is reaching new levels of sophistication and advancements. One such technology is the Global Positioning System (GPS) which has evolved to be more precise since its inception. GPS has powered up two significant features in today's mobile devices: Geofencing and Location Based Tracking (or Geolocation).

Let's first understand what is Geofencing and Geolocation and why are they important for modern businesses.

Geofencing is the technique of defining a virtual fence around a geographical location using GPS, RFID (Radio Frequency Identification), Wi-fi network proximity and cellular data, to trigger a pre-programmed action when a mobile object enters (or exits) the fenced periphery. These actions could be to push notifications, alerts, or SMS. Geofencing applications allow administrators to draw virtual boundaries based on a satellite view of a geographic location or using longitude and latitude of a location or using user created web-based maps.

Location Tracking is a precursor to geofencing and helps you track the physical location of a GPS enabled object (a mobile device, vehicle, human carrier with a GPS chip, etc.).

Geofencing and location tracking together eliminate the tedious task of constantly monitoring the location of an object of interest, as it alerts (or notifies) the system when it is within (or breaches) the pre-programmed boundary.

Red-zone Identification

Geofencing API from Android is used to create virtual boundaries or fences around geographical locations (Create and monitor geofences 2020). The developers can add geofences at different locations by providing the latitudes and longitudes along with radius to define the virtual boundary at that location. Geofencing technology senses the user's current location and checks whether the location is inside any of the geofences created. A broadcast receiver receives intent contained in a pending intent (an android API) sent by the location services when the user has entered, and can initiate a background work or send a notification. The geofence transitions events include enter, exit, and dwell and multiple transition events can be set for the geofences. In this application, the dwell transition is set for the containment zones with a loitering delay of 5 seconds and an expiration duration set to never expire. The broadcast receiver is set to initiate a notification by the notification manager upon receiving an intent. Once the geofences are set, the user would receive notification on entering and dwelling inside a containment zone.

RESTful API

Representational State Transfer (REST) API or RESTful web services are architectural styles for communications often used in web services development (RESTful API 2020). These APIs use less bandwidth than the Simple Object Access Protocol (SOAP) and hence they are useful for cloud applications. The RESTful API uses the HTTP methodologies which are defined by the RFC 2616 protocol. The information stored in a RESTful API are resources which can be read, updated, or deleted using resource methods like GET, POST or DELETE. The resources are accessed using Uniform Resource Identifiers (URIs). In this application, we have used a RESTful API from COVID19 India API (COVID19 India API 2020) and the resource that we have used is the Country and State wise data. We have used the GET method to receive the data of Telangana as a JSON object in the application

Alert Notification

Geofencing combines awareness of the user's current location with awareness of the user's proximity to locations that may be of interest. To mark a location of interest, you specify its latitude and longitude. To adjust the proximity for the location, you add a radius. The latitude, longitude, and radius define a geofence, creating a circular area, or fence, around the location of interest.

You can have multiple active geofences, with a limit of 100 per app, per device user. For each geofence, you can ask Location Services to send you entrance and exit events, or you can specify a duration within the geofence area to wait or dwell before triggering an event. You can limit the duration of any geofence by specifying an expiration duration in milliseconds. After the geofence expires, Location Services automatically removes it.

Set For Geofence Monitoring

The first step in requesting geofence monitoring is to request the necessary permissions. To use geofencing, your app must request the following:

- ACCESS_FINE_LOCATION
- ACCESS_BACKGROUND_LOCATION if your app targets Android 10 (API level 29) or higher

To learn more, see the guide on how to request location permissions.

If you want to use a Broadcast Receiver to listen for geofence transitions, add an element specifying the service name. This element must be a child of the <application> element:

```
<application
  android:allowBackup="true">
  ...
  <receiver android:name="GeofenceBroadcastReceiver"/>
<application/>
```

To access the location APIs, you need to create an instance of the Geofencing client. To learn how to connect your client:

Create and add geofences

Your app needs to create and add geofences using the location API's builder class for creating Geofence objects, and the convenience class for adding them. Also, to handle the intents sent from Location Services when geofence transitions occur, you can define a Pending Intent as shown in this section.

Note: On single-user devices, there is a limit of 100 geofences per app. For multi-user devices, the limit is 100 geofences per app per device user.

Create geofence objects

First, use Geofence Builder to create a geofence, setting the desired radius, duration, and transition types for the geofence. For example, to populate a list object:

This example pulls data from a constants file. In actual practice, apps might dynamically create geofences based on the user's location.

Specify geofences and initial triggers

The following snippet uses the Geofencing Request class and its nested Geofencing Request Builder class to specify the geofences to monitor and to set how related geofence events are triggered:

This example shows the use of two geofence triggers. The GEOFENCE_TRANSITION_ENTER transition triggers when a device enters a geofence, and the GEOFENCE_TRANSITION_EXIT transition triggers when a device exits a geofence. Specifying INITIAL_TRIGGER_ENTER tells Location services that GEOFENCE_TRANSITION_ENTER should be triggered if the device is already inside the geofence.

In many cases, it may be preferable to use instead INITIAL_TRIGGER_DWELL, which triggers events only when the user stops for a defined duration within a geofence. This approach can help reduce "alert spam" resulting from large numbers notifications when a device briefly enters and exits geofences. Another strategy for getting best results from your geofences is to set a minimum radius of 100 meters. This helps account for the location accuracy of typical Wi-Fi networks, and also helps reduce device power consumption.

Define a broadcast receiver for geofence transitions

An Intent sent from Location Services can trigger various actions in your app, but you should not have it start an activity or fragment, because components should only become visible in response to a user action. In many cases, a Broadcast Receiver is a good way to handle a geofence transition. A Broadcast Receiver gets updates when an event occurs, such as a transition into or out of a geofence, and can start long-running background work.

The following snippet shows how to define a Pending Intent that starts a Broadcast Receiver:

Add geofences

To add geofences, use the GeofencingClient.addGeofences() method. Provide the GeofencingRequest object, and the Pending Intent. The following snippet demonstrates processing the results:

Handle geofence transitions

When Location Services detects that the user has entered or exited a geofence, it sends out the Intent contained in the Pending Intent you included in the request to add geofences. A broadcast receiver like GeofenceBroadcastReceiver notices that the Intent was invoked and can then obtain the geofencing event from the intent, determine the type of Geofence transition(s), and determine which of the defined geofences was triggered. The broadcast receiver can direct an app to start performing background work or, if desired, send a notification as output.

CHAPTER 5

5. PROJECT SYSTEM DESIGN

5.1 DATAFLOW DIAGRAMS

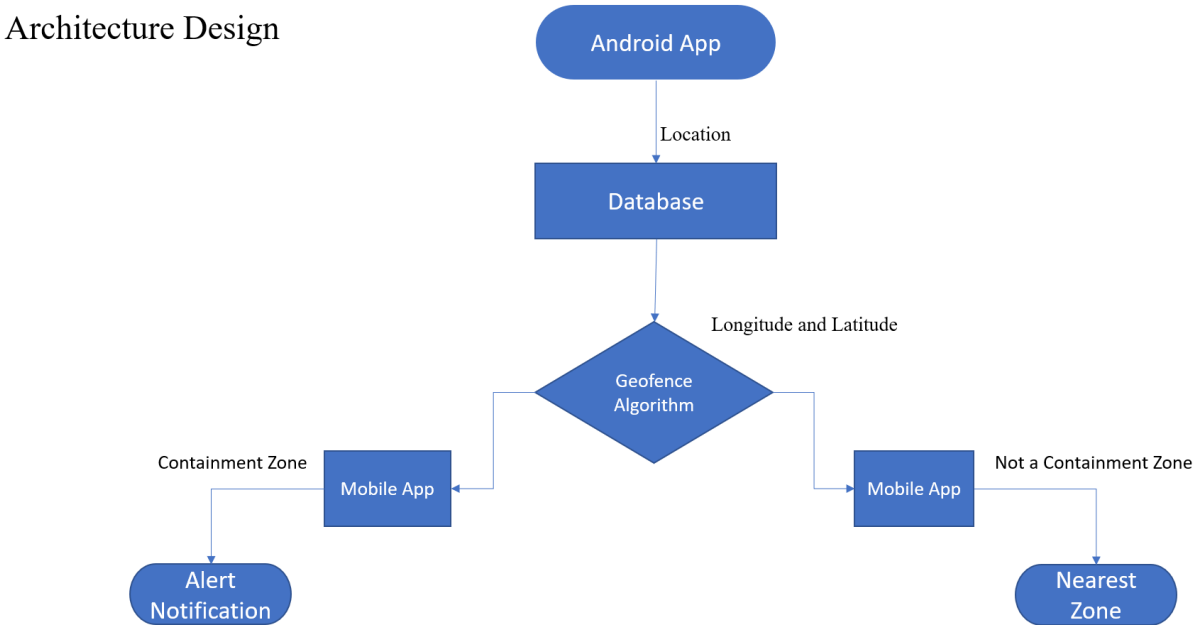


Figure 5.1: Architecture Design of the System

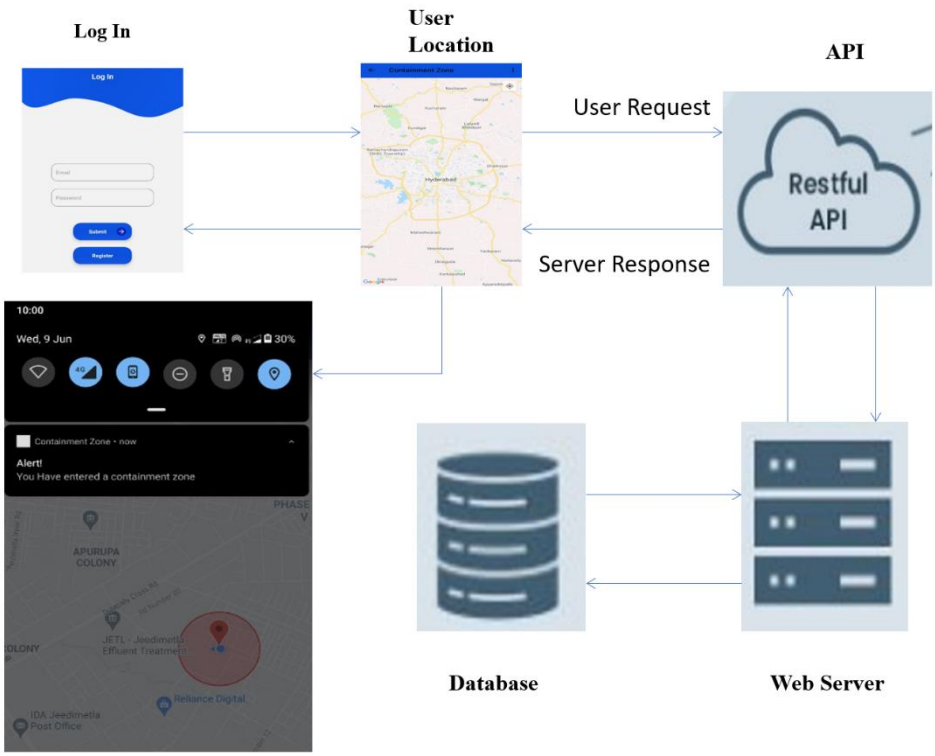


Figure 5.2: Architecture Desing of the User Application

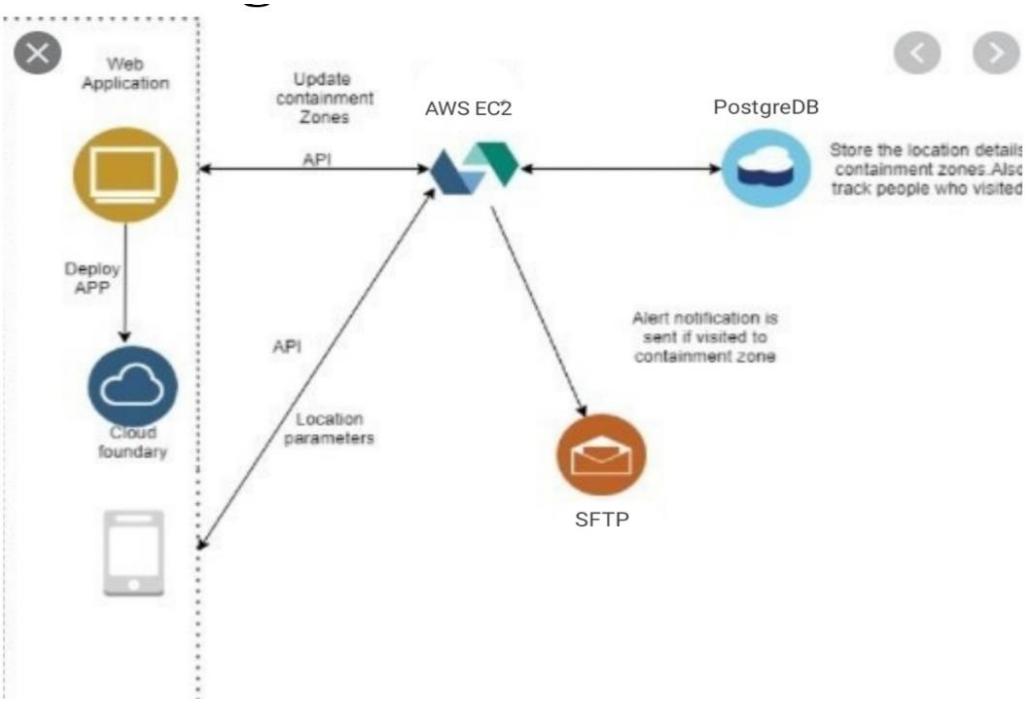


Figure 5.3: Detailed Design of the System

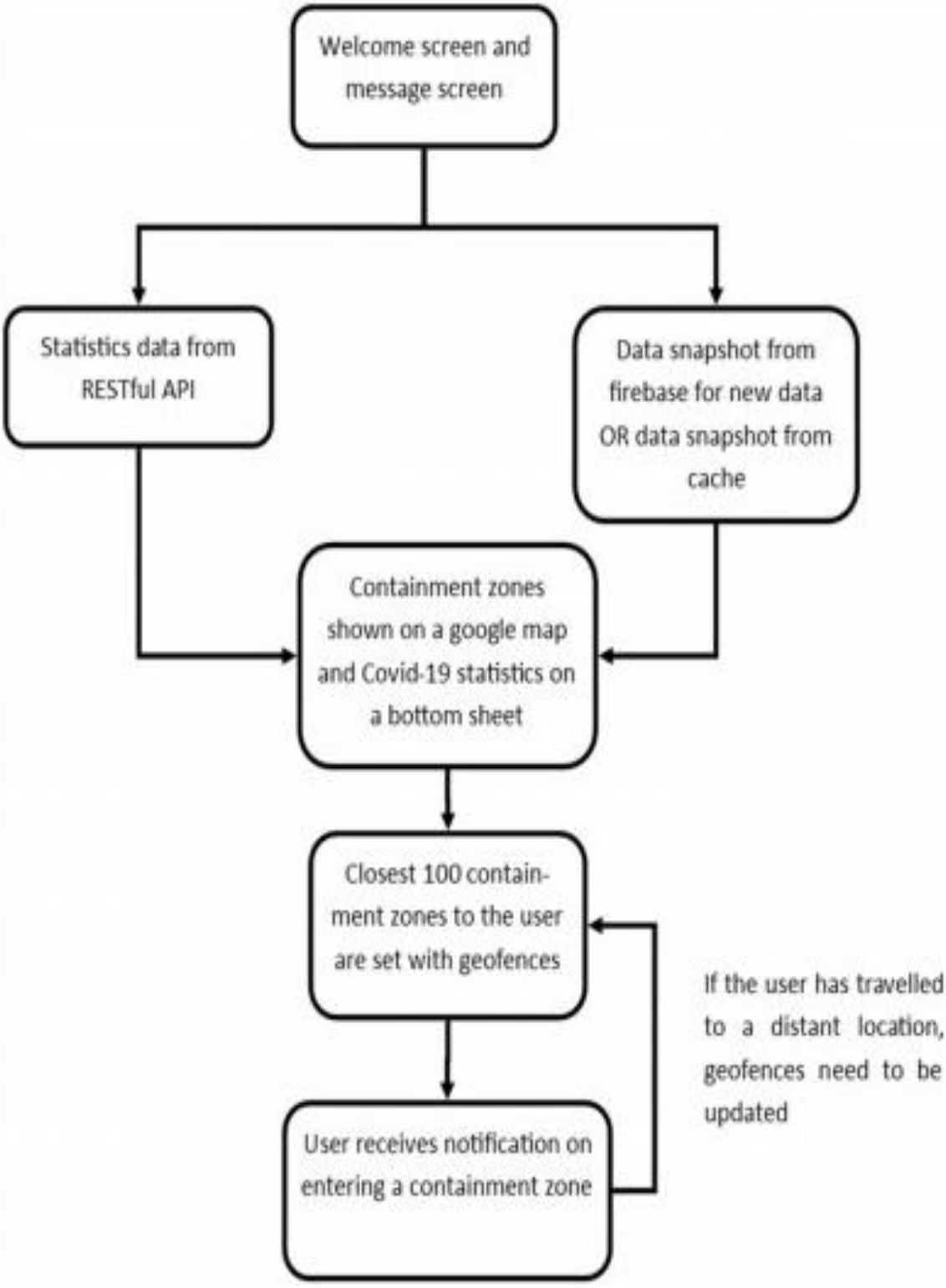
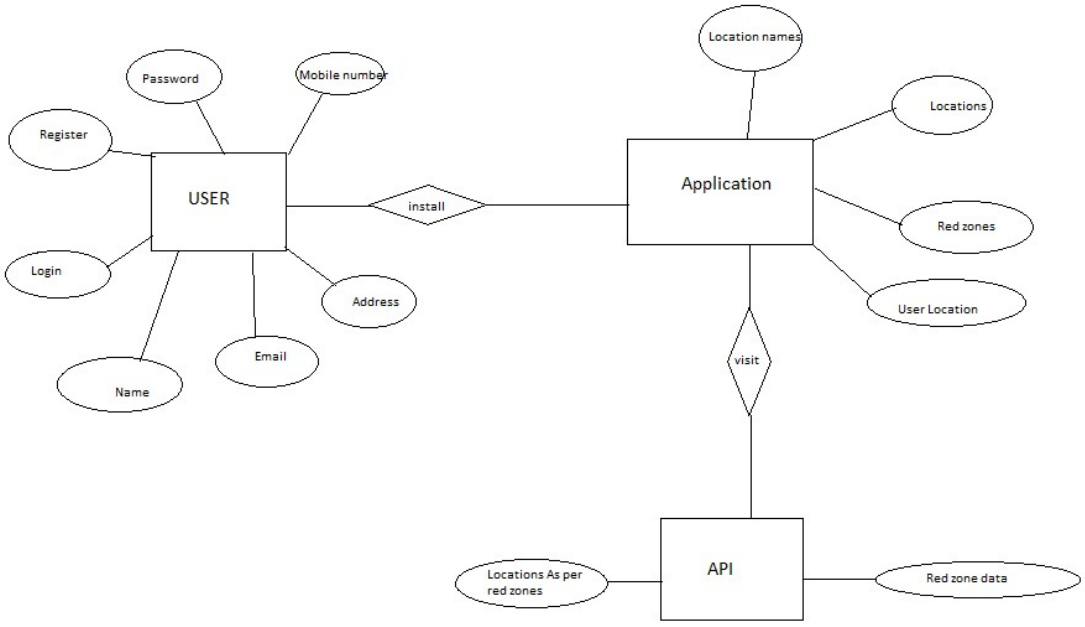


Figure 5.4: Data Flow Diagram of Working Application

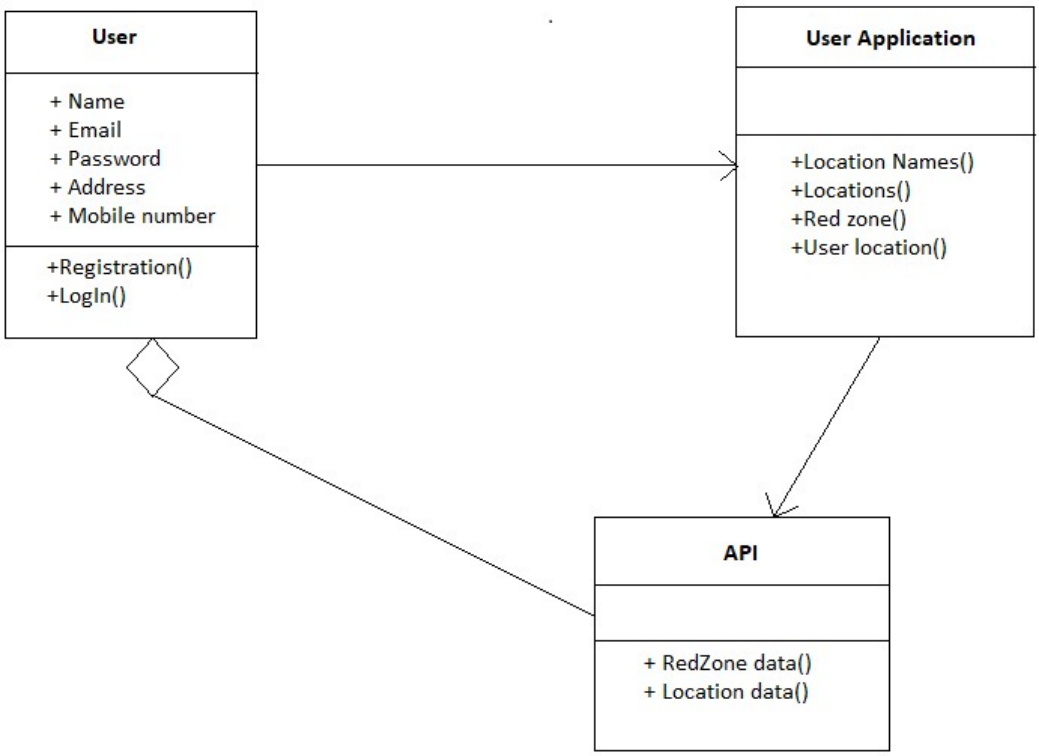
5.2 ENTITY RELATIONSHIP DIAGRAM

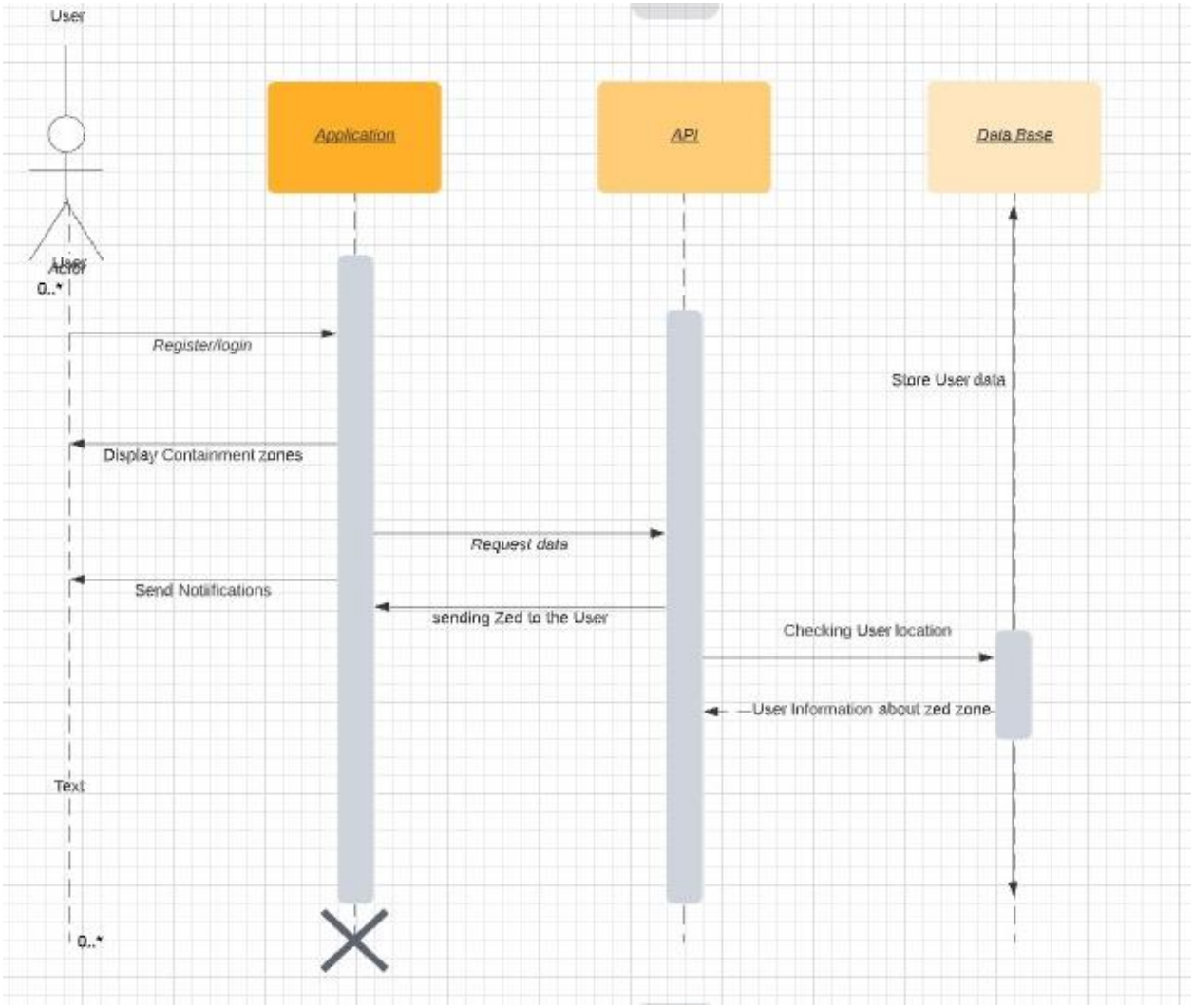


E-R Diagram of the system

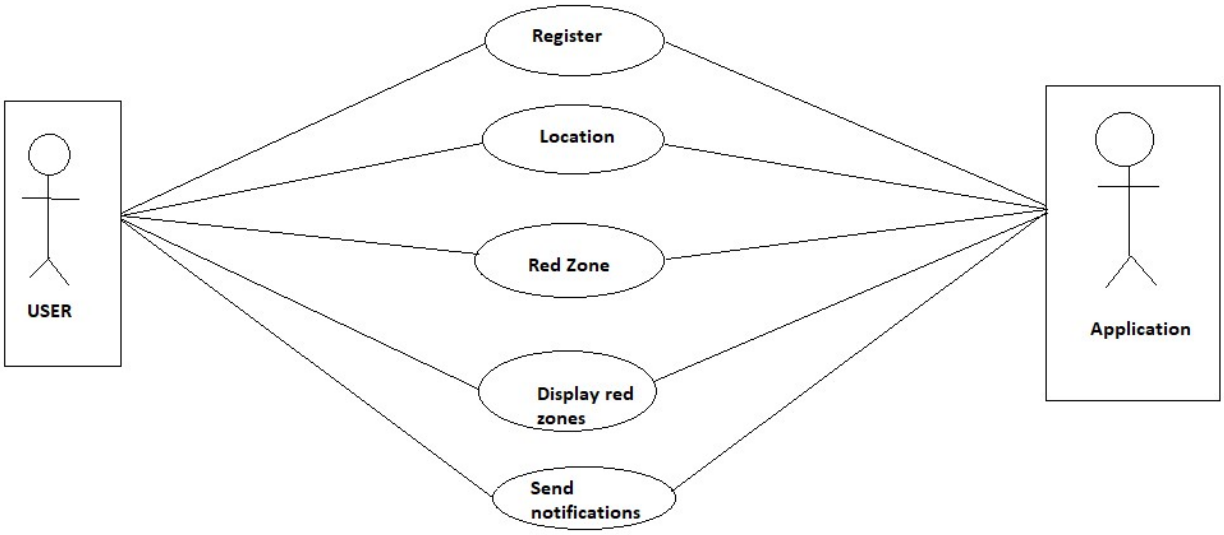
5.3 UNIFIED MODEL LANGUAGE DIAHRAMS

Class Diagram





Sequence Diagram



Use Case Diagram

CHAPTER 6

6. PROJECT CODING

6.1 CODE TEMPLATES

Geo - fencing Implementation:

```
<?xml version="1.0" encoding="UTF-8"?>
<module external.linked.project.id="GeoFence"
external.linked.project.path="$MODULE_DIR$"
external.root.project.path="$MODULE_DIR$" external.system.id="GRADLE"
type="JAVA_MODULE" version="4">
  <component name="FacetManager">
    <facet type="java-gradle" name="Java-Gradle">
      <conFigureuration>
        <option name="BUILD_FOLDER_PATH" value="$MODULE_DIR$/build" />
        <option name="BUILDABLE" value="false" />
      </conFigureuration>
    </facet>
  </component>
  <component name="NewModuleRootManager" LANGUAGE_LEVEL="JDK_1_7" inherit-
compiler-output="true">
    <exclude-output />
    <content url="file://$MODULE_DIR$">
      <excludeFolder url="file://$MODULE_DIR$/.gradle" />
    </content>
    <orderEntry type="inheritedJdk" />
    <orderEntry type="sourceFolder" forTests="false" />
  </component>
</module>
package com.example.geofence;

import android.content.Context;

import androidx.test.platform.app.InstrumentationRegistry;
import androidx.test.ext.junit.runners.AndroidJUnit4;

import org.junit.Test;
import org.junit.runner.RunWith;

import static org.junit.Assert.*;

/**
 * Instrumented test, which will execute on an Android device.
 *
 * @see <a href="http://d.android.com/tools/testing">Testing documentation</a>
 */
@RunWith(AndroidJUnit4.class)
public class ExampleInstrumentedTest {
    @Test
    public void useAppContext() {
        // Context of the app under test.
        Context appContext =
InstrumentationRegistry.getInstrumentation().getTargetContext();
```

```
        assertEquals("com.example.geofence", appContext.getPackageName());
    }
}
```

Geofence Broadcast Receiver:

```
package com.example.geofence;
import android.content.BroadcastReceiver;
import android.content.Context;
import android.content.Intent;
import android.util.Log;
import android.widget.Toast;

import com.google.android.gms.location.Geofence;
import com.google.android.gms.location.GeofencingEvent;

import java.util.List;

public class GeofenceBroadcastReceiver extends BroadcastReceiver {

    private static final String TAG = "GeofenceBroadcastReceiv";

    @Override
    public void onReceive(Context context, Intent intent) {
        // TODO: This method is called when the BroadcastReceiver is receiving
        // an Intent broadcast.
//        Toast.makeText(context, "Geofence triggered...",
Toast.LENGTH_SHORT).show();

        NotificationHelper notificationHelper = new NotificationHelper(context);

        GeofencingEvent geofencingEvent = GeofencingEvent.fromIntent(intent);

        if (geofencingEvent.hasError()) {
            Log.d(TAG, "onReceive: Error receiving geofence event...");
            return;
        }

        List<Geofence> geofenceList = geofencingEvent.getTriggeringGeofences();
        for (Geofence geofence: geofenceList) {
            Log.d(TAG, "onReceive: " + geofence.getRequestId());
        }
//        Location location = geofencingEvent.getTriggeringLocation();
        int transitionType = geofencingEvent.getGeofenceTransition();

        switch (transitionType) {
            case Geofence.GEOFENCE_TRANSITION_ENTER:
                Toast.makeText(context, "GEOFENCE_TRANSITION_ENTER",
Toast.LENGTH_SHORT).show();
                notificationHelper.sendHighPriorityNotification("GEOFENCE_TRANSI
TION_ENTER", "", MainActivity.class);
                break; case Geofence.GEOFENCE_TRANSITION_EXIT:
                Toast.makeText(context, "GEOFENCE_TRANSITION_EXIT",
Toast.LENGTH_SHORT).show();
                notificationHelper.sendHighPriorityNotification("GEOFENCE_TRANSI
TION_EXIT", "", MainActivity.class);
                break;
```

```
        }

    }
}
```

Geofence Helper:

```
package com.example.geofence;
import android.app.PendingIntent;
import android.content.Context;
import android.content.ContextWrapper;
import android.content.Intent;

import com.google.android.gms.common.api.ApiException;
import com.google.android.gms.location.Geofence;
import com.google.android.gms.location.GeofenceStatusCodes;
import com.google.android.gms.location.GeofencingRequest;
import com.google.android.gms.maps.model.LatLng;

public class GeofenceHelper extends ContextWrapper {

    private static final String TAG = "GeofenceHelper";
    PendingIntent pendingIntent;

    public GeofenceHelper(Context base) {
        super(base);
    }

    public GeofencingRequest getGeofencingRequest(Geofence geofence) {
        return new GeofencingRequest.Builder()
                .addGeofence(geofence)
                .setInitialTrigger(GeofencingRequest.INITIAL_TRIGGER_ENTER)
                .build();
    }

    public Geofence getGeofence(String ID, LatLng latLng, float radius, int
transitionTypes) {
        return new Geofence.Builder()
                .setCircularRegion(latLng.latitude, latLng.longitude, radius)
                .setRequestId(ID)
                .setTransitionTypes(transitionTypes)
                .setLoiteringDelay(5000)
                .setExpirationDuration(Geofence.NEVER_EXPIRE)
                .build();
    }

    public PendingIntent getPendingIntent() {
        if (pendingIntent != null) {
            return pendingIntent;
        }
        Intent intent = new Intent(this, GeofenceBroadcastReceiver.class);
        pendingIntent = PendingIntent.getBroadcast(this, 2607, intent,
PendingIntent.FLAG_UPDATE_CURRENT);

        return pendingIntent;
    }

    public String getErrorString(Exception e) {
        if (e instanceof ApiException) {
```

```
            ApiException apiException = (ApiException) e;
            switch (apiException.getStatusCode()) {
                case GeofenceStatusCodes
                        .GEOFENCE_NOT_AVAILABLE:
                    return "GEOFENCE_NOT_AVAILABLE";
                case GeofenceStatusCodes
                        .GEOFENCE_TOO_MANY_GEOFENCES:
                    return "GEOFENCE_TOO_MANY_GEOFENCES";
                case GeofenceStatusCodes
                        .GEOFENCE_TOO_MANY_PENDING_INTENTS:
                    return "GEOFENCE_TOO_MANY_PENDING_INTENTS";
            }
        }
        return e.getLocalizedMessage();
    }
}
```

Notification Helper:

```
package com.example.geofence;
import android.app.Notification;
import android.app.NotificationChannel;
import android.app.NotificationManager;
import android.app.PendingIntent;
import android.content.Context;
import android.content.ContextWrapper;
import android.content.Intent;
import android.graphics.Color;
import android.os.Build;

import androidx.annotation.RequiresApi;
import androidx.core.app.NotificationCompat;
import androidx.core.app.NotificationManagerCompat;

import java.util.Random;

public class NotificationHelper extends ContextWrapper {

    private static final String TAG = "NotificationHelper";

    public NotificationHelper(Context base) {
        super(base);
        if (Build.VERSION.SDK_INT >= Build.VERSION_CODES.O) {
            createChannels();
        }
    }

    private String CHANNEL_NAME = "High priority channel";
    private String CHANNEL_ID = "com.example.notifications" + CHANNEL_NAME;

    @RequiresApi(api = Build.VERSION_CODES.O)
    private void createChannels() {
        NotificationChannel notificationChannel = new
NotificationChannel(CHANNEL_ID, CHANNEL_NAME,
NotificationManager.IMPORTANCE_HIGH);
        notificationChannel.enableLights(true);
        notificationChannel.enableVibration(true);
        notificationChannel.setDescription("this is the description of the
channel.");
```

```
        notificationChannel.setLightColor(Color.RED);
        notificationChannel.setLockscreenVisibility(Notification.VISIBILITY_PUBL
IC);
        NotificationManager manager = (NotificationManager)
getSystemService(Context.NOTIFICATION_SERVICE);
        manager.createNotificationChannel(notificationChannel);
    }

    public void sendHighPriorityNotification(String title, String body, Class
activityName) {

        Intent intent = new Intent(this, activityName);
        PendingIntent pendingIntent = PendingIntent.getActivity(this, 267,
intent, PendingIntent.FLAG_UPDATE_CURRENT);

        Notification notification = new NotificationCompat.Builder(this,
CHANNEL_ID)
//                .setContentTitle(title)
//                .setContentText(body)
                .setSmallIcon(R.drawable.ic_launcher_background)
                .setPriority(NotificationCompat.PRIORITY_HIGH)
                .setStyle(new
NotificationCompat.BigTextStyle().setSummaryText("summary").setBigContentTitle(t
itle).bigText(body))
                .setContentIntent(pendingIntent)
                .setAutoCancel(true)
                .build();

        NotificationManagerCompat.from(this).notify(new Random().nextInt(),
notification);


    }

}
```

Login:

```
package com.example.geofence;

import androidx.appcompat.app.AppCompatActivity;

import android.content.Context;
import android.content.Intent;
import android.content.SharedPreferences;
import android.graphics.Color;
import android.graphics.drawable.ColorDrawable;
import android.os.AsyncTask;
import android.os.Bundle;
import android.speech.tts.TextToSpeech;
import android.util.Log;
import android.view.LayoutInflater;
import android.view.View;
import android.widget.EditText;
import android.widget.TextView;
import android.widget.Toast;

import org.json.JSONException;
import org.json.JSONObject;
```

```
import java.io.BufferedReader;
import java.io.DataOutputStream;
import java.io.IOException;
import java.io.InputStreamReader;
import java.net.HttpURLConnection;
import java.net.MalformedURLException;
import java.net.ProtocolException;
import java.net.URL;
import java.net.URLConnection;
import java.util.Timer;
import java.util.TimerTask;

public class Login extends AppCompatActivity {
    EditText email;
    EditText password;
    TextView submit,register;
    String emailid;
    String pass;
    SharedPreferences sharedpreferences ;
    SharedPreferences.Editor editor;


    @Override
    protected void onCreate(Bundle savedInstanceState) {
        super.onCreate(savedInstanceState);
        setContentView(R.layout.activity_login);
        email=(EditText)findViewById(R.id.email);
        password=(EditText)findViewById(R.id.password);
        sharedpreferences= getSharedPreferences("mydata", Context.MODE_PRIVATE);

        submit=(TextView) findViewById(R.id.submit);
        register=(TextView) findViewById(R.id.register);
        submit.setOnClickListener(new View.OnClickListener() {
            @Override
            public void onClick(View v) {
                 emailid=email.getText().toString();
                 pass=password.getText().toString();
                if(!emailid.isEmpty() || !pass.isEmpty()) {
                    sendLogin login=new sendLogin();
                    login.execute();
                }
                else{
                    email.setError("Please Enter Email Id");
                    password.setError("Please Enter Password");

                }

            }
        });

        register.setOnClickListener(new View.OnClickListener() {
            @Override
            public void onClick(View v) {
                Intent in=new Intent(Login.this, Registration.class);
                startActivity(in);
            }
        });
```

```
    }


    class sendLogin extends AsyncTask<Void,Void,String>{

        @Override
        protected String doInBackground(Void... voids) {

            try {
                //Initializing the URL
                URL url = new URL("http://18.207.120.122:3002/signin");
                HttpURLConnection connection = (HttpURLConnection)
url.openConnection();
                connection.setDoOutput(true);
                connection.setDoInput(true);
                connection.setInstanceFollowRedirects(false);
                connection.setRequestMethod("POST");
                connection.setRequestProperty("Content-Type",
"application/json");
                connection.setRequestProperty("charset", "utf-8");
                connection.setUseCaches (false);
                //Sending the Data to the server
                DataOutputStream wr = new
DataOutputStream(connection.getOutputStream ());
                JSONObject jsonParam = new JSONObject();
                jsonParam.put("email", emailid);
                jsonParam.put("password1", pass);
                wr.writeBytes(jsonParam.toString());
                wr.flush();
                wr.close();
                //Receving the data from the server
                BufferedReader bufferedReader = new BufferedReader(new
InputStreamReader(connection.getInputStream()));
                StringBuilder stringBuilder = new StringBuilder();
                String line;
                System.out.println(" bufferedreader response :" +
bufferedReader);

                while ((line = bufferedReader.readLine()) != null) {
                    stringBuilder.append(line);


                }
                bufferedReader.close();
                Log.e("res", stringBuilder.toString());
                return stringBuilder.toString();
            } catch (Exception e) {
                Log.e("ERROR", e.getMessage(), e);
                return null;
            }

        }

        @Override
        protected void onPostExecute(String s) {
            super.onPostExecute(s);
            Log.e("Output--1",""+s);
            if(s.contentEquals("valid")){
```

```
                editor = sharedpreferences.edit();
                Toast.makeText(Login.this, "Login Success",
Toast.LENGTH_SHORT).show();
                Intent in=new Intent(Login.this, MainActivity.class);
                editor.putString("email", emailid);
                editor.commit();
                editor.apply();
                startActivity(in);
                finish();
            }
            else{
                Toast.makeText(Login.this, "Login Failed",
Toast.LENGTH_SHORT).show();
            }
        }
    }
}
```

Registration:

```
package com.example.geofence;

import androidx.appcompat.app.AppCompatActivity;

import android.content.Intent;
import android.os.AsyncTask;
import android.os.Bundle;
import android.util.Log;
import android.view.View;
import android.widget.EditText;
import android.widget.TextView;

import org.json.JSONObject;

import java.io.BufferedReader;
import java.io.DataOutputStream;
import java.io.InputStreamReader;
import java.net.HttpURLConnection;
import java.net.URL;

public class Registration extends AppCompatActivity {
    EditText email;
    EditText password;
    EditText address;
    EditText mobile;
    EditText name;
    TextView submit,login;
    String emailid;
    String pass,mobileNum,addr,fname;
    @Override
    protected void onCreate(Bundle savedInstanceState) {
        super.onCreate(savedInstanceState);
        setContentView(R.layout.activity_registration);
        email=(EditText)findViewById(R.id.email);
        password=(EditText)findViewById(R.id.password);
        name=(EditText)findViewById(R.id.name);
        address=(EditText)findViewById(R.id.address);
        mobile=(EditText)findViewById(R.id.mobile);
```

```
        submit=(TextView) findViewById(R.id.submit);
        login=(TextView) findViewById(R.id.login);
        submit.setOnClickListener(new View.OnClickListener() {
            @Override
            public void onClick(View v) {
                emailid=email.getText().toString();
                pass=password.getText().toString();
                mobileNum=mobile.getText().toString();
                addr=address.getText().toString();
                fname=name.getText().toString();
                if(!emailid.isEmpty() || !pass.isEmpty()||!mobileNum.isEmpty()
|| !addr.isEmpty()||!fname.isEmpty()) {
                    register reg=new register();
                    reg.execute();
                }
                else{
                    email.setError("Please Enter Email Id");
                    password.setError("Please Enter Password");

                }

            }
        });


        login.setOnClickListener(new View.OnClickListener() {
            @Override
            public void onClick(View v) {
                Intent in=new Intent(Registration.this, Login.class);
                startActivity(in);
            }
        });
    }

    class register extends AsyncTask<Void,Void,String> {

        @Override
        protected String doInBackground(Void... voids) {

            try {
                //Initializing the URL
                URL url = new URL("http://18.207.120.122:3002/signup");
                HttpURLConnection connection = (HttpURLConnection)
url.openConnection();
                connection.setDoOutput(true);
                connection.setDoInput(true);
                connection.setInstanceFollowRedirects(false);
                connection.setRequestMethod("POST");
                connection.setRequestProperty("Content-Type",
"application/json");
                connection.setRequestProperty("charset", "utf-8");
                connection.setUseCaches (false);
                DataOutputStream wr = new
DataOutputStream(connection.getOutputStream ());
                JSONObject jsonParam = new JSONObject();
                jsonParam.put("email", emailid);
                jsonParam.put("password1", pass);
```

```
                jsonParam.put("address", addr);
                jsonParam.put("mobile", mobileNum);
                jsonParam.put("name", fname);
                wr.writeBytes(jsonParam.toString());
                wr.flush();
                wr.close();

                BufferedReader bufferedReader = new BufferedReader(new
InputStreamReader(connection.getInputStream()));
                StringBuilder stringBuilder = new StringBuilder();
                String line;
                System.out.println(" bufferedreader response :" +
bufferedReader);

                while ((line = bufferedReader.readLine()) != null) {
                    stringBuilder.append(line);


                }
                bufferedReader.close();
                Log.e("res", stringBuilder.toString());
                return stringBuilder.toString();
            } catch (Exception e) {
                Log.e("ERROR", e.getMessage(), e);
                return null;
            }



        }

        @Override
        protected void onPostExecute(String s) {
            super.onPostExecute(s);
            Log.e("Output--1",""+s);

        }
    }
}
```

6.2 OUTLINE FOR VARIOUS FILES

Latitude and Longitude:

Geofencing is often used tool in Geographic data science, especially in marketing, security and zoning applications. The example in the above GIF shows an app that alerts vehicles based on their location and London's Congestion Charge Zone (CCZ). The application calculates the congestion charge and tracks the number of vehicles inside the congestion area at a given time.

The concept of geofencing is straightforward, yet it is a powerful technique that enhances location applications. Simply put, a geofence is a defining virtual boundary around geographic objects or an area, so that every time a user enters or leaves the boundary perimeters, actions or notifications can be triggered. With the increased use of smartphones, GPS, and location services, geofencing becomes an indispensable tool in location data analytics and intelligence.

In this tutorial, we use a GPS Trajectory dataset. It contains GPS

points (latitude and longitude) with timestamps. It also provides unique track_id for each trajectory. Let us read the data with the Pandas library.

https://www.dropbox.com/s/ejev7z29lzirbo5/GPSTrajectory.zip

The first five rows of the dataset are shown below. We have latitude and longitude as well as track_id and time.

#	Country	State	Pincode	Latitude	Longitude
1	india	andhra	507894	27.809	76.345
2	India	Tamilnadu	505124	18.123	79.345
3	India	Karnataka	500456	28.945	79.567
4	India	Telanagana	505185	27.204	77.49
5	India	Telanagana	505185	27	77

Table: 6.2 Containment Zone Allocation

Trajectory DataFrame

We convert the Dataframe into Geodataframe, which allows us to perform geofencing. Converting the data frame to Geodataframe with Geopandas is straightforward.

```
px.set_mapbox_access_token("pk.eyJ1Ijoic2hha2Fzb20iLCJhIjoiY2plMWg1NGFpMXZ5NjJxbjhlM2ttN3AwbiJ9.RtGYHmreKiyBfHuElgYq_w")
px.scatter_mapbox(gdf, lat="latitude", lon="longitude" ,size_max=6, zoom=8, width=1200, height=800)
```

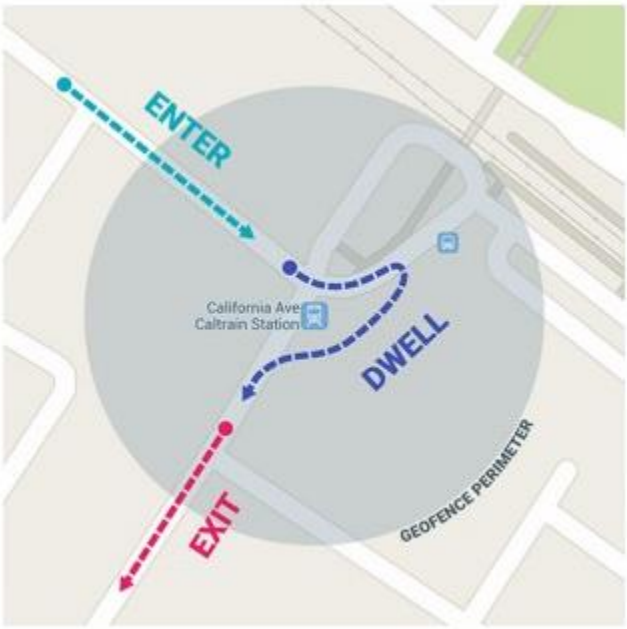


Figure 6.1: User Boundary details or outliers of zones

Geofence trigger events

Working of the Application The application gets data from the Amazon Web Services database. A collection is created in Amazon Web Services with containment zones as documents. Each document has four fields: latitude, longitude, location name and radius. Accordingly, a Java object is created which can get the data from the document. In the map's activity, the EC2 instance and collection references are created to which a snapshot listener is attached. The snapshot listener retrieves the document snapshots which are then converted into the Java object mentioned earlier. With the help of getters each data from the document is retrieved and are converted to string. Markers and circles are set using the location coordinates and radius and tags are given by the location names. The google map gets populated with these markers surrounded by circles which represents the containment zones. A JSON request is made with the get method to the REST API URL which returns the West Bengal Covid-19 case data as a response. The response is converted to a JSON object and the information is extracted. The google map shows all the containment zones in West

Bengal along with the location of the user using “set my location” enabled (Google map API). The Geofencing API can create up to 100 geofences per device and the number of containment zones are more than 1000 in West Bengal. The solution to this problem is to create100 geofences on the 100 nearest containment zones. Once the map is loaded and populated with containment zones and user’s location, the user can press a button to add the geofences on the closest 100 containment zones. The snapshot listener returns the documents containing the locations of the containment zones from AWS which are not sorted according to the distance between the user and containment zone. The distance between the user and each containment zone is measured using distance between method of the location manager (Android Developer-Locations 2020). This distance is then used as a key and is stored along with the document in a Tree map. Likewise, all the containment zones with their distance from the user get stored in the tree map and get sorted according to the distance or key. First 100 entries from the tree map are retrieved and geofences are created on these 100 containment zones as shown in Figure. 8. Once the geofences are set, the user can get notification on entering the containment zones

6.3 CLASS WITH FUNCTIONALITY

Create and add geofences

Your app needs to create and add geofences using the location API's builder class for creating Geofence objects, and the convenience class for adding them. Also, to handle the intents sent from Location Services when geofence transitions occur, you can define a Pending Intent as shown in this section**.**

Create geofence objects

First, use Geofence Builder to create a geofence, setting the desired radius, duration, and transition types for the geofence. For example, to populate a list object:

```
geofenceList.add(new Geofence.Builder()
    // Set the request ID of the geofence. This is a string to identify this
    // geofence.
    .setRequestId(entry.getKey())

    .setCircularRegion(
            entry.getValue().latitude,
            entry.getValue().longitude,
            Constants.GEOFENCE_RADIUS_IN_METERS
    )
    .setExpirationDuration(Constants.GEOFENCE_EXPIRATION_IN_MILLISECONDS)
    .setTransitionTypes(Geofence.GEOFENCE_TRANSITION_ENTER |
            Geofence.GEOFENCE_TRANSITION_EXIT)
    .build());
```

Specify geofences and initial triggers

The following snippet uses the GeofencingRequest class and its nested GeofencingRequestBuilder class to specify the geofences to monitor and to set how related geofence events are triggered:

```
private GeofencingRequest getGeofencingRequest() {
    GeofencingRequest.Builder builder = new GeofencingRequest.Builder();
    builder.setInitialTrigger(GeofencingRequest.INITIAL_TRIGGER_ENTER);
    builder.addGeofences(geofenceList);
```

```
    return builder.build();
}
```

Define a broadcast receiver for geofence transitions

An Intent sent from Location Services can trigger various actions in your app, but you should *not* have it start an activity or fragment, because components should only become visible in response to a user action. In many cases, a Broadcast Receiver is a good way to handle a geofence transition. A Broadcast Receiver gets updates when an event occurs, such as a transition into or out of a geofence, and can start long-running background work.

The following snippet shows how to define a Pending Intent that starts a Broadcast Receiver:

```
public class Main Activity extends App Compat Activity {

    // ...

    private PendingIntent getGeofencePendingIntent() {
        // Reuse the PendingIntent if we already have it.
        if (geofencePendingIntent != null) {
            return geofencePendingIntent;
        }
        Intent intent = new Intent(this, GeofenceBroadcastReceiver.class);
        // We use FLAG_UPDATE_CURRENT so that we get the same pending intent back
when
        // calling addGeofences() and removeGeofences().
        geofencePendingIntent = PendingIntent.getBroadcast(this, 0, intent, PendingIntent.
                FLAG_UPDATE_CURRENT);
        return geofencePendingIntent;
    }
```

Add geofences

To add geofences, use the GeofencingClient.addGeofences() method. Provide the GeofencingRequest object, and the PendingIntent. The following snippet demonstrates processing the results:

```
geofencingClient.addGeofences(getGeofencingRequest(), getGeofencePendingIntent())
        .addOnSuccessListener(this, new OnSuccessListener<Void>() {
            @Override
            public void onSuccess(Void aVoid) {
                // Geofences added
                // ...
            }
        })
        .addOnFailureListener(this, new OnFailureListener() {
            @Override
            public void onFailure(@NonNull Exception e) {
                // Failed to add geofences
                // ...
            }
        });
```

Handle geofence transitions

When Location Services detects that the user has entered or exited a geofence, it sends out the Intent contained in the Pending Intent you included in the request to add geofences. A broadcast receiver like GeofenceBroadcastReceiver notices that the Intent was invoked and can then obtain the geofencing event from the intent, determine the type of Geofence transition(s), and determine which of the defined geofences was triggered. The broadcast receiver can direct an app to start performing background work or, if desired, send a notification as output.

```
public class GeofenceBroadcastReceiver extends Broadcast Receiver {
    // ...
    protected void onReceive(Context context, Intent intent) {
        GeofencingEvent geofencingEvent = GeofencingEvent.fromIntent(intent);
        if (geofencingEvent.hasError()) {
            String errorMessage = GeofenceStatusCodes
```

```
                .getStatusCodeString(geofencingEvent.getErrorCode());
        Log.e(TAG, errorMessage);
        return;
    }

    // Get the transition type.
    int geofenceTransition = geofencingEvent.getGeofenceTransition();

    // Test that the reported transition was of interest.
    if (geofenceTransition == Geofence.GEOFENCE_TRANSITION_ENTER ||
            geofenceTransition == Geofence.GEOFENCE_TRANSITION_EXIT) {

        // Get the geofences that were triggered. A single event can trigger
        // multiple geofences.
        List<Geofence> triggeringGeofences = geofencingEvent.getTriggeringGeofences();

        // Get the transition details as a String.
        String geofenceTransitionDetails = getGeofenceTransitionDetails(
                this,
                geofenceTransition,
                triggeringGeofences
        );

        // Send notification and log the transition details.
        sendNotification(geofenceTransitionDetails);
        Log.i(TAG, geofenceTransitionDetails);
    } else {
        // Log the error.
        Log.e(TAG, getString(R.string.geofence_transition_invalid_type,
                geofenceTransition));
    }
  }
}
```

Stop geofence monitoring

Stopping geofence monitoring when it is no longer needed or desired can help save battery power and CPU cycles on the device. You can stop geofence monitoring in the main activity used to add and remove geofences; removing a geofence stops it immediately. The API provides methods to remove geofences either by request IDs, or by removing geofences associated with a given Pending Intent.

The following snippet removes geofences by Pending Intent, stopping all further notification when the device enters or exits previously added geofences:

```
geofencingClient.removeGeofences(getGeofencePendingIntent())
        .addOnSuccessListener(this, new OnSuccessListener<Void>() {
            @Override
            public void onSuccess(Void aVoid) {
                // Geofences removed
                // ...
            }
        })
        .addOnFailureListener(this, new OnFailureListener() {
            @Override
            public void onFailure(@NonNull Exception e) {
                // Failed to remove geofences
                // ...
            }
        });
```

6.4 METHODS INPUT AND OUTPUT PARAMETERS

Input Parameters

User Registration, User login, User name, User email, User password, User location, User mobile number.

Output Parameters

Tests have been carried out in various containment zones across Telangana for the validation of the Android application. The identified containment zones chosen for the testing of the application were visited one by one. Shows various containment zones identified for conducting the test, the date, time of entry, time of receiving the notification alerts upon entering. From, it is highlighted that the application sends notification alerts within 5–8 seconds on entering.

CHAPTER 7

7. PROJECT TESTING

7.1 VARIOUS TEST CASES

COVID-19 tracking tools or contact-tracing apps are getting developed at a rapid pace by different governments in their respective countries. This study explores one such tool called Aarogya setu, developed by the Government of India. It is a mobile application developed under the Health Ministry, as a part of the E-Governance initiative, to track and sensitize the citizens of India in a joint battle against COVID-19 spread. The study aims to understand various useful features of this tool and to present different concepts of data science applied within the application along with its importance in managing the ongoing pandemic. The App uses Bluetooth and GPS technologies to alert a user when they are nearby a COVID-19 infected person. The application uses various Data Science concepts such as Classification, Association Rule Mining, and Clustering to analyse COVID-19 spread in India. The study also shows potential upgradations in the application, which includes usage of Artificial Intelligence and Computer Vision to detect COVID-19 patients. The study would be useful for mobile technology professionals, data science professionals, medical practitioners, health-related frontline workers, public administrators, and government officials.

7.2 BLACK BOX TESTING

Black Box Testing is a software testing method in which the functionalities of software applications are tested without having knowledge of internal code structure, implementation details and internal paths. Black Box Testing mainly focuses on input and output of software applications and it is entirely based on software requirements and specifications. It is also known as Behavioural Testing.

Figure 7.1: Demonstration of black box testing

The above Black-Box can be any software system you want to test. For Example, an operating system like Windows, a website like Google, a database like Oracle or even your own custom application. Under Black Box Testing, you can test these applications by just focusing on the inputs and outputs without knowing their internal code implementation.

7.3 WHITE BOX TESTING

White Box Testing is software testing technique in which internal structure, design and coding of software are tested to verify flow of input-output and to improve design, usability and security. In white box testing, code is visible to testers so it is also called Clear box testing, Open box testing, Transparent box testing, Code-based testing and Glass box testing. It is one of two parts of the Box Testing approach to software testing. Its counterpart, Blackbox testing, involves testing from an external or end-user type perspective. On the other hand, White box testing in software engineering is based on the inner workings of an application and revolves around internal testing. The term "WhiteBox" was used because of the see-through box concept. The clear box or WhiteBox name symbolizes the ability to see through the software's outer shell (or "box") into its inner workings. Likewise, the "black box" in "Black Box Testing" symbolizes not being able to see the inner workings of the software so that only the end-user experience can be tested.

STEP 1) UNDERSTAND THE SOURCE CODE:

The first thing a tester will often do is learn and understand the source code of the application. Since white box testing involves the testing of the inner workings of an application, the tester must be very knowledgeable in the programming languages used in the applications they are testing. Also, the testing person must be highly aware of secure coding practices. Security is often one of the primary objectives of testing software. The tester should be able to find security issues and prevent attacks from hackers and naive users who might inject malicious code into the application either knowingly or unknowingly.

STEP 2) CREATE TEST CASES AND EXECUTE

The second basic step to white box testing involves testing the application's source code for proper flow and structure. One way is by writing more code to test the application's source code. The tester will develop little tests for each process or series of processes in the application. This method requires that the tester must have intimate knowledge of the code and is often done by the developer.

CHAPTER 8

8. OUTPUT SCREENS

8.1 USER INTERFACE

The app should have a user registration and login. After the user logged into the app it will track the user location and update the database with the current location. If the user is visiting the Containment Zone, he will get an alert notification.

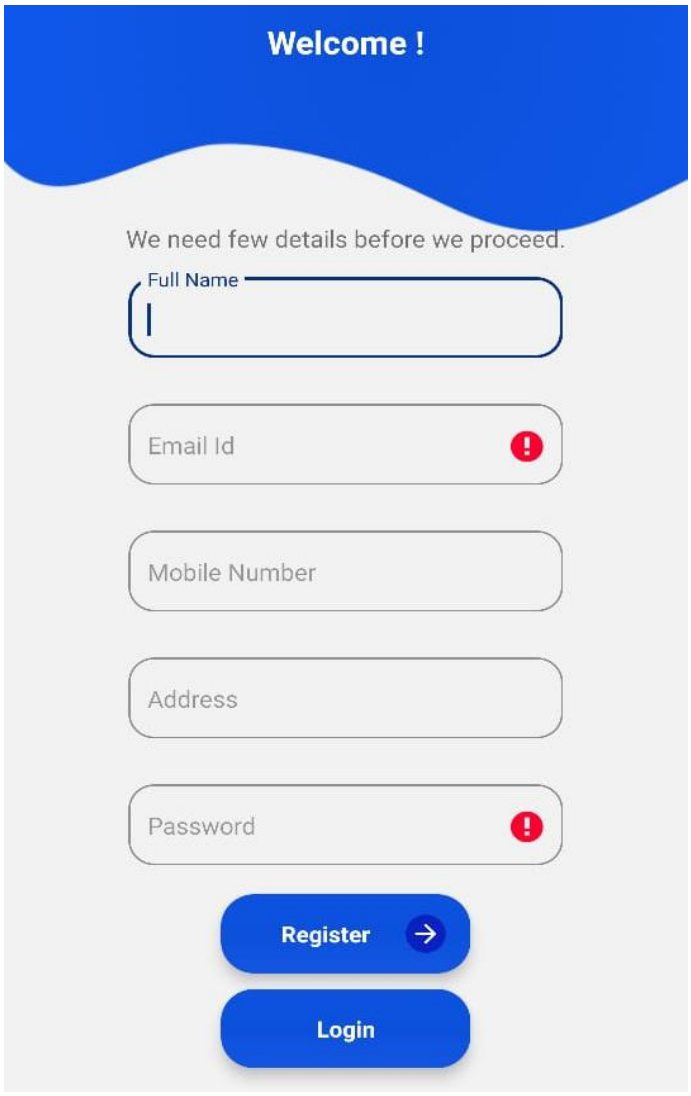


Figure 8.1: Registering Page interface

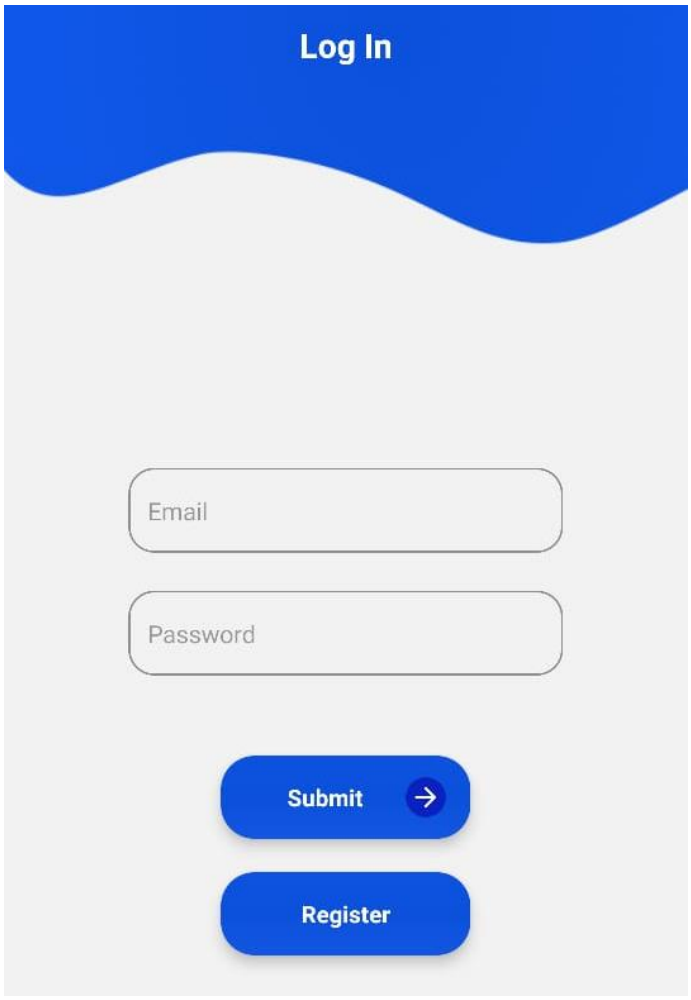


Figure 8.2: Login Page Interface

Figure 8.3: Location Page Interface

8.2 OUTPUT SCREENS

The app should have a user registration and login. After the user logged into the app it will track the user location and update the database with the current location. If the user is visiting the containment zone, he will get an alert notification

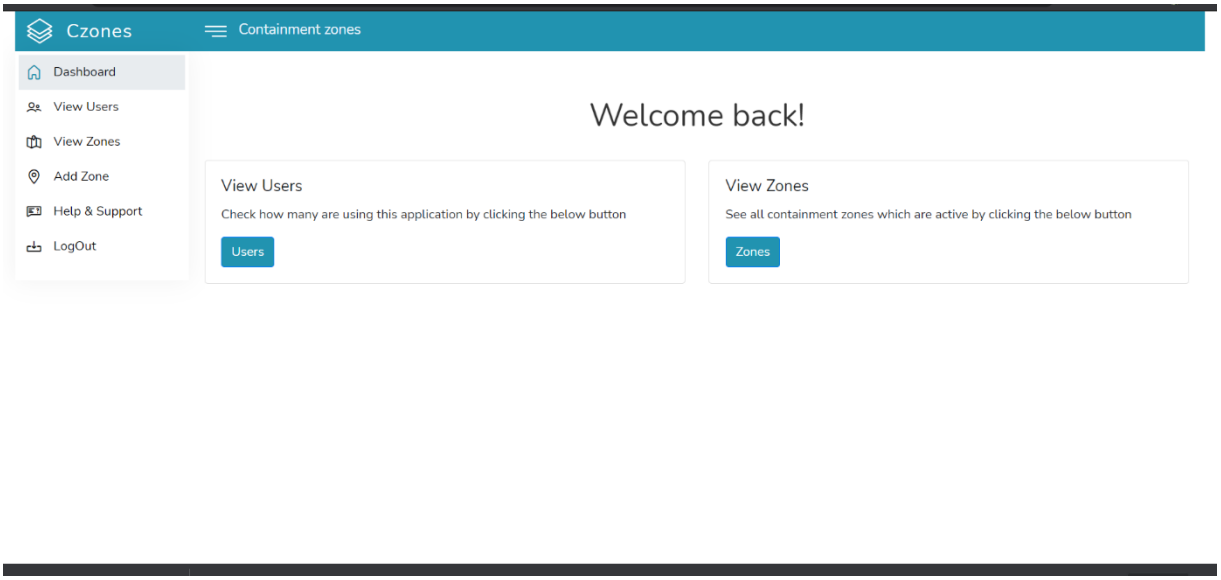


Figure 8.4: Welcome Page Interface at Admin

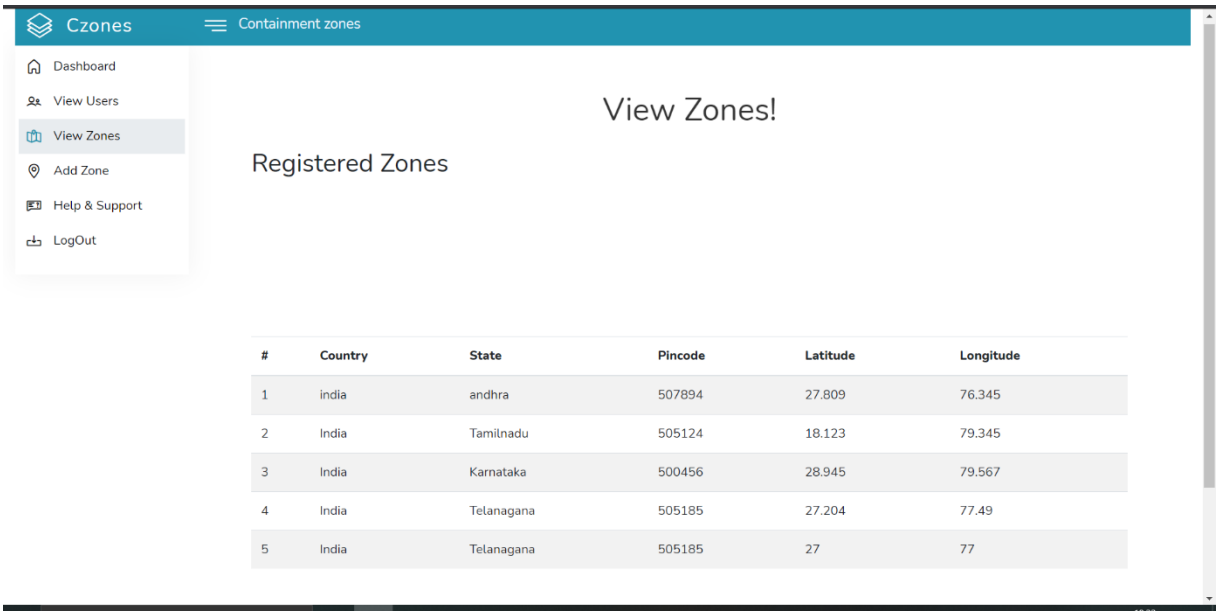


Figure 8.5: View Zone Page Interface

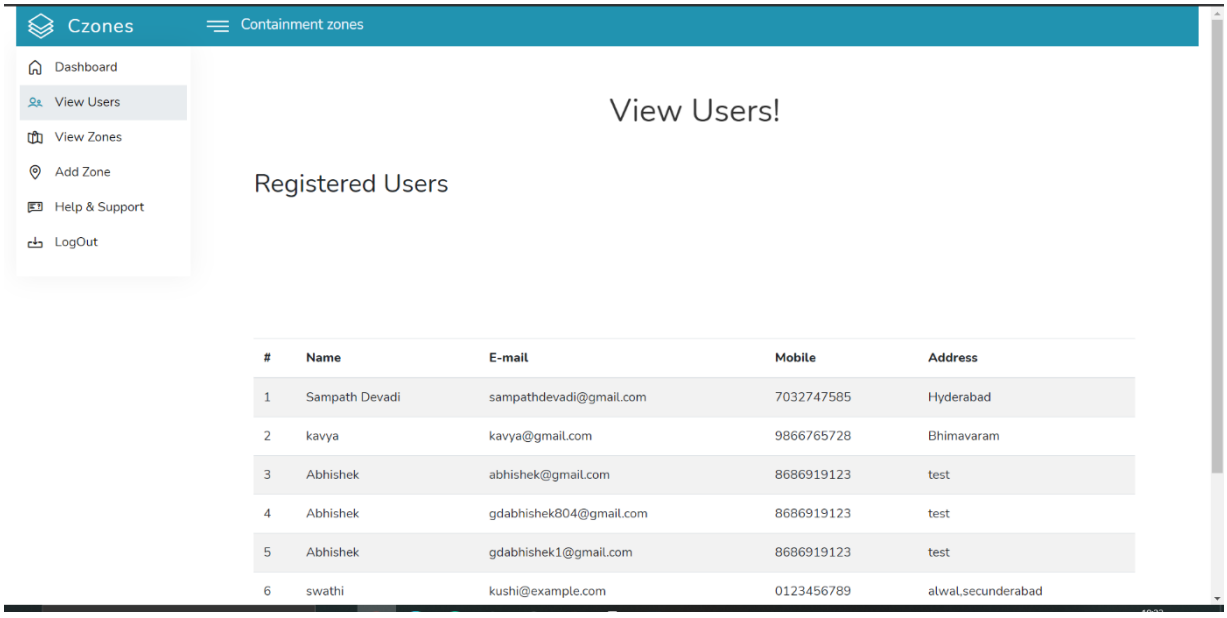


Figure 8.6: View Users Page Interface

CHAPTER 9

9. EXPERIMENTAL RESULTS

It is highlighted that the application sends notification alerts within 5–8 seconds on entering. It has been observed that the application receives around 23 KB of data during the start of the Maps-Activity when the application is initiated and receives around 10 KB when it has been running in the background. This data includes both APP-store data and Google maps data. The application provides an efficient way of showing the identified Covid-19 containment zones to the users in a Google map. This application further tracks the user's location and checks whether it is present in the list of identified containment zones. It sends separate notification alerts to the user on entering.

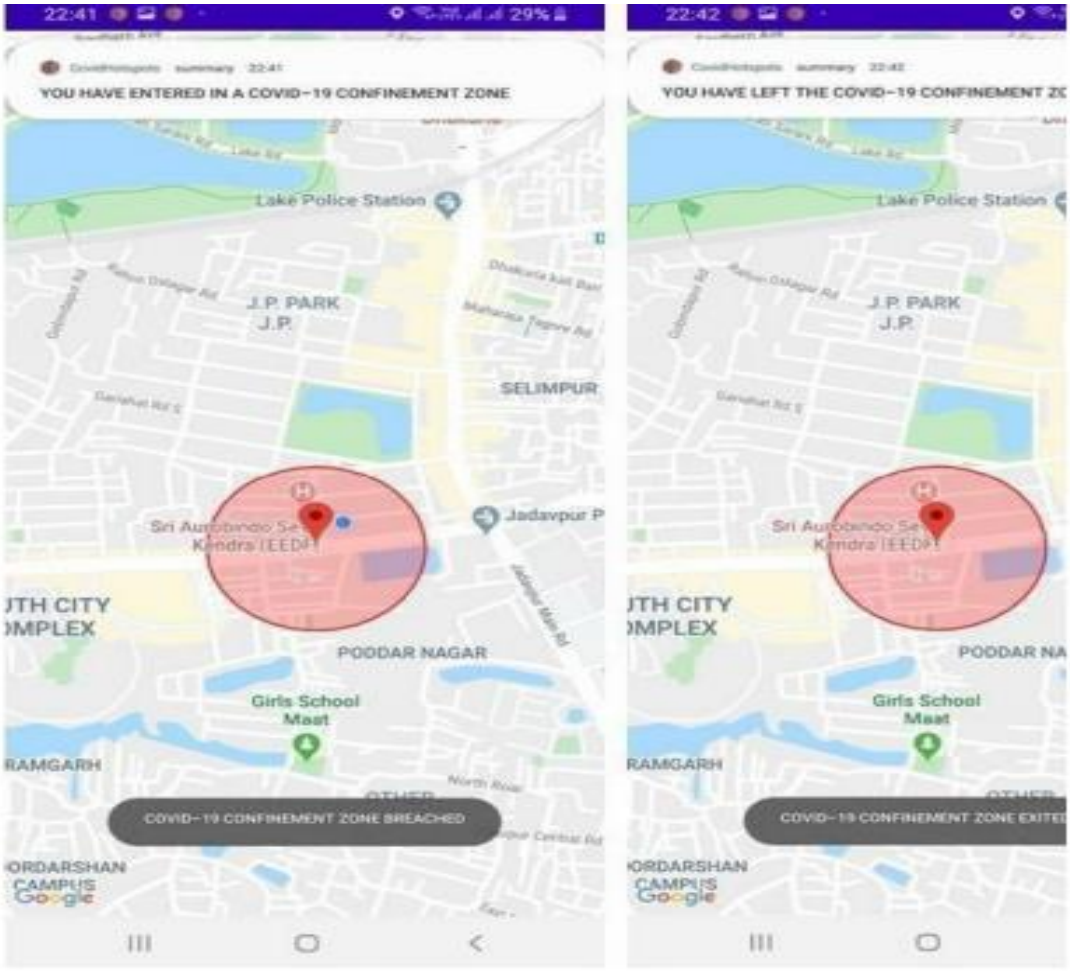


Figure 9.1: Geofencing Circle of Containment Zone

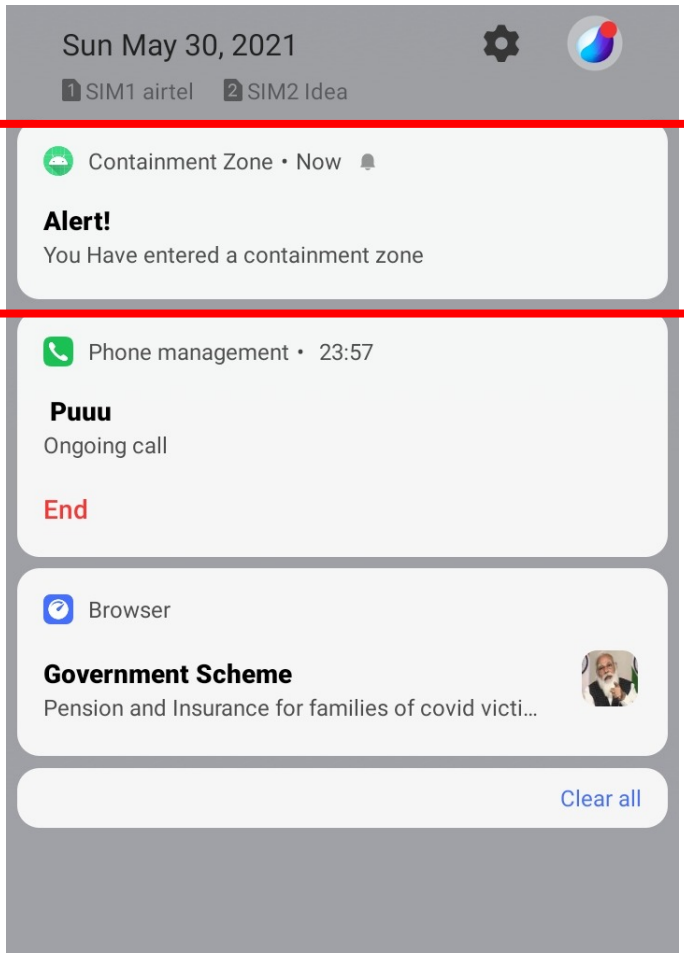


Figure 9.2: Notification Alert on Mobile

CHAPTER 10

10.1 CONCLUSION

The application provides an efficient way of showing the identified Covid-19 containment zones to the users in a Google map. With the alarming increase of Covid-19 affected cases throughout the world, this developed application can be employed as a tool for creating further social awareness among the people. The application further tracks the user's location and provides notification alert if the user has entered a containment zone. The application also provides daily Covid-19 case statistics to the users to keep them updated. The application is developed on Android SDK and uses to store the location data. Android's geofencing client is used to create geofences around the containment zones and notification manager is used to provide notifications. The Application has show great results. The notifications are sent with in a second. In the worst case possibility the application takes 5-8 seconds to send messages.

10.2 FUTURE ENHANCEMENT

The developed android application further extracts the IMEI Number of the trespasser in the containment zones which can be useful to the local police to track and identify people who are frequently trespassing the containment zones. Thereby this application identifies the containment zones and highlights the need for taking further precautionary measures for combating Covid-19. The application has been tested in various locations and has been found to yield good results. The application can be further used for many purposes like maritime and forest safety to prevent users from entering restricted areas.

REFERENCES

[1] M. Cascella, M. Rajnik, A. Cuomo, S. C. Dulebohn, and R. Di Napoli, ‘‘Features, evaluation and treatment coronavirus (COVID-19) [updated 2020 Apr 6],’’ in StatPearls [Internet]. Treasure Island, FL, USA: StatPearls Publishing, Jan. 2020. [Online]. Available: https:// www.ncbi.nlm.nih.gov/books/NBK554776/

[2] World Health Organization. Coronavirus disease (COVID-19) Pandemic. Accessed: Apr. 30, 2020. [Online]. Available: https:// www.who.int/emergencies/diseases/novel-coronavirus-2019

[3] WHO. (Apr. 2020). (WHO Situation Report 101). [Online]. Available: https://www.who.int/docs/defaultsource/coronaviruse/situation- reports/%20200430-sitrep-101-covid-19.pdf?sfvrsn=2ba4e093_2

[4] T. Singhal, ‘‘A review of coronavirus disease-2019 (COVID-19),’’ Indian J. Pediatrics, vol. 87, no. 4, pp. 281–286, Apr. 2020

[5] Wawrzyniak and T. Hyla, "Application of Geofencing Technology for the Purpose of Spatial Analyses in Inland Mobile Navigation," *2016 Baltic Geodetic Congress (BGC Geomatics)*, Gdansk, 2016, pp. 34-39. https://doi.org/10.1109/BGC.Geomatics.2016.15

[6] Cloud Firestore Data model https://firebase.google.com/docs/firestore/data-model.

[7] Namiot, “Geofence services”, International Journal of Open Information Technologies 9/2013.

[8] Kupper, U. Bareth, B. Freese, “Geofencing and background tracking–the next features in LBSs”, Proceedings of 41th annual conference on Geselltschaft fur Informatics, 2011.

[9] Freepik. *Timeline Flat Design Infographic—Designed by Freepik*. Accessed: Apr. 8, 2020. [Online]. Available: https://www.freepik.com/ free-vector/timeline-flat-design-inforaphic_628%2049.htm#query=timeline&position=49

C. Sohrabi, Z. Alsafi, N. O’Neill, M. Khan, A. Kerwan, A. Al-Jabir, C. Iosifidis, and R. Agha, ‘‘World health organization declares global emergency: A review of the 2019 novel coronavirus (COVID-19),’’ *Int. J. Surgery*, vol. 76, pp. 71–76, Apr. 2020.

[10] WHO. (Feb. 2020). *Report WHO-China Joint Mission Coronavirus Disease 2019 (COVID-19)*. [Online]. Available: https://www.who.int/docs/default-source/coronaviruse/who-china-joint-mi%ssion-on-covid-19-final-report.pdf

[11] People Who Are at Higher Risk for Severe Illness. (Apr. 2020). *Centers for Disease*

Control Prevention (CDC). [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/need-extra- precautions/people%-at-higher-risk.html

[11] WHO. (Feb. 2020). Report WHO-China Joint Mission Coronavirus Disease 2019 (COVID-19). [Online]. Available: https://www.who.int/docs/default-source/coronaviruse/who-chinajoint-mi%ssion-on-covid-19-final-report.pdf

[12] People Who Are at Higher Risk for Severe Illness. (Apr. 2020). Centers for Disease Control Prevention (CDC). [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/need-extraprecautions/people%-at-higher-risk.htm

[13] World Health Organization. Modes of Transmission of Virus Causing COVID-19: Implications for IPC Precaution Recommendations. Accessed: Apr. 20, 2020. [Online]. Available: https://www.who.int/newsroom/commentaries/detail/modes-of-transmission%-of-virus-causingcovid-19-implications-for-ipc-precaution-recommendations

[14] National Institutes Health. (Mar. 2020). Study Suggests New Coronavirus May Remain on Surfaces for Days. [Online]. Available: https://www.nih.gov/news-events/nih-research-matters/study-suggestsnew%-coronavirus-may-remain-surfaces-days

[15] P. Belluck, ‘‘What does the coronavirus do to the body?’’ The New York Times, Mar. 2020. [Online]. Available: https://www.nytimes.com/article/ coronavirus-body-symptoms.html?searchRe%sultPosition=10

[16] L. Fang, G. Karakiulakis, and M. Roth, ‘‘Are patients with hypertension and diabetes mellitus at increased risk for COVID-19 infection?’’ Lancet. Respiratory Med., vol. 8, no. 4, p. e21, 2020.

[17] S. H. Wong, R. N. S. Lui, and J. J. Y. Sung, ‘‘Covid-19 and the digestive system,’’ J. Gastroenterol. Hepatol., 2020.

[18] R. Baldwin and E. Tomiura, ‘‘Thinking ahead about the trade impact of COVID-19,’’ Economics in the Time COVID-19, 2020, p. 59.

[19] V. Surveillances, ‘‘The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (COVID-19) China, 2020,’’ China CDC Weekly, vol. 2, no. 8, pp. 113–122, 2020.

[20] H. Chen, J. Guo, C. Wang, F. Luo, X. Yu, W. Zhang, J. Li, D. Zhao, D. Xu, Q. Gong, J. Liao, H. Yang, W. Hou, and Y. Zhang, ‘‘Clinical characteristics and intrauterine vertical transmission potential of COVID-19 infection in nine pregnant women: A retrospective review of medical records,’’ Lancet, vol. 395, no. 10226, pp. 809–815, Mar. 2020.

PUBLICATIONS

This project has been presented by us in the International Conference named *Online Mega International Conference "*Innovations in Computers Networks, Computational Intelligence and IOT" (ICICCI-21)*" On 25th & 26th June, 2021.*

Devadi Sampath Rao is currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. He completed her intermediate from Sri Chaitanya Junior Kalashala and 10th class from TVR Model High School. He is one of the members of Cyber Security Club in our college. His responsibilities in that group include taking up the Network Security Challenges as a serious hobby. His technical skills include Python and Java. He also has a basic understanding of C and C++. He took part in Employability Skill development Program conducted by Zensar. He is also a student of Smart Interviews. Her participations include: National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020, Faculty Development Program Workshop on "AI and ML". He also participated as Intern in 8-week Internship in "Python and ML" in GoalStreet Internships. He has done projects such as "Real Time Chat Application", "Sentiment Analysis Project", and currently working on "AI Chatbot for Banking sector". His areas of interest are Python, Artificial Intelligence, Machine Learning and Full Stack Development. He completed few certification courses from online platforms like Coursera, HackerRank and SoloLearn. He also has Seven-star rating in Problem Solving in HackerRank and Two-star Rating in Codecheff.

Anil Kumar Reddy is currently pursuing her Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. He completed her intermediate from Narayana Junior College and 10th class from Vivekananda School. He is one of the members of Coders Club in our college. His responsibilities in that group include mentoring and motivating students to take coding as a serious hobby. His technical skills include C, Python and Java. She also has a basic understanding of C++. He participations include: National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020, "Know More - Teach More ", the Global Webinar on Cloud & Big Data – Changing the way we work which was conducted Global Education & Careers Forum (GECF) on 12th August 2020, "Know More - Teach More ", the Global Webinar on Cyber Threats and Defence Techniques conducted by GECF on 22nd July 2020, "One Day Webinar on Internet of Things and Its Applications" conducted by Anand Institute of Higher Technology on 21st May 2020 and IIC Online Sessions conducted by Institution's Innovation Council (IIC) of MHRD's Innovation Cell, New Delhi to promote Innovation, IPR, Entrepreneurship, and Start-ups among HEIs from 28th

April to 22nd May 2020. His areas of interest are Python, Artificial Intelligence, Machine Learning and Deep Learning. He completed few certification courses from online platforms like Coursera, CursaApp and SoloLearn.

G Vijay chandu is currently pursuing her Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. His completed her intermediate from Narayana Junior College and 10th class from Geetanjali model School. He also has a C language certificate from SSSIT institution and also has java certificate. Her areas of interest are Python, Artificial Intelligence, Machine Learning and Deep Learning. He completed few certification courses from online platforms like Coursera, CursaApp and SoloLearn.

Jansi Lingapuram Neha is currently pursuing her Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. she completed her intermediate from Sri Chaitanya Junior College and 10th class from Sri Chaitanya Techno School. She is one of the members of Coders Club in our college. Her participations include: Ongoing industrial Live Project Training with "Techkriti IIT-Kanpur in collaboration with 1stop" conducted from June 15th-16th 2020 , participated in certified for a course called "Getting started with python" on course era in December 20th-26th 2020 , participated actively in a virtual workshop called " women in cyber security and privacy in 2020" conducted in 20th-23th October 2020 and Undertook virtual Training at "Smart interviews institute " conducted three months 2020 and participated actively in a started National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020. Her areas of interest are Python, Artificial Intelligence, Machine Learning and Deep Learning. She completed few certification courses from online platforms like Coursera, CursaApp and SoloLearn.

A

PROJECT REPORT

On

HEART DISEASE PREDICTION

Submitted by

Mr. P. Rohith(17K81A05P7) **Mr. G. Aravind (17K81A05L2)**

Ms. D. Madhuritha(17K81A05K6) **Mr. S. Sandeep Sai(17K81A05P9)**

in partial fulfillment for the award of the

degree of

BACHELOR OF TECHNOLOGY

IN

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Under the Guidance of

Dr. K. Srinivas

Associate Professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ST. MARTIN'S ENGINEERING COLLEGE
An Autonomous Institute

Dhulapally, Secunderabad – 500 100

JUNE 2021

BONAFIDE CERTIFICATE

This is to certify that the project entitled **Heart Disease Prediction**, is being submitted by **P. Rohith(17K81A05P7), G. Aravind(17K81A05L2), D. Madhuritha(17K81A05K6), S. Sandeep Sai(17K81A05P9)** in partial fulfillment of the requirement for the award of the degree of **BACHELOR OF TECHNOLOGY** in **Computer Science of Engineering** is recorded of bonafide work carried out by them. The result embodied in this report have been verified and found satisfactory.

.

Guide

Dr. K. SRINIVAS

Department of CSE

Head of the Department

Dr. M. NARAYANAN

Department of CSE

Internal Examiner

External Examiner

Place:

Date:

DECLARATION

We, the student of **Bachelor of Technology** in Department of Computer Science and Engineering', session: 2017 – 2021, St. Martin's Engineering College, Dhulapally, Kompally, Secunderabad, hereby declare that work presented in this Project Work entitled Heart Disease Prediction is the outcome of our own bonafide work and is correct to the best of our knowledge and this work has been undertaken taking care of Engineering Ethics. This result embodied in this project report has not been submitted in any university for award of any degree.

P. Rohith	17K81A05P7
G. Aravind	17K81A05L2
D. Madhuritha	17K81A05K6
S. Sandeep Sai	17K81A05P9

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible and whose encouragements and guidance have crowded effects with success.

We extended our deep sense of gratitude to Principal**,** **Dr. P. SANTOSH KUMAR PATRA**, St. Martin's Engineering College, Dhulapally, for permitting us to undertake this project.

We are also thankful to **Dr. M. NARAYANAN**, Head of the Department, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, forhis support and guidance throughout our project.

We would like to thank **Dr. T. POONGOTHAI,** Department of Computer Science and Engineering (AI & ML) for her encouragement and insightful comments and as well as our project coordinators **Dr. B. RAJALINGAM**, Associate Professor and **Dr. GOVINDA RAJULU. G** Associate Professor, in Department of Computer Science and Engineering for their valuable support.

We would like to express our sincere gratitude and indebtedness to our projectsupervisor **Dr. SRINIVAS. K** Associate Professor, Computer Science andEngineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

Finally, we express thanks to all those who have helped us successfully to completing this project. Furthermore, we would like to thank our family and friends for their moral support and encouragement.

We express thanks to all those who have helped us in successfully completing the project.

P. Rohith 17K81A05P7
G. Aravind 17K81A05L2
D. Madhuritha 17K81A05K6
S. Sandeep Sai 17K81A05P9

ABSTRACT

Heart related diseases or cardiovascular diseases (CVDs) are the main reason for a huge number of deaths in the world over the last few decades and has emerged as the most life-threatening disease, not only in India but in the whole world. So, there is a need of reliable, accurate and feasible system to diagnose such diseases in time for proper treatment. Machine Learning algorithms and techniques have been applied to various medical datasets to automate the analysis of large and complex data. Many researchers, in recent times, have been using several machine learning techniques to help the health care industry and the professionals in the diagnosis of heart related diseases. This paper presents a survey of various models based on such algorithms and techniques and analyze their performance. Models based on supervised learning algorithms such as Support Vector Machines (SVM), K-Nearest Neighbour (KNN), Naïve Bayes, Decision Trees (DT), Random Forest (RF) and ensemble models are found very popular among the researchers.

TABLE OF CONTENTS

LIST OF FIGURES

Table 1. List of Figures

LIST OF OUTPUT SCREENS

Table 2. List of Output screens

LIST OF ACRONYMS

CPU	Central Processing Unit
GB	Giga Bytes
OS	Operating System
GUI	Graphical User Interface
CVD	Cardiovascular Diseases
AUT	Application Under Test
ER	Entity Relationship
SDLC	Software Development Life Cycle
UML	Unified Modeling Language
RAM	Random Access Memory
SUT	System Under Test
PDF	Portable Document Format
OOPS	Object-Oriented Programming System

Table 3. List of Acronyms

1.INTRODUCTION

Heart is an important organ of the human body. It pumps blood to every part of our anatomy. If it fails to function correctly, then the brain and various other organs will stop working, and within few minutes, the person will die. Change in lifestyle, work related stress and bad food habits contribute to the increase in rate of several heart related diseases. Heart diseases have emerged as one of the most prominent cause of death all around the world. According to World Health Organization, heart related diseases are responsible for the taking 17.7 million lives every year, 31% of all global deaths. In India too, heart related diseases have become the leading cause of mortality [1]. Heart diseases have killed 1.7 million Indians in 2016, according to the 2016 Global Burden of Disease Report, released on September 15,2017. Heart related diseases increase the spending on health care and also reduce the productivity of an individual. Estimates made by the World Health Organization (WHO), suggest that India have lost up to $237 billion, from 2005-2015, due to heart related or cardiovascular diseases [2]. Thus, feasible and accurate prediction of heart related diseases is very important. Medical organizations, all around the world, collect data on various health related issues. These data can be exploited using various machine learning techniques to gain useful insights. But the data collected is very massive and, many a times, this data can be very noisy. These datasets, which are too overwhelming for human minds to comprehend, can be easily explored using various machine learning techniques. Thus, these algorithms have become very useful, in recent times, to predict the presence or absence of heart related diseases accurately.

1.1. PROJECT OVERVIEW

Heart Disease Prediction system is used to predict whether the person is suffering from heart diseases or not in very easy manner. It produces the great effort to deal to remove the barriers of biasness. The system can discover and extract hidden knowledge associated with diseases from a historical heart data set heart disease prediction system aims to exploit data mining techniques on medical data set to assist in the prediction of the heart diseases.

1.2. PROJECT OBJECTIVE

Heart Disease Prediction provides new approach to concealed patterns in the data. The objective of the project is that integration of clinical decision support with computer-based patient records could reduce medical errors, enhance patient safety, decrease unwanted practice variation, and improve patient outcome. It helps to avoid human biasness. To implement Naïve Bayes Classifier that classifies the disease as per the input of the user. This system also reduces the cost of medical tests.

1.3. ORGANIZATION OF CHAPTERS

This documentation consists of 10 different chapter and they are:

1. Introduction – This chapter covers the overview of our project and its objectives.

2. Literature Survey – This includes the details of our survey.

3. Software and Hardware Requirements – We specify our software and hardware requirements here.

4. Software Development Analysis – This section includes the problem definition and details of the modules we used in our project.

5. Project System Design – This chapter includes the design part of our project which includes uml diagrams.

6. Project Coding – This section contains the details of our project code.

7. Project Testing – The details of test cases and testing are included in this chapter.

8. Output Screens – This contains the screenshots of how our project looks like when executed.

9. Experimental Results – This chapter contains the screenshots of our results.

10. Conclusion and Future Enhancements – This covers the conclusion of our project and the possible future developments.

2.LITERATURE SURVEY

A literature survey or a literature review in a project report is that section which shows the various analysis and research made in the field of your interest and the results already published, considering the various parameters of the project and the extent of the project.

It is the most important part of our report as it gave us a direction in our research. It helped us set a goal for our analysis - thus giving us our problem statement.

2.1 SURVEY ON BACKGROUND

1. Responding to the threat of chronic diseases in India

AUTHORS: Bela Shah, Cherian Varghese, Anbumani Ramadoss

At the present stage of India's health transition, chronic diseases contribute to an estimated 53% of deaths and 44% of disability-adjusted life-years lost. Cardiovascular diseases and diabetes are highly prevalent in urban areas. Demographic and socioeconomic factors are hastening the health transition, with sharp escalation of chronic disease burdens expected over the next 20 years. A national cancer control programme, initiated in 1975, has established 13 registries and increased the capacity for treatment. A comprehensive law for tobacco control was enacted in 2003. An integrated national programme for the prevention and control of cardiovascular diseases and diabetes is under development. There is a need to increase resource allocation, coordinate multisectoral policy interventions, and enhance the engagement of the health system in activities related to chronic disease prevention and control.

2. Study of Machine Learning Algorithms for Special Disease Prediction.

AUTHORS: Dhomse Kanchan B, Mahale Kishor M

The worldwide study on causes of death due to heart disease/syndrome has been observed that it is the major cause of death. If recent trends are allowed to continue, 23.6 million people will die from heart disease in coming 2030. The healthcare industry collects large amounts of heart disease data which unfortunately are not “mined” to discover hidden information for effective decision making. In this paper, study of PCA has been done which finds the minimum number of attributes required to enhance the precision of various supervised machine learning algorithms. The purpose of this research is to study supervised machine learning algorithms to predict heart disease. Data mining has number of important techniques like categorization, preprocessing. Diabetic is a life threatening disease which prevent in several urbanized as well as emergent countries like India. The data categorization is diabetic patients datasets which is developed by collecting data from hospital repository consists of 1865 instances with dissimilar attributes..

3. Cardiovascular risk prediction method based on CFS subset evaluation and random forest classification framework

AUTHORS: Shan Xu; Zhen Zhang; Daoxian Wang; Junfeng Hu; Xiaohui Duan; Tiangang Zhu

Cardiovascular Disease (CVD) is a highly significant contributor to loss of quality and quantity of life all over the world. Early detection and risk prediction is very important for patients' treatment and doctors' diagnose. This paper focus on establishing a more accurate and practical risk prediction system based on data mining techniques to provide auxiliary medical service. In order to be practically used for collecting and analyzing patients' data in healthcare industries, the system consists of four parts: data interface, data preparation, feature selection and classification. Data interface response to obtain hospitals' raw data from hospital; data preprocessing is needed for data integration, data cleaning and rating mapping etc. Key features were then selected by CFS Subset Evaluation combined with Best-First-Search method to reduce dimensionality. Random forest was inducted as basic classifier to identify risk level, which is a prior trial in CVD risk prediction field. Cleveland Heart-Disease Database (CHDD) and Cardiology inpatient dataset of PKU People's Hospital were both tested to confirm accuracy as well as practicality. In CHDD test, our system has a significantly higher accuracy of 91.6% than other methods. In People's Hospital dataset test, it achieves an accuracy of 97%, which is better than most of other classifiers except SVM (98.9%), however random forest only take half of time than SVM. Comprehensively considering the risk prediction system shows great significance in accuracy and practical use for patients' treatment and doctors' diagnose.

4. Prediction of heart disease using hybrid technique for selecting features

AUTHOR: Kanika Pahwa; Ravinder Kumar.

Generally Healthcare industry is known to be `information rich', but woefully all the data required to discover hidden patterns are not mined. For effective decision making in field of medical, advanced techniques of data mining are used. This paper proposed a prediction of heart disease using random forest and naive bayes. In addition, approach is proposed to select features before classification in order to improve performance of models. For feature selection, SVM-RFE and gain ratio algorithms are applied to dataset which in results assigns weight to each feature. This approach helps to improve accuracy and reduce computational time. Experimental results shows that proposed approach of selecting feature increases accuracy for both models.

5. Utilizing ECG-Based Heartbeat Classification for Hypertrophic Cardiomyopathy Identification.

AUTHORS: Sima, D., Schmuck, B., Szöll″osi, S., & Miklós, Á.

Hypertrophic cardiomyopathy (HCM) is a cardiovascular disease where the heart muscle is partially thickened and blood flow is (potentially fatally) obstructed. A test based on electrocardiograms (ECG) that record the heart electrical activity can help in early detection of HCM patients. This paper presents a cardiovascular-patient classifier we developed to identify HCM patients using standard 10-second, 12-lead ECG signals. Patients are classified as having HCM if the majority of their recorded heartbeats are recognized as characteristic of HCM. Thus, the classifier's underlying task is to recognize individual heartbeats segmented from 12-lead ECG signals as HCM beats, where heartbeats from non-HCM cardiovascular patients are used as controls. We extracted 504 morphological and temporal features—both commonly used and newly-developed ones—from ECG signals for heartbeat classification. To assess classification performance, we trained and tested a random forest classifier and a support vector machine classifier using 5-fold cross validation. We also compared the performance of these two classifiers to that obtained by a logistic regression classifier, and the first two methods performed better than logistic regression.

2.2. CONCLUSION ON SURVEY

In this paper, we surveyed on two things: the first part of the study is finding important factors affecting the heart disease, the important attributes and their minimum support values for no heart disease. Then use of machine learning algorithm for prediction of heart disease. From risk factors we have selected number of attributes and their minimum value for normal and diseased person. Values of the attributes more than the minimum value means you have a risk of heart disease. Second part of the studies of supervised machine algorithms for prediction of heart disease. In this various algorithm studied are Support vector machine, Decision tree, Random Forest, Linear regression and Naive Bayes classifier. Lots of work is done in this area. The dataset is quite old and has no new attributes added in it. There is no cleaning and pruning of data. Uncleaned and missing values in the dataset has no use for classification and prediction. Moreover, no one has worked on the size of the dataset. The small size of the data set is a problem for machine learning algorithms. Large size of the dataset is needed for better prediction.

3. SOFTWARE AND HARDWARE REQUIREMENT

Requirement is a condition or capability possessed by the software or system component in order to solve a real-world problem. The problems can be to automate a part of a system, to correct shortcomings of an existing system, to control a device, and so on.

Requirements describe how a system should act, appear or perform. For this, when users request for software, they provide an approximation of what the new system should be capable of doing. Requirements differ from one user to another and from one business process to another.

The purpose of the requirements document is to provide a basis for the mutual understanding between the users and the designers of the initial definition of the software development life cycle (SDLC) including the requirements, operating environment and development plan.

Requirements help to understand the behavior of a system, which is described by various tasks of the system. For example, some of the tasks of a system are to provide a response to input values, determine the state of data objects, and so on. Note that requirements are considered prior to the development of the software. The requirements, which are commonly considered, are classified into three categories, namely, functional requirements, non-functional requirements, and domain requirements.

The functional requirements should be complete and consistent. Completeness implies that all the user requirements are defined. Consistency implies that all requirements are specified clearly without any contradictory definition. Generally, it is observed that completeness and consistency cannot be achieved in large software or in a complex system due to the problems that arise while defining the functional requirements of these systems. The different needs of stakeholders also prevent the achievement of completeness and consistency. Due to these reasons, requirements may not be obvious when they are,'first specified and may further lead to inconsistencies in the requirements specification.

The non-functional requirements (also known as **quality requirements)** are related to system attributes such as reliability and response time. Non-functional requirements arise due to user requirements, budget constraints, organizational policies, and so on. These requirements are not related directly to any particular function provided by the system.

Non-functional requirements should be accomplished in software to make it perform efficiently. For example, if an aero plane is unable to fulfill reliability requirements, it is not approved for safe operation. Similarly, if a real time control system is ineffective in accomplishing non-functional requirements, the control functions cannot operate correctly.

System requirements are the configuration that a system must have in order for a hardware or software application to run smoothly and efficiently. Failure to meet these requirements can result in installation problems or performance problems. The former may prevent a device or application from getting installed, whereas the latter may cause a product to malfunction or perform below expectation or even to hang or crash.

System requirements are also known as minimum system requirements.

Hardware system requirements often specify the operating system version, processor type, memory size, available disk space and additional peripherals, if any, needed. Software system requirements, in addition to the requirements, may also specify additional software dependencies (e.g., libraries, driver version, framework version). Some hardware/software manufacturers provide an upgrade assistant program that users can download and run to determine whether their system meets a product's requirements.

Some products include both minimum and recommended system requirements. A video game, for instance, may function with the minimum required CPU and GPU, but it will perform better with the recommended hardware. A more powerful processor and graphics card may produce improved graphics and faster frame rates (FPS).
Some system requirements are not flexible, such as the operating system(s) and disk space required for software installation. Others, such as CPU, GPU, and RAM requirements may vary significantly between the minimum and recommended requirements. When buying or upgrading a software program, it is often wise to make sure your system has close to the recommended requirements to ensure a good user experience.

3.1 SOFTWARE REQUIREMENTS

- Operating System : Windows XP.
- Platform : PYTHON TECHNOLOGY
- Tool : Spyder, Python 3.5
- Front End : Anaconda
- Back End : Python Anaconda Script

3.2 HARDWARE REQUIREMENTS

- System : Pentium IV 2.4 GHz.
- Hard Disk : 40 GB.
- Monitor : 15inch VGA Color.
- Mouse : Logitech Mouse.
- Ram : 512 MB
- Keyboard : Standard Keyboard

4. SOFTWARE DEVELOPMENT ANALYSIS

Software development is a process of writing and maintaining the source code, but in a broader sense, it includes all that is involved between the conception of the desired software through to the final manifestation of the software, sometimes in a planned and structured process. Therefore, software development may include research, new development, prototyping, modification, reuse, re- engineering, maintenance, or any other activities that result in software products.

4.1. OVERVIEW OF PROBLEM

The Main disadvantage is the Prediction of cardiovascular disease results is not accurate. On the other-hand the Data mining techniques does not help to provide effective decision making**.** And it cannot handle enormous datasets for patient records.

4.2. DEFINE THE PROBLEM

Heart diseases have killed 1.7 million Indians in 2016, according to the 2016 Global Burden of Disease Report, released on September 15,2017. Heart related diseases increase the spending on health care and also reduce the productivity of an individual. Estimates made by the World Health Organization (WHO), suggest that India have lost up to $237 billion, from 2005-2015, due to heart related or cardiovascular diseases [2]. Thus, feasible and accurate prediction of heart related diseases is very important.

4.3. MODULES OVERVIEW

This project has one user i.e., Patient and four modules are designed for the interactions between user and application. Each Module has its own functionality. A module allows us to logically organize the code. Grouping related code into a module makes the code easier to understand and use.

4.4. DEFINE THE MODULES

This application has four modules which are listed in the following.

1. Home Page
2. Login Page
3. Registration Page
4. User Module

4.5 MODULE FUNCTIONALITY

In this project there are three modules to achieve our expected result. These are the major functionalities of the project. The registration and login process are important to access the project for user. There is only one user (person who wants to check).

1. Home Page

The patient opens the home page and check for options for registering and logging in. After opening the home page, the person could see abstract of the project, this gives him clear idea on what to do thereby knowing how the project works and what is it all about. The abstract had a brief of heart diseases and factors leading to heart diseases. The person can also see register page and log in page. After getting a brief about the project, now the person will now hover to registration page.

2. Register Page

After hovering to register page the person could see the registration form where he/she needs to fill in the details. The new user signup screen appears where the user should enter all their details. The details to be entered are username, password, contact number, email id, and address and then click on register button. The user is now registered and is ready to hover to next page which is log in page.

3. Login Page

Now the user enters to login page where he/she needs to enter the login credentials which were used during the registration process. After entering the login details the user is hovered to main page where the screen shows Welcome along with user name and abstract of the project is again mentioned here just to make sure that the user should know the process and gain knowledge about the heart diseases. Now the user can see two options named Prediction your heart condition and logout. To know about the heart disease the user has to click on Predict my heart condition and enter the thirteen important attributes which are required to predict the heart disease. The attributes namely are Age, Gender, Chest pain type, Resting blood pressure, Cholesterol, Fasting blood sugar, Resting electrocardiographic results, Max heartrate achieved, Exercise induced angina, Old peak, Slope peak exercise, Total major vessel, and Thal. After entering all these mentioned values, the user clicks on submit button. The details are submitted and analyzed the algorithms and result is obtained on the screen. The result shows whether the person is detected with heart disease or not. Now after getting to know the results, the user logs out of the page.

4. User Module

The main aim of this analysis is to develop a prototype Health Care Prediction System using, Naive Bayes. The System will discover and extract hidden data related to diseases (heart attack, cancer and diabetes) from a historical heart disease database. It will answer complicated queries for diagnosing sickness and so assist care practitioners to form intelligent clinical selections which ancient call support systems cannot. By providing effective treatments, it conjointly helps to reduce treatment prices. To reinforce visualization and easy interpretation, it displays the results in tabular and PDF forms.

5. PROJECT SYSTEM DESIGN

SYSTEM ARCHITECTURE

The architecture of the system disease prediction using machine learning consist of various datasets through which we will compare the symptoms of the user and predicts it, then the datasets are transformed into the smaller sets and from there it gets classified based on the classification algorithms later on the classified data is then processed into the machine learning technologies through which the data gets processed and goes in to the disease prediction model using all the inputs from the user that is mentioned above. Then after user entering the above information and overall processed data combines and compares in the prediction model of the system and finally predicts the disease. An architecture diagram is a graphical representation of a set of concepts, that are part of an architecture, including their principles, elements and components. The diagram explains about the system software in perception of overview of the system.

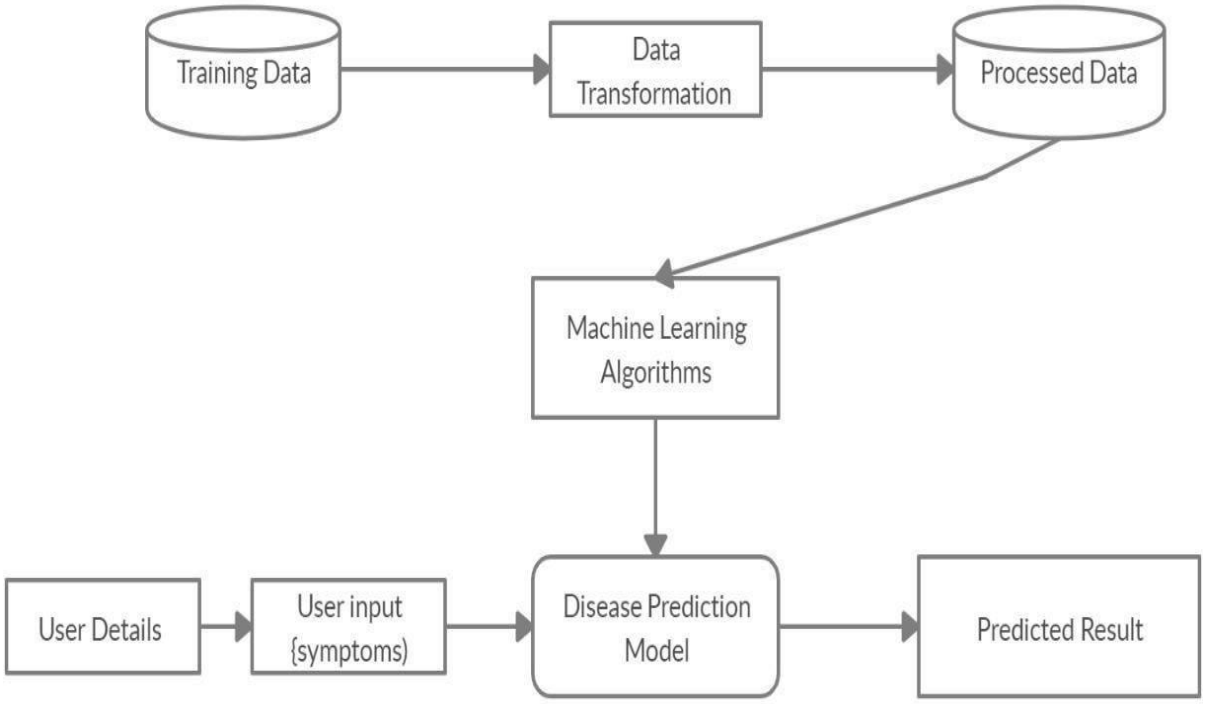


Fig 5.1 System Architecture

5.1 DATA FLOW DIAGRAMS
CONTEXT LEVEL DIAGRAM

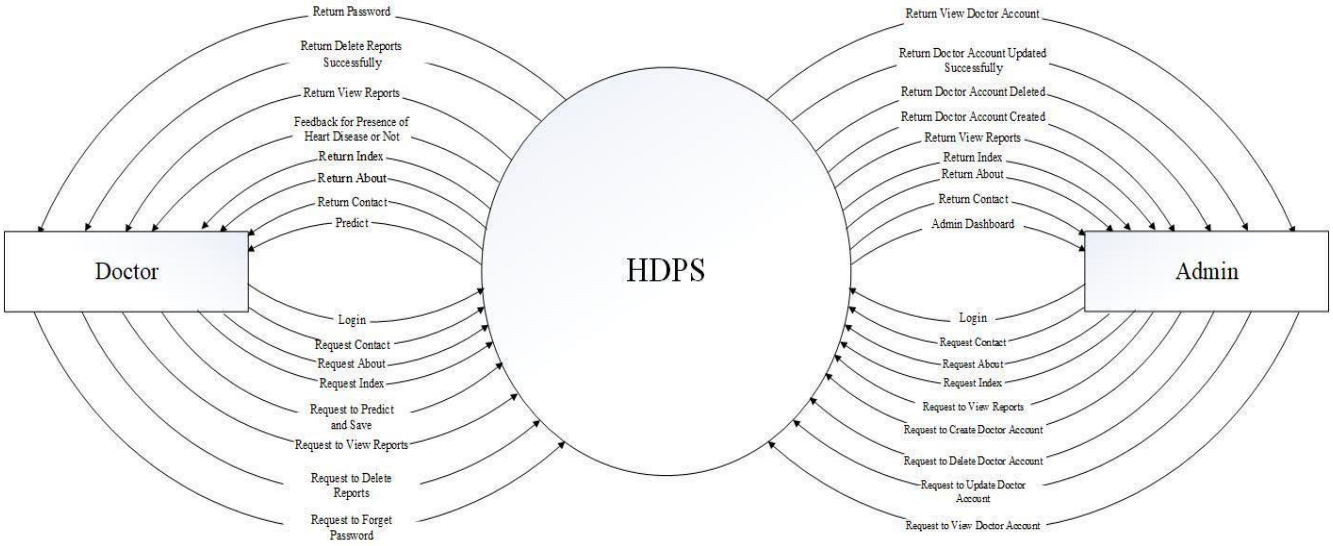


Fig 5.2 Context Level Diagram

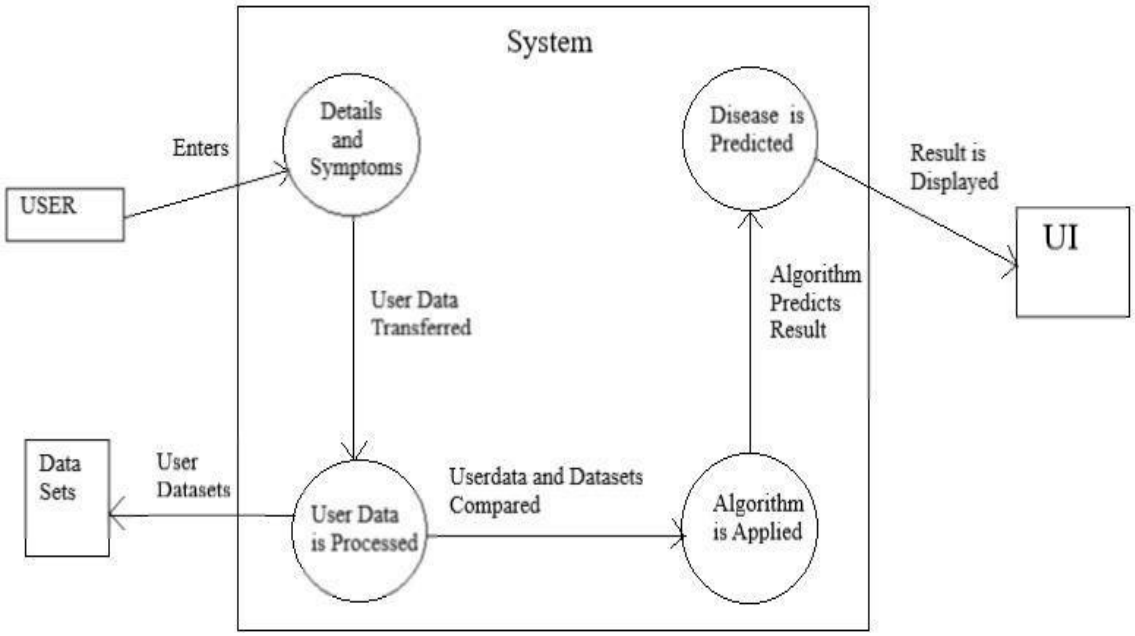


Fig 5.3 Data Flow Diagram

The dataflow diagram of the project disease prediction using machine learning consist of all the various aspects a normal flow diagram requires. This dataflow diagram shows how from starting the model flows from one step to another, like he enter into the system then enters all the information's and all other general information along with the symptoms that goes into the system, compares with the prediction model and if true is predicts the appropriate results otherwise it shows the details where the user if gone wrong while entering the information

5.2 ENTITY-RELATIONSHIP DIAGRAM

An entity relationship model is a high-level conceptual model which describes data in terms of entities, their attributes and their relationships (Riccardi, 2002). The entity relationship diagram shows how is represented and organized in the database schema without specifying the actual data (Pagh, 2006).

The system administrator has the user id attribute as the primary key. The relationship between the system administrator and the user is one to many. This shows that one system administrator can manage more than one entity user.

The user entity has username attribute as the primary key. The entities manager and authorized party borrow attributes from user. These borrowed attributes include name, username and password. The two have "ISA" relationship with the entity user. In addition to these attributes, authorized party entity has location and organization attributes.

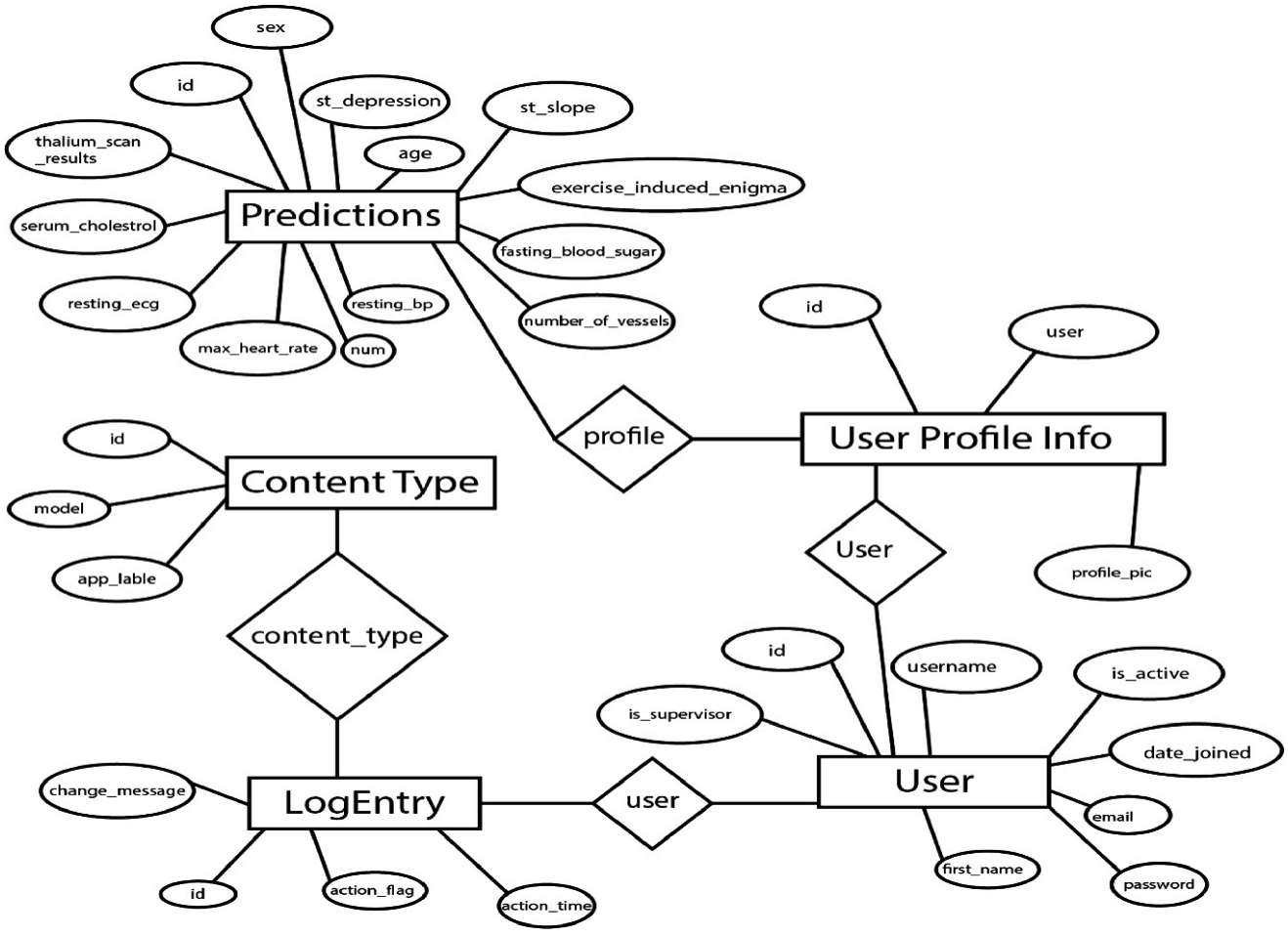


Fig 5.4 E-R Diagram

5.3 UML DIAGRAMS

UML stands for Unified Modeling Language. UML is a standardized general-purpose modeling language in the field of object-oriented software engineering. The standard is managed, and was created by, the Object Management Group.

The goal is for UML to become a common language for creating models of object-oriented computer software. In its current form UML is comprised of two major components: a Meta-model and a notation. In the future, some form of method or process may also be added to; or associated with, UML.

The Unified Modeling Language is a standard language for specifying, Visualization, Constructing and documenting the artifacts of software system, as well as for business modeling and other non-software systems.

The UML represents a collection of best engineering practices that have proven successful in the modeling of large and complex systems.

The UML is a very important part of developing objects-oriented software and the software development process. The UML uses mostly graphical notations to express the design of software projects.

USE CASE DIAGRAM:

The Use Case diagram of the project disease prediction using machine learning consist of all the various aspects a normal use case diagram requires. This use case diagram shows how from starting the model flows from one step to another, like he enters into the system then enters all the information's and all other general information along with the symptoms that goes into the system, compares with the prediction model and if true is predicts the appropriate results otherwise it shows the details where the user if gone wrong while entering the information's and it also shows the appropriate precautionary measure for the user to follow. Here the use case diagram of all the entities is linked to each other where the user gets started with the system.

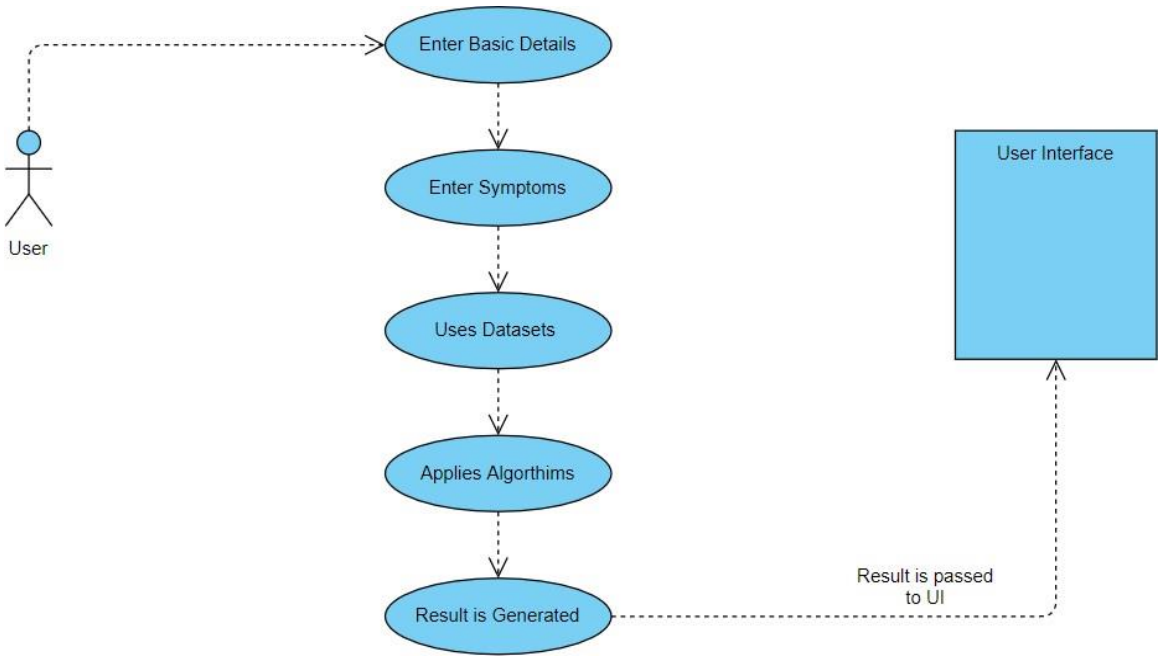


Fig. 5.5 Use Case Diagram

CLASS DIAGRAM:

Disease prediction using machine learning consist of class diagram that all the other application that consists the basic class diagram, here the class diagram is the basic entity that is required in order to carry on with the project. Class diagram consist information about all the classes that is used and all the related datasets, and all the other necessary attributes and their relationships with other entities, all these information is necessary in order to use the concept of the prediction, where the user will enter all necessary information such as user name, email, phone number, and many more attributes that is required in order to login into the system and using the files concept we will store the information of the users who are registering into the system and retrieves those information later while logging into the system.

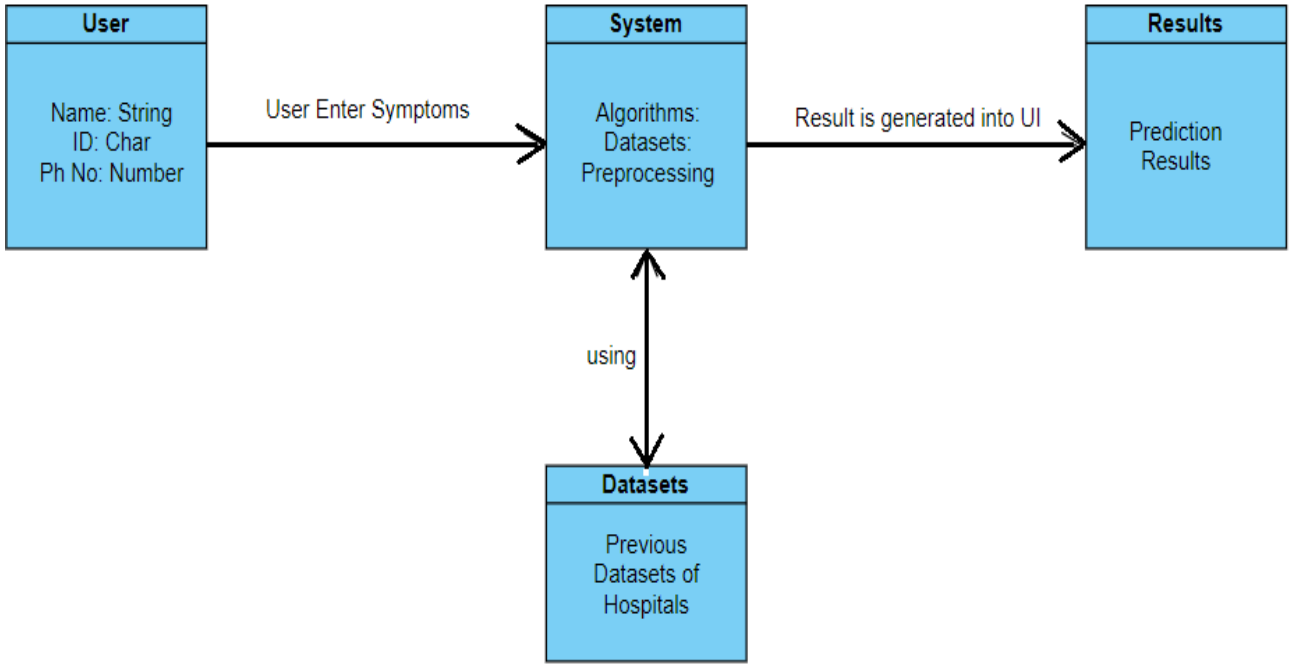


Fig. 5.6 Class Diagram

SEQUENCE DIAGRAM:

The Sequence diagram of the project disease prediction using machine learning consist of all the various aspects a normal sequence diagram requires. This sequence diagram shows how from starting the model flows from one step to another, like he enters into the system then enters all the information's and all other general information along with the symptoms that goes into the system, compares with the prediction model and if true is predicts the appropriate results otherwise it shows the details where the user if gone wrong while entering the information's and it also shows the appropriate precautionary measure for the user to follow. Here the sequence of all the entities is linked to each other where the user gets started with the system.

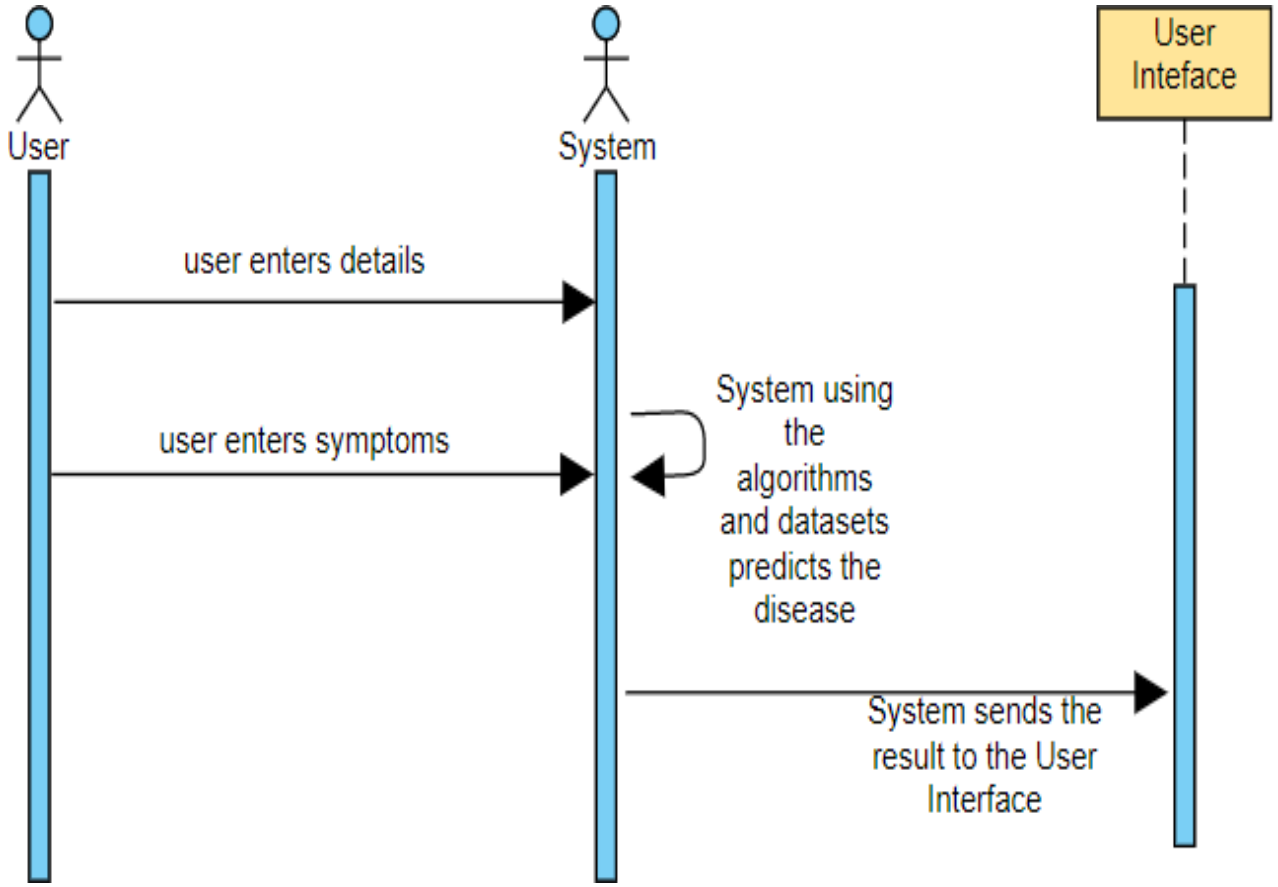


Fig. 5.7 Sequence Diagram

ACTIVITY DIAGRAM:

Activity diagram is another important diagram in UML to describe the dynamic aspects of the system. Activity diagram is basically a flowchart to represent the flow from one activity to another activity. The activity can be described as an operation of the system. The control flow is drawn from one operation to another. Here in this diagram the activity starts from user where the user registers into the system then login using the credentials and then the credentials are matched in the system and if it's true, then the user proceeds to the prediction phase where the prediction happens. Then finally after processing the data from datasets the analysis will happen then the correct result will be displayed that is nothing but the Output.

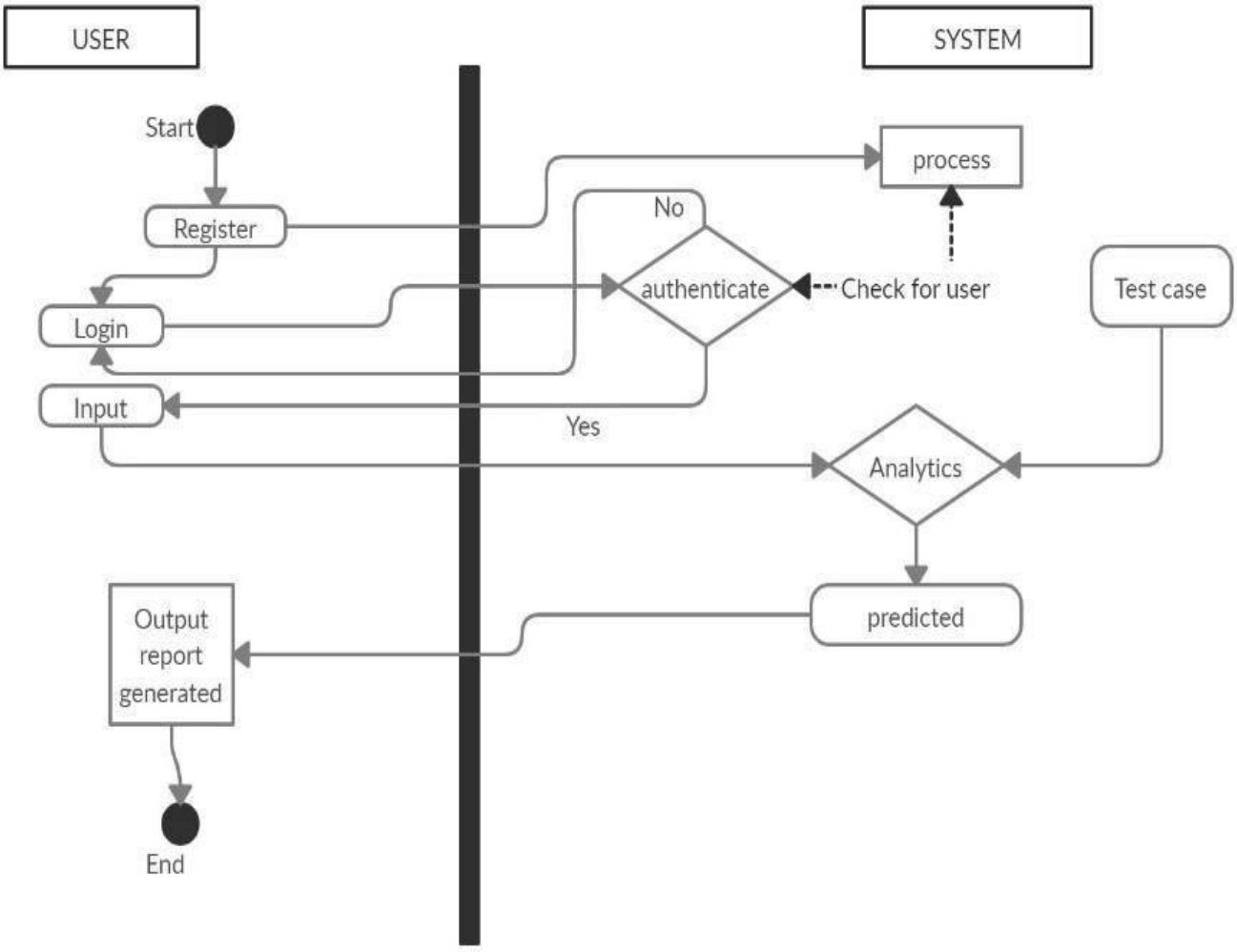


Fig. 5.8 Activity Diagram

COMPONENT DIAGRAM:

A component diagram, also known as a UML component diagram, describes the organization and wiring of the physical components in a system. Component diagrams are often drawn to help model implementation details and double-check that every aspect of the system's required function is covered by planned development. Here component diagram consists of all major components that is used to build a system. So, design, Algorithm, File System and Datasets all are linked to one another. Datasets are used to compare the results and algorithm is used to process those results and give a correct accuracy and design UI is used to show the result in an appropriate way in the system and file system is used to store the user data. So, like this all components are interlinked to each other.

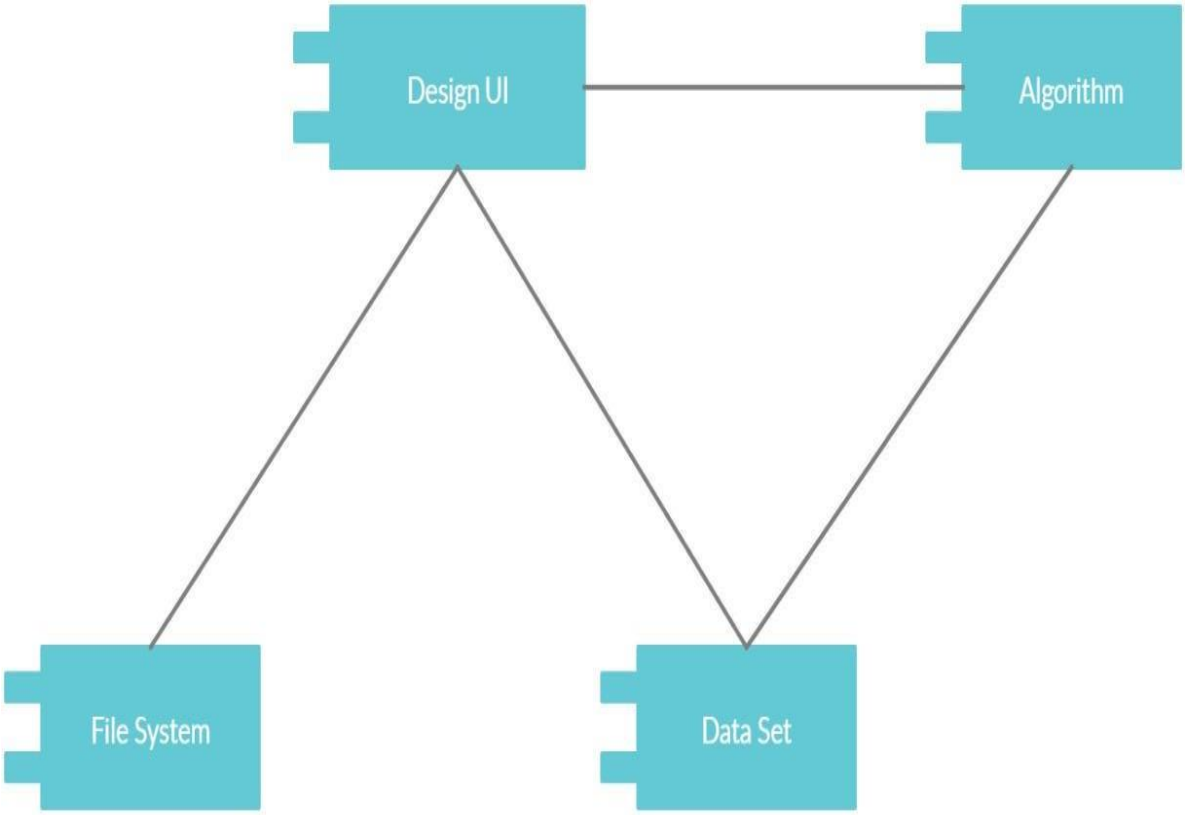


Fig 5.9 Component Diagram

STATE CHART DIAGRAM:

A State chart diagram describes the behavior of a single object in response to a series of events in a system. Sometimes it's also known as a Harel state chart or a state machine diagram. This UML diagram models the dynamic flow of control from state to state of a particular object within a system. It is similar to activity diagram but here there are only few rules like how it starts and how it ends all are denoted with the help of the symbol given below, the system starts with the registration and then login comes, if the login is successful then it will go to the next step and if login is incorrect then comes back to same page stating incorrect details. After the successful login the user needs to enter the symptoms and then press the prediction button, at the same time the backend process will do their work and the correct result is predicted.

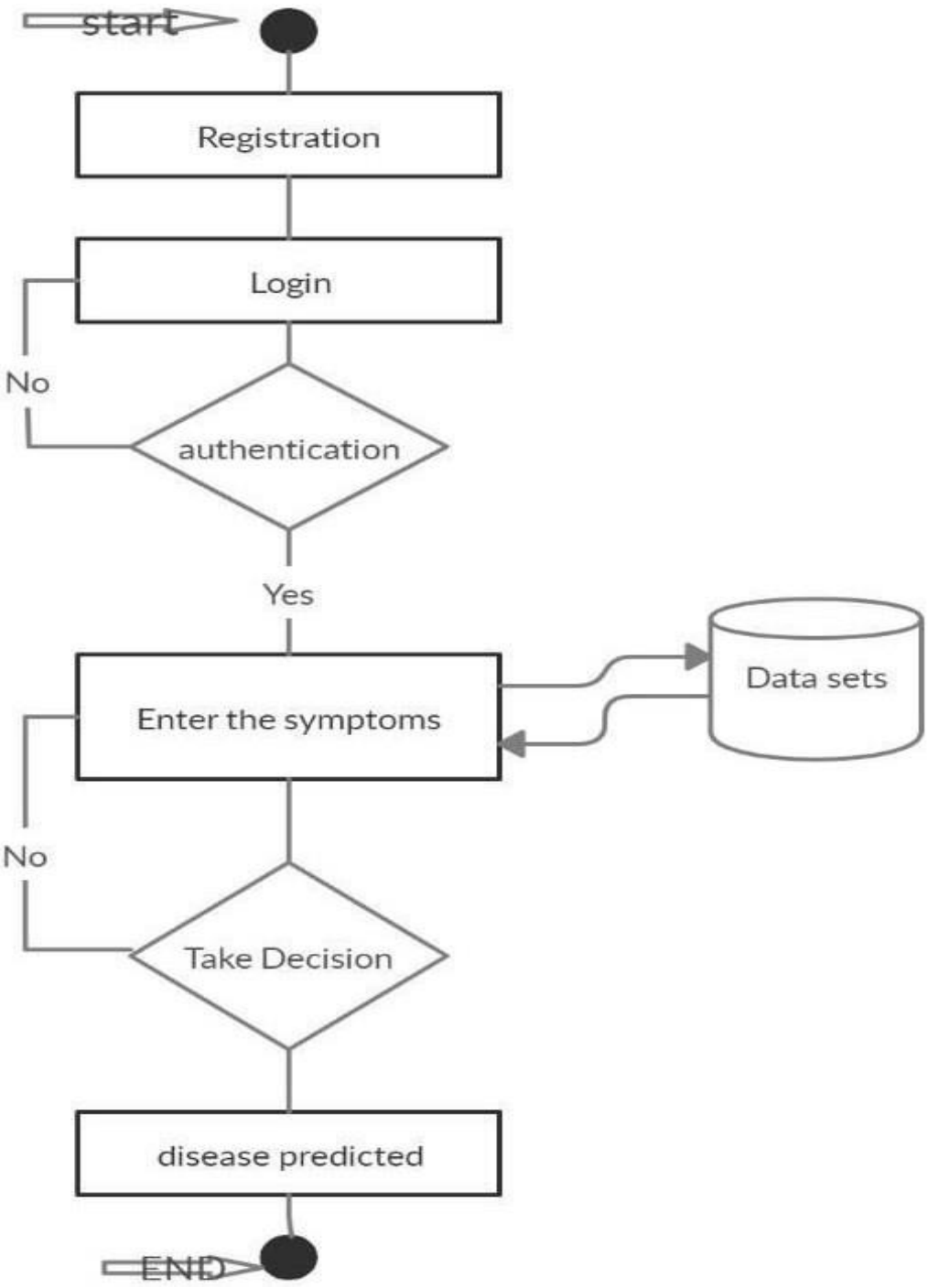
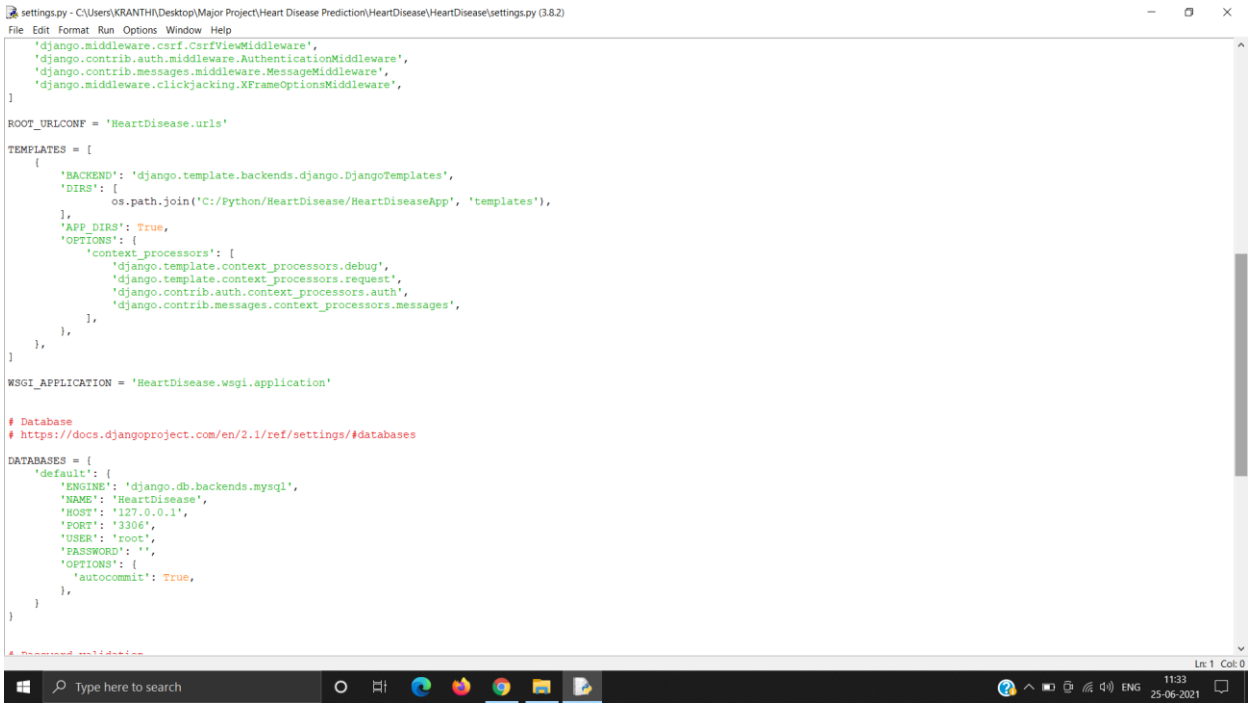


Fig. 5.10 State Chart Diagram

6. PROJECT CODING

Project Coding is the process of designing and building an executable program code to accomplish a specific computing result or to churn out a particular prototype or product. Programming involves tasks such as: analysis, generating algorithms, profiling algorithms' accuracy and resource consumption, and the implementation of algorithms in a chosen programming language (commonly referred to as coding). The source code of a program is written in one or more languages that are intelligible to programmers rather than machine code which is directly executed by the CPU. The purpose of programming is to find a sequence of instructions that will automate the performance of a task (which can be as complex as an OS) on a computer often for solving a given problem. Proficient programming thus often requires expertise in several different subjects, including knowledge of the application domain, specialized algorithms, and formal logic.

6.1. CODE TEMPLATES



settings.py - C:\Users\KRANTHI\Desktop\Major Project\Heart Disease Prediction\HeartDisease\HeartDisease\settings.py (3.8.2)

File Edit Format Run Options Window Help

```
        'USER': 'root',
        'PASSWORD': '',
        'OPTIONS': {
          'autocommit': True,
        },
    }
}


# Password validation
# https://docs.djangoproject.com/en/2.1/ref/settings/#auth-password-validators

AUTH_PASSWORD_VALIDATORS = [
    {
        'NAME': 'django.contrib.auth.password_validation.UserAttributeSimilarityValidator',
    },
    {
        'NAME': 'django.contrib.auth.password_validation.MinimumLengthValidator',
    },
    {
        'NAME': 'django.contrib.auth.password_validation.CommonPasswordValidator',
    },
    {
        'NAME': 'django.contrib.auth.password_validation.NumericPasswordValidator',
    },
]


# Internationalization
# https://docs.djangoproject.com/en/2.1/topics/i18n/

LANGUAGE_CODE = 'en-us'

TIME_ZONE = 'UTC'

USE_I18N = True

USE_L10N = True

USE_TZ = True


# Static files (CSS, JavaScript, Images)
# https://docs.djangoproject.com/en/2.1/howto/static-files/

STATIC_URL = '/static/'
```

Ln: 1 Col: 0

Type here to search ENG 11:33 25-06-2021

urls.py - C:\Users\KRANTHI\Desktop\Major Project\Heart Disease Prediction\HeartDisease\HeartDisease\urls.py (3.8.2)

File Edit Format Run Options Window Help

```
"""HeartDisease URL Configuration

The `urlpatterns` list routes URLs to views. For more information please see:
    https://docs.djangoproject.com/en/2.1/topics/http/urls/
Examples:
Function views
    1. Add an import:  from my_app import views
    2. Add a URL to urlpatterns:  path('', views.home, name='home')
Class-based views
    1. Add an import:  from other_app.views import Home
    2. Add a URL to urlpatterns:  path('', Home.as_view(), name='home')
Including another URLconf
    1. Import the include() function: from django.urls import include, path
    2. Add a URL to urlpatterns:  path('blog/', include('blog.urls'))
"""
from django.contrib import admin
from django.urls import path, include

urlpatterns = [
    path('admin/', admin.site.urls),
    path('', include('HeartDiseaseApp.urls')),
]
```

Ln: 1 Col: 0

Type here to search ENG 11:35 25-06-2021

wsgi.py - C:\Users\KRANTHI\Desktop\Major Project\Heart Disease Prediction\HeartDisease\HeartDisease\wsgi.py (3.8.2)

File Edit Format Run Options Window Help

```
"""
WSGI config for HeartDisease project.

It exposes the WSGI callable as a module-level variable named ``application``.

For more information on this file, see
https://docs.djangoproject.com/en/2.1/howto/deployment/wsgi/
"""

import os

from django.core.wsgi import get_wsgi_application

os.environ.setdefault('DJANGO_SETTINGS_MODULE', 'HeartDisease.settings')

application = get_wsgi_application()
```

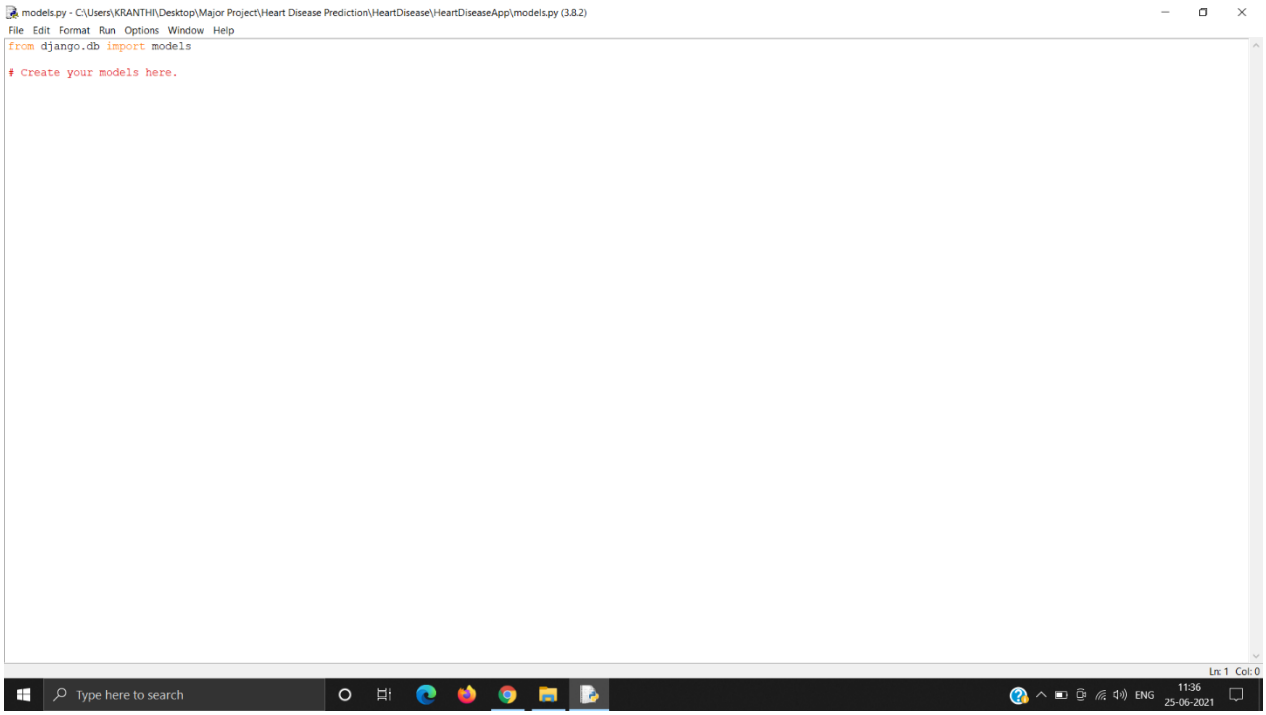
Ln: 1 Col: 0

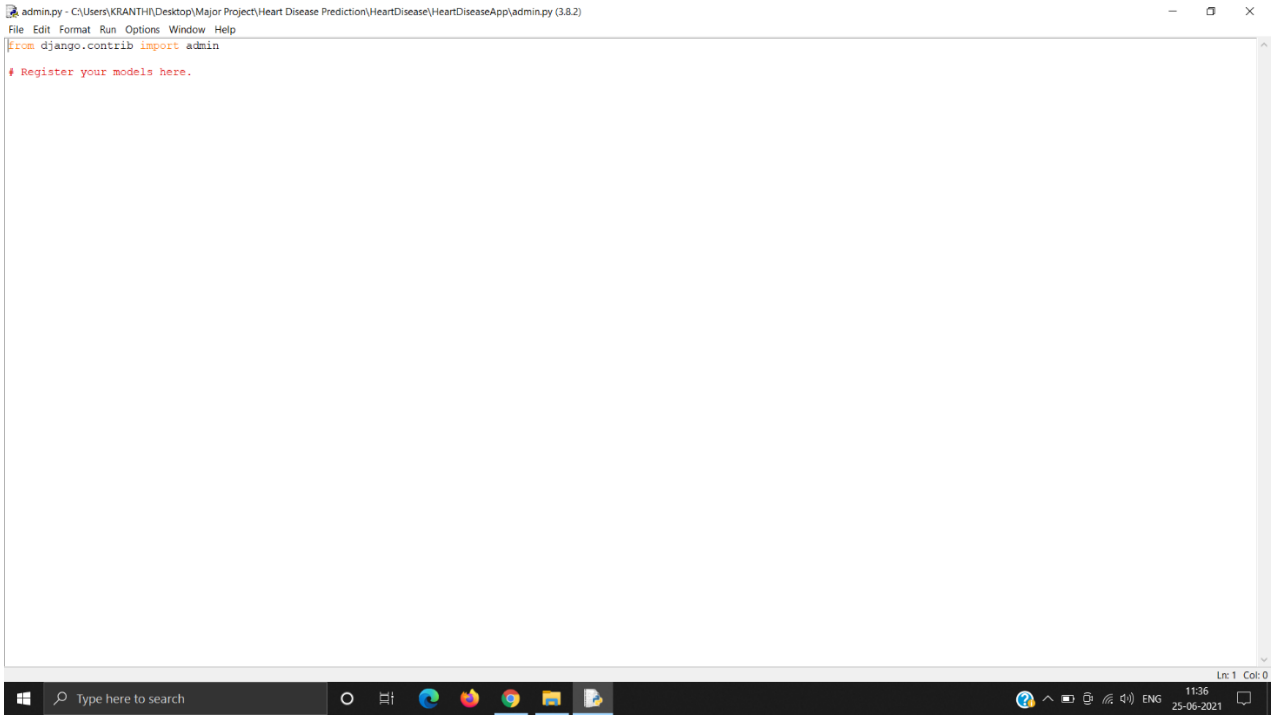
views.py - C:\Users\KRANTHI\Desktop\Major Project\Heart Disease Prediction\HeartDisease\HeartDiseaseApp\views.py (3.8.2)

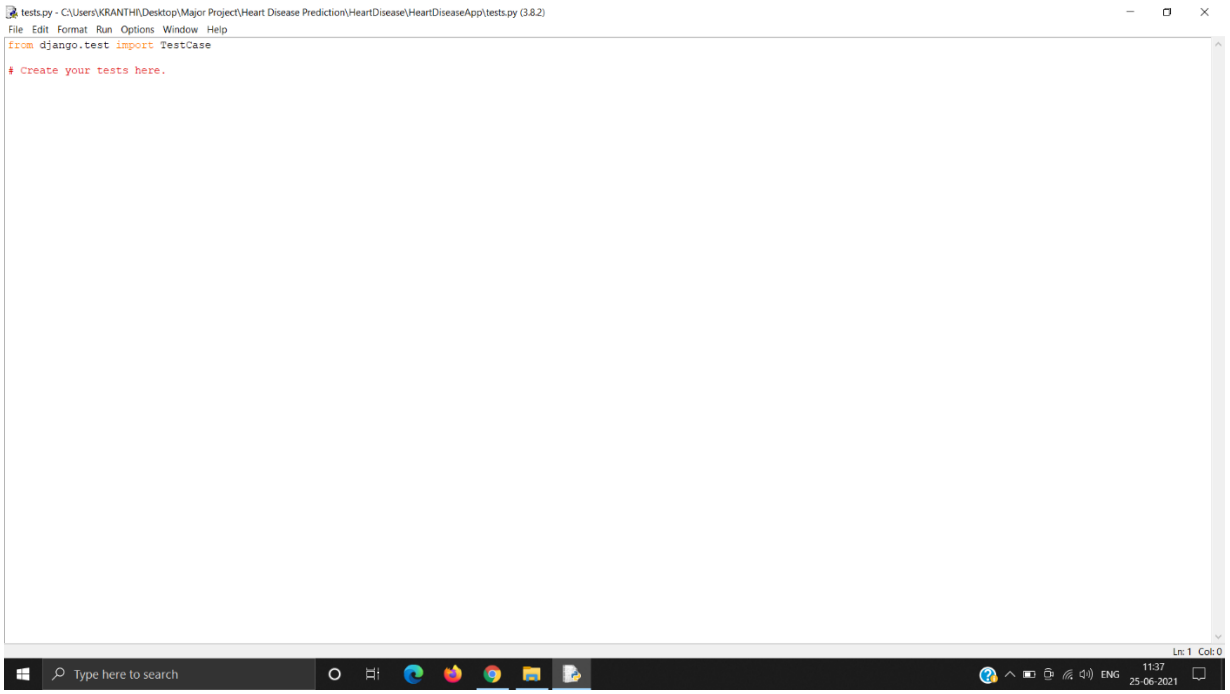
File Edit Format Run Options Window Help

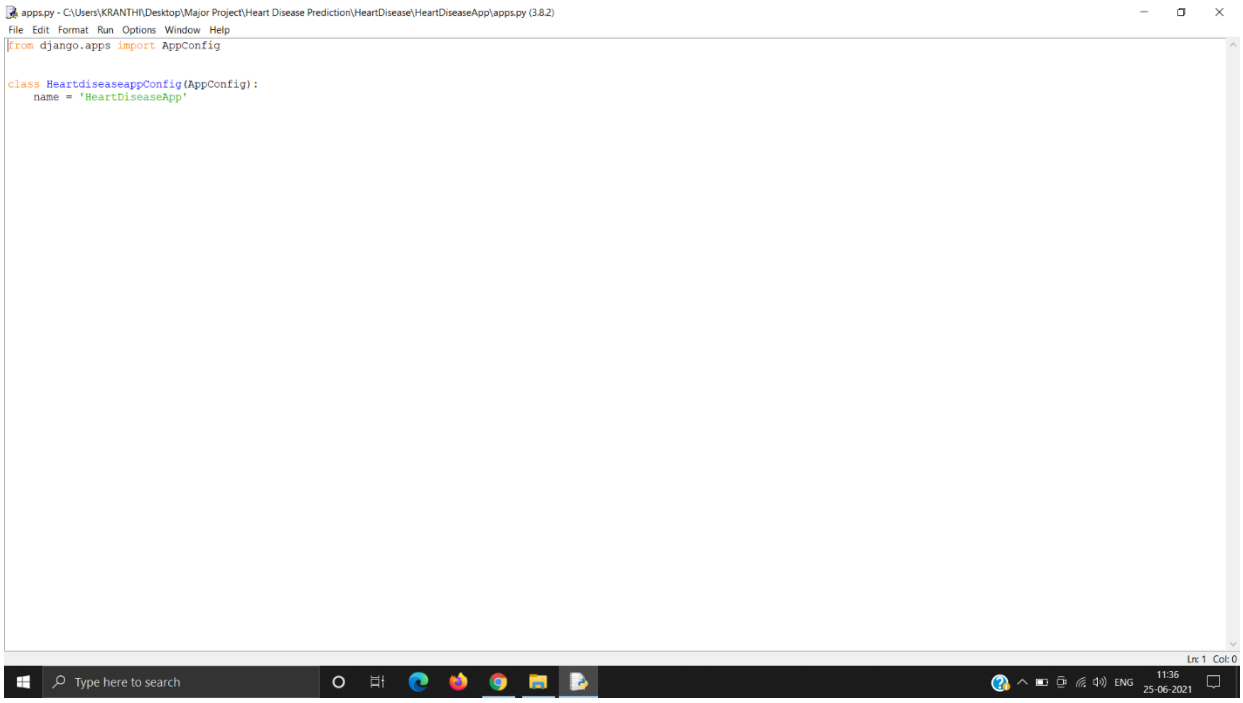
```
from django.shortcuts import render
from django.template import RequestContext
from django.contrib import messages
import pymysql
from django.http import HttpResponse
from sklearn.naive_bayes import GaussianNB
from sklearn.model_selection import train_test_split
import pandas as pd

def index(request):
    if request.method == 'GET':
       return render(request, 'index.html', {})

def Login(request):
    if request.method == 'GET':
       return render(request, 'Login.html', {})

def Register(request):
    if request.method == 'GET':
       return render(request, 'Register.html', {})

def Predict(request):
    if request.method == 'GET':
       return render(request, 'Predict.html', {})


def PredictHeartCondition(request):
    if request.method == 'POST':
        age = request.POST.get('age', False)
        gender = request.POST.get('gender', False)
        cp = request.POST.get('cp', False)
        bps = request.POST.get('trestbps', False)
        chol = request.POST.get('chol', False)
        fbs = request.POST.get('fbs', False)
        ecg = request.POST.get('restecg', False)
        thalach = request.POST.get('thalach', False)
        exang = request.POST.get('exang', False)
        oldpeak = request.POST.get('oldpeak', False)
        slope = request.POST.get('slope', False)
        ca = request.POST.get('ca', False)
        thal = request.POST.get('thal', False)

        data = 'age,sex,cp,trestbps,chol,fbs,restecg,thalach,exang,oldpeak,slope,ca,thal\n'
        data+=age+","+gender+","+cp+","+bps+","+chol+","+fbs+","+ecg+","+thalach+","+exang+","+oldpeak+","+slope+","+ca+","+thal

        file = open('testdata.txt','w')
        file.write(data)
```

Ln: 1 Col: 0


models.py - C:\Users\KRANTHI\Desktop\Major Project\Heart Disease Prediction\HeartDisease\HeartDiseaseApp\models.py (3.8.2)
File Edit Format Run Options Window Help
from django.db import models
# Create your models here.
Ln: 1 Col: 0
Type here to search
ENG
11:36
25-06-2021


admin.py - C:\Users\KRANTHI\Desktop\Major Project\Heart Disease Prediction\HeartDisease\HeartDiseaseApp\admin.py (3.8.2)
File Edit Format Run Options Window Help
from django.contrib import admin
# Register your models here.
Ln: 1 Col: 0
Type here to search
ENG
11:36
25-06-2021


tests.py - C:\Users\KRANTHI\Desktop\Major Project\Heart Disease Prediction\HeartDisease\HeartDiseaseApp\tests.py (3.8.2)
File Edit Format Run Options Window Help
from django.test import TestCase
# Create your tests here.
Ln: 1 Col: 0
Type here to search
ENG
11:37
25-06-2021


apps.py - C:\Users\KRANTHI\Desktop\Major Project\Heart Disease Prediction\HeartDisease\HeartDiseaseApp\apps.py (3.8.2)
File Edit Format Run Options Window Help
from django.apps import AppConfig
class HeartdiseaseappConfig(AppConfig):
name = 'HeartDiseaseApp'
Ln: 1 Col: 0
Type here to search
ENG
11:36
25-06-2021

settings.py - C:\Users\KRANTHI\Desktop\Major Project\Heart Disease Prediction\HeartDisease\HeartDisease\settings.py (3.8.2)

File Edit Format Run Options Window Help

```
"""
Django settings for HeartDisease project.

Generated by 'django-admin startproject' using Django 2.1.7.

For more information on this file, see
https://docs.djangoproject.com/en/2.1/topics/settings/

For the full list of settings and their values, see
https://docs.djangoproject.com/en/2.1/ref/settings/
"""

import os

# Build paths inside the project like this: os.path.join(BASE_DIR, ...)
BASE_DIR = os.path.dirname(os.path.dirname(os.path.abspath(__file__)))


# Quick-start development settings - unsuitable for production
# See https://docs.djangoproject.com/en/2.1/howto/deployment/checklist/

# SECURITY WARNING: keep the secret key used in production secret!
SECRET_KEY = '9&7)au1p_8-cz_p2k9aee5&zl!6zku3+2e&6f%ll)a#e-zn2@*'

# SECURITY WARNING: don't run with debug turned on in production!
DEBUG = True

ALLOWED_HOSTS = []


# Application definition

INSTALLED_APPS = [
    'django.contrib.admin',
    'django.contrib.auth',
    'django.contrib.contenttypes',
    'django.contrib.sessions',
    'django.contrib.messages',
    'django.contrib.staticfiles',
    'HeartDiseaseApp'
]

MIDDLEWARE = [
    'django.middleware.security.SecurityMiddleware',
    'django.contrib.sessions.middleware.SessionMiddleware',
    'django.middleware.common.CommonMiddleware',
    'django.middleware.csrf.CsrfViewMiddleware',
```

Ln: 1 Col: 0

Type here to search ENG 11:32 25-06-2021

views.py - C:\Users\KRANTHI\Desktop\Major Project\Heart Disease Prediction\HeartDisease\HeartDiseaseApp\views.py (3.8.2)

File Edit Format Run Options Window Help

```
        file.write(data)
        file.close()

        train = pd.read_csv('dataset')
        X = train.values[:, 0:13]
        Y = train.values[:, 13]
        #X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size = 0.2, random_state = 0)

        cls = GaussianNB()
        cls.fit(X, Y)

        test = pd.read_csv('testdata.txt')
        test = test.values[:, 0:13]
        y_pred = cls.predict(test)

        result = ''
        print(y_pred)
        for i in range(len(test)):
            if str(y_pred[i]) == '0.0':
                result = str(test[i])+"<br/>Result =  No Heart Disease Detected"
            if str(y_pred[i]) == '1.0':
                result = str(test[i])+"<br/>Result =  Heart Disease Detected"

        context= {'data':result}
        return render(request, 'Result.html', context)


def Signup(request):
    if request.method == 'POST':
      #user_ip = getClientIP(request)
      #reader = geoip2.database.Reader('C:/Python/PlantDisease/GeoLite2-City.mmdb')
      #response = reader.city('103.48.68.11')
      #print(user_ip)
      #print(response.location.latitude)
      #print(response.location.longitude)
      username = request.POST.get('username', False)
      password = request.POST.get('password', False)
      contact = request.POST.get('contact', False)
      email = request.POST.get('email', False)
      address = request.POST.get('address', False)

      db_connection = pymysql.connect(host='127.0.0.1',port = 3306,user = 'root', password = '', database = 'HeartDisease',charset='utf8')
      db_cursor = db_connection.cursor()
      student_sql_query = "INSERT INTO register(username,password,contact,email,address) VALUES('"+username+"','"+password+"','"+contact+"','"+email+"','"+address+"')"
      db_cursor.execute(student_sql_query)
      db_connection.commit()
```

Ln: 1 Col: 0

Type here to search ENG 11:37 25-06-2021

views.py - C:\Users\KRANTHI\Desktop\Major Project\Heart Disease Prediction\HeartDisease\HeartDiseaseApp\views.py (3.8.2)

File Edit Format Run Options Window Help

```
        #print(user_ip)
        #print(response.location.latitude)
        #print(response.location.longitude)
        username = request.POST.get('username', False)
        password = request.POST.get('password', False)
        contact = request.POST.get('contact', False)
        email = request.POST.get('email', False)
        address = request.POST.get('address', False)

        db_connection = pymysql.connect(host='127.0.0.1',port = 3306,user = 'root', password = '', database = 'HeartDisease',charset='utf8')
        db_cursor = db_connection.cursor()
        student_sql_query = "INSERT INTO register(username,password,contact,email,address) VALUES('"+username+"','"+password+"','"+contact+"','"+email+"','"+address+"')"
        db_cursor.execute(student_sql_query)
        db_connection.commit()
        print(db_cursor.rowcount, "Record Inserted")
        if db_cursor.rowcount == 1:
         context= {'data':'Signup Process Completed'}
         return render(request, 'Register.html', context)
        else:
         context= {'data':'Error in signup process'}
         return render(request, 'Register.html', context)

def UserLogin(request):
    if request.method == 'POST':
        username = request.POST.get('username', False)
        password = request.POST.get('password', False)
        utype = 'none'
        con = pymysql.connect(host='127.0.0.1',port = 3306,user = 'root', password = '', database = 'HeartDisease',charset='utf8')
        with con:
            cur = con.cursor()
            cur.execute("select * FROM register")
            rows = cur.fetchall()
            for row in rows:
                if row[0] == username and row[1] == password:
                    utype = 'success'
                    break
        if utype == 'success':
            file = open('session.txt','w')
            file.write(username)
            file.close()
            context= {'data':'welcome '+username}
            return render(request, 'UserScreen.html', context)
        if utype == 'none':
            context= {'data':'Invalid login details'}
            return render(request, 'Login.html', context)
```

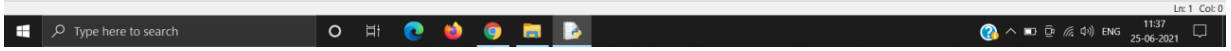
Ln: 1 Col: 0

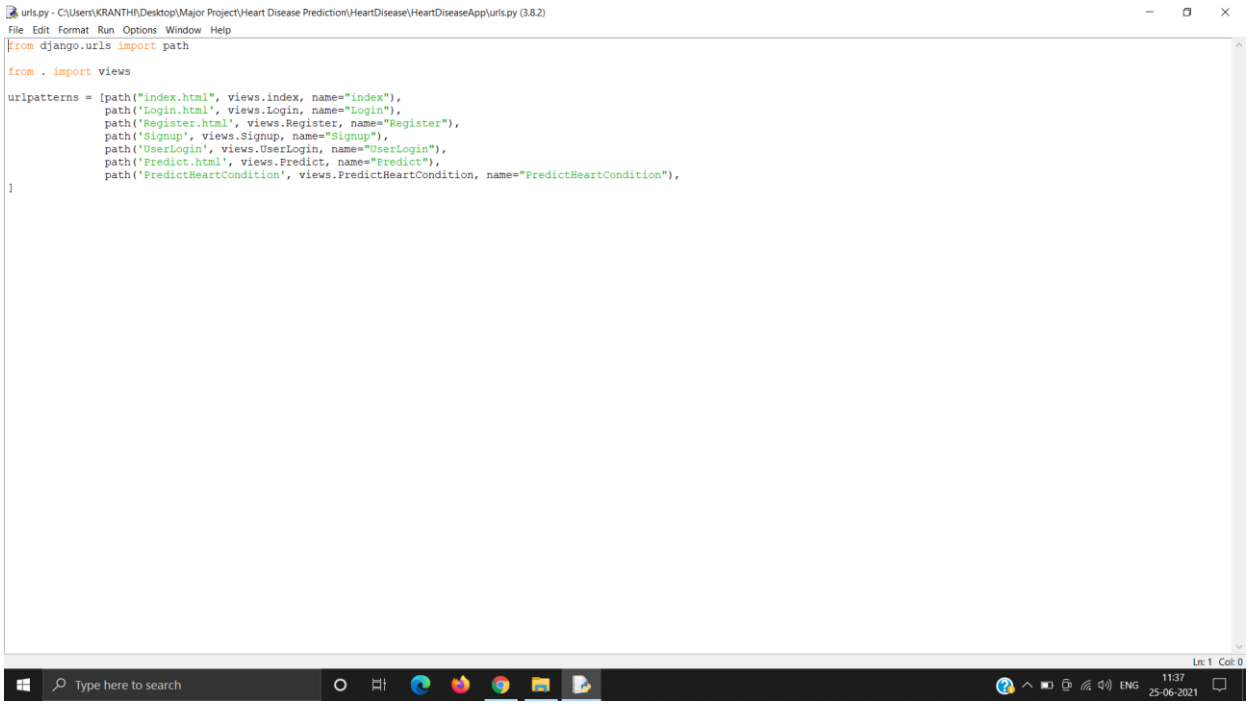
urls.py - C:\Users\KRANTHI\Desktop\Major Project\Heart Disease Prediction\HeartDisease\HeartDiseaseApp\urls.py (3.8.2)

File Edit Format Run Options Window Help

```
from django.urls import path

from . import views

urlpatterns = [path("index.html", views.index, name="index"),
               path('Login.html', views.Login, name="Login"),
               path('Register.html', views.Register, name="Register"),
               path('Signup', views.Signup, name="Signup"),
               path('UserLogin', views.UserLogin, name="UserLogin"),
               path('Predict.html', views.Predict, name="Predict"),
               path('PredictHeartCondition', views.PredictHeartCondition, name="PredictHeartCondition"),
]
```

Ln: 1 Col: 0





6.2. OUTLINE FOR VARIOUS FILES

We used Python programming to implement our project. A single python file is used to implement our code. This file consists of various modules that we have used. Our project modules are – Home page, Register page, Login page, User. We also used various pythonmodules like tkinter, matplotlib, NumPy, imutils, os, cv2.

6.3. CLASS WITH FUNCTIONALITY

There are multiple classes in our code, some of which are:

1) Templates: They handle the backend context processors for requests, debugging, authentication and messaging.

2) Middleware: The middleware class handles the view, sessions, security and frame options.

3) Django Installed Apps: These handle the content types and static files through the code.

4) Validator: They handle the validation of multiple users in and out of the website based on the credentials that the user would provide.

5) Database Linking: This handles the backend linking of the entire code to the MySQL server in order to facilitate the final output.

6.4. METHODS INPUT AND OUTPUT PARAMETERS

We implemented multiple methods, few of which are :

1. enterDetails()
2. submitDetails()
3. login()
4. predict()
5. logout() , etc.

Our first method enterDetails() takes person details as an input. Second submitDetails() takes input from enterDetails() module and does the processing and gives the result. login() takes username and password as parameters and shows the user interface for faculty as a result. predict() method shows the screen in which user has to enter the details and submit it. logout() method will close the interface and displays home page.

7.PROJECT TESTING

Project Testing is a method to check whether the actual software product matches expected requirements and to ensure that software product is Defect free. It involves execution of software/system components using manual or automated tools to evaluate one or more properties of interest. The purpose of software testing is to identify errors, gaps or missing requirements in contrast to actual requirements.

Some prefer saying Software testing as a White Box and Black Box Testing. In simple terms, Software Testing means the Verification of Application Under Test (AUT). This tutorial introduces testing software to the audience and justifies its importance.

Project testing is important because, if there are any bugs or errors in the software, it can be identified early and can be solved before delivery of the software product. Properly tested software product ensures reliability, security and high performance which further results in time saving, cost effectiveness and customer satisfaction.

Typically Testing is classified into three categories.

- Functional Testing
- Non-Functional Testing or Performance Testing
- Maintenance (Regression and Maintenance)

7.1. VARIOUS TEST CASES

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub-assemblies, assemblies and/or a finished product. It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of tests. Each test type addresses a specific testing requirement.

We have performed multiple tests under the broad categorization of white-box and black-box testing which further include unit, integration, boundary value and statement covering etc.

ID	Case	Expected Outcomes	Comments
1.0	**Login**		
1.1	Password or username left out	Error Dialog Box	Pass
1.2	Wrong password or username entered	Error Dialog Box	Pass
2.0	**User Registration**		
2.1	Leaving out a required field	Error Dialog Box	Pass
3.0	**Predict your heart disease**		
3.1	Leaving out a required field	Error Dialog Box	Pass

Table 6. Test Cases Tabulation

Unit testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration.

Integration testing

Integration tests are designed to test integrated software components to determine if they actually runas one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

Functional test

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.
Functional testing is centered on the following items:

Valid Input	: identified classes of valid input must be accepted.
Invalid Input	: identified classes of invalid input must be rejected.
Functions	: identified functions must be exercised.
Output	: identified classes of application outputs must be exercised.
Systems/Procedures	: interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows;data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

System Test

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration-oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

7.2. BLACK BOX

Black Box Testing is a software testing method in which the functionalities of software applications are tested without having knowledge of internal code structure, implementation details and internal paths. Black Box Testing mainly focuses on input and output of software applications and it is entirely based on software requirements and specifications. It is also known as Behavioral Testing.

Here are the generic steps followed to carry out any type of Black Box Testing.

- Initially, the requirements and specifications of the system are examined.
- Tester chooses valid inputs (positive test scenario) to check whether SUT processes them correctly. Also, some invalid inputs (negative test scenario) are chosen to verify that the SUT is able to detect them.
- Tester determines expected outputs for all those inputs.
- Software tester constructs test cases with the selected inputs.
- The test cases are executed.
- Software tester compares the actual outputs with the expected outputs.
- Defects if any are fixed and re-tested.

Types of Black Box Testing

There are many types of Black Box Testing but the following are the prominent ones -

- Functional testing - This black box testing type is related to the functional requirements of a system; it is done by software testers.
- Non-functional testing - This type of black box testing is not related to testing of specific functionality, but non-functional requirements such as performance, scalability, usability.
- Regression testing - Regression Testing is done after code fixes, upgrades or any other system maintenance to check the new code has not affected the existing code.

Black Box Testing Techniques

Following is the prominent Test Strategy amongst the many used in Black box Testing

- Equivalence Class Testing: It is used to minimize the number of possible test cases to an optimum level while maintains reasonable test coverage.
- Boundary Value Testing: Boundary value testing is focused on the values at boundaries. This technique determines whether a certain range of values are acceptable by the system or not. It is very useful in reducing the number of test cases. It is most suitable for the systems where an input is within certain ranges.
- Decision Table Testing: A decision table puts causes and their effects in a matrix. There is a unique combination in each column.

7.3. WHITE BOX TESTING

White Box Testing is software testing technique in which internal structure, design and coding of software are tested to verify flow of input-output and to improve design, usability and security. In white box testing, code is visible to testers so it is also called Clear box testing, Open box testing, Transparent box testing, Code-based testing and Glass box testing.

It is one of two parts of the Box Testing approach to software testing. Its counterpart, Blackbox testing, involves testing from an external or end-user type perspective. On the other hand, White box testing in software engineering is based on the inner workings of an application and revolves around internal testing.

The term "White Box" was used because of the see-through box concept. The clear box or White Box name symbolizes the ability to see through the software's outer shell (or "box") into its inner workings. Likewise, the "black box" in "Black Box Testing" symbolizes not being able to see the inner workings of the software so that only the end-user experience can be tested.

White box testing involves the testing of the software code for the following:

- Internal security holes
- Broken or poorly structured paths in the coding processes
- The flow of specific inputs through the code
- Expected output
- The functionality of conditional loops
- Testing of each statement, object, and function on an individual basis

The testing can be done at system, integration and unit levels of software development. One of the basic goals of white box testing is to verify a working flow for an application. It involves testing a series of predefined inputs against expected or desired outputs so that when a specific input does not result in the expected output, you have encountered a bug.

To give you a simplified explanation of white box testing, we have divided it into two basic steps. This is what we do when testing an application using the white box testing technique:

STEP 1) UNDERSTAND THE SOURCE CODE

The first thing a tester will often do is learn and understand the source code of the application. Since white box testing involves the testing of the inner workings of an application, the tester must be very knowledgeable in the programming languages used in the applications they are testing. Also, the testing person must be highly aware of secure coding practices. Security is often one of the primary objectives of testing software. The tester should be able to find security issues and prevent attacks from hackers and naive users who might inject malicious code into the application either knowingly or unknowingly.

Step 2) CREATE TEST CASES AND EXECUTE

The second basic step to white box testing involves testing the application's source code for proper flow and structure. One way is by writing more code to test the application's source code. The tester will develop little tests for each process or series of processes in the application. This method requires that the tester must have intimate knowledge of the code and is often done by the developer.

The goal of White Box testing in software engineering is to verify all the decision branches, loops, statements in the code.

A major White box testing technique is Code Coverage analysis. Code Coverage analysis eliminates gaps in a Test Case suite. It identifies areas of a program that are not exercised by a set of test cases. Once gaps are identified, you create test cases to verify untested parts of the code, thereby increasing the quality of the software product

Following is important White Box Testing Techniques:

- Statement Coverage
- Decision Coverage
- Branch Coverage
- Condition Coverage
- Multiple Condition Coverage
- Finite State Machine Coverage
- Path Coverage
- Control flow testing
- Data flow testing

8. OUTPUT SCREENS

An output screen is a device used to display output. An output screen could be a separate monitor or another display device used only to display the output being received from the computer or other devices.

Here, in the screen prints given below, we can see that the user interface screens consist of the home page which describes the purpose of the portal, and the public access screen which allows users to upload the credentials of the child that they found in order to check whether the child exists in the repository or not.

8.1 USER INTERFACE

1) HOME PAGE

This allows the user to look at the various sections of the website like the home page, official login, registration, so that they can navigate through the web page in an easier way.

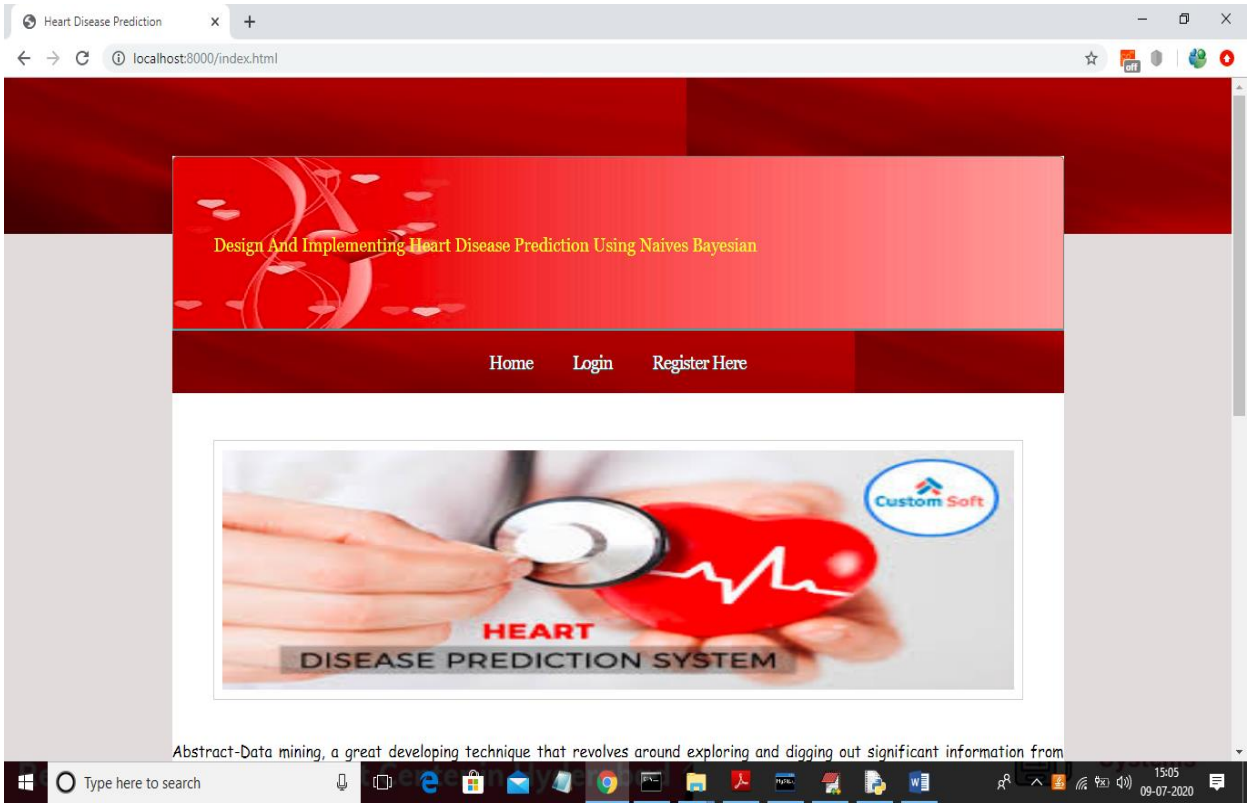


Fig 8.1 User Interface

8.2 OUTPUT SCREENS

1) The official register screen helps the user to check the activity going on in the web page. The user upload the details for registering.

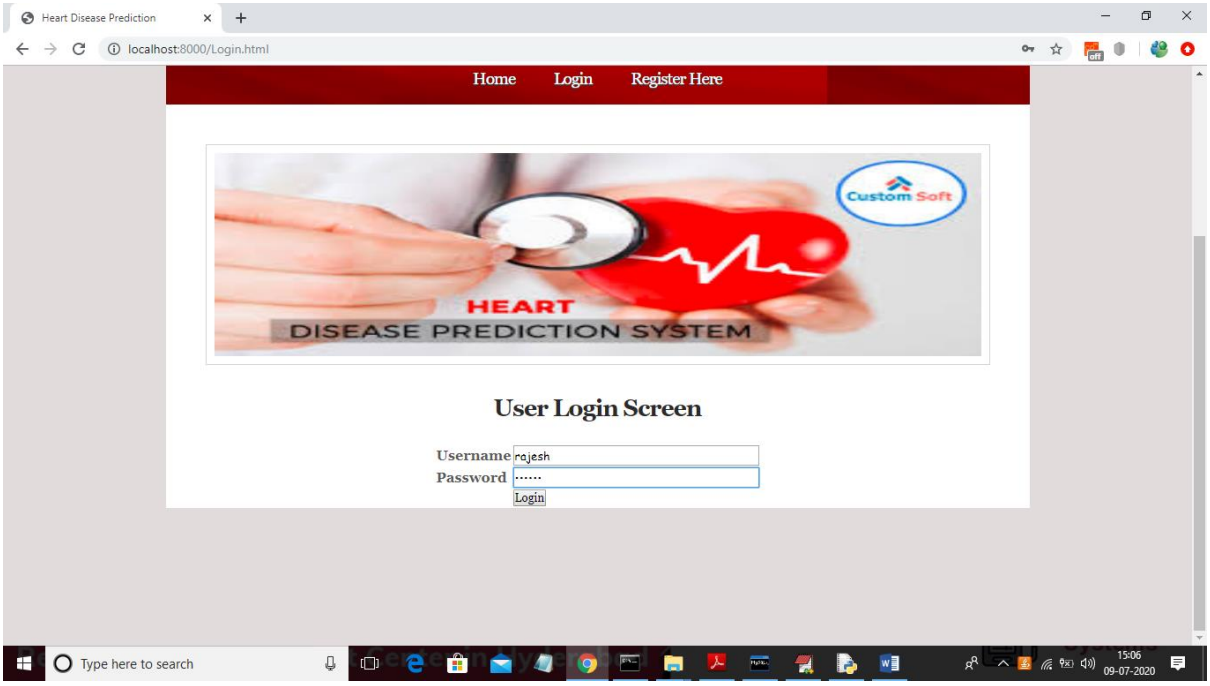


Fig 8.2 User login Screen

9. EXPERIMENTAL RESULT

1) Now the user clicks on login link to log user

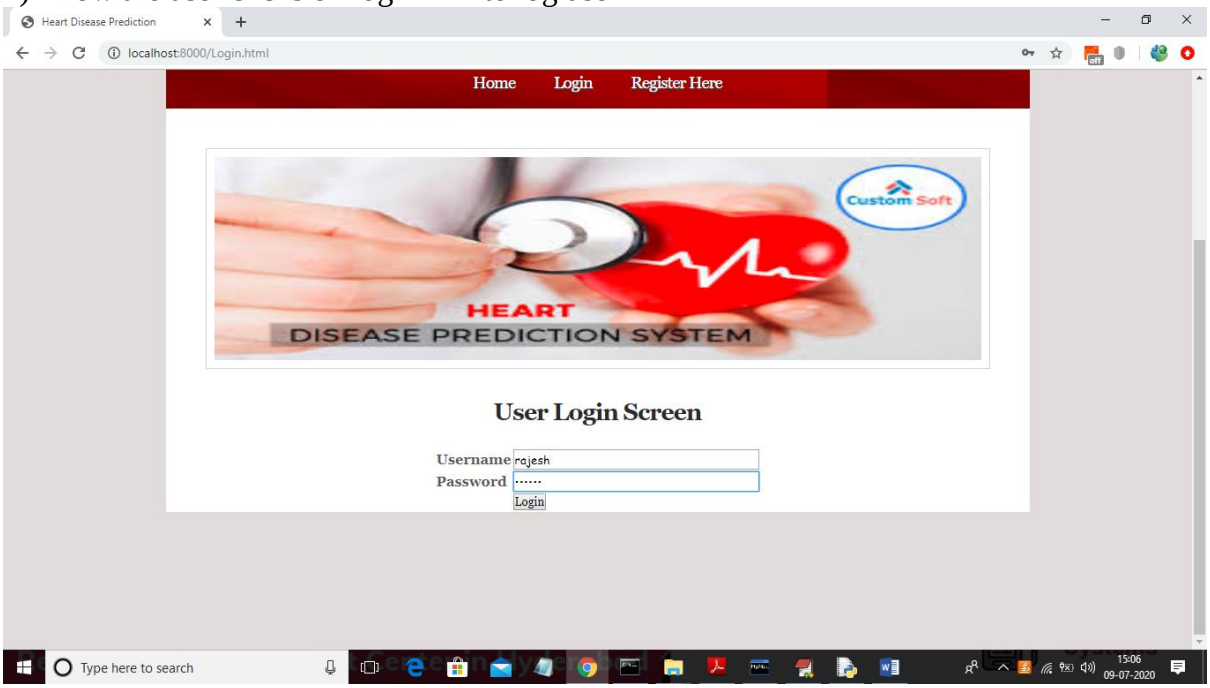


Fig 9.1 User login Screen

2) After login will get in above screen. Now click on 'Predict Your Heart Condition' link to input details in the below screen.

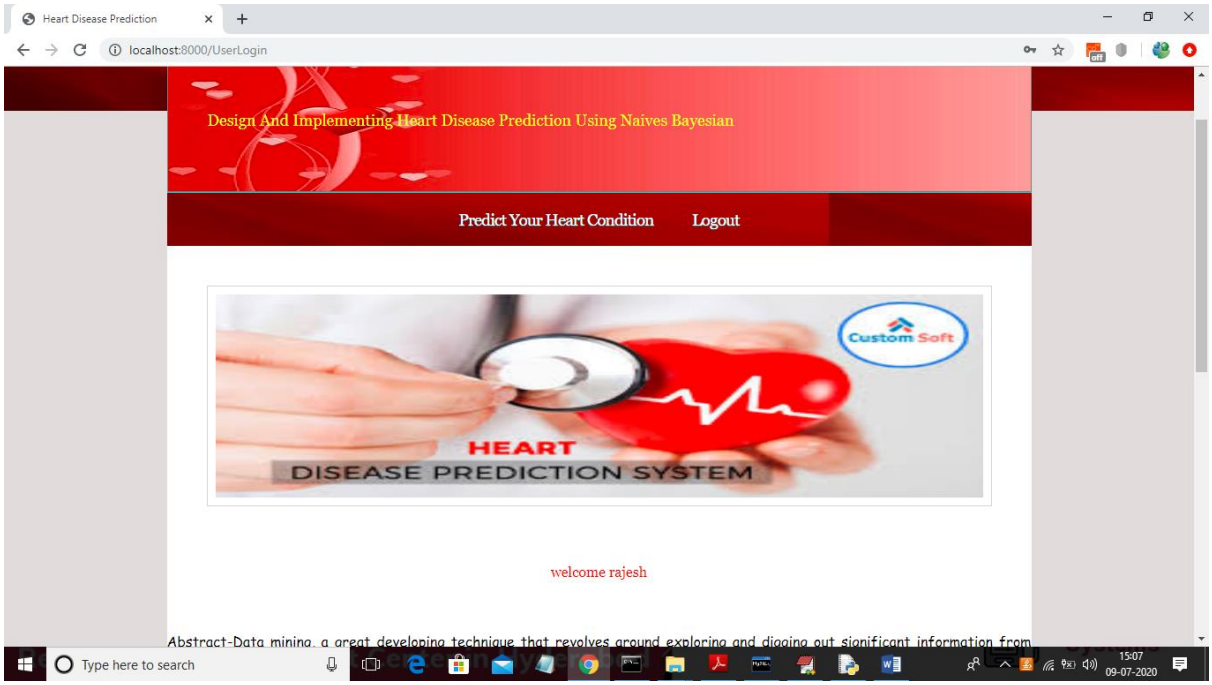


Fig 9.2 Test Screen

3) In below screen user will enter details about his heart and application will predict disease

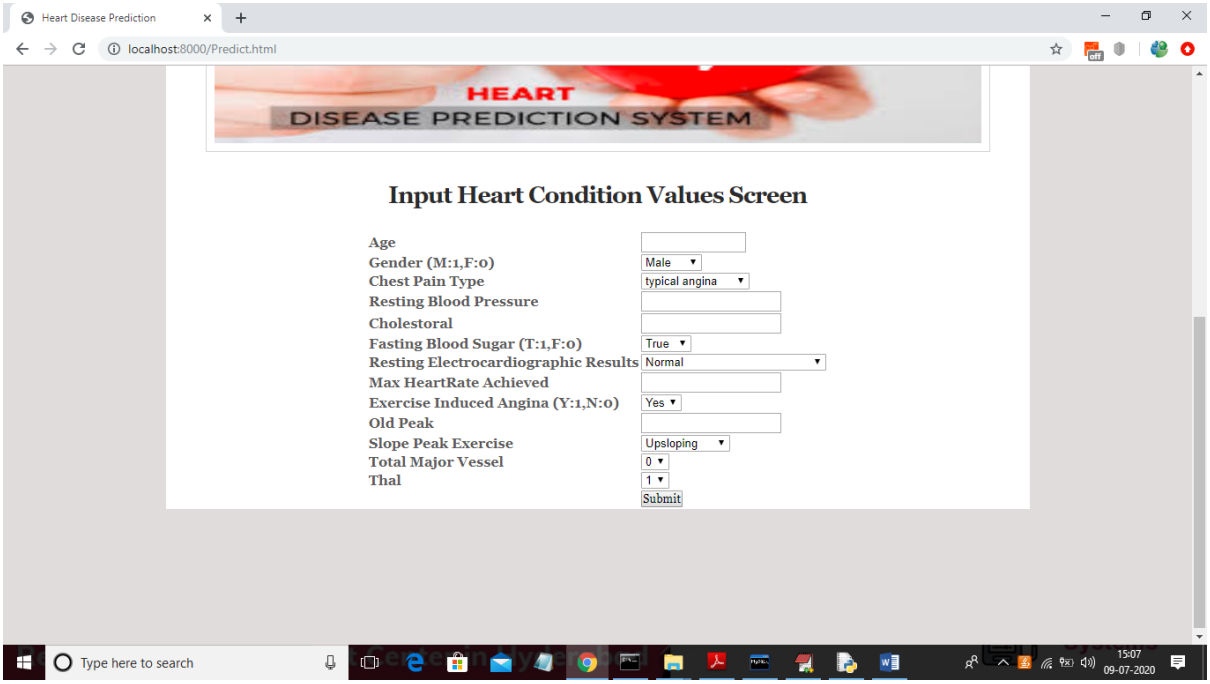


Fig 9.3 Input heart condition values Screen

4) After entering details will get below screen of prediction and click on submit.

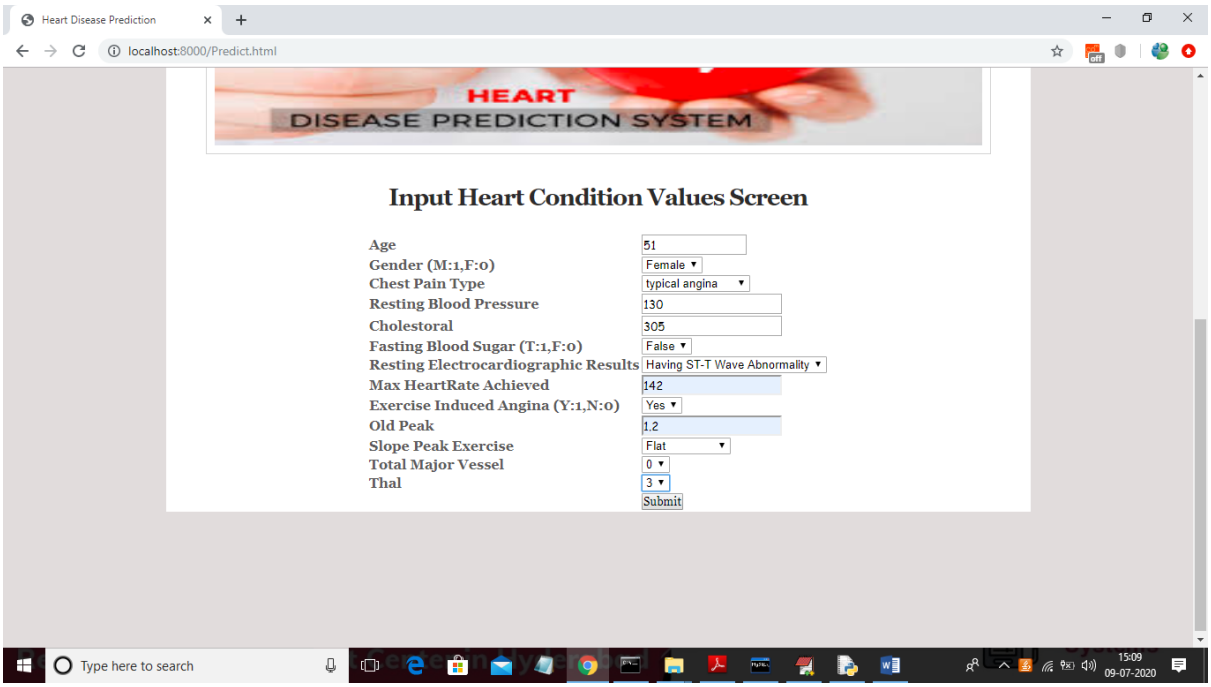


Fig 9.4 Input heart condition with values Screen

5) The result page is displayed which shows whether the user is detected with heart disease or not.

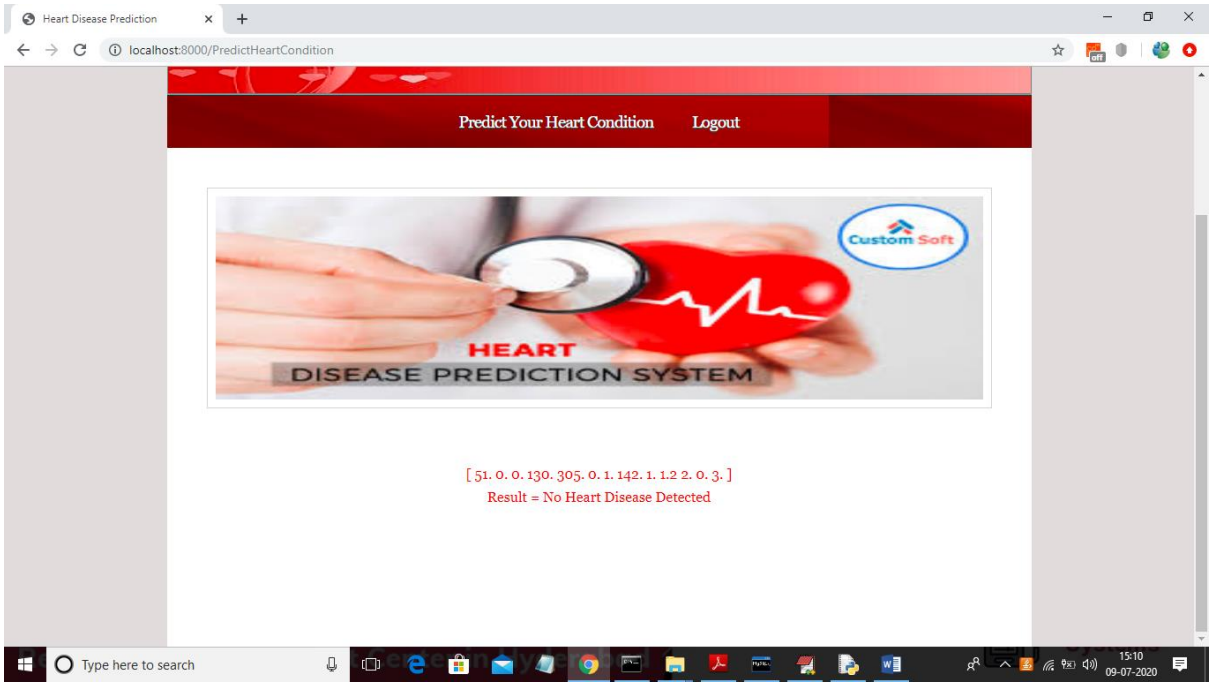


Fig 9.5 Output Screen

10. CONCLUSION AND FUTURE ENHANCEMENT

The proposed system is GUI-based, user-friendly, scalable, reliable and an expandable system. The proposed working model can also help in reducing treatment costs by providing Initial diagnostics in time. The model can also serve the purpose of training tool for medical students and will be a soft diagnostic tool available for physician and cardiologist. Prediction of cardiovascular disease results is not accurate. Data mining techniques does not help to provide effective decision making. Cannot handle enormous datasets for patient records. As an extension to this work, and some sort of limitation to the work performed here, different types of classifiers can be included in the analysis and more in-depth sensitivity analysis can be performed on these classifiers, also an extension can be made by applying same analysis to other bioinformatics diseases' datasets, and see the performance of these classifiers to classify and predict these diseases**.**

11. REFERENCES

[1] Ramadoss and Shah B et al. “A. Responding to the threat of chronic diseases in India”. Lancet. 2005; 366:1744–1749. doi: 10.1016/S0140-6736(05)67343-6.

[2] Global Atlas on Cardiovascular Disease Prevention and Control. Geneva, Switzerland: World Health Organization, 2011.

[3] Dhomse Kanchan B and Mahale Kishor M. et al. “Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analysis”, 2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication.

[4] R. Kavitha and Kannan et al. “An Efficient Framework for Heart Disease Classification using Feature Extraction and Feature Selection Technique in Data Mining “, 2016.

[5] Shan Xu, Tiangang Zhu, Zhen Zang, Daoxian Wang, Junfeng Hu and Xiaohui Duan et al. “Cardiovascular Risk Prediction Method Based on CFS Subset Evaluation and Random Forest Classification Framework”, 2017 IEEE 2nd International Conference on Big Data Analysis.

[6] Manpreet Singh, Levi Monteiro Martins, Patrick Joanis and Vijay K. Mago et al. “Building a Cardiovascular Disease Predictive Model using Structural Equation Model & Fuzzy Cognitive Map”, 978-1-5090-0626-7/16/$31.00 c 2016 IEEE.

[7] Kanika Pahwa and Ravinder Kumar et al. “Prediction of Heart Disease Using Hybrid Technique For Selecting Features”, 2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON).

[8] Seyedamin Pouriyeh, Sara Vahid, Giovanna Sannino, Giuseppe De Pietro, Hamid Arabnia, Juan Gutierrez et al. “A Comprehensive Investigation and Comparison of Machine Learning Techniques in the Domain of Heart Disease”,22nd IEEE Symposium on Computers and Communication (ISCC 2017): Workshops - ICTS4eHealth 2017

[9] Hanen Bouali and Jalel Akaichi et al. “Comparative study of Different classification techniques, heart diseases use Case.”, 2014 13th International Conference on Machine Learning and Applications

[10] Seyedamin Pouriyeh, Sara Vahid, Giovanna Sannino, Giuseppe De Pietro, Hamid Arabnia, Juan Gutierrez et al. “A Comprehensive Investigation and Comparison of Machine Learning Techniques in the Domain of Heart Disease”, 22nd IEEE Symposium on Computers and Communication (ISCC 2017): Workshops - ICTS4eHealth 2017

[11] Houda Mezrigui, Foued Thaliana and Kaouther Laabidi et al. “Decision Support System for Medical Diagnosis Using a KernelBased Approach”, ICCAD’17, Hammamet - Tunisia, January 19- 21, 2017.

12. PUBLICATIONS

CONFERENCE:

International Conference on “Innovations in Computers Networks, Computational Intelligence and IoT” (ICICCI – 21)

Paper ID: ICICCI – 21 – 0156

13. STUDENTS PROFILE

P. Rohith is currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. He has completed his Secondary Education from Gautami Techno School and Higher Secondary Education from Sri Chaitanya Junior College. His technical skills include Data Science, Microsoft Excel, Tableau, SQL, Statistics, Data Science using Python, Machine Learning, Power BI and C. He attended E-summit, Entrepreneurship carnival, which was hosted by EDC – MLRIT in association with Nucleus Tech and SUMVN in the year 2017. He has completed online courses (AWS Fundamentals: Going Cloud Native, Data Science Math Skill, AI for everyone, Leadership and Emotional Intelligence, Managing Project Risks and Fundamentals, Python) from Coursera and Cursa App. He has also done a 6 months Data Science course from reputed online institute Board Infinity and completed projects on every module related with Data Science. He has attended a Virtual training on mastering Data Science in 11hrs conducted by ExcelR and TASK. He is passionate towards learning new technologies and aspires to become a Data Scientist in the near future. He is fond of data analytics and looking forward to show passion in stock market analysis. His personal projects include "Secure Tensor Decomposition using fully homomorphic encryption scheme" which is 3-month internship based on cloud-based security service under the guidance of Lasya Infotech Institution.

G. Aravind is currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. He has completed his Secondary Education from Alphores High School and Higher Secondary Education from Pragathi Junior College. His technical skills include python and Sql. He attended 5 days online International Hands-on-Training in python programming,3-day online workshop on "AI & ML in speech and audio processing" in the year 2020. He has completed online courses (AWS Fundamentals: Going Cloud-Native, AI for everyone, Data science math skill, Leadership and emotional intelligence, Managing project risks and changes) from Coursera and (Python, Sql, Machine learning, Artificial intelligence, Web development) from Cursa.

D. Madhuritha is currently pursuing her Bachelor of technology in the stream of Computer Science and Engineering at St. Martin's Engineering college. She completed her intermediate from Sri Gayatri Junior College and 10th class from St. Martin's High school. Her technical skills include C,C++ and Python. She also has basic understanding of Java. She took part in E-Summit program conducted at Marri Laxman Reddy Institute of Technology in 2018. And also completed few certification courses from online platforms like Coursera and Solo learn app.

S. Sandeep Sai is currently pursuing his Bachelor of technology in the stream of Computer Science and Engineering at St. Martin's Engineering college. He completed his intermediate from Abhyaas Junior College and 10th class from Spr school of excellence. His technical skills include C, C++ and Python. He also has basic understanding of Java. He took part in E-Summit program conducted at Marri Laxman Reddy Institute of Technology in 2018. And also completed few certification courses from online platforms like Coursera and Solo learn app.

A

PROJECT REPORT

On

Sliding Window Blockchain Architecture for Internet of Things

Submitted by

1) T.Amith Krishna(17K81A05N9) 2) B.Sidharth (17K81A05J6)
3) T.Naveen (17K81A05P0) 4) V.Ramya (17K81A05P3)

in partial fulfilment for the award of the

degree of

BACHELOR OF TECHNOLOGY

IN

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Under the Guidance

of

Dr. B.Rajalingam

(B.Tech., M.E., Ph. D)

Associate Professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ST.MARTIN'S ENGINEERING COLLEGE

An Autonomous Institute Dhulapally, Secunderabad – 500100

JUNE 2021

BONAFIDE CERTIFICATE

This is to certify that the project entitled "Sliding window blockchain architecture for internet of things**"**, is being submitted by **1. Mr. T.Amith Krishna 17K81A05N9, 2. Mr. B.Sidharth 17K81A05J6, 3. Mr. T.Naveen 17K81A05P0, 4. Ms. V.Ramya 17K81A05P3** in partial fulfillment of the requirement for the award of the degree of **BACHELOR OF TECHNOLOGY IN COMPUTER SCIENCE AND ENGINEERING** is recorded of bonafide work carried out by them. The result embodied in this report have been verified and found satisfactory.

Assistant Professor

Dr.B.Rajalingam

Department of CSE

Head of the Department

Dr.M.NARAYANAN

Department of CSE

Internal Examiner

External Examiner

Place:

Date:

DECLARATION

We, the student of **Bachelor of Technology** in Department of **Computer Science and Engineering'**, session: 2017 – 2021, St. Martin's Engineering College, Dhulapally, Kompally, Secunderabad, hereby declare that work presented in this Project Work entitled **'Sliding Window Blockchain Architecture for Internet of Things'** is the outcome of our own bonafide work and is correct to the best of our knowledge and this work has been undertaken taking care of Engineering Ethics. This result embodied in this project report has not been submitted in any university for award of any degree.

T.Amith Krishna	17K81A05N9
B.Sidharth	17K81A05J6
T.Naveen	17K81A05P0
V.Ramya	17K81A05P3

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible and whose encouragements and guidance have crowded effects with success.

We extended our deep sense of gratitude to Principal**, Dr. P. SANTOSH KUMARPATRA**, St. Martin's Engineering College, Dhulapally, for permitting us to undertake this project.

We are also thankful to **Dr. M.NARAYANAN**, Head of the Department, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

We would like to thank **Dr. T. POONGOTHAI,** Department of Computer Science and Engineering (AI & ML) for her encouragement and insightful comments and as well as our project coordinators **Dr. B.RAJALINGAM**, Associate Professor and **Dr. R.SANTHOSHKUMAR**, Associate Professor, in Department of Computer Science and Engineering for their valuable support.

We would like to express our sincere gratitude and indebtedness to our project supervisor Dr. B.RAJALINGAM, Associate Professor, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

Finally, we express thanks to all those who have helped us successfully to completing this project. Furthermore, we would like to thank our family and friends for their moral support and encouragement.

We express thanks to all those who have helped us in successfully completing the project.

T.Amith Krishna 17K81A05N9
B.Sidharth 17K81A05J6
T.Naveen 17K81A05P0
V.Ramya 17K81A05P3

ABSTRACT

Internet of Things (IoT) refers to the concept of enabling Internet connectivity and associated services to non-traditional computers formed by integrating essential computing and communication capability to physical things for everyday usage. Security and privacy are two of the major challenges in IoT. The essential security requirements of IoT cannot be ensured by the existing security frameworks due to the constraints in CPU, memory, and energy resources of the IoT devices. Also, the centralized security architectures are not suitable for IoT because they are subjected to single point of attacks. Defending against targeted attacks on centralized resources is expensive.

Therefore, the security architecture for IoT needs to be decentralized and designed to meet the limitations in resources. Blockchain is a decentralized security framework suitable for a variety of applications. However, blockchain in its original form is not suitable for IoT, due to its high computational complexity and low scalability. In this paper, we propose a sliding window blockchain (SWBC) architecture that modifies the traditional blockchain architecture to suit IoT applications. The proposed sliding window blockchain uses previous $(n - 1)$ blocks to form the next block hash with limited difficulty in Proof-of-Work.

The performance of SWBC is analyzed on a real-time data stream generated from a smart home testbed. The results show that the proposed blockchain architecture increases security and minimizes memory overhead while consuming fewer resources.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF OUTPUT SCREENS

LIST OF ACRONYMS

IOT	Internet of Things
SWBC	Sliding Window Blockchain
POW	Proof-of-Work
SDLC	Software Development Life Cycle
FPS	Frames per second
UML	Unified Modelling Language

1. INTRODUCTION

Blockchain is a distributed ledger used to record transactions between two or more parties. Unlike relational database systems, blockchain is a data structure where new entries get appended at the end of the ledger, and there exist no administrator permissions within a blockchain which allow modification of the data. Also, the addition of a new block to the chain needs to be verified by all other parties through a consensus algorithm. Since there exists a distributed control over the blockchain, it is difficult for attackers to modify the data compared to a relational database system. Relational databases are primarily designed for centralized data storage and blockchain are specifically designed for decentralized data storage.

There exist two types of blockchains: (i) Permissioned (ii) Permissionless.

A permissioned blockchain is a private blockchain which requires pre-verification of the participants within the network who are assumed to know each other whereas, a permissionless blockchain is a public blockchain. Traditional blockchain approach is not suitable for IoT with real-time data streams due to their computationally complex Proof-of-Work (PoW) . As the computational time increases, blockchain security becomes infeasible to be used for IoT.

The two major challenges involved in applying blockchain to IoT environments include:

(i) Computational complexity and (ii) scalability.

The computational complexity depends on difficulty level and Merkle tree size. Merkle tree is a tree in which every leaf node is labelled with the hash of a transaction data and every non-leaf node is labelled with the cryptographic hash of the labels of its child nodes. Merkle tree grows with the number of transactions made and, thereby, increasing the time consumed for Proof-of-Work, which is less favourable for an IoT network. Scalability refers to the limits on the number of transactions a blockchain can process within a specific time period. Bitcoin is a popular example of a blockchain. Bitcoin blockchain is a payment system that does not rely on a central authority to secure and control its money supply. Each block in a Bitcoin blockchain has limited block size. In Bitcoin, the block size is limited to 1 MB and a block is mined every ten minutes. Interestingly, the existing literature suggests blockchain as one of the data security and privacy algorithms that can be implemented for IoT applications due to its distributed architecture. The concept of smart homes plays an important role in the planning of future housing-based models of health care. Smart homes use IoT as their ambient networking environment. IoT is a network of

things embedded with sensors and are connected to the Internet . IoT helps to connect the resources of a smart home, both physical and virtual things, that are embedded with electronics, sensors, actuators, and software, to collect and exchange data. MavHome is one of the earliest smart home projects, which created a home that acts as a rational agent. The agent seeks to maximize inhabitant comfort and minimize the operational cost. To achieve the goals, the agent predicts the mobility pattern and the device usage of the inhabitants. As the smart home concept becomes important, it is essential to provide an adequate level of protection against cyberattacks for residential customers. Traditional security solutions tend to be expensive for IoT in terms of processing and memory overhead due to the limited computational power and memory size of IoT devices.

Therefore, the resource constrained nature of IoT devices involved in a smart home makes the standard security solutions infeasible. As a result, smart homes are prone to security vulnerabilities. The major challenges in adopting conventional security mechanisms in IoT include: (i) resource constraints, (ii) heterogeneous communication protocols, (iii) unreliable communications, and (iv) energy constraints. It identified that the traditional security and privacy policies based on asymmetric encryption schemes are difficult to implement in an IoT environment due to its centralized key management system. In such a context, blockchain technologies help to track, coordinate, carry out transactions, and store information from a large number of devices and, thereby, enabling the creation of applications that do not require a centralized cloud. Blockchain forms a decentralized network which enables all parties to make transactions. The blockchain approach has been widely applied in fields including finance, insurance, manufacturing, and health-care. Kshetri demonstrated that blockchain based identity and access management systems have the ability to significantly strengthen IoT security. Dorri etal proposed a hierarchical architecture that uses a centralized private immutable ledger operating at local IoT network level within a smart home to reduce the overhead and a decentralized public blockchain at higher-end devices for better trust. In

, Shen et al. applied blockchain to a smart home system to ensure the security and privacy of information. Christidis et al. used smart contracts in IoT to facilitate the sharing of service resources as well as to automate the process in a cryptographically verifiable manner. Table I provides a comparison of different blockchain architectures proposed for IoT applications.

1.1 PROJECT OVERVIEW

In this paper, we propose new blockchain architecture for IoT environments, especially in the context of smart home applications. A smart home monitors, analyzes, and reports the state of the home. Smart homes use devices connected to IoT to automate and monitor in-home systems. Smart home can be considered as the smallest unit of a smart city. The security standardization of a smart home supports a smart city and vice versa.

In a smart home, the real-time data streams are generated by sensors which help us to monitor the current status of the home, analyze energy consumption, and investigate any accidents inside a smart home. The volume of data generated by a smart home depends on the number of sensors deployed and the frequency of data acquisition. Therefore, proper sampling of sensor data is required to produce meaningful information which can be later stored in the blockchain. The volume of data stored in a blockchain decides the packet overhead, memory overhead, and computational overhead. In this context, our proposed sliding window blockchain architecture tries to improve the security and reduce the memory overhead of IoT in a smart home environment.

IoT is revolutionizing the living environments, therefore, it is necessary to provide security for the data generated from IoT. However, the limitations of CPU, memory, and energy resources of IoT devices make the traditional centralized security algorithms infeasible. Therefore, we propose a sliding window blockchain, which is a decentralized security architecture, that provides security while considering the limitations of IoT devices.

1.2 PROJECT OBJECTIVE

1) A novel Sliding Window Blockchain architecture for IoT is proposed to provide security while considering the limitations of IoT devices.

2) A smart home testbed is set up to implement and analyze the performance of the proposed architecture.

3) Analysis of the performance and security of the proposed sliding window blockchain architecture is carried out on the smart home testbed.

1.3. ORGANIZATION OF CHAPTERS

This documentation consists of 10 different chapter and they are:

1. Introduction – This chapter covers the overview of our project and its objectives.

2. Literature Survey – This includes the details of our survey.

3. Software and Hardware Requirements – We specify our software and hardware requirements here.

4. Software Development Analysis – This section includes the problem definition and details of the modules we used in our project.

5. Project System Design – This chapter includes the design part of our project which includes UML diagrams.

6. Project Coding – This section contains the details of our project code.

7. Project Testing – The details of test cases and testing are included in this chapter.

8. Output Screens – This contains the screenshots of how our project looks like when executed.

9. Experimental Results – This chapter contains the screenshots of our results.

10. Conclusion and Future Enhancements – This covers the conclusion of our project and the possible future developments.

2. LITERATURE SURVEY

Literature survey is the most important step in software development process. Before developing the tool it is necessary to determine the time factor, economy and company strength. Once these things are satisfied, ten next steps are to determine which operating system and language can be used for developing the tool. Once the programmers start building the tool the programmers need lot of external support. This support can be obtained from senior programmers, from book or from websites. Before building the system the above consideration are taken into account for developing the proposed system.

.

3. SOFTWARE AND HARDWARE REQUIREMENTS

Requirement is a condition or capability possessed by the software or system component in order to solve a real world problem. The problems can be to automate a part of a system, to correct shortcomings of an existing system, to control a device, and so on. Requirements describe how a system should act, appear or perform. For this, when users request for software, they provide an approximation of what the new system should be capable of doing. Requirements differ from one user to another and from one business process to another. The purpose of the requirements document is to provide a basis for the mutual understanding between the users and the designers of the initial definition of the software development life cycle (SDLC) including the requirements, operating environment and development plan. Requirements help to understand the behaviour of a system, which is described by various tasks of the system. For example, some of the tasks of a system are to provide a response to input values, determine the state of data objects, and so on. Note that requirements are considered prior to the development of the software. The requirements, which are commonly considered, are classified into three categories, namely, functional requirements, non-functional requirements, and domain requirements. The functional requirements should be complete and consistent. Completeness implies that all the user requirements are defined. Consistency implies that all requirements are specified clearly without any contradictory definition. Generally, it is observed that completeness and consistency cannot be achieved in large software or in a complex system due to the problems that arise while defining the functional requirements of these systems. The different needs of stakeholders also prevent the achievement of completeness and consistency. Due to these reasons, requirements may not be obvious when they are,first specified and may further lead to inconsistencies in the requirements specification. The non-functional requirements (also known as quality requirements) are related to system attributes such as reliability and response time. Non-functional requirements arise due to user requirements, budget constraints, organizational policies, and so on. These requirements are not related directly to any particular function provided by the system. 8 Non-functional requirements should be accomplished in software to make it perform efficiently. For example, if an aeroplane is unable to fulfil reliability requirements, it is not approved for safe operation. Similarly, if a real time control system is ineffective in accomplishing non-functional requirements, the control functions cannot operate correctly. System requirements are the configuration that a system must have in order for a hardware or software application to run smoothly and efficiently. Failure to meet these requirements can result in installation problems or performance problems. The former may prevent a device or application from getting installed, whereas the latter may cause a product to malfunction or perform below expectation or even to

hang or crash. System requirements are also known as minimum system requirements. Hardware system requirements often specify the operating system version, processor type, memory size, available disk space and additional peripherals, if any, needed. Software system requirements, in addition to the requirements, may also specify additional software dependencies (e.g., libraries, driver version, framework version). Some hardware/software manufacturers provide an upgrade assistant program that users can download and run to determine whether their system meets a product's requirements. Some products include both minimum and recommended system requirements. A video game, for instance, may function with the minimum required CPU and GPU, but it will perform better with the recommended hardware. A more powerful processor and graphics card may produce improved graphics and faster frame rates (FPS). Some system requirements are not flexible, such as the operating system(s) and disk space required for software installation. Others, such as CPU, GPU, and RAM requirements may vary significantly between the minimum and recommended requirements. When buying or upgrading a software program, it is often wise to make sure your system has close to the recommended requirements to ensure a good user experience.

3.1 SOFTWARE REQUIREMENTS

• Operating System: Windows XP.

• Platform: PYTHON TECHNOLOGY

• Tool: Spyder, Python 3.5

• Front End: Anaconda

• Back End: Python Anaconda Script

3.2 HARDWARE REQUIREMENTS

• System: Pentium IV 2.4 GHz.

• Hard Disk: 40 GB.

• Monitor: 15 inch VGA Color.

• Mouse: Logitech Mouse.

• Ram: 512 MB

• Keyboard: Standard Keyboard

4. SOFTWARE DEVELOPMENT ANALYSIS

Software development is a process of writing and maintaining the source code, but in a broader sense, it includes all that is involved between the conception of the desired software through to the final manifestation of the software, sometimes in a planned and structured process. Therefore, software development may include research, new development, prototyping, modification, reuse, re-engineering, maintenance, or any other activities that result in software products.

4.1 OVERVIEW OF THE PROBLEM

Blockchain uses an immutable distributed ledger which replicates a copy of the ledger across the authorized parties and, thereby, preventing the single point of vulnerability that is prone to be exploited. However, blockchain faces critical challenges for its application in an IoT environment. The Proof-of-Work calculation is computationally intensive and time-consuming. Since the majority of IoT devices are resource-constrained and most IoT applications need low latency, application of traditional blockchain becomes infeasible. The Merkle tree implementation becomes a bottleneck for IoT due to the existence of numerous sensors in typical IoT deployment. The underlying blockchain protocols create significant network overhead, which is not suitable for communication among IoT devices.

4.2 DEFINING THE PROBLEM

The objective of this paper is to propose a blockchain architecture to increase the IoT security, reduce the memory overhead and network overhead, and also to evaluate the performance of the blockchain architecture on a smart home. The prototype of a smart home environment has a heterogeneous network consisting of Wi-Fi, Zigbee, and Bluetooth technologies. Different organizations such as hospital, health insurance, police, and NGO are connected to the smart home for smart management. The smart home owner chooses the organization that needs to be connected with the home. The organizations act as miners. In our prototype, sliding window blockchain is used to securely store the state of home as well as the transactions between the organizations through a secure channel. The blockchain used is private and permissioned. The data generated by the smart home are encrypted before it is being stored in the blockchain and the key for encryption is shared only with the concerned organization. The owner and organizations together form smart contracts and blocks are added to the chain only if all the group members of the blockchain validate the block.

5. PROJECT SYSTEM DESIGN

A blockchain is a growing list of records, called blocks that are linked using cryptographically generated hashes. Figure 1 shows the basic architecture of a simplified blockchain. Each block contains a cryptographic hash of the previous block, chaining the blocks together. Chaining blocks together makes it impossible to modify transactions included in any block without modifying all subsequent blocks. As a result, the cost of modifying a particular block increases with every new block added to the blockchain, magnifying the effect of the Proof of-Work.

A block consists of timestamp, hash of the previous block, nonce (value representing the iteration for which the Proof of-Work gets solved), and the transaction data (represented as a Merkle tree root). The first block of a blockchain is called genesis block and has no previous block hash. Each block in the blockchain is added through the process of mining (Proof-of-Work) which validates whether the transactions are legal. When the majority of the miners validate the block, a consensus is sent to miners to add the block to the blockchain.

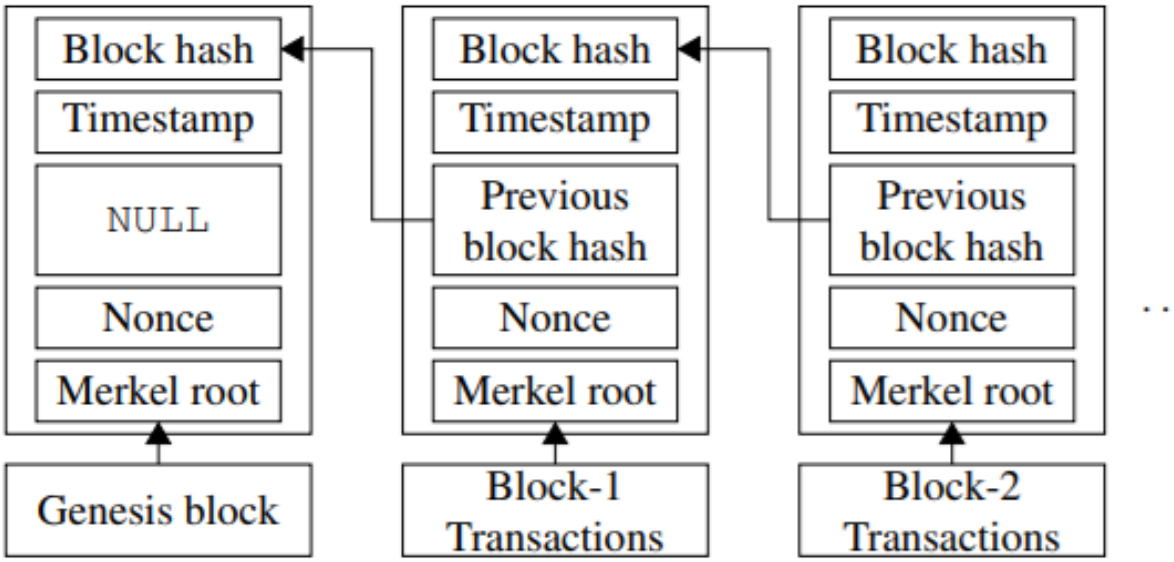


Figure 1: Blockchain architecture.

The Sliding Window Blockchain (SWBC) utilizes a window that slides through the blockchain for every block addition. The window initially consists of one block and increases up to n blocks as defined by the window size. The blocks in the sliding window are used while creating a new block. In the proposed SWBC architecture, the block hash is generated by hashing the blocks in the window as shown in Figure 2. The size of the sliding window determines the number of recent past blocks used to perform the hash update function. The sliding window blockchain has a computational overhead of O(n) for a constant difficulty of mining, where n is the number of blocks in the window used for the hash updatefunction.Sliding window improves the immutability of the blockchain records.

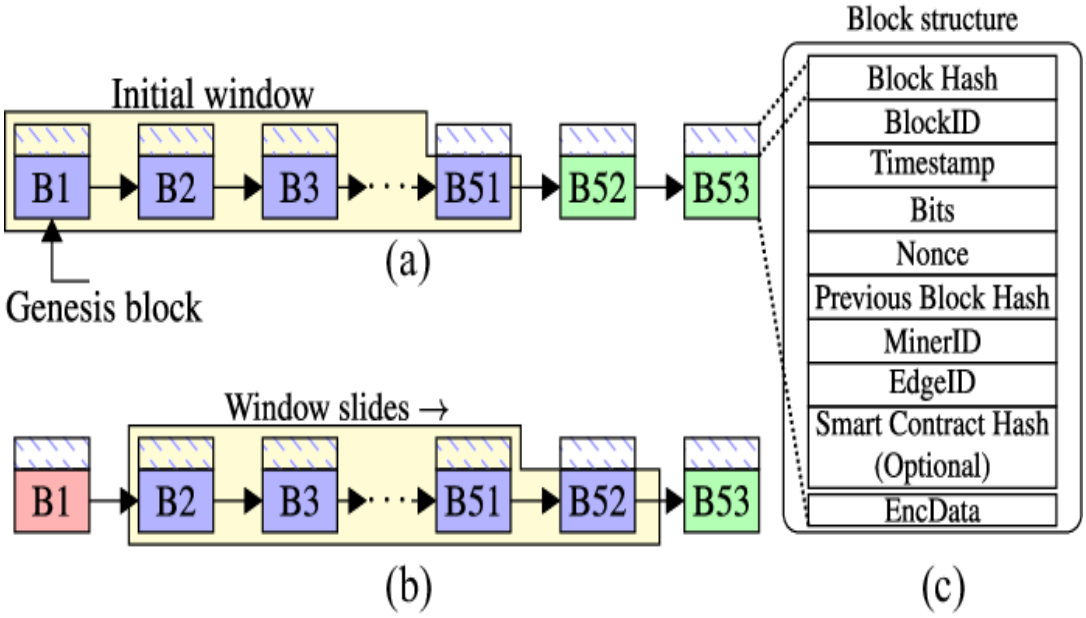


Figure 2: Sliding window blockchain

A false miner requires previous (n − 1) blocks and the window size n to mine a block. The window size is kept secret and sent only to the miners along with the genesis block. The limited part of the chain, i.e., the recent n blocks is stored in the memory of IoT device and the whole blockchain is stored in a private cloud. When the window slides, the older block comes out of the window and is deleted from the IoT device memory. Therefore, the memory overhead to store the blocks in IoT device is reduced. The SWBC structure and its comparison with a Bitcoin blockchain are discussed in the following sections.

5. DATA FLOW DIAGRAMS

5.1 CONTEXT-LEVEL DIAGRAM:

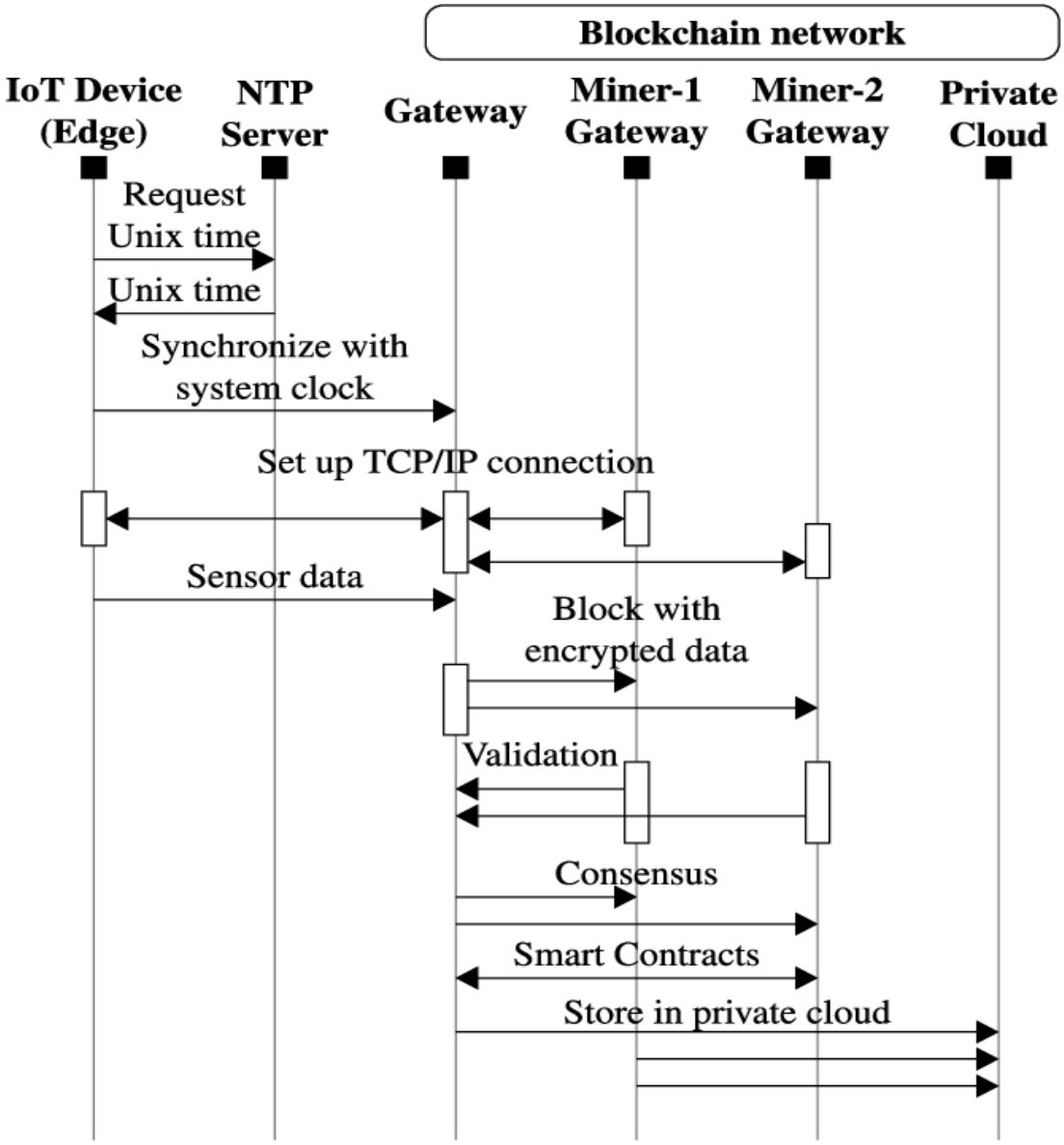


Figure 5.1: Context-level diagram

When the validation is received from the miners, the owner sends the consensus message to add the block. All the miners have the privilege to create a block and send it to the group for validation. The access permissions and privileges of the miners are registered on the smart contracts formed by the group. The communication between IoT device and blockchain is shown in Figure 5.1.1.

5.2 ENTITY-RELATIONSHIP DIAGRAM

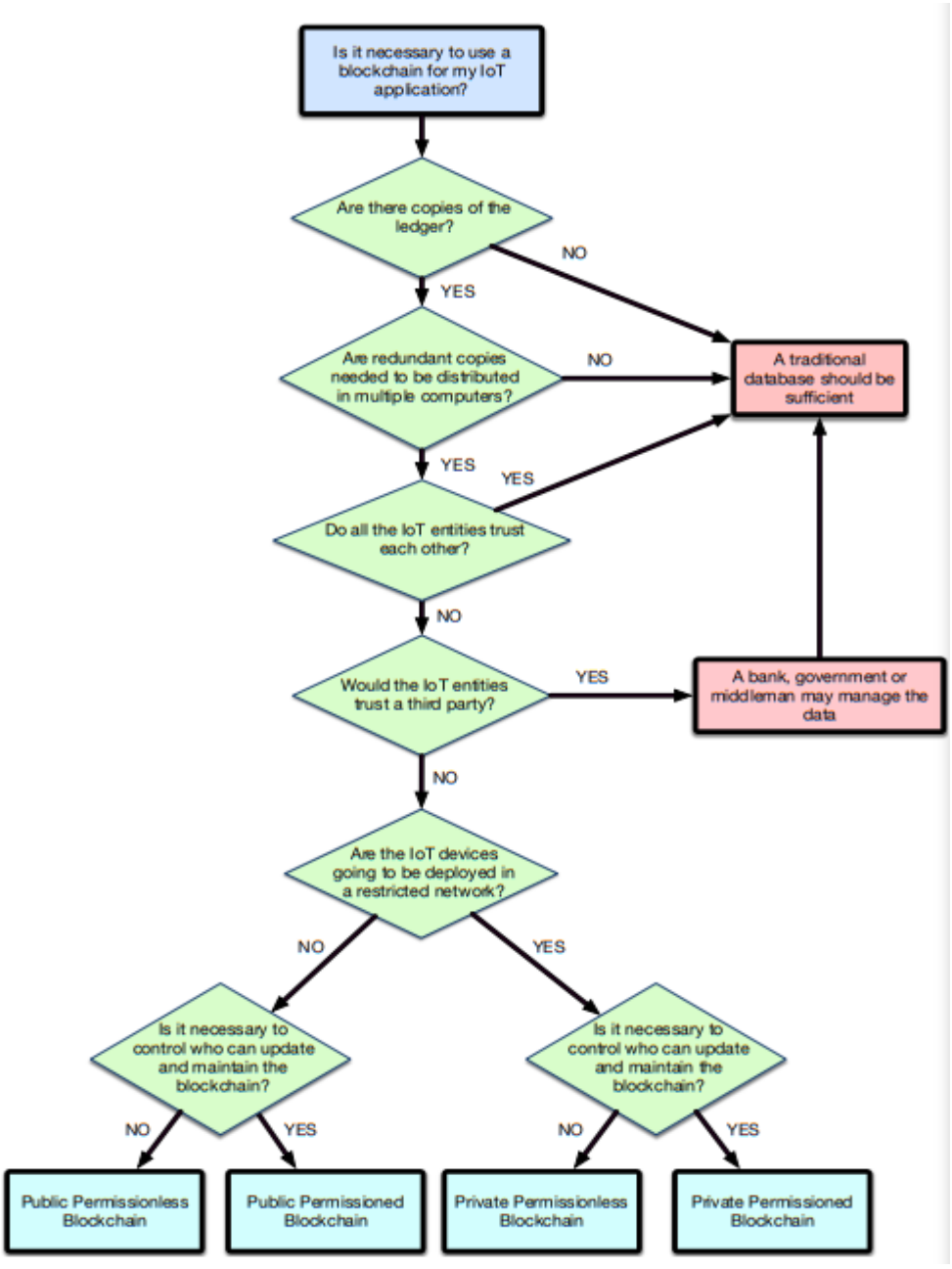


Figure 5.2: ER Diagram

5.3 UML DIAGRAMS

UML stands for Unified Modeling Language. UML is a standardized general-purpose modeling language in the field of object-oriented software engineering. The standard is managed, and was created by, the Object Management Group.

The goal is for UML to become a common language for creating models of object oriented computer software. In its current form UML is comprised of two major components: a Meta-model and a notation. In the future, some form of method or process may also be added to; or associated with, UML.

5.3.1 USE CASE DIAGRAM:

A use case diagram in the Unified Modelling Language (UML) is a type of behavioural diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.

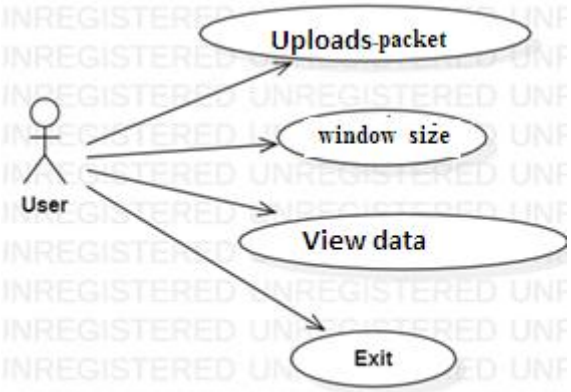


Figure 5.3.1: Use case Diagram

5.3.2 CLASS DIAGRAM:

In software engineering, a class diagram in the Unified Modelling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.

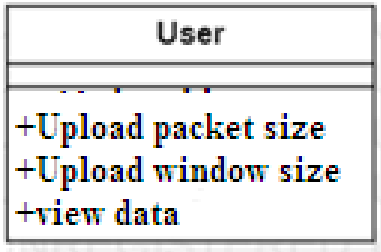


Figure 5.3.2 : Class Diagram

5.3.3 SEQUENCE DIAGRAM:

A sequence diagram in Unified Modelling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.

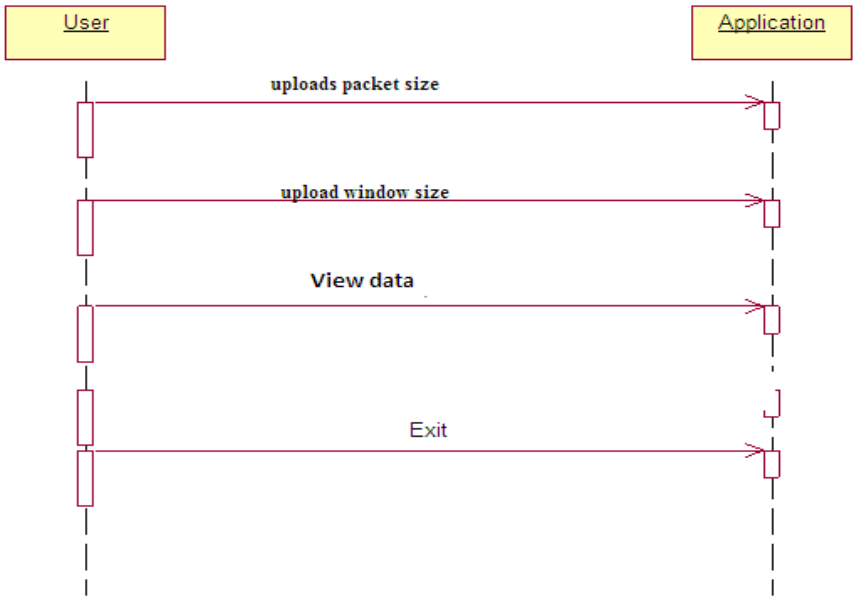


Figure 5.3.3: Sequence Diagram

5.3.4 ACTIVITY DIAGRAM:

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modelling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.

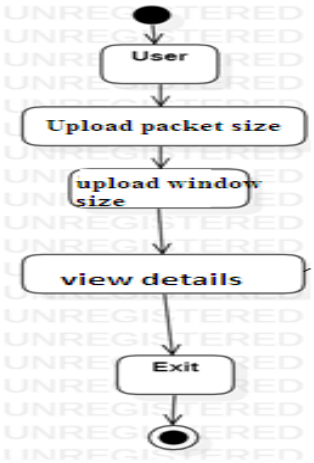


Figure 5.3.4: Activity Diagram

6. PROJECT CODING

Project Coding is the process of designing and building an executable computer program to accomplish a specific computing result or to churn out a particular prototype or product. Programming involves tasks such as: analysis, generating algorithms, profiling algorithms' accuracy and resource consumption, and the implementation of algorithms in a chosen programming language (commonly referred to as coding). The source code of a program is written in one or more languages that are intelligible to programmers, rather than machine code, which is directly executed by the central processing unit. The purpose of programming is to find a sequence of instructions that will automate the performance of a task (which can be as complex as an operating system) on a computer, often for solving a given problem. Proficient programming thus often requires expertise in several different subjects, including knowledge of the application domain, specialized algorithms, and formal logic.

6.1 CODING TEMPLATES

1) Below screen shot of code showing how encryption and block chain hash generation take place

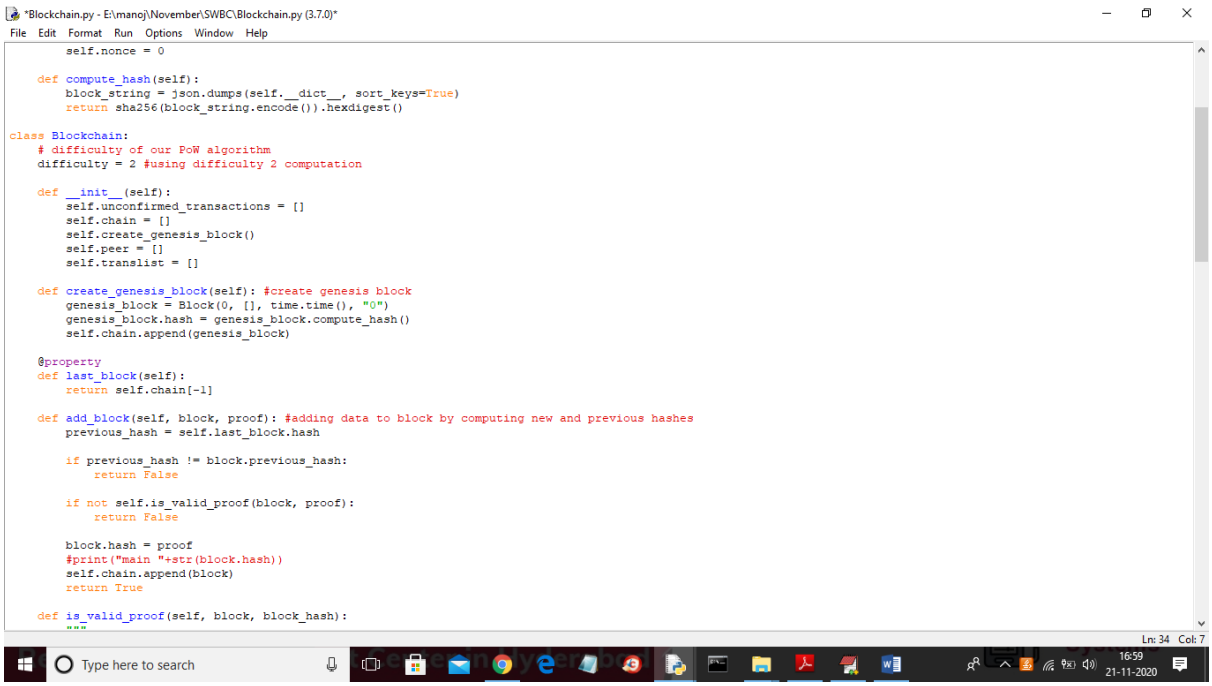


Figure 6.1:Code Template

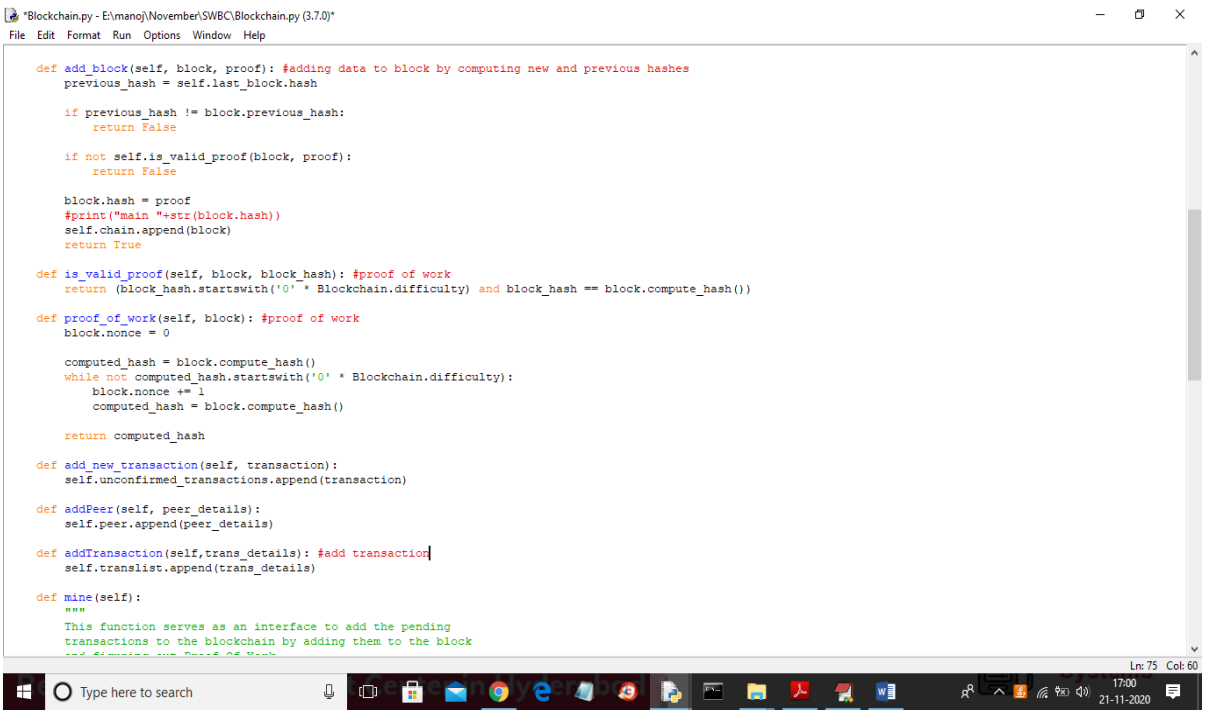


Figure 6.2: Code Template

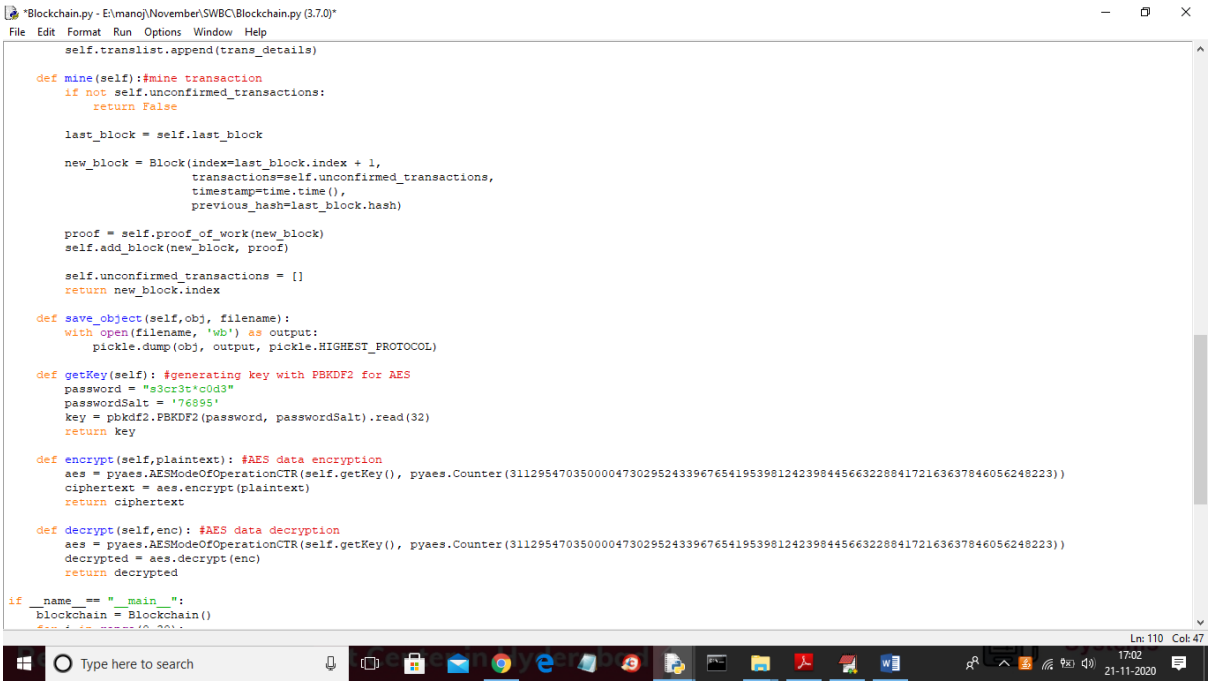


Figure 6.3: Code Template

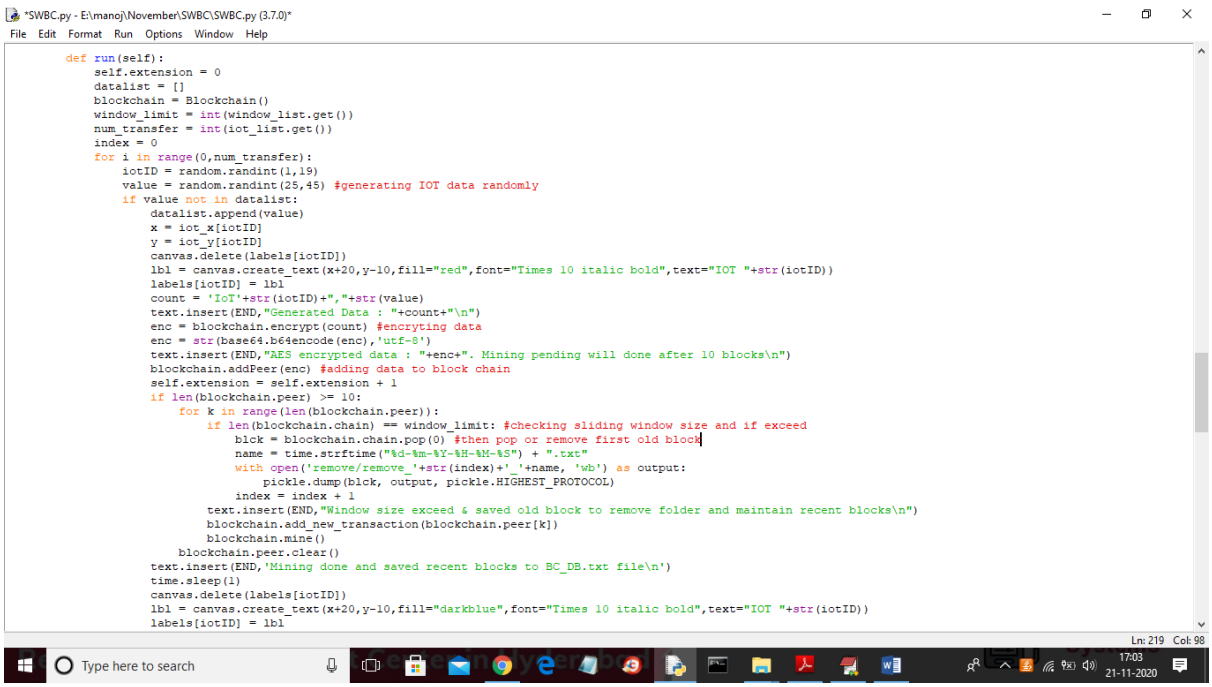


Figure 6.4: Code Template

6.2 OUTLINE FOR VARIOUS FILE

We used Python programming to implement our project. A single python file is used to implement ourcode.

This file consists of various modules that we have used. Our project modules are – Home page, Register page, Login page, User.

We also used various python modules like tkinter, matplotlib, NumPy, imutils, os, cv2.

6.3 CLASS WITH FUNCTIONALITY

There are multiple classes in our code, some of which are:

1) Templates: They handle the backend context processors for requests, debugging, authentication and messaging.

2) Middleware: The middleware class handles the view, sessions, security and frame options.

3) Django Installed Apps: These handle the content types and static files through the code.

4) Validator: They handle the validation of multiple users in and out of the website based on the credentials that the user would provide.

5) Database Linking: This handlesthe backend linking ofthe entire code to the MySQL server in order to facilitate the final output.

6.4. METHODS INPUT AND OUTPUT PARAMETERS

We implemented multiple methods, few of which are :

1. enterframesize()

2. enterpacketsize()

3. runSWBC()

4. runextensionalgorithm()

5. executetable()

6.comparegraph(), etc

7. PROJECT TESTING

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

TYPES OF TESTS

Unit testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

Integration testing

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

Functional test

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centred on the following items:

Valid Input : identified classes of valid input must be accepted.

Invalid Input : identified classes of invalid input must be rejected.

Functions : identified functions must be exercised.

Output : identified classes of application outputs must be exercised.

Systems/Procedures : interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

System Test

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

White Box Testing

White Box Testing is a testing in which in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is purpose. It is used to test areas that cannot be reached from a black box level.

Black Box Testing

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box .you cannot “see” into it. The test provides inputs and responds to outputs without considering how the software works.

Unit Testing

Unit testing is usually conducted as part of a combined code and unit test phase of the software lifecycle, although it is not uncommon for coding and unit testing to be conducted as two distinct phases.

Test strategy and approach

Field testing will be performed manually and functional tests will be written in detail.

Test objectives

- All field entries must work properly.
- Pages must be activated from the identified link.
- The entry screen, messages and responses must not be delayed.

Features to be tested

- Verify that the entries are of the correct format
- No duplicate entries should be allowed
- All links should take the user to the correct page.

Integration Testing

Software integration testing is the incremental integration testing of two or more integrated software components on a single platform to produce failures caused by interface defects.

The task of the integration test is to check that components or software applications, e.g. components in a software system or – one step up – software applications at the company level – interact without error.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

Acceptance Testing

User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

8. OUTPUT SCREENS

An output screen is a device used to display output. An output screen could be a separate monitor or another display device used only to display the output being received from the computer or other devices.

8.1 USER INTERFACE

In below screen select number of packet transfer and then select window size as 5, 10 or 15 and then click on ‘Create Smart Home IOT Network’ button to get below screen

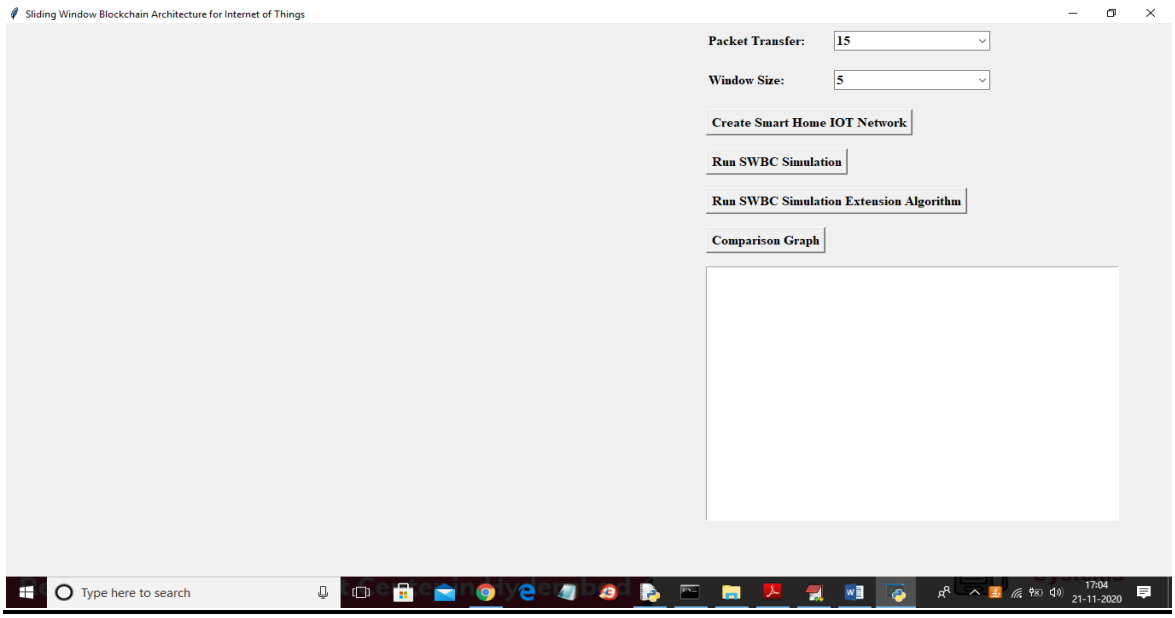


Figure 8.1: Home page of Smart Home IOT Network

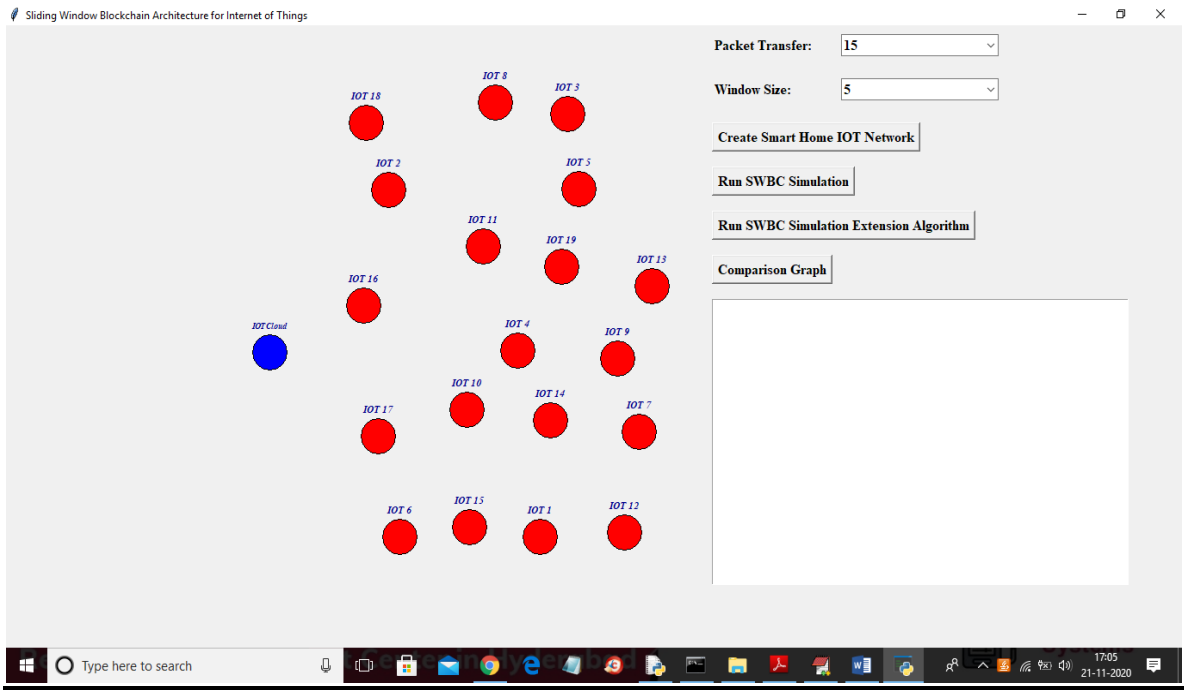


Figure 8.1.1: Displaying IoT devices

In above screen , selected number of packet transfer as 15 and window size is 5 and block chain can store data up to 5 blocks and if exceed then old block remove out and send to cloud for storage and new block will store in IOT memory. In above screen all red colour circles are home IOT and blue colour circle is the IOT cloud which will receive data from IOT upon IOT window full. Now click on 'Run SWBC Simulation' button to allow each circle to sense data randomly and while sensing circle label will change to red colour

8.2 OUTPUT SCREENS

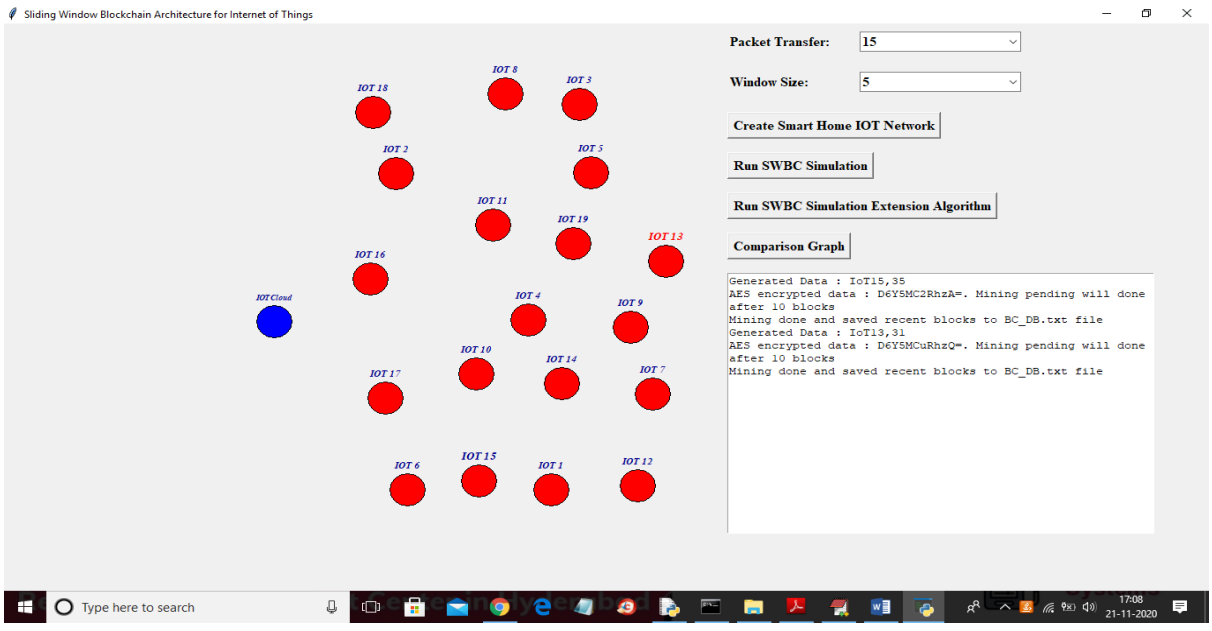


Figure 8.2: Sensing data packets from each IoT device

In above screen IOT13 label is sensing data and its label colour change to red colour and this simulation will run for 15 packets transfer and for each transfer sensor will be chosen randomly. In text area we can see which IOT is sensing data and its sense value separated by comma symbol. In next line displaying AES encrypted data and then displaying mining is done or not and after simulation will get below screen

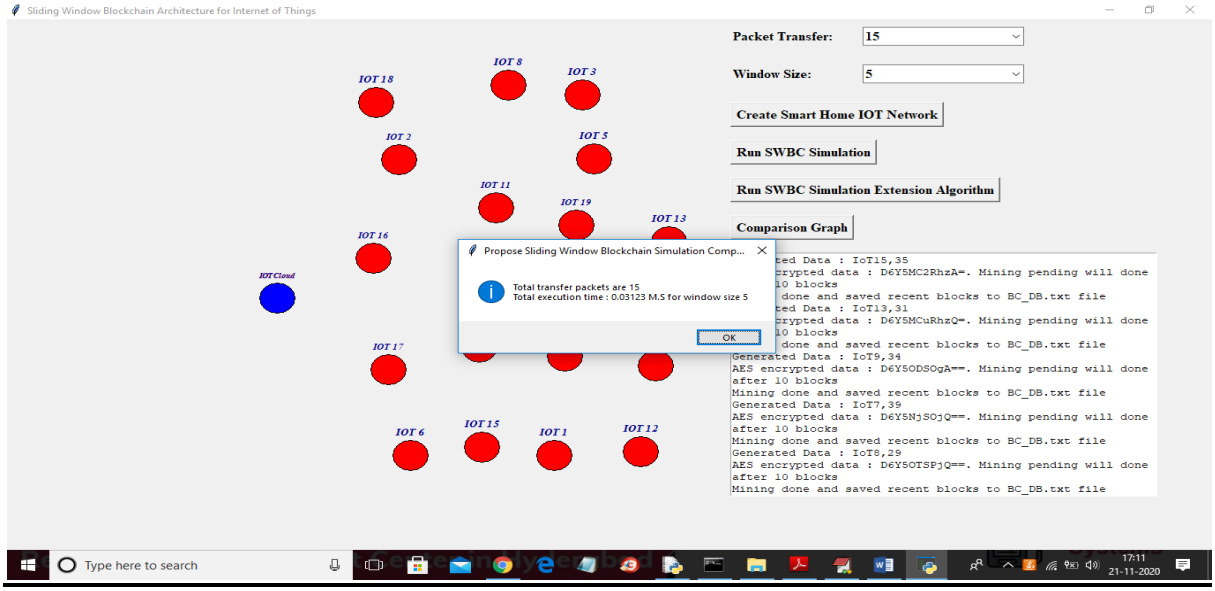


Figure 8.3: Displaying packets in dialog box

In above screen after sending packets we will get above dialog box with total packets sense and send and it will display how much time it took to process that window size 5 and displaying total sense and send packets as 15. In below screen we can see latest recent blocks store at IOT memory

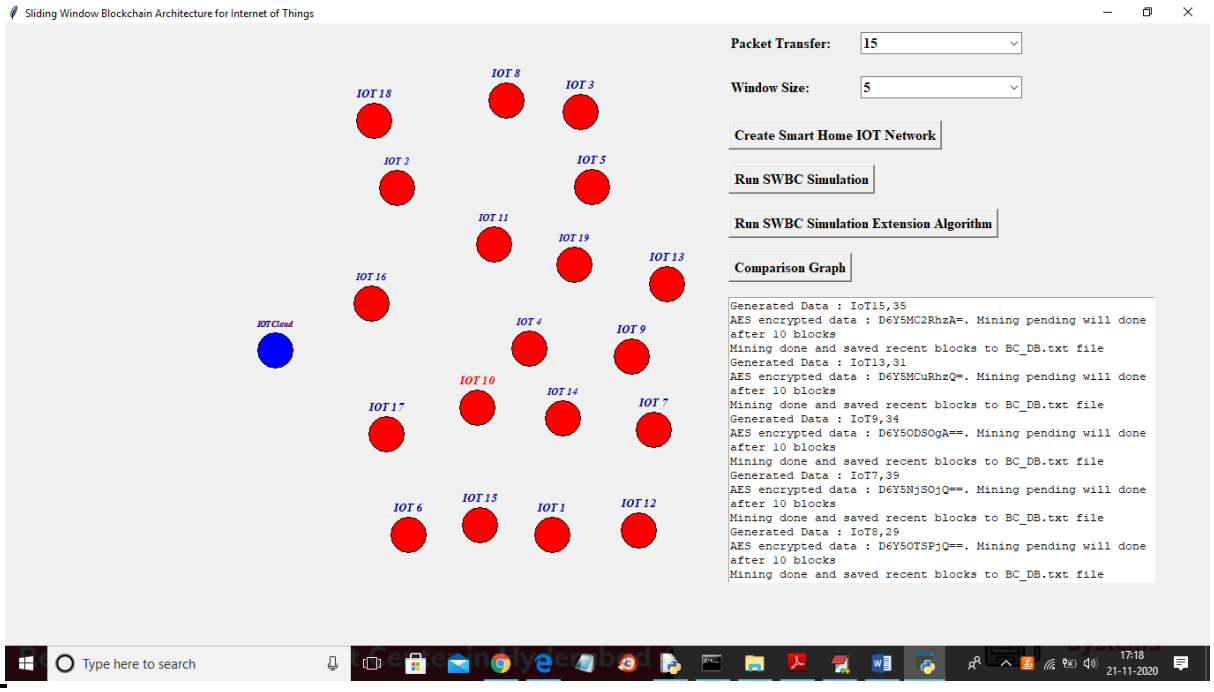


Figure 8.4: Detecting duplicate data packets

In above screen also IOT start sensing and sending packets and in above screen IOT10 is changed to red colour which means its sensing and sending data and after all 15 packets transfer will get below screen

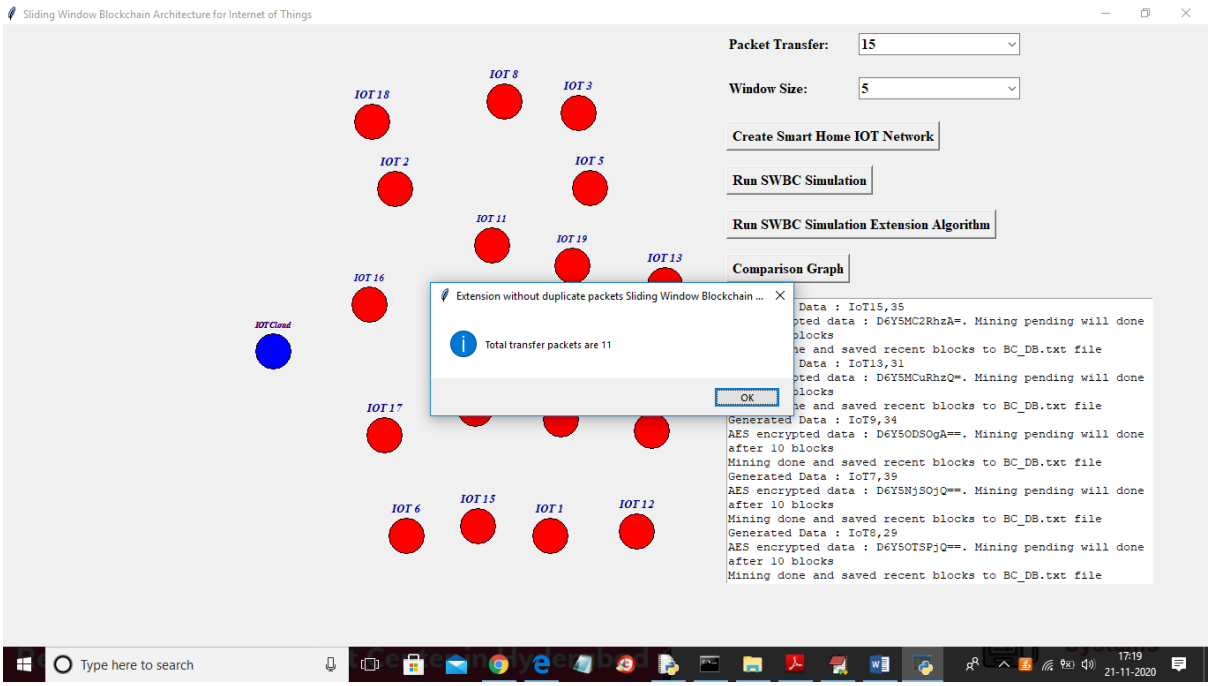
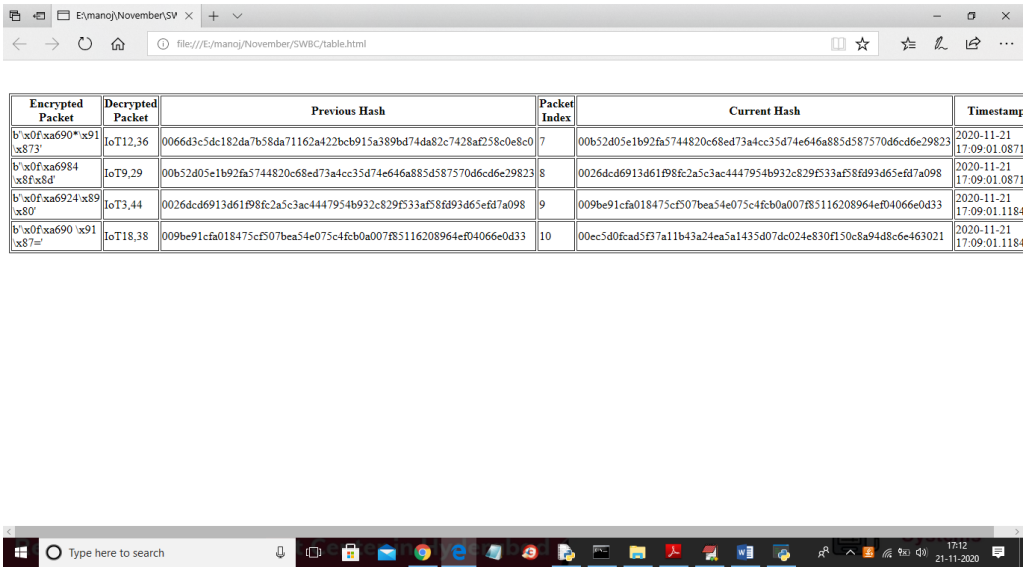


Figure 8.5: Displaying number of transferred data packets without duplication

In above screen with extension work from 15 packets we process only 11 packets and 4 duplicate packets avoided and this 4 packets energy consumption will be saved. Now click on ‘Comparison Graph’ button to get below graph

9. EXPERIMENTAL RESULTS



Encrypted Packet	Decrypted Packet	Previous Hash	Packet Index	Current Hash	Timestamp
b'\x0f\xa690*\x91 \x873'	IoT12,36	0066d3c5dc182da7b58da71162a422bcb915a389bd74da82c7428af258c0e8c0	7	00b52d05e1b92fa5744820c68ed73a4cc35d74e646a885d587570d6cd6e29823	2020-11-21 17:09:01.0871
b'\x0f\xa6984 \x8f\x8d'	IoT9,29	00b52d05e1b92fa5744820c68ed73a4cc35d74e646a885d587570d6cd6e29823	8	0026dcd6913d61f98fc2a5c3ac4447954b932c829f533af58fd93d65efd7a098	2020-11-21 17:09:01.0871
b'\x0f\xa6924\x89 \x80'	IoT3,44	0026dcd6913d61f98fc2a5c3ac4447954b932c829f533af58fd93d65efd7a098	9	009be91cfa018475cf507bea54e075c4fcb0a007f85116208964ef04066e0d33	2020-11-21 17:09:01.1184
b'\x0f\xa690 \x91 \x87='	IoT18,38	009be91cfa018475cf507bea54e075c4fcb0a007f85116208964ef04066e0d33	10	00ec5d0fcad5f37a11b43a24ea5a1435d07dc024e830f150c8a94d8c6e463021	2020-11-21 17:09:01.1184

Figure 9.1: Displaying latest four records of IoT devices

In above screen as our window size was 5 and first block is empty for genesis and latest 4 records of IOT are displaying and in above screen in first column showing encrypted data then in second column decrypted data and then showing previous hash value and then block chain index value and then current hash value with time. In above screen we can see that block chain verify previous and current hash value. In above screen we can see current hash of first row is matched with previous hash of second row. In above screen with propose work we sense and store 15 packets and sometime IOT sensor will sense same data as temperature will not change for some intervals and if we send same data again and again then it waste processing time and increase overhead. We can avoid this overhead by monitoring data and If same data generate again then we will not process in extension work. Now click on 'Run SWBC Simulation Extension Algorithm' button to avoid duplicate processing

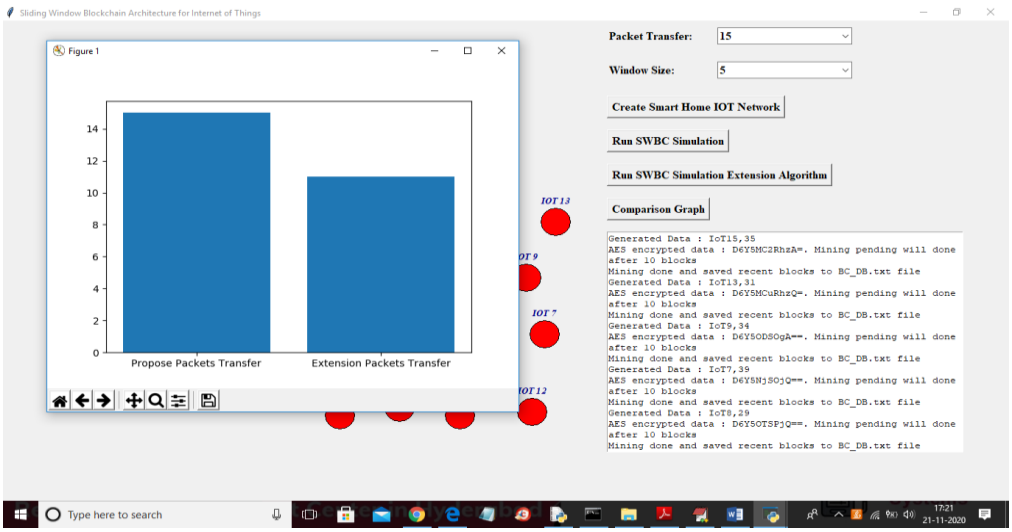


Figure 9.2: Graph between proposed and extension packet transfer

In above graph x-axis represents algorithm name and y-axis represents number of packets transfer and with extension work application process only 11 packets and can save energy of 4 packets.

10. CONCLUSION AND FUTURE ENHANCEMENTS

IoT devices face constraints on resources such as computational capability, energy sources, and memory. Therefore, the standard security algorithms are not feasible for IoT. We proposed a sliding window blockchain that meets the requirements of a resource constrained IoT network by reducing the memory overhead and limiting the computational overhead. The memory overhead is reduced by storing only a limited part of the blockchain, as defined by the sliding window size in the IoT device and maintaining the whole blockchain in the private cloud. Computational overhead is limited by using the difficulty level between 1 and 5 and by eliminating the Merkle tree. The security is increased by generating the block hash using the properties of n blocks in the sliding window. A false miner cannot mine a block unless he gets the previous (n−1) blocks and the window size information.

From the experimental results, we observed the following:

(i)The computational time of PoW for each level of difficulty increases exponentially.

(ii) The total block addition time increases with the increase in the number of miners in the group.

(iii)As the window size increases, the hash computation time increases linearly.

(iv) A random selection of difficulty for each block in a blockchain reduces the total block addition time.

Future work can be carried out to analyse the impact of a variable size sliding window. New consensus algorithms can be developed to suit the IoT environment. Furthermore, energy consumption of the blockchain can also be analysed to draw more insights on energy resources required for an IoT device.

11. REFERENCES

[1] S. Kulkarni, “The beauty of the blockchain,” Open Source for You, vol. 06, pp. 22–24, June 2018.

[2] T. M. F. Carames and P. F. Lamas, “A review on the use of blockchain for the Internet of Things,” IEEE Access, vol. 6, pp. 32 979–33 001, May 2018.

[3] A. Dorri, S. S. Kanhere, and R. Jurdak, “Blockchain in Internet of Things: challenges and solutions,” arXiv preprint arXiv:1608.05187, August 2016.

[4] IoT Agenda, “Smart home or building,” April 2018. [Online]. Available: https://internetofthingsagenda.techtarget.com/definition/ smart-home-or-building

[5] L. Jiang, D. Y. Liu, and B. Yang, “Smart home research,” in Proceedings of 2004 International Conference on Machine Learning and Cybernetics, vol. 2, August 2004, pp. 659–663.

[6] theinstitute.ieee.org, “Towards a definition of the Internet of Things (IoT),” May 2015. [Online]. Available: https://iot.ieee.org/images/files/pdf/IEEE IoT Towards Definition Internet of Things Revision1 27MAY15.pdf

[7] J. Wan, X. Gu, L. Chen, and J. Wang, “Internet of Things for ambient assisted living: Challenges and future opportunities,” in International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), October 2017, pp. 354–357.

[8] D. Abbasinezhad-Mood, A. Ostad-Sharif, and M. Nikooghadam, “Novel anonymous key establishment protocol for isolated smart meters,” IEEE Transactions on Industrial Electronics, vol. 67, no. 4, pp. 2844–2851, April 2020.

[9] S. K. Das, D. J. Cook, A. Battacharya, E. O. Heierman, and T. Y. Lin, “The role of prediction algorithms in the MavHome smart home architecture,” IEEE Wireless Communications, vol. 9, no. 6, pp. 77–84, December 2002.

[10] C. Qu, M. Tao, J. Zhang, X. Hong, and R. Yuan, “Blockchain based credibility verification method for IoT entities,” Security and Communication Networks, vol. 2018, pp. 1–11, June 2018.

[11] C. Lee, L. Zappaterra, K. Choi, and H. A. Choi, “Securing smart home: Technologies, security challenges, and security requirements,” in IEEE Conference on Communications and Network Security, October 2014, pp. 67–72.

[12] P. Treleaven, R. G. Brown, and D. Yang, “Blockchain technology in finance,” Computer, vol. 50, no. 9, pp. 14–17, September 2017.

[13] V. Gatteschi, F. Lamberti, C. Demartini, C. Pranteda, and V. Santamar´ıa, “To blockchain or not to blockchain: That is the question,” IT Professional, vol. 20, no. 2, pp.

62–74, March 2018.

[14] P. A. Laplante and B. Amaba, “Introducing the Internet of Things department,” IT Professional, vol. 20, no. 1, pp. 15–18, January 2018.

[15] C. Esposito, A. D. Santis, G. Tortora, H. Chang, and K. R. Choo, “Blockchain: A panacea for healthcare cloud-based data security and privacy,” IEEE Cloud Computing, vol. 5, no. 1, pp. 31–37, January 2018.

[16] N. Kshetri, “Can blockchain strengthen the Internet of Things,” IT Professional, vol. 19, no. 4, pp. 68–72, July/August 2017.

[17] A. Dorri, S. S. Kanhere, and R. Jurdak, “Towards an optimized blockchain for IoT,” in Proceedings of the Second International Conference on Internet of Things Design and Implementation. ACM, August 2017, pp. 173–178.

[18] J. Shen, C. Wang, T. Li, X. Chen, X. Huang, and Z. H. Zhan, “Secure data uploading scheme for a smart home system,” Information Sciences, vol. 453, pp. 186–197, July 2018.

[19] K. Christidis and M. Devetsikiotis, “Blockchains and smart contracts for the Internet of Things,” IEEE Access, vol. 4, pp. 2292–2303, January 2016.

[20] G. Zyskind, O. Nathan, and A. Pentland, “Enigma: Decentralized computation platform with guaranteed privacy,” arXiv preprint arXiv:1506.03471, 2015.

[21] B. Liu, X. L. Yu, S. Chen, X. Xu, and L. Zhu, “Blockchain based data integrity service framework for IoT data,” in 2017 IEEE International Conference on Web Services (ICWS), June 2017, pp. 468–475.

[22] A. Bahga and V. K. Madisetti, “Blockchain platform for industrial Internet of Things,” Journal of Software Engineering and Applications, vol. 9, no. 10, p. 533, October 2016.

[23] A. Boudguiga, N. Bouzerna, L. Granboulan, A. Olivereau, F. Quesnel, A. Roger, and R. Sirdey, “Towards better availability and accountability for IoT updates by means of a blockchain,” in 2017 IEEE European Symposium on Security and Privacy Workshops (EuroS PW), April 2017, pp. 50–58.

[24] R. Di Pietro, X. Salleras, M. Signorini, and E. Waisbard, “A blockchainbased trust system for the Internet of Things,” in Proceedings of the 23nd ACM on Symposium on Access Control Models and Technologies. ACM, June 2018, pp. 77–83.

[25] P. Otte, M. de Vos, and J. Pouwelse, “Trustchain: A sybil-resistant scalable blockchain,” Future Generation Computer Systems, pp. 12–23, July 2017.

[26] Feng Tian, “A supply chain traceability system for food safety based on HACCP, blockchain & Internet of Things,” in 2017 International Conference on Service Systems and Service Management, June 2017, pp. 1–6.

[27] T. Bocek, B. B. Rodrigues, T. Strasser, and B. Stiller, “Blockchains everywhere - a use-case of blockchains in the pharma supply-chain,” in 2017 IFIP/IEEE Symposium on Integrated Network and Service Management (IM), May 2017, pp. 772–777.

[28] A. Dorri, S. S. Kanhere, R. Jurdak, and P. Gauravaram, “Blockchain for IoT security and privacy: The case study of a smart home,” in 2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), March 2017, pp. 618–623.

[29] M. Samaniego and R. Deters, “Blockchain as a service for IoT,” in 2016 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), December 2016, pp. 433–436.

[30] B. Kaliski, “Password-based cryptography specification version 2.0,” Network Working Group, RSA Laboratories, pp. 1–34, September 2000.

[31] Z. Zheng, S. Xie, H. Dai, X. Chen, and H. Wang, “An overview of blockchain technology: Architecture, consensus, and future trends,” in 2017 IEEE International Congress on Big Data (BigData Congress), June 2017, pp. 557–564.

[32] D. Abbasinezhad-Mood and M. Nikooghadam, “Design and extensive hardware performance analysis of an efficient pairwise key generation scheme for smart grid,” International Journal of Communication Systems, vol. 31, no. 5, p. e3507, 2018.

[33] H. Wang, Y. Wang, Z. Cao, Z. Li, and G. Xiong, “An overview of blockchain security analysis,” in China Cyber Security Annual Conference. Springer, February 2018, pp. 55–72.

12. PUBLICATIONS

CONFERENCE:

International Conference on “Innovations in Computers Networks, Computational Intelligence and IoT” (ICICCI – 21)

Paper ID : ICICCI – 21 – 0114

13. STUDENT PROFILES

T. Naveen is currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. He has completed his Secondary Education from Bhagavathi High School and Higher Secondary Education from SR Junior College. His technical skills include Python, Data Science, Microsoft Excel, Tableau, SQL, Statistics, Data Science using Python, Machine Learning. He attended E-summit, Entrepreneurship carnival, which was hosted by EDC – MLRIT in association with Nucleus Tech and SUMVN in the year 2017. He has completed online courses (Python core, Data Science Math Skill, React for beginners from new boston, HTML and CSS, Managing Project Risks and Fundamentals) from Coursera, Sololearn and Cursa App. He has also done a 6 week Data Science course from reputed online institute Internshaala and completed projects on every module related with Data Science. He has attended a Virtual training on top trending technologies conducted by ccbp 4.0, an industrial revolution. He is a coding enthusiast and executed many programs in various coding platforms such as hackerrank, codechef, codeforces etc. He is passionate towards learning new technologies and aspires to become a Software Developer in the near future. His personal projects include "Unwanted Message Filtering System from OSN User walls" which is 3-month internship based on Machine-learning text classifiers under the guidance of Lasya Infotech Institution.

B.Sidharth is currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. He has completed his Secondary Education from Kakatiya High School and Higher Secondary Education from Sri Chaitanya Junior College. His technical skills include python,C,C++ and Sql. He attended E-summit, Entrepreneurship carnival, which was hosted by EDC – MLRIT in association with Nucleus Tech and SUMVN in the year 2017.He attended 5 days online International Hands-on-Training in python programming,3-day online workshop on "AI & ML in speech and audio processing" in the year 2020. He has attended a Virtual training on mastering Data Science in 11hrs conducted by ExcelR and TASK. He is passionate towards learning new technologies. He has completed online courses (AWS Fundamentals: Going Cloud-Native, AI for everyone, Data science math skill, Leadership and emotional intelligence, Managing project risks and changes) from Coursera and (Python, Sql, Machine learning, Artificial intelligence, Web development) from Cursa. His personal projects include "A System which filter Unwanted messages from OSN Userwalls" which is 3-month internship based on cloud-based security service under the guidance of Lasya Infotech Institution

T. Amith Krishna is currently pursuing his Bachelor of Technology with a specialization in Computer Science and Engineering at St. Martin's Engineering College. He has completed his Secondary Education from Nalanda Vidya Bhavan High School and Higher Secondary Education from Sri Chaitanya Junior College. His technical skills include Python, Microsoft Excel, SQL, Statistics, Data Science using Python, Machine Learning, C language. He attended E-summit, Entrepreneurship carnival, which was hosted by EDC – MLRIT in association with Nucleus Tech and SUMVN in the year 2017. He has completed online courses (AWS Fundamentals: Going Cloud Native, Data Science Math Skill, AI for everyone, Leadership and Emotional Intelligence, Managing Project Risks and Fundamentals, Python) from Coursera App. He has also done a 6 months Python from reputed online institute Udemy and completed projects on every module related. He has attended a Virtual training on mastering Data Science in 11hrs conducted by ExcelR and TASK. He is passionate towards learning new technologies and aspires to become a Developer in the near future. He is fond of data analytics and looking forward to show passion in stock market analysis. His personal projects include "A System which filter Unwanted messages from OSN Userwalls" which is 3-month internship based on cloud-based security service under the guidance of Lasya Infotech Institution.

V.Ramya is currently pursuing her Bachelor of technology in the stream of Computer Science and Engineering at St. Martin's Engineering college. She completed her intermediate from Sri Chaithanya Junior College and 10th class from Bhashyam High School. Her technical skills include C, C++ and Python. She also has basic understanding of Java. She took part in E-Summit program conducted at Marri Laxman Reddy Institute of Technology in 2018. She has completed online courses (AWS Fundamentals: Going Cloud Native, Data Science Math Skill, AI for everyone, Leadership and Emotional Intelligence, Managing Project Risks and Fundamentals, Python) from Coursera App

A

PROJECT REPORT

On

EMOTION BASED MUSIC RECOMMENDATION SYSTEM USING WEARABLE PHYSIOLOGICAL SENSORS

Submitted by

1) Mr. B. Nikhil Reddy	**17K81A05J7**
2) Mr. E. Sriman Goud	**17K81A05K7**
3) Mr. B. Vamshi	**17841A0559**
4) Mr. A. Siddhartha Reddy	**17K81A05J4**

in partial fulfillment for the

award of the degree of

BACHELOR OF TECHNOLOGY

IN

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Under the Guidance of

Mr. V.L. Kartheek

Assistant Professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ST. MARTIN'S ENGINEERING COLLEGE
An Autonomous Institute

Dhulapally, Secunderabad – 500 100

JUNE 2021

BONAFIDE CERTIFICATE

This is to certify that the project entitled Emotion based music recommendation system using wearable physiological sensors, is being submitted by B. Nikhil Reddy Regd. No. 17K81A05J7, A. Siddhartha Reddy Regd. No. 17K81A05J4, E. Sriman Goud Regd. No. 17K81A05K7, B. Vamshi Regd. No. 17841A0559 in partial fulfillment of the requirement for the award of the degree of **BACHELOR OF TECHNOLOGY IN Computer Science and Engineering** is recorded of bonafide work carried out by them. The result embodied in this report have been verified and found satisfactory.

Guide Signature
Mr. V.L. Kartheek
Department of CSE

Head of the Department
Dr. M. NARAYANAN
Department of CSE

Internal Examiner

External Examiner

Place:

Date:

DECLARATION

We, the students of **Bachelor of Technology** in **Department of Computer Science and Engineering**, session: <2017 – 2021>, St. Martin's Engineering College, Dhulapally, Kompally, Secunderabad, hereby declare that work presented in this Project Work entitled **Emotion based music recommendation system using wearable physiological sensors** is the outcome of our own bonafide work and is correct to the best of our knowledge and this work has been undertaken taking care of Engineering Ethics. This result embodied in this project report has not been submitted in any university for award of any degree.

1) Mr. B. Nikhil Reddy	17K81A05J7
2) Mr. E. Sriman Goud	17K81A05K7
3) Mr. B. Vamshi	17841A0559
4) Mr. A. Siddhartha Reddy	17K81A05J4

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible and whose encouragements and guidance have crowded effects with success.

We extended our deep sense of gratitude to Principal**, Dr. P. SANTOSH KUMAR PATRA**, St. Martin's Engineering College, Dhulapally, for permitting us to undertake this project.

We are also thankful to **Dr. M. NARAYANAN**, Head of the Department, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

We would like to thank **Dr. T. POONGOTHAI,** Department of Computer Science and Engineering (AI & ML) for her encouragement and insightful comments and as well as our project coordinators **Dr. B. RAJALINGAM**, Associate Professor and **Mr. J. SUDHAKAR**, Assistant Professor, in Department of Computer Science and Engineering for their valuable support.

We would like to express our sincere gratitude and indebtedness to our project supervisor **Mr. V.L. KARTHEEK,** Assistant Professor, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

Finally, we express thanks to all those who have helped us successfully to completing this project. Furthermore, we would like to thank our family and friends for their moral support and encouragement.

We express thanks to all those who have helped us in successfully completing the project.

1) Mr. B. Nikhil Reddy 17K81A05J7
2) Mr. E. Sriman Goud 17K81A05K7
3) Mr. B. Vamshi 17841A0559
4) Mr. A. Siddhartha Reddy 17K81A05J4

ABSTRACT

Most of the existing music recommendation systems use collaborative or content-based recommendation engines. However, the music choice of a user is not only dependent to the historical preferences or music contents. But also dependent to the mood of that user. This paper proposes an emotion-based music recommendation framework that learns the emotion of a user from the signals obtained via wearable physiological sensors.

In particular, the emotion of a user is classified by a wearable computing device which is integrated with a galvanic skin response (GSR) and photo plethysmography (PPG) physiological sensors. This emotion information is feed to any collaborative or content-based recommendation engine as a supplementary data.

Thus, existing recommendation engine performances can be increased using these data. Therefore, in this paper emotion recognition problem is considered as arousal and valence prediction from multi-channel physiological signals. Experimental results are obtained on 32 subjects' GSR and PPG signal data with/out feature fusion using decision tree, random forest, support vector machine and k-nearest neighbours' algorithms.

The results of comprehensive experiments on real data confirm the accuracy of the proposed emotion classification system that can be integrated to any recommendation engine.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF OUTPUT SCREENS

LIST OF ACRONYMS

<AVI>	Audio Video Interlace
<BMP>	Bitmap
<CPU>	Central Processing Unit
<GB>	Giga Bytes
<GUI>	Graphical User Interface
<GSR>	Galvanic Skin Response

1.INTRODUCTION

Stress is a physiological response to the mental, emotional, or physical challenge and it can be defined as the reaction of a person to the environmental requests or influences (Sun et al., 2010). Stress conditions can cause physical and emotional exhaustion that leads to symptoms such as headaches, stomach complaints and difficulties in sleeping. A study conducted by the American Institute of Stress (Statistic Brain Research Institute, NY) has shown as in 2015 the 48% of people feels that their stress condition has increased over the past five years. 77% of people regularly experiences physical symptoms caused by stress with a negative impact on their personal and professional life (Statistic Brain, 2015).

The influence of stress and its consequences on society concerns also the economic aspect. According to the recent EU-funded project 2013, the cost to Europe of work-related stress and depression was estimated to be €617 billion annually. The total amount includes loss of productivity, health care costs and social welfare costs (EU-OSHA, 2016). The early detection of stress can positively affect personal wellbeing and society affluence.

Traditionally, the level of personal stress has been established using some psycho-metric instruments and scales (Ulstein et al., 2007), which are subjective. Subsequently the correlation between the variation of the physiological signals and stress was investigated in order to make the measurement more objective.

1.1 Project Overview

In this paper author is describing concept to recommend music to user by detecting moods of user. Existing technique were using collaboration technique which will use previous user data (input: set of images pre-defined to train the algorithm) to recommend music to user, if there is no input from previous user then this technique will not useful. This existing technique requires lots of manual work to arrange different music to different categories such as happy, sad or angry etc.

To overcome from above issue author is using ‘Wearable Physiological Sensors ’ and this sensors will send signals to application regarding user current status and then this application using SVM (support vector machine) and deep learning neural network algorithms will classify/predict the mood by extracting features from signal.

1.2 Project Objectives

The main objective of our music recommendation system is to provide suggestions to the users that fit the user's preferences. The analysis of the facial expression/user emotion may lead to understanding the current emotional or mental state of the user. Music and videos are one region where there is a significant chance to prescribe abundant choices to clients in light of their inclinations and also recorded information.

It is well known that humans make use of facial expressions to express more clearly what they want to say and the context in which they meant their words. More than 60 percent of the users believe that at a certain point of time the number of songs present in their songs library is so large that they are unable to figure out the song which they have to play.

By developing a recommendation system, it could assist a user to make a decision regarding which music one should listen to helping the user to reduce his/her stress levels. The user would not have to waste any time in searching or to look up for songs and the best track matching the user's mood is detected, and songs would be shown to the user according to his/her mood. The image of the user is captured with the help of a webcam or Electrodes.

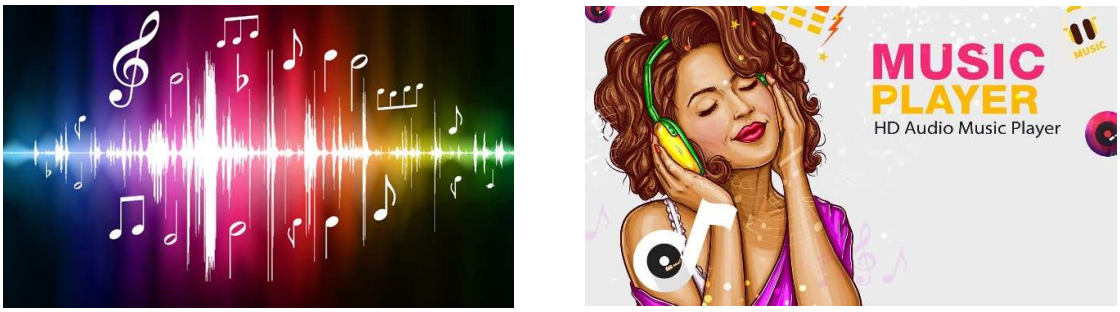


Fig: Music playing based on mood.

2. LITERATURE SURVEY

2.1 Survey on Background

1. A machine learning model for emotion recognition from physiological signals.

Emotions are affective states related to physiological responses. This study proposes a model for recognition [1] of three emotions: amusement, sadness, and neutral from physiological signals with the purpose of developing a reliable methodology for emotion recognition using wearable devices. Target emotions are captured by photoplethysmography [5], which provides information about heart rate, and galvanic skin response. These signals were analyzed in frequency and time domains [2] to obtain a set of features. Several feature selection techniques and classifiers were evaluated. The best model was obtained with random forest recursive feature elimination, for feature selection, and a support vector machine for classification. The results show that it is possible to detect amusement, sadness, and neutral emotions using only galvanic skin response [3] features. 100% accuracy.

2. Anxiety Level Recognition for Virtual Reality Therapy System Using Physiological Signals.

Virtual reality exposure therapy [11] (VRET) can have a significant impact towards assessing and potentially treating various anxiety disorders. One of the main strengths of VRET systems is that they provide an opportunity for a psychologist [4] to interact with virtual 3D [6] environments and change therapy scenarios according to the individual patient's needs. Therefore, in order to fully use all advantages provided by the VRET system, a mental stress detection system is needed. The patient's physiological signals can be collected with wearable biofeedback sensors. Signals like blood volume pressure [12] (BVP), galvanic skin response (GSR), and skin temperature can be processed and used to train the anxiety level classification models. The acquired data were used to train a four-level anxiety recognition model [7] (where each level of 'low', 'mild', 'moderate', and 'high' refer to the levels of anxiety rather than to separate classes of the anxiety disorder). We achieved 86.3% accuracy with the signal fusion-based support vector machine (SVM) [19] classifier.

3. An emotional recommender system for music.

Nowadays, Recommender Systems have become essential to users for finding "what they need" [2] within large collections of items. Meanwhile, recent studies have demonstrated as user

personality can effectively provide a more valuable information to significantly improve recommenders' [8] performance, especially considering behavioural [9] data captured from social network logs. In this work, we describe a novel music recommendation technique based on the identification of personality traits, moods and emotions of a single user, starting from solid psychological observations recognized by the analysis of user behavior within a social environment. In particular, users personality and mood have been embedded within a content-based filtering approach to obtain more accurate and dynamic results. Several experiments are then reported to show effectiveness [7] of user personality and mood recognition recommendation, thus encouraging research in this direction.

4. Improved curriculum learning using SSM for facial expression recognition.

Facial expression recognition [15] is an important research issue in the pattern recognition field. However, the generalization of the model still remains a challenging task. we apply a strategy of curriculum learning to facial expression recognition during the stage of training. And a novel curriculum [5] design method is proposed. The system first employs the unsupervised density–distance clustering method to determine the clustering center of each category. Then, the dataset is divided into three subsets of various complexity according to the distance [7] from each sample to the clustering center in the feature space. To solve the problem that the model has a poor recognition accuracy for anger, fear and sadness, a self-selection mechanism [12] is introduced in the test stage to make further judgment on the result of the main model.

5. Using CNN for facial expression recognition: a study of the effects of kernel size and number of filters on accuracy.

Facial expression recognition is a challenging problem in image classification. This has led to increased efforts in solving the problem of facial expression recognition using convolutional neural networks [14] (CNNs). A simple architecture is fast to train and easy to implement. An effective architecture achieves good accuracy on the test data. CNN architectures [8] are black boxes to us. VGGNet, AlexNet and Inception are well-known CNN architectures. These architectures have strongly influenced CNN model designs for new datasets. This work tries to overcome this limitation by using FER-2013 dataset [16] as starting point to design new CNN models. In this work, the effect of CNN [22] parameters namely kernel size and number of filters on the classification accuracy is investigated using FER-2013 dataset. Our major contribution is a thorough evaluation of different kernel sizes and number of filters to propose two novel CNN architectures [17] which achieve a human-like accuracy of 65%.

6. An Automatic Emotion Recognition System for Annotating *Spotify*'s Songs.

The recognition of emotions for annotating large-size music datasets [12] is still an open challenge [21]. The problem lies in that most of the solutions require the audio of the songs and user/expert intervention during certain phases of the recognition process [10]. It consists of a heterogeneous set of machine learning models that have been developed from *Spotify*'s Web data services [17] and miner tools. In order to improve the accuracy of resulting annotations, each model is specialized in recognizing a class of emotions. These models have been validated by using the *AcousticBrainz* database [7] and have been exported to be integrated into a music emotion recognition system. It has been used to emotionally annotate the *Spotify* music database [5] which is composed by more than 30 million songs.

7. DJ-Running: Wearables and Emotions for Improving Running Performance.

Music can have a positive influence on long-distance runners' motivation and performance. It requires selecting the most suitable music by considering the runner's physiological data [18], the type of training session and the geographical [23] and environmental conditions under which the activity is done. We are interested in studying the runners' emotions [11] during the training sessions and in using these emotions to recommend personalized music that increases their motivation and performance [6]. More specifically, in this paper we present an adapted glove that integrates different sensors for collecting data [3], which help to determine the runner's emotional state, and the changes that it experiences. Preliminary results about the interpretation of these data and emotions are discussed and a prototype [15] of recommendation system based on Spotify is sketched.

3. SOFTWARE AND HARDWARE REQUIREMENTS

The project involved analysing the design of few applications so as to make the application more users friendly. To do so, it was really important to keep the navigations from one screen to the other well-ordered and at the same time reducing the amount of typing the user needs to do. In order to make the application more accessible, the browser version had to be chosen so that it is compatible with most of the Browsers.

REQUIREMENT SPECIFICATION

3.0 Functional Requirements

- Graphical User interface with the User.

3.1 Software Requirements

For developing the application the following are the Software Requirements:

1. Python
2. Django
3. Mysql
4. Wamp server

Operating Systems supported

1. Windows 7
2. Windows XP
3. Windows 8

Technologies and Languages used to Develop

1. Python

Debugger and Emulator

- Any Browser (Particularly Chrome)

3.2 Hardware Requirements

For developing the application of the following are the Hardware Requirements:

- Processor: Pentium IV or higher
- RAM: 256 MB
- Space on Hard Disk: minimum 512MB

4. SOFTWARE DEVELOPMENT ANALYSIS

MATERIALS, MODULES AND METHODS

In this section the sensor devices used for the acquisition of physiological signals, the experimental protocol developed and adopted and the methodology chosen for data analysis are described in detail.

4.1 Instrumentation

The choice of the wearable sensor devices to be included into the test has been per-formed according to two criteria: accuracy of measurements and unobtrusiveness of the sensors. There are several devices on the market that claim the measurement of cardiac and electro-dermal activity in a unobtrusive way. Unfortunately, not all these devices are accurate enough for a reliable assessment of stress conditions. In order to find a reasonable trade-off, we selected two devices: Zephyr BioHarness™3 and Shimmer GSR Sensor (Fig. 1).

Figure 1. Zephyr Bio Harness™ 3on the left and Shimmer GSR Sensor on the right

Zephyr BioHarness™3 (BH3) (Medtronic, 2015) is a Bluetooth chest belt capable of retrieving signals derived from the ECG such Heart Rate and R-R Intervals. The ECG signal is sampled at 250 Hz. Moreover, the BH3 is able to collect other signals such as breathing rate, posture information and skin temperature. For the data analysis Giorgia Acerbi and Erika Rovini and the development of the stress detection algorithm the Inter-Beat-Interval data provided by the device has been used. The GSR Module developed by Shimmer (Shim-mer, 2016) is a wearable sensor composed by two special finger electrodes and a main unit that streams data related to the galvanic skin response with a sample frequency of 51.2 Hz using a Bluetooth connection.

4.2 Participants

Twelve voluntary students (3 men, 9 women) with a mean age of 26.0 years old (SD= 4.8 years, range = 21-30 years old) participated on purpose in this study. All the partic-ipants did not meet the exclusion criteria that consisted in neurological disorders that made unable the subjects to complete the mental tasks proposed or cardiac diseases that could deface the physiological response in electro cardiac activity.

Participants completed the experimental session in the Scuola Superiore Sant'Anna (Pisa, Italy) and in the Telecom Italia WHITE Joint Open Lab (Pisa, Italy). Written in-formed consent was obtained from all the participants before starting the tests.

4.3 Experimental Protocol

The experimental protocol was intended to put the subjects in a state of emotional and cognitive stress, in order to measure the variations of their physiological parameters induced by stress.

The experimentation consisted in three phases: a baseline, a stress induction and a recovery stage. During baseline the subjects relaxed in a separate room, for 10 minutes, without using mobile phone, without music or external sounds, without stimuli and without closing their eyes. This phase was indispensable in order to acquire the personal baseline of each subject, since physiological parameters show a wide intersubjects variability. At the end of baseline recording, the psychologist administered psychometric instruments to the participants to obtain a subjective perception about the level of stress, anxiety and drowsiness. Then the subjects performed the stress phase, during about 15-20 minutes, completing a series of extremely demanding cognitive tests handed out by the psychologist in order to induce the stress. People were not aware that this phase was part of the experiment: the psychologist indeed pretend-ed to be sent by university to detect the intelligence quotient (IQ) for a poll. The investigator assumed a very aggressive behaviour towards the subject, behaving rude and correcting the person even when the he accomplished the task properly. Further-more, the user performed the required tasks by listening a noisy sound in background.

A wearable system for stress detection through physiological data analysis that simulates high intensity traffic jam. At the end of this phase, the subjects filled out the psychometric instruments again. Afterwards a recovery period of 10 minutes was performed, in the same conditions as in the baseline phase.

During the whole experimental session (baseline, stress and recovery phases), the tested subjects wore the kit of wearable sensors described in par.2.1, in order to record electro cardiac and electrodermal activities.

4.3.1 Tests for stress induction

The aim of this experimental protocol was to arouse stress in tested subjects that would produce major changes in the level of physiological signals. For this reason, in the experimental protocol the stress induction phase consisted of two paths:

(i) The use of validated neuropsychological tests that caused a great cognitive effort;

(ii) The creation of a stressful social situation that would put the subject under pressure causing a strong emotional reaction.

The five different following tasks (Fig. 2), were executed by the tested subjects:

- **Digit Span:** it is a common measure of short-term memory to evaluate working memory's number storage capacity. In the test of Reverse Digit Span a list of random numbers was read out loud to the person who had to immediately repeat it in a backward order.
- **Stroop Colour Test:** it is a common test to measure selective and divided attention, cognitive flexibility and processing speed (Lansbergen et al., 2007). This test is a demonstration of interference in the reaction time of a task in which the subject was asked to read out loud and as fast as possible either the written word or the ink col-or.
- **Corsi Reverse**: The Corsi block-tapping test is a psychological test that assesses visual-spatial short-term memory. The experiment is done typically by using a wooden base where nine identical spatially separated blocks are present. In the Reverse Corsi Test (Gillet, 2007) the experimenter indicated a sequence of blocks by tapping them and the subject was requested to reproduce the spatial succession of boxes in the reverse way.
- **Kohs Block Design Test:** this is a performance test designed to be an IQ test and to measure visual-spatial skills (Barbeau, 1980). The subject was asked to replicate the patterns displayed on a series of test cards by using coloured cubes (each side has a single colour or two colours divided by a diagonal line).
- **Tower of Hanoi**: this is a mathematical game, common to test problem solving and executive capacity of the subject (Miyake et al., 2000). It is composed by three rods and a number of disks of different sizes which can slide onto any rod. The subject had to move the entire stack to another rod, following simple rules: only one disk

A wearable system for stress detection through physiological data analysis that simulates high intensity traffic jam. At the end of this phase, the subjects filled out the psychometric instruments again. Afterwards a recovery period of 10 minutes was performed, in the same conditions as in the baseline phase.

During the whole experimental session (baseline, stress and recovery phases), the tested subjects wore the kit of wearable sensors described in par.2.1, in order to record electro cardiac and electrodermal activities.

Giorgia Acerbi and Erika Rovini could be moved at a time; only the upper disk from one of the stacks could be moved and placed on top of another stack; no disk could be placed on top of a smaller disk

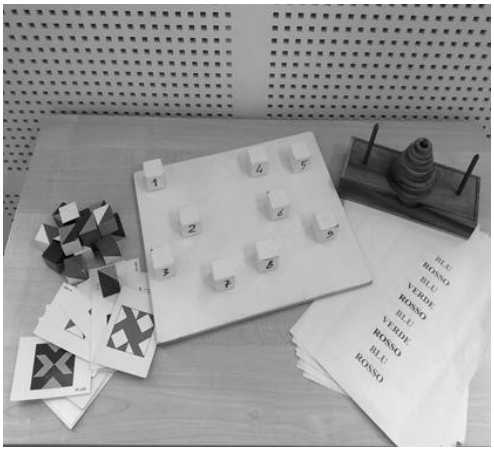


Figure 2. Stress induction test set administered during the experimental session

4.3.2 Psychometric Instruments

In order to measure the emotional state and the level of stress of the subjects, the following psychometric instruments were administered before and after stress induction phase:

- **State-Trait Anxiety Inventory (STAI)**: this scale is one of the most frequently, re-liable and sensitive used measures of anxiety in applied psychology research. In this study the short-form of the STAI scale was used, consisting of only six items (STAI-6) since the objective was to establish the level of stress and anxiety produced during the stress phase (Marteau et al., 1992). Higher STAI scores suggest higher levels of anxiety.
- **Karolinska Sleepiness Scale (KSS):** it is one of the most common sleepiness state tests and it is a 9-point Likert scale based on a self-reported assessment of the per-son's level of drowsiness at the moment (Åkerstedt & Gillberg, 1990). The subject had to choose his level of sleepiness from 1="very alert" to 9="very sleepy". KSS was originally developed to constitute a one-dimensional scale of sleepiness and was validated against alpha and theta electroencephalographic activity (Kaida et al., 2006).

- **Shortened State Stress Questionnaire (SSSQ):** The 24-item SSSQ (Helton, 2004), based on the 90 Question Dundee Stress State Questionnaire (DSSQ), provides a rapid, reliable, self-report assessment of the three primary stress dimensions: distress, task engagement and worry (Pfaff et al., 2012).

4.3.3 Data Analysis

The physiological data acquired during the whole experimentation have been offline analysed using Matlab® R2012a.

The acquired data have been examined for baseline phase (10 minutes of record-ing), stress phase (ranging from 15 to 20 minutes, depending from the attitude and be-haviour of the tested subject). Thus, for each phase, we obtained a dataset composed by a set of GSR features for each participant and another dataset consisting of features extracted from HRV signal. All these data were analysed in order to investigate varia-tions in physiological parameters that could be attributed to stress statutes of the tested subjects.

MODULES

4.4 Galvanic Skin Response (GSR)

The EDA has been recorded using Shimmer GSR sensor which provides as output the galvanic resistance, that has been converted into galvanic skin conductance. In the feature extraction algorithm, the signal has been analysed with temporal windows of 2 minutes, after a filtering process, using a moving average filter. The features extraction algorithm is based on startle detection that can lead to a set of computable features. The method used is referred to the scoring multiples response method of Boucsein (Boucsein, 2012), that establishes a local baseline at the level of the onset of the second response and measures the distance from that baseline to the following peak. The detection algorithm identifies all the occurrences of when the first derivative exceeded to a certain threshold.

It was empirically determined as 0.005 μS. Given the variability of GSR signal among subjects, this threshold is not absolute but it has found to be adequate for the 12 subjects analyzed. Furthermore, to ensure to not con-sider subsequent startles, a minimum distance has been chosen as in (Shumm et al., 2008, considering that a startle event is expected to last about 1-3 s. Once the response was detected, the zero-crossing of the derivative preceding and following the response were identified as the onset and end of the startle (Haley & Picard, 2000). Starting from the startle detection, the following parameters have been calculated (Table 1):

Table 1. Features extracted from GSR signal calculated within a temporal window.

Feature Name	Description
Num_Startle	Number of the stressors
Sum_Amplitude	Sum of the amplitude of the stressors
Sum_RiseTime	Sum of the rise duration of the stressors
Sum_RecTime	Sum of the decrease duration of the stressors
Rise_Rate	Mean value of the rise duration of the stressors
Decay_Rate	Mean value of the decrease duration of the stressors
Area_GSR	Mean of the area under each stressor
Mean_GSR	Mean value of GSR signal
Std_GSR	Standard deviation of GSR signal

.

4.4.1 Electro cardiac activity

Electro cardiac activity has been recorded using the chest belt Zephyr Bio Harness™ BH3. The device provides as output the raw ECG signal and the HRV data that specifies the temporal distance between a beat and the following one. Starting from Inter-Beat-Interval (IBI), the algorithm to extract the main features has been developed. The IBI signal has been modified identifying and correcting ectopic rhythm, which is an irregular heart rhythm due to a premature heartbeat. The analysis of cardiac signal has been structured investigating both the time domain and the frequency domain. Regarding the time domain, the following parameters have been selected and computed (see Table 2)

Table 2. Features extracted from HRV signal in the temporal domain.

Feature Name	Description
IBI_mean	Mean of Inter-Beat-Interval corresponding to R-to-R interval
SDNN	Standard deviation of all Normal RR intervals (NN intervals)

HR_mean	Mean of Heart Rate
SDHR	Standard deviation of the Heart Rate
RMSSD	Square root of the mean of the squared differences between adjacent normal RR intervals
pNN50	Percentage of differences between adjacent normal RR intervals exceeding 50 ms
#ECT	Number of ectopic intervals (abnormal RR intervals)
%ECT	Percentage of ectopic intervals on the total number of RR intervals

By identifying and correcting ectopic rhythm, a Normal-to-Normal (NN) interval sequence appropriate for HRV analysis is obtained. Since the NN interval sequence is an irregularly sampled time sequence, for spectral analysis it had to be therefore con-verted to an equidistantly sampled sequence (Mali et al., 2014). After a smoothing of the signal, the NN interval sequence has been resampled at 4Hz. For the analysis in frequency domain, the following parameters have been computed (see Table 3):

Table 3. Features extracted from HRV signal in the frequency domain.

Feature Name	**Description**
Peak VLF	Frequency peak in very low frequency (VLF) range (0.04–0.15 Hz)
Area VLF	Signal power by Power Spectral Density (PSD) in VLF
%VLF	Percentage of signal power in the VLF respect to the total signal power
Peak LF	Frequency peak in low frequency (LF) range (0.04–0.15 Hz)
Area LF	Signal power by PSD in LF
%LF	Percentage of signal power in the LF respect to the total signal power
Peak HF	Frequency peak in high frequency (HF) range (0.15–0.4 Hz)

Area HF	Signal power by PSD in HF
%HF	Percentage of signal power in the HF respect to the total signal power
LF/HF	Ratio between LF and HF powers

4.4.2 Data processing and Statistical Analysis

After extracting features from physiological signals, Kolmogorov-Smirnov test was applied in order to verify the normal distribution of data. A non-parametric statistical analysis was used because the test showed data were not normally distributed. Then, Kruskal-Wallis (KW) test was used for comparing data acquired in baseline phase and those recorded during stress phase in order to verify a significant difference (p-value<0.05) on the basis of the extracted parameters. Furthermore, the linear correlation between the significant parameters was calculated using the Pearson's coefficient. If the value of correlation between two features was at least rho=0.8, the less significant one was deleted. Then, the remaining features were used for Principal Component Analysis (PCA) in order to identify how the groups investigated, related to different phases of the experimental protocol, could be visualized and separated in the space of the principal components (PCs). Finally, the most important PCs, that included more than 80% of the overall variance of data, were taken into account in order to train and test a Support Vector Machine (SVM) classifier which had to be able to correctly classify a subject as stressed or not-stressed.

Regarding the analysis of the psychometric instruments, a T-test has been conduct-ed in order to assess if significant differences between after and before the stress induction phase could be revealed.

Finally, a linear regression analysis has been implemented with the aim to look for a correlation between the results obtained by the psychometric instruments administered and the physiological parameters measured.

Results and Discussion

In this section the results obtained from both the analysis of physiological data and the psychometric instruments are reported and widely discussed, examining the most important features extracted, the evaluation of the psychometric instruments and the algorithm for data classification.

4.4.5 Physiological Parameters Assessment

Features extracted by physiological parameters are reported in Tables 4-5 both for baseline phase and stress phase as mean values and standard deviations. Furthermore p-values, calculated with KW test for non parametric data, are also disclosed because they represent if there are significant differences between the two investigated groups.

Table 4. Features extracted from GSR signal: mean values ± standard deviations and significance.

Parameters	Baseline		S		p-va
Num_Startle (#)	1	4.7	17	3	0.
Sum_amplit (μS)	1	12.4	11	6	0.4
Sum_RiseTi (s)	3	6.94	41	5	0.(
Sum_RecTi (s)	6	8.7	64	2	0.4
Rise_Rate (μS/s)	3	0.9(	3	0	0.
Decay_Rate (μS/s)	9	4.3	5	1	0.(
Area_GSR (s·μS)	2	1.9	2	1	0.1
Mean_GSR (μS)	1	5.34	16	5	0.(
Std_GSR (μ	1	1.3	1	0	0.3

* Significant difference between groups ($p<0.05$)

Significant differences are observed in some parameters, both for features extracted by electro dermal and electro cardiac activities, representing a concrete variation in physiological response to a psychological stress induction.

In particular for the first signal, Sum_RiseTime, Decay_Rate and Mean_GSR are the significant parameters. For the second signal IBI_mean, HR_mean, SDHR, RMSSD and pNN50 are the significant features in the temporal domain, whereas Peak VLF, %VLF and Peak LF are the ones in the frequency domain.

Discussing significant parameters derived by electrodermal activity, Sum_RiseTime is a parameter that gives an indication of how the global GSR level is varying as time progresses. If the sympathetic branch of the ANS is highly aroused, then sweat gland activity also increases. This fact leads to an increase of skin conductance, that can be then a measure of emotional and sympathetic responses. A significant variation of this parameter from baseline to stress phase can be explained as an increase of arousal level of the subject, probably due to an increment of stress level during the execution of the stressor tasks. A significant variation has been observed from baseline phase to stress phase for other two parameters: Decay_Rate and Mean_GSR. Regarding the mean value of GSR, it reflects the variation of the signals in terms of arousal, cognitive load and stress in general. So, an increase of cognitive load corresponds to an increase of the mean value of the signal, related to a bigger sweat gland activity that modifies SC. Finally, a considerable variation in decay rate, which represents an indirect measure of the relaxation pattern experienced by the subject (Singh et al., 2012) could mean that when the arousal level is high, the GSR needs more time to assume values similar to baseline ones. So, it is reasonable to have a variation of the time needed to obtain a relaxation, during a stress phase, respect to the baseline.

Regarding electro cardiac activity variations in the mean values of IBI and HR from baseline to stress phase are absolutely congruent with an increase of stress level: the number of beats in a minute increases, with a related reduction of the time between a heartbeat and the following one. According to Orsila et al. (2008) in which RMSSD parameter changed its values among different phases of the experimental session de-scribed, this parameter presents a variation from baseline to stress phase. The lower value in the stress stage may suggests the subjects' perceived stress was effectively higher during this phase of the protocol. The difference between baseline and stress conditions in pNN50 was expected, as in (Taelman et al, 2009). It is probably due to the short-term variability, which is lower with a cognitive task than during rest. Also, SDHR changes between the phases, being a measure for long term variability. Analysing frequency domain parameters, it is known that sympathetic and parasympathetic activities are reflected into LF and HF power, so a variation in one of the parameters linked to these frequency contributions is justified. The activation of SNS is indeed reflected in the variation of peak LF, peak VLF and %VLF.

4.4.6 Psychometric Instruments Evaluation

A comparison between the scores of the state tests administered before and after stress induction has been performed (Table 6) using the T-test. The KSS scores did not show significant differences between before and after stress induction phase, whereas the STAI-6 scores showed statistically significant differences between the two phases ($p<0.05$) indicating a recognisable level of anxiety in the tested subjects. The SSSQ scores also showed significance differences between pre and post stress induction tests ($p<0.01$) In particular, a highly statistically significant result ($p<0.01$) emerged from a subscale of SSSQ called "distress" that is the most important factor of SSSQ measuring the negative effect of the situation (Helton, 2004). A statistically significant result related to the variation of this subscale could mean that the stress induction phase effectively provided a negative effect on participants.

Table 6. Questionnaires results: mean values ± standard deviations and t-test significance.

Scale	Baseline	Stress	p-value
KSS	3.8 ± 1.3	3.3 ± 0.6	0.089
STAI-6	10.9 ± 2.2	13.7 ± 4.1	0.031 *
SSSQ	93.4 ± 19.5	117.6 ± 34.2	0.001 *
Distress	19.9 ± 10.9	39.5 ± 20.1	0.001 *
Task Management	36.5 ± 4.1	35.8 ± 6.1	0.639
Worry	37.0 ± 12.9	42.3 ± 19.5	0.174

Data Classification

According to the aim of the paper, a classifier was implemented in order to identify the status of the subjects on the basis of the measured physiological signals. Basically, the classifier should be able to distinguish if a person is stressed or not.

For this purpose, the datasets acquired both in baseline and stress phases were used and, in particular, the parameters resulted significant at the KW test in distinguishing between the two phases have been taken into account (see par. 3.1).

The linear correlation between the significant parameters was calculated using the Pearson's coefficient.

If the value of correlation between two features was at least 0.80, the less signifi-cant one was deleted.

Thus, a reduced number of eight parameters has been selected and used for Princi-pal Component Analysis (PCA) that allowed to visualize the separation between sub-jects in baseline and stress phases in the space of the PCs as shown in fig. 3.

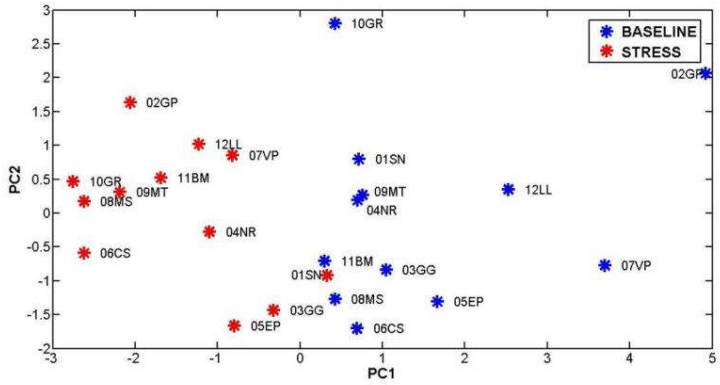


Figure 3. PCA synthesizes the differences in physiological parameters between baseline (blue markers) and stress phase (red markers). The first four PCs contains the 87.8% of the overall variance.

5. PROJECT SYSTEM DESIGN

5.1 UML DIAGRAMS

UML stands for Unified Modeling Language. UML is a standardized general-purpose modeling language in the field of object-oriented software engineering. The standard is managed, and was created by, the Object Management Group.

The goal is for UML to become a common language for creating models of object-oriented computer software. In its current form UML is comprised of two major components: a Meta-model and a notation. In the future, some form of method or process may also be added to; or associated with, UML.

The Unified Modeling Language is a standard language for specifying, Visualization, Constructing and documenting the artifacts of software system, as well as for business modeling and other non-software systems.

The UML represents a collection of best engineering practices that have proven successful in the modeling of large and complex systems.

The UML is a very important part of developing objects-oriented software and the software development process. The UML uses mostly graphical notations to express the design of software projects.

GOALS:

The Primary goals in the design of the UML are as follows:

1. Provide users a ready-to-use, expressive visual modeling Language so that they can develop and exchange meaningful models.
2. Provide extendibility and specialization mechanisms to extend the core concepts.
3. Be independent of particular programming languages and development process.
4. Provide a formal basis for understanding the modeling language.

5. Encourage the growth of OO tools market.
6. Support higher level development concepts such as collaborations, frameworks, patterns and components.
7. Integrate best practices.

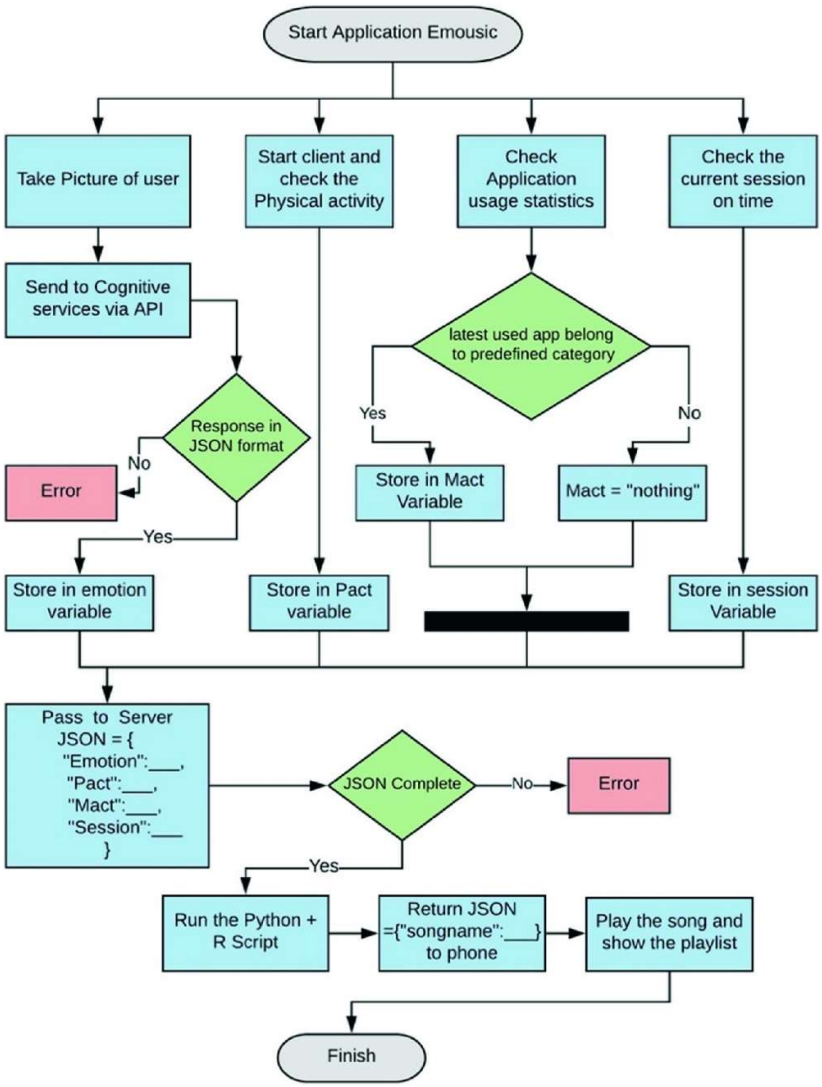


Fig: The flowchart showing the process flow of the Algorithm and the Playing the Music

USE CASE DIAGRAM:

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.

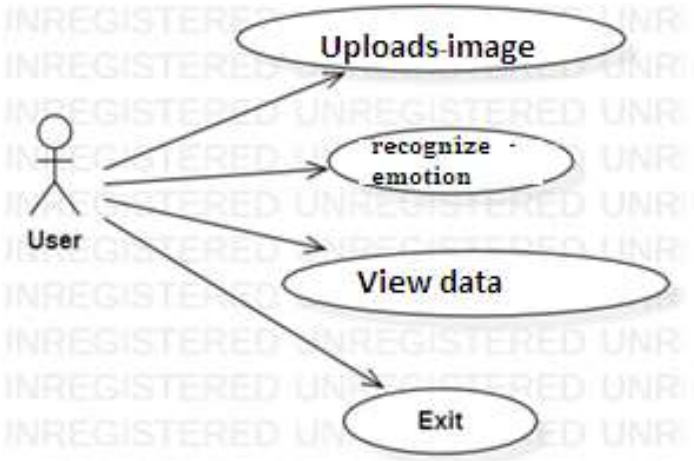


Fig: Use-case Diagram

CLASS DIAGRAM:

In software engineering, a class diagram in the Unified Modelling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.

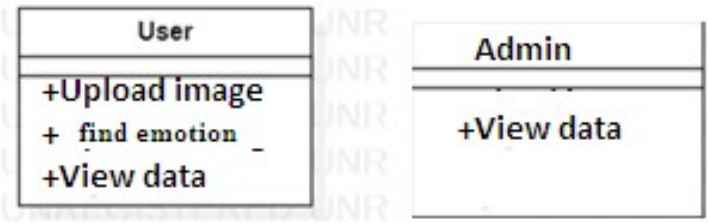


Fig: Class Diagram.

SEQUENCE DIAGRAM:

A sequence diagram in Unified Modelling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.

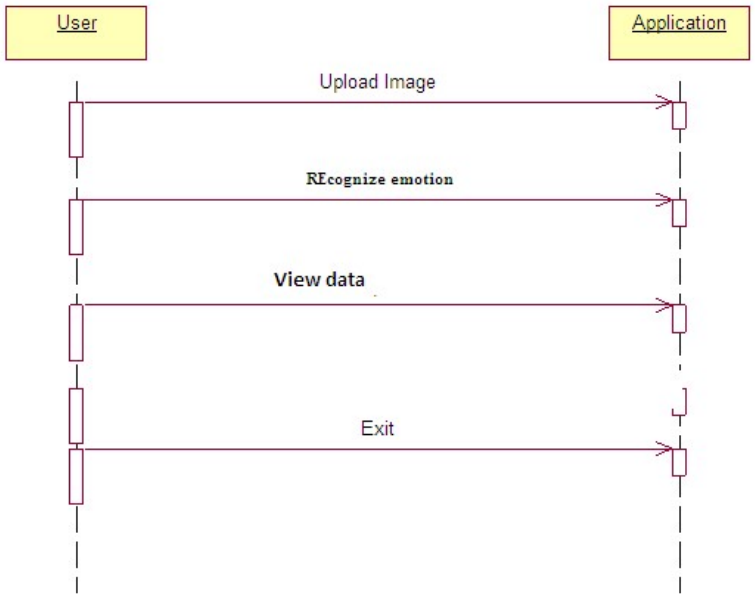


Fig: Sequence Diagram.

ACTIVITY DIAGRAM:

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modeling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.

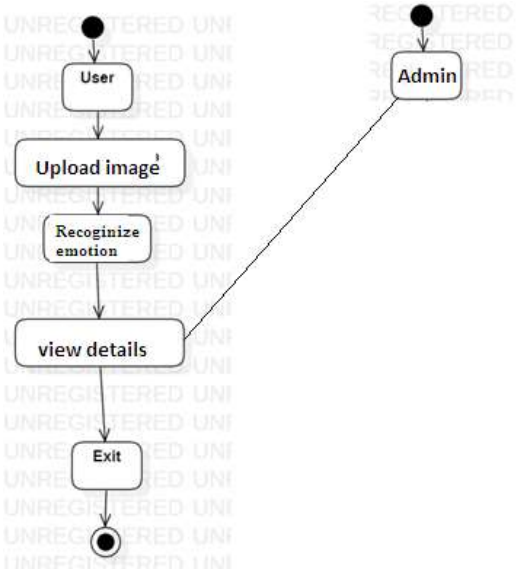


Fig: Activity Diagram.

6. PROJECT CODING

6.1 Code Templates

EmotionReader.py

Import modules:

1. Pandas
2. Numpy

#Declare global variables

Def authentication handler ():

  # Insert OAuth token

Def PushEmotionCSV ():

  #Reads Expressions and writes them to Emotion.csv

MusicReccomendation.py

#import modules

Import:

1. Tkinter
2. Numpy
3. Threading
4. Pandas
5. Keras

#Declare global variables

Def upload ():

  # Import's csv files

Def readDataset ():

  #Reads Dataset from csv files

Def getEmotions():

  # Performs Expression analysis on Emotions data

```
Def collaborativeFilter ():
  # Performs collabirative filtering on Genre
Def contentFilter ():
  # Performs Content Based Filtering on Genre
Def Recommendation ():
#Builds Recomended Music to the user
```

6.2. OUTLINE FOR VARIOUS FILES

We used Python programming to implement our project. A single python file is used to implement our code. This file consists of various modules that we have used. Our project modules are – User Login, Admin Login, Add Emotions, Recommend Music. We also used various python modules like tkinter, keras, Sleep, numpy, Threading.

6.3. METHODS INPUT AND OUTPUT PARAMETERS

In our project code, we implemented nine methods. They are:

1. Authenticaion Handler()
2. pushSongsCSV()
3. upload()
4. readDataset()
5. getEmotion()
6. collaborativeFiltering()
7. contentFiltering()
8. recommendation()
9. graph()
10. Cascadeclassifier()
11. Videocapture()

Our EmotionReader.py file script contains two methods Authintication Handler and PushEmotionCSV this methods perform fetching of music files from database and writing them to .csv file.

Our Music Reccomendation.py file script contains all the other methods which selects a Music for Recommendation and Graph to showcase the trend of upcoming Genres in Music.

File Edit Format Run Options Window Help

```
from keras.models import load_model
from time import sleep
from keras.preprocessing.image import img_to_array
from keras.preprocessing import image
import cv2
from playsound import playsound
import numpy as np
import threading as T

face_classifier = cv2.CascadeClassifier('./haarcascade_frontalface_default.xml')
classifier =load_model('./Emotion_Detection.h5')

class_labels = ['Angry','Happy','Neutral','Sad','Surprise']

cap = cv2.VideoCapture(0)
count=0
```

Fig: Introducing all the Libraries which are useful.

```
def player(emotion):
    emotion=emotion.lower()
    cap2 = cv2.VideoCapture(0)
    font                   = cv2.FONT_HERSHEY_SIMPLEX
    bottomLeftCornerOfText = (10,380)
    bottomLeftCornerOfText2 = (10,430)
    fontScale              = 1
    fontColor              = (255,255,255)
    lineType               = 2
    global s
    def songplayer():
        global s
        if emotion == 'neutral':
            s = 'Nee Chepakallu.mp3'
            playsound('SONGS/neutral/Nee Chepakallu.mp3')

        if emotion == 'scared':
            s = 'Chandramukhi.mp3'
            playsound('SONGS/scared/Chandramukhi.mp3')

        if emotion == 'surprised':
            s = 'Seheri.mp3'
            playsound('SONGS/surprised/Seheri.mp3')

        if emotion == 'angry':
            s = 'Harima Harima.mp3'
            playsound('SONGS/angry/Harima Harima.mp3')

        if emotion == 'fear':
            s = 'Tanha Tanha.mp3'
            playsound('SONGS/fear/Tanha Tanha.mp3')

        if emotion == 'happy':
            s = 'Ay Pilla.mp3'
            playsound('SONGS/happy/Ay Pilla.mp3')

        if emotion == 'sad':
            s = 'Yetu Pone.mp3'
            playsound('SONGS/sad/Yetu Pone.mp3')
```

Fig: Defining the players Emotions for Which Genre of the music should be played.

```
x = T.Thread(target=songplayer)
x.start()
while True:
    ret, frame = cap2.read()
    labels = []
    gray = cv2.cvtColor(frame,cv2.COLOR_BGR2GRAY)
    faces = face_classifier.detectMultiScale(gray,1.3,5)

    for (x,y,w,h) in faces:
        cv2.rectangle(frame,(x,y),(x+w,y+h),(255,0,0),2)
        roi_gray = gray[y:y+h,x:x+w]
        roi_gray = cv2.resize(roi_gray,(48,48),interpolation=cv2.INTER_AREA)
    t = "Detected Emotion = "+emotion
    t2 = 'Playing = '+s
    cv2.putText(frame,t,
        bottomLeftCornerOfText,
        font,
        fontScale,
        fontColor,
        lineType)
    cv2.putText(frame,t2,
        bottomLeftCornerOfText2,
        font,
        fontScale,
        fontColor,
        lineType)
    cv2.imshow('Emotion Detector',frame)
    if cv2.waitKey(1) & 0xFF == ord('q'):
        break
cap2.release()
```

Fig: Defining loops for Displaying the emotions played in the webcam by the user.

```
while True:
    # Grab a single frame of video
    ret, frame = cap.read()
    labels = []
    gray = cv2.cvtColor(frame,cv2.COLOR_BGR2GRAY)
    faces = face_classifier.detectMultiScale(gray,1.3,5)

    for (x,y,w,h) in faces:
        cv2.rectangle(frame,(x,y),(x+w,y+h),(255,0,0),2)
        roi_gray = gray[y:y+h,x:x+w]
        roi_gray = cv2.resize(roi_gray,(48,48),interpolation=cv2.INTER_AREA)


        if np.sum([roi_gray])!=0:
            roi = roi_gray.astype('float')/255.0
            roi = img_to_array(roi)
            roi = np.expand_dims(roi,axis=0)

        # make a prediction on the ROI, then lookup the class

            preds = classifier.predict(roi)[0]
            print("\nprediction = ",preds)
            label=class_labels[preds.argmax()]
            print("\nprediction max = ",preds.argmax())
            print("\nlabel = ",label)
            label_position = (x,y)
            cv2.putText(frame,label,label_position,cv2.FONT_HERSHEY_SIMPLEX,2,(0,255,0),3)
            count+=1
            if count==10:
                player(label)
                break

        else:
            cv2.putText(frame,'No Face Found',(20,60),cv2.FONT_HERSHEY_SIMPLEX,2,(0,255,0),3
        print("\n\n")
    cv2.imshow('Emotion Detector',frame)
    if cv2.waitKey(1) & 0xFF == ord('q'):
        break

cap.release()
```

Fig: Defining the loops inside the conditions for capturing the frames and emotion of the user and predicting the type of music to be played.

6.2 CLASS WITH FUNCTIONALITY

Requirements Specification

Use this Requirements Specification template to document the requirements for your product or service, including priority and approval. Tailor the specification to suit your project, organizing the applicable sections in a way that works best, and use the checklist to record the decisions about what is applicable and what isn't.

The format of the requirements depends on what works best for your project.

This document contains instructions and examples which are for the benefit of the person writing the document and should be removed before the document is finalized.

To regenerate the TOC, select all (CTL-A) and press F9.

Executive Summary

Project Overview

Describe this project or product and its intended audience, or provide a link or reference to the project charter.

Purpose and Scope of this Specification

Describe the purpose of this specification and its intended audience. Include a description of what is within the scope what is outside of the scope of these specifications. For example:

In scope

This document addresses requirements related to phase 2 of Project A:

modification of Classification Processing to meet legislative mandate ABC.

modification of Labor Relations Processing to meet legislative mandate ABC.

Out of Scope

The following items in phase 3 of Project A are out of scope:

modification of Classification Processing to meet legislative mandate XYZ.

modification of Labor Relations Processing to meet legislative mandate XYZ.

(Phase 3 will be considered in the development of the requirements for Phase 2, but the Phase 3 requirements will be documented separately.)

Product/Service Description

In this section, describe the general factors that affect the product and its requirements. This section should contain background information, not state specific requirements (provide the reasons why certain specific requirements are later specified).

Product Context

How does this product relate to other products? Is it independent and self-contained? Does it interface with a variety of related systems? Describe these relationships or use a diagram to show the major components of the larger system, interconnections, and external interfaces.

User Characteristics

Create general customer profiles for each type of user who will be using the product. Profiles should include:

Student/faculty/staff/other

experience

technical expertise

other general characteristics that may influence the product

Assumptions

List any assumptions that affect the requirements, for example, equipment availability, user expertise, etc. For example, a specific operating system is assumed to be available; if the operating system is not available, the Requirements Specification would then have to change accordingly.

Constraints

Describe any items that will constrain the design options, including

parallel operation with an old system

audit functions (audit trail, log files, etc.)

access, management and security

criticality of the application

system resource constraints (e.g., limits on disk space or other hardware limitations)

other design constraints (e.g., design or other standards, such as programming language or framework)

Dependencies

List dependencies that affect the requirements. Examples:

This new product will require a daily download of data from X,

Module X needs to be completed before this module can be built.

Requirements

Describe all system requirements in enough detail for designers to design a system satisfying the requirements and testers to verify that the system satisfies requirements.

Organize these requirements in a way that works best for your project. See **Error! Reference source not found.Error! Reference source not found., Error! Reference source not found.** for different ways to organize these requirements.

Describe every input into the system, every output from the system, and every function performed by the system in response to an input or in support of an output. (Specify what functions are to be performed on what data to produce what results at what location for whom.)

Each requirement should be numbered (or uniquely identifiable) and prioritized.

See the sample requirements in Functional Requirements, and **Error! Reference source not found.**, as well as these example priority definitions:

Priority Definitions

The following definitions are intended as a guideline to prioritize requirements.

Priority 1 – The requirement is a "must have" as outlined by policy/law

Priority 2 – The requirement is needed for improved processing, and the fulfillment of the requirement will create immediate benefits

Priority 3 – The requirement is a "nice to have" which may include new functionality

It may be helpful to phrase the requirement in terms of its priority, e.g., "The value of the employee status sent to DIS **must be** either A or I" or "It **would be nice** if the application warned the user that the expiration date was 3 business days away". Another approach would be to group requirements by priority category.

A good requirement is:

Correct

Unambiguous (all statements have exactly one interpretation)

Complete (where TBDs are absolutely necessary, document why the information is unknown, who is responsible for resolution, and the deadline)

Consistent

Ranked for importance and/or stability

Verifiable (avoid soft descriptions like "works well", "is user friendly"; use concrete terms and specify measurable quantities)

Modifiable (evolve the Requirements Specification only via a formal change process, preserving a complete audit trail of changes)

Does not specify any particular design

Traceable (cross-reference with source documents and spawned documents).

Functional Requirements

In the example below, the requirement numbering has a scheme - BR_LR_0## (BR for Business Requirement, LR for Labor Relations). For small projects simply BR-## would suffice. Keep in mind that if no prefix is used, the traceability matrix may be difficult to create (e.g., no differentiation between '02' as a business requirement vs. a test case)

Non-functional requirement

In systems engineering and requirements engineering, a non-functional requirement is a requirement that specifies criteria that can be used to judge the operation of a system, rather than specific behaviors. They are contrasted with functional requirements that define specific behavior or functions. **Non-functional requirements** add tremendous value to business analysis. It is commonly misunderstood by a lot of people. It is important for business stakeholders, and Clients to clearly explain the requirements and their expectations in measurable terms. If the non-functional requirements are not measurable then they should be revised or rewritten to gain better clarity. For example, User stories help in mitigating the gap between developers and the user community in Agile Methodology.

Usability:

Prioritize the important functions of the system based on usage patterns. **Frequently used functions should be tested for usability**, as should complex and critical functions. Be sure to create a requirement for this.

Reliability:

Reliability defines the trust in the system that is developed after using it for a period of time. It defines the likeability of the software to work without failure for a given time period.

The number of bugs in the code, hardware failures, and problems can reduce the reliability of the software.

Your goal should be a long MTBF (mean time between failures). It is defined as the average period of time the system runs before failing.

Create a requirement that data created in the system will be retained for a number of years without the data being changed by the system.

It's a good idea to also include requirements that make it easier to monitor system performance.

Performance:

What should system response times be, as measured from any point, under what circumstances? Are there specific peak times when the load on the system will be unusually high?

Think of stress periods, for example, at the end of the month or in conjunction with payroll disbursement.

Supportability:

The system needs to be **cost-effective to maintain**.

Maintainability requirements may cover diverse levels of documentation, such as system documentation, as well as test documentation, e.g. which test cases and test plans will accompany the system.

6.3 INPUT AND OUTPUT DESIGN

Input design

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the following things:

- What data should be given as input?
- How the data should be arranged or coded?
- The dialog to guide the operating personnel in providing input.
- Methods for preparing input validations and steps to follow when error occur.

OBJECTIVES

1.Input Design is the process of converting a user-oriented description of the input into a computer-based system. This design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerized system.

2. It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The goal of designing input is to make data entry easier and to be free from errors. The data entry screen is designed in such a way that all the data manipulates can be performed. It also provides record viewing facilities.

3.When the data is entered it will check for its validity. Data can be entered with the help of screens. Appropriate messages are provided as when needed so that the user will not be in maize of instant. Thus the objective of input design is to create an input layout that is easy to follow

Output design

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making.

1. Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is

designed so that people will find the system can use easily and effectively. When analysis design computer output, they should Identify the specific output that is needed to meet the requirements.

2.Select methods for presenting information.

3.Create document, report, or other formats that contain information produced by the system.

The output form of an information system should accomplish one or more of the following objectives.

- Convey information about past activities, current status or projections of the
- Future.
- Signal important events, opportunities, problems, or warnings.
- Trigger an action.
- Confirm an action.

STUDIES AND ANALYSIS

1 FEASIBILITY STUDY:

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are

- ECONOMICAL FEASIBILITY
- TECHNICAL FEASIBILITY
- SOCIAL FEASIBILITY

ECONOMICAL FEASIBILITY:

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

TECHNICAL FEASIBILITY:

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

5.1.3 SOCIAL FEASIBILITY:

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

7.PROJECT TESTING

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

7.1 Various Test Cases

Unit testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

Integration testing

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

Functional test

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items:

Valid Input : identified classes of valid input must be accepted.

Invalid Input : identified classes of invalid input must be rejected.

Functions : identified functions must be exercised.

Output : identified classes of application outputs must be exercised.

Systems/Procedures : interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

System Test

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

7.2 White Box Testing

White Box Testing is a testing in which in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is purpose. It is used to test areas that cannot be reached from a black box level.

7.3 Black Box Testing

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box .you cannot “see” into it. The test provides inputs and responds to outputs without considering how the software works.

Unit Testing

Unit testing is usually conducted as part of a combined code and unit test phase of the software lifecycle, although it is not uncommon for coding and unit testing to be conducted as two distinct phases.

Test strategy and approach

Field testing will be performed manually and functional tests will be written in detail.

Test objectives

- All field entries must work properly.
- Pages must be activated from the identified link.
- The entry screen, messages and responses must not be delayed.

Features to be tested

- Verify that the entries are of the correct format
- No duplicate entries should be allowed
- All links should take the user to the correct page.

Integration Testing

Software integration testing is the incremental integration testing of two or more integrated software components on a single platform to produce failures caused by interface defects.

The task of the integration test is to check that components or software applications, e.g. components in a software system or – one step up – software applications at the company level – interact without error.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.8

Acceptance Testing

User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements.

Test Results*:* All the test cases mentioned above passed successfully. No defects encountered.

8. OUTPUT SCREENS

8.1 USER INTERFACES AND OUTPUT INTERFACE.

In this article, we have proposed an image that is used and an expression of certain emotion for the algorithm to process and detect the expression in image and based on that a music playlist will be recommended by the algorithm. It also recommends music playlist based on emotion or a real time image using web cam and from databases, web Databases, spotify, Music applications etc,. The proposed system used support vector machine (SVM) algorithm which helped to improve the recommendations. Based on our experiments, the algorithm has worked accordingly to the expressions being put. In the future, we plan to include the camera within the wearables that are really accurate and work free to further improve the mobility and increase the ease of efficient music track playing while user is busy.

1. To run project install below package
2. Pip install playsound.
3. Double click on 'run.bat' file to get below screen.

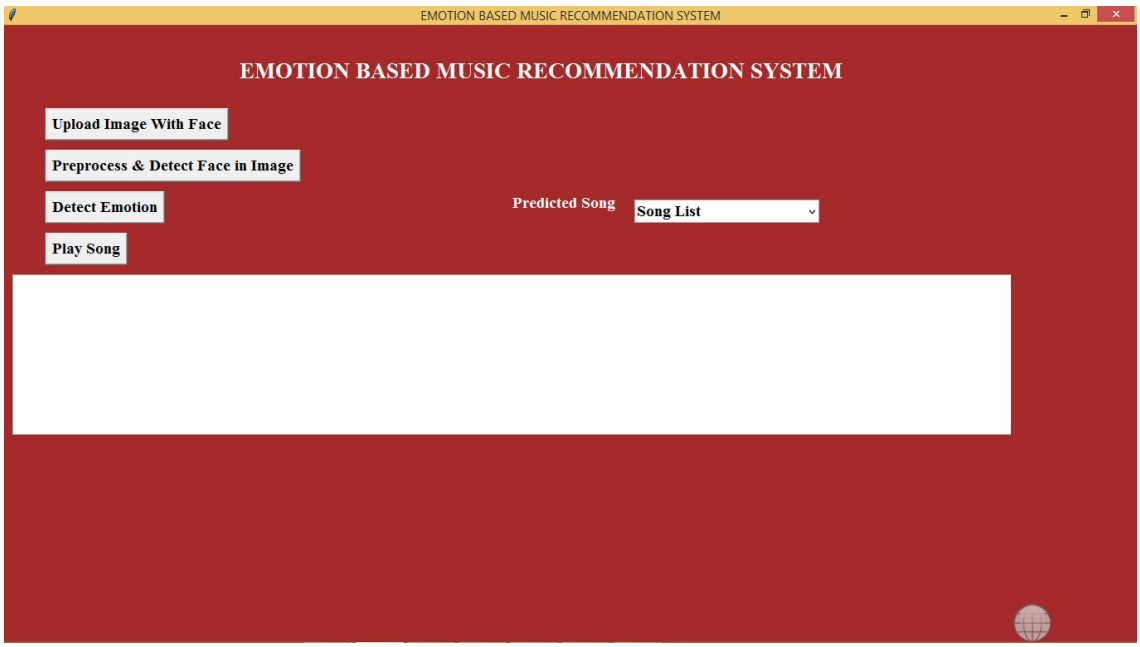


Fig1: Emotion-based music recommendation system window.

4. In the fig.1, there is a click button ‘Upload Image With Face’ to upload image that opens up a new window of your computer files.

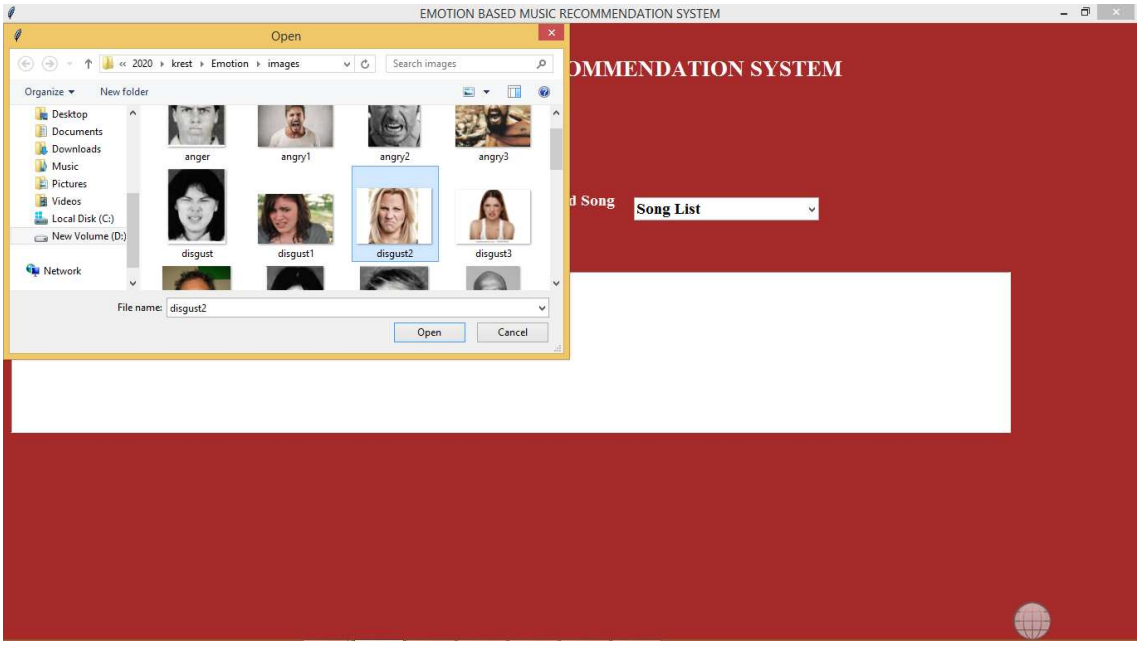


Fig 2: selecting the images having facial expressions from computer.

1. In the fig.2, the user selects the image from his folder in computer to project or propose an emotion for the recommender to work.

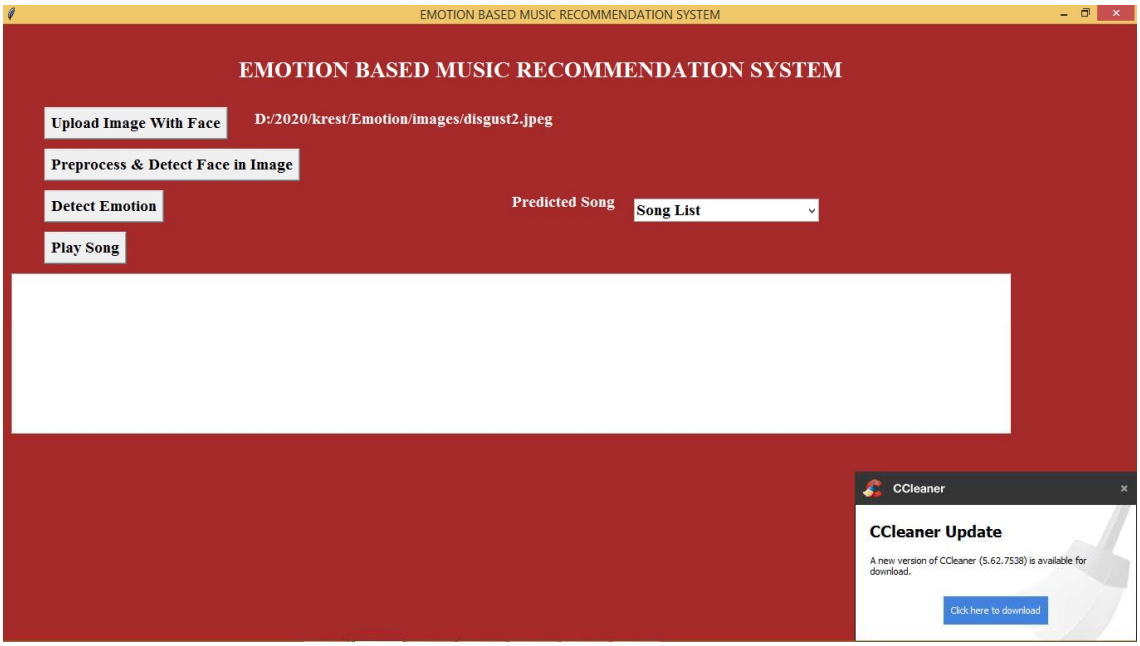


Fig3: clicking the Pre-process and Detect Face in Image.

2. In the fig.3, I am selecting one ‘disgust’ image. Now click on ‘Pre-process & Detect Face in Image’ button to perform pre-processing and to extract face from images.

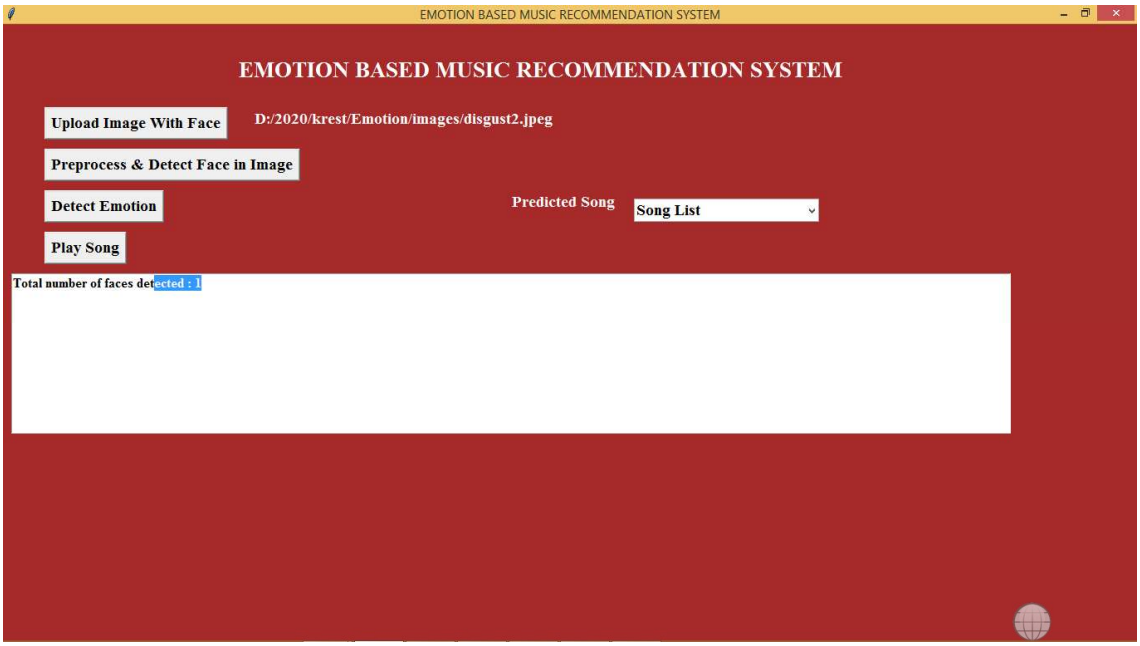


Fig 4: The detection of the Emotion Is performed to obtain results.

1. In the fig.4, we can see in uploaded image one face is detected. Now click on Detect Emotion button to detect emotion.

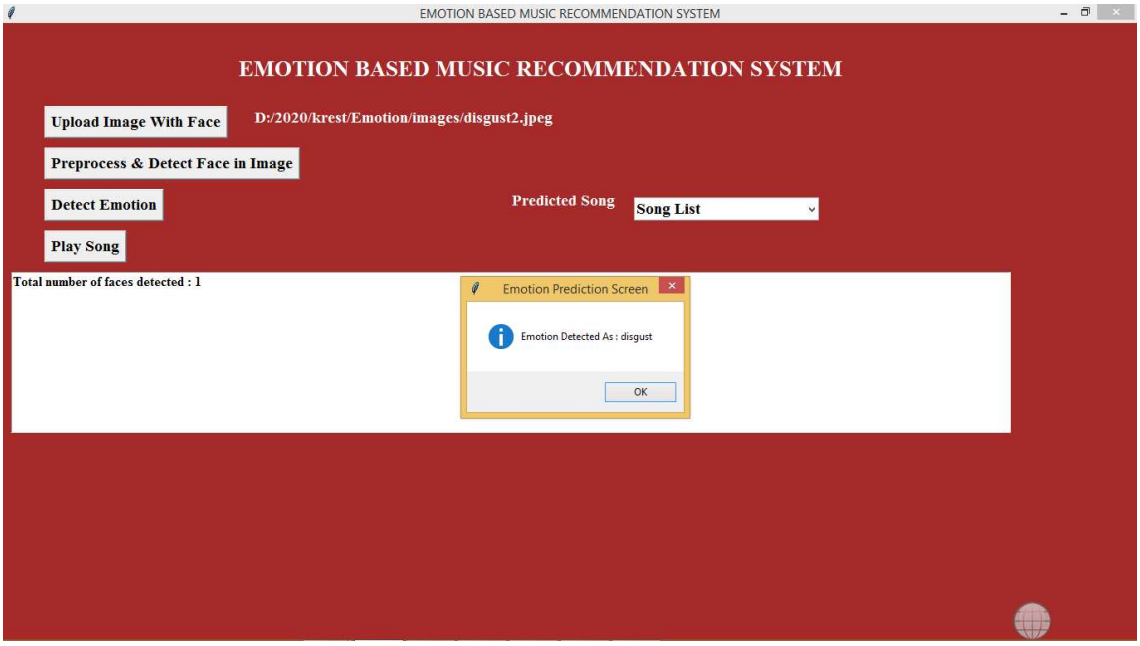


Fig 5: Separate window appeared showing emotion that has been detected.

2. In the fig.5, we can see emotion disgust is detected and now click on drop down arrow link to get all disgust songs list.

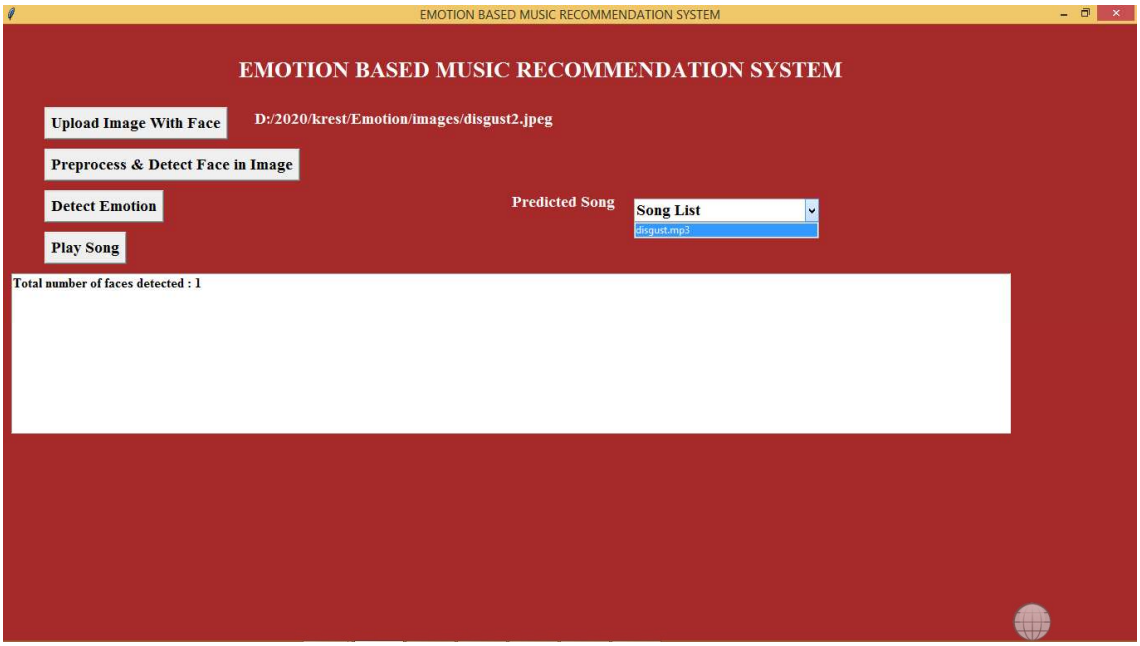


Fig 6: The Song or music file recommended in the list.

3. In fig.6, In drop down box we can see ‘disgust.mp3’ songs is showing, select that song and click on ‘Play Song’ button to play song.

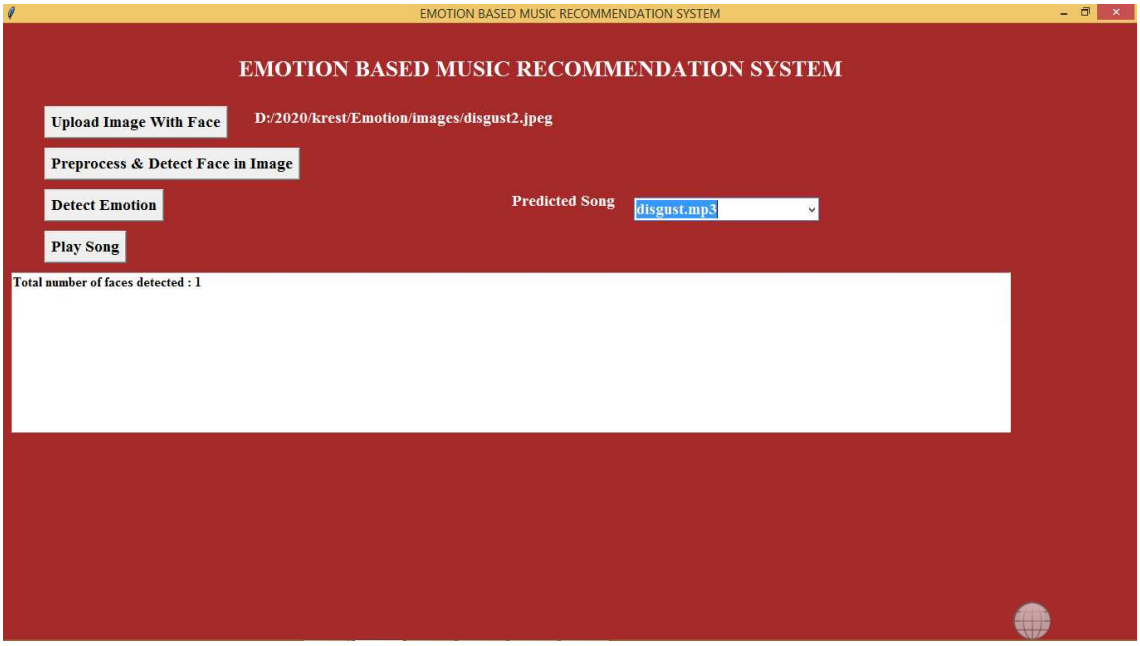


Fig 7: Music or Song automatically playing in the background.

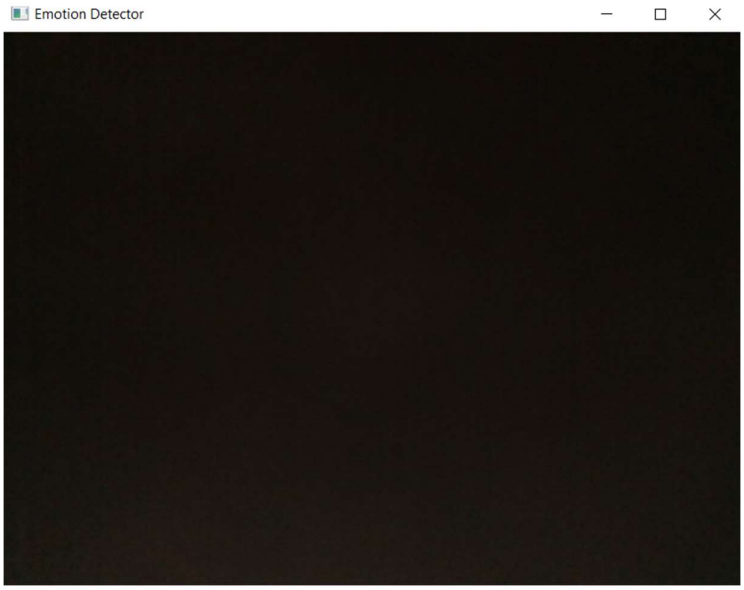
4. The fig.7 is showing, If your system has audio driver then u can hear song.

Note: no technologies can detect 100% emotion from images but this project can detect up to 90%.

9. EXPERIMENTAL RESULTS

9.1 Results

1. The algorithm opens up a dialog box and a window which starts capturing the emotions of the user.



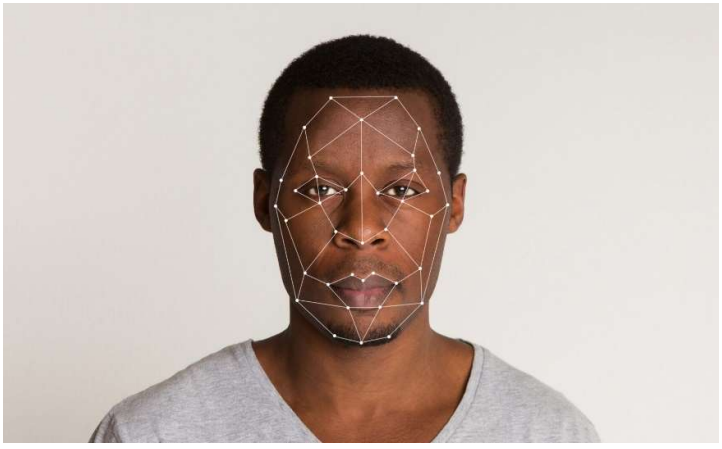


Fig: The window opened for capturing the face.

2. The window captures the expressions of a person and now predicts the expressions according to the Expressions dataset using which it trained.

Fig: The prediction of the Human facial expressions.

3. The Algorithm then starts working and the appropriate music starts playing According to the mood of the user and the Music is selected from web database and also there can be a user preference also like a separate personal folder of music files.

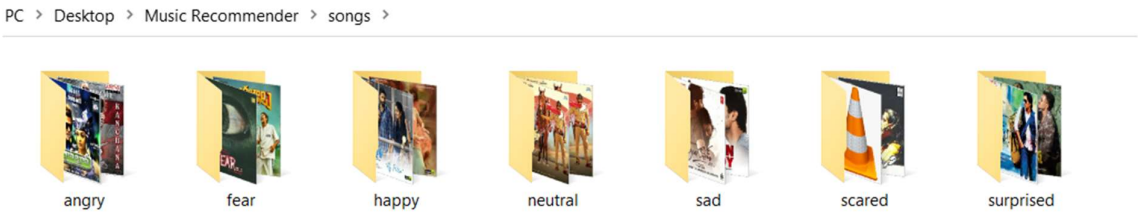


Fig: Music folders segregated according to Mood.

4. The music gets automatically played after the prediction of the face is completed.

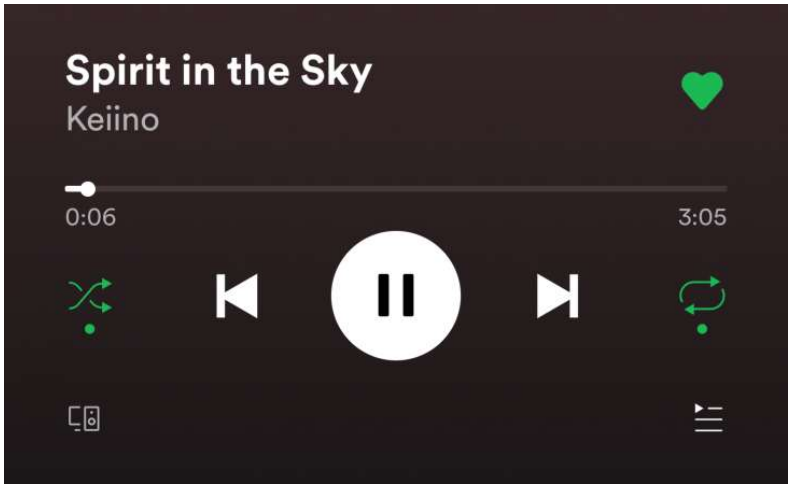


Fig: Music Playing in the Background.

10. CONCLUSION

In this project, we presented a model to recommend a music-based on the emotion based detected from the facial expression. This project proposed designed & developed an emotion-based music recommendation system using wearable physiological (face recognition) sensors System. Music is the one that has the power to heal any stress or any kind of emotions. Recent development promises a wide scope in developing emotion-based music recommendation system. Thus, the proposed system presents Face based emotion recognition system to detect the emotions and play music from the emotion detected.

At the second stage, the system provides generalized playlists which are created with respect to the mood and activity context for an appropriate audience group. When the system learns more about the user, it continuously classifies tracks and tunes the personalized recommendation model. which effects on further playlist creation. Trial experiments showed that the core music attributes such as energy, valence, tempo, and loudness in generated playlists have sufficient matching with attributes of music tracks from MuPsych datasets for the particular mood and activity contexts. Therefore, we can conclude that the system selects music with similar attributes to those, which were explored during emotional state transitions during listening. There are needed further elaborations, comprehensive experiments and validations of the system to judge how recommendations effect mood and match preferences.

Paying attention to various factors, such as particular context, personal parameters, feelings and emotions, is highly important to a decision-making process of recommendations. Contemporary music recommendation systems face the gap in personalization, human feelings, contextual preferences and emotional factors while suggesting music. In this paper, we proposed emotion-driven recommendation system with respect to personalized preferences and particular life and activity contexts.

The approach presented in this study is targeted to provide maximum benefits for people from the music listening experience. It is important to make the system aware of how it is doing the recommendations, to continuously improve the music selection. By feeding the data from various sources, the system is aimed to listen to each particular user and understand their purposes of listening, feelings and contextual preferences to select the best-suited music pieces for them. We observed what kind of data is needed for the recommendation system and how it can be fetched. Main data processing tools are clarified in the scope of this paper and the experimental prototype has been elaborated. However, to achieve maximum accuracy in predictions and make them more or less relevant, machine learning systems require a large amount of the data to train the models. At this moment the data collection is in active process.

FUTUREWORK & ENHANCMENT

The music player based on facial recognition system is highly essential for all the person in modern day life ecology. This system is further enhanced with benefit able features for upgrading in future. The methodology of enhancement in the automatic play of songs are done by detection of the facial expression. The facial expression is detected by programming interface with the RPI camera. An alternative method, based on additional emotions which is excluded in our system as disgust and fear. On this emotion included to support the playing of music automatically.

At the same time this kind of system requires significant clinical research and collaboration with psychologists to tune and test the model for real recommendations and reduce possible associated risks. Further work on the implementation and testing of the recommendation engine, empirical experiments and impact evaluations are considered for the next step when the appropriate amount of the data will be collected. Music creation by artificially intelligent systems with particular music attributes to move states of human emotions can be considered as the further elaboration work in this context.

PUBLICATIONS

Emotion Based Music Recommendation System Using Wearable Sensors

B. Nikhil Reddy[1] , E. Sriman Goud [2], A. Siddharth Reddy[3], B. Vamshi[4], V.L.Kartheek[5]
[1,2,3,4]UG Scholar, [5]Assistant Professor
Department of Computer Science & Engineering
St. Martin's Engineering College,
Near Forest Academy, Dhulapally, Secunderabad, Telangana, India-500014
Email-id: esriman92@gmail.com[1], nikhilreddy0803@gmail.com[2] , vamc526@gmail.com[3], siddu199809@gmail.com[4], kartheekv999@gmail.com[5]

ABSTRACT:

Most of the existing music recommendation systems use collaborative or content-based recommendation engines. However, the music choice of a user is not only dependent to the historical preferences or music contents. But also dependent to the mood of that user. This paper proposes an emotion-based music recommendation framework that learns the emotion of a user from the signals obtained via wearable Accessories which include sensors. In particular, the emotion of a user is classified by a wearable computing device which is integrated with a galvanic skin response (GSR) and photo plethysmography (PPG) physiological sensors and an optical sensor like camera. This emotion information is feed to any collaborative or content- based recommendation engine as a supplementary data. Thus, existing recommendation engine performances can be increased using these data. Therefore, in this paper emotion recognition problem is considered as arousal and valence prediction from multi-channel physiological signals. Experimental results are obtained on 32 subjects' GSR and PPG signal data with/out feature fusion using decision tree, random forest, support vector machine and k-nearest neighbour algorithms. The results of comprehensive experiments on real data confirm the accuracy of the proposed emotion classification system that can be integrated to any recommendation engine.

Keywords: Content-Based Emotion Facial Recognition Galvanic-skin-Response Photo-Plethysmography Physiological Sensors Skin Valence Wearable

I. INTRODUCTION:

Stress is a physiological response to the mental, emotional, or physical challenge and it can be defined as the reaction of a person to the environmental requests or influences (Sun et al., 2010). Stress conditions can cause physical and emotional exhaustion that leads to symptoms such as headaches, stomach complaints and difficulties in sleeping. A study conducted by the American Institute of Stress (Statistic Brain Research Institute, NY) has shown as in 2015 the 48% of people feels that their stress condition has increased over the past five years. 77% of people regularly experiences physical symptoms caused by stress with a negative impact on their personal and professional life (Statistic Brain, 2015). The influence of stress and its

consequences on society concerns also the economic aspect. According to the recent EU-funded project 2013, the cost to Europe of work-related stress and depression was estimated to be €617 billion annually. The total amount includes loss of productivity, health care costs and social welfare costs (EU-OSHA, 2016). The early detection of stress can positively affect personal wellbeing and society affluence.

Traditionally, the level of personal stress has been established using some psycho-metric instruments and scales (Ulstein et al., 2007), which are subjective. Subsequently the correlation between the variation of the physiological signals and stress was investigated in order to make the measurement more objective.

II. LITERATURE SURVEY:

Emotions are affective states related to physiological responses. This study proposes a model for recognition [1] of three emotions: amusement, sadness, and neutral from physiological signals with the purpose of developing a reliable methodology for emotion recognition using wearable devices. Target emotions are captured by photoplethysmography, which provides information about heart rate, and galvanic skin response. These signals were analyzed in frequency [1] and time domains to obtain a set of features. Several feature selection techniques and classifiers were evaluated. The best model was obtained with random forest recursive feature [14] elimination, for feature selection, and a support vector machine for classification.

Virtual reality exposure therapy (VRET) [2] can have a significant impact towards assessing and potentially treating various anxiety disorders. One of the main strengths of VRET systems is that they provide an opportunity for a psychologist to interact with virtual 3D environments [2] and change therapy scenarios according to the individual patient's need. Therefore, in order to fully use all advantages provided by the VRET system [13], a mental stress detection system is needed. The patient's physiological signals can be collected with wearable biofeedback sensors [2]. Signals like blood volume pressure (BVP), galvanic skin response (GSR), and skin temperature can be processed and used to train the anxiety level classification models. The acquired data were used to train a four-level anxiety recognition model [11].

Meanwhile, recent studies have demonstrated as user personality can effectively provide a more valuable information to significantly improve recommenders' [3] performance, especially considering behavioral data captured from social network logs. In this work, we describe a novel music recommendation technique based on the identification of personality traits, moods and emotions of a single user, starting from solid psychological observations [3] recognized by the analysis of user behavior within a social environment [11]. In particular, users personality and mood have been embedded within a content-based filtering approach to obtain more accurate and dynamic results

Facial expression recognition is an important research issue in the pattern recognition field. However, the generalization of the model still remains a challenging task. We apply a strategy of curriculum learning to facial expression [4] recognition during the stages of training. And a novel curriculum design method is proposed. The system first employs the unsupervised density–distance clustering [15] method to determine the clustering center of each category. Then, the dataset is divided into three subsets of various complexity [4] according to the distance from each sample to the clustering center in the feature space.

Facial expression recognition is a challenging problem in image classification. This has led to increased efforts in solving the problem of facial expression recognition using convolutional neural networks [5] (CNNs). A simple architecture is fast to train [16] and easy to implement. An effective architecture achieves good accuracy on the test data. CNN architectures are black boxes to us. VGG Net, Alex Net and Inception are well-known CNN architectures. These architectures have strongly influenced CNN model designs for new datasets. This work tries to overcome this limitation by using FER-2013 [5] dataset as starting point to design new CNN models. In this work, the effect of CNN parameters namely kernel size and number of filters on the classification accuracy is investigated using FER-2013 dataset.

The recognition of emotions for annotating large-size music datasets is still an open challenge [6]. The problem lies in that most of the solutions require the audio of the songs and user/expert intervention during certain phases of the recognition process [16]. It consists of a heterogeneous set of machine learning models that have been developed from Spotify's Web data services and miner tools. In order to improve the accuracy of resulting annotations, each model is specialized in recognizing a class of emotions [8]. These models have been validated by using the Acoustic Brainz [6] database and have been exported to be integrated into a music emotion recognition system.

Music can have a positive influence on long-distance runners' motivation and performance. It requires selecting the most suitable music by considering the runner's physiological data, the type of training session and the geographical and environmental conditions under which the activity is done. We are interested in studying the runners' emotions during the training sessions and in using these emotions to recommend [7] personalized music that increases their motivation and performance. More specifically, in this paper we present an adapted glove that integrates different sensors for collecting data, which help to determine the runner's emotional state, and the changes that it experiences [9]. Preliminary results about the interpretation of these data and emotions are discussed and a prototype of recommendation system based on Spotify [10] is sketched.

III. PROPOSED METHODOLOGY:

The Algorithm used for the Emotion based music recommender is Support Vector Machine a Machine Learning Algorithm. An interpreted language, Python has a design philosophy that emphasizes code readability (notably using whitespace indentation to delimit code blocks rather than curly brackets or keywords), and a syntax that allows programmers to express concepts in fewer lines of code than might be used in languages such

as C++or Java. It provides constructs that enable clear programming on both small and large scales. Python interpreters are available for many operating systems.

Effective in high dimensional spaces.
Still effective in cases where number of dimensions is greater than the number of samples.
Uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.
Versatile: different Kernel functions can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels.

Electrodermal activity

Psycho physiological measures have been recently used in HRI studies, in which, in addition to HRV, Galvanic Skin Response (GSR) has been used. The neural mechanism and pathways involved in the central control of electrodermal activity are numerous and complex. EDA is related to the level of arousal elicited by an extended range of psychological and emotional states with either positive or negative valence. Different studies investigating anxiety, anger, fear and also joy experiences report increased EDA (Ritz et al., 2000; Stemmler et al., 2001). It is also an indicator of the cognitive load, stress and arousal (Park, 2009; Haapalainen et al., 2010), because of the variation of the skin electrical resistance in response to various emotional stimuli. When a subject is under mental stress, sweat gland activity is activated and increases skin conductance (SC). Since the sweat glands are also controlled by the SNS, SC acts as an indicator for sympathetic activation due to the stress reaction (Sun et al., 2010).

GSR has already been used in previous works in combination with other physiological parameters. For example, SC has been combined with electro cardiac activity, electromyographic activity and respiration activity in order to monitor drivers' behaviors through open roads (Haley & Picard, 2005). In particular, the parameters provided were the number of stressors in a given temporal window, the sum of the amplitude of all the stressors counted in that temporal window, the sum of the response durations and the sum of the areas under the peaks counted as stressors. Finally, the integration of GSR, HRV and accelerometer data has been implemented in the work of Sun et al. (2010), with the aim to differentiate between physical activity and mental stress. In particular, electrodermal activity has been analyzed through three main parameters: the number of the stressor, the related amplitude and the sum of the duration of the responses.

Experimental Protocol

The experimental protocol was intended to put the subjects in a state of emotional and cognitive stress, in order to measure the variations of their physiological parameters induced by stress. The experimentation consisted in three phases: a baseline, a stress induction and a recovery stage. During baseline the subjects relaxed in a separate room, for 10 minutes, without using mobile phone, without music or external sounds, without stimuli and without closing their eyes. This phase was indispensable in order to acquire the personal baseline of each subject, since physiological parameters show a wide inter-subjects variability. At the end of baseline recording, the psychologist administered psychometric instruments to the participants to obtain a subjective perception about the level of stress, anxiety and drowsiness. Then the subjects

performed the stress phase, during about 15-20 minutes, completing a series of extremely demanding cognitive tests handed out by the psychologist in order to induce the stress. People were not aware that this phase was part of the experiment: the psychologist indeed pretend-ed to be sent by university to detect the intelligence quotient (IQ) for a poll. The investigator assumed a very aggressive behavior towards the subject, behaving rude and correcting the person even when the he accomplished the task properly. Further-more, the user performed the required tasks by listening a noisy sound in background

Galvanic Skin Response (GSR)

The EDA has been recorded using Shimmer GSR sensor which provides as output the galvanic resistance, that has been converted into galvanic skin conductance. In the feature extraction algorithm, the signal has been analyzed with temporal windows of 2 minutes, after a filtering process, using a moving average filter. The features extraction algorithm is based on startle detection that can lead to a set of computable features. The method used is referred to the scoring multiples response method of Boucsein (Boucsein, 2012), that establishes a local baseline at the level of the onset of the second response and measures the distance from that baseline to the following peak. The detection algorithm identifies all the occurrences of when the first derivative exceeded to a certain threshold. It was empirically determined as 0.005 μS. Given the variability of GSR signal among subjects, this threshold is not absolute but it has found to be adequate for the 12 subjects analyzed. Furthermore, to ensure to not con-sider subsequent startles, a minimum distance has been chosen as in (Shumm et al., 2008, considering that a startle event is expected to last about 1-3 s. Once the response was detected, the zero-crossing of the derivative preceding and following the response were identified as the onset and end of the startle (Haley & Picard, 2000).

Electro cardiac activity

Electro cardiac activity has been recorded using the chest belt Zephyr Bio-Harness ™ BH3. The device provides as output the raw ECG signal and the HRV data that specifies the temporal distance between a beat and the following one. Starting from Inter-Beat-Interval (IBI), the algorithm to extract the main features has been developed. The IBI signal has been modified identifying and correcting ectopic rhythm, which is an irregular heart rhythm due to a premature heartbeat. The analysis of cardiac signal has been structured investigating both the time domain and the frequency domain.

Data processing and Statistical Analysis

After extracting features from physiological signals, Kolmogorov-Smirnov test was applied in order to verify the normal distribution of data. A non-parametric statistical analysis was used because the test showed data were not normally distributed. Then, Kruskal-Wallis (KW) test was used for comparing data acquired in baseline phase and those recorded during stress phase in order to verify a significant difference ($p\text{-value}<0.05$) on the basis of the extracted

parameters. Furthermore, the linear correlation between the significant parameters was calculated using the Pearson's coefficient. If the value of correlation between two features was at least rho=0.8, the less significant one was deleted. Then, the remaining features were used for Principal Component Analysis (PCA) in order to identify how the groups investigated, related to different phases of the experimental protocol, could be visualized and separated in the space of the principal components (PCs). Finally, the most important PCs, that included more than 80% of the overall variance of data, were taken into account in order to train and test a Support Vector Machine (SVM) classifier which had to be able to correctly classify a subject as stressed or not-stressed. Regarding the analysis of the psychometric instruments, a T-test has been conduct-ed in order to assess if significant differences between after and before the stress in-duction phase could be revealed. Finally, a linear regression analysis has been implemented with the aim to look for a correlation between the results obtained by the psychometric instruments administered and the physiological parameters measured.

IV. SYSTEM ARCHITECTURE:

This is a simple flowchart representing a process for dealing with a Music recommender System. A flowchart is a type of diagram that represents a workflow or process. Flowchart can also be defined as a diagrammatic representation of an algorithm, a step-by-step approach to solving a task.

The flowchart shows the steps as boxes of various kinds, and their order by connecting the boxes with arrows. This diagrammatic representation illustrates a solution model to the given problem. Flowcharts are used in analyzing, designing, documenting or managing the processing various fields.

REFERENCES

[1] J.A. Domínguez-Jiménez, K.C. Campo-Landines, J.C. Martínez-Santos, E.J. Delahoz, S.H. Contreras-Ortiz "A machine learning model for emotion recognition from physiological signals". Berlin, Germany: 2018.

[2] Justas Šalkevicius, Robertas Damaševičius, Ilona Laukienė, "Anxiety Level Recognition for Virtual Reality Therapy System Using Physiological Signals" (Kaunas University of Technology)., 2019.

[3] Vincenzo Moscato; Antonio Picariello; Giancarlo Sperli., "An emotional recommender system for music". (IEEE institute) 2020.

[4] Xiaoqian_Liu & Fengyu Zhou , "Improved curriculum learning using SSM for facial expression recognition", (The Visual Computer) 2020.

[5] Abhinav_Agrawal & Namita_Mittal , "Using CNN for facial expression recognition: a study of the effects of kernel size and number of filters on accuracy", (the Visual Computer team) 2020.

[6] J. García de Quirós, S. Baldassarri, J. R. Beltrán, A. Guiu, P. Álvarez, "An Automatic Emotion Recognition System for Annotating *Spotify*'s Songs", (University of Zaragoza, Zaragoza, Spain) 2018.

[7] José Ramón Beltrán, Sandra Baldassarri, "DJ-Running: Wearables and Emotions for Improving Running Performance", (Department of Electronic Engineering and Communications, Aragón Institute of Engineering Research) 2019.

[8] Kelly_Brooks and Kristal_Brooks, "Enhancing sports performance through the use of music", (American Society of Exercise Physiologists) 2018.

[9] Anagha S. Dhavalikar and Dr. R. K. Kulkarni, "Face Detection and Facial Expression Recognition System" 2018 Interntional Conference on Electronics

[10] Yong-Hwan Lee, Woori Han and Young seop Kim, "Emotional Recognition from Facial Expression Analysis using Bezier Curve Fitting", 2017 14th International Conference on Network-Based Information Systems.

[11] Arto Lehtiniemi and Jukka Holm, "Using Animated Mood Pictures in Music Recommendation", 2018 16th International Conference on Information Visualisation.

[12] F. Abdat, C. Maaoui and A. Pruski, "Human computer interaction using emotion recognition from facial expression". 2019.

[13] T.-H. Wang and J.-J.J. Lien, “Facial Expression Recognition System Based on Rigid and Non-Rigid Motion Separation and 3D Pose Estimation,”, 2020.

[14] Renuka R. Londhe, Dr. Vrushshen P. Pawar, “Analysis of Facial Expression and Recognition Based On Statistical Approach”, International Journal of Soft Computing and Engineering, 2019.

[15] Anukriti Dureha “An Accurate Algorithm for Generating a Music Playlist based on Facial Expressions”, 2020.

[16] Bruce Ferwerda and Markus Schedl “Enhancing Music Recommender Systems with Personality Information and Emotional States”, 2017.

[17] Agnes Savill. Music & Letters. Oxford University Pres. 2018.

[18] Fancourt, Daisy & Ockelford, Adam & Belai, Abi. (2020). The Psychoneuroimmunological Effects of Music: A Systematic Review and A New Model. Brain, behavior, and immunity.

[19] Musliu, Arian & Berisha, Blerta & Latifi, Diellza & Peci, Djellon. (2019). The Impact of Music in Memory. European Journal of Social Sciences Education and Research.

[20] Erkkilä, Jaakko & Punkanen, Marko & Fachner, Jörg & Ala-Ruona, Esa & Pöntiö, Inga & Tervaniemi, Mari & Vanhala, Mauno & Gold, Christian. (2020). Individual music therapy for depression – Randomised Controlled Trial. The British journal of psychiatry: the journal of mental science.

[21] Erkkilä, Jaakko & Brabant, Olivier & Saarikallio, Suvi & Ala-Ruona, Esa & Hartmann, Martin & Letule, Nerdinga & Geretsegger, Monika & Gold, Christian. (2019).

[22] Helfenstein S., Kaikova O., Khriyenko O., Terziyan V., Emotional Business Intelligence: Enabling Experience-Centric Business with the Feelings Explorer, Proceedings of the 7th International Conference on Human System Interaction (2020).

[23] Vempala, Naresh & Russo, Frank. (2018). Modeling Music Emotion Judgments Using Machine Learning Method

PROFILES OF STUDENTS

EMAGOUDA SRIMAN GOUD is currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. He completed his intermediate from Sri Gayatri Junior College and 10th class at Bhashyam Educational Institutions. His technical skills include C, Python and Java. He also has a good grip on the Machine Learning Hands-on Python. He took part an Internship at Goal Street Internship and Worked on a machine learning Project. He is also a student of Smart Interviews. His participations include: National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020, "Know More - Teach More ", the Global Webinar on Cloud & Big Data – Changing the way we work which was conducted Global Education & Careers Forum (GECF) on 12th August 2020. And National Level Two-Day Symposium "Automatic Accident Detection System" Project, IIC Online Sessions conducted by Institution's Innovation Council (IIC) of MHRD's Innovation Cell, New Delhi to promote Innovation, IPR, Entrepreneurship, and Start-ups among HEIs from 28th.He is also participated in webinar on Digital Transformation in Education Sector Post-Covid era. He is also participated in many sports event in college.

His areas of interest are Python, Artificial Intelligence, Machine Learning and Data science, big data and Cybersecurity. He completed few certification courses from online platforms like Coursera, CursaApp.

BANDA NIKHIL REDDY is currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. She completed her intermediate from Narayana Junior College and 10th class from Balaji Techno School. His technical skills include C, Python and Java. He also has a basic understanding of C++. He took part in Employability Skill development Program conducted by Zensar. He is also a student of Smart Interviews. His participations include: National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020, "Five days Online International Hands-on Certification Training in Python Programming" from 20th to 24th August 2020, "IIC Online Sessions conducted by Institution's Innovation Council (IIC) of MHRD's Innovation Cell, New Delhi to promote Innovation, IPR, Entrepreneurship, and Start-ups among HEIs from 28th

April to 22nd May 2020, "Rough Set Theory and it's Applications" organized by Department of CSE, Vignan's foundation for Science, Technology and research on 11th June 2021 and Two Day National Level SEMINAR on "Recent Trends Cloud Computing, Fog and Edge Computing " conducted from 18th June to 19th June 2021. His areas of interest are Python, Artificial Intelligence, Machine Learning and Deep Learning. He completed few certification courses from online platforms like Coursera, CursaApp and SoloLearn, Edapt, Skill Up,Learn mall.

ANTHAREDDY SIDDHARTHA is currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. He completed his intermediate from Urbane Junior College and 10th class from Sacred heart academy. His technical skills include C, Python and Java. He also has a basic understanding of C++. He took part in Employability Skill development Program conducted by Zensar. He is also a student of Smart Interviews. His participations include: National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020, "Know More - Teach More ", the Global Webinar on Cloud & Big Data – Changing the way we work which was conducted Global Education & Careers Forum (GECF) on 12th August 2020.IIC Online Sessions conducted by Institution's Innovation Council (IIC) of MHRD's Innovation Cell, New Delhi to promote Innovation, IPR, Entrepreneurship, and Start-ups among HEIs from 28th.He is also participated in webinar on Digital Transformation in Education Sector Post-Covid era. He is also participated in many sports event in college.

His areas of interest are Python, Artificial Intelligence, Machine Learning and Data science, big data. He completed few certification courses from online platforms like Coursera, CursaApp and SoloLearn.

BAIRAGONI VAMSHI GOUD is currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. he completed his intermediate from Narayana Junior College and 10th class from Bhashyam High School. he is one of the members of Coders Club in our college. His responsibilities in that group include mentoring and motivating students to take coding as a serious hobby. His technical skills include C, Python and Java. he also has a basic understanding of C++. he took part in Employability Skill development Program conducted by Zensar. he is also a student of Smart Interviews. His participations include: National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020, Two-day workshop on Machine Learning and Ethical Hacking which was conducted in the month of February year 2019 and 2020 respectively. "Know More - Teach More ", the Global Webinar on Cloud & Big Data – Changing the way we work which was conducted Global Education & Careers Forum (GECF) on 12th August 2020, Women online workshop on "Women in Cyber Security and Privacy in 2020" which was conducted from 6th to 10th July 2020, "Know More - Teach More ", the Global Webinar on Cyber Threats and Defense Techniques conducted by GECF on 22nd July 2020, "One Day Webinar on Internet of Things and Its Applications" conducted by Anand Institute of Higher Technology on 21st May 2020 and IIC Online Sessions conducted by Institution's Innovation Council (IIC) of MHRD's Innovation Cell, New Delhi to promote Innovation, IPR, Entrepreneurship, and Start-ups among HEIs from 28th

April to 22nd May 2020. His areas of interest are Python, Artificial Intelligence, Machine Learning and Deep Learning. he has completed few certification courses from online platforms like Coursera, CursaApp and SoloLearn.

APPENDICES

Code Part

```
from keras.models import load_model
from time import sleep
from keras.preprocessing.image import img_to_array
from keras.preprocessing import image
import cv2
from playsound import playsound
import numpy as np
import threading as T

face_classifier = cv2.CascadeClassifier('./haarcascade_frontalface_default.xml')
classifier =load_model('./Emotion_Detection.h5')

class_labels = ['Angry','Happy','Neutral','Sad','Surprise']

cap = cv2.VideoCapture(0)
count=0

def player(emotion):
    emotion=emotion.lower()
    cap2 = cv2.VideoCapture(0)
    font                   = cv2.FONT_HERSHEY_SIMPLEX
    bottomLeftCornerOfText = (10,380)
    bottomLeftCornerOfText2 = (10,430)
    fontScale              = 1
    fontColor              = (255,255,255)
    lineType               = 2
    global s
    def songplayer():
        global s
        if emotion == 'neutral':
            s = 'Nee Chepakallu.mp3'
            playsound('SONGS/neutral/Nee Chepakallu.mp3')
```

```
    if emotion == 'scared':
        s = 'Chandramukhi.mp3'
        playsound('SONGS/scared/Chandramukhi.mp3')

    if emotion == 'surprised':
        s = 'Seheri.mp3'
        playsound('SONGS/surprised/Seheri.mp3')

    if emotion == 'angry':
        s = 'Harima Harima.mp3'
        playsound('SONGS/angry/Harima Harima.mp3')

    if emotion == 'fear':
        s = 'Tanha Tanha.mp3'
        playsound('SONGS/fear/Tanha Tanha.mp3')

    if emotion == 'happy':
        s = 'Ay Pilla.mp3'
        playsound('SONGS/happy/Ay Pilla.mp3')

    if emotion == 'sad':
        s = 'Yetu Pone.mp3'
        playsound('SONGS/sad/Yetu Pone.mp3')

x = T.Thread(target=songplayer)
x.start()
while True:
    ret, frame = cap2.read()
    labels = []
    gray = cv2.cvtColor(frame,cv2.COLOR_BGR2GRAY)
    faces = face_classifier.detectMultiScale(gray,1.3,5)

    for (x,y,w,h) in faces:
```

```
            cv2.rectangle(frame,(x,y),(x+w,y+h),(255,0,0),2)
            roi_gray = gray[y:y+h,x:x+w]
            roi_gray = cv2.resize(roi_gray,(48,48),interpolation=cv2.INTER_AREA)
        t = "Detected Emotion = "+emotion
        t2 = 'Playing = '+s
        cv2.putText(frame,t,
            bottomLeftCornerOfText,
            font,
            fontScale,
            fontColor,
            lineType)
        cv2.putText(frame,t2,
            bottomLeftCornerOfText2,
            font,
            fontScale,
            fontColor,
            lineType)
        cv2.imshow('Emotion Detector',frame)
        if cv2.waitKey(1) & 0xFF == ord('q'):
            break
    cap2.release()

while True:
    # Grab a single frame of video
    ret, frame = cap.read()
    labels = []
    gray = cv2.cvtColor(frame,cv2.COLOR_BGR2GRAY)
    faces = face_classifier.detectMultiScale(gray,1.3,5)

    for (x,y,w,h) in faces:
        cv2.rectangle(frame,(x,y),(x+w,y+h),(255,0,0),2)
        roi_gray = gray[y:y+h,x:x+w]
        roi_gray = cv2.resize(roi_gray,(48,48),interpolation=cv2.INTER_AREA)
```

```
        if np.sum([roi_gray])!=0:
            roi = roi_gray.astype('float')/255.0
            roi = img_to_array(roi)
            roi = np.expand_dims(roi,axis=0)

        # make a prediction on the ROI, then lookup the class

            preds = classifier.predict(roi)[0]
            print("\nprediction = ",preds)
            label=class_labels[preds.argmax()]
            print("\nprediction max = ",preds.argmax())
            print("\nlabel = ",label)
            label_position = (x,y)

cv2.putText(frame,label,label_position,cv2.FONT_HERSHEY_SIMPLEX,2,(0,255,0),3)
            count+=1
            if count==10:
                player(label)
                break

        else:
            cv2.putText(frame,'No Face
Found',(20,60),cv2.FONT_HERSHEY_SIMPLEX,2,(0,255,0),3)
        print("\n\n")
    cv2.imshow('Emotion Detector',frame)
    if cv2.waitKey(1) & 0xFF == ord('q'):
        break

cap.release()
```

A

PROJECT REPORT

On

SUPERMARKET BILLING MACHINE USING WEBCAM

Submitted by

Ms.B.AKHILA (17K81A05J9)

Ms.P.LAVANYA (17K81A05M7)

Ms.V.POOJITHA (17K81A05P2)

Ms.R.SINDHUJA (17K81A05N3)

in partial fulfillment for the award of the degree

of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

Under The Guidance of

Dr. K. SRINIVAS

ASSISTANT PROFESSOR

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ST.MARTIN'S ENGINEERING COLLEGE

An Autonomous Institute

Dhulapally, Secunderabad – 500100

JUNE2021

BONAFIDE CERTIFICATE

This is to certify that the project entitled **SUPERMARKET BILLING SYSTEM USING WEBCAM**,isbeing submitted by **B.AKHILA (17K81A05J9),P.LAVANYA(17K81A0M7), V.POOJITHA (17K81A05P2), R.SINDHUJA(17K81A05N3)**inpartial fulfillment of the requirement for the award of thedegree of**BACHELOROF TECHNOLOGYIN** COMPUTER SCIENCE AND ENGINEERINGis recordedofbonafide work carried out by them. The result embodied in this report have been verified and found satisfactory.

Project Guide
Dr. K.Srinivas
Department ofCSE

Head of the Department
Dr.M.Narayanan
Department of CSE

InternalExaminer

External Examiner

Place:

Date:

DECLARATION

We, the students of **Bachelor of Technology** in Department of Computer Science and Engineering, session: 2017 – 2021, St. Martin's Engineering College, Dhulapally, Kompally, Secunderabad, hereby declare that work presented in this Project Work entitled SUPERMARKET BILLING MACHINE USING WEBCAM is the outcome of our own bonafide work and is correct to the best of our knowledge and this work has been undertaken taking care of Engineering Ethics. This result embodied in this project report has not been submitted in any university for award of anydegree.

B.AKHILA	**17K81A05J9**
P.LAVANYA	**17K81A05M7**
V.POOJITHA	**17K81A05P2**
R.SINDHUJA	**17K81A05N3**

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible and whose encouragements and guidance have crowded effects withsuccess.

We extended our deep sense of gratitude to Principal**, Dr. P. SANTOSH KUMAR PATRA**, St. Martin's Engineering College, Dhulapally, for permitting us to undertake this project.

We are also thankful to **Dr.M.Narayanan**, Head of the Department, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project and as well as our project coordinator **Mr.Santhosh Kumar**, Associate Professor, Department of Computer Science and Engineering for his valuable support.

We would like to express our sincere gratitude and indebtedness to our project supervisor **DR. K SRINIVAS**, Assistant Professor, Department of Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

Finally, we express thanks to all those who have helped us successfully to completing this project. Furthermore, we would like to thank our family and friends for their moral support and encouragement.

We express thanks to all those who have helped us in successfully completing the project.

B.AKHILA17K81A05J9

P.LAVANYA 17K81A05M7

V.POOJITHA17K81A05P2

R.SINDHUJA 17K81A05N3

INDEX

ABSTRACT

Supermarket is the place where customers come to purchase their daily essential products and pay for them through traditional billing. Nowadays, if a consumer would like to buy something at a shopping mall, consumers need to take the particular items from the display shelf and then queue up and wait for their turn to make payment. Problem will surely arise when the size of a shopping mall is relatively huge and sometimes consumers don't even know where certain items are placed. Besides, consumers also need to queue for a long time at the cashier to wait for turn to make payment, Weneed to calculate how many products are sold and generate the bill for the customer. We propose a automated billing system that also helps the customer to have a hassle-free shopping experience. We have 2 users in the system. First one is the administrator who will decide the taxes and commissions on the products and can see the report of any product. He is the one who will decide the products available for customers. The second one is the customer or the billing manager who can purchase the items available or can make the bill for the customers. This can also be used for online purchasing as the customer can access it easily. To develop a supermarket basket that assists the customer to locate and select products & inform them on the products details in the shopping arena. Additionally, with each product identified uniquely and support billing and inventory updates. We develop smart shopping system for the customer that assists the customer to locate the shelves where the product. Also by using the concept of market basket analysis we can solve the problems of the customers to find the items related to thatproduct.

LIST OF FIGURES

1. INTRODUCTION

Supermarket is the place where customers come to purchase their daily essential products and pay for them through traditional billing. We need to calculate how many products are sold and generate the bill for the customer. We have 2 users in the system. First one is the administrator who will decide the taxes and commissions on the products and can see the report of any product. He is the one who will decide the products available for customers. The second one is the customer or the billing manager who can purchase the items available or can make the bill for the customers. This can also be used for online purchasing as the customer can access it easily. Nowadays, if a consumer would like to buy something at a shopping mall, consumers need to take the particular items from the display shelf and then queue up and wait for their turn to make payment. Problem will surely arise when the size of a shopping mallis relatively huge and sometimes consumers dont even know where certain items are placed. Besides, consumers also need to queue for a long time at the cashier to wait for turn to make payment, The time taken for consumers to wait for the consumers in front of the queue to scan every single item and then followed by making payment will definitely take plenty of time. This condition will surely become worst during the season of big sales or if the shopping mall still uses the conventional way to key in the price of every item by hand to the cash register. On the other hand, consumers often have to worry about plenty of things when going to the shopping mall. While doing survey we found that most of the people prefer to leave the shopping mall instead of waiting in long queues to buy a few products. People find it difficult to locate the product they wanted to buy, after selecting product they need to stand in a long queue for billing and payment. To try to solve the problems previously identified, recent years have seen the appearance of several technological solutions for hypermarket assistance. All such solutions share the same objectives to save consumers. In the present scenario, it is essential to have an automatic billing system for shopping malls, supermarket and other wholesale & retail stores. Numerous billing systems like barcode scanning mechanism-based systems or tag-based systems are available in the market. It is important to replace such existing system with better and robust systems so hereby we proposed “Supermarket billing systemusing

webcam". In this system, the basic fundamental is barcode scanning for products, but we replace the conventional barcode scanner for faster and better results.

1.1 PROJECT OVERVIEW

The projectis employed for automating the billing system in supermarkets, the database of this project will consists of some predefines shapes. The camera will capture the image of goods, it will find the objects which are predefined then it compares with the database, the software part will calculate the amount of bill. Now there will be two sorts of customers registered and non-registered, if the customer is registered then the amount of bill can directly be debited from his account. Barcodes are widely used in many grocery supermarkets like Hypermarket, D-mart, etc..In our prototype, the android phone is being used as a barcode scanner for simple, better and portable barcode scanner. This scanner is connected wirelessly to MCU via Bluetooth module. MCU is also connected to PC/Laptop for creating the database of all customers, their products, and bills. This database also tracks the total sale and number of goods sold per day. In addition, RFID technology is implemented in this system for payment through card-based system. Simulation and hardware-based results are proposed in this paper. Unstaffed retail shop has been emerging in the past few years and significantly affected conventional shopping styles. In this field unmanned retail container plays an important role, it can highly influence the user shopping experience, the traditional method on weighing sensors cannot sense what the customer is taking. This paper proposes a smart unstaffed retail shop scheme based on image processing & open CV python which aiming at exploring the feasibility of implementing the unstaffed retail shopping style. The merits of this project are that it use open CV python which gives more than 98% accuracy & It is better than manual counting while the demerits are employability will be decreased because one machine can do work of many persons. To try to solve the problems previously identified, recent years have seen the appearance of several technological solutions for hypermarket assistance. All such solutions share the same objectives to save consumers.

1.2 PROJECT OBJECTIVES

The objectives for the smart shopping cart system project is to make the shopping easy for the customer in the supermarket and can save the time of the customer waiting in the queue as the bill is already made in the customer's screen by individually scanning their product and add into their cart. We always see that in a big Shoppe the customer fond to be hard to find the products they need to ask for the helper or the owner of the Shoppe and also, they need hold up in the line in the billing counter. Sometimes might be finding products is easy than waiting in the billing queue because it consumes more time of the customer. So now by taking the motivation of this scenario which was regularly done in all the Shoppe we are designing this system which can be benefited for the customer. To provide faster service at the checkouts this in the advantage for shop owners is that they will require fewer cashiers, which will result in a huge reduction in their cost. To develop a system which allows customer to pre decided budget and only buys the essential commodities actually needed by him, also the system aids. To remove the long queues at the billing counter. To develop the profitable system for the shopping centers this reduces the number of billing counters and in turn will help in reducing employee costssignificantly.

1.3 SCOPE OF THEPROJECT

To develop a supermarket basket that assists the customer to locate and select products & inform them on the products details in the shopping arena. Additionally, with each product identified uniquely and support billing and inventory updates. We develop smart shopping system for the customer that assists the customer to locate the shelves where the product. Alsoby using the concept of market basket analysis we can solve the problems of the customers to find the items related to that product. The best and most useful example of this market basket analysis is that if a customer purchases bread then he will also purchases the related items that is better using these concepts we can make customer to purchase the relatedproducts.

The scope of the project is described as follows.

- Calculate thebill.
- Give the bill to thecustomers.
- Store how many products aresold.
- Store products and their prices with the information.
- Set the rate of taxes and commission onproducts.
- Can see the report of the product in a fixed period oftime.
- Change the Graphical User Interface of thesystem.

1.4 ORGANISATION OFCHAPTERS

1.4.1 INTRODUCTION

Nowadays, if a consumer would like to buy something at a shopping mall, consumers need to take the particular items from the display shelf and then queue up and wait for their turn to make payment. Problem will surely arise when the size of a shopping mall is relatively huge and sometimes consumers dont even know where certain items are placed. Besides, consumers also need to queue for a long time at the cashier to wait for turn to make payment, The time taken for consumers to wait for the consumers in front of the queue to scan every single item and then followed by making payment will definitely take plenty of time. This condition will surely become worst during the season of big sales or if the shopping mall still uses the conventional way to key in the price of every item by hand to the cash register. On the other hand, consumers often have to worry about plenty of things when going to the shopping mall. While doing survey we found that most of the people prefer to leave the shopping mall instead of waiting in long queues to buy a few products. People find it difficult to locate the product they wanted to buy, after selecting product they need to stand in a long queue for billing and payment. To try to solve the problems previously identified, recent years have seen the appearance of several technological solutions for hypermarket assistance. All such solutions share the same objectives to save consumers.

1.4.2 LITERATURE SURVEY

A number of methods are proposed by researchers in this domain. B. Ananthabarathi proposed High Speed Billing System in which RF detector is placed inside the shopping cart which is linked to the server for billing [3]. R.Rajeshkumar, R.Mohanraj, M.Varatharaj proposed Smart Trolley in which they have used RFID cards for each product and RFID reader with MCU on each trolley for calculating the bills while shopping [4]. P. Chandrasekar, T. Sangeetha have proposed Smart Shopping Cart with Zigbee and RFID in which they utilize RFID cards for each product along with Product Identification Device (PID) for the trolley which is used for calculation of products and bill. This approach used Zigbee for transmitting the billing details to central billing system [1]. Few more researchers have proposed system for billing management but most of the methods are similar in nature and used MCU plus communication module based system for each trolley [5], [6],[7].

1.4.3 REQUIREMENTSSPECIFICATION

SOFTWAREREQUIREMENTS

- Operating system : Windows Family.
- CodingLanguage : Python.
- FrontEnd : Python.
- Designing : HTML, CSS, JavaScript
- IDE :PyCharm.
- Data Base :MySQL.

HARDWARE REQUIREMENTS

- Processor : Any UpdateProcessor.
- RAM : Min.4GB.
- HARD DISK : Min.100GB.

1.4.4 SOFTWARE DEVELOPMENTANALYSIS

MachineLearning

Machine learning is often categorized as a subfield of artificial intelligence, but I find that categorization can often be misleading at first brush. The study of machine learning certainly arose from research in this context, but in the data science application of machine learning methods, it's more helpful to think of machine learning as a means of building models of data. Fundamentally, machine learning involves building mathematical models to help understand data. "Learning" enters the fray when we give these models tunable parameters that can be adapted to observed data; in this way the program can be considered to be "learning" from the data. Once these models have been fit to previously seen data, they can be used to predict and understand aspects of newly observed data. I'll leave to the reader the more philosophical digression regarding the extent to which this type of mathematical, model-based "learning" is similar to the "learning" exhibited by the human brain. Understanding the problem setting in machine learning is essential to using these tools effectively, and so we will start with some broad categorizations of the types of approaches we'll discuss here. At the most fundamental level, machine learning can be categorized into two main types: supervisedlearning and unsupervised learning.

Supervised learning involves somehow modeling the relationship between measured features of data and some label associated with the data; once this model is determined, it can be used to apply labels to new, unknown data. This is further subdivided into classificatio*n* tasks and regression tasks: in classification, the labels are discrete categories, while in regression, the labels are continuous quantities.We

will see examples of both types of supervised learning in the following section.
Unsupervised learning involves modeling the features of a dataset withoutreference toanylabel,andisoftendescribedas"lettingthedatasetspeakforitself."These models include tasks such as clustering and dimensionality reduction. Clustering algorithms identify distinct groups of data, while dimensionality reduction algorithms search for more succinct representations of the data. We will see examples of both types of unsupervised learning in the following section.

1.4.5 PROJECT SYSTEMDESIGN

- **Add Product Details:** To build project I used some sample products image to train product identification models
- **Train Model:** In this Module screen train model generated with100% accuracy and now show product to webcam.
- **Add/Remove Product from basket**: To allow application to identify product image and then show in text area and if we again show same product then application will remove from textarea

1.4.6 PROJECTCODING

Python is an interpreted high-level programming language for general-purpose programming. Created by Guido van Rossum and first released in 1991, Python has a design philosophy that emphasizes code readability, notably using significant whitespace. Python features a dynamic type system and automatic memory management. It supports multiple programming paradigms, including object-oriented, imperative, functional and procedural, and has a large and comprehensive standard library. Python is currently the most widely used multi-purpose, high-level programming language. Python allows programming in Object-Oriented and Procedural paradigms. Python programsgenerally are smaller than other programming languages like Java. Programmers have to type relatively less and indentation requirement of the language, makes them readable all the time.Python

language is being used by almost all tech-giant companies like – Google, Amazon, Facebook, Instagram, Dropbox, Uber… etc. The biggest strength of Python is huge collection of standard library which can be used for the following-

- MachineLearning
- GUI Applications (like Kivy, Tkinter, PyQtetc.)
- Web frameworks like Django (used by YouTube, Instagram,Dropbox)
- Image processing (like Opencv,Pillow)
- Web scraping (like Scrapy, BeautifulSoup,Selenium)
- Testframeworks
- Multimedia

The libraries used in the project are:

TKinter: Tkinter is a standard GUI (graphical user interface) package. Tkinter is Python's default GUI module and also the most common way that is used for GUI programming in Python.

Matplotlib: Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms.

Numpy: Numpy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays.

TensorFlow: TensorFlow is a free and open-source software library for dataflow anddifferentiable programming across a range of tasks.

1.4.7 PROJECTTESTING

The purpose of testing is to discover errors. Testing is the process of trying to discovereveryconceivablefaultorweaknessinaworkproduct.Itprovidesawayto

check the functionality of components, sub assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

Types of tests

Unit testing: Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated.

Integration testing: Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields.

Function testing: Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

System testing: System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results.

White Box testing: White Box Testing is a testing in which in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is purpose. It is used to test areas that cannot be reached from a black box level.

Black Box testing: Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document.

Unit testing: Unit testing is usually conducted as part of a combined code and unit test phase of the software lifecycle, although it is not uncommon for coding and unit testing to be conducted as two distinctphases.

Integration testing: Integration testing is to check that components or software applications, e.g. components in a software system or – one step up – software applications at the company level – interact without error.

User Acceptance testing: Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements.

1.4.8 INPUTSCREENS

The Input Screen allows users to search for transactions using three search options, Quick Search, Passenger Search, and Transaction Search. Common Search Options are applied to these search options to refine the searchresults.

Objectives:

- Input Design is the process of converting a user-oriented description of the input into a computer-based system. This design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerizedsystem.
- It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The goal of designing input is to make data entry easier and to be free from errors. The data entry screen is designed in such a way that all the data manipulates can be performed. It also provides record viewingfacilities.
- When the data is entered it will check for its validity. Data can be entered with the help of screens. Appropriate messages are provided as when needed so that the user will not be in maize of instant. Thus the objective of input design is to create an input layout that is easy tofollow

1.4.9 OUTPUTSCREENS

The design of output is the most important task of any system. During output design, developers identify the type of outputs needed, and consider the necessary output controls and prototype report layouts.

Objectives:

- Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is designed so that people will find the system can use easily and effectively.
- When analysis design computer output, they should Identify the specific output that is needed to meet therequirements.
- Select methods for presentinginformation.
- Create document, report, or other formats that contain information produced by the system.
- Convey information about past activities, current status or projections of the Future.
- Signal important events, opportunities, problems, orwarnings.
- Trigger anaction.
- Confirm anaction.

1.4.10 CONCLUSION

In this Python project, the users are also provided an option to purchase items from the supermarket. The user can view items and then purchase the items which they need. To buy an item, the user needs to enter the product name and then click enter to confirm. This system then displays a message saying the user to pay the price of the item in the counter. In the modern era, people have more income to spend and lesser time to spend, so they typically opt for supermarkets for grocery. Truly the client is ina position & absolves to opt for product from large on the market varieties which attract the large customers mainly in big cities thus therefore long queues ofshoppers

are seen at these stores. In several cases, the barcode is either broken or there is also downside in reading barcode because of lighting effects, low resolution etc.

2. LITERATURESURVEY

2.1 SURVEY ONBACKGROUND

"Image Processing System for Automatic Segmentation and Yield Prediction of fruits using Open CV." (2018)

This paper proposes an image processing system for automatic segmentation and yieldprediction offruitsisproposedonthebasisofcolorandshapefeaturesisbeing performed. Initially the preprocessing is done on input fruit tree images. Then itis converted from RGB to HSV color space to detect the fruit region from its background. Color thresholding is used to mask the desired colors. Gaussian filter is used to remove noise. The contour of the image is taken. Then these imagesare processed by image processing algorithm. Color and shape based counting of fruit is presented at the output. The edge detection and combination of a circular fitting algorithm is applied for the automatic segmentation and automatic counting of fruits in the image. Different types of fruits (orange/tangerine, pomegranate, apple, lemon, mango, cherry) are used for automatic counting. Open CV Python software is used to perform the required image processing operations.

"Object detection and recognition of intelligent service robot based on deep learning." (2018)

This study aims at the accuracy and real-time performance of object detection and recognition of service robot in complex scenes, an end to end object detection and recognition algorithm based on deep learning is proposed. Firstly, the local multi branch deep convolution neural network is adopted to enhance the feature representation capability of the model by enhancing the convolution module function. Then, combining the anchor point mechanism, the object class and position regression prediction is carried out on the multi-layer feature map. When the local features and the global features are fully fused, the natural multi-scale detection and recognition is realized on multiple receptive fields.

"Object Detection and Recognition for Assistive Robots." (2017)

This study presents a vision system for assistive robots that is able to detect and recognize objects from a visual input in ordinary environments in real time. The system computes color, motion, and shape cues, combining them in a probabilistic manner to accurately achieve object detection and recognition, taking some inspiration from vision science. In addition, with the purpose of processing the input visual data in real time, a graphical processing unit (GPU) has been employed. The presented approach has been implemented and evaluated on a humanoid robot torso located at realistic scenarios. For further experimental validation, a publicimage repository for object recognition has been used, allowing a quantitative comparison with respect to other state-of-the-art techniques when realworld scenes are considered.

"New Object Detection, Tracking, and Recognition Approaches for Video Surveillance Over Camera Network." (2015)

This paper proposes a framework for achieving these tasks in a non-overlapping multiple camera network. A new object detection algorithm using mean shift (MS) segmentation is introduced, and occluded objects are further separated with the help of depth information derived from stereo vision. The detected objects are then tracked by a new object tracking algorithm using a novel Bayesian Kalman filter with simplified Gaussian mixture (BKF-SGM). It employs a Gaussian mixture (GM) representation of the state and noise densities and a novel direct density simplifying algorithm for avoiding the exponential complexity growth of conventional Kalman filters (KFs) using GM.

"Fast and Lightweight Object Detection Network: Detection and recognition on resource constrained devices." (2018)

The intrinsic ability of humans to rapidly detect, differentiate, and classify objects allows us to make quick decisions in regards to what we see. Several appliances can make use of fast and lightweight automated object detection for images or videos. Throughout the last five years, the technology industry has constantly introduced computational and hardware solutions, such as devices with impressive processing

and storage capabilities. However, object detection methods usually require either high processing power or large storage availability, making it hard for resource constrained devices to perform the detection in real-time without a connection to a powerful server. The model presented in this paper requires only 95 megabytes of storage and took 113 msin average per image running on a laptop CPU, making it suitable for standalone devices that can be used on thego.

“High Speed Billing System in Departmental Stores.” (2012)

The aim of this study was to make vending of goods in shops completely automatic billing in order to save time. The main idea of the project is to create a complete self functioning rapid billing and dispatch system in a super-market. The system basically consists of a cart fitted with an RF-detector linked to a billing server. When goods in departmental stores are added to the cart, the RF detector automatically detects the item type, quantity and sends the information to the billing server. The billing server simultaneously starts billing for the particular cart. In case the customer removes an item from the cart, the server immediately recognizes it and aligns the billing accordingly. Once the customer acknowledges the server to provide the bill, the server provides the gross amount. The customer on paying the bill is provided an acknowledgement by the server and the goods are dispatched to the customer by the time.

2.2 CONCLUSIONS ON SURVEY

The study is aimed to understand a system employed for automating the billing system in supermarkets, the database will consists of some predefines shapes. The study revealed that Python is the most effective and relevant technology to develop the intended aUtomated system along with support of it’swide range of libraries like OpenCv and Haar Cascade. The challenge is to program the system using OpenCvand Haar Cascade algorithms which enable us to perform image processing, object detection and image tracing functions and to use to camera will capture the image of goods, it will find the objects which are predefined then it compares with the database, the software part will calculate the amount of bill. The most feasible option istouseaandroidphone,whichwillbeusedasabarcodescannerforsimple,better

and portable barcode scanner. RFID technology is the most effective and efficient mechanism to be used in our proposed system. Unstaffed retail shop has been emerging in the past few years and significantly affected conventional shopping styles. In this field unmanned retail container plays an important role, it can highly influence the user shopping experience, the traditional method on weighing sensors cannot sense what the customer is taking. This study also helped us to conclude that a smart unstaffed retail shop scheme based on image processing & open CV python which aiming at exploring the feasibility of implementing the unstaffed retail shopping style. The merits using open CV python which gives more than 98% accuracy &Itis better than manual counting while the demerits are employability will bedecreasedbecauseonemachinecandoworkofmanypersons.Totrytosolvethe problems previously identified, recent years have seen the appearance ofseveral technological solutions for hypermarket assistance.

3. SOFTWARE AND HARDWAREREQUIREMENTS

The project involved analyzing the design of few applications so as to make the application more users friendly. To do so, it was really important to keep the navigations from one screen to the other well ordered and at the same time reducing the amount of typing the user needs to do. In order to make the application more accessible, the browser version had to be chosen so that it is compatible with most of the Browsers.

3.1 SOFTWAREREQUIREMENTS

- Operating system : Windows Family.
- CodingLanguage : Python.
- FrontEnd : Python.
- Designing : HTML, CSS, JavaScript
- IDE :PyCharm.
- Data Base :MySQL.

3.2 HARDWAREREQUIREMENTS

- Processor : Any UpdateProcessor.
- RAM : Min.4GB.
- Harddisk : Min.100GB.

Functional Requirements

- Graphical User interface with theUser.

Debugger and Emulator

- Any Browser (ParticularlyChrome)

4. SOFTWARE DEVELOPMENTANALYSIS

4.1 OVERVIEW OF THEPROBLEM

The existing system is a traditional billing system, the billing is done by barcode scanner we need to detect every barcode attached to every item in purchased item list. When all the items get scanned the price and quantity of items is automatically get into the system and then the bill is get generated. Customers can pay bill through credit/debit cards or by cash. But it is a time consuming process for the billing purpose, so that the waiting time to pay the bill is increased. To overcome the time consuming process the RFID based smart trolley is proposed, along with the help of aurdino and pythontechnologies.

4.2 DEFINE THE PROBLEM

The aim of the proposed system is to To develop a supermarket basket that assists the customer to locate and select products &inform them on the products details in the shopping arena. Additionally, with each product identified uniquely and support billing and inventory updates. We develop smart shopping system for the customer that assists the customer to locate the shelves where the product. Alsoby using the concept of market basket analysis we can solve the problems of the customers to find the items related to that product. The best and most useful example of this market basket analysis is that if a customer purchases bread then he will also purchases the related items that is better than using these traditional billing concepts and we can provide the customer with better purchase experience and help him find the related products. We have two users- admin and customer, and there are two types of customers registered and non-registered customers. Registered customers are provided with a membership card which can be used to access the application and pay the bill instead of credit or debit card. The proposed system is developed using python technology, using the relevant libraries and algorithms like OpenCV and Haar cascade.

4.3 MODULESOVERVIEW

The project we built contains three modules, each module having unique and essential functionality. Our project is to develop a automatic supermarket billing system which uses a web cam to perform the task of image processing, object detection, training the system to detect the added products and displaying the details of the product added to the basket. The three modules included in the system performs the intended operations. The first module, “Add Product Details” is the first module which is used to add a product details to the database, the system opens the camera when you click on this module. The second module, “Train Model” is used to train the system to check the accuracy of the image we added using the webcam, the image must be scanned and added to the module at least 8times to reach the intended 100% accuracy. The final module, “Add/ Remove Product from Basket” enables us to update the listof products, we either add or remove the products according to the availability in stock and personal preference of thecustomer.

4.4 MODULEDEFINITIONS

- **Add Product Details:** To build the project we used some sample products image to train product identification models. This module enables us to details of the detected productimage.
- **Train Model:** In this Module screen train model generated with 100% accuracy and then show the product to webcam.
- **Add/Remove Product from basket**: To allow application to identify product image and then show in text area and if we again show same product then application will remove from text area.

4.5 MODULEFUNCTIONALITIES

The projectis employed for automating the billing system in supermarkets, the database of this project will consists of some predefines shapes. The camerawill

capture the image of goods, it will find the objects which are predefined then it compares with the database, the software part will calculate the amount of bill.

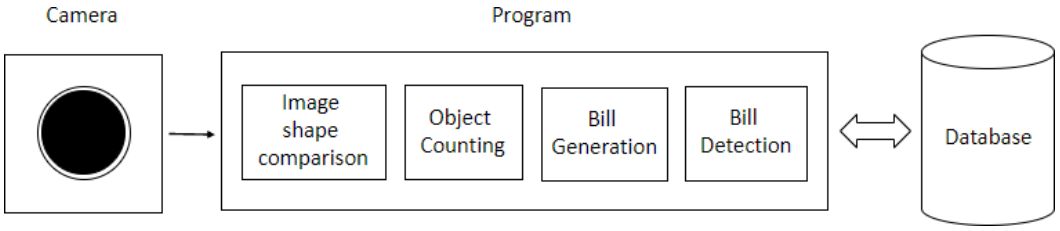


FIG. 4.2.1 SYSTEM ARCHITECTURE

Now there will be two sorts of customers registered and non-registered, if the customer is registered then the amount of bill can directly be debited from his account. Barcodes are widely used in many grocery supermarkets like Hypermarket , D-mart, etc.. In our prototype, the android phone is being used as a barcode scanner for simple, better and portable barcode scanner. This scanner is connected wirelessly to MCU via Bluetooth module. MCU is also connected to PC/Laptop for creating the database of all customers, their products, and bills. This database also tracks the total sale and number of goods sold per day. In addition, RFID technology is implemented in this system for payment through card-based system.

5. PROJECT SYSTEMDESIGN

5.1 DATA FLOWDIAGRAMS

Also known as DFD, Data flow diagrams are used to graphically represent the flow of data in a business information system. DFD describes the processes that are involved in a system to transfer data from the input to the file storage and reports generation. Data flow diagrams can be divided into logical and physical. The logical data flow diagram describes flow of data through a system to perform certain functionality of a business. The physical data flow diagram describes the implementation of the logical data flow.

GOALS:

- To graphically represent the functions, or processes, which capture, manipulate, store, and distribute data between a system and its environment and between components of asystem.
- The visual representation makes it a good communication tool between User and System designer. Structure of DFD allows starting from a broad overview and expand it to a hierarchy of detailed diagrams. DFD has often been used due to the followingreasons:
- To represent logical information flow of the system
- Helps in determination of physical system constructionrequirements
- Simplicity ofnotation
- To establish manual and automated systemsrequirements

5.1.1 LEVEL 0DFD

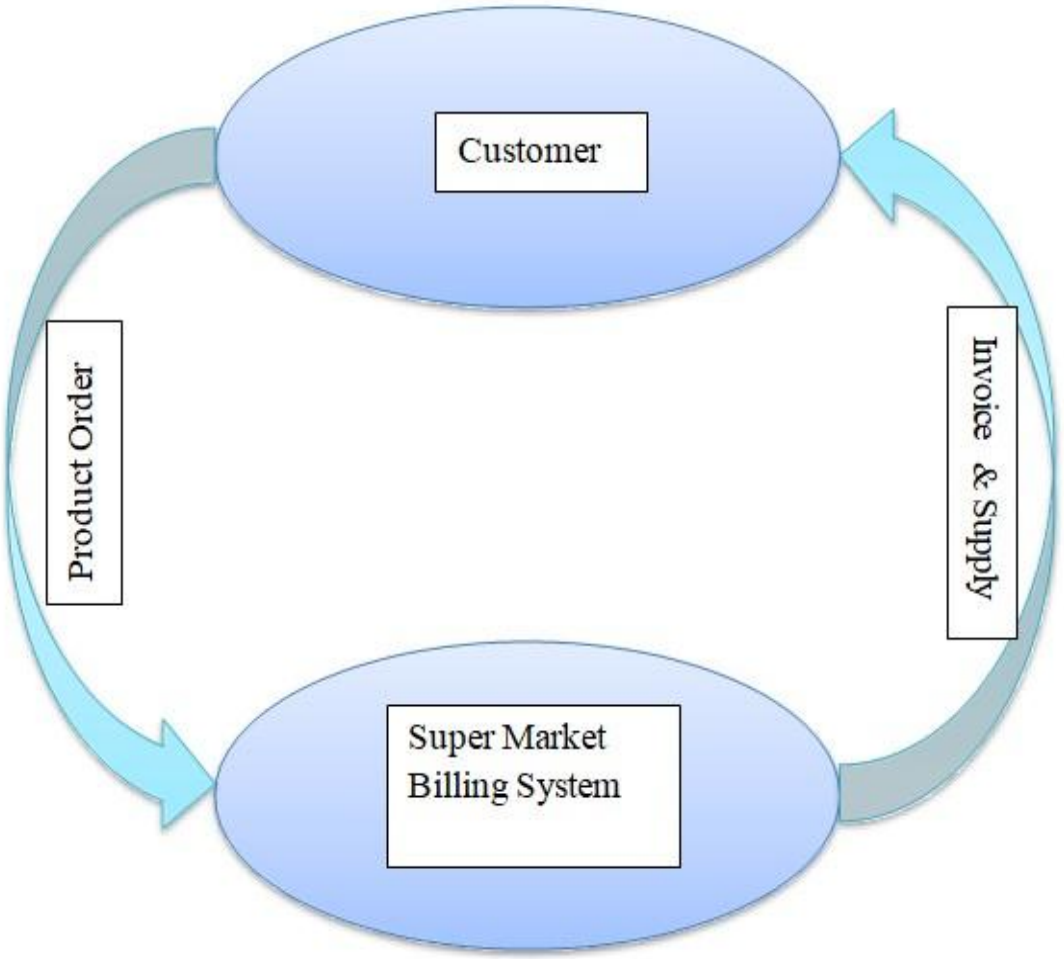


FIG. 5.1.1 LEVEL 0 DFD

5.1.2 LEVEL 1DFD

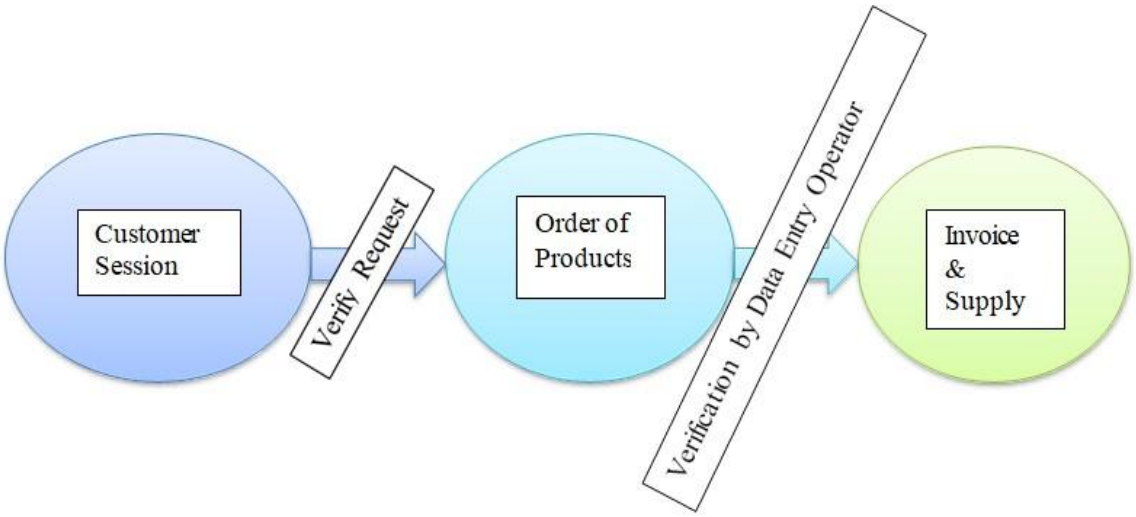


FIG. 5.1.2 LEVEL 1 DFD

5.1.3 LEVEL 2DFD

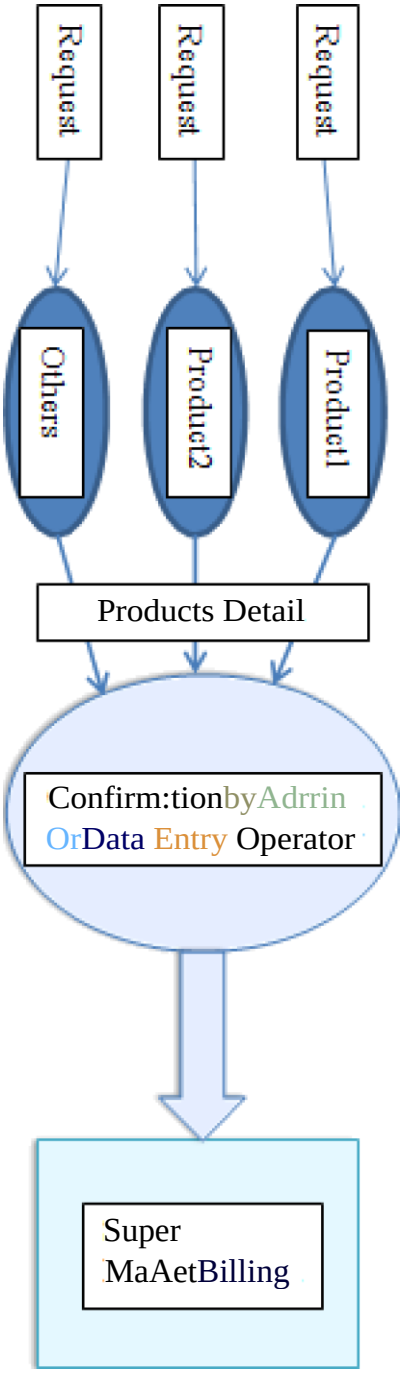


FIG. 5.1.3 LEVEL 2 DFD

5.2 E-RDIAGRAMS

ER Diagram stands for Entity Relationship Diagram, also known as ERD is a diagram that displays the relationship of entity sets stored in a database. In other words, ER diagrams help to explain the logical structure of databases. ER diagrams are created based on three basic concepts: entities, attributes and relationships. ER Diagrams contain different symbols that use rectangles to represent entities, ovals to define attributes and diamond shapes to represent relationships.

The primary reasons for using ER diagrams- Helps you to define terms related to entity relationship modeling Provide a preview of how all your tables should connect, what fields are going to be on each table. Helps to describe entities, attributes, relationships. ER diagrams are translatable into relational tables which allows you to build databases quickly. ER diagrams can be used by database designers as a blueprint for implementing data in specific software applications. The database designer gains a better understanding of the information to be contained in the database with the help of ERP diagram. ERD Diagram allows you to communicate with the logical structure of the database tousers

Entity Relationship Diagram Symbols & Notations:

It mainly contains three basic symbols which are rectangle, oval and diamond to represent relationships between elements, entities and attributes. There are some sub-elements which are based on main elements in ERD Diagram. ER Diagram is a visual representation of data that describes how data is related to each other using different ERD Symbols and Notations.

Following are the main components and its symbols in ER Diagrams:

- **Rectangles:** This Entity Relationship Diagram symbol represents entitytypes
- **Ellipses :**Symbol representattributes
- **Diamonds:** This symbol represents relationshiptypes

- **Lines:** It links attributes to entity types and entity types with other relationship types
- **Primary key:** attributes areunderlined
- **Double Ellipses:** Represent multi-valuedattributes

Components of ER diagrams:

This model is based on three basic components:

- Entities
- Attributes
- Relationships

GOALS:

- Systematically analyze data requirements to produce a well-designeddatabase.
- Representing real-world entities and the relationships betweenthem.
- Creating an ER Model in DBMS is considered as a best practice before implementing yourdatabase.
- Analyze data requirements systematically to produce a well-designed database. So, it is considered a best practice to complete ER modeling before implementing yourdatabase.

5.2.1 E-R DIAGRAM

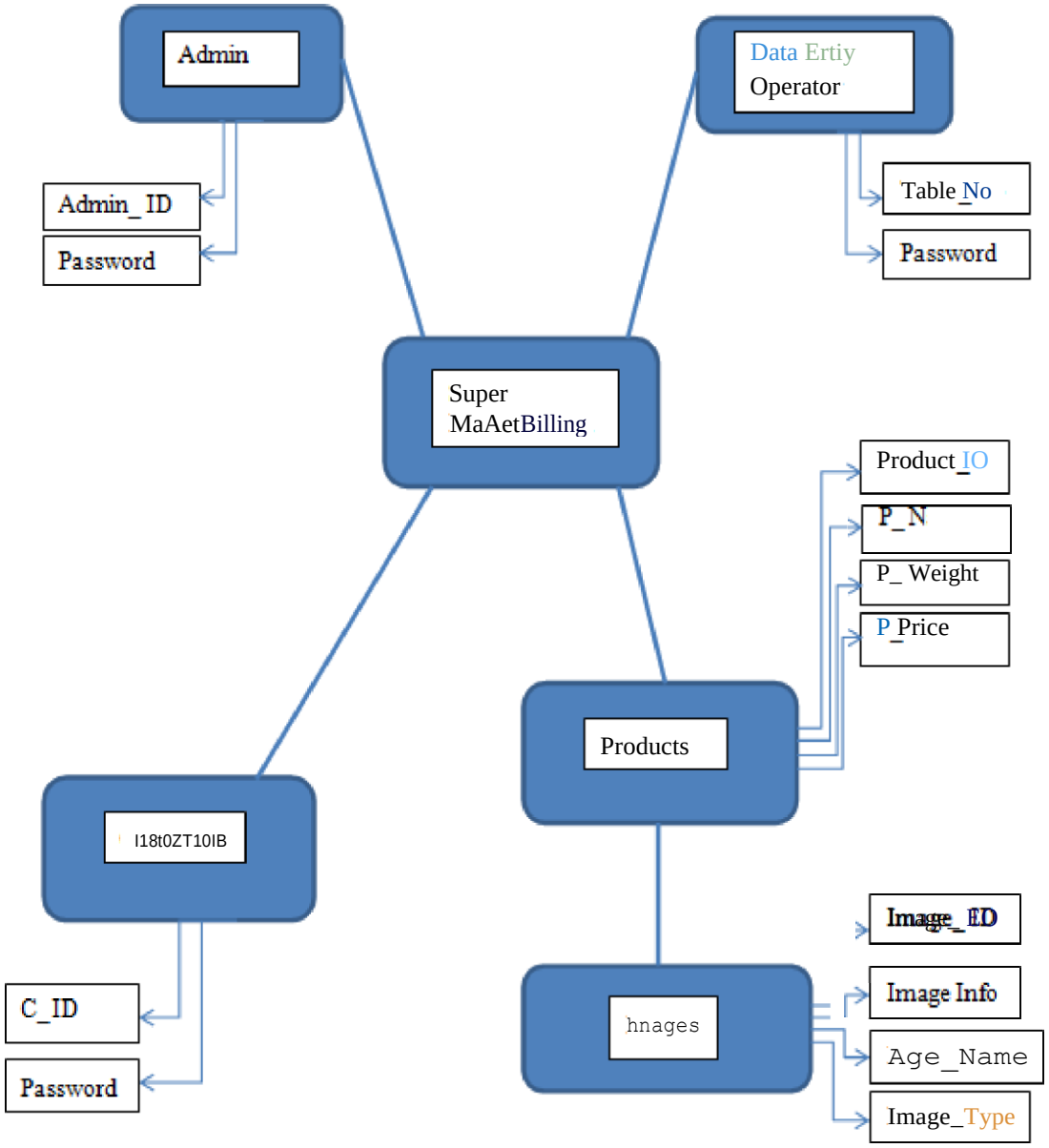


FIG. 5.2.1 E-R DIAGRAM

5.3 UMLDIAGRAMS

UML stands for Unified Modeling Language. UML is a standardized general-purpose modeling language in the field of object-oriented software engineering. The standard is managed, and was created by, the Object ManagementGroup.

The goal is for UML to become a common language for creating models of object oriented computer software. In its current form UML is comprised of two major components: a Meta-model and a notation. In the future, some form of method or process may also be added to; or associated with, UML.

The Unified Modeling Language is a standard language for specifying, Visualization, Constructing and documenting the artifacts of software system, as well as for business modeling and other non-software systems.

The UML represents a collection of best engineering practices that have proven successful in the modeling of large and complex systems.

The UML is a very important part of developing objects oriented software and the software development process. The UML uses mostly graphical notations to express the design of software projects.

GOALS:

- Provide users a ready-to-use, expressive visual modeling Language so that they can develop and exchange meaningfulmodels.
- Provide extendibility and specialization mechanisms to extend the coreconcepts.
- Be independent of particular programming languages and developmentprocess.
- Provide a formal basis for understanding the modelinglanguage.
- Encourage the growth of OO toolsmarket.
- Support higher level development concepts such as collaborations, frameworks, patterns andcomponents.
- Integrate bestpractices.

5.3.1 USE CASEDIAGRAM

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.

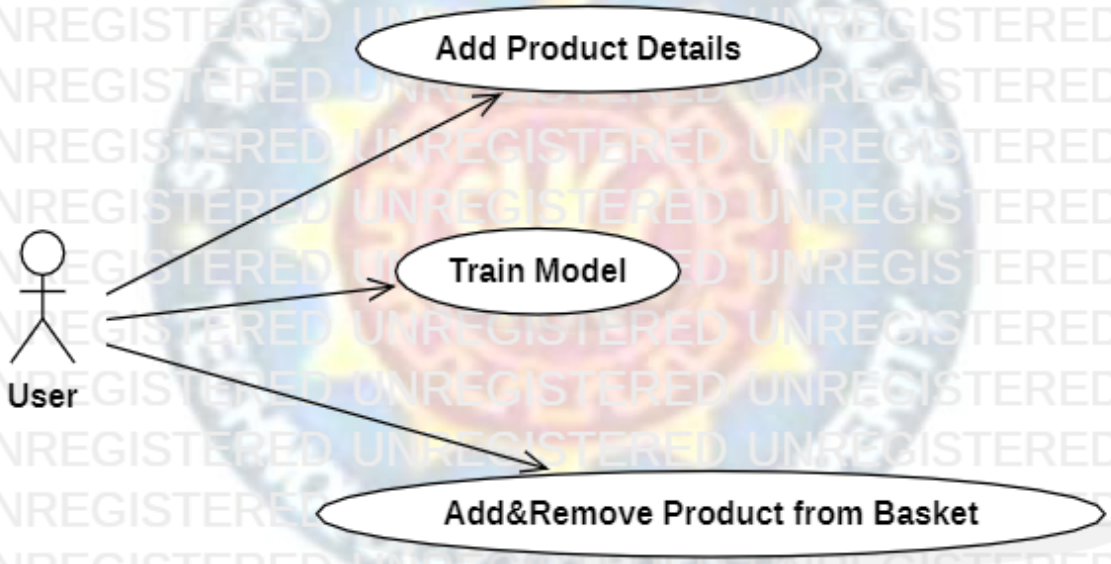


FIG. 5.3.1 USE CASE DIAGRAM

5.3.2 CLASSDIAGRAM

In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.

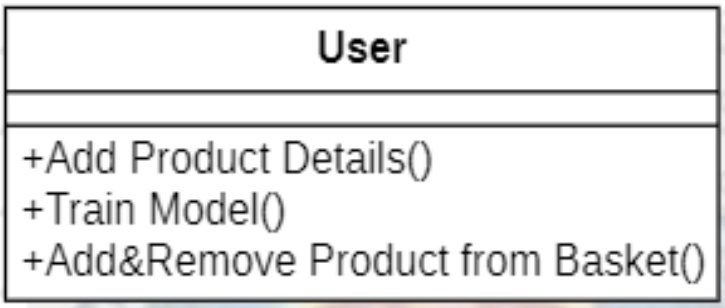


FIG. 5.3.2 CLASS DIAGRAM

5.3.3 SEQUENCEDIAGRAM

A sequence diagram in Unified Modeling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.

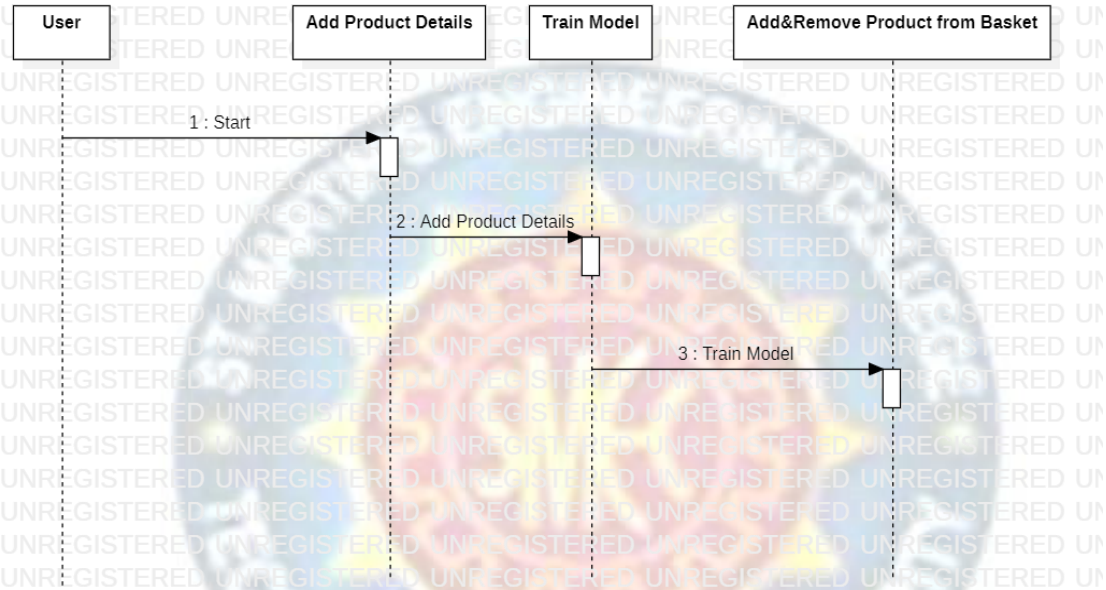


FIG. 5.3.3 SEQUENCE DIAGRAM

5.3.4 ACTIVITYDIAGRAM

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modeling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.

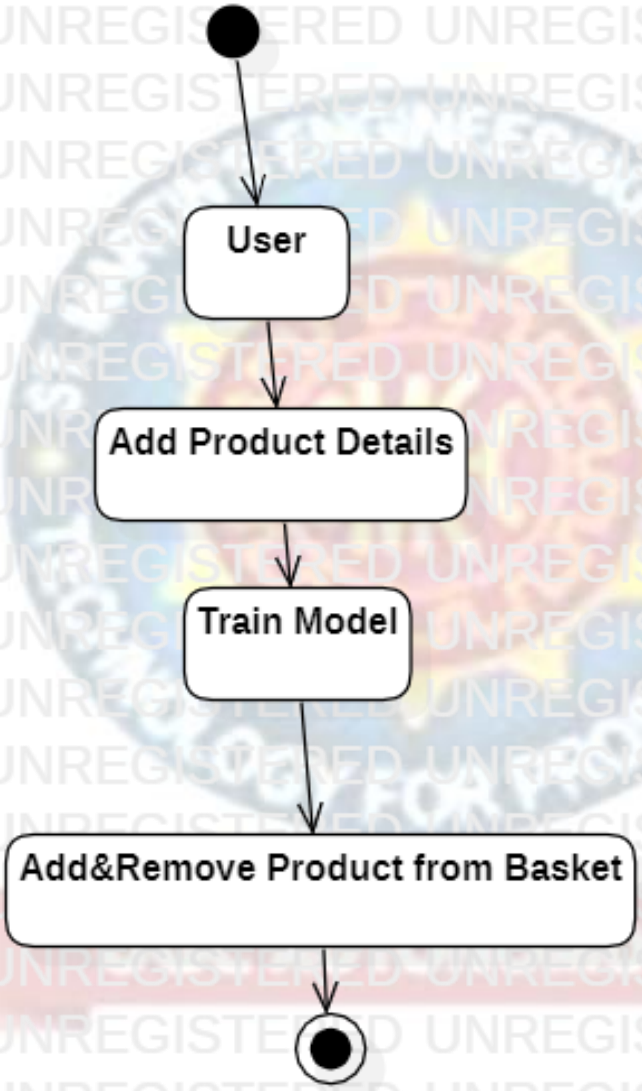


FIG. 5.3.4 ACTIVITY DIAGARAM

6. PROJECTCODING

6.1 TECHNOLOGIES

PYTHON

Python is a general-purpose interpreted, interactive, object-oriented, and high-level programming language. An interpreted language, Python has a design philosophy that emphasizes code readability (notably using whitespace indentation to delimit codeblocks rather than curly brackets or keywords), and a syntax that allows programmers to express concepts in fewer lines of code than might be used in languages such as C++or Java. It provides constructs that enable clear programming on both small and large scales. Python interpreters are available for many operating systems. CPython, the reference implementation of Python, is open source software and has a community-based development model, as do nearly all of its variant implementations. C Python is managed by the non-profit Python Software Foundation. Python features a dynamic type system and automatic memorymanagement.

Python is currently the most widely used multi-purpose, high-level programming language.

Python allows programming in Object-Oriented and Procedural paradigms. Python programs generally are smaller than other programming languages like Java.

Programmers have to type relatively less and indentation requirement of the language, makes them readable all the time.

Python language is being used by almost all tech-giant companies like – Google, Amazon, Facebook, Instagram, Dropbox, Uber… etc.

The biggest strength of Python is huge collection of standard library which can be used for thefollowing:

- Machine Learning
- GUI Applications (like Kivy, Tkinter, PyQtetc.)
- Web frameworks like Django (used by YouTube, Instagram,Dropbox)
- Image processing (like Opencv,Pillow)
- Web scraping (like Scrapy, BeautifulSoup,Selenium)
- Testframeworks
- Multimedia

Advantages of Python:

Let's see how Python dominates over other languages.

1. Extensive Libraries

Python downloads with an extensive library and it contain code for various purposes like regular expressions, documentation-generation, unit-testing, web browsers, threading, databases, CGI, email, image manipulation, and more. So, we don't have to write the complete code for that manually.

2. Extensible

As we have seen earlier, Python can be extended to other languages. You can write some of your code in languages like C++ or C. This comes in handy, especially in projects.

3. Embeddable

Complimentary to extensibility, Python is embeddable as well. You can put your Python code in your source code of a different language, like C++. This lets us add scripting capabilities to our code in the otherlanguage.

4. ImprovedProductivity

The language's simplicity and extensive libraries render programmers more productive than languages like Java and C++ do. Also, the fact that you need to write less and get more thingsdone.

5. IOTOpportunities

Since Python forms the basis of new platforms like Raspberry Pi, it finds the future bright for the Internet Of Things. This is a way to connect the language with the real world.

6. Simple andEasy

When working with Java, you may have to create a class to print 'Hello World'. But inPython, just a print statement will do. It is also quite easy to learn, understand, and code. This is why when people pick up Python, they have a hard time adjusting to other more verbose languages likeJava.

7. Readable

Because it is not such a verbose language, reading Python is much like reading English. This is the reason why it is so easy to learn, understand, and code. It alsodoes not need curly braces to define blocks, and indentation is mandatory. This further aids the readability of thecode.

8. Object-Oriented

This language supports both the procedural and object-oriented programming paradigms. While functions help us with code reusability, classes and objects let us model the real world. A class allows the encapsulation of data and functions into one.

9. Free andOpen-Source

Like we said earlier, Python is freely available. But not only can you downloadPython for free, but you can also download its source code, make changes to it, and even distribute it. It downloads with an extensive collection of libraries to help you with your tasks.

10. Portable

When you code your project in a language like C++, you may need to make some changes to it if you want to run it on another platform. But it isn't the same with Python. Here, you need to code only once, and you can run it anywhere. This is

called Write Once Run Anywhere (WORA). However, you need to be careful enough not to include any system-dependent features.

11. Interpreted

Lastly, we will say that it is an interpreted language. Since statements are executed one by one, debugging is easier than in compiled languages.

Advantages of Python Over Other Languages

1. LessCoding

Almost all of the tasks done in Python requires less coding when the same task is done in other languages. Python also has an awesome standard library support, so you don't have to search for any third-party libraries to get your job done. This is the reason that many people suggest learning Python tobeginners.

2. Affordable

Python is free therefore individuals, small companies or big organizations can leverage the free available resources to build applications. Python is popular and widely used so it gives you better community support.

The 2019 Github annual survey showed us that Python has overtaken Java in the most popular programming language category.

3. Python is forEveryone

Python code can run on any machine whether it is Linux, Mac or Windows. Programmers need to learn different languages for different jobs but with Python, you can professionally build web apps, perform data analysis and machine learning, automate things, do web scraping and also build games and powerful visualizations. It is an all-rounder programming language.

Disadvantages of Python

So far, we've seen why Python is a great choice for your project. But if you choose it, you should be aware of its consequences as well. Let's now see the downsides of choosing Python over another language.

1. SpeedLimitations

We have seen that Python code is executed line by line. But since Python is interpreted, it often results in slow execution. This, however, isn't a problem unless speed is a focal point for the project. In other words, unless high speed is a requirement, the benefits offered by Python are enough to distract us from its speed limitations.

2. Weak in Mobile Computing andBrowsers

While it serves as an excellent server-side language, Python is much rarely seen on the client-side. Besides that, it is rarely ever used to implement smartphone-based applications. One such application is calledCarbonnelle.

The reason it is not so famous despite the existence of Brython is that it isn't that secure.

3. DesignRestrictions

As you know, Python is dynamically-typed. This means that you don't need to declare the type of variable while writing the code. It uses duck-typing. But wait, what's that? Well, it just means that if it looks like a duck, it must be a duck. While this is easy on the programmers during coding, it can raise run-time errors.

4. Underdeveloped Database AccessLayers

Compared to more widely used technologies like JDBC (Java DataBase Connectivity) and ODBC (Open DataBase Connectivity), Python's database access layers are a bitunderdeveloped. Consequently, it is less often applied in huge enterprises.

Why Python

- Python works on different platforms (Windows, Mac, Linux, Raspberry Pi,etc).
- Python has a simple syntax similar to the Englishlanguage.
- Python has syntax that allows developers to write programs with fewer lines than some other programminglanguages.

- Python runs on an interpreter system, meaning that code can be executed as soon as it is written. This means that prototyping can be veryquick.
- Python can be treated in a procedural way, an object-orientated way or a functionalway.

Good to know

- The most recent major version of Python is Python 3, which we shall be using in this tutorial. However, Python 2, although not being updated with anything other than security updates, is still quite popular.
- In this tutorial Python will be written in a text editor. It is possible to write Python in an Integrated Development Environment, such as Thonny, Pycharm, Netbeans or Eclipse which are particularly useful when managing larger collections of Pythonfiles.

Python Syntax compared to other programming languages

- Python was designed for readability, and has some similarities to the English language with influence frommathematics.
- Python uses new lines to complete a command, as opposed to other programming languages which often use semicolons orparentheses.
- Python relies on indentation, using whitespace, to define scope; such as the scope of loops, functions and classes. Other programming languages often use curly-brackets for thispurpose.

Python Installation

To check if you have python installed on a Windows PC, search in the start bar for Python or run the following on the Command Line (cmd.exe):

C:\Users\Your Name>python --version

To check if you have python installed on a Linux or Mac, then on linux open the command line or on Mac open the Terminal and type:

python --version

If you find that you do not have python installed on your computer, then you can download it for free from the following website: https://www.python.org/

Python Quickstart

Python is an interpreted programming language, this means that as a developer you write Python (.py) files in a text editor and then put those files into the python interpreter to be executed.

The way to run a python file is like this on the command line:

C:\Users\Your Name>python helloworld.py

Where "helloworld.py" is the name of your python file.

Let's write our first Python file, called helloworld.py, which can be done in any text editor.

helloworld.py

print("Hello, World!")

Simple as that. Save your file. Open your command line, navigate to the directory where you saved your file, and run:

C:\Users\Your Name>python helloworld.py

The output should read:

Hello, World!

Congratulations, you have written and executed your first Python program.

The Python Command Line

To test a short amount of code in python sometimes it is quickest and easiest not to write the code in a file. This is made possible because Python can be run as a command lineitself.

Type the following on the Windows, Mac or Linux command line:

C:\Users\Your Name>python

Or, if the "python" command did not work, you can try "py":

Virtual Environments and Packages

Introduction

Python applications will often use packages and modules that don't come as part of the standard library. Applications will sometimes need a specific version of a library, because the application may require that a particular bug has been fixed or the application may be written using an obsolete version of the library'sinterface.

This means it may not be possible for one Python installation to meet the requirements of every application. If application A needs version 1.0 of a particular module but application B needs version 2.0, then the requirements are in conflict and installing either version 1.0 or 2.0 will leave one application unable torun.

The solution for this problem is to create a virtual environment, a self-contained directory tree that contains a Python installation for a particular version of Python, plus a number of additional packages.

Different applications can then use different virtual environments. To resolve the earlier example of conflicting requirements, application A can have its own virtual environment with version 1.0 installed while application B has another virtual environment with version 2.0. If application B requires a library be upgraded to version 3.0, this will not affect application A's environment.

Creating Virtual Environments

The module used to create and manage virtual environments is called venv. venv will usually install the most recent version of Python that you have available. If you have multiple versions of Python on your system, you can select a specific Python version by running python3 or whichever version you want.

To create a virtual environment, decide upon a directory where you want to place it, and run the venv module as a script with the directory path:

python3 -m venv tutorial-env

This will create the tutorial-env directory if it doesn't exist, and also create directories inside it containing a copy of the Python interpreter, the standard library, and various supporting files.

A common directory location for a virtual environment is .venv. This name keeps the directory typically hidden in your shell and thus out of the way while giving it a name that explains why the directory exists. It also prevents clashing with .env environment variable definition files that some tooling supports.

Once you've created a virtual environment, you may activate it.

On Windows, run:

tutorial-env\Scripts\activate.bat

On Unix or MacOS, run:

source tutorial-env/bin/activate

(This script is written for the bash shell. If you use the csh or fish shells, there are alternate activate.csh and activate.fish scripts you should use instead.)

Activating the virtual environment will change your shell's prompt to show what virtual environment you're using, and modify the environment so that running python will get you that particular version and installation of Python. For example:

$ source ~/envs/tutorial-env/bin/activate

(tutorial-env) $ python

Python 3.5.1 (default, May 6 2016, 10:59:36)

...

>>> import sys

>>>sys.path

['', '/usr/local/lib/python35.zip', ...,

'~/envs/tutorial-env/lib/python3.5/site-packages']

>>>

MACHINE LEARNING

Machine learning is a subfield of artificial intelligence (AI). The goal of machine learning generally is to understand the structure of data and fit that data into models that can be understood and utilized by people. Although machine learning is a field within computer science, it differs from traditional computational approaches. In traditional computing, algorithms are sets of explicitly programmed instructions used by computers to calculate or problem solve. Machine learning algorithms instead allow for computers to train on data inputs and use statistical analysis in order to output values that fall within a specific range. Because of this, machine learning facilitates computers in building models from sample data in order to automate decision-making processes based on data inputs. Any technology user today has benefitted from machine learning. Facial recognition technology allows social media platforms to help users tag and share photos of friends. Optical character recognition (OCR) technology converts images of text into movable type. Recommendation engines, powered by machine learning, suggest what movies or television shows to watch next based on user preferences. Self-driving cars that rely on machine learning to navigate may soon be available to consumers. Machine learning is a continuously developing field. Because of this, there are some considerations to keep in mind as you work with machine learning methodologies, or analyze the impact of machine learning processes. In this tutorial, we'll look into the common machine learning methods of supervised and unsupervised learning, and common algorithmic approaches in machine learning, including the k-nearest neighbor algorithm, decision tree learning, and deep learning. We'll explore which programming languages are most used in machine learning, providing you with some of the positive and negative attributes of each. Additionally, we'll discuss biases that are perpetuatedby machine

learning algorithms, and consider what can be kept in mind to prevent these biases when building algorithms.

Machine Learning Methods

In machine learning, tasks are generally classified into broad categories. These categories are based on how learning is received or how feedback on the learning is given to the system developed. Two of the most widely adopted machine learning methods are “Supervised learning” which trains algorithms based on example input and output data that is labeled by humans, and “Unsupervised learning” which provides the algorithm with no labeled data in order to allow it to find structure within its input data. Let’s explore these methods in more detail.

Supervised Learning

In supervised learning, the computer is provided with example inputs that are labeled with their desired outputs. The purpose of this method is for the algorithm to be able to “learn” by comparing its actual output with the “taught” outputs to find errors, and modify the model accordingly. Supervised learning therefore uses patterns to predict label values on additional unlabeled data. For example, with supervised learning, an algorithm may be fed data with images of sharks labeled as fish and images of oceans labeled as water. By being trained on this data, the supervised learning algorithm should be able to later identify unlabeled shark images as fish and unlabeled ocean images as water. A common use case of supervised learning is to use historical data to predict statistically likely future events. It may use historical stock market information to anticipate upcoming fluctuations, or be employed to filter out spam emails. In supervised learning, tagged photos of dogs can be used as input data to classify untagged photos ofdogs.

UnsupervisedLearning

In unsupervised learning, data is unlabeled, so the learning algorithm is left to find commonalities among its input data. As unlabeled data are more abundant than labeled data, machine learning methods that facilitate unsupervised learning are particularlyvaluable.

The goal of unsupervised learning may be as straightforward as discovering hidden patterns within a dataset, but it may also have a goal of feature learning, which allows the computational machine to automatically discover the representations that are needed to classify raw data.

Unsupervised learning is commonly used for transactional data. You may have a large dataset of customers and their purchases, but as a human you will likely not be able to make sense of what similar attributes can be drawn from customer profiles and their types of purchases. With this data fed into an unsupervised learning algorithm, it may be determined that women of a certain age range who buy unscented soaps are likely to be pregnant, and therefore a marketing campaign related to pregnancy and baby products can be targeted to this audience in order to increase their number of purchases. Without being told a "correct" answer, unsupervised learning methods can look at complex data that is more expansive and seemingly unrelated in order to organize it in potentially meaningful ways. Unsupervised learning is often used for anomaly detection including for fraudulent credit card purchases, and recommender systems that recommend what products to buy next. In unsupervised learning, untagged photos of dogs can be used as input data for the algorithm to find likenesses and classify dog photos together.

Challenges in Machines Learning

While Machine Learning is rapidly evolving, making significant strides with cybersecurity and autonomous cars, this segment of AI as whole still has a long way to go. The reason behind is that ML has not been able to overcome number of challenges. The challenges that ML is facing currentlyare:

Quality of data: Having good-quality data for ML algorithms is one of the biggest challenges. Use of low-quality data leads to the problems related to data preprocessing and featureextraction.

Time-Consuming task: Another challenge faced by ML models is the consumption of time especially for data acquisition, feature extraction andretrieval.

Lack of specialist persons: As ML technology is still in its infancy stage, availability of expert resources is a tough job.

No clear objective for formulating business problems: Having no clear objective and well-defined goal for business problems is another key challenge for ML because this technology is not that mature yet.

Issue of Over-fitting & Under-fitting: If the model is over-fitting or under-fitting, it cannot be represented well for the problem.

Curse of dimensionality: Another challenge ML model faces is too many features of data points. This can be a real hindrance.

Difficulty in deployment: Complexity of the ML model makes it quite difficult to be deployed in real life.

Applications of Machines Learning

Machine Learning is the most rapidly growing technology and according to researchers we are in the golden year of AI and ML. It is used to solve many real-world complex problems which cannot be solved with traditional approach. Following are some real-world applications ofML:

- Emotionanalysis
- Sentimentanalysis
- Error detection andprevention
- Weather forecasting andprediction
- Stock market analysis andforecasting
- Speechsynthesis
- Speechrecognition
- Customersegmentation

- Objectrecognition
- Frauddetection
- Fraudprevention
- Recommendation of products to customer in online shopping

OpenCV

OpenCV (Open Source Computer Vision Library) is an open source computer vision and machine learning software library. OpenCV was built to provide a common infrastructure for computer vision applications and to accelerate the use of machine perception in the commercial products. The library has more than 2500 optimized algorithms, which includes a comprehensive set of both classic and state-of-the-art computer vision and machine learning algorithms. These algorithms can be used to detect and recognize faces, identify objects, classify human actions in videos, track camera movements, track moving objects, extract 3D models of objects, produce 3D point clouds from stereo cameras, stitch images together to produce a high resolution image of an entire scene.

Features of OpenCV Library

Using OpenCV library, you can-

- Read and writeimages
- Capture and savevideos
- Process images (filter,transform)
- Perform featuredetection
- Detect specific objects such as faces, eyes, cars, in the videos orimages.
- Analyze the video, i.e., estimate the motion in it, subtract the background, and track objects init.

OpenCV was originally developed in C++. In addition to it, Python and Java bindings were provided. OpenCV runs on various Operating Systems such as windows, Linux, OSx, FreeBSD, Net BSD, Open BSD, etc.

OpenCV Library Modules

Following are the main library modules of the OpenCV library.

Core Functionality

This module covers the basic data structures such as Scalar, Point, Range, etc., that are used to build OpenCV applications. In addition to these, it also includes the multidimensional array Mat, which is used to store the images. In the Java library of OpenCV, this module is included as a package with the nameorg.opencv.core.

Image Processing

This module covers various image processing operations such as image filtering, geometrical image transformations, color space conversion, histograms, etc. In the Java library of OpenCV, this module is included as a package with the name org.opencv.imgproc.

Video

This module covers the video analysis concepts such as motion estimation, background subtraction, and object tracking. In the Java library of OpenCV, this module is included as a package with the name org.opencv.video.

Video I/O

This module explains the video capturing and video codecs using OpenCV library. In the Java library of OpenCV, this module is included as a package with the name org.opencv.videoio.

calib3d

This module includes algorithms regarding basic multiple-view geometry algorithms, single and stereo camera calibration, object pose estimation, stereo correspondence and elements of 3D reconstruction. In the Java library of OpenCV, this module is included as a package with the name org.opencv.calib3d.

features2d

This module includes the concepts of feature detection and description. In the Java library of OpenCV, this module is included as a package with the name org.opencv.features2d.

Objdetect

This module includes the detection of objects and instances of the predefined classes such as faces, eyes, mugs, people, cars, etc. In the Java library of OpenCV, this module is included as a package with the name org.opencv.objdetect.

Highgui

This is an easy-to-use interface with simple UI capabilities. In the Java library of OpenCV, the features of this moduleis included in two different packages namely, org.opencv.imgcodecs andorg.opencv.videoio.

6.2 CODING

SUPERMARKET.PY

```
import tkinter
import cv2
import PIL.Image, PIL.ImageTk
from tkinter importsimpledialog
importtime
from tkinter import messagebox
import os
from keras.utils.np_utils import to_categorical
import numpy as np
from keras.layers import MaxPooling2D
from keras.layers import Dense, Dropout, Activation, Flatten
from keras.layers import Convolution2D
from keras.models import Sequential
from keras.models import model_from_json
```

```
import pickle
from tkinter import *
import random

class App:
    global classifier
    global labels
    global X_train
    global Y_train
    global prices
    global cart
    global text
    global person_id
    global img_canvas
    global cascPath
    global faceCascade
    global pid

    definit(self, window, window_title, video_source=0):
        global cart
        global text
        cart = []
        self.window = window
        self.window.title(window_title)
        self.window.geometry("1300x1200")
        self.video_source = video_source
        self.vid = MyVideoCapture(self.video_source)
        self.canvas = tkinter.Canvas(window, width = self.vid.width, height =
self.vid.height)
        self.canvas.pack()
        self.font1 = ('times', 13, 'bold')
        self.btn_snapshot=tkinter.Button(window, text="Add Product Details",
command=self.snapshot)
        self.btn_snapshot.place(x=10,y=50)
```

```
        self.btn_snapshot.config(font=self.font1)
        self.btn_train=tkinter.Button(window, text="Train Model",
command=self.trainmodel)
        self.btn_train.place(x=10,y=100)
        self.btn_train.config(font=self.font1)
        self.btn_predict=tkinter.Button(window, text="Add/Remove Product from
Basket", command=self.predict)
        self.btn_predict.place(x=10,y=150)
        self.btn_predict.config(font=self.font1)

        self.btn_person=tkinter.Button(window, text="Capture Person",
command=self.capturePerson)
        self.btn_person.place(x=10,y=200)
        self.btn_person.config(font=self.font1)

        self.img_canvas = tkinter.Canvas(window, width = 200, height = 200)
        self.img_canvas.place(x=10,y=250)

        self.text=Text(window,height=35,width=45)
        scroll=Scrollbar(self.text)
        self.text.configure(yscrollcommand=scroll.set)
        self.text.place(x=1000,y=50)
        self.text.config(font=self.font1)

        self.cascPath = "haarcascade_frontalface_default.xml"
        self.faceCascade = cv2.CascadeClassifier(self.cascPath)

        self.delay = 15
        self.update()
        self.window.mainloop()

    def getID(self,name):
        index = 0
        for i in range(len(labels)):
```

```
            if labels[i] == name:
                index = i
                break
        return index

    def capturePerson(self):
        option = 0
        ret, frame = self.vid.get_frame()
        img = frame

        gray = cv2.cvtColor(frame, cv2.COLOR_BGR2GRAY)
        faces = self.faceCascade.detectMultiScale(gray,1.3,5)
        print("Found {0} faces!".format(len(faces)))
        for (x, y, w, h) in faces:
            cv2.rectangle(frame, (x, y), (x+w, y+h), (0, 255, 0), 2)
            img = frame[y:y + h, x:x + w]
            img = cv2.resize(img,(500,500))
            option = 1
        if option == 1:
            self.pid = random.randint(1000, 100000)
            cv2.imwrite("images/"+str(self.pid)+".jpg",img);
            cv2.imshow("Person ID : "+str(self.pid)+".jpg",img)
            c2.waitKey(0)
        else:
            messagebox.showinfo("Face or person not detected","Face or person not
detected")

    def snapshot(self):
        ret, frame = self.vid.get_frame()
        if ret:
            img = cv2.cvtColor(frame, cv2.COLOR_RGB2BGR)
            pname = simpledialog.askstring("Enter Product Name", "Enter Product
Name",parent=self.window)
```

```
        price = simpledialog.askfloat("Enter Product Price", "Enter Product Price",
parent=self.window, minvalue=1.0, maxvalue=100000.0)
        if not os.path.exists('Product/'+pname):
            os.makedirs('Product/'+pname)
        img_name = time.strftime("%d-%m-%Y-%H-%M-%S") + ".jpg"
        cv2.imwrite('Product/'+pname+'/'+img_name,img)
        f = open("details.txt", "a+")
        f.write(pname+","+str(price)+","+img_name+"\n")
        f.close()
        messagebox.showinfo("Product details saved","Product details saved")

    deftrainmodel(self):
        global labels
        global X_train
        global Y_train
        global classifier
        global prices
        labels = []
        X_train = []
        Y_train = []
        prices =[]
        path = 'Product'
        for root, dirs, directory in os.walk(path):
            for j in range(len(directory)):
                name = os.path.basename(root)
                if name not in labels:
                    labels.append(name)

        for i in range(len(labels)):
            cost = '0'
            with open("details.txt", "r") as file:
                for line in file:
                    line = line.strip('\n')
                    line = line.strip()
```

```
            arr = line.split(",")
            if arr[0] == labels[i] and cost == '0':
                cost =arr[1]
        file.close()
        prices.append(cost)

    for root, dirs, directory in os.walk(path):
        for j in range(len(directory)):
            name = os.path.basename(root)
            img = cv2.imread(root+"/"+directory[j])

            img = cv2.resize(img, (256,256))
            im2arr = np.array(img)
            im2arr = im2arr.reshape(256,256,3)
            X_train.append(im2arr)
            Y_train.append(self.getID(name))
    X_train  =np.asarray(X_train)
    Y_train  =np.asarray(Y_train)
    print(Y_train)
    print(labels)
    print(prices)
    X_train = X_train.astype('float32')
    X_train = X_train/255

    test = X_train[3]
    cv2.imshow("aa",test)
    cv2.waitKey(0)
    indices = np.arange(X_train.shape[0])
    np.random.shuffle(indices)
    X_train = X_train[indices]
    Y_train = Y_train[indices]
    Y_train = to_categorical(Y_train)

    if os.path.exists('Model/model.json'):
```

```
        with open('Model/model.json', "r") as json_file:
            loaded_model_json = json_file.read()
            classifier = model_from_json(loaded_model_json)

        classifier.load_weights("Model/model_weights.h5")
        classifier.make_predict_function()
        print(classifier.summary())
        f = open('Model/history.pckl', 'rb')
        data = pickle.load(f)
        f.close()

        acc = data['accuracy']
        accuracy = acc[9] * 100
        messagebox.showinfo("Training model accuracy","Training Model Accuracy
= "+str(accuracy))
    else:
        classifier = Sequential()
        classifier.add(Convolution2D(32, 3, 3, input_shape = (256, 256, 3), activation
= 'relu'))
        classifier.add(MaxPooling2D(pool_size = (2, 2)))
        classifier.add(Convolution2D(32, 3, 3, activation = 'relu'))
        classifier.add(MaxPooling2D(pool_size = (2, 2)))
        classifier.add(Flatten())
        classifier.add(Dense(output_dim = 256, activation = 'relu'))
        classifier.add(Dense(output_dim = 4, activation = 'softmax'))
        print(classifier.summary())
        classifier.compile(optimizer = 'adam', loss = 'categorical_crossentropy',
metrics = ['accuracy'])
        hist = classifier.fit(X_train, Y_train, batch_size=16, epochs=10, shuffle=True,
verbose=2)
        classifier.save_weights('Model/model_weights.h5')
        model_json = classifier.to_json()
        with open("Model/model.json", "w") as json_file:
            json_file.write(model_json)
```

```
        f = open('Model/history.pckl', 'wb')
        pickle.dump(hist.history, f)
        f.close()
        f = open('Model/history.pckl', 'rb')
        data = pickle.load(f)
        f.close()
        acc = data['accuracy']
        accuracy = acc[9] * 100
        messagebox.showinfo("Training model accuracy","Training Model Accuracy
= "+str(accuracy))

    def predict(self):
        ret, frame = self.vid.get_frame()
        img = cv2.cvtColor(frame, cv2.COLOR_RGB2BGR)
        img = cv2.resize(img, (256,256))
        im2arr = np.array(img)
        im2arr = im2arr.reshape(1,256,256,3)
        image = np.asarray(im2arr)
        image = image.astype('float32')
        image = image/255
        preds = classifier.predict(image)
        predict = np.argmax(preds)
        pname = labels[predict]
        print(str(pname)+" "+str(np.amax(preds)))
        if np.amax(preds) >= 0.85:
            cost = prices[predict]
            if pname in cart:
                cart.remove(pname)
            else:
                cart.append(pname)
            self.text.delete('1.0', END)
            total_amt = 0
            for i in range(len(cart)):
                for k in range(len(labels)):
```

```
                if labels[k] == cart[i]:
                    cost = prices[k]
                    k = len(labels)
            total_amt = total_amt + float(cost)
            self.text.insert(END,"Product Name : "+cart[i]+"\n")
            self.text.insert(END,"Product Cost : "+cost+"\n\n")
        self.text.insert(END,"Total Amount :"+str(total_amt)+"\n\n")
    else:
        messagebox.showinfo("Unable to recognized product","Unable torecognized
product")



  def update(self):
    ret, frame = self.vid.get_frame()
    if ret:
      self.photo = PIL.ImageTk.PhotoImage(image = PIL.Image.fromarray(frame))
      self.canvas.create_image(0, 0, image = self.photo, anchor = tkinter.NW)
      self.window.after(self.delay, self.update)



class MyVideoCapture:
  definit(self, video_source=0):

    self.vid = cv2.VideoCapture(video_source)
    if not self.vid.isOpened():
      raise ValueError("Unable to open video source", video_source)
    self.width = self.vid.get(cv2.CAP_PROP_FRAME_WIDTH)
    self.height = self.vid.get(cv2.CAP_PROP_FRAME_HEIGHT)
    self.pid = 0

  def get_frame(self):
    if self.vid.isOpened():
      ret, frame = self.vid.read()
      if ret:
```

```
            return (ret, cv2.cvtColor(frame, cv2.COLOR_BGR2RGB))
        else:
            return (ret, None)
    else:
        return (ret, None)

defdel(self):
    if self.vid.isOpened():
        self.vid.release()
App(tkinter.Tk(), "Tkinter and OpenCV")
```

TEST.PY

```
import tkinter
import cv2
import PIL.Image, PIL.ImageTk
from tkinter importsimpledialog
importtime
from tkinter import messagebox
import os
from keras.utils.np_utils import to_categorical
import numpy as np
from keras.layers importMaxPooling2D
from keras.layers import Dense, Dropout, Activation, Flatten
from keras.layers import Convolution2D
from keras.models import Sequential
from keras.models import model_from_json
import pickle
from tkinter import *
class App:
    global classifier
    global labels
    global X_train
    global Y_train
```

```
    global prices
    global cart
    global text

    definit(self, window, window_title, video_source=0):
        global cart
        global text
        cart = []
        self.window = window
        self.window.title(window_title)
        self.window.geometry("1300x1200")
        self.video_source = video_source
        self.vid = MyVideoCapture(self.video_source)
        self.canvas = tkinter.Canvas(window, width = self.vid.width, height =
self.vid.height)
        self.canvas.pack()
        self.font1 = ('times', 13, 'bold')
        self.btn_snapshot=tkinter.Button(window, text="Add Product Details",
command=self.snapshot)
        self.btn_snapshot.place(x=10,y=50)
        self.btn_snapshot.config(font=self.font1)
        self.btn_train=tkinter.Button(window, text="Train Model",
command=self.trainmodel)
        self.btn_train.place(x=10,y=100)
        self.btn_train.config(font=self.font1)
        self.btn_predict=tkinter.Button(window, text="Add/Remove Product from
Basket", command=self.predict)
        self.btn_predict.place(x=10,y=150)
        self.btn_predict.config(font=self.font1)

        self.text=Text(window,height=35,width=45)
        scroll=Scrollbar(self.text)
        self.text.configure(yscrollcommand=scroll.set)
        self.text.place(x=1000,y=50)
```

```
        self.text.config(font=self.font1)

        self.delay = 15
        self.update()
        self.window.mainloop()

    def getID(self,name):
        index = 0
        for i in range(len(labels)):
            if labels[i] == name:
                index = i
                break
        return index

    def snapshot(self):
        ret, frame = self.vid.get_frame()
        if ret:
            img = cv2.cvtColor(frame, cv2.COLOR_RGB2BGR)
            pname = simpledialog.askstring("Enter Product Name", "Enter Product
Name",parent=self.window)
            price = simpledialog.askfloat("Enter Product Price", "Enter Product Price",
parent=self.window, minvalue=1.0, maxvalue=100000.0)
            if not os.path.exists('Product/'+pname):
                os.makedirs('Product/'+pname)
            img_name = time.strftime("%d-%m-%Y-%H-%M-%S") + ".jpg"
            cv2.imwrite('Product/'+pname+'/'+img_name,img)
            f = open("details.txt", "a+")
            f.write(pname+","+str(price)+","+img_name+"\n")
            f.close()
            messagebox.showinfo("Product details saved","Product details saved")

    def trainmodel(self):
        global labels
        global X_train
```

```
global Y_train
global classifier
global prices
labels = []
X_train = []
Y_train = []
prices = []
path = 'Product'
for root, dirs, directory in os.walk(path):
    for j in range(len(directory)):
        name = os.path.basename(root)
        if name not in labels:
            labels.append(name)

for i in range(len(labels)):
    cost = '0'
    with open("details.txt", "r") as file:
        for line in file:
            line = line.strip('\n')
            line = line.strip()
            arr =line.split(",")
            if arr[0] == labels[i] and cost == '0':
                cost =arr[1]
    file.close()
    prices.append(cost)

for root, dirs, directory in os.walk(path):
    for j in range(len(directory)):
        name = os.path.basename(root)
        img= cv2.imread(root+"/"+directory[j])
        img= cv2.resize(img, (256,256)) im2arr
        =np.array(img)
        im2arr = im2arr.reshape(256,256,3)
        X_train.append(im2arr)
```

```
            Y_train.append(self.getID(name))
        X_train = np.asarray(X_train)
        Y_train = np.asarray(Y_train)
        print(Y_train)
        print(labels)
        print(prices)
        X_train = X_train.astype('float32')
        X_train = X_train/255

        test = X_train[3]
        cv2.imshow("aa",test)
        cv2.waitKey(0)
        indices = np.arange(X_train.shape[0])
        np.random.shuffle(indices)
        X_train = X_train[indices]
        Y_train = Y_train[indices]
        Y_train = to_categorical(Y_train)

        if os.path.exists('Model/model.json'):
            with open('Model/model.json', "r") as json_file:
                loaded_model_json = json_file.read()
                classifier = model_from_json(loaded_model_json)

            classifier.load_weights("Model/model_weights.h5")
            classifier._make_predict_function()
            print(classifier.summary())
            f = open('Model/history.pckl', 'rb')
            data = pickle.load(f)
            f.close()
            acc = data['accuracy']
            accuracy = acc[9] * 100
            messagebox.showinfo("Training model accuracy","Training Model Accuracy
= "+str(accuracy))
        else:
```

```
        classifier = Sequential()
        classifier.add(Convolution2D(32, 3, 3, input_shape = (256, 256, 3), activation
= 'relu'))
        classifier.add(MaxPooling2D(pool_size = (2, 2)))
        classifier.add(Convolution2D(32, 3, 3, activation = 'relu'))
        classifier.add(MaxPooling2D(pool_size = (2, 2)))
        classifier.add(Flatten())
        classifier.add(Dense(output_dim = 256, activation = 'relu'))
        classifier.add(Dense(output_dim = 4, activation = 'softmax'))
        print(classifier.summary())
        classifier.compile(optimizer = 'adam', loss = 'categorical_crossentropy',
metrics = ['accuracy'])
        hist = classifier.fit(X_train, Y_train, batch_size=16, epochs=10, shuffle=True,
verbose=2)
        classifier.save_weights('Model/model_weights.h5')
        model_json = classifier.to_json()
        with open("Model/model.json", "w") as json_file:
            json_file.write(model_json)
        f = open('Model/history.pckl', 'wb')
        pickle.dump(hist.history, f)
        f.close()
        f = open('Model/history.pckl', 'rb')
        data = pickle.load(f)
        f.close()
        acc = data['accuracy']
        accuracy = acc[9] * 100
        messagebox.showinfo("Training model accuracy","Training Model Accuracy
= "+str(accuracy))


    def predict(self):
        ret, frame = self.vid.get_frame()
        img = cv2.cvtColor(frame, cv2.COLOR_RGB2BGR)
        img = cv2.resize(img, (256,256))
```

```
        im2arr = np.array(img)
        im2arr = im2arr.reshape(1,256,256,3)
        image = np.asarray(im2arr)
        image = image.astype('float32')
        image = image/255
        preds = classifier.predict(image)
        predict = np.argmax(preds)
        pname = labels[predict]
        print(str(pname)+" "+str(np.amax(preds)))
        cost = prices[predict]
        if pname in cart:
            cart.remove(pname)
        else:
            cart.append(pname)
        self.text.delete('1.0', END)
        total_amt = 0
        for i in range(len(cart)):
            for k in range(len(labels)):
                if labels[k] == cart[i]:
                    cost = prices[k]
                    total_amt = total_amt + float(cost)
                    k = len(labels)
            self.text.insert(END,"Product Name : "+cart[i]+"\n")
            self.text.insert(END,"Product Cost : "+cost+"\n\n")
        self.text.insert(END,"\nTotal Amount : "+str(total_amt))
    def update(self):
        ret, frame = self.vid.get_frame()
        if ret:
            self.photo = PIL.ImageTk.PhotoImage(image = PIL.Image.fromarray(frame))
            self.canvas.create_image(0, 0, image = self.photo, anchor = tkinter.NW)
            self.window.after(self.delay, self.update)


class MyVideoCapture:
```

```
    definit(self, video_source=0):
        self.vid = cv2.VideoCapture(video_source)
        if not self.vid.isOpened():
            raise ValueError("Unable to open video source", video_source)
        self.width = self.vid.get(cv2.CAP_PROP_FRAME_WIDTH)
        self.height = self.vid.get(cv2.CAP_PROP_FRAME_HEIGHT)

    def get_frame(self):
        if self.vid.isOpened():
            ret, frame = self.vid.read()
            if ret:
                return (ret, cv2.cvtColor(frame, cv2.COLOR_BGR2RGB))
            else:
                return (ret, None)
        else:
            return (ret, None)

    defdel(self):
        if self.vid.isOpened():
            self.vid.release()
App(tkinter.Tk(), "Tkinter and OpenCV")
```

TEST1.PY

```
from keras.utils.np_utils import to_categorical
import os
import cv2
import numpy as np
from keras.layers importMaxPooling2D
from keras.layers import Dense, Dropout, Activation, Flatten
from keras.layers import Convolution2D
from keras.models import Sequential
from keras.models import model_from_json
import pickle
```

```
path = 'Product'
labels = []
X_train = []
Y_train =[]
def getID(name):
    index =0
    for i in range(len(labels)):
        if labels[i] == name:
            index = i
            break
    return index

for root, dirs, directory in os.walk(path):
    for j in range(len(directory)):
        name = os.path.basename(root)
        if name not in labels:
            labels.append(name)

for root, dirs, directory in os.walk(path):
    for j in range(len(directory)):
        name = os.path.basename(root)
        img= cv2.imread(root+"/"+directory[j])
        img= cv2.resize(img, (256,256)) im2arr
        =np.array(img)
        im2arr = im2arr.reshape(256,256,3)
        X_train.append(im2arr)
        Y_train.append(getID(name))
X_train  =np.asarray(X_train)
Y_train  =np.asarray(Y_train)
print(Y_train)

X_train = X_train.astype('float32')
X_train = X_train/255
```

```
test = X_train[3]
cv2.imshow("aa",test)
cv2.waitKey(0)
indices = np.arange(X_train.shape[0])

np.random.shuffle(indices)
X_train = X_train[indices]
Y_train = Y_train[indices]
Y_train = to_categorical(Y_train)



classifier = Sequential()
classifier.add(Convolution2D(32, 3, 3, input_shape = (256, 256, 3), activation =
'relu'))
classifier.add(MaxPooling2D(pool_size = (2, 2)))
classifier.add(Convolution2D(32, 3, 3, activation = 'relu'))
classifier.add(MaxPooling2D(pool_size = (2, 2)))
classifier.add(Flatten())
classifier.add(Dense(output_dim = 256, activation = 'relu'))
classifier.add(Dense(output_dim = 4, activation = 'softmax'))
print(classifier.summary())
classifier.compile(optimizer = 'adam', loss = 'categorical_crossentropy', metrics =
['accuracy'])
hist = classifier.fit(X_train, Y_train, batch_size=16, epochs=10, shuffle=True,
verbose=2)
classifier.save_weights('Model/model_weights.h5')
model_json = classifier.to_json()
with open("Model/model.json", "w") as json_file:
    json_file.write(model_json)
f = open('Model/history.pckl', 'wb')
pickle.dump(hist.history, f)
f.close()
f = open('Model/history.pckl', 'rb')
data = pickle.load(f)
```

```
f.close()
acc = data['accuracy']
accuracy = acc[9] * 100
print(accuracy)
```

6.3 METHODS

The following methods are used in the project code:

TKinter: Tkinter is a standard GUI (graphical user interface) package. Tkinter is Python's default GUI module and also the most common way that is used for GUI programming in Python.

Matplotlib: Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms.

Numpy: Numpy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays.

TensorFlow: TensorFlow is a free and open-source software library for dataflow anddifferentiable programming across a range of tasks.

7. PROJECTTESTING

The purpose of testing is to discover errors. Testing is the process of trying todiscover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testingrequirement.

7.1 VARIOUS TESTCASES

Test Case1:

We upload the images of the products and add the details of the products

Test Case 2:

We remove the products from the basket

Unit testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

Integration testing

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

Functional test

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.
Functional testing is centered on the following items:

Valid Input - identified classes of valid input must be accepted.
Invalid Input - identified classes of invalid input must be rejected.
Functions - identified functions must be exercised.
Output - identified classes of application outputsmustbe exercised.
Systems/Procedures - interfacing systems or procedures must beinvoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

System Test

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration oriented system integration test. System testing is based on

process descriptions and flows, emphasizing pre-driven process links and integration points.

7.2 WHITE BOXTESTING

White Box Testing is a testing in which in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is purpose. It is used to test areas that cannot be reached from a black box level.

7.3 BLACK BOXTESTING

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box .you cannot “see” into it. The test provides inputs and responds to outputs without considering how the software works.

Unit Testing

Unit testing is usually conducted as part of a combined code and unit test phase of the software lifecycle, although it is not uncommon for coding and unit testing to be conducted as two distinct phases.

Test strategy and approach

Field testing will be performed manually and functional tests will be written in detail.

Test objectives

- All field entries must workproperly.
- Pages must be activated from the identifiedlink.
- The entry screen, messages and responses must not bedelayed.

Features to be tested

- Verify that the entries are of the correctformat
- No duplicate entries should beallowed
- All links should take the user to the correctpage.

Integration Testing

Software integration testing is the incremental integration testing of two or more integrated software components on a single platform to produce failures caused by interface defects.

The task of the integration test is to check that components or software applications, e.g. components in a software system or – one step up – software applications at the company level – interact without error.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

Acceptance Testing

User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

8. OUTPUTSCREENS

To build supermarket basket project we used some sample products image to train product identification models and below are some products details screenshots.

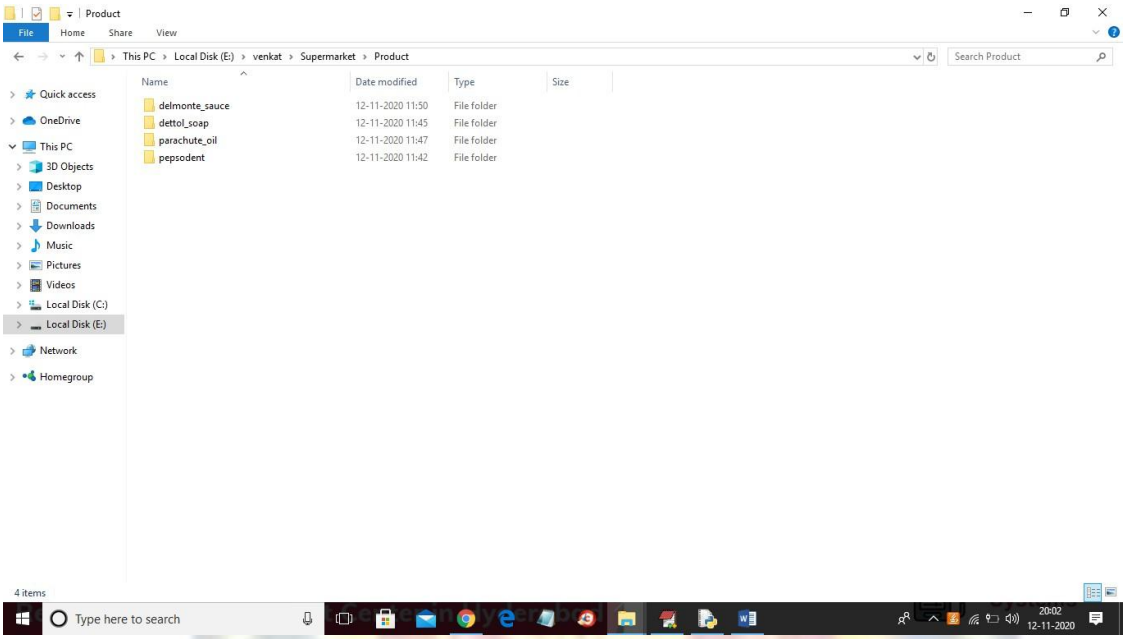


FIG. 8.1 SAMPLE PRODUCT IMAGE 1

In above screen I took 4 products folders and each folder contains images of those products. For example below is the images of Dettol_soap folder

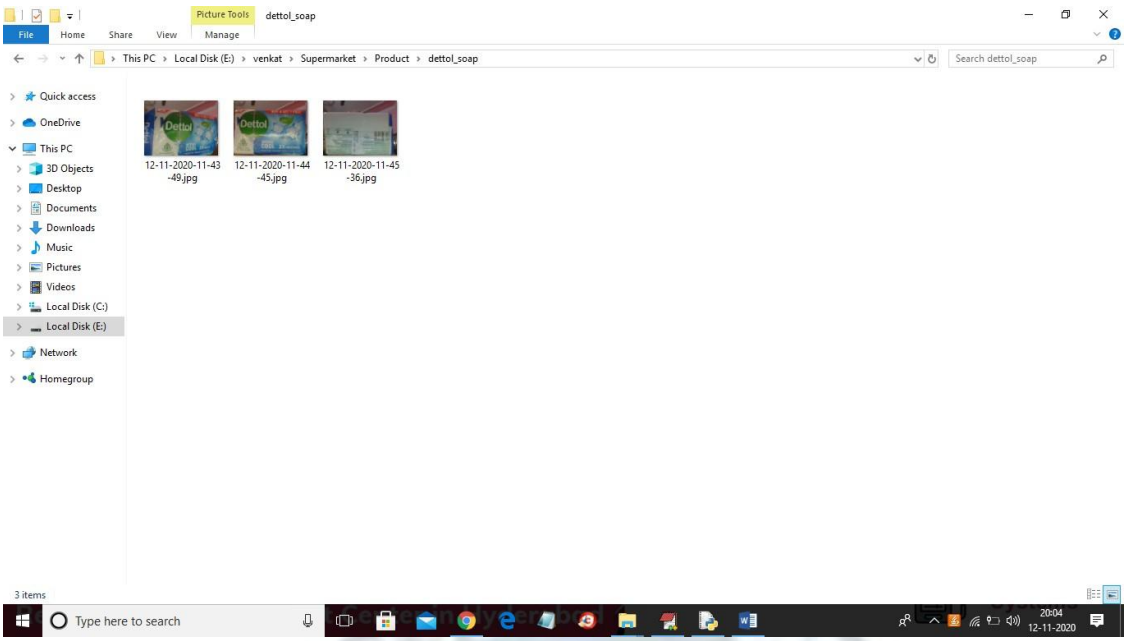


FIG. 8.2 SAMPLE PRODUCT IMAGE 2

In above screens we can see Dettol images and now to identify products run the project by double click on ‘run.bat’ file to get below screen.

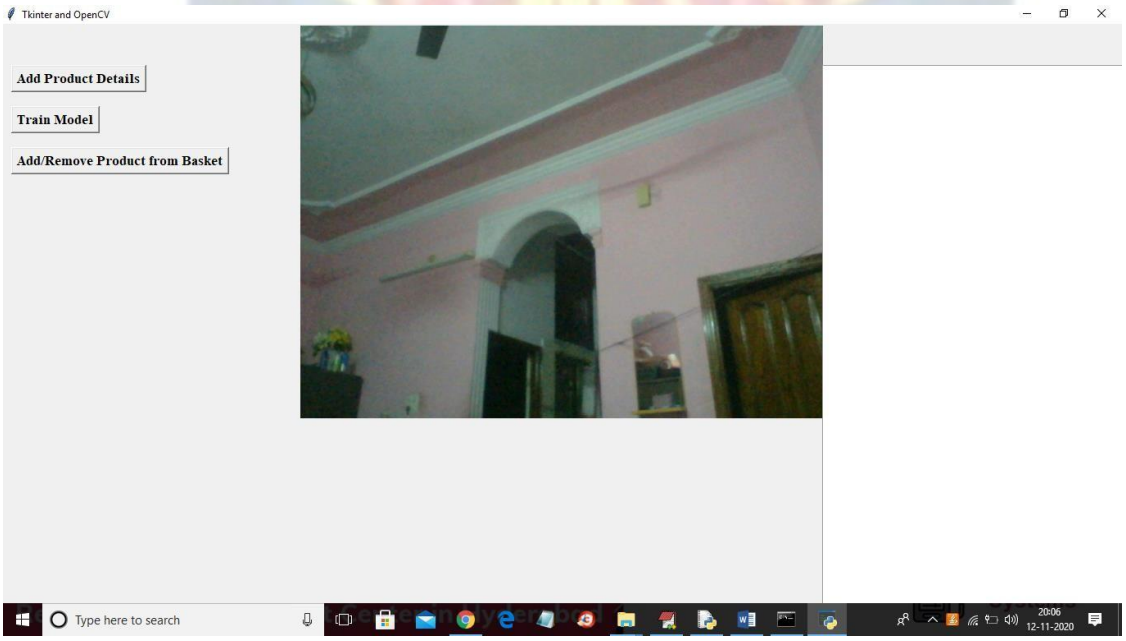


FIG. 8.3 OPENING WEBCAM

In above screen we can see application connected to web cam and now click on ‘Train Model’ button to train model with images.

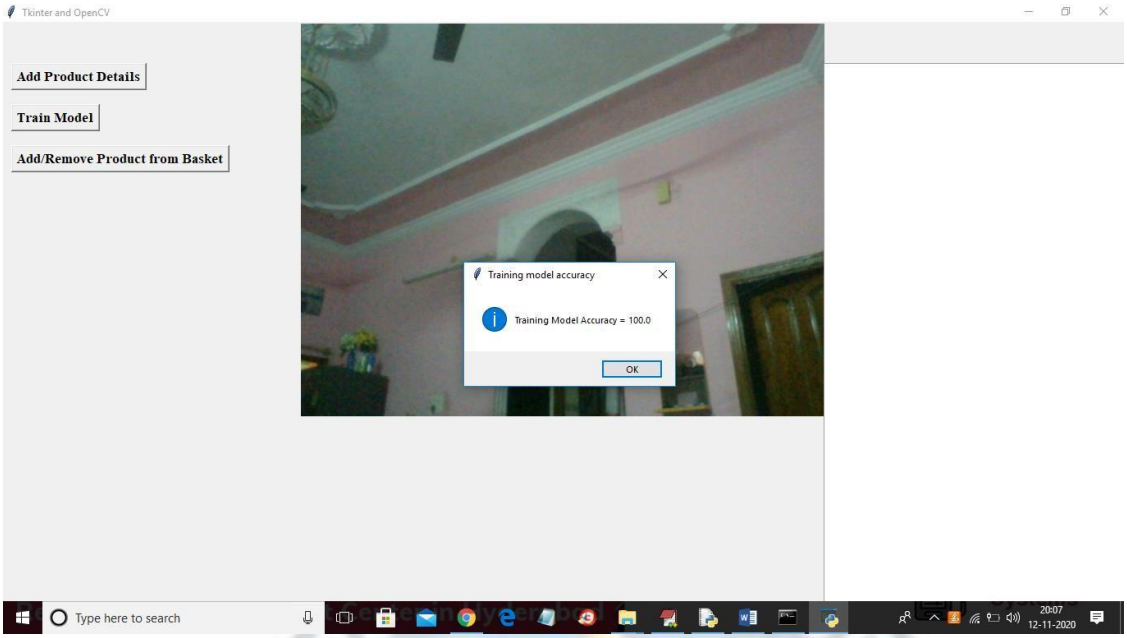


FIG. 8.4 TRAIN MODEL

In above screen train model generated with 100% accuracy and now show product to web cam and click on ‘Add/Remove Product from Basket’ button to allow application to identify product image and then show in text area and if we again show same product then application will remove from text area.

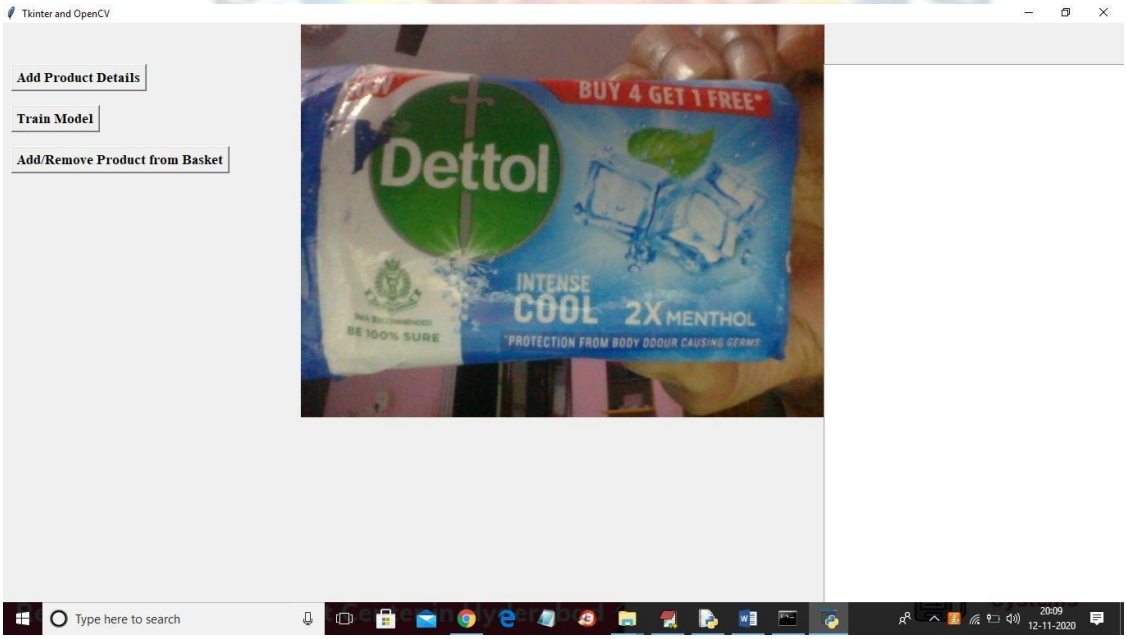


FIG. 8.5 ADDING NEW PRODUCT 1

In above screen I am showing one product and after clicking on ‘Add/Remove Product from Basket’ button will get belowresult.

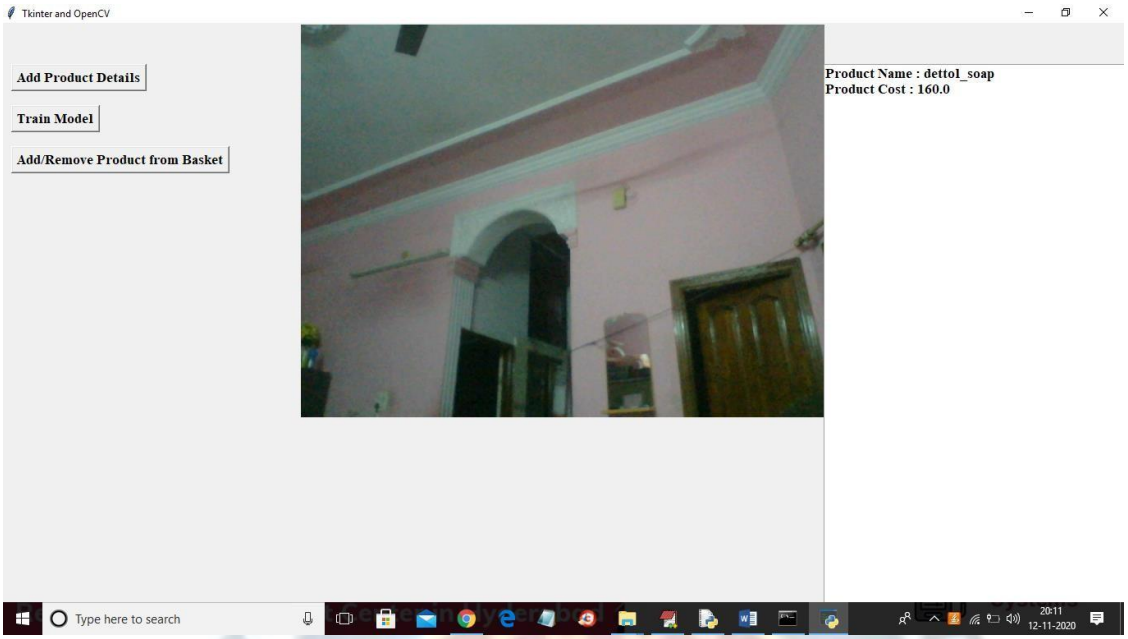


FIG. 8.6 ADDING PRODUCT 1 DETAILS

In above screen in text area we can consider as basket and the name of product and cost is displaying and now try with other product.

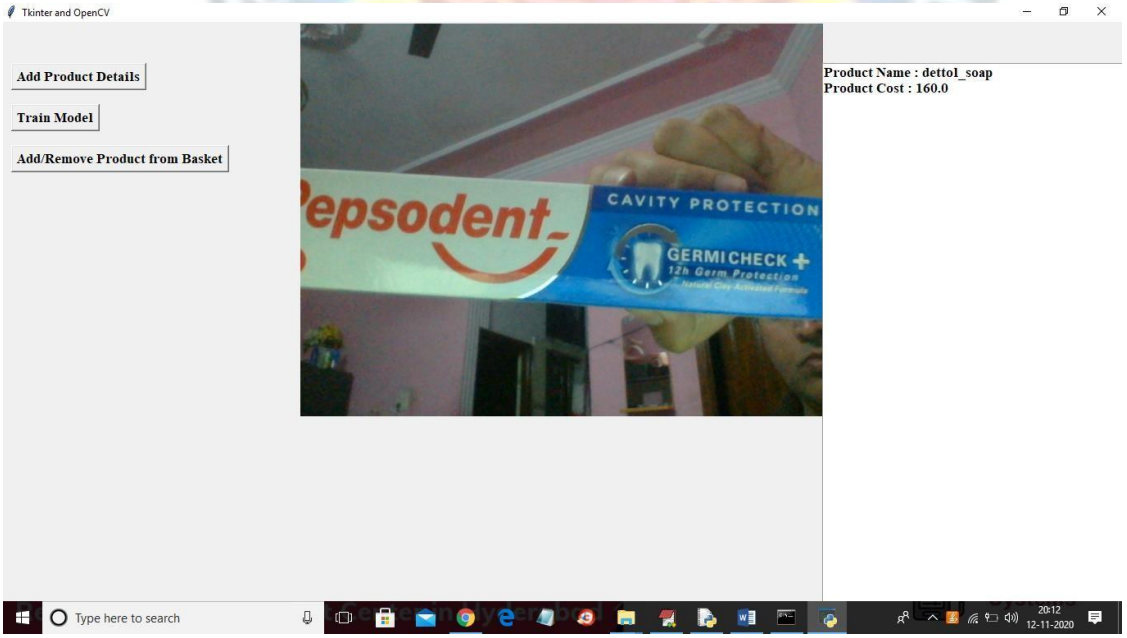


FIG. 8.7 ADDING NEW PRODUCT 2

In above screen showing another image and after clicking on ‘Add/Remove Product from basket’ button will get below screen.

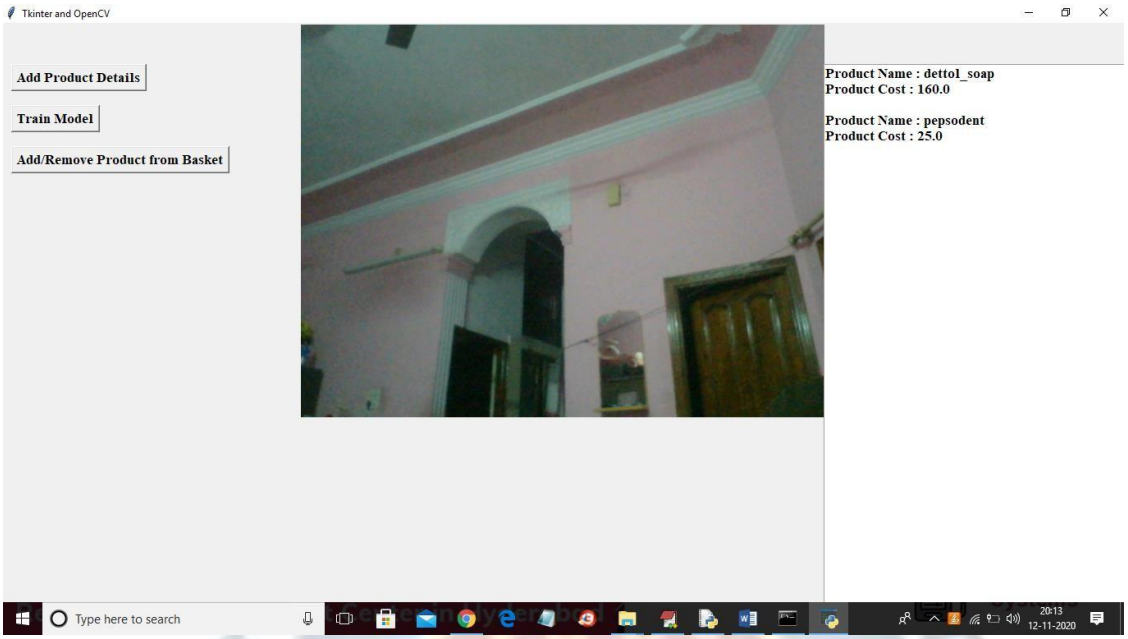


FIG. 8.8 ADDING PRODUCT 2 DETAILS

In above screen we can see two products added to basket and now show same product again to remove from basket.

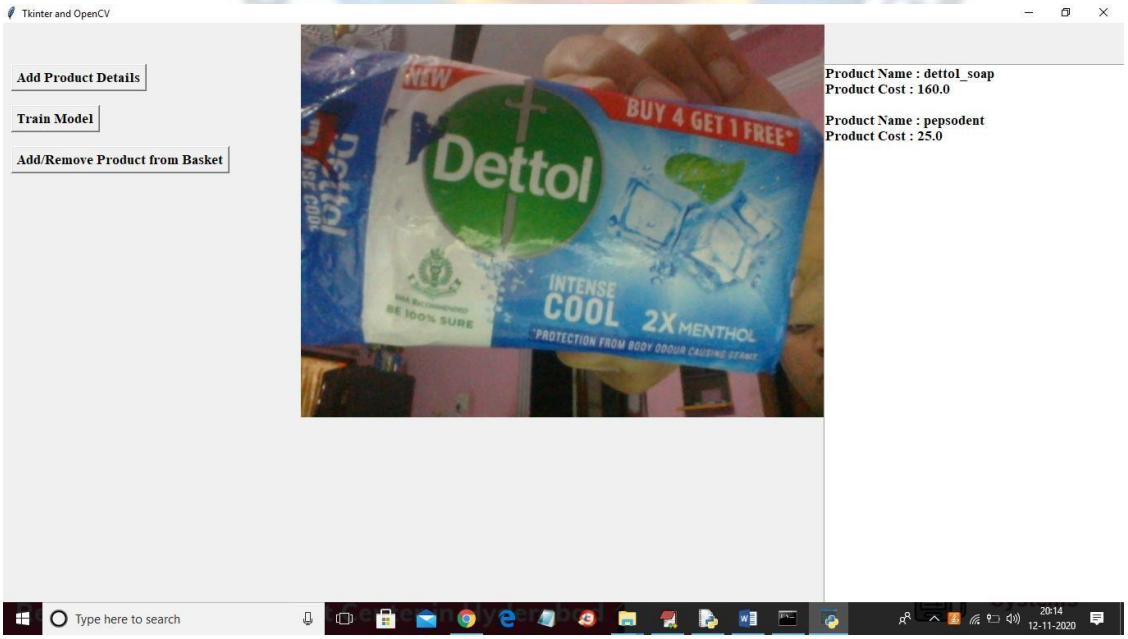


FIG. 8.9 REMOVING PRODUCT 1

In above screen I am showing same product again and then application identified this item from basket and removed it and see the below output screen after removing item.

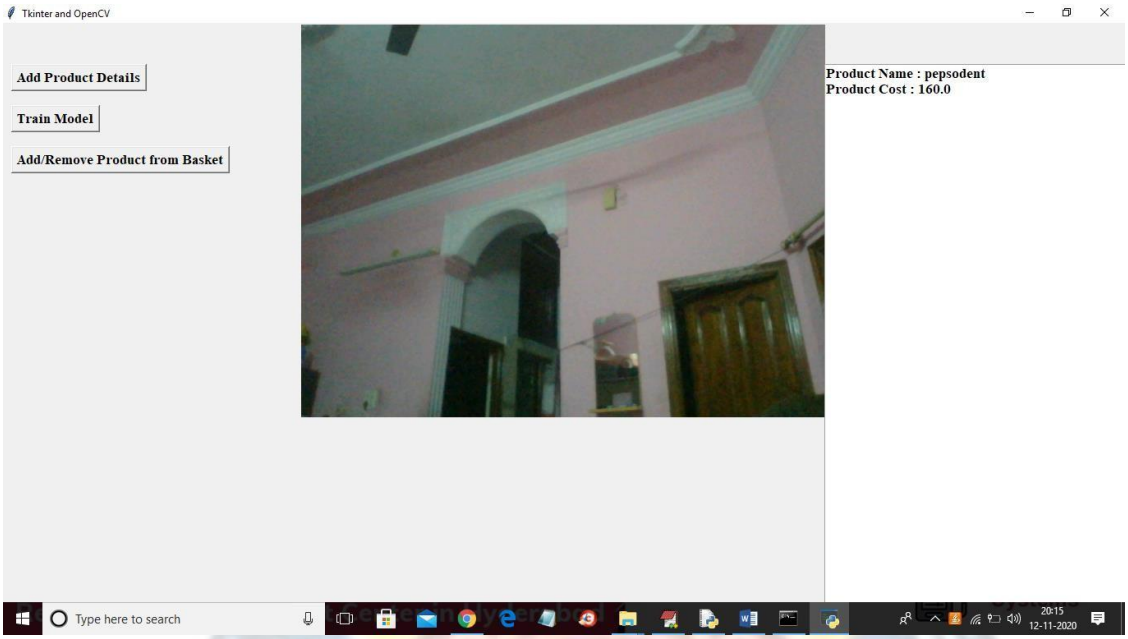


FIG. 8.10 AFTER REMOVING PRODUCT 1

Similarly u can test with other products also and if u have new images then send to us we will rebuild model as per your image and send.

9. CONCLUSION

In this Python project, the users are also provided an option to purchase items from the supermarket. The user can view items and then purchase the items which they need. To buy an item, the user needs to enter the product name and then click enter to confirm. This system then displays a message saying the user to pay the price of the item in the counter. In the modern era, people have more income to spend and lesser time to spend, so they typically opt for supermarkets for grocery. Truly the client is ina position & absolves to opt for product from large on the market varieties which attract the large customers mainly in big cities thus therefore long queues of shoppers are seen at these stores. In several cases, the barcode is either broken or there is also downside in reading barcode because of lighting effects, low resolutionetc.

10. FUTUREENHANCEMENT

In the modern era, the people have more income to spend and lesser time to spend, so they generally typically opt for supermarkets for grocery. Truly the client is in a position & absolves to opt for product from large on the market varieties which attract the large customers mainly in big cities thus therefore long queues of shoppers are seen at these stores. In several cases, the barcode is either broken or there is also downside in reading barcode because of lighting effects, low resolution etc. A bar code based billing system is also expensive as it requires bar coding of all products. The planet is moving towards an era of automation and creature is a great asset which should be utilized in additional intellectual works instead of manual, monotonous works.

11. REFERENCES

[1] Atzori, L., Iera, A., & Morabito, G, “The internet of things: A survey,” Computer Networks, vol. 54, no. 15, 2010, pp.2787–2805

[2] Lizheng Liu1, Bo Zhou2, Zhuo Zou1, Shih-Ching Yeh1, Lirong Zheng" Image Processing System for Automatic Segmentation and Yield Prediction of fruits using Open CV." International Conference on Emerging Trends & Innovations in Engineering and Technological Research(2018).

[3] Sarvini T, Sneha T, Sukanya Gowthami G S, Sushmita S & R Kumar "Performance Comparison of weed Detection Algorithm" IEEE, International Conference on Communication and Signal Processing, April 4-6, India.2019.

[4] GorbunovVladimir(&) , IonovEvgen(&) , and Naing Lin Aung " Automatic Detection & classification of weaving fabric defects based on digital image processing." Second International Conference on green computing(2019).

[5] https://www.amazon.com/b?ie=UTF8&node=16008589011

[6] Zhang, Yanan, H. Wang, and F. Xu. "Object detection and recognition of intelligent service robot based on deep learning." IEEE International Conference on Cybernetics and Intelligent Systems IEEE,2018.

[7] Martinez-Martin, Ester, and A. P. D. Pobil. "Object Detection and Recognition for Assistive Robots." IEEE Robotics & Automation MagazinePP.99(2017):1-1.

[8] Zhang, Shuai, et al."New Object Detection, Tracking, and Recognition Approaches for Video Surveillance Over Camera Network." IEEE Sensors Journal 15.5(2015):2679-2691.

[9] Oliveira, Bernardo A. G. De, F. Magalhaes, and C. A. P. D. S. Martins. "Fast and Lightweight Object Detection Network: Detection and recognition on resource constrained devices." IEEE Access PP.99(2018):1-1. [10] Ren, S., He, K., Girshick, R., Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. In: NIPS.2

[10] Jerry B, Andrea C. Bitcoin: A Primer for Policymakers. Mercatus Center, George Mason University, 2013.

[11] Mayur Chaudhari, Amit Gore, Rajendra Kale and S.H. Patil, “Intelligent ShoppingCartwith Goods Management Using Sensors”, International Research Journal of Engineeringand Technology (IR-JET), Volume 3 Issue 05 May 2016.

[12] “Scan to Arduino”, Lund Software Tools, Google Play Store.https://play.google.com/store/apps/details?id=se.LundSoftwares.ScanToArduino&hl.Lastvisit:25.01.2018 .

[13] B. Ananthabarathi, “High Speed Billing System in Departmental Stores” Middle-East Journal of Scientific Research, pp. 1828-1832, 2012.

[14] R.Rajeshkumar, R.Mohanraj and M.Varatharaj, “Automatic Barcode Based Bill Calculation by Using Smart Trolley”, International Jour-nal of Engineering Science and Computing, Volume 6, Issue 3, 2016.

[15] Janhavi Iyer, Harshad Dhabu and Sudeep K. Mohanty, “Smart Trol-ley System for Automated Billing using RFID and ZIGBEE”, Inter-national Journal of Emerging Technology and Advanced Engineer-ing, Volume 5, Issue 10, October 2015.

[16] P. Chandrasekar and T. Sangeetha, “Smart Shopping Cart with Au-tomatic Billing System through RFID and ZigBee”, In proceedings of IEEE International Conference on Information, Communication and Embedded System (ICICES) pp. 1-4 , 2014.

[17] 7.S.Rohith and C Madhusudan, “Easy Billing System at Shopping Mall Using Hitech Trolly”, International Journal & Magazine of Engineering, Technology, Management and Research, Volume 2, Issue 7, July 2015

12. PROFILES

Akhila Bodige is currently pursuing her Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. She completed her intermediate from Narayana Junior College and 10th class from SRI Saraswathi High School.Hertechnical skills include C,C++,Python and MySQL. She also has a basic understanding of Java.She is also a student of Smart Interviews. Her participations include: National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020,"MHRDInnovation cell On Startup policy",Women online workshop on "Women in Cyber Security and Privacy in 2020" which was conducted from 6th to 10th July 2020.Her areas of interest are JavaScirpt ,Python ,Machine Learning.She completed few certification courses from online platforms like Coursera, CursaApp and SoloLearn.

Prodduturi Lavanyais currently pursuing her Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. She completed her intermediate from Sri Chaitanya panineeya MV Junior College and 10th class from Sri Chaitanya H S TBP School .Hertechnical skills include C,C++,Python and MySQL. She also has a basic understanding of Java.She is also a student of Smart Interviews. Her participations include: National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020,"MHRDInnovation cell On Startup policy",Women online workshop on "Women in Cyber Security and Privacy in 2020" which was conducted from 6th to 10th July 2020.Her areas of interest are JavaScirpt ,Python ,Machine Learning.She completed few certification courses from online platforms like Coursera, CursaApp and SoloLearn.

Poojitha Vadde is currently pursuing her Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. She completed her intermediate from Narayana Junior College and 10th class from Universal Vidyalayam School .Hertechnical skills include C,C++,Python and MySQL. She also has a basic understanding of Java.She is also a student of Smart Interviews. Her participations include: National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020,"MHRDInnovation cell On Startup policy",Women online workshop on "Women in Cyber Security and Privacy in 2020" which was conducted from 6th to 10th July 2020.Her areas of interest are JavaScirpt ,Python ,Machine Learning.She completed few certification courses from online platforms like Coursera, CursaApp and SoloLearn.

Sindhuja R is currently pursuing her Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. She completed her intermediate from Prathibha Junior College and 10th class from Jawahar Navodaya Vidyalaya .Hertechnical skills include C,C++,Python and MySQL. She also has a basic understanding of Java.She is also a student of Smart Interviews. Her participations include: National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020,"MHRDInnovation cell On Startup policy",Women online workshop on "Women in Cyber Security and Privacy in 2020" which was conducted from 6th to 10th July 2020.Her areas of interest are JavaScirpt ,Python ,Machine Learning.She completed few certification courses from online platforms like Coursera, CursaApp and SoloLearn.

A

PROJECT REPORT

On

MACHINE LEARNING BASED APPROCHES FOR DETECTING COVID-19 USING CLINICAL TEXTDATA

Submitted by

1. Ms. V. Athiksha (17K81A05P4)

2. Ms. D. Chitra (17K81A05K5)

3.Mr. G. Snehith (17K81A05L0)

4.Mr. A. Akhil Reddy (17K81A05J5)

in partial fulfillment for the award of the

degree of

BACHELOR OF TECHNOLOGY

IN

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Under the Guidance of

Mr. U. Nagaiah

Assistant Professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ST. MARTIN'S ENGINEERING COLLEGE
An Autonomous Institute

Dhulapally, Secunderabad – 500100

JUNE 2021

BONAFIDE CERTIFICATE

This is to certify that the project entitled "Machine learning based approaches for detecting COVID-19 using clinical text data", is being submitted by **Venkannagari Athiksha (17K81A05P4), Diddi Chitra (17K81A05K5), G. Venkata Snehith (17K81A05L0) and A. Akhil Reddy (17K81A05J5)** in partial fulfillment of the requirement for the award of the degree of **BACHELOR OF TECHNOLOGY IN COMPUTER SCIENCE AND ENGINEERING** is recorded of Bonafede work carried out by them. The results embodied in this report have been verified and found satisfactory.

Assistant Professor

Mr.U.Nagaiah

Department of CSE

Head of the Department

Dr.M.NARAYANAN

Department of CSE

Internal Examiner

External Examiner

Place:

Date:

DECLARATION

We, the students of **Bachelor of Technology** in the Department of **'Computer Science and Engineering'**, session: 2017 – 2021, St. Martin's Engineering College, Dhulapally, Kompally, Secunderabad, hereby declare that work presented in this Project Work entitled Machine learning based approaches for detecting COVID-19 using clinical text data is the outcome of our own Bonafede work and is correct to the best of our knowledge and this work has been undertaken taking care of Engineering Ethics. This result embodied in this project report has not been submitted in any university for award of any degree.

Ms. V. Athiksha (17K81A05P4)
Ms. D. Chitra (17K81A05K5)
Mr. G. Snehith (17K81A05L0)
Mr. A. Akhil Reddy (17K81A05J5)

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible and whose encouragements and guidance have crowded effects with success.

We extended our deep sense of gratitude to our respected Principal**,** **Dr. P. SANTOSH KUMAR PATRA**, St. Martin's Engineering College, Dhulapally, for permitting us to undertake this project. We are also thankful to **Dr. M. NARAYANAN**, Head of the Department, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

We would like to thank **Dr. T. POONGOTHAI,** Department of Computer Science and Engineering (AI & ML) for her encouragement and insightful comments and as well as our project coordinators **Dr. B. RAJALINGAM**, Associate Professor and **Mr. J. SUDHAKAR**, Assistant Professor, in Department of Computer Science and Engineering for their valuable support.

We would like to express our sincere gratitude and indebtedness to our project supervisor Mr. UPPULA NAGAIAH, Assistant Professor, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for her support and guidance throughout our project.

Finally, we express thanks to all those who have helped us successfully to completing this project. Furthermore, we would like to thank our family and friends for their moral support and encouragement.

We express thanks to all those who have helped us in successfully completing the project.

Ms. V. Athiksha (17K81A05P4)
Ms. D. Chitra (17K81A05K5)
Mr. G. Snehith (17K81A05L0)
Mr. A. Akhil Reddy (17K81A05J5)

ABSTRACT

Technology advancements have a rapid effect on every field of life, be it medical field or any other field. Artificial intelligence has shown the promising results in health care through its decision making by analyzing the data. COVID-19 has affected more than 100 countries in a matter of no time. People all over the world are vulnerable to its consequences in future. It is imperative to develop a control system that will detect the coronavirus. One of the solutions to control the current havoc can be the diagnosis of disease with the help of various AI tools. In this paper, we classified textual clinical reports into four classes by using classical and ensemble machine learning algorithms. Feature engineering was performed using techniques like Term frequency/inverse document frequency (TF/IDF), Bag of words (BOW) and report length. These features were supplied to traditional and ensemble machine learning classifiers. Logistic regression and Multinomial Naïve Bayes showed better results than other ML algorithms by having 96.2% testing accuracy. In future recurrent neural network can be used for better accuracy.

TABLE OF CONTENTS

LIST OF TABLES

Table 1. List of Tables

LIST OF FIGURES

Table 2. List of Figures

LIST OF OUTPUT SCREENS

Table 3. List of output screens

LIST OF ABBREVIATIONS

ARDS	Acute Respiratory Distress Syndrome
SVM	Support Vector Machine
UML	Unified Modelling Language
GUI	Graphical User Interface
SARS	Severe Acute Respiratory Syndrome
TF	Term Frequency
IDF	Inverse Document Frequency
BOW	Bag of Words
MNB	Multinomial Naïve Bayes
ML	Machine Learning
ANN	Artificial Neural Networks
MERS	Middle East Respiratory Syndrome
RAM	Random Access Memory
CPU	Central Processing Unit
ER	Entity-Relationship
DFD	Data Flow Diagram

Table 4. List of abbreviations

GLOSSARY OF TERMS

TERM	MEANING
Machine Learning	Machine learning is a method of data analysis that automates analytical model building and these systems can learn from data, identify patterns and make decisions with minimal human intervention.
Ensemble Machine Learning	Ensemble learning helps improve machine learning results by combining several models. Ensemble methods are meta-algorithms that combine several machine learning
Support Vector Machine	Support vector machines (SVMs) are linear classifiers which are based on the margin maximization principle.
Logistic Regression	Logistic regression predicts the class of numerical variable based on its relationship with the label
Multinomial Naïve Bayes	MNB computes class probabilities of a given text by using Bayes rule
Decision trees	An alternative approach for classification it partitions the input space into regions and classifies every region independently
Bagging	An ensemble machine learning algorithm which improves the performance of other classification and regression machine learning algorithms
AdaBoost	AdaBoost ensemble learning algorithm works with those instances of the dataset, which are weighted

Table 5. Glossary of Terms

1. INTRODUCTION

In December 2019, the novel coronavirus appeared in the Wuhan city of China and was reported to the World Health Organization (W.H.O) on 31st December 2019. The virus created a global threat and was named as COVID-19 by W.H.O on 11th February 2020. The COVID-19 is the family of viruses including SARS, ARDS. W.H.O declared this outbreak as a public health emergency and mentioned the following: the virus is being transmitted via the respiratory tract when a healthy person comes in contact with the infected person.

The virus may transmit between persons through other roots which are currently unclear. The infected person shows symptoms within 2–14 days, depending on the incubation period of the middle east respiratory syndrome (MERS), and the severe acute respiratory syndrome (SARS).

According to W.H.O the signs and symptoms of mild to moderate cases are dry cough, fatigue and fever while as in severe cases dyspnea (shortness of breath), Fever and tiredness may occur. The persons having other diseases like asthma, diabetes, and heart disease are more vulnerable to the virus and may become severely ill. The person is diagnoses based on symptoms and his travel history. Vital signs are being observed keenly of the client having symptoms.

No specific treatment has been discovered as on 10th April 2020, and patients are being treated symptomatically. The drugs like hydroxychloroquine, antipyretic, anti-viral are used for the symptomatic treatment.

Currently, no such vaccine is developed for preventing this deadly disease, and we may take some precautions to prevent this disease. By washing hands regularly with soap for 20 s and avoiding close contact with others by keeping the distance of about 1 m may reduce the chances of getting affected by this virus.

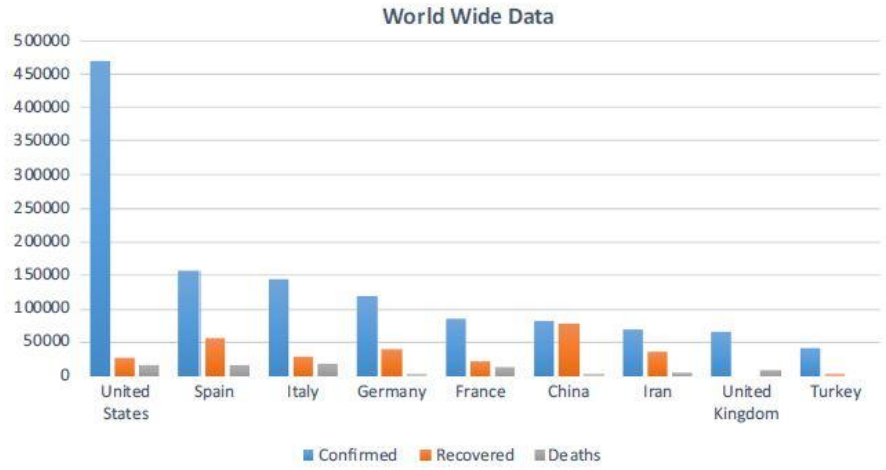
While sneezing, Covering the mouth and nose with the help of disposable tissue and avoiding the contact with the nose, ear and mouth can help in its prevention. SARS is an airborne disease that appeared in 2003 in China and affected 26 countries by having 8 K cases in the same year and transferred from person to person.

The signs and symptoms of SARS are fever, cold, diarrhea, shivering, malaise, myalgia and dyspnea The ARDS (acute respiratory distress syndrome) is characterized by rapid onset of inflammation in lungs which leads to respiratory failure and its signs and symptoms are bluish skin color, fatigue and shortness of breath. ARDS is diagnosed by PaO2/FiO2 ratio of less than 300 mm Hg. Till 10th of April 2020, almost 1.6 million confirmed cases of coronavirus are detected around the globe.

Almost 97 K persons have died, and 364 K persons have recovered from this deadly virus [5]. Figure 1 shows the worldwide data regarding coronavirus. Since no drug or vaccine is made for curing the COVID-19. Various paramedical companies have claimed of developing a vaccine for this virus. Less testing has also given rise to this disease as we lack the medical resources due to pandemic. Since thousands and thousands are being tested positive day by day around the globe, it is not possible to test all the persons who show symptoms.

Apart from clinical procedures, machine learning provides a lot of support in identifying the disease with the help of image and textual data. Machine learning can be used for the identification of novel coronavirus. It can also forecast the nature of the virus across the globe. However, machine learning requires a huge amount of data for classifying or predicting diseases.

Supervised machine learning algorithms need annotated data for classifying the text or image into different categories. From the past decade, a huge amount of progress is being made in this area for resolving some critical projects. Recent pandemic has attracted many researchers around the globe to solve this problem. Data provided by John Hopkins University in the form of X-ray images and various researchers build a model of machine learning that classifies X-ray image into COVID-19 or not. Since the latest data published by Johns Hopkins gives the metadata of these images.



1.1 PROJECT OVERVIEW

We propose a methodology for detecting COVID -19 virus type where the data consists of clinical reports in the form of text in this project, we are classifying that text into four different categories of diseases such that it can help in detecting coronavirus from earlier clinical symptoms.

We used supervised machine learning techniques and ensemble learning techniques for classifying the text into four different categories COVID, SARS, ARDS and Both (COVID, ARDS).

1.2 PROJECT OBJECTIVE

The main objective of this project is It is imperative to develop a control system that will detect the coronavirus. One of the solutions to control the current havoc can be the diagnosis of disease with the help of various AI tools. The central focus of this project is to deploy the system with much higher accuracy which in result assembles the classification for further consequences.

1.3 ORGANIZATION OF CHAPTERS

This documentation consists of 10 different chapter, and they are:

1. Introduction – This chapter covers the overview of our project and its objectives.
2. Literature Survey – This includes the details of our survey.
3. Software and Hardware Requirements – We specify our software and hardware requirements here.
4. Software Development Analysis – This section includes the problem definition and details of the modules we used in our project.

5. Project System Design – This chapter includes the design part of our project which includes UML diagrams.
6. Project Coding – This section contains the details of our project code.
7. Project Testing – The details of test cases and testing are included in this chapter.
8. Output Screens – This contains the screenshots of how our project looks like when executed.
9. Experimental Results – This chapter contains the screenshots of our results.
10. Conclusion and Future Enhancements – This covers the conclusion of our project.

2. LITERATURE SURVEY

Machine learning and natural language processing use big data-based models for pattern recognition, explanation, and prediction.

NLP has gained much interest in recent years, mostly in the field of text analytics, Classification is one of the major task in text mining and can be performed using different algorithms [1]. Kumar et al. [2] performed a SWOT analysis of various supervised and unsupervised text classification algorithms for mining the unstructured data. The various applications of text classification are sentiment analysis, fraud detection, and spam detection etc. Opinion mining is majorly being used for elections, advertisement, business etc. Verma et al. [3] analysed Sentiments of Indian government projects with the help of the lexicon-based dictionary. The machine learning has changed the perspective of diagnosis by giving great results to diseases like diabetes and epilepsy.

Chakraborti et al. [4] detected epilepsy using machine learning approaches, electroencephalogram (EEG) signals are used for detecting normal and epileptic conditions using artificial neural networks (ANN). Sarwar et al. [5] diagnosis diabetes using machine learning and ensemble learning techniques result indicated that ensemble technique assured accuracy of 98.60%. These purposes can be beneficial to diagnose and predict COVID-19. Firm and exact diagnosis of COVID-19 can save millions of lives and can produce a massive amount of data on which a machine learning (ML) models can be trained. ML may provide useful input in this regard, in making diagnoses based on clinical text, radiography Images etc.

According to Bullock et al. [6], Machine learning and deep learning can replace humans by giving an accurate diagnosis. The perfect diagnosis can save radiologists' time and can be cost-effective than standard tests for COVID-19. X-rays and computed tomography (CT) scans can be used for training the machine learning model. Several initiatives are underway in this regard.

Wang and Wong [7] developed COVID-Net, which is a deep convolutional neural network, which can diagnose COVID-19 from chest radiography images. Once the COVID-19 is detected in a person, the question is whether and how intensively that person will be significantly, this provides a greater accuracy to the overall classification process. The mindset of this task is to analyze various machine learning techniques for binary classification concerning with illness i.e., to diagnose whether a subject is suffering from disease or not. affected. Not all COVID-19 positive patients will need rigorous attention. Being able to prognosis who will be affected more severely can help in directing assistance and planning medical resource allocation and utilization.

Yan et al. [8] used machine learning to develop a prognostic prediction algorithm to predict the mortality risk of a person that has been infected, using data from (only) 29 patients at Tongji Hospital in Wuhan, China. Jiang et al. [9] proposed a machine learning model that can predict a person affected with COVID-19 and has the possibility to develop acute respiratory distress syndrome (ARDS). The proposed model resulted in 80% of accuracy. The samples of 53 patients were used for training their model and are restricted to two Chinese hospitals. ML can be used to diagnose COVID-19 which needs a lot of research effort but is not yet widely operational. Since less work is being done on diagnosis and predicting using text, we used machine learning and ensemble learning models to classify the clinical reports into four categories of viruses.

The virus may transmit between persons through other roots which are currently unclear. The infected person shows symptoms within 2–14 days, depending on the incubation period of the middle east respiratory syndrome (MERS), and the severe acute respiratory syndrome (SARS). According to W.H.O the signs and symptoms of mild to moderate cases are dry cough, fatigue and fever while as in severe cases dyspnea (shortness of breath), Fever and tiredness may occur [10]. The persons having other diseases like asthma, diabetes, and heart disease are more vulnerable to the virus and may become severely ill. The person is diagnoses based on symptoms and his travel history. Vital signs are being observed keenly of the client having symptoms. No specific treatment has been discovered as on 10th April 2020, and patients are being treated symptomatically. The drugs like hydroxychloriquine, antipyretic, anti-virals are used for the symptomatic treatment.

Apart from clinical procedures, machine learning provides a lot of support in identifying the disease with the help of image and textual data. Machine learning can be used for the identification of novel coronavirus. It can also forecast the nature of the virus across the globe. However, machine learning requires a huge amount of data for classifying or predicting diseases. Supervised machine learning algorithms need annotated data for classifying the text or image into different categories. From the past decade, a huge amount of progress is being made in this area for resolving some critical projects.

2.2 CONCLUSIONS ON SURVEY

From [1] These works provide basic background information about various techniques and algorithms in AI tool learning and machine learning in order to provide with data processing, accuracy, classification.From [2] They use multiple algorithms like K-NN Algorithm, nearest Neighbor Algorithms, Support vector machine (SVM) etc [3,4]. under various contexts.From[5] Deep learning and Machine Learning algorithms are a major part of classification involved projects since they are the fundamental aspect to these projects[7].

From [8.9] Without effective and necessary knowledge about these concepts in detail, it would have been extremely difficult for us to choose and apply the required algorithm to make our project's output reliable. From[10] Every reference gave an example to how to use ensemble learning and machine learning to work on our project and produce the existing project.

3. SOFTWARE AND HARDWARE REQUIREMENTS

Requirement is a condition or capability possessed by the software or system component in order to solve a real-world problem. The problems can be to automate a part of a system, to correct shortcomings of an existing system, to control a device, and so on.

Requirements describe how a system should act, appear or perform. For this, when users request for software, they provide an approximation of what the new system should be capable of doing. Requirements differ from one user to another and from one business process to another.

The purpose of the requirements document is to provide a basis for the mutual understanding between the users and the designers of the initial definition of the software development life cycle (SDLC) including the requirements, operating environment and development plan.

Requirements help to understand the behavior of a system, which is described by various tasks of the system. For example, some of the tasks of a system are to provide a response to input values, determine the state of data objects, and so on. Note that requirements are considered prior to the development of the software. The requirements, which are commonly considered, are classified into three categories, namely, functional requirements, non-functional requirements, and domain requirements.

The functional requirements should be complete and consistent. Completeness implies that all the user requirements are defined. Consistency implies that all requirements are specified clearly without any contradictory definition. Generally, it is observed that completeness and consistency cannot be achieved in large software or in a complex system due to the problems that arise while defining the functional requirements of these systems. The different needs of stakeholders also prevent the achievement of completeness and consistency. Due to these reasons, requirements may not be obvious when they are, ‘first specified and may further lead to inconsistencies in the requirements specification.

The non-functional requirements (also known as **quality requirements)** are related to system attributes such as reliability and response time. Non-functional requirements arise due to user requirements, budget constraints, organizational policies, and so on. These requirements are not related directly to any particular function provided by the system.

Non-functional requirements should be accomplished in software to make it perform efficiently. For example, if an airplane is unable to fulfill reliability requirements, it is not approved for safe operation. Similarly, if a real time control system is ineffective in accomplishing non-functional requirements, the control functions cannot operate correctly.

System requirements are the configuration that a system must have for a hardware or software application to run smoothly and efficiently. Failure to meet these requirements can result in installation problems or performance problems. The former may prevent a device or application from getting installed, whereas the latter may cause a product to malfunction or perform below expectation or even to hang or crash.

System requirements are also known as minimum system requirements. Hardware system requirements often specify the operating system version, processor type, memory size, available disk space and additional peripherals, if any, needed. Software system requirements, in addition to the requirements, may also specify additional software dependencies (e.g., libraries, driver version, framework version). Some hardware/software manufacturers provide an upgrade assistant program that users can download and run to determine whether their system meets a product's requirements.

Some products include both minimum and recommended system requirements. A video game, for instance, may function with the minimum required CPU and GPU, but it will perform better with the recommended hardware. A more powerful processor and graphics card may produce improved graphics and faster frame rates (FPS).

Some system requirements are not flexible, such as the operating system(s) and disk space required for software installation. Others, such as CPU, GPU, and RAM requirements may vary significantly between the minimum and recommended requirements. When buying or upgrading a software program, it is often wise to make sure your system has close to the recommended requirements to ensure a good user experience.

3. SOFTWARE AND HARDWARE REQUIREMENTS

3.1 SOFTWARE REQUIREMENTS

- Operating System:Windows XP.
- Platform: PYTHON TECHNOLOGY
- Tool: Python 3.7

3.2 HARDWARE REQUIREMENTS

- Processor: core i3 or above
- Ram: 4GB

4. SOFTWARE DEVELOPMENT ANALYSIS

Software development is a process of writing and maintaining the source code, but in a broader sense, it includes all that is involved between the conception of the desired software through to the final manifestation of the software, sometimes in a planned and structured process. Therefore, software development may include research, new development, prototyping, modification, reuse, re- engineering, maintenance, or any other activities that result in software products.

4.1 OVERVIEW OF THE PROBLEM

Technology advancements have a rapid effect on every field of life, be it medical field or any other field. Artificial intelligence has shown the promising results in health care through its decision making by analyzing the data. COVID-19 has affected more than 100 countries in a matter of no time. People all over the world are vulnerable to its consequences in future.

It is imperative to develop a control system that will detect the coronavirus. One of the solutions to control the current havoc can be the diagnosis of disease with the help of various AI tools.

4.2 DEFINING THE PROBLEM

Machine Learning can be used to diagnose COVID-19 which needs a lot of research effort but is not yet widely operational. Since less work is being done on diagnosis and predicting using text, we used machine learning and ensemble learning models to classify the clinical reports into four categories of viruses.

The machine learning has changed the perspective of diagnosis by giving great results to diseases like diabetes and epilepsy. These purposes can be beneficial to diagnose and predict COVID-19. Firm and exact diagnosis of COVID-19 can save millions of lives and can produce a massive amount of data on which a machine learning (ML) models can be trained. ML may provide useful input in this regard, in particular in making diagnoses based on clinical text, radiography Images etc.

4.3 MODULES OVERVIEW

1) Data Collection

As W.H.O declared Coronavirus pandemic as Health Emergency. The researchers and hospitals give open access to the data regarding this pandemic. We have collected from an open-source data repository GitHub.1 In which about 212 patient's data is stored which have shown symptoms of corona virus and other viruses. Data consists of about 24 attributes namely patient id, offset, sex, age, finding, survival, intubated, went_icu, needed_supplemental, O2, extubated, temperature, pO2_saturation, leukocyte_ count, neutrophil count, lymphocyte count, view, modality, date, location, folder, filename, DOI, URL. License. Clinical notes and other notes.

2) Relevant Dataset

Since our work is regarding text mining so we extracted clinical notes and findings. Clinical notes consist of text while as the attribute finding consist label of the corresponding text. About 212 reports were used and their length was calculated. We consider only those reports that are written in the English language. Figure 3 gives the length distribution of clinical reports that are written in English. The clinical reports are labelled to their corresponding classes. In our dataset, we have four classes COVID, ARDS, SARS and Both (COVID, ARDS). Figure 4 shows the different classes in which clinical text is being categorized and corresponding report length.

3) Preprocessing

The text is unstructured so it needed to be refined such that machine learning can be done. Various steps are being followed in this phase; the text is being cleaned by removing unnecessary text. Punctuation and lemmatization are being done such that the data is refined in a better way. Stop words, symbols, URL's, links are removed such that classification can be achieved with better accuracy.

4) Feature Engineering

From the preprocessed clinical reports, various features are extracted as per the semantics and are converted into probabilistic values. We use TF//IDF technique for extracting relevant features. Bag of words was also taken into consideration, unigrams, bigrams were also extracted. We identified 40 relevant features by which the classification can be achieved. By giving the corresponding weight to the feature and the same input is being supplied to machine learning algorithms.

4.4 DEFINING THE MODULES

The project mainly consists of 5 modules:

1) Data collection module.

2) Data processing module

3) Feature Engineering module

4) Classification module

5) Result module

4.5 MODULE FUNCTIONALITY

1) Data collection Module – The data collection module enables the user to upload the selective dataset file from local computer.
2) Data Processing Module – The uploaded data we extract all text data from dataset and now in above screen text in first sentence we have ‘on’ stop words and many numbers of numerical values and to remove those stop words and to clean data preprocess all stop words removed out and in above ‘on’ stop word removed out.
3) Feature Engineering module – Using this module, from the preprocessed clinical reports, various features are extracted as per the semantics and are converted into probabilistic values.
4) Classification module – Using this module we can run all traditional algorithms on features data and to calculate accuracy displaying accuracy, precision, recall and F Score for each algorithm.
5) Result module – This module enables we can see accuracy, precision, recall and f score for each algorithm in group bar chart and in above graph x-axis represents algorithm name y-axis represents values.

5. PROJECT SYSTEM DESIGN

Systems design is the process of defining elements of a system like modules, architecture, components and their interfaces and data for a system based on the specified requirements. It is the process of defining, developing and designing systems which satisfies the specific needs and requirements of a business or organization. A systemic approach is required for a coherent and well-running system. Bottom-Up or Top-Down approach is required to take into account all related variables of the system. A designer uses the modelling languages to express the information and knowledge in a structure of system that is defined by a consistent set of rules and definitions. The designs can be defined in graphical or textual modelling languages.

Unified Modelling Language has been used by us to describe software both structurally and behaviorally with notations.

SYSTEM ARCHITECTURE

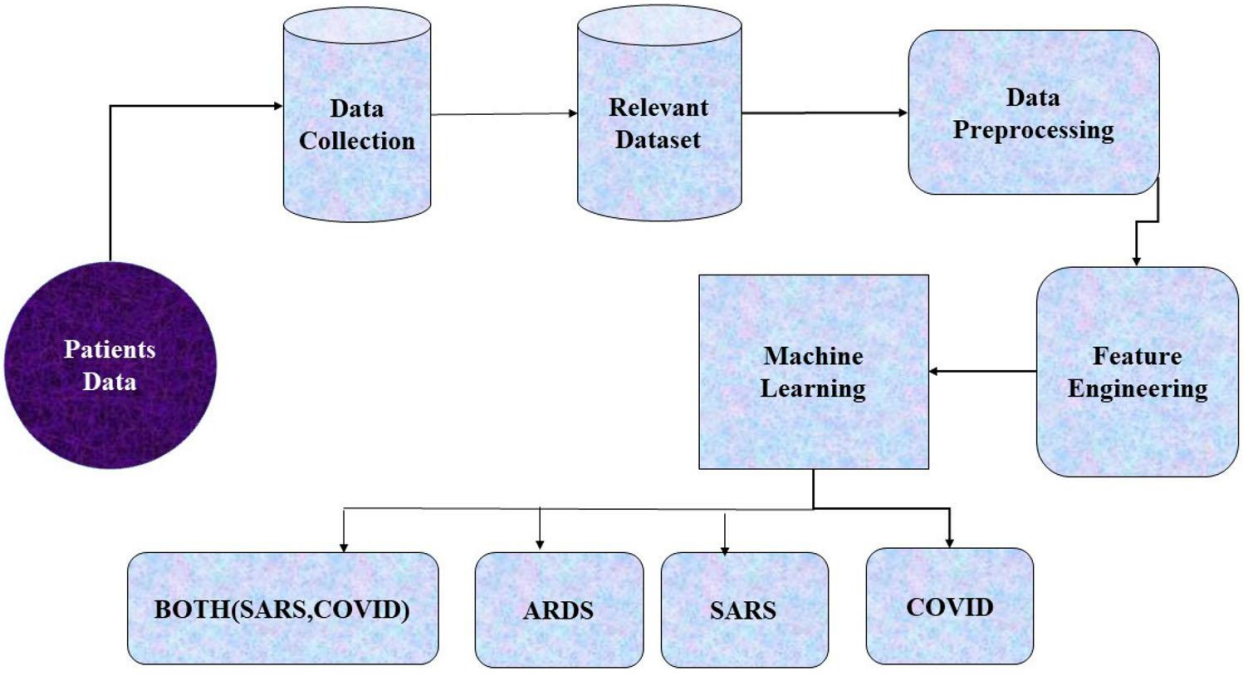


Figure 5.1 System Architecture

The system called the ewheelz uses the 2-tier architecture. The 1st tier is the GUI, which is said to be front-end, and the 2nd tier is the database, which uses My-Sql, which is the back end.

The front-end can be run on different systems (clients). The database will be running at the server. Users access these forms by using the user-ids and the passwords.

5.1 DATA FLOW DIAGRAMS

CONTEXT-LEVEL DIAGRAM:

The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly.

But, using the classic algorithms of machine learning, text is considered as a sequence of keywords; instead, an approach based on semantic analysis mimics the human ability to understand the meaning of a text**.**

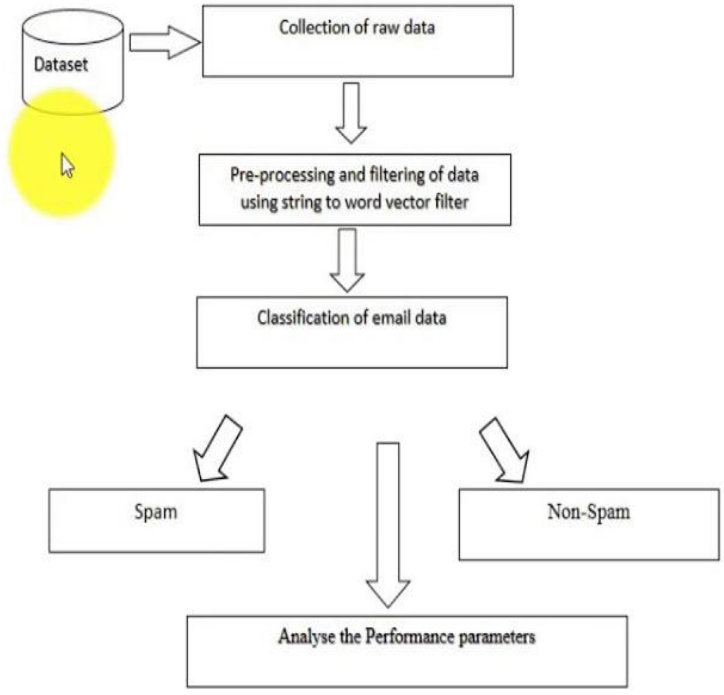


Figure 5.2 Context Level Diagram

LEVEL-0 DIAGRAM

As illustrated in the context diagram, the major processes involved in detecting the COVID - 19 type includes:

Uploading of datasets:

The users should upload the clinical text reports using dataset file from their local computer. The data collection is carried out by uploading the datasets into the model.

Preprocessing of Data:

The text is unstructured so it needed to be refined such that machine learning can be done. Various steps are being followed in this phase; the text is being cleaned by removing unnecessary text. Punctuation and lemmatization are being done such that the data is refined in a better way. Stop words, symbols, URL's, links are removed such that classification can be achieved with better accuracy.

Running probabilistic and statistical tactics:

Since our work is regarding text mining so we extracted clinical notes and findings. Clinical notes consist of text while as the attribute finding consist label of the corresponding text. About 212 reports were used and their length was calculated. We consider only those reports that are written in the English language. Figure 3 gives the length distribution of clinical reports that are written in English. The clinical reports are labelled to their corresponding classes. In our dataset, we have four classes COVID, ARDS, SARS and Both (COVID, ARDS). It shows the different classes in which clinical text is being categorized and corresponding report length.

Classification of virus:

The data consists of clinical reports in the form of text in this paper, we are classifying that text into four different categories of diseases such that it can help in detecting coronavirus from earlier clinical symptoms. We used supervised machine learning techniques for classifying the text into four different categories COVID, SARS, ARDS and Both (COVID, ARDS). We are also using ensemble learning techniques for classification.

Mechanism for feature extraction:

We identified 40 relevant features by which the classification can be achieved. By giving the corresponding weight to the feature and the same input is being supplied to machine learning algorithms.

All these processes have been summarized and diagrammatically represented in the level-0 diagram below:

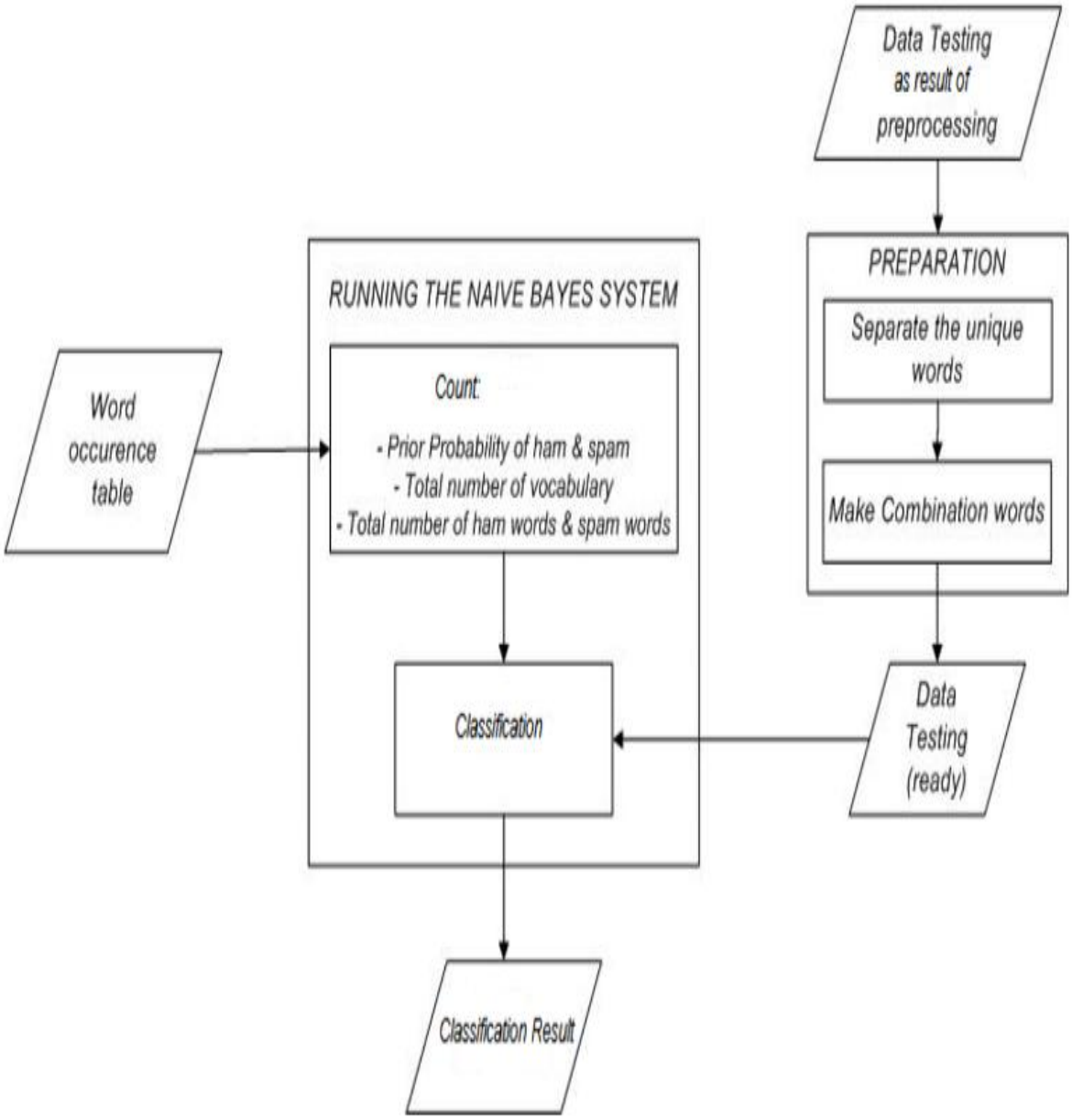


Figure 5.3 Level-0 DFD

5.2 ENTITY-RELATIONSHIP DIAGRAM

An entity relationship model is a high-level conceptual model which describes data in terms of entities, their attributes and their relationships (Riccardi, 2002). The entity relationship diagram shows how is represented and organized in the database schema without specifying the actual data (Pagh, 2006).

The user entity can choose to upload the selective dataset files into the system. The user datafile indicates as the single attribute that has one to many key relationships with other logistics and statistical values. The classifications are enabled in accordance to the key entities of the datasets uploaded.

5.3 UML DIAGRAMS

CLASS DIAGRAM

A class diagram provides a pictorial representation of all the classes in an object-oriented system; their attributes and methods; their connections; their interactions and inheritances if any. In simpler terms, classes represent objects whose roles are similar and to what extend the objects of the classes “know” about each other (Felicia, 2011).

The user can choose any of their respective clinical text report as in their dataset file from their local computer .The classes which define their interactions with user namely involve the uploading the dataset , preprocessing the dataset ,to regulate the future enhancement the key constraint class Future engineering is used, the classification of the data is enabled in due course with the relative classes Run logistic regression , Naïve bayes ,Decision tree &SVM , and ensemble learning algorithms Bagging , Adaboost, Random forest , Granit boosting .The resultant report is showed in the final comparative graph.

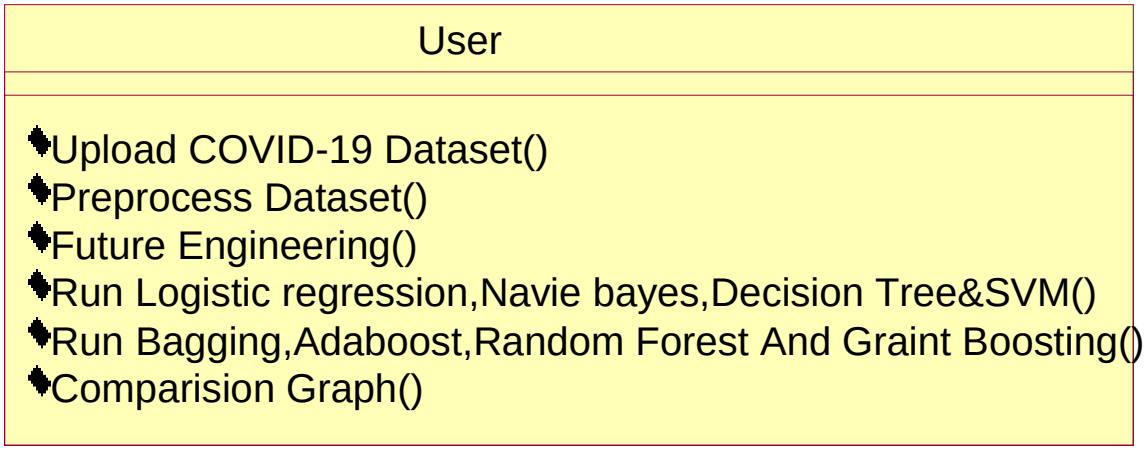


Figure 5.5 Class Diagram

USECASE DIAGRAM

A use case is simply a list of actions which typically define the interactions between an actor and the system with an aim of achieving a certain goal. Each interaction is a single unit of work and captures a “contract” for the behavior of the system under discussion to deliver a single goal (Kettenis, 2007). Most of the functional requirements are captured by the use case. The diagram is displayed below:

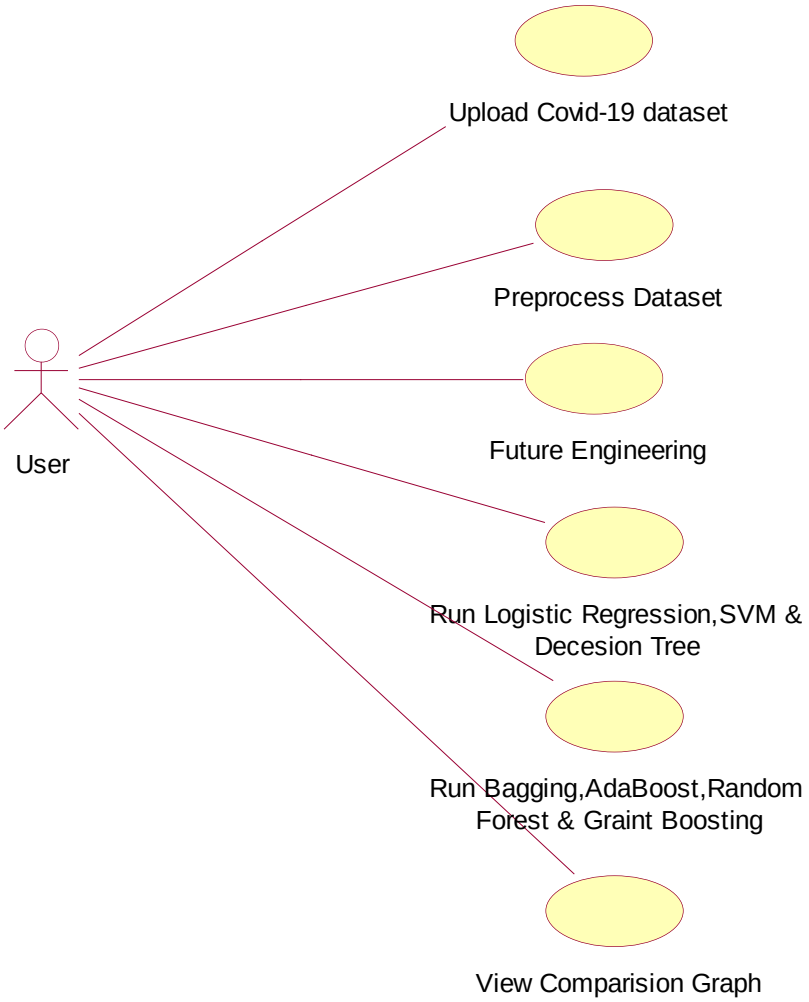


Figure 5.6 Use-case Diagram

SEQUENCE DIAGRAM

The sequence diagram in this case provides a visual representation the object interactions during the searching process. This includes the actor and the objects the actors interact with throughout the execution of the search.

The user first uploads the respective COVID-19 dataset into the application. From the below fundamentals of executing the modules the dataset file is set to undergo preprocessing of the data. The feature engineering is operated from the preprocessed clinical reports, various features are extracted as per the semantics and are converted into probabilities values. The following sequence is carried out in accordance to the user option which enables the key to regulate statistical values.

The result report is show in the comparative graph as the final output.

The following diagram depicts the entire scenario described above in further detail:

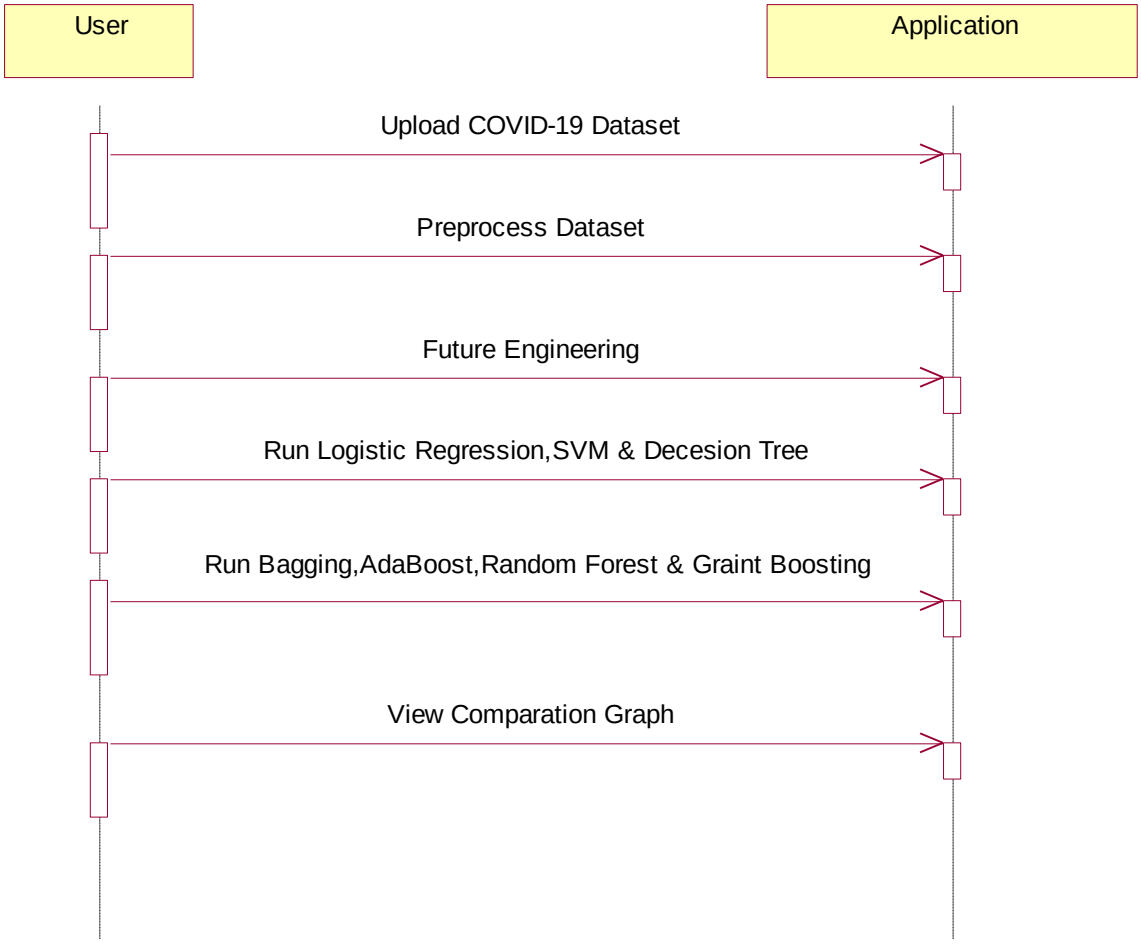


Figure 5.7 Sequence Diagram

COLLABORATION DIAGRAM

A collaboration diagram is an illustration of the relationships and interactions among software objects in the UML. These diagrams can be used to portray the dynamic behavior of a particular use case and define the role of each object.

Collaboration diagrams are created by first identifying the structural elements required to carry out the functionality of an interaction. These diagrams are used to visualize the structural organization of objects and their interactions. Sequence diagrams, on the other hand, focus on the order of messages that flow between objects. However, in most scenarios, a single figure is not sufficient in describing the behavior of a system and both figures are required.

A model is then built using the relationships between those elements. Several vendors offer software for creating and editing collaboration diagrams.

A collaboration diagram resembles a flowchart that portrays the roles, functionality and behavior of individual objects as well as the overall operation of the system in real time.

The below diagram depicts the relationship between elements of our model:

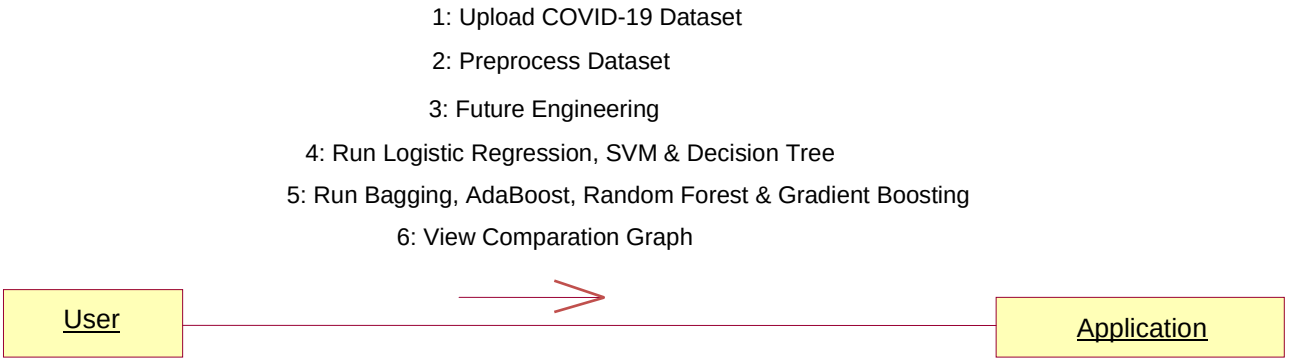


Figure 5.8: Collaboration diagram

6. PROJECT CODING

Project Coding is the process of designing and building an executable computer program to accomplish a specific computing result or to churn out a particular prototype or product. Programming involves tasks such as: analysis, generating algorithms, profiling algorithms' accuracy and resource consumption, and the implementation of algorithms in a chosen programming language (commonly referred to as coding). The source code of a program is written in one or more languages that are intelligible to programmers, rather than machine code, which is directly executed by the central processing unit. The purpose of programming is to find a sequence of instructions that will automate the performance of a task (which can be as complex as an operating system) on a computer, often for solving a given problem. Proficient programming thus often requires expertise in several different subjects, including knowledge of the application domain, specialized algorithms, and formal logic.

+

6.1 CODE TEMPLATES

1) This template contains the overview of multiple Django Server tools and packages that we have imported for authentication, content type verification, message display, session checking, security and context processors for debugging, requests etc.

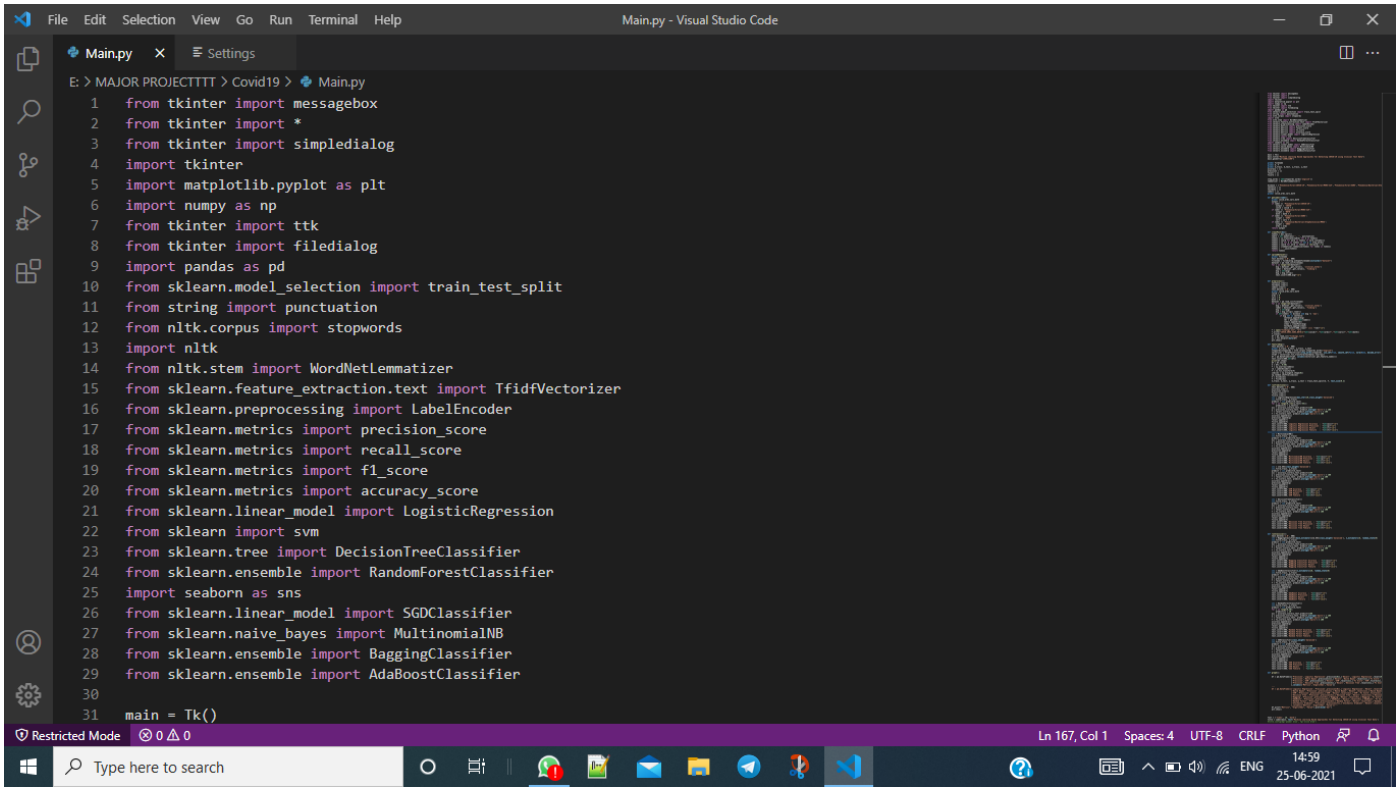


Figure 6.1 Code Template (1)

2) This template displays the details to access the uploading of the dataset file from local computer.

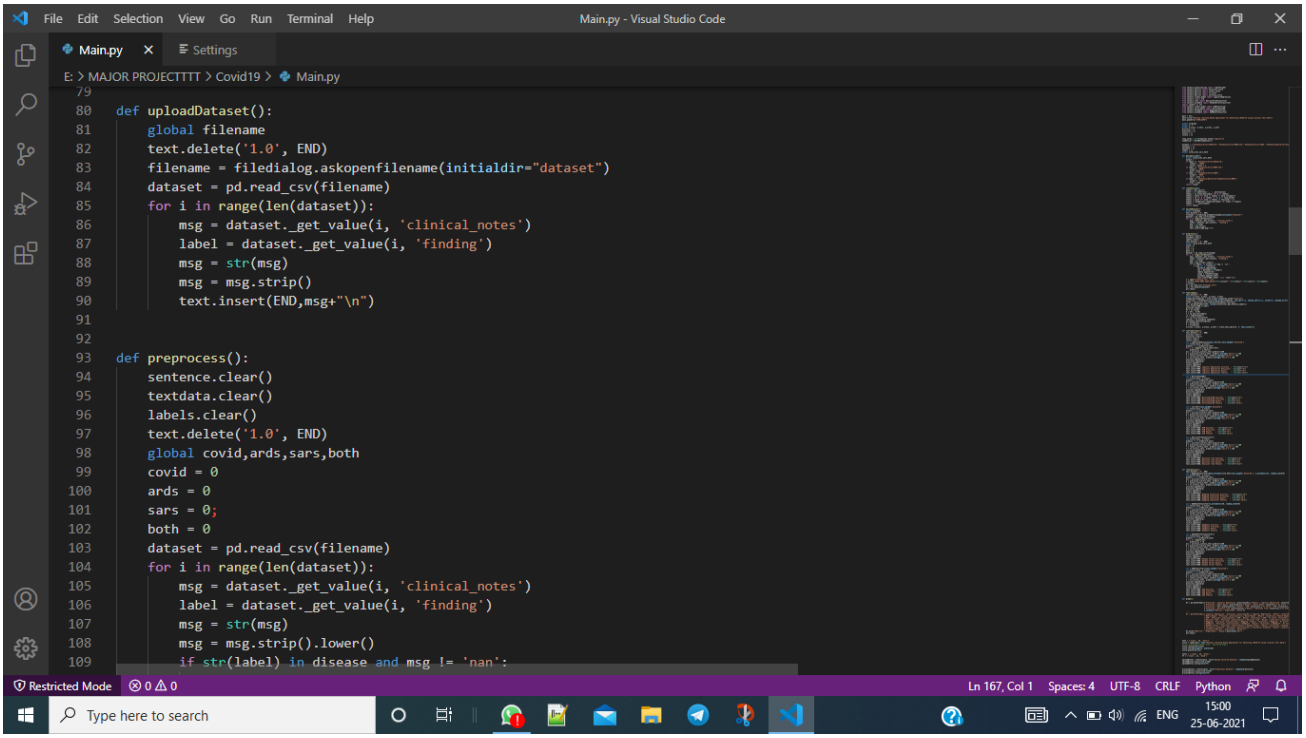


Figure 6.2 Code Template(2)

6.2 OUTLINE FOR VARIOUS FILES

We used Python programming to implement our project. We also used HTML and CSS to develop our webpage. This dataset contains more than 30 columns, but we are extracting two column values such as ‘clinical notes’ and ‘finding’. clinical_notes column contains medical text data, and this text data is preprocess using NLTK library to remove stop words, special symbols and then apply lemmatize to remove ‘Ing, tion etc.’ from text. After preprocess text we will apply TF-IDF to extract top 40 features from dataset. Below is the dataset screen shots and this dataset saved inside ‘dataset’ folder.

6.3 CLASS WITH FUNCTIONALITY

There are multiple classes in our code, some of which are:

1) Upload Covid-19 Dataset’ button and then upload dataset.

selecting and uploading ‘dataset.csv’ file and then click on ‘Open’ button to load dataset.

2) We extract all text data from dataset and now in above screen text in first sentence we have ‘on’ stop words and many numbers of numerical values and to remove those stop words and

to clean data.

3) Preprocess all stop words removed out and in above ‘on’ stop word removed out graph showing count/finding of each label and now close above graph and then click on ‘Feature Engineering’ button to apply TF-IDF on above text data and to get below features.
4) All text data converted to above TF-IDF features and now click on ‘Run Logistic Regression, Naive Bayes, SVM & Decision Tree’ to run all traditional algorithms on features data and to calculate accuracy.
5) Displaying accuracy, precision, recall and F Score for each algorithm and now click on ‘Run Bagging, Adaboost, Random Forest & Gradient Boosting’ button to calculate accuracy of classical algorithms.
6) Showing classical algorithms accuracy and other metrics values and now click on ‘Comparative Analysis Graph’ button to get graph we can see accuracy, precision, recall and f score for each algorithm in group bar chart and in above graph x-axis represents algorithm name y-axis represents values.

6.4 METHODS INPUT AND OUTPUT PARAMETERS We implemented multiple methods, few of which are:

1. Upload_Dataset()
2. getLabel()
3. clean_Post()
4. preprocess()
5. featureEng()
6. run_classical()
7. run_traditional(), etc.

7. PROJECT TESTING

Project Testing is a method to check whether the actual software product matches expected requirements and to ensure that software product is Defect free. It involves execution of software/system components using manual or automated tools to evaluate one or more properties of interest. The purpose of software testing is to identify errors, gaps or missing requirements in contrast to actual requirements.

Some prefer saying Software testing as a White Box and Black Box Testing. In simple terms, Software Testing means the Verification of Application Under Test (AUT). This tutorial introduces testing software to the audience and justifies its importance.

Project testing is important because, if there are any bugs or errors in the software, it can be identified early and can be solved before delivery of the software product. Properly tested software product ensures reliability, security and high performance which further results in time saving, cost effectiveness and customer satisfaction.

Typically Testing is classified into three categories.

- Functional Testing
- Non-Functional Testing or Performance Testing
- Maintenance (Regression and Maintenance)

7.1 VARIOUS TEST CASES

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub-assemblies, assemblies and/or a finished product. It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of tests. Each test type addresses a specific testing requirement.

We have performed multiple tests under the broad categorization of white-box and black-box testing which further include unit, integration, boundary value and statement covering etc.

Test Id	Test Name	Input	Output	Expected Result	Status
1	Covid-19 Collect dataset	File	Data stored	Data stored	PASS
	Master browsing data	data	Data is not Present in directory	Data stored	FAIL
2	Analysis Algorithm	File Name with Uploaded data	File Attacked or Safe	Attacked or Safe	PASS
4	Comparative Analysis Graph'	No Inputs	No Spam to Compare File	Comparative Analysis Graph'	PASS

Table 6. Test Cases Tabulation

7.2 BLACK-BOX TESTING

Black Box Testing is a software testing method in which the functionalities of software applications are tested without having knowledge of internal code structure, implementation details and internal paths. Black Box Testing mainly focuses on input and output of software applications and it is entirely based on software requirements and specifications. It is also known as Behavioral Testing.

Here are the generic steps followed to carry out any type of Black Box Testing.

- Initially, the requirements and specifications of the system are examined.
- Tester chooses valid inputs (positive test scenario) to check whether SUT processes them correctly. Also, some invalid inputs (negative test scenario) are chosen to verify that the SUT is able to detect them.

- Tester determines expected outputs for all those inputs.
- Software tester constructs test cases with the selected inputs.
- The test cases are executed.
- Software tester compares the actual outputs with the expected outputs.
- Defects if any are fixed and re-tested.

Types of Black Box Testing

There are many types of Black Box Testing, but the following are the prominent ones -

- Functional testing - This black box testing type is related to the functional requirements of a system; it is done by software testers.
- Non-functional testing - This type of black box testing is not related to testing of specific functionality, but non-functional requirements such as performance, scalability, usability.
- Regression testing - Regression Testing is done after code fixes, upgrades or any other system maintenance to check the new code has not affected the existing code.

Black Box Testing Techniques

Following is the prominent Test Strategy amongst the many used in Black box Testing.

- Equivalence Class Testing: It is used to minimize the number of possible test cases to an optimum level while maintains reasonable test coverage.
- Boundary Value Testing: Boundary value testing is focused on the values at boundaries. This technique determines whether a certain range of values are acceptable by the system or not. It is very useful in reducing the number of test cases. It is most suitable for the systems where an input is within certain ranges.
- Decision Table Testing: A decision table puts causes and their effects in a matrix. There is a unique combination in each column.

7.3 WHITE-BOX TESTING

White Box Testing is software testing technique in which internal structure, design and coding of software are tested to verify flow of input-output and to improve design, usability and security. In white box testing, code is visible to testers, so it is also called Clear box testing, Open box testing, Transparent box testing, Code-based testing and Glass box testing.

It is one of two parts of the Box Testing approach to software testing. Its counterpart, Blackbox testing, involves testing from an external or end-user type perspective. On the other hand, White box testing in software engineering is based on the inner workings of an application and revolves around internal testing.

The term "Whitebox" was used because of the see-through box concept. The clear box or White Box name symbolizes the ability to see through the software's outer shell (or "box") into its inner workings. Likewise, the "black box" in "Black Box Testing" symbolizes not being able

to see the inner workings of the software so that only the end-user experience can be tested.

White box testing involves the testing of the software code for the following:

- Internal security holes
- Broken or poorly structured paths in the coding processes.
- The flow of specific inputs through the code
- Expected output.
- The functionality of conditional loops
- Testing of each statement, object, and function on an individual basis

The testing can be done at system, integration and unit levels of software development. One of the basic goals of Whitebox testing is to verify a working flow for an application. It involves testing a series of predefined inputs against expected or desired outputs so that when a specific input does not result in the expected output, you have encountered a bug.

To give you a simplified explanation of white box testing, we have divided it into two basic steps. This is what we do when testing an application using the white box testing technique:

STEP 1) UNDERSTAND THE SOURCE CODE

The first thing a tester will often do is learn and understand the source code of the application. Since white box testing involves the testing of the inner workings of an application, the tester must be very knowledgeable in the programming languages used in the applications they are testing. Also, the testing person must be highly aware of secure coding practices. Security is often one of the primary objectives of testing software. The tester should be able to find security issues and prevent.

attacks from hackers and naive users who might inject malicious code into the application either knowingly or unknowingly.

Step 2) CREATE TEST CASES AND EXECUTE

The second basic step to white box testing involves testing the application's source code for proper flow and structure. One way is by writing more code to test the application's source code. The tester will develop little tests for each process or series of processes in the application. This method requires that the tester must have intimate knowledge of the code and is often done by the developer.

The goal of Whitebox testing in software engineering is to verify all the decision branches, loops, statements in the code.

A major White box testing technique is Code Coverage analysis. Code Coverage analysis eliminates gaps in a Test Case suite. It identifies areas of a program that are not exercised by a set of test cases. Once gaps are identified, you create test cases to verify untested parts of the code, thereby increasing the quality of the software product.

There are automated tools available to perform Code coverage analysis. Below are a few coverage analysis techniques a box tester can use:

Statement Coverage: - This technique requires every possible statement in the code to be tested at least once during the testing process of software engineering.

Branch Coverage - This technique checks every possible path (if-else and other conditional loops) of a software application.

Apart from above, there are numerous coverage types such as Condition Coverage, Multiple Condition Coverage, Path Coverage, Function Coverage etc. Each technique has its own merits and attempts to test (cover) all parts of software code. Using Statement and Branch coverage you generally attain 80-90% code coverage which is sufficient.

8. OUTPUT SCREENS

An output screen is a device used to display output. An output screen could be a separate monitor, or another display device used only to display the output being received from the computer or other devices.

Here, in the screen prints given below, we can see that the user interface screens consist of the home page which describes the purpose of the portal, and the public access screen which allows users to upload the credentials of the child that they found in order to check whether the child exists in the repository or not.

8.1 USER INTERFACE

1) HOME PAGE

This homepage displays the representation of functionalities of the model.

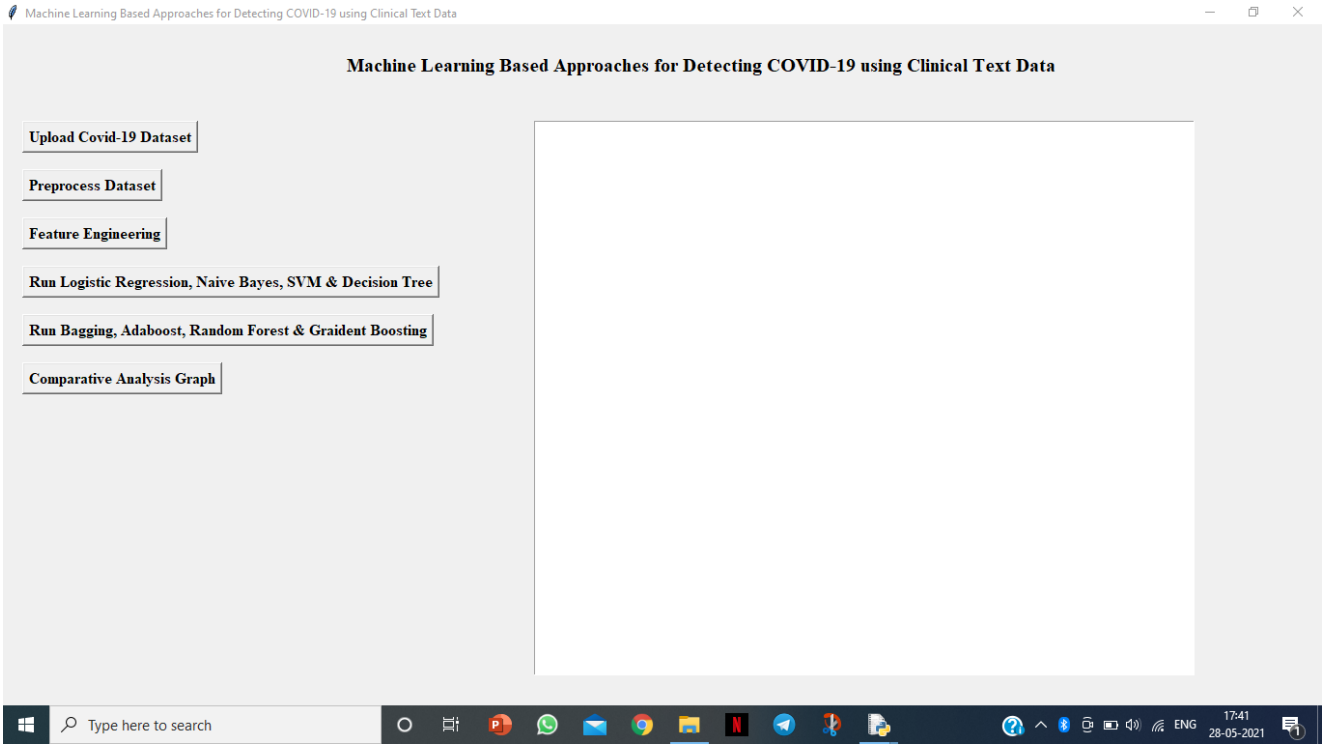


Figure 8.1 User Interface

8 .2 OUTPUT SCREENS

1) The homepage enables the user to ‘Upload Covid-19 Dataset’ button and then upload dataset selecting and uploading ‘dataset.csv’ file and then click on ‘Open’ button to load dataset.

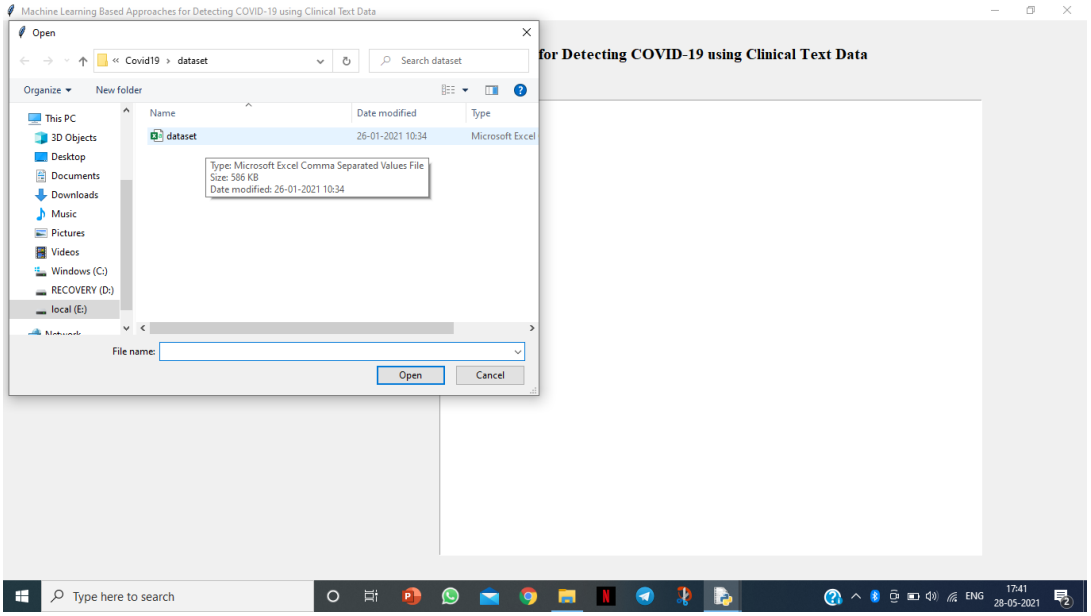


Fig 8.2 Uploading dataset Screen.

2) The uploaded data results into we can see accuracy, precision, recall and f score for each algorithm in group bar chart and in above graph x-axis represents algorithm name y-axis represents values.

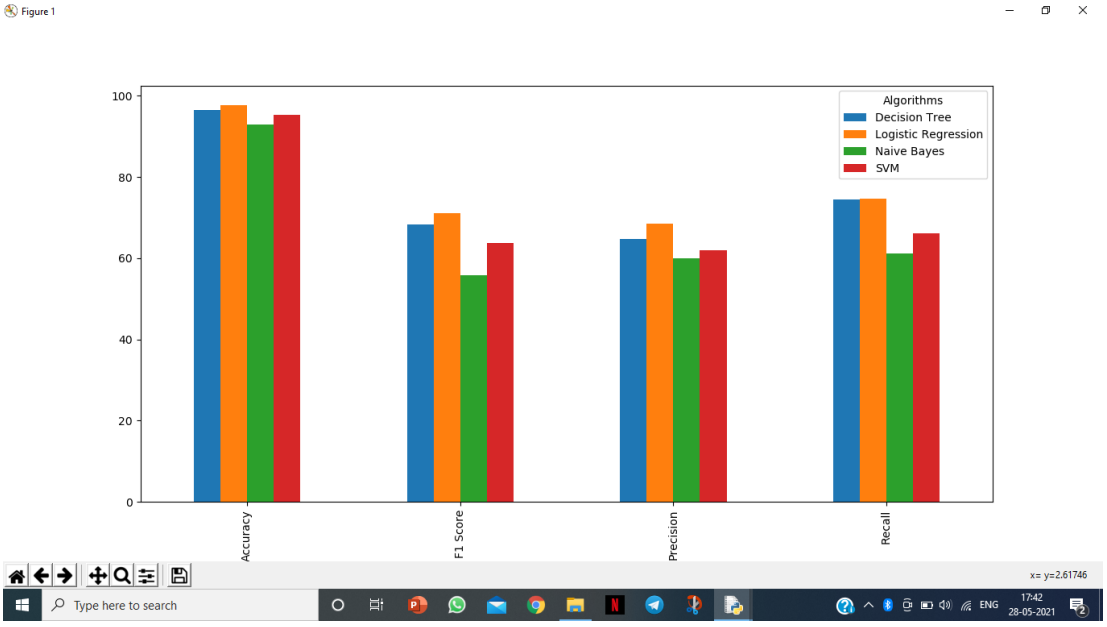


Figure 8.3 Results Screen

9. EXPERIMENTAL RESULT

1)The uploaded data is screening all text data converted to above TF-IDF features and now click on ‘Run Logistic Regression, Naive Bayes, SVM & Decision Tree’ to run all traditional algorithms on features data and to calculate accuracy displaying accuracy, precision, recall and

F Score for each algorithm.

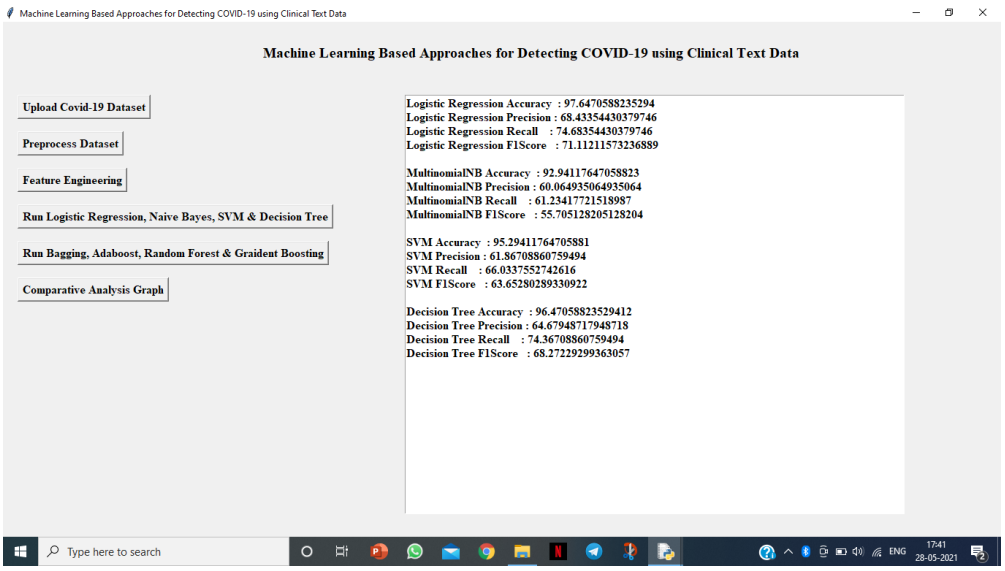


Figure 9.1 Traditional algorithms result

2)Now click on ‘Run Bagging, Adaboost, Random Forest & Gradient Boosting’ button to calculate accuracy of classical algorithms showing classical algorithms accuracy and other metrics values.

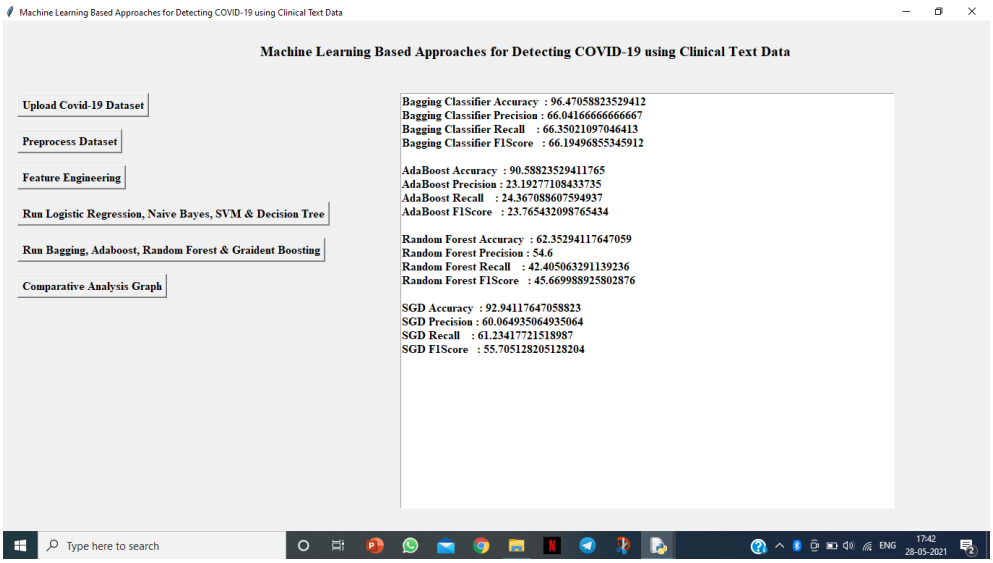


Figure 9.2 Classical algorithms result

3)Now click on ‘Comparative Analysis Graph’ button to get below graph we can see accuracy, precision, recall and f score for each algorithm in group bar chart and in above graph x-axis represents algorithm name y-axis represents values.

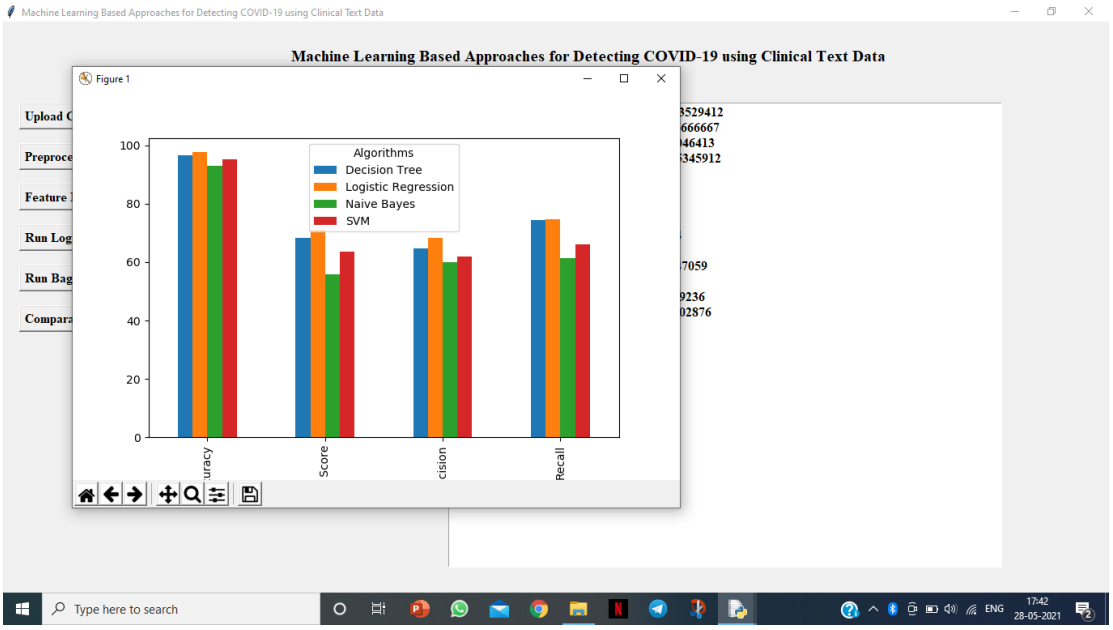


Figure 9.3 Activity Tracking

10. CONCLUSION AND FUTURE ENHANCEMENT

COVID-19 has shocked the world due to its non-availability of vaccine or drug. Various researchers are working for conquering this deadly virus. We used 212 clinical reports which are labelled in four classes namely COVID, SARS, ARDS and both (COVID, ARDS).

Various features like TF/IDF, bag of words is being extracted from these clinical reports. The machine learning algorithms are used for classifying clinical reports into four different classes. After performing classification, it was revealed that logistic regression and multinomial Naïve Bayesian classifier gives excellent results by having 94% precision, 96% recall, 95% f1 score and accuracy 96.2%.

To get the real accuracy of the model we experimented it in two stages. In the first stage, we took 75% of the available data and it shows less accuracy as compared to the stage in which whole data was used for experimentation. So we can conclude that if more data is supplied to these algorithms, there are chances of improvement in performance. As we are facing a severe challenge in tackling the deadly virus, our work will somehow help the community by analyzing the clinical reports and take necessary actions.

Various other machine learning algorithms that showed better results were random forest, stochastic gradient boosting, decision trees and boosting. The efficiency of models can be improved by increasing the amount of data. Also, the disease can be classified on the gender-based such that we can get information about whether male are affected more or females. More feature engineering is needed for better results and deep learning approach can be used in future.

11. REFERENCES

[1]. Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, Hu Y, Tao ZW, Tian JH, Pei YY, Yuan ML, Zhang YL, Dai FH, Liu Y, Wang QM, Zheng JJ, Xu L, Holmes EC, Zhang YZ (2020) A new coronavirus associated with human respiratory disease in china.

[2]. Medscape Medical News, The WHO declares public health emergency for novel coronavirus (2020) https://www.medscape. com/viewarticle/924596

[3]. Chen N, Zhou M, Dong X, Qu J, Gong F, Han Y, Qiu Y, Wang J, Liu Y, Wei Y, Xia J, Yu T, Zhang X, Zhang L (2020) Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study.

[4]. World health organization: https://www.who.int/new-room/g-adetail/ q-a-corronaviruses#:/text=symptoms. Accessed 10 Apr 2020

[5]. Wikipedia coronavirus Pandemic data: https://en.m.wikipedia. org/wiki/Template:2019%E2%80%9320_coronavirus_pan demic_data. Accessed 10 Apr 2020

[6]. Khanday, A.M.U.D., Amin, A., Manzoor, I., & Bashir, R., ‘‘Face Recognition Techniques: A Critical Review’’ 2018

[7]. Kumar A, Dabas V, Hooda P (2018) Text classification algorithms for mining unstructured data: a SWOT analysis. Int J Inf Technol. https://doi.org/10.1007/s41870-017-0072-1

[8]. Verma P, Khanday AMUD, Rabani ST, Mir MH, Jamwal S (2019) Twitter Sentiment Analysis on Indian Government Project using R. Int J Recent Tech Eng. https://doi.org/10.35940/ijrte. C6612.098319

[9]. Chakraborti S, Choudhary A, Singh A et al (2018) A machine learning based method to detect epilepsy. Int J Inf Technol 10:257–263. https://doi.org/10.1007/s41870-018-0088-1

[10]. Sarwar A, Ali M, Manhas J et al (2018) Diagnosis of diabetes type-II using hybrid machine learning based ensemble model. Int J Inf Technol. https://doi.org/10.1007/s41870-018-0270-5

[11]. Bullock J, Luccioni A, Pham KH, Lam CSN, Luengo-Oroz M (2020) Mapping the Landscape of artificial intelligence applications against COVID-19

[12]. Wang L, Wong A (2020) COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 Cases from chest radiography images. https://arxiv.org/abs/2003.09871

[13]. Yan L, Zhang H-T, Xiao Y, Wang M, Sun C, Liang J, Li S, Zhang M, Guo Y, Xiao Y, Tang X, Cao H, Tan X, Huang N, Amd A, Luo BJ, Cao Z, Xu H, Yuan Y (2020) Prediction of criticality in patients with severe covid-19 Infection using three clinical features: a machine learning-based prognostic model with clinical data in Wuhan. medRxiv. https://doi.org/10.1101/2020.02.27. 20028027

[14]. Jiang X, Coffee M, Bari A, Wang J, Jiang X, Huang J, Shi J, Dai J, Cai J, Zhang T, Wu Z, He G, Huang Y (2020) Towards an artificial intelligence framework for data-driven prediction of coronavirus clinical severity. Compu Mater Contin 63(1):537–551

[15]. Description of Logistic Regression Algorithm. https://machine learningmastery.com/logistic-regression-for-machine-learning/. Accessed 15 May 2019

[16]. Description of Multinomial Naı¨ve Bayes Algorithm https://www. 3pillarglobal.com/insights/document-classification-using-multi nomial-naive-bayes-classifier. Accessed 15 May 2019

[17]. Khanday AMUD, Khan QR, Rabani ST. SVMBPI: support vector machine-based propaganda identification. SN Appl. Sci. (accepted)

[18]. Description of Decision Tree Algorithm: https://dataspirant.com/ 2017/01/30/how_decision_tree_algorithm_works/. Accessed 10 July 2019

[19]. Description of Boosting Algorithm: https://towardsdatascience. com/boosting. Accessed 10 July 2019

[20]. Description of Adaboost Algorithm: https://towardsdatascience. com/boosting-algorithm-adaboost-b673719ee60c. Accessed 10 July 2019

12. PUBLICATIONS

CONFERENCE:

International Conference on “Innovations in Computers Networks, Computational Intelligence and IoT” (ICICCI – 21)

Paper ID: ICICCI – 21 – 0098

13. STUDENT PROFILES

Venkannagari Athiksha is a Bachelor of Technology student at St. Martin's Engineering College studying Computer Science and Engineering. She finished her schooling in Wisdom High school, and intermediate education from Narayana Junior college. Python and Java are among her technical skills. Her participations include: a National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" from the 10th to the 12th of December 2020. She spends her free time taking online certification courses related to her field of study as well as personal interests from platforms such as Coursera and does take part in coding learning platforms such as hacker rank while she accomplished: Silver badge in hacker rank python, golden badge in hacker rank problem solving.

Diddi Chitra is currently pursuing her Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. She completed her intermediate from Sri Chaitanya Junior College and 10th class from Vijaya high School. Her technical skills include C, Java and Python. She also has a basic understanding of C++. Her participations include: a National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" from the 10th to the 12th of December 2020, Internet of things workshop in BITS Pilani Hyderabad Campus. In accordance with academic courses, she has completed few online courses on platforms like Coursera (Basics on python, Python data structures.)

G. Venkata Snehith is currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. He has completed his Secondary Education from St. Triveni Talent School and Higher Secondary Education from Narayana Junior College. His technical skills include Java, C, Python. He attended a National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" from the 10th to the 12th of December 2020, participated in "TECHNOVATION-2018" organized by our college. He has completed online courses on Coursera platform in HTML, Python, Leadership and Emotional Intelligence. Apart from his academic interests, he's a keen badminton player and participated in few college level badminton championships. He participated in Inter College Badminton Competition and stood as Runner in college level Badminton Competition.

A. Akhil Reddy is currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. He has completed his Secondary Education from Krishnaveni Talent School and Higher Secondary Education from Narayana Junior College. His technical skills include Java, C, Python. He attended E-summit, Entrepreneurship carnival, which was hosted by EDC – MLRIT in association with Nucleus Tech and SUMVN in the year 2017. He has completed online courses (Data Science Math Skill, Leadership and Emotional Intelligence, Managing Project Risks and Fundamentals, Python) from Coursera and CursaApp.

10. APPENDICES

Appendix A: User Requirements Questionnaire

User Requirements

Questionnaire

This research will be used for academic purpose only. Its main objective is to collect the user requirements to create a child tracing prototype. Kindly provide your honest answers inthe following questions. Please note that your responses will be treated as private and confidential.

1. COVID-19 detection is thoroughly inspected in this model.

- ϒ Strongly Agree
- ϒ Agree
- ϒ Neutral
- ϒ Disagree
- ϒ Strongly Disagree

2. The process of collecting data is cost - efficient and time -effective.

- ϒ Strongly Agree
- ϒ Agree
- ϒ Neutral
- ϒ Disagree
- ϒ Strongly Disagree

3. The preprocessing and refining of the data enable the smooth classification of the data in result provide with accurate reports.

- ϒ Strongly Agree
- ϒ Agree
- ϒ Neutral
- ϒ Disagree
- ϒ Strongly Disagree

4. This model is highly efficient for updating the overall effect of COVID-19 globally.

ϒ Strongly Agree
ϒ Agree
ϒ Neutral
ϒ Disagree
ϒ Strongly Disagree

5. I believe that the scope of the model can be increased by implementing Neural technology.

ϒ Strongly Agree
ϒ Agree
ϒ Neutral
ϒ Disagree
ϒ Strongly Disagree

6. It provides the user with the much-needed accurate reports of clinical reports.

ϒ Strongly Agree
ϒ Agree
ϒ Neutral
ϒ Disagree
ϒ Strongly Disagree

Appendix B: System Usability Questionnaire

This research will be used for academic purpose only. Its main objective is to find out users' experience in using the child tracing prototype. Kindly provide your honest opinion on the same. Please note that your responses will be treated as private and confidential.

System Usability Scale

Kindly rate the Data classification prototype about the following:

1. The user interface is very user friendly.

ϒ Strongly Agree

ϒ Agree

ϒ Neutral

ϒ Disagree

ϒ Strongly Disagree

2. I can use this prototype with the minimum training.

ϒ Strongly Agree

ϒ Agree

ϒ Neutral

ϒ Disagree

ϒ Strongly Disagree

3. Updating the global COVID-19 reports using this system will take a shorter duration as compared to the current methods.

ϒ Strongly Agree

ϒ Agree

ϒ Neutral

ϒ Disagree

ϒ Strongly Disagree

4. This question is practical and aims at testing the accuracy of the prototype. Kindly provide a list of different clinical text reports to the researcher. After the researcher inputs them into the system, try classification them and note down your findings (Among them, how many were most accurate?)

……..

5. The system provides a convenient way of resulting reports with accurate reports.

ϒ Strongly Agree
ϒ Agree
ϒ Neutral
ϒ Disagree
ϒ Strongly Disagree

6. This provides as a data interpretating model for quick analysis of data with accurate results.

ϒ Strongly Agree
ϒ Agree
ϒ Neutral
ϒ Disagree
ϒ Strongly Disagree

7. How likely are you to recommend this system to the other users?

ϒ Very Likely
ϒ Likely
ϒ Neutral
ϒ Not Likely
ϒ Not Likely at All

8. Any Comments

……
……
……
……
……
……
……
……
……
……
……
……

Appendix C: Interview Questions

Interview Questions

This research will be used for academic purpose only. Its main objective is to find out users' experience in using the child tracing prototype. Kindly provide your honest opinion on the same. Please note that your responses will be treated as private and confidential.

Interviewee: ……. **Location:**…….

Medium: ……. **Date:** ……

1. What do u think are the consequences of COVID-19 virus?

   ……
   ……
   ……
   ……
   ……
   ……
   ……
   ……
   ……

2. How can we lead to the analysis of aftereffects of COVID effected people?

   ……
   ……
   ……
   ……
   ……
   ……
   ……
   ……
   ……
   ……

3. What is the accurate percentage effect of COVID -19 onto countries?

……
……
……
……
……
……
……
……

4. How do you handle huge data sets while collecting the clinical text reports?

……
……
……
……
……
……
……
……

5. How do u regulate the type of COVID-19 virus and their overall effect on specific domain?

……
……
……
……
……
……
……
……

6. What are the challenges faced while treating the effected patients with other simultaneous diseases?

……
……
……
……
……
……
……
……

7. In your opinion, what do you think needs to be improved in the process of:

    a. In increasing the scope of collecting versatile clinical text reports.

……
……
……
……
……
……
……
……
……
……
……

    b. Refining the data while being consistent with the accuracy?

……
……
……
……
……
……
……
……
……
……

A

PROJECT REPORT

On

SUSPICIOUS ACTIVITY DETECTION

Submitted by

1)K. Vinay Kumar (17K81A05L6) 2) S. Akhila (17K81A05N4)

3)T. Sai Vardhan (17K81A05N6) 4) A. Hima Varsha (17K81A05J1)

In partial fulfillment for the award of the

degree of

BACHELOR OF TECHNOLOGY

IN

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Under the Guidance of

Dr. N. Satheesh

B.E, M.E, Ph.D.

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ST. MARTIN'S ENGINEERING COLLEGE
An Autonomous Institute

Dhulapally, Secunderabad – 500 100

JUNE 2021

BONAFIDE CERTIFICATE

This is to certify that the project entitled **Suspicious Activity Detection**, is being submitted by **T. Sai Vardhan (17K81A05N6), A. Hima Varsha (17K8A05J1), K. Vinay Kumar (17K81A05L6), S. Akhila (17K81A05N4)** in partial fulfillment of the requirement for the award of the degree of **BACHELOR OF TECHNOLOGY IN COMPUTER SCIENCE AND ENGINEERING** is recorded of bonafide work carried out by them. The result embodied in this report have been verified and found satisfactory.

Guide Signature
Dr. N. Satheesh
Department of CSE

Head of the Department
Dr. M. NARAYANAN
Department of CSE

Internal Examiner

External Examiner

Place:

Date:

DECLARATION

We, the student of **Bachelor of Technology** in Department of Computer Science and Engineering', session: 2017 – 2021, St. Martin's Engineering College, Dhulapally, Kompally, Secunderabad, hereby declare that work presented in this Project Work entitled **Suspicious Activity Detection** is the outcome of our own bonafide work and is correct to the best of our knowledge and this work has been undertaken taking care of Engineering Ethics. This result embodied in this project report has not been submitted in any university for award of any degree.

T. Sai Vardhan (17K81A05N6)
A. Hima Varsha (17K81A05J1)
K. Vinay Kumar (17K81A05L6)
S. Akhila (17K81A05N4)

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible and whose encouragements and guidance have crowded effects with success.

We extended our deep sense of gratitude to Principal**, Dr. P. SANTOSH KUMAR PATRA**, St. Martin's Engineering College, Dhulapally, for permitting us to undertake this project.

We are also thankful to **Dr. M.NARAYANAN**, Head of the Department, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

We would like to thank **Dr. T. POONGOTHAI,** Department of Computer Science and Engineering (AI & ML) for her encouragement and insightful comments and as well as our project coordinators **Dr. B.RAJALINGAM**, Associate Professor and **Dr. R. SANTHOSHKUMAR**, Associate Professor, in Department of Computer Science and Engineering for their valuable support.

We would like to express our sincere gratitude and indebtedness to our project supervisor Dr. N. Satheesh, Professor, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

Finally, we express thanks to all those who have helped us successfully to completing this project. Furthermore, we would like to thank our family and friends for their moral support and encouragement.

We express thanks to all those who have helped us in successfully completing the project.

T. Sai Vardhan 17K81A05N6
A. Hima Varsha 17K81A05J1
K. Vinay Kumar 17K81A05L6
S. Akhila 17K81A05N4

ABSTRACT

With the increasing in the number of anti-social activities that have been taking place, security has been given utmost importance lately. Many Organizations have installed CCTVs for constant Monitoring of people and their interactions. For a developed Country with a population of 64 million, every person is captured by a camera 30 times a day. A lot of video data generated and stored for a certain time duration. A 704x576 resolution image recorded at 25fps will generate roughly 20GB per day. Constant Monitoring of data by humans to judge if the events are abnormal is near impossible task as requires a workforce and their constant attention. This creates a need to automate the same. Also, there is need to show in which frame and which part of it contain the unusual activity which aid the faster judgment of the unusual activity being abnormal. This is done by converting video into frames and analysing the persons and their activities from the processed frame. Machine Learning and Deep Learning Algorithms and techniques support us in a wide accept to make Possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF OUPUT SCREENS

LIST OF ABBREVATIONS

AVI	Audio Video Interlace
BMP	Bitmap
CPU	Central Processing Unit
GB	Giga Bytes
GUI	Graphical User Interface
CNN	Convolutional Neural Network
HTTP	Hyper Text Mark Up Language
CSS	Cascading Style Sheets
DFD	Data Flow Diagram
ER	Entity-Relationship
OS	Operating System

GLOSSARY OF TERMS

TERM	MEANING
Machine Learning	Machine learning is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention.
Deep Learning	Deep learning is part of a broader family of machine learning methods based on artificial neural networks with representation learning.
CNN	A convolutional neural network (CNN, or ConvNet) is a class of deep neural network, most commonly applied to analyse visual imagery.
Suspicious Activity	Suspicious behaviour or activity can be any action that is out of place and does not fit into the usual day-to-day activity of our campus community.
Background Extraction	Background extraction is a fast and efficient moving object segmentation algorithm.
Foreground Object Extraction	Foreground object extraction from the video is an initial and important step of suspicious human activity recognition.

1.INTRODUCTION

1.1 PROJECT OVERVIEW:

Human face and human behavioural pattern play an important role in person identification. Visual information is a key source for such identifications. Surveillance videos provide such visual information which can be viewed as live videos, or it can be played back for future references. The recent trend of 'automation' has its impact even in the field of video analytics. Video analytics can be used for a wide variety of applications like motion detection, human activity prediction, person identification, abnormal activity recognition, vehicle counting, people counting at crowded places, etc. In this domain, the two factors which are used for person identification are technically termed as face recognition and gait recognition, respectively. Among these two techniques, face recognition is more versatile for automated person identification through surveillance videos. Face recognition can be used to predict the orientation of a person's head, which in turn will help to predict a person's behaviour. Motion recognition with face recognition is very useful in many applications such as verification of a person, identification of a person and detecting presence or absence of a person at a specific place and time. In addition, human interactions such as subtle contact among two individuals, head motion detection, hand gesture recognition and estimation are used to devise a system that can identify and recognize suspicious behaviour among pupil in an examination hall successfully. This paper provides a methodology for suspicious human activity detection through face recognition.

Video processing is used in two main domains such as security and research. Such a technology uses intelligent algorithms to monitor live videos. Computational complexities and time complexities are some of the key factors while designing a real-time system. The system which uses an algorithm with a relatively lower time complexity, using less hardware resources and which produces good results will be more useful for time-critical applications like bank robbery detection, patient monitoring system, detecting and reporting suspicious activities at the railway station, etc

Manual monitoring of exam hall through invigilators and manual monitoring of exam hall through surveillance videos is performed throughout the world. Monitoring an examination hall is a very challenging task in terms of manpower. Manual monitoring of examination halls may be prone to error during human supervision. Such a system when implemented as an 'automatic suspicious activity detection system' will not only help in detecting suspicious activities but also helps in minimizing such activities. Moreover, the probability of error will be much lesser. This system will serve as a useful surveillance system for educational institution

This paper describes a technology in which real time videos are analysed and are used for human activity analysis in an examination hall, thus helping to classify whether the particular person's activity is suspicious or not. The system developed identifies abnormal head motions, thereby prohibiting copying. It also identifies a student moving out of his place or swapping his position with another student. Finally, the system detects contact between students and hence prevents passing incriminating material among students. In our research, we have contributed upon a system that will intellectually process live video of examination halls with students and classify their activities as suspicious or not. This research proposes an intelligent algorithm that can monitor and analyse the activities of students in an examination hall and can alert the educational institute's administration on account of any malpractices/suspicious activities.

The Suspicious Human Activity Detection system aims to identify the students who indulge in malpractices/suspicious activities during the course of an examination. The system automatically detects suspicious activities and alerts administration.

1.2. PROJECT OBJECTIVE:

Suspicious Human Activity Recognition from Video Surveillance is an active research area of image processing and computer vision which involves recognition of human activity and categorizes them into normal and abnormal activities. Abnormal activities are the unusual or suspicious activities rarely performed by the human at public places, such as left luggage for Explosive attacks, theft, running crowd, fights and attacks, vandalism and crossing borders. Normal activities are the usual activities performed by the human at public places, such as running, boxing, jogging and walking, hand waving and clapping. Now-a-days, use of video surveillance is increasing day by day to monitor the human activity which prevents the
suspicious activities of the human.

An important chore of the video surveillance is to analyze the captured video frames for identifying unusual or suspicious activities in security-sensitive region of any country such as banks, parking lots, department stores, government buildings, prisons, military bases. Video Surveillance captures images of moving objects in order to watch assault and fraud, comings and goings, prevent theft, as well as manage crowd movements and incidents. In public places, human performs normal (usual) and abnormal (suspicious or unusual) activities. Normal activities are the usual activities that are not dangerous for the human world, but abnormal activities may be dangerous for all over the world. Therefore, an intelligent surveillance system is required that can recognize all the activities and identify the more dangerous and suspicious activities performed by a human being.

The scope of the project is to describe a technology in which real time videos are analyzed and are used for human activity analysis in real time scenario, thus helping to classify whether the particular person's activity is suspicious or not. The system which uses an algorithm with a relatively lower time complexity, using less hardware resources and which produces accurate results will be more useful for time-critical applications like bank robbery detection, patient monitoring system, detecting and reporting suspicious activities at the railway station, etc.

The Goal of the video surveillance is to develop an intelligent video surveillance to replace the traditional passive video surveillance so that abnormal activities performed by human being can be captured.

After analyzing, a frame with highest probability which contain a suspicious activity will be displayed.

1.3 ORGANIZATION OF CHAPTERS

This documentation consists of 10 different chapter and they are:

1. Introduction – This chapter covers the overview of our project and its objectives.

2. Literature Survey – This includes the details of our survey.

3. Software and Hardware Requirements – We specify our software and hardware requirements here.

4. Software Development Analysis – This section includes the problem definition and details of the modules we used in our project.

5. Project System Design – This chapter includes the design part of our project which includes uml diagrams.

6. Project Coding – This section contains the details of our project code.

7. Project Testing – The details of test cases and testing are included in this chapter.

8. Output Screens – This contains the screenshots of how our project looks like when executed.

9. Experimental Results – This chapter contains the screenshots of our results.

10. Conclusion and Future Enhancements – This covers the conclusion of our project and the possible future developments.

2. LITERATURE SURVEY

2.1 SURVEY ON BACKGROUND

Deep neural network model that can detect [1] handguns in images and a machine learning and computer vision pipeline that detects abandoned luggage so that we could identify potential gun-based crime and abandoned luggage situations in surveillance footage. [2] In this it consists of six abnormal activities such as abandoned object detection, theft detection, fall detection, accidents and illegal parking detection on road, violence activity detection, and fire detection. It aims to detect suspicious activities such as object exchange, [3] entry of a new person, peeping into other's answer sheet and person exchange from the video captured by a surveillance camera during examinations. This requires the process of face recognition, hand recognition and detecting the contact between the face and hands of the same person and that among different persons.

Improves robustness for activity detection by providing intelligent control and failover mechanisms, built on top of low-level motion detection algorithms such as frame differencing and feature correlation. For activity recognition[4], we propose an efficient representation of human activities that enables recognition of different interaction patterns among a group of people based on simple statistics computed on the tracked trajectories, without building complicated Markov chain, hidden Markov models (HMM), or coupled hidden Markov models (CHMM). The proposed work aims in developing a system that [5] analyze and detect the suspicious activity that are often occurring in a classroom environment. Video Analytics provides an optimal solution for this as it helps in pointing out an event and retrieves the relevant information from the video recorded.

This study presents an approach, called SARDBN (Suspicious Activity Reporting using Dynamic Bayesian Network) [6], that employs a combination of clustering and dynamic Bayesian network (DBN) to identify anomalies in sequence of transactions. [7] Detection of suspicious activity and estimate of risk from human behavior shot by surveillance camera extracts Motion Region from moving person, and measures Motion Quantity for measuring his/her active state. And this proposal method finds the detecting point of suspicious activity, and estimates the degree of risk of the suspicious activity.

Anticipated method for automatically detecting the suspicious or violent activities of a person from the surveillance video. We train the SVM classifier with the HOG features extracted from the video frames of two types: frames showing no [8] violent activities and those showing violent activities like kicking, pushing, punching, etc. In the testing phase, the frames from the surveillance video are read and processed to classify them as violent or normal frames. Employs a one-class support vector machine (SVM) that is trained on commonly available normal activities, [9] which filters out the activities that have a very high probability of being normal.

We then derive abnormal activity models from a general normal model via a kernel nonlinear regression (KNLR) to reduce false positive rate in an unsupervised manner. The hierarchical approach is used to detect the different suspicious activities such as loitering, fainting, unauthorized entry etc. [10] this approach is based on the motion features between the different objects. Forensic surveillance strategy by introducing an Instant Suspicious Activity identification at the Edge (I-SAFE) [11] using fuzzy decision making. A fuzzy control system is proposed to mimic the decision-making process of a security officer. **P**roposes a system for detection of the sentiments in online messages and comments exchanged over social networking and blogging [12] sites, which

monitors communication between users and identify the messages of individuals who exhibits anomalies behaviour over time. A framework is proposed to detect suspicious human behavior as well as tracking of human who is doing some [13] unusual activity such as fighting and threatening actions and also distinguishing the human normal activities from the suspicious behavior. A solution to combine logging, and network-based intrusion detection and prevention system. The system has been developed [14] considering the Software Engineering framework of requirements analysis, design, implementation, and testing. [15] Monitor the events taking place in frame of camera using image processing. In this proposed method, we are using a Raspberry Pi as our main processor to which camera will be interfaced.

2.2 CONCLUSION ON SURVEY

These works provide basic background information about various techniques and algorithms in deep learning and machine learning in order to improve face recognition, feature comparison and feature matching. Deep learning and Machine Learning algorithms are the major part of image recognition involved projects since they are the fundamental aspect to these projects. Recent decade witnessed a good number of publications in the field of visual surveillance to recognize the abnormal activities. Furthermore, a few surveys can be seen in the literature for the different abnormal activity recognition, but none of them have addressed different abnormal activities in a review. In general, we have discussed all the steps those have been followed to recognize the human activity from the surveillance videos in the literature, such as foreground object extraction, object detection based on tracking or non-tracking methods, feature extraction, classification, activity analysis and recognition. From the survey we have concluded that not only detecting human activities is important, the objects present in the video should also be detected.

3.SOFTWARE AND HARDWARE REQUIREMENTS

PYTHON

Python is a general-purpose interpreted, interactive, object-oriented, and high-level programming language. An interpreted language, Python has a design philosophy that emphasizes code readability (notably using whitespace indentation to delimit code blocks rather than curly brackets or keywords), and a syntax that allows programmers to express concepts in fewer lines of code than might be used in languages such as C++ or JAVA. It provides constructs that enable clear programming on both small and large scales. Python interpreters are available for many operating systems. CPython, the reference implementation of Python, is open source software and has a community-based development model, as do nearly all of its variant implementations. CPython is managed by the non-profit Python Software Foundation. Python features a dynamic type system and automatic memory management. It supports multiple programming paradigms, including object-oriented, imperative, functional and procedural, and has a large and comprehensive standard library.

DJANGO

Django is a high-level Python Web framework that encourages rapid development and clean, pragmatic design. Built by experienced developers, it takes care of much of the hassle of Web development, so you can focus on writing your app without needing to reinvent the wheel. It’s free and open source.

Django's primary goal is to ease the creation of complex, database-driven websites. Django emphasizes reusability and "pluggability" of components, rapid development, and the principle of don't repeat yourself. Python is used throughout, even for settings files and data models.

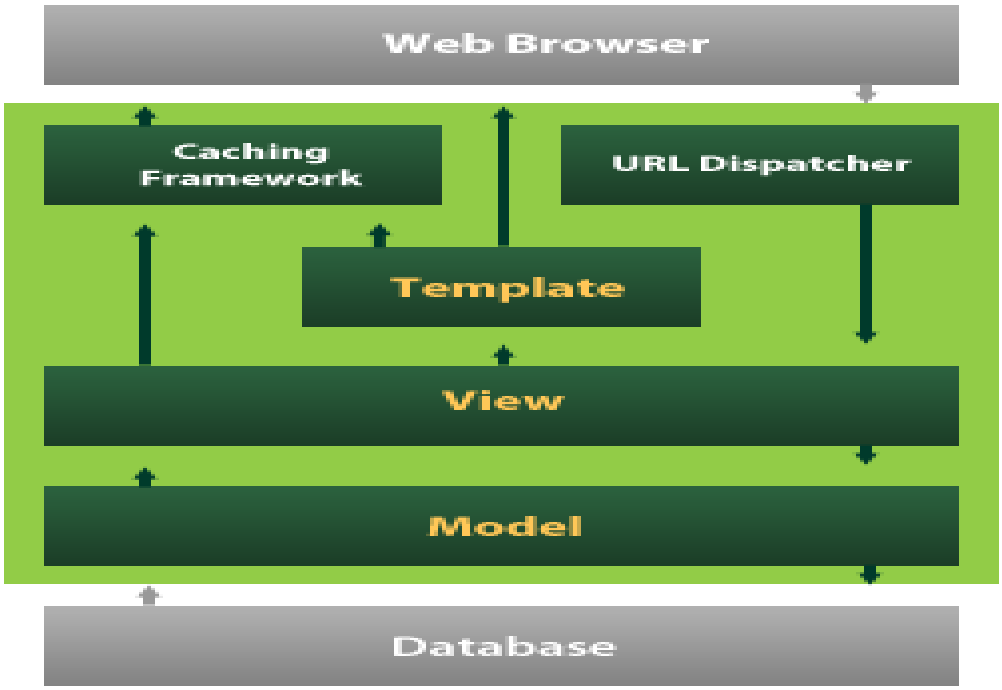


Fig 3.1 Django Model

Django also provides an optional administrative create, read, update and delete interface that is generated dynamically through introspection and configured via admin models.

Why Django Framework?

- Excellent documentation and high scalability.
- Used by Top MNCs and Companies, such as Instagram, Disqus, Spotify, Youtube, Bitbucket, Dropbox, etc. and the list is never-ending.
- Easiest Framework to learn, rapid development and Batteries fully included.
- The last but not least reason to learn Django is Python, Python has huge library and features such as Web Scrapping, Machine Learning, Image Processing, Scientific Computing, etc. One can integrate it all this with web application and do lots and lots of advance stuff.

FUNCTIONAL REQUIREMENTS:

These are the requirements that the end user specifically demands as basic facilities that the system should offer. All these functionalities need to be necessarily incorporated into the system as a part of the contract. These are represented or stated in the form of input to be given to the system, the operation performed and the output expected. They are basically the requirements stated by the user which one can see directly in the final product, unlike the non-functional requirements.

Benefits of Functional Requirement

Here, are the pros/advantages of creating a typical functional requirement document-

- Helps you to check whether the application is providing all the functionalities that were mentioned in the functional requirement of that application
- A functional requirement document helps you to define the functionality of a system or one of its subsystems.
- Functional requirements along with requirement analysis help identify missing requirements. They help clearly define the expected system service and behavior.
- Errors caught in the Functional requirement gathering stage are the cheapest to fix.

NON-FUNCTIONAL REQUIRMENTS:

In systems engineering and requirements engineering, a non-functional requirement is a requirement that specifies criteria that can be used to judge the operation of a system, rather than specific behaviors. They are contrasted with functional requirements that define specific behavior or functions. **Non-functional requirements** add tremendous value to business analysis. It is commonly misunderstood by a lot of people. It is important for business stakeholders, and Clients to clearly explain the requirements and their expectations in measurable terms. If the non-functional requirements are not measurable then they should be revised or rewritten to gain better clarity. For example, User stories help in mitigating the gap between developers and the user community in Agile Methodology.

Usability:

Prioritize the important functions of the system based on usage patterns. **Frequently used functions should be tested for usability**, as should complex and critical functions. Be sure to create a requirement for this.

Reliability:

Reliability defines the trust in the system that is developed after using it for a period of time. It defines the likeability of the software to work without failure for a given time period.

The number of bugs in the code, hardware failures, and problems can reduce the reliability of the software.

Your goal should be a long MTBF (mean time between failures). It is defined as the average period of time the system runs before failing.

Create a requirement that data created in the system will be retained for a number of years without the data being changed by the system.

It's a good idea to also include requirements that make it easier to monitor system performance.

Performance:

What should system response times be, as measured from any point, under what circumstances? Are there specific peak times when the load on the system will be unusually high?

Think of stress periods, for example, at the end of the month or in conjunction with payroll disbursement.

Supportability:

The system needs to be **cost-effective to maintain**.

Maintainability requirements may cover diverse levels of documentation, such as system documentation, as well as test documentation, e.g. which test cases and test plans will accompany the system.

3.1 SOFTWARE REQUIREMENTS:

- **Operating system** : Windows 7 Ultimate.
- **Coding Language** : Python.
- **Front-End** : Python.
- **Designing** : Html, CSS, and JavaScript.
- **Data Base** : MySQL.

3.2 HARDWARE REQUIREMENTS:

- **System** : I3 or Higher.
- **Hard Disk** : 40 GB.
- **Floppy Drive** : 1.44 Mb.
- **Monitor** : 14’ Colour Monitor.
- **Mouse** : Optical Mouse.
- **Ram** : 4GB

SYSTEM STUDY

FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are,

- ECONOMICAL FEASIBILITY
- TECHNICAL FEASIBILITY
- SOCIAL FEASIBILITY

ECONOMICAL FEASIBILITY

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified.

Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

TECHNICAL FEASIBILITY

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources.

This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

SOCIAL FEASIBILITY

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity.

The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

4. SOFTWARE DEVELOPMENT ANALYASIS

4.1 OVERVIEW OF PROBLEM

Suspicious human activity recognition from surveillance video is an active research area of image processing and computer vision. Through the visual surveillance, human activities can be monitored in sensitive and public areas such as bus stations, railway stations, airports, banks, shopping malls, school and colleges, parking lots, roads, etc. to prevent terrorism, theft, accidents and illegal parking, vandalism, fighting, chain snatching, crime and other suspicious activities. It is very difficult to watch public places continuously, therefore an intelligent video surveillance is required that can monitor the human activities in real-time and categorize them as usual and unusual activities; and can generate an alert.

4.2 DEFINE THE PROBLEM

In this project we need to detect person behaviour as suspicious or not, now a day's everywhere CCTV cameras are installed which capture videos and store at centralized server and manually scanning those videos to detect suspicious activity from human required lots of human efforts and time. To overcome from such issue author is asking to automate such process using Machine Learning Algorithms.

4.3 MODULES OVERVIEW:

We developed this system with the help of following modules:

- Foreground Extraction
- Object Tracking
- Feature Extraction
- Classification and Activity Recognition

4.4 DEFINE THE MODULES

- **Foreground Extraction:**

  Foreground object extraction from the video is an initial and important step of suspicious human activity recognition. Background subtraction is a powerful mechanism to detect the change in the sequence of frames and to extract foreground objects. Foreground objects consists of moving objects and newly arrived objects in a video which becomes stationary after some time such as left luggage. But moving objects are considered as the foreground objects while static objects are considered as background of the video in background subtraction techniques. This concept simplifies the moving object detection from a video of static camera but difficult to detect newly arrived stationary objects.

- **Object Tracking**:

  It is an important and challenging chore in the field of computer vision. It helps in generating the trajectory of an object over time with the tracing its position in consecutive frames of surveillance video to analyze the human behavior. Object shape representations employed for tracking are points, object contour, object silhouette, primitive geometric shapes, articulated shapes and skeletal models. Sometimes, tracking of an object becomes difficult due to noise in the image, partial or full occlusion of objects, complex object shapes, illumination changes, complex object motion, and deformable objects.

- **Feature Extraction:**

  Selecting appropriate features plays an important role in an automatic recognition of abnormal activities from video surveillance. The main objective of feature extraction is to find the most promising information in the recorded video.

- **Classification and Activity Recognition:**

  After finding moving or stationary foreground objects in a frame, the object classification step is applied for the recognition of normal or abnormal behavior. Object classification distinguishes to the static human from static abandoned object, fighting from boxing, face from skin color objects, fire from flashlight, sun light, and any artificial light, falling human pose from laying human pose etc.

4.5 MODULES FUNCTIONALITY

- **Foreground object detection**: Foreground object extraction from the video is an initial and important step of suspicious human activity recognition.

  a. Moving Object Detection: Moving object detection can be performed based on two approaches- background modeling and change detection-based approaches. The change detection approaches find the difference between two consecutive frames to recover motion and apply post processing methods to recover the complete object. These methods are faster in respect to execution while lacking in accuracy. A reasonably correct background model for the background can help to extract the foreground objects much effectively in comparison to the previous class of methods.

  b. Stationary Foreground Object Detection: Suspicious activity recognition includes abandoned object detection to prevent the explosive attacks performed by terrorists. In video surveillance, background techniques consider moving objects as a foreground object and static object as a background. Therefore, when a newly arrived object becomes static then it is absorbed in the background. Several authors used different background subtraction techniques with dual background approach with different learning rate to extract the two foreground objects for detecting the stationary objects of the video.

  c. Noise removal, shadow removal and illumination handling methods: Detecting the foreground objects without noise, illumination effect, and shadow is a very challenging in area of computer vision. Noise creates problem in the identification of the object, illumination effect causes the false detection, and shadow changes the appearance of the object due to that object tracking becomes very difficult.

- **Object tracking:** Object tracking is an important and challenging chore in the field of computer vision. It helps in generating the trajectory of an object over time with the tracing its position in consecutive frames of surveillance video to analyze the human behavior. Object shape representations employed for tracking are points, object contour, object silhouette, primitive geometric shapes, articulated shapes and skeletal models.

- **Feature extraction:** Selecting appropriate features plays an important role in an automatic recognition of abnormal activities from video surveillance. The main objective of feature extraction is to find the most promising information in the recorded video.

a. **Feature extraction for abandoned object detection/theft detection:** To detect the static objects in the video is very complex task. Therefore, some features of objects are extracted from video to make distinction between moving and stationary objects. Dual foreground with different learning rate: In Porikli (2007), Porikli et al. (2008), dual foreground technique has been employed with two different longterm and shortterm learning 123 R. K. Tripathi et al. rates. With these two different learning rates, two foreground masks FL and FS are created. If (FL; FS) = (1, 0), then object is static. Centroid, height and width of an object Centroid is defined as an average of the pixels in x and y coordinates belonging to the object that can be calculated through the following formula:

$$Cx = \frac{\sum_{i=1}^{n} Xi}{N}$$
$$Cy = \frac{\sum_{i=1}^{n} Yi}{N}$$

b. Feature extraction for falling detection: Point features extraction-Centroids, orientation and distance: In Chua et al. (2013), three points are drawn on human shape with the help of bounding box around the human. A bounding box is computed around the human, and then bounding box is divided into three portions which represent upper, mid and lower body part. The starting and end point of the bounding box are used to calculate the centroids of the three regions. The coordinates of the centroids are computed by the following formula:

$$C_{Xi} = \frac{1}{N_{Ri}} \sum_{i=1}^{N_{Ri}} X_i \quad i = 1, 2, 3$$
$$C_{Yi} = \frac{1}{N_{Ri}} \sum_{i=1}^{N_{Ri}} Y_i \quad i = 1, 2, 3$$

b. **Feature extraction for abnormal activity detection on road traffic:** Estimate motion vectors of vehicles Once a vehicle region leaves the slit, its shape is updated along the time sequence by algorithm. Histogram of flow gradients (HFG) Histogram of Flow Gradients algorithm (Sadeky et al. 2010) is similar to the HOG, but differs in that HFG locally runs on optical flow field in motion scenes. HFG can be implemented computationally faster than that of HOG. The angle and magnitude of the optical flow required to construct HFG are determined by the following formula:

$$\theta = \tan^{-1}\left(\frac{u}{v}\right), \rho = (u^2 + v^2)^{1/2}$$

d. Feature extraction for violence detection: Shape and texture features Shape and texture features are extracted in Kausalya and Chitrakala (2012) for tracking the moving object. Curvelet mainly extracts the features from images and use to compute the similarity values between images so that efficient geometric shape structure-based image retrieval is possible. Edge detection map is used to detect the edge features.

e. Feature extraction for fire and smoke detection: Flame detection through HSV and YCbCr color model In Seebamrungsat et al. (2014), properties of the YCbCr and HSV color models are used to differentiate the flame colors from the 123 Suspicious human activity recognition: a review background. The HSV color model is applied to detect information related to brightness and color. Through the YCbCr color model, information regarding brightness can be extracted due to its more capability to distinguish bright images efficiently than other color models.

- **Classification and activity recognition:** After finding moving or stationary foreground objects in a frame, the object classification step is applied for the recognition of normal or abnormal behavior. Object classification distinguishes to the static human from static abandoned object, fighting from boxing, face from skin color objects, fire from flashlight, sun light, and any artificial light, falling human pose from laying human pose etc. In general, there are three- feature based, motion based and shape-based classification methods. Several researchers have utilized the different features with different classifiers such as SVM, k-Nearest Neighbor, Multi-SVM, Cascade classifier, Neural Network, and HAR to analyze the human behavior and recognition of abnormal activities.

5. PROJECT SYSTEM DESIGN

INPUT DESIGN

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the following things:

- What data should be given as input?
- How the data should be arranged or coded?
- The dialog to guide the operating personnel in providing input.
- Methods for preparing input validations and steps to follow when error occur.

OBJECTIVES

1.Input Design is the process of converting a user-oriented description of the input into a computer-based system. This design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerized system.

2. It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The goal of designing input is to make data entry easier and to be free from errors. The data entry screen is designed in such a way that all the data manipulates can be performed. It also provides record viewing facilities.

3. When the data is entered it will check for its validity. Data can be entered with the help of screens. Appropriate messages are provided as when needed so that the user

will not be in maize of instant. Thus the objective of input design is to create an input layout that is easy to follow

OUTPUT DESIGN

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making.

1. Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is designed so that people will find the system can use easily and effectively. When analysis design computer output, they should Identify the specific output that is needed to meet the requirements.

2.Select methods for presenting information.

3.Create document, report, or other formats that contain information produced by the system.

The output form of an information system should accomplish one or more of the following objectives.

- Convey information about past activities, current status or projections of the
- Future.
- Signal important events, opportunities, problems, or warnings.
- Trigger an action.
- Confirm an action.

5.1 Data-Flow-Diagram

Also known as DFD, Data flow diagrams are used to graphically represent the flow of data in a business information system. DFD describes the processes that are involved in a system to transfer data from the input to the file storage and reports generation.

Data flow diagrams can be divided into logical and physical. The logical data flow diagram describes flow of data through a system to perform certain functionality of a business. The physical data flow diagram describes the implementation of the logical data flow.

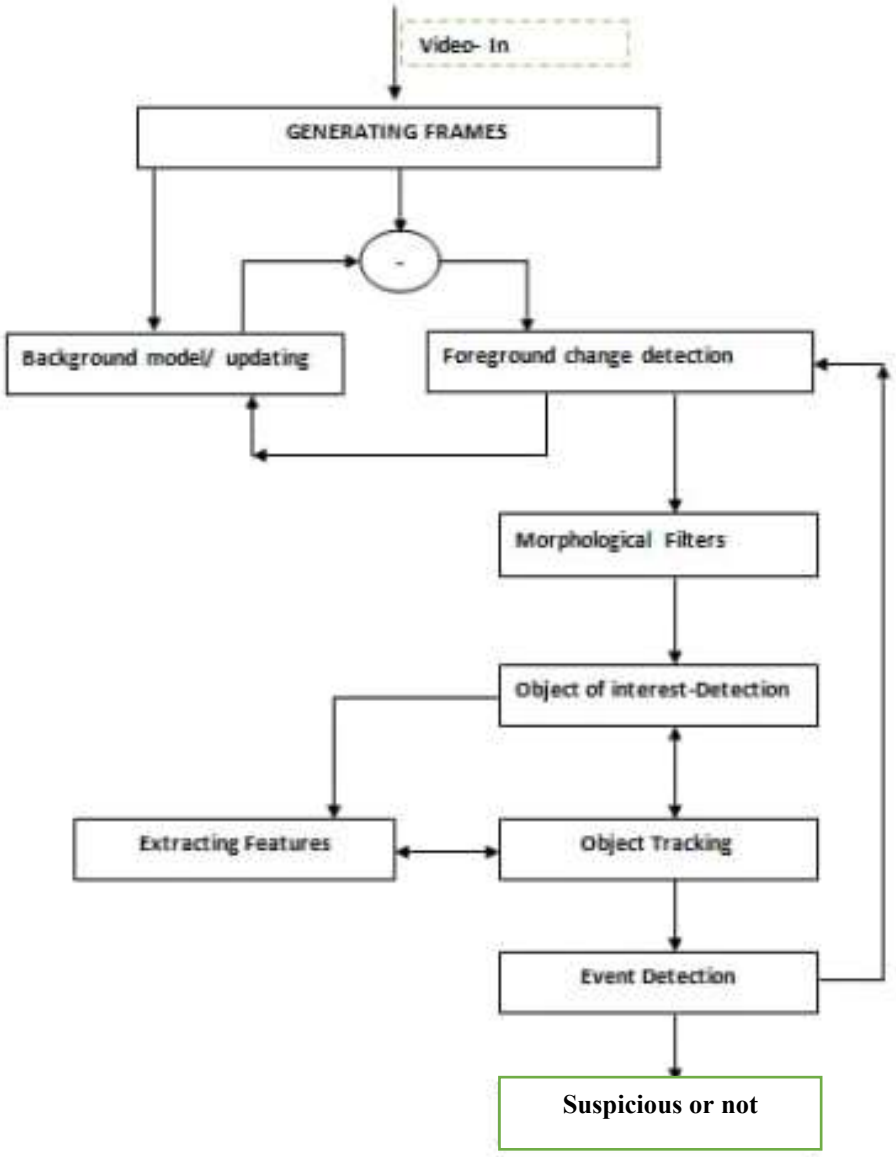


Fig 5.1.1 DFD

Video Analytics, also referred to as Video Content Analysis (VCA), is a generic term used to describe computerized processing and analysis of video streams. Before detecting the human activities from a video input, it is important to detect the objects present in the video. The object detection can be done by frame difference, background elimination and background registration techniques. Post processing techniques are used to increase the clarity of the frame obtained. The noises in the frames can be removed by using the Gaussian filter approach. Background subtraction and histogram of gradients technique is used for detecting the presence or absence of a human. The detection of human motion can be done by several other methods. One such method is the motion estimation map. It is used to estimate the motion of humans in a surveillance video. The huge advantage of this method is the ability of it to identify motion in daylight as well as night time. Adaptive Background modeling is another method which is used to detect humans in a crowded environment. The faces of the persons in the processed frame have to be identified. This can be done by using the Haar features approach or the Surf features approach.

After Background detection and foreground detection, feature extraction and classification of video is done.

After analyzing the video, frames extracted. Frames having suspicious probability greater than 85% is detected.

5.2 ER DIAGRAM

ER Diagram stands for Entity Relationship Diagram, also known as ERD is a diagram that displays the relationship of entity sets stored in a database. In other words, ER diagrams help to explain the logical structure of databases. ER diagrams are created based on three basic concepts: entities, attributes and relationships.

ER Diagrams contain different symbols that use rectangles to represent entities, ovals to define attributes and diamond shapes to represent relationships.

At first look, an ER diagram looks very similar to the flowchart. However, ER Diagram includes many specialized symbols, and its meanings make this model unique. The purpose of ER Diagram is to represent the entity framework infrastructure

ER Modelling helps you to analyze data requirements systematically to produce a well-designed database. So, it is considered a best practice to complete ER modeling before implementing your database.

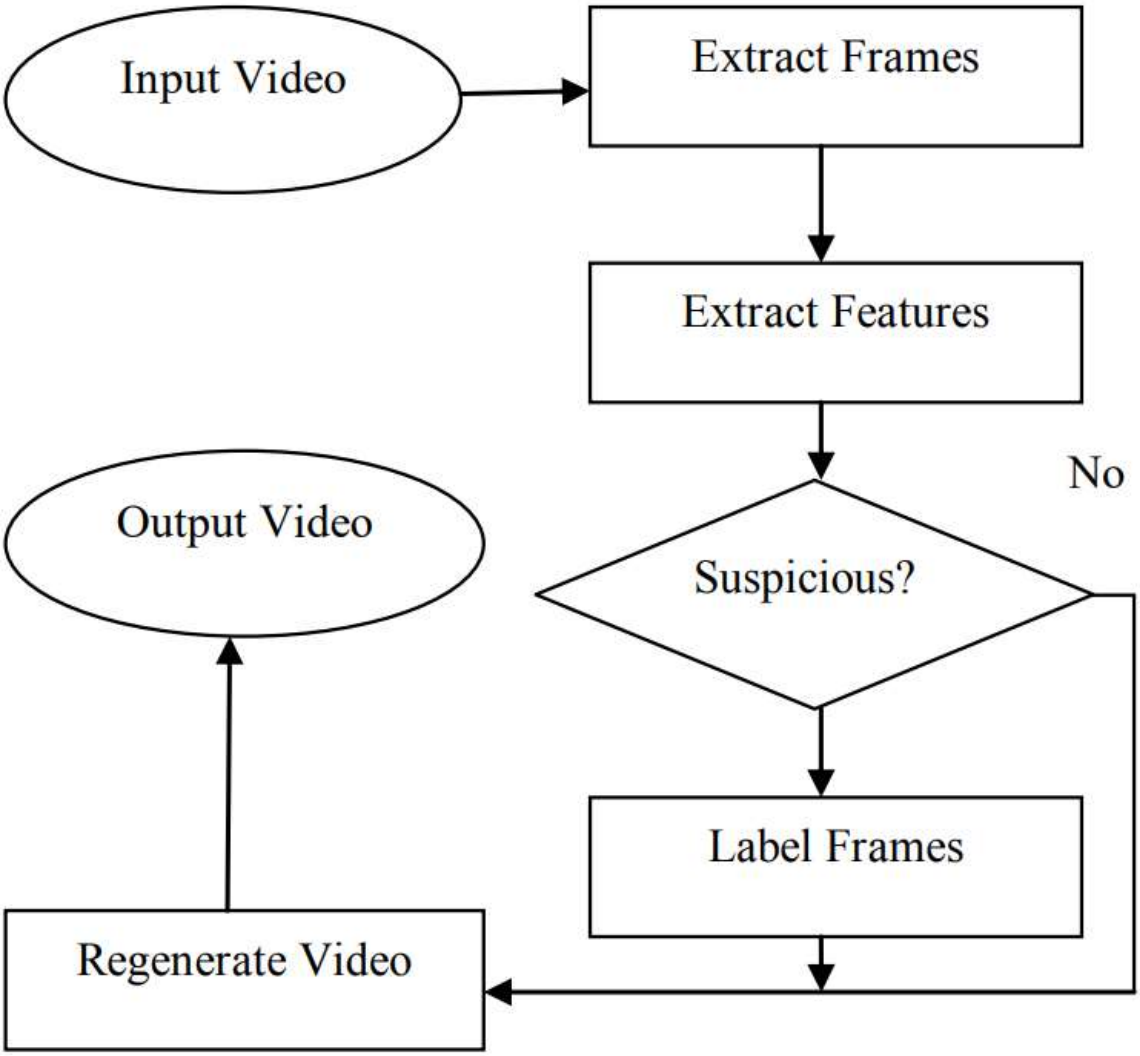


Fig 5.2.1 E.R. diagram

5.3 UML DIAGRAMS

Class Diagram:

A class diagram provides a pictorial representation of all the classes in an object-oriented system; their attributes and methods; their connections; their interactions and inheritances if any. In simpler terms, classes represent objects whose roles are similar and to what extend the objects of the classes "know" about each other**.**

- tkinter class is GUI library for python used to create widget

- matplotlib class is used for data visualization and graphical plotting library in python and numeric extension of NumPy.

  -pyplot is a plotting library used for 2D graphics in python.

- imageai class is a library used for computer visualization containing classes and functions to perform video object detection, tracking and video analysis.

  -videoCapture is used to run detection tasks and analyze videos, live video feeds from device cameras and IP cameras.

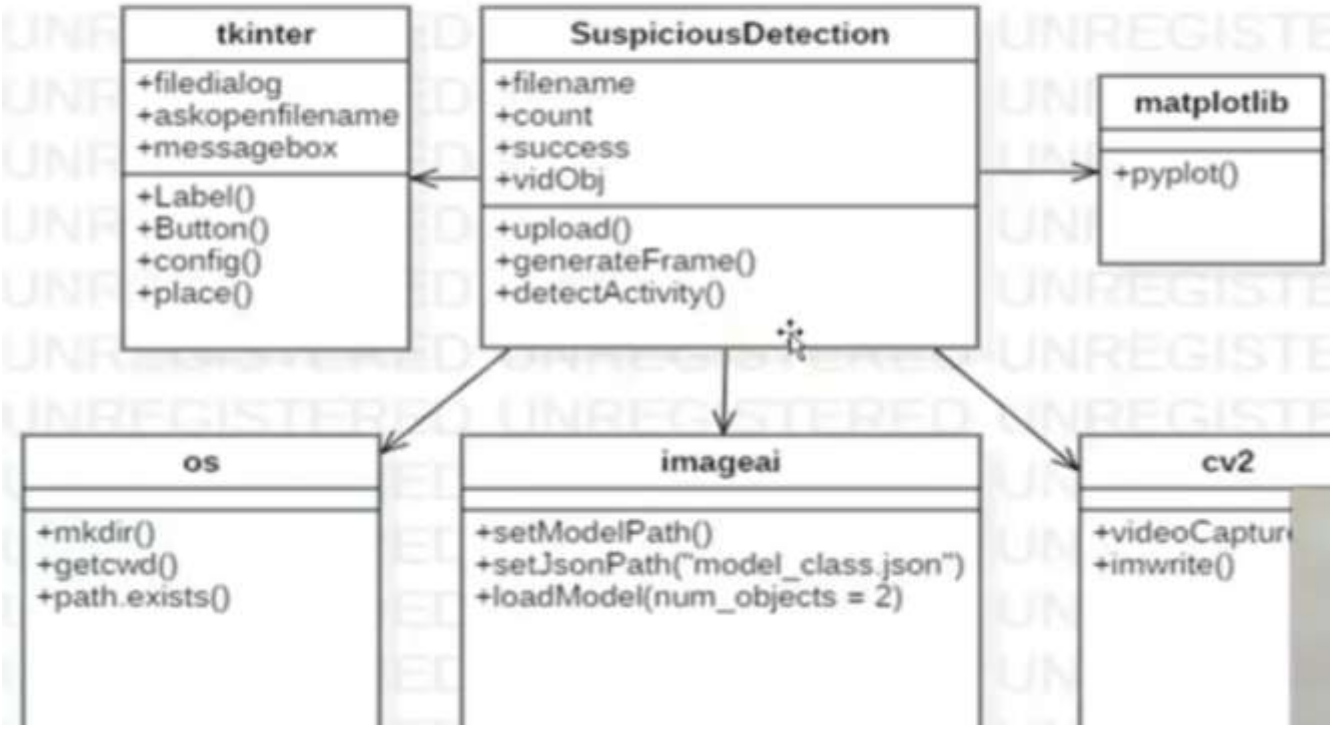


Fig 5.3.1 Class Diagram

Use-Case Diagram:

A use case is simply a list of actions which typically define the interactions between an actor and the system with an aim of achieving a certain goal. Each interaction is a single unit of work and captures a "contract" for the behaviour of the system under discussion to deliver a single goal (Kettenis, 2007). Most of the functional requirements are captured by the use case.

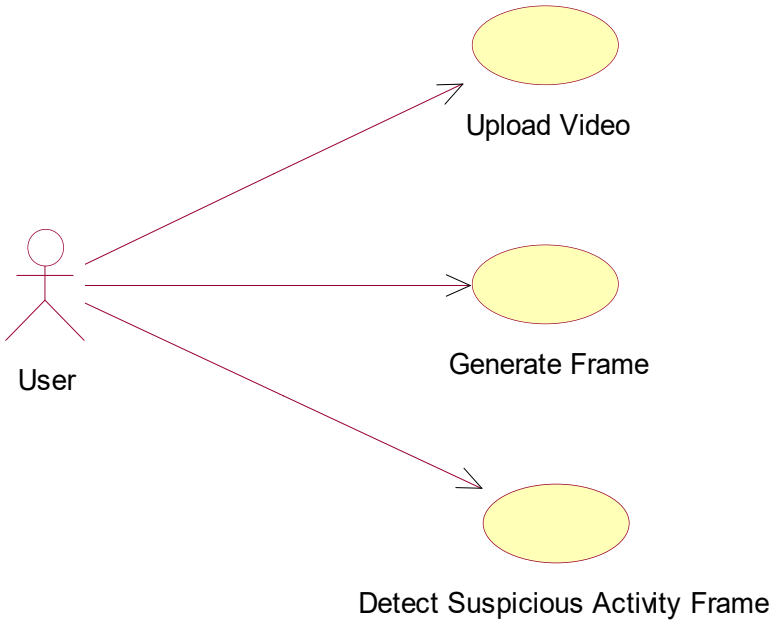


Fig 5.3.2 Use-Case Diagram

Here user can perform 3 actions:

Upload Video: User can input the video

Generate Frames: User can extract the frames from the input video

Detect Suspicious Activity Frame: User will get to know about the unusual activity frame along with frame number.

Sequence Diagram:

The sequence diagram in this case provides a visual representation the object interactions during the searching process. This includes the actor and the objects the actors interact with throughout the execution of the search.

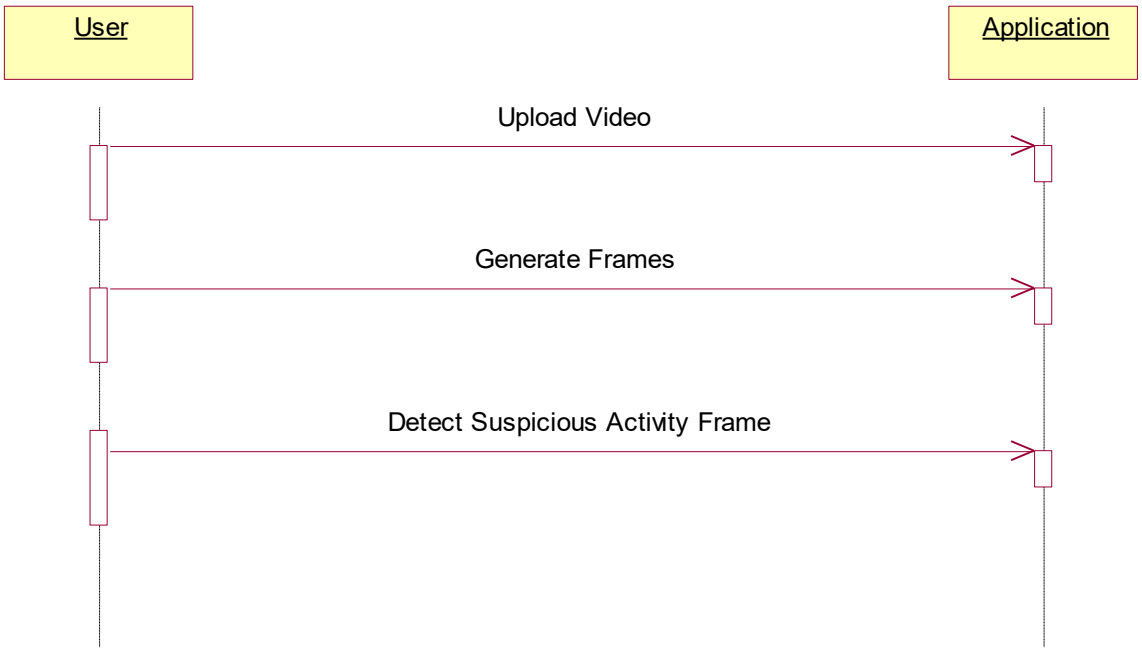


Fig 5.3.3 Sequence Diagram

COLLABORATION DIAGRAM:

A collaboration diagram, also known as a communication diagram, is an illustration of the relationships and interactions among software objects in the Unified Modelling Language (UML). These diagrams can be used to portray the dynamic behaviour of a particular use case and define the role of each object.

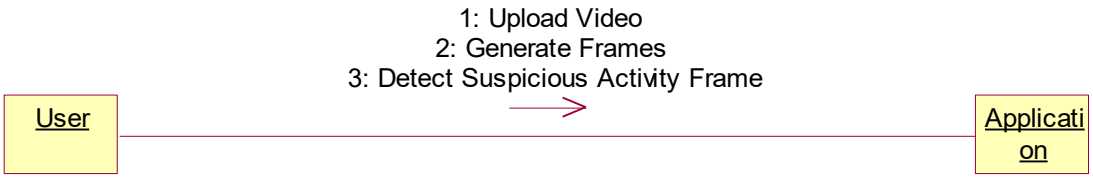


Fig 5.3.4 Collaboration Diagram

ACTIVITY DIAGRAM:

Activity diagram is another important diagram in UML to describe the dynamic aspects of the system. Activity diagram is basically a flowchart to represent the flow from one **activity** to another activity. The activity can be described as an operation of the system. The control flow is drawn from one operation to another.

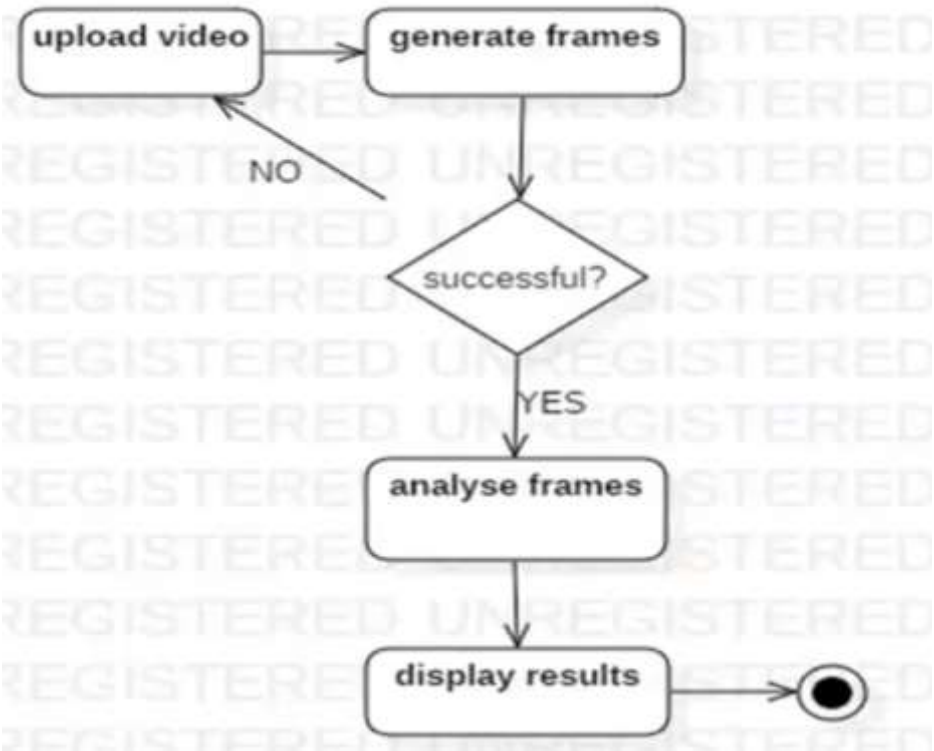


Fig 5.3.5 Activity Diagram

Firstly, we have to upload the video. After uploading the video, we have to extract the frames from the input video. If the frames are generated successfully then we have to analyse the frames. If the frames are not generated successfully then the first step is repeated. After analysing the frames, the results will be displayed whether it is suspicious or not.

COMPONENT DIAGRAM:

Component diagram is a special kind of diagram in UML. The purpose is also different from all other diagrams discussed so far. It does not describe the functionality of the system but it describes the components used to make those functionalities.

Thus from that point of view, component diagrams are used to visualize the physical components in a system. These components are libraries, packages, files, etc.

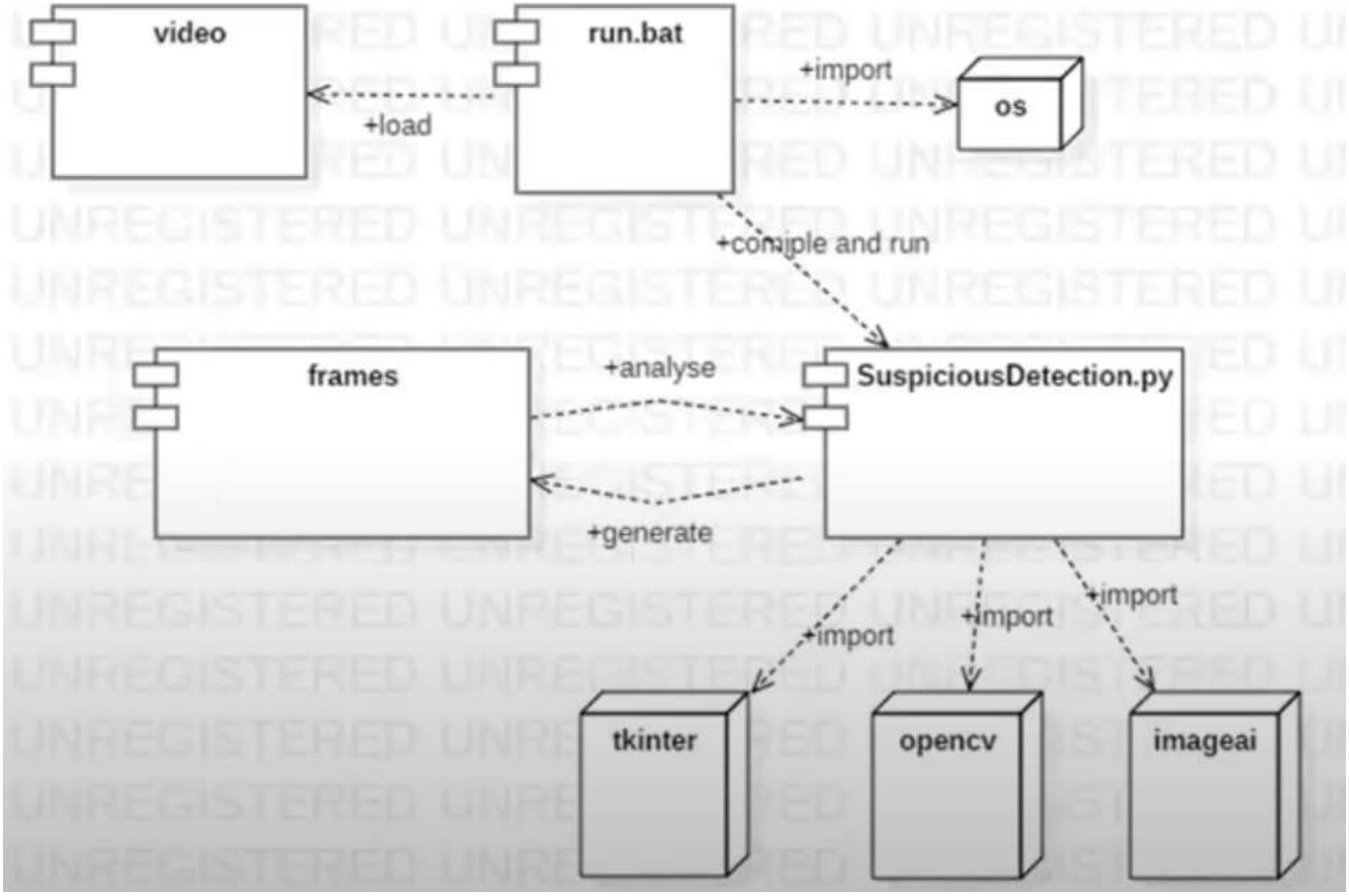


Fig 5.3.6 Component Diagram

6. PROJECT CODING

6.1 CODING TEMPLATES

```
def upload():
   global filename
   filename = askopenfilename(initialdir = "videos")
   pathlabel.config(text=filename)
#This function is called when a user needs to upload a video sample.


def generateFrame():
   global filename
   text.delete('1.0', END)
# This is function is called after the user uploads the video,
#The main objective of this function is to segment the video and
 #Generate frames for every second. After Generating,
#The Frames are stored in a specified Frames folder.


def detectActivity():
   imagePaths = sorted(list(paths.list_images("frames")))
#This function detects whether the video contains a suspicious activity
 #From the frames that have a high probability and returns the frame numbers.
#If the video does not contain any suspicious activity, this function returns
#"No suspicious activity is detected".
```

6.2 OUTLINE FOR VARIOUS FILES

- **Tkinter**

  Creates a GUI interface consisting of buttons performing specific actions.

- **Matplotlib**

  It is a plotting library for the python and its numerical mathematics extension NumPy.

- **TensorFlow**

  It is used to create Deep Learning models directly or by using wrapper libraries that simplify the process built on top of TensorFlow.

- **cv2**

  This module is used for video Segmentation

- **shutil**

  This module helps in automating process of removal of files and directories.

- **OS**

  The OS module in Python provides functions for interacting with the operating system.

- **Imageai**

  This is a machine learning library that simplifies AI training and object detection in images.

- **PixelLib**

  This library used for easy implementation of semantic and instance segmentation of objects in images and videos.

6.3 CLASS WITH FUNCTIONALITY

generateFrame() :

```
def generateFrame():
    global filename
    text.delete('1.0', END)
    if not os.path.exists('frames'):
        os.mkdir('frames')
    else:
        shutil.rmtree('frames')
        os.mkdir('frames')
    vidObj = cv2.VideoCapture(filename)
    count = 0
    success = 1
    while success:
        success, image = vidObj.read()
        if count < 500:
            cv2.imwrite("frames/frame%d.jpg" % count, image)
            text.insert(END,"frames/frame."+str(count)+" saved\n")
            print("frames/frame."+str(count)+" saved")
            #pathlabel.config(text="frames/frame."+str(count)+" saved")
        else:
            break
        count += 1
```

detectActivity():

```
def detectActivity():
    imagePaths = sorted(list(paths.list_images("frames")))
    count = 0
    option = 0;
    text1.delete('1.0', END)
    for imagePath in imagePaths:
        predictions, probabilities = prediction.predictImage(imagePath, result_count=1)
        for eachPrediction, eachProbability in zip(predictions, probabilities):
            if float(eachProbability) > 80:
                count = count + 1;
            if float(eachProbability) < 80:
                count = 0
            if count > 10:
                option = 1
                print(imagePath+" is predicted as "+eachPrediction+" with probability :
                      " +str(eachProbability))
                text1.insert(END,imagePath+" is predicted as "+eachPrediction+" with probability :
                             " +str(eachProbability)+"\n\n")
                count = 0;
        print(imagePath+" processed")
    if option == 0:
        text1.insert(END,"No suspicious activity found in given footage")
```

6.4 METHODS INPUT AND OUTPUT PARAMETERS.

- upload ()
- generateframe ()
- detectActivity ()
- imagepath ()
- imwrite()

- upload() method is used to input a video sample from specific path which needs to be segmented.
- generateframe() method is used to extract frames per second and stored in specific location.
- detectActivity() method is used to detect specific activity from the frames generated and it describes whether the frame contains any suspicious activity or not.
- In this project, first we have to input, and the video is taken by upload() method. After uploading the video, we must extract frames from video using generateframe() method. The main functionality of this method is to segment the video and generate frames.
- detectActivity() method is used to detect whether the frame applied on trained model is Suspicious or not.
- imwrite(A , filename) writes image data A to the file specified by filename , inferring the file format from the extension. imwrite creates the new file in your current folder. The bit depth of the output image depends on the data type of A and the file format.
- Relative path where image is located in the same directory as python code file, an absolute path can be used as well.

7. PROJECT TESTING

7.1 VARIOUS TEST CASES

Unit testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

Integration testing

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

Functional test

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centred on the following items:

Valid Input : identified classes of valid input must be accepted.

Invalid Input : identified classes of invalid input must be rejected.

Functions : identified functions must be exercised.

Output : identified classes of application outputs must be exercised.

Systems/Procedures : interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases.

In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing.

Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

System Test

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

7.2 BLACK-BOX TESTING

Black box testing is a technique of software testing which examines the functionality of software without peering into its internal structure or coding. The primary source of black box testing is a specification of requirements that is stated by the customer.In this method, tester selects a function and gives input value to examine its functionality, and checks whether the function is giving expected output or not. If the function produces correct output, then it is passed in testing, otherwise failed. The test team reports the result to the development team and then tests the next function. After completing testing of all functions if there are severe problems, then it is given back to the development team for correction

7.3 WHITE – BOX TESTING

White box testing which also known as glass box is testing, structural testing, clear box testing, open box testing and transparent box testing. It tests internal coding and infrastructure of a software focus on checking of predefined inputs against expected and desired outputs. It is based on inner workings of an application and revolves around internal structure testing. In this type of testing programming skills are required to design test cases. The primary goal of white box testing is to focus on the flow of inputs and outputs through the software and strengthening the security of the software.

Developers do white box testing. In this, the developer will test every line of the code of the program. The developers perform the White-box testing and then send the application or the software to the testing team, where they will perform the black box testing and verify the application along with the requirements and identify the bugs and sends it to the developer.

8. OUTPUT SCREENS

8.1 USER INTERFACES

Home Page GUI:

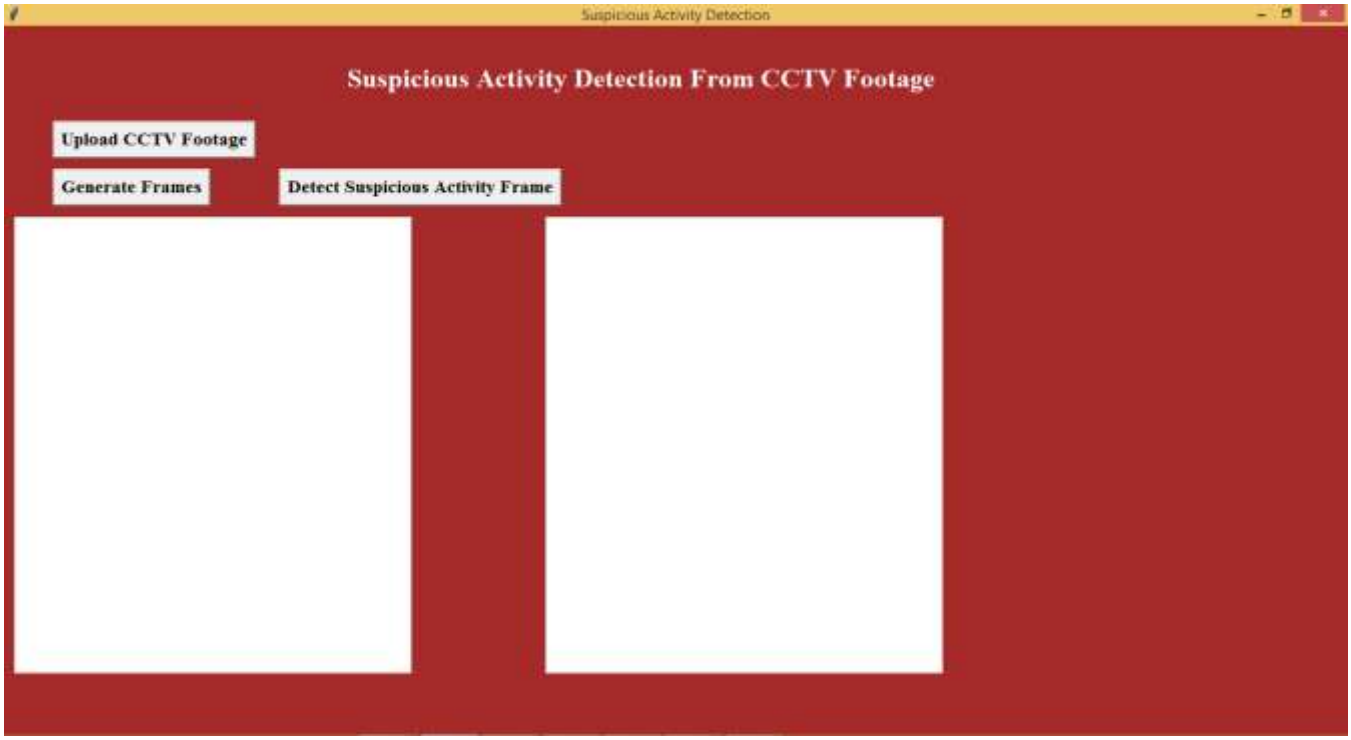


Fig 8.1.1 User Interface

User can interact the GUI application and the UI consists of three buttons:

- upload CCTV footage: users can input the video by using this input button.
- Generate frames: extraction of frames from video is done and stored in frames folder.
- Detect Suspicious Activity Frame: The suspicious activity frame is detected from the extracted video frames.

8.2 OUTPUT SCREENS

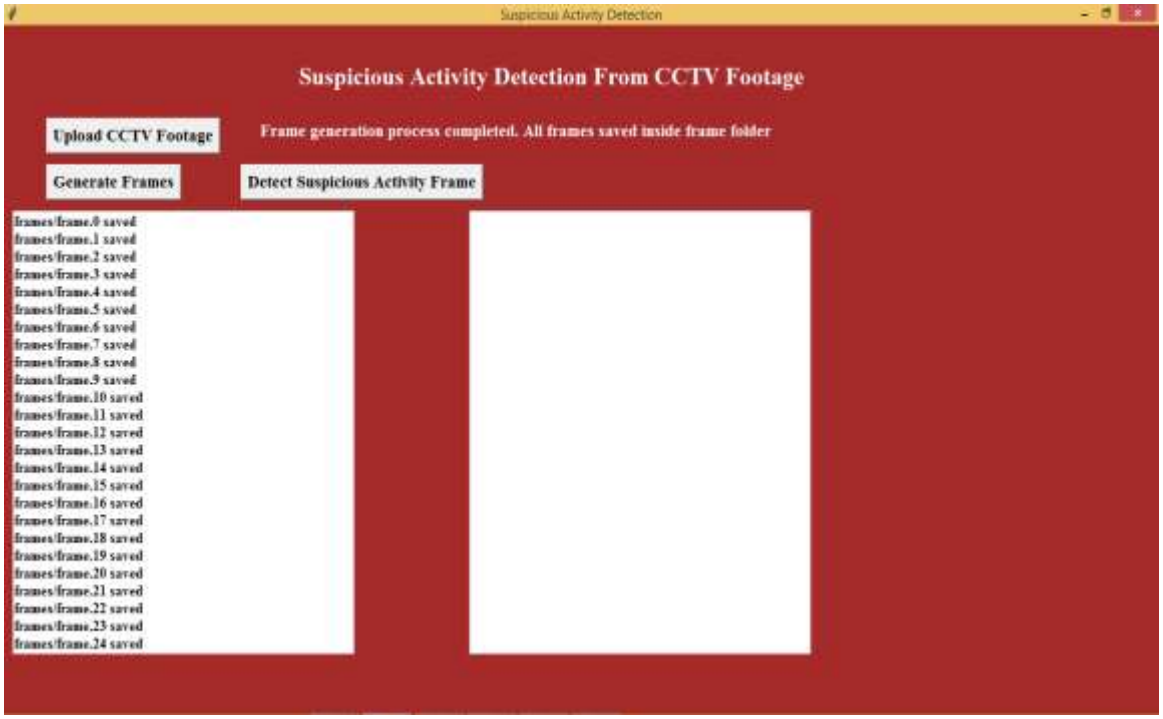


Fig8.2.1 For Generating Frames

- After the video is uploaded and upon clicking generate frames, frames start extracting from video and are saved in frames folder. Suspicious frame is detected with frame number displayed in message box.

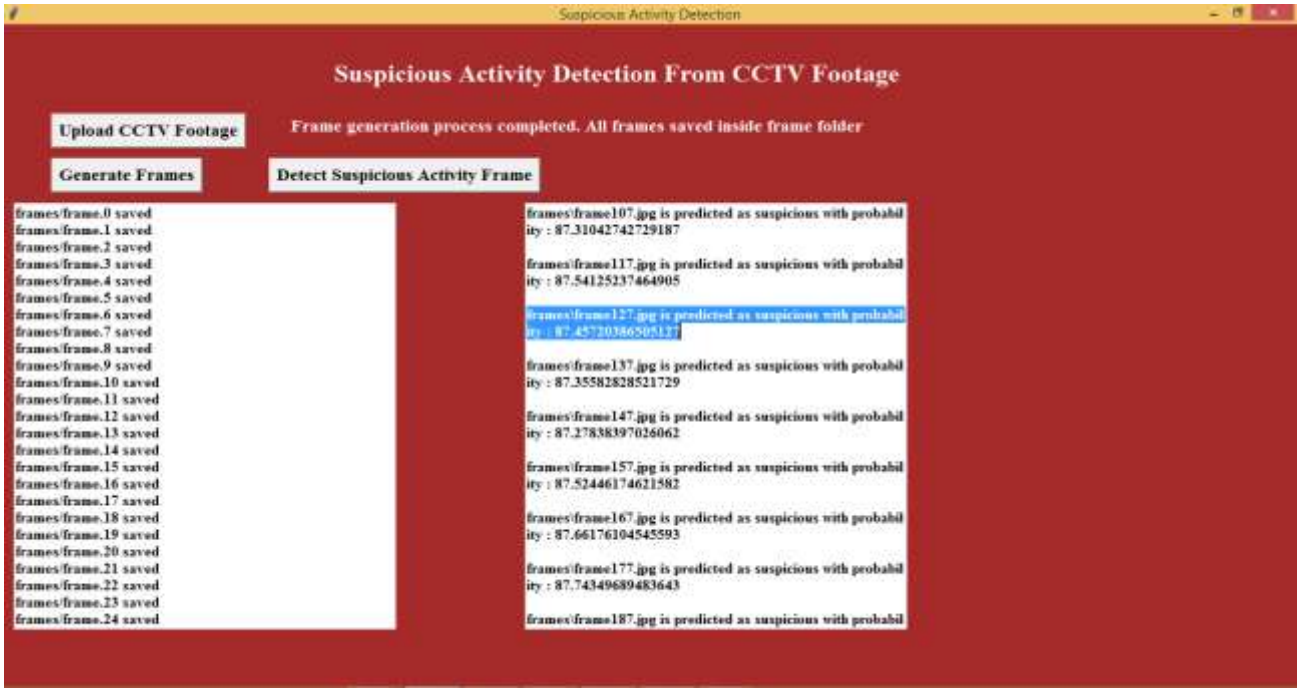


Fig 8.2.2 For Detecting Suspicious Activity

9. EXPERIMENTAL RESULTS

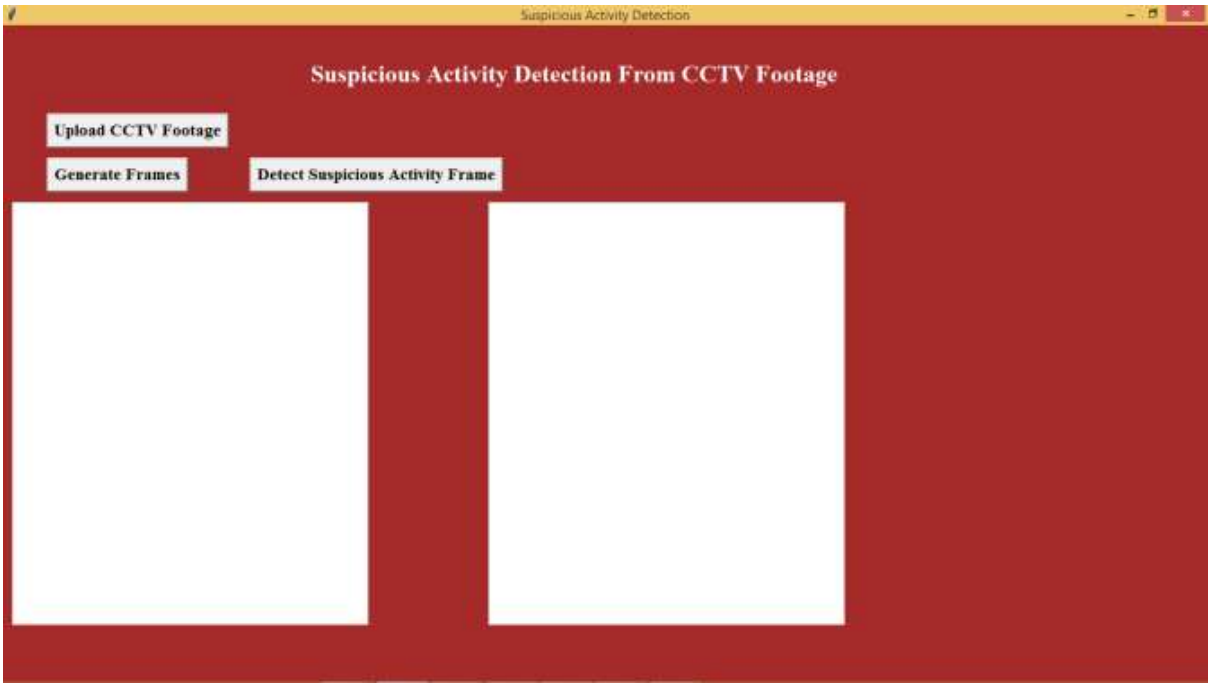


Fig 9.1 GUI

- Click on 'Upload CCTV Footage' button to upload video. In below screen,

  Video is uploaded and then click on 'Generate Frames' button to generate frame.

Fig 9.2 Uploading Video

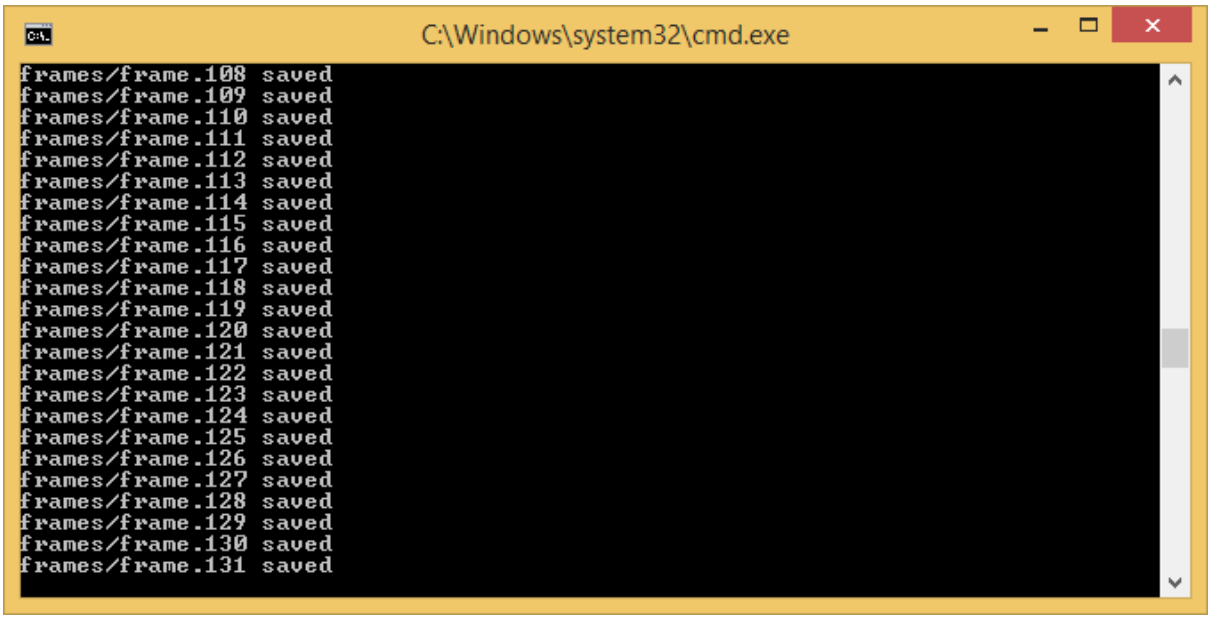


Fig 9.3 Generating Frames

- After clicking on generate frames, the frames which are extracted from video are saved into frames folder.

Fig 9.4 Saved Frames

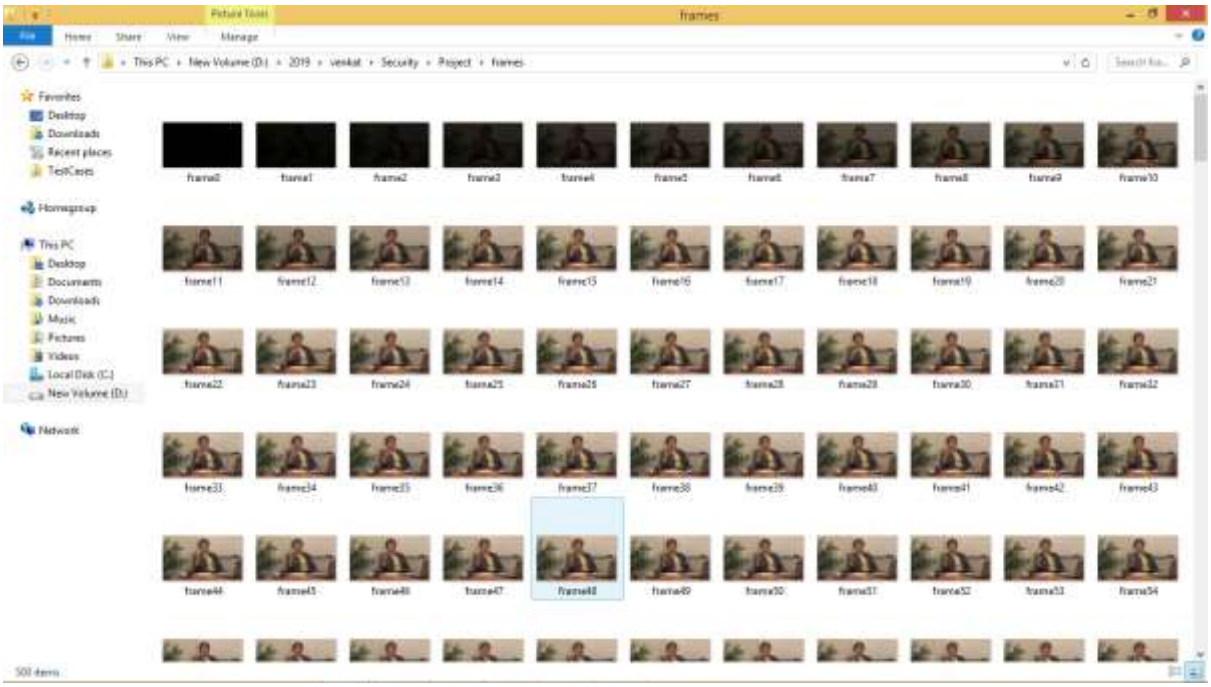


Fig 9.5 Frames Folder

- Saved frames are then processed for suspicious activity and the result is displayed in message box.

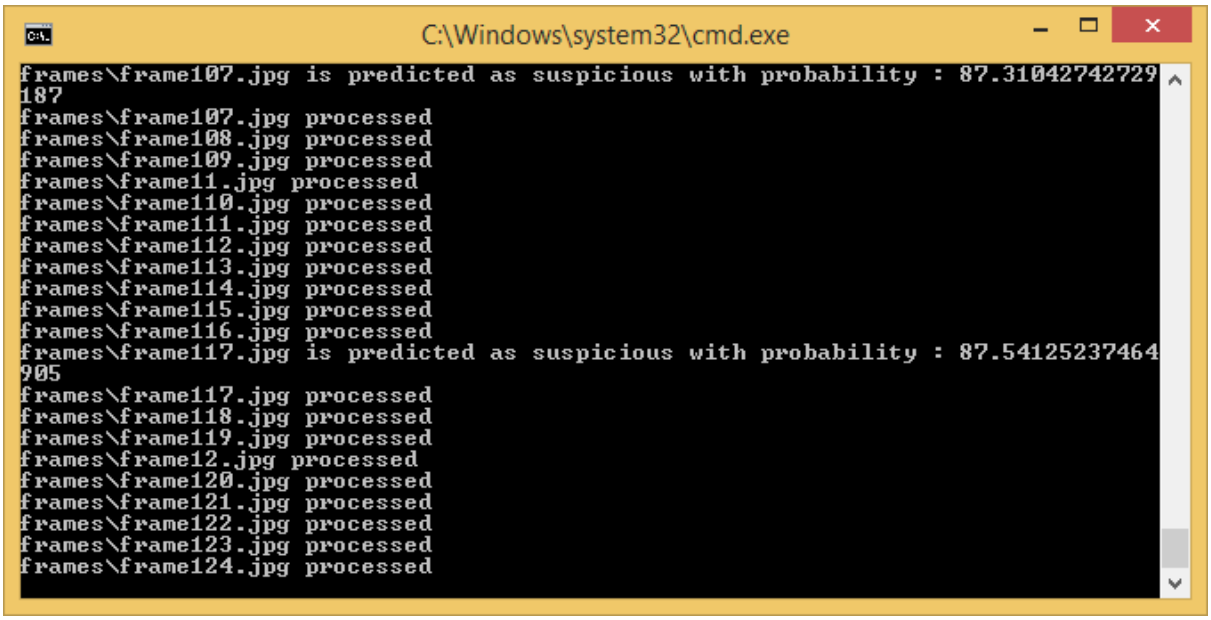


Fig 9.6 Processed Frames

Fig 9.7 Detecting Suspicious Activity

- Frame containing suspicious probability greater than 85% is detected as suspicious activity and displayed in message box along with frame number.

10.CONCLUSION AND FUTURE ENHANCEMENT

In this survey paper, we have discussed the various techniques related to abandoned object detection, theft detection, falling detection, accidents and illegal parking detection, violence detection and fire detection for the foreground object extraction, tracking, feature extraction and classification. In past decades, several researchers proposed novel approaches with noise removal, illumination handling, and occlusion handling methods to reduce the false object detection. Many researchers have also worked for making real-time intelligent surveillance system but processing rate of the video frames is not as good as required and there is no such system that has been developed with 100% detection accuracy and 0% false detection rate for videos having complex background. Much of the attention is required in the following suspicious activities detection: Abandoned object detection and theft detection Majority of the works have been done for the abandoned object detection from surveillance videos captured by static cameras. fewworks detected the static human as an abandoned object. To resolve such problems, human detection method should be very effective and system should check the presence of the owner in the scene, if owner is invisible in the scene for long duration then alarm should be raised. To resolve the problem of theft or object removal, face of the person who is picking up the static object, should match with the owner otherwise an alarm must be raised to alert the security. Future work may also resolve the low contrast situation i.e. similar color problem such as black bag and black background which lead to miss detections. Future improvements may be integration of intensity and depth cues in the form of 3D aggregation of evidence and occlusion analysis in detail. Spatial-temporal features can be extended to 3-dimensional space for the improvement of abandoned object detection method for various complex environments. Thresholding based future works can improve the performance of the surveillance system by using adaptive or hysteresis thresholding approaches. Few works have been also proposed for abandoned object detection from the multiple views captured by multiple cameras. To incorporate these multiple views to infer the information about abandoned object can also be improved. There is a large scope to detect abandoned object from videos captured by moving cameras. Falling detection Most of the works have been done for fall detection of single person in indoor videos based on human shape analysis, posture estimation analysis and motion based analysis. Future works can include the integration of multiple elderly monitoring which is able to monitor more than one person in the indoor scene. Many elder people go for morning walk everyday in public areas such as parks; to monitor these elder people, a future work can include one or more than one human fall detection from outdoor surveillance videos.

Accidents, illegal parking, and rule breaking traffic detection Several researchers have presented accidents detection, illegal parking detection and illegal U-turn detections from static video surveillance. These systems become incapable to detect these abnormal activities in more crowded traffic on roads. Future works should be based on unsupervised learning of transportation system because of no standard dataset is available for the training. Violence detection Several research works have been done for the prevention of violence activities such as vandalism, fighting, shooting, punching, and hitting. To detect such violence activities, single view static video camera has been used but sometimes this system fails in occlusion handling. Therefore, a multi-view system has been proposed by few researchers to resolve this problem but it requires important cooperation between all views at the low level steps for abnormal activity detection. Future work may be automatic surveillance system for moving videos. Improvements are required in accuracy, false alarm reduction, and frame rate to develop an intelligent surveillance system for the road traffic monitoring. Fire detection Future work can include more improvement in accuracy, frame rate, false alarms reduction and also it can be improved to detect far distant small fire covered by dense smoke.

11. REFERENCES

[1] Loganathan, S., Kariyawasam, G., & Sumathipala, P. (2019, November). Suspicious Activity Detection in Surveillance Footage. In 2019 International Conference on Electrical and Computing Technologies and Applications (ICECTA) (pp. 1-4). IEEE.

[2] Nikouei, S. Y., Chen, Y., Aved, A., Blasch, E., & Faughnan, T. R. (2019, November). I-safe: Instant suspicious activity identification at the edge using fuzzy decision making. In Proceedings of the 4th ACM/IEEE Symposium on Edge Computing (pp. 101-112).

[3] Ramachandran, S., & Palivela, L. H. (2019). An intelligent system to detect human suspicious activity using deep neural networks. Journal of Intelligent & Fuzzy Systems, 36(5), 4507-4518.

[4] Sarang, S., Shinde, H., Raut, V., Sonje, S., & Phadke, G. (2019, June). Real-Time Suspicious Activity Detection. In International Conference on Soft Computing and Signal Processing (pp. 459-466). Springer, Singapore

[5] Tripathi, R. K., Jalal, A. S., & Agrawal, S. C. (2018). Suspicious human activity recognition: a review. Artificial Intelligence Review, 50(2), 283-339.

[6] Roy, P. K., & Om, H. (2018). Suspicious and violent activity detection of humans using HOG features and SVM classifier in surveillance videos. In Advances in Soft Computing and Machine Learning in Image Processing (pp. 277-294). Springer, Cham.

[7] Kamthe, U. M., & Patil, C. G. (2018, August). Suspicious Activity Recognition in Video Surveillance System. In 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA) (pp. 1-6). IEEE.

[8] Senthilkumar, T., & Narmatha, G. (2016). Suspicious Human Activity Detection in Classroom Examination. In Computational Intelligence, Cyber Security and Computational Models (pp. 99-108). Springer, Singapore.

[9] Kenkre, P. S., Pai, A., & Colaco, L. (2015). Real-time intrusion detection and prevention system. In Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2014 (pp. 405-411). Springer, Cham.

[10] Kumar, A. S., & Singh, S. (2013, December). Detection of user clusters with suspicious activity in online social networking sites. In 2013 2nd International Conference on Advanced Computing, Networking and Security (pp. 220-225). IEEE.

[11] Gowsikhaa, D., & Abirami, S. (2012). Suspicious Human Activity Detection from Surveillance Videos. International Journal on Internet & Distributed Computing Systems.

[12] Raza, S., & Haider, S. (2011). Suspicious activity reporting using dynamic bayesian networks. Procedia Computer Science, 3, 987-991.

[13] Takai, M. (2010, December). Detection of suspicious activity and estimate of risk from human behavior shot by the surveillance camera. In 2010 Second World Congress on Nature and Biologically Inspired Computing (NaBIC) (pp. 298-304). IEEE.

[14] Yin, J., Yang, Q., & Pan, J. J. (2008). Sensor-based abnormal human-activity detection. IEEE Transactions on Knowledge and Data Engineering, 20(8), 1082-1090.

[15] Niu, W., Long, J., Han, D., & Wang, Y. F. (2004, June). Human activity detection and recognition for video surveillance. In 2004 IEEE International Conference on Multimedia and Expo (ICME) (IEEE Cat. No. 04TH8763) (Vol. 1, pp. 719-722). IEEE

12. PUBLICATIONS

CONFERENCE:

International Conference on “Innovations in Computers Networks, Computational Intelligence and IoT” (ICICCI – 21)

Paper ID: ICICCI – 21 – 0085

13.STUDENT PROFILE

Abbagoni Hima Varsha is currently pursuing her Bachelors of Technology in the stream of Computer Science and Engineering at St.Martin's Engineering College. She Completed her 12th standard from Sri Chaitanya Junior College and 10th from Divyanjali High School. She is well trained in C, Java and have basic knowledge on Python. She was the part of Employability skill development program conducted by Zensor. She is also a student in the Smart Interviews. Her participations include: Women Online Five Days Workshop on "Women in Cyber Security and Privacy in 2020" which was conducted on 6th to 10th July, a webinar on "Digital Transformation in Education Sector Post-Covid era" conducted on 11th June 2021 conducted collegedunia Two days National Level Seminar on "Recent Trends in Cloud Computing and Fog, Edge Computing" in online mode during the period from 18th June to 19 June, 2021conducted by St.Martin's Engineering College. She completed a summer internship program conducted by Goal Street On Machine Learning using python from June 2020 to August 2020 and also done a capstone project-"To preict whether the cance is Benign or Malignant".She completed few certification courses from various platforms such as Coursera, Udemy, CursaApp and Solo Learn.

Karnati Vinay kumar is currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. He pursued his intermediate from Sri Chaitanya Junior College and 10th class from P.N.M High School. His technical skills include C, Python, and Java. He also has basic knowledge of C++. He took part in the Employability Skill development Program conducted by Zensar. He is also a student of Smart Interviews. He Participated in events conducted by College Sympo Aagna 2020, A two-day national Level Technical Symposium" which was conducted from 30th to 31st January 2020, and won first prize in the poster presentation. He had Completed Six Months Summer Internship and done a mini project On Machine Learning Using Python. During his first year, his team won the first prize in the Micro project. He attended an Online Internship program For Engineering students on 3rd May 2020 and attended Python Training from 20th to 24th August. His area of interest is Python. He completed few certification courses from online platforms like Coursera and Solo Learn.

Akhila Srigadha is currently pursuing her Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. She completed her Intermediate from Sree Sandeepani junior College and 10th class from Vignan High School. Her technical skills include c and python. She also has a basic understanding of C++ and Java. She took part in Employability Skill development program conducted by Zensar. She is also a student of Smart Interviews. Her participations include: National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020, "women online workshop on "Women in Cyber Security and Privacy in 2020" which was conducted from 6th to 10th July 2020, "One Day Webinar on Internet of Things and Its Applications" conducted by Anand Institute of Higher Technology on 21st May 2020. Her areas of interest are Python, Java and Machine Learning. She completed few certification courses from online platforms like Coursera, Udemy and SoloLearn.

Tanniru Sai Vardhan is currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. He completed his intermediate from Sri Chaitanya Junior College and SSC from Bethany Academy. He is one of the members of Coders Club in our college and as well as Cyber Security learner's club. His technical skills include C, Python and Java. He also has a basic understanding of C++. He took part in Employability Skill development Program conducted by Zensar. He is also a student of Smart Interviews. His participations include: Sympo Aagna 2020, A two-day national Level Technical Symposium" which was conducted from 30th to 31st January 2020, and won first prize in poster presentation for an IOT project. He completed a summer internship in Machine Learning in association with Goal Street, from June 2020 to August 2020, and completed capstone project- "predicting whether cancer is benign or malignant". He's an active member in Martin's sports club.

His areas of interest are Python, Artificial Intelligence, Machine Learning and Deep Learning. He completed few certification courses from online platforms like Coursera, CursaApp and SoloLearn.

14.APPENDICES

Suspicious human activity recognition from surveillance video is an active research area of image processing and computer vision. Through the visual surveillance, human activities can be monitored in sensitive and public areas such as bus stations, railway stations, airports, banks, shopping malls, school and colleges, parking lots, roads, etc. to prevent terrorism, theft, accidents and illegal parking, vandalism, fighting, chain snatching, crime and other suspicious activities. It is very difficult to watch public places continuously, therefore an intelligent video surveillance is required that can monitor the human activities in real-time and categorize them as usual and unusual activities; and can generate an alert.

In today's insecure world the video surveillance plays an important role for the security of the indoor as well as outdoor places. The components of video surveillance system such as behaviour recognition, understanding and classifying the activity as normal or suspicious can be used for real time applications. In this paper the hierarchical approach is used to detect the different suspicious activities such as loitering, fainting, unauthorized entry etc. This approach is based on the motion features between the different objects. First of all the different suspicious activities are defined using semantic approach. Then the object detection is done using background subtraction. The detected objects are then classified as living (human) or non living (bag). These objects are required to be tracked which is done using correlation technique. Finally using the motion features & temporal information the events are classified as normal or suspicious. As the semantic based approach is used computational complexity is less and the efficiency of the approach is more.

A

PROJECT REPORT

On

CYBER CRIME AND SECURITY

Submitted by

1)Ms.P.Shilpa (17K81A05M3) 2)Ms.K.Subhashini(17K81A05L4)

3)Mr.K.Sumanth(16K81A05M1) 4)Mr.R.Jagan(16K81A05L8)

in partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Under the Guidance of

R.V.Sudhakar

Associate Professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ST.MARTIN'S ENGINEERING COLLEGE
An Autonomous Institute

Dhulapally, Secunderabad – 500 100

JUNE 2021

BONAFIDE CERTIFICATE

This is to certify that the project entitled CYBER CRIME AND SECURITY, is being submitted by **P.SHILPA (17K81A05M3), K.SUBHASHINI(17K81A05L4), K.SUMANTH (16K81A05M1) , R.JAGAN (16K81A05L8)** in partial fulfillment of the requirement for the award of the degree of **BACHELOR OF TECHNOLOGY IN COMPUTER SCIENCE ENGINEERING** is recorded of bonafide work carried out by them. The result embodied in this report have been verified and found satisfactory.

<Signature>
R.V.SUDHAKAR
Department of CSE

Head of the Department
Dr.M.NARAYANAN
Department of CSE

Internal Examiner

External Examiner

Place:

Date:

DECLARATION

We, the student of **Bachelor of Technology** in Department of Computer Science and Engineering', session: 2017 – 2021, St. Martin's Engineering College, Dhulapally, Kompally, Secunderabad, hereby declare that work presented in this Project Work entitled **Cyber crime and security** is the outcome of our own bonafide work and is correct to the best of our knowledge and this work has been undertaken taking care of Engineering Ethics. This result embodied in this project report has not been submitted in any university for award of any degree.

P. SHILPA (17K81A05M3)

K. SUBHASHINI (17K81A05L4)

K. SUMANTH (16K81A05M1)

R. JAGAN (16K81A05L8)

ABSTRACT

Cybercrime is inescapable, ubiquitous and increasingly linked with different parts and areas of criminal environs. This evolution and network gave rise to cyber space which controls and manages to provide equal opportunities and facilities to all the people to access any kind of information. Due to gradually increase of the internet users and netizens, abusage of technology is broadening gradually which tends to cyber crimes. Cyber crime is basically an unlawful act that leads to criminal activity. Cyber Security, a mechanism by which computer information and the equipments are protected from unauthorized and illegal access. This paper illustrates and focuses on cybercrime, its impact on society, types of threats, and cyber security. Nowadays Computer crime issues and thefts have become tremendously high-profile, particularly those surrounding copyright infringement, hacking, child pornography, child grooming, and spoofing.

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible and whose encouragements and guidance have crowded effects with success.

We extended our deep sense of gratitude to Principal**,** **Dr. P. SANTOSH KUMAR PATRA**, St. Martin's Engineering College, Dhulapally, for permitting us to undertake this project.

We are also thankful to **Dr. M.NARAYANAN**, Head of the Department, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

We would like to thank **Dr. T. POONGOTHAI,** Department of Computer Science and Engineering (AI & ML) for her encouragement and insightful comments and as well as our project coordinators **Dr. B.RAJALINGAM**, Associate Professor and **Dr. R.SANTHOSHKUMAR**, Associate Professor, in Department of Computer Science and Engineering for their valuable support.

We would like to express our sincere gratitude and indebtedness to our project supervisor **R.V.SUDHAKAR**, Associate professor, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

Finally, we express thanks to all those who have helped us successfully to completing this project. Furthermore, we would like to thank our family and friends for their moral support and encouragement.

We express thanks to all those who have helped us in successfully completing the project.

P.SHILPA (17K81A05M3)

K.SUBHASHINI (17K81A05L4)

K.SUMANTH (16K81A05M1)

R.JAGAN (16K81A05L8)

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

<AVI>	Audio Video Interlace
<BMP>	Bitmap
<CPU>	Central Processing Unit
<GB>	Giga Bytes
<GUI>	Graphical User Interface

GLOSSARY OF TERMS

TERM	MEANING
Machine Learning	Machine learning is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention.
Deep Learning	Deep learning is part of a broader family of machine learning methods based on artificial neural networks with representation learning.
CNN	A convolutional neural network (CNN, or ConvNet) is a class of deep neural network, most commonly applied to analyse visual imagery.
Suspicious Activity	*Suspicious behaviour* or activity can be any action that is out of place and does not fit into the usual day-to-day activity of our campus community.
Background Extraction	*Background extraction* is a fast and efficient moving object segmentation algorithm.
Foreground Object Extraction	Foreground object extraction from the video is an initial and important step of suspicious human activity recognition.

1.INTRODUCTION

Computer fraud can be a untrustworthy misrepresentation of the fact proposed to prompt another to abstain from doing something that causes loss. Computer crime can be summarized as a criminal activity which involves information technology infrastructure, in addition to unauthorized access, illegal interception, any data interference, computer or systems interference, abusage of devices, forgery, blackmail, embezzlement, and some electronic fraud. There exits privacy issues whenever any confidential information or data is hijack or lost, either lawfully or otherwise. The very first crime that was recorded, took place in 1820 in France, Joseph-Marie Jack quad, a textile manufacturer, produced a device namely loom which allowed the continuous repetition of series of steps involved in the weaving of some special fabrics. This leads to a kind of fear in employee‟s minds and they committed sabotage. Cyber crime cells are there in states basically to handle these crimes, and to expel or punish the netizens or criminals committing any of the cyber crime[1]. It basically ranges from theft of an individual‟s identity entire disruption of a particular country‟s Internet and network connectivity due to massive attacks across its networking resources. In this digital age, online communication now become a norm, the internet users and the government are at a enlarged risk of becoming the bull's-eye of the cyber attacks. Cyber crime can cause harm to any organisation.

To fight the fast-spreading cybercrime, governments and businesses must have collaboration globally basically to develop any impressive model that somehow controls the threat. The internet is basically used for the betterment of life, to make people aware of world-wide activities, enhances the speed of life as well and makes users technically strong and up-to-the-mark. As the use of technology is increasing day-by-day, the crime is also increasing gradually. It covers all the forms of crimes and thefts related to computer networks. Some of the criminals are technically expert and educated having deeper and remarkable knowledge regarding the technology. Hacking of the ATM password, transferring the money by hacking the bank account details of the victim's account to theirs, some pornography issues etc are some of the thefts that are handled by educated people [2]. There is an urge to implement some of the rules and regulations, to tackle and handle these crimes governing cyber space particularly known as Cyber Law.

1.1 PROJECT OVERVIEW:

Cyber crime on the rise As per the cyber crime data maintained by the National Crime Records Bureau (NCRB), a total of 217, 288, 420 and 966 Cyber Crime cases were registered under the Information Technology Act, 2000 during 2007, 2008, 2009 and 2010 respectively. A few cyber crime against individuals are: Email spoofing: This technique is a forgery b)of an email header.

General Public. We can say that it is an unlawful acts wherein the computer either a tool or target or both. Organizations and Cyber crime: An Analysis of ... Alternatively, it could be a short-lived project driven by a group that undertakes a specific online crime and/or targets a particular victim or group. This term has nowhere been defined in any statute/Act passed or enacted by the Indian Parliament

1.2 PROJECT OBJECTIVES:

Objective 1: Safeguard national critical information infrastructure (CII) **Objective** 2: Respond to, resolve, and recover from cyber incidents and attacks through timely information sharing, collaboration, and action. **Objective** 3: Establish a legal and regulatory framework to enable a safe and vibrant cyberspace.

This paper's approach is to use historical crime records to provide an understanding and analysis which helps law makers and law protector to take preventive measures by using the inferences drawn from the outputs of crime records. This paper used data processing technologies which are Hadoop and its ecosystem for analysis and processing of the data quickly and parallely and applied various techniques like classification, attribute search etc. for extracting the required information.

1.3 ORGANIZATIONS OF CHAPTERS:

Cyber organized criminals have engaged in a variety of cybercrimes, including fraud, hacking, malware creation and distribution, DDoS attacks, blackmail, and intellectual property crime (see Cybercrime Module 2 on General Types of Cybercrime and Cybercrime Module 11 on Cyber-Enabled Intellectual Property Crime), such as the sale of counterfeit or falsified trademarked products (e.g., apparel, accessories, shoes, electronics, medical products, automobile parts, etc.) and the labels, packages, and any other identifying designs of these products (Albanese, 2018; Europol, 2018; Broadhurst et al., 2018; Maras, 2016). These types of cybercrimes cause financial, psychological, economic, and even physical harm (especially counterfeit electronics and automobile parts, as well as falsified medical products, defined by the World Health Organization as "deliberately/fraudulently misrepresent their identity, composition or source," see WHO, 2017), and have been used to fund other forms of serious crime, such as terrorism (Binder, 2016).

Organized criminal groups have also profited and/or otherwise benefited from illicit products and services offered online. For example, the creator of the Butterfly Bot advertised this malware online as capable of taking control of Windows and Linux computers (BBC News, 2013). The creator of the Butterfly Bot also sold plug-ins that modified the functions of the malware, and also offered to create customized versions of the malware for paying customers (FBI, 2010). Various online criminal networks deployed the Butterfly Bot, the largest application of this malware resulted in the Mariposa botnet, which infected 12.7 million computers around the world (BBC News, 2013).

2. LITERATURE SURVEY

1) REVIEW ON CYBER CRIME AND SECURITY

Cybercrimes is defined as the criminal activities carried out by means of using digital devices like computers through the internet. Basically, a crime committed by using the internet is called a cyber-crime. Now a day's, information is wealth and also to earn money in an illegal way, cyber-attacks are happening, and data is been stolen from the servers or money is been stolen in an illegal way. So, this paper describes the list of cyber threats happened around the world until now and its prevention mechanisms. Also, the cyber threat predictions in the upcoming year is also discussed in the final section and cyber threat analysis for January 2019 is also been discussed

2) Cybercrime: A threat to Network Security

Digital technology is encompassing in all walks of life, all over the world and has brought the real meaning of globalisation. At the one end cyber system provides opportunity to communicate and at the other end some individuals or community exploit its power for criminal purposes. Criminals exploit the Internet and other network communications which are international in scope. Situation is alarming; Cybercrime is an upcoming and is talk of the town in every field of the society/system. Theoretically and practically this is a new subject for researchers and is growing exponentially. Lot of work has been done and endless has to be go because the invention or up gradation of new technology leads to the technical crime i.e. the digital or we can say the cybercrime or e-crime. This is because every day a new technique is being developed for doing the cybercrime and many times we are not having the proper investigating method/model/technique to tackle that newly cybercrime. In the present day world, India has witnessed an unprecedented index of cybercrimes whether they pertain to Trojan attacks, salami attacks, e-mail bombing, DOS attacks, information theft, or the most common offence of hacking. Despite technological measures being adopted by corporate organizations and individuals, we have witnessed that the frequency of cybercrimes has increased over the last decade any information easily within

a few seconds by using internet which is the medium for huge information and a large base of communications around the world. Certain precautionary measures should be taken by all of us while using the internet which will assist in challenging this major threat cybercrime. In this paper, we have discussed various categories of cybercrime and cybercrime as a threat to person, property, government and society and we have suggested various preventive measures to be taken to snub the cybercrime

2) Cyber security: challenges for society- literature review

Infrastructure, it Cyber security is the activity of protecting information and information systems (networks, computers, data bases, data centers and applications) with appropriate procedural and technological security measures. Firewalls, antivirus software, and other technological solutions for safeguarding personal data and computer networks are essential but not sufficient to ensure security. As the authors' nation rapidly building its Cyber-is equally important that they educate their population to work properly with this infrastructure. Cyber-Ethics, Cyber-Safety, and Cyber-Security issues need to be integrated in the educational process beginning at an early age.

3) A Literature Review on Cyber Security Automation for Controlling Distributed Data

In the today computer engineering era data protection and cyber security are becoming challenge for scientist. Cyber security is the activity of protection of information and information systems like network, computers, data base, data centre and application. Most of the government and private organizations are trying to protect our data and information from cyber terrorist or hackers. Cybersecurity plays important role in information system as well as data sharing. For the protection of important information and data most of the software was developed by many organization using different techniques. In this paper we are collecting literature data regarding cyber security as well as different automation software for securing our data. For developing software many

techniques are used like one time password, event log analysis, malicious attack

detection, and virtualization.demonstrate that the proposed scheme provides efficient searching to good peers while penalizing the malicious peers by increasing their search times. Keywords: P2P network–topology adaptation–trust management–semantic community–malicious peer.

2.1 SURVEY ON BACKGROUND:

The Internet connection was available to general public in 1989 and the first ever website was launched in 1991. Today, there are more than a billion websites and the number of internet users is increasing with each passing day. Figure 1 shows exponential growth of the internet users from 1995 to 2017. There will be 6 billion internet users by 2022 which equals 75 percent of the estimated world population of 8 billion. This prevalent and dominant nature of computers and interenet in our life has made cybercrimes more prominent (Morgan, 2017).

2.2 CONCLUSIONS ON SURVEY:

Even large organizations with top talent and significant resources devoted to cybersecurity have suffered major cybersecurity compromises, and organizations that do not have such levels of talent or resources face even greater challenges. More highly skilled workers in cybersecurity roles would help the nation respond more robustly to the cybersecurity problems it faces. All organizations need to understand their threat environment and the risks they face, address their cybersecurity problems, and hire the most appropriate people to do that work.

3. SOFTWARE AND HARDWARE REQUIREMENTS

3.1 SOFTWARE REQUIREMENTS:

- Operating system : - Windows XP/7.
- Coding Language : ASP.NET, C#.NET
- Data Base : MS SQL SERVER 2005

3.2 HARDWARE REQUIREMENTS:

- System : Pentium IV 2.4 GHz.
- Hard Disk : 40 GB.
- Floppy Drive : 1.44 Mb.
- Monitor : 15 VGA Colour.
- Mouse : Logitech.
- Ram : 1GB

4 . SOFTWARE DEVELOPMENT ANALYASIS

4.1 OVERVIEW OF PROBLEM:

Cybercrimes are unlikely to fade any time soon. Instead, many speculate that cybercrimes will outnumber traditional crimes soon. With that said, cybercriminals continue to prey upon multiple targets. These primary targets fall into four main categories: federal agencies (i.e., the White House, Congress, Department of Homeland Security, and other government agencies), the military, businesses/corporations, and civilians. An understanding of the targets of cybercrime is just as important as an understanding of the motivations of the cybercriminal. Motivations may be based on power reassurance (i.e., the crime provides self-confidence), asserting power, anger or retaliation, instilling fear and intimidation or sadistic behaviors, and lastly, profit (Turvey, 2004). The next part of this chapter focuses on some of the current cybercrimes and impending threats regarding hacking, identity theft and fraud, child sexual exploitation and online pornography, and cyberstalking and digital harassment. Included is a description of the extent to which these crimes exist, the primary targets as well as some of the cybercriminal's methods and motivations.

4.2 DEFINE THE PROBLEM:

New technologies create new criminal opportunities but few new types of crime. What distinguishes cybercrime from traditional criminal activity? Obviously, one difference is the use of the digital computer, but technology alone is insufficient for any distinction that might exist between different realms of criminal activity. Criminals do not need a computer to commit fraud, traffic in child pornography and intellectual property, steal an identity, or violate someone's privacy. All those activities existed before the "cyber" prefix became ubiquitous. Cybercrime, especially involving the Internet, represents an extension of existing criminal behaviour alongside some novel illegal activities.

4.3 MODULES OVERVIEW:

MODULES:

1. Network Monitor
2. Server
3. User

1. Network Monitor:

In this Module All network transaction can able to monitor. We can able see the Attacks graph also.

2.Server:

In This Module how many nodes are connected in the network .when the data was upload into server.

When the data was attacked by the attacker.

3. User:

In This modules user can able simulate the network. He can able to create a network how many nodes he want in the network. Upload data into the server. He can to download data from the server.

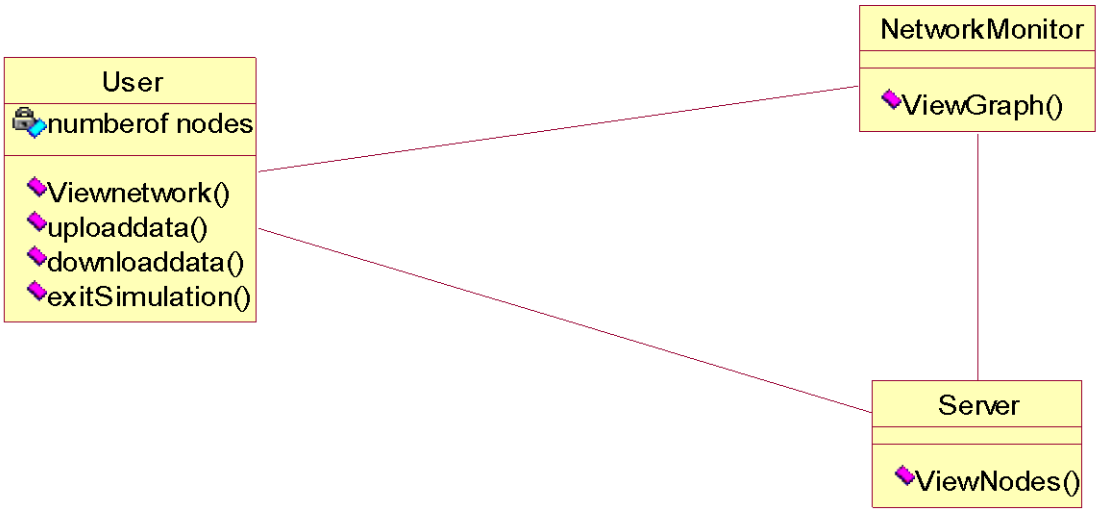
To implement this project we have designed 3 modules

1) Server module: In this module server will run in infinite loop and wait for client request and if client send any data then it will receive and store it and if client send request to download a file then server will send that file back to client.
2) Network Monitor module: This is a MapReduce application which will inspect all incoming packets and then analyse it and if all packets are server capacity limit then it will forward packets to server for storage else discard it.
3) User's module: This is a simulation module where users will send request to server for file upload or download

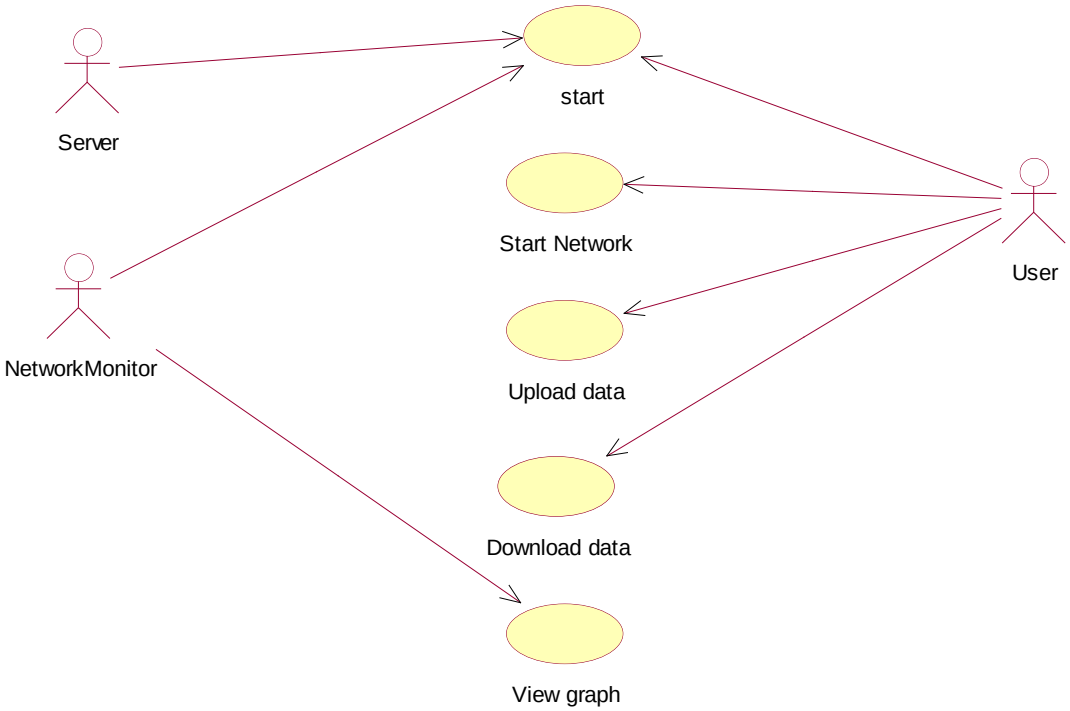
5.PROJECT SYSTEM DESIGN

5.1 UML DIAGRAMS:

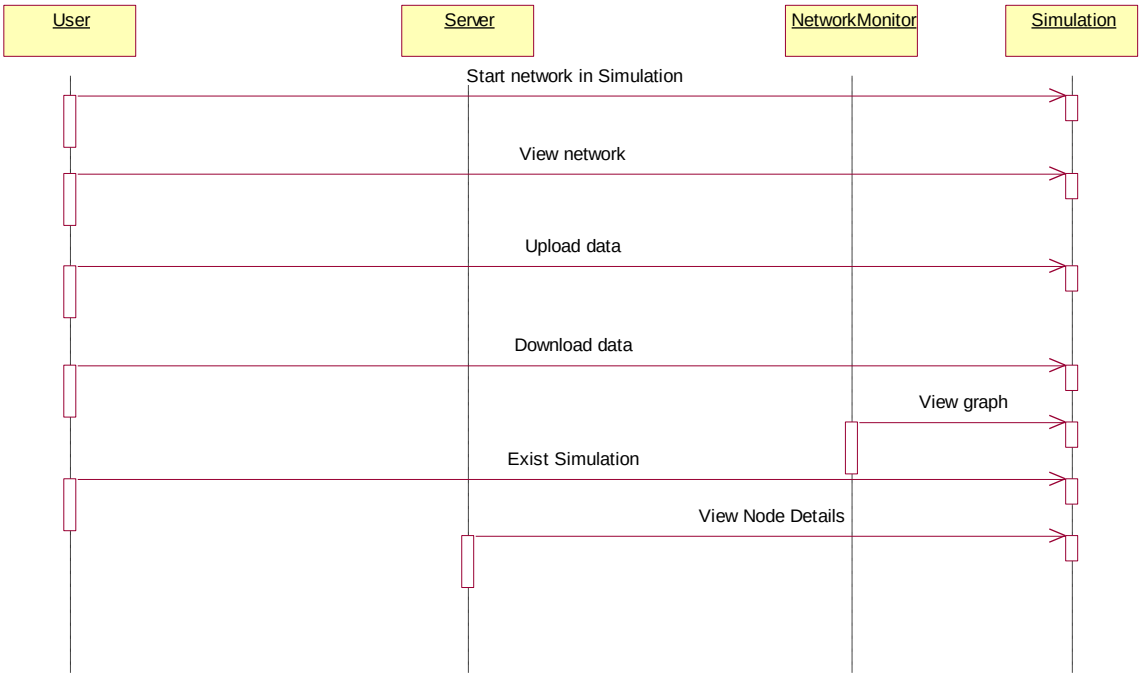
Class Diagram:



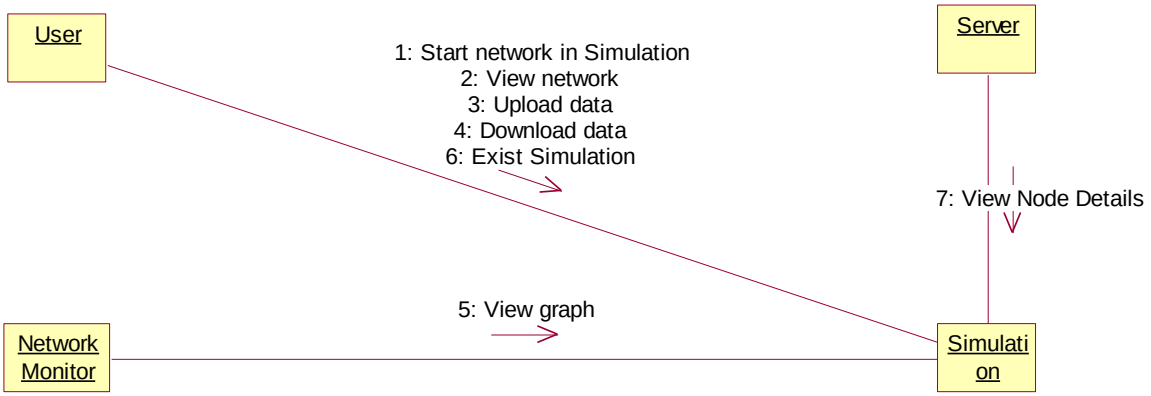
Use Case Diagram:



Sequence Diagram:



Collaboration Diagram:



6. PROJECT CODING

6.1 INPUT DESIGN AND OUTPUT DESIGN:

INPUT DESIGN:

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the following things:

- What data should be given as input?
- How the data should be arranged or coded?
- The dialog to guide the operating personnel in providing input.
- Methods for preparing input validations and steps to follow when error occur.

OBJECTIVES:

1. Input Design is the process of converting a user-oriented description of the input into a computer-based system. This design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information.

2. It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The goal of designing input is to make data entry easier and to be free from errors. The data entry screen is designed in such a way that all the data manipulates can be performed. It also provides record viewing facilities.

3. When the data is entered it will check for its validity. Data can be entered with the help of screens. Appropriate messages are provided as when needed so that the user will not be in maize of instant. Thus the objective of input design is to create an input layout that is easy to follow

OUTPUT DESIGN:

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making.

1. Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is designed so that people will find the system can use easily and effectively. When analysis design computer output, they should Identify the specific output that is needed to meet the requirements.

2. Select methods for presenting information.

3. Create document, report, or other formats that contain information produced by the system.

The output form of an information system should accomplish one or more of the following objectives.

- ❖ Convey information about past activities, current status or projections of the
- ❖ Future.
- ❖ Signal important events, opportunities, problems, or warnings.
- ❖ Trigger an action.
- ❖ Confirm an action.

7. PROJECT TESTING

7.1 SYSTEM TESTING:

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirements

TYPES OF TESTS:

Unit testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration specifications and contains clearly defined inputs and expected results.

Integration testing

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

Functional test

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items:

Valid Input : identified classes of valid input must be accepted.

Invalid Input : identified classes of invalid input must be rejected.

Functions : identified functions must be exercised.

Output : identified classes of application outputs must be exercised.

Systems/Procedures: interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

Unit Testing:

Unit testing is usually conducted as part of a combined code and unit test phase of the software lifecycle, although it is not uncommon for coding and unit testing to be conducted as two distinct phases.

Test strategy and approach

Field testing will be performed manually and functional tests will be written in detail.

Test objectives

- All field entries must work properly.
- Pages must be activated from the identified link.
- The entry screen, messages and responses must not be delayed.

Features to be tested

- Verify that the entries are of the correct format
- No duplicate entries should be allowed
- All links should take the user to the correct page.

Integration Testing

Software integration testing is the incremental integration testing of two or more integrated software components on a single platform to produce failures caused by interface defects.

The task of the integration test is to check that components or software applications, e.g. components in a software system or – one step up – software applications at the company level – interact without error.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

Acceptance Testing

User Acceptance Testing is a critical phase of any project and requires significant

participation by the end user. It also ensures that the system meets the functional requirements.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

7.2 White Box Testing:

White Box Testing is a testing in which in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is purpose. It is used to test areas that cannot be reached from a black box level.

7.3 Black Box Testing:

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box .you cannot “see” into it. The test provides

inputs and responds to outputs without considering how the software

8. OUTPUT SCREENS

8.1 USER INTERFACES:

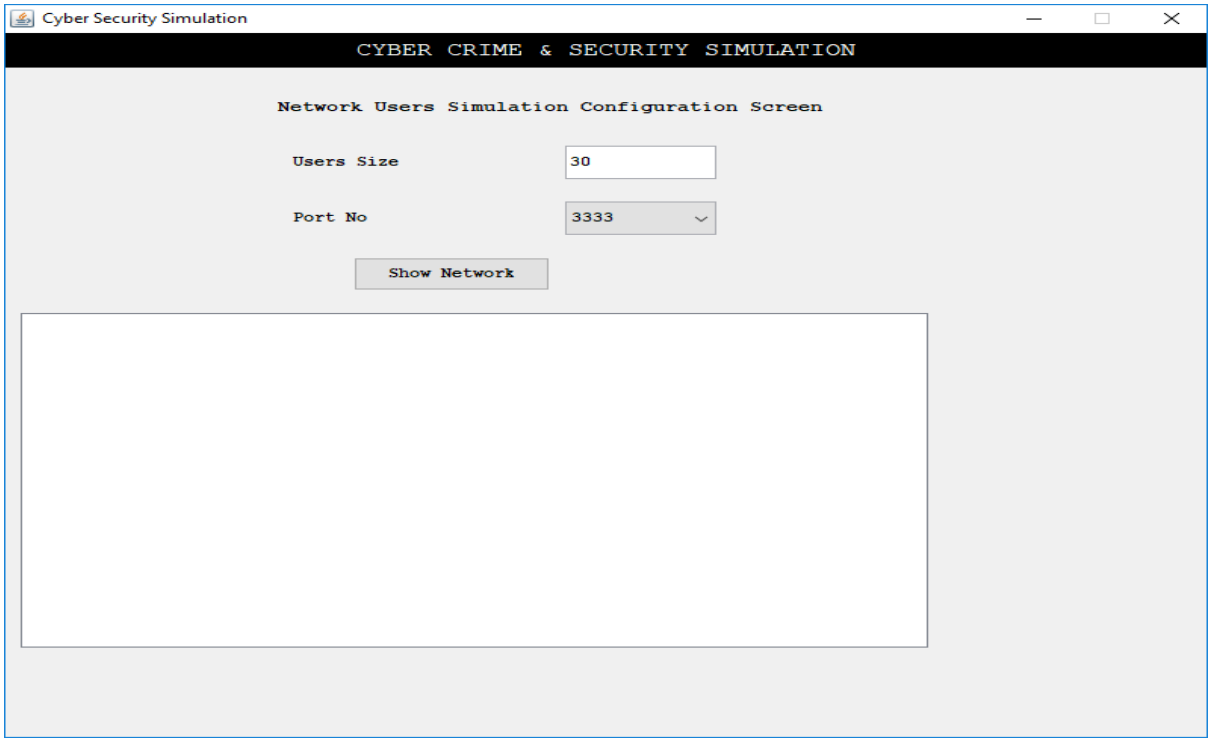


Fig:8.1.1 User interface

The user interface (UI) comprises everything a user may interact with when navigating an application or website, from the menu bar and toolbar to windows, buttons, and other controls.

A well-designed UI will enhance the user experience (UX), allowing the user to interact with software in a natural and intuitive way and enabling them to perform the actions they desire without difficulty.

Thanks to its ability to positively impact the UX, the UI has grown to become an integral component of user-facing IT. But, as a constant feature within the browser, its sustained presence has garnered the attention of hackers who can target it to perpetrate cyber attacks.

8.2 OUTPUT SCREENS:

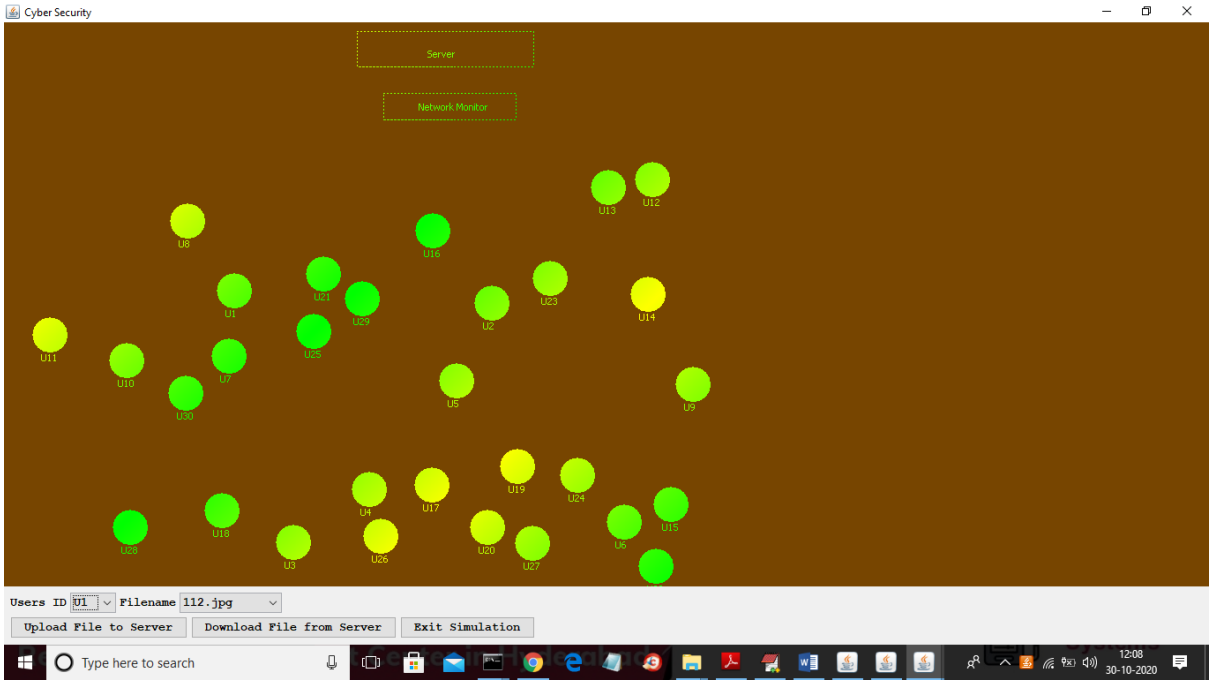


Fig:8.2.1 Users screens

9.EXPERIMENTAL RESULT

To run project first double click on ‘run.bat’ file from Server folder to start server and to get below screen

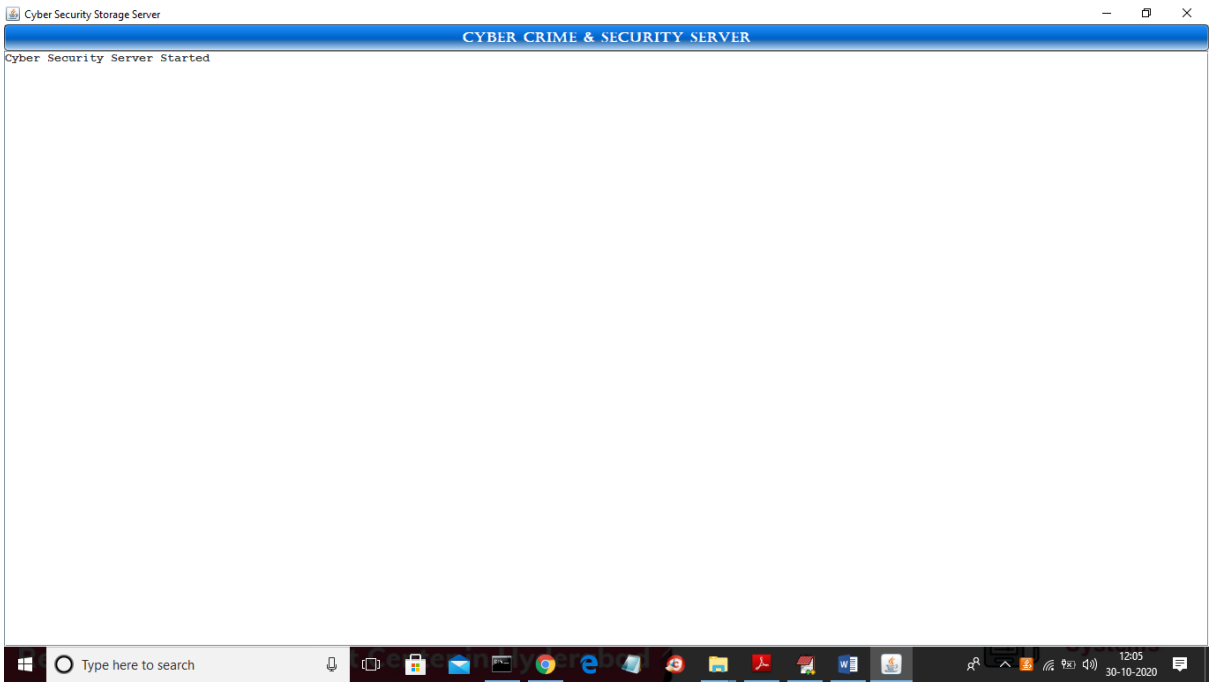


Fig:9.1 server

In above screen server started and now double click on ‘run.bat’ file from ‘NetworkMonitor’ folder to start map reduce monitoring

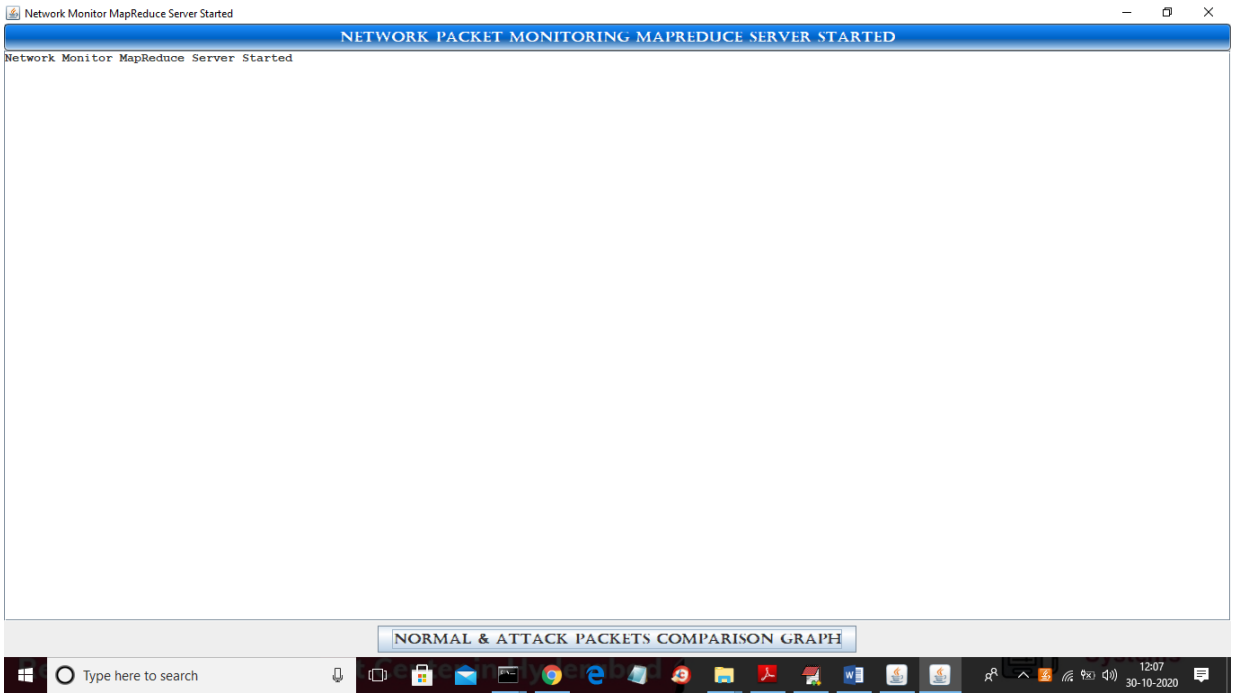


Fig:9.2 Network monitor

In above screen monitor server also starts and now double click on ‘run.bat’ file from Users folder to get below screen

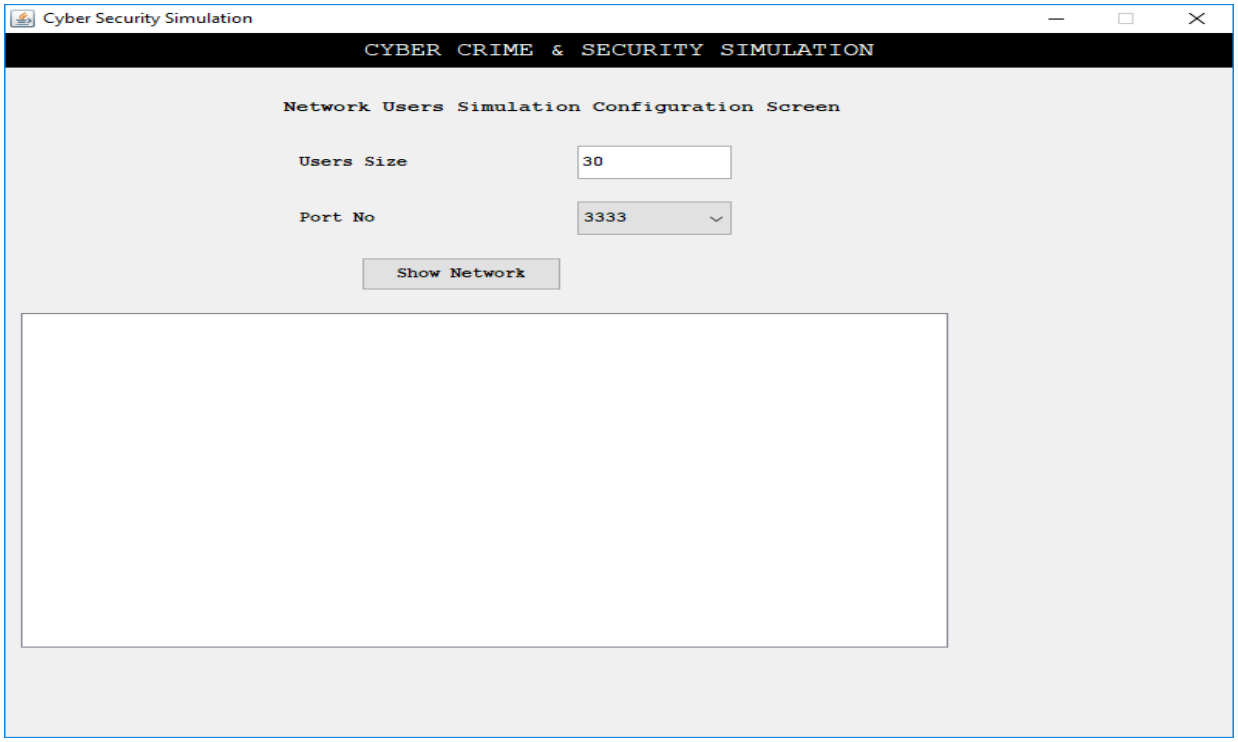


Fig:9.3 User size

In above screen I entered users size 30 and server port number 3333 and now click on ‘Show Network’ button to get below screen of simulation

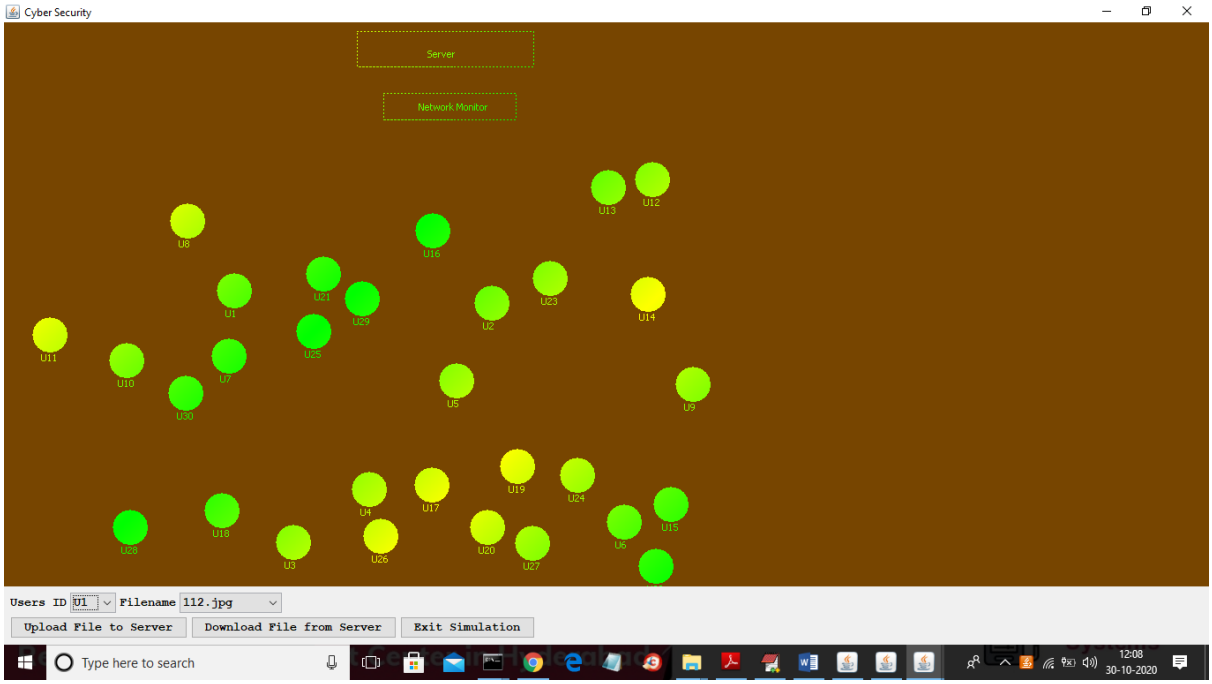


Fig: 9.4 Users screen

In above screen each circle represents one users and rectangle represents monitor and server and in above screen select any user from dropdown box and then click on ‘Upload File to Server’ button to upload file

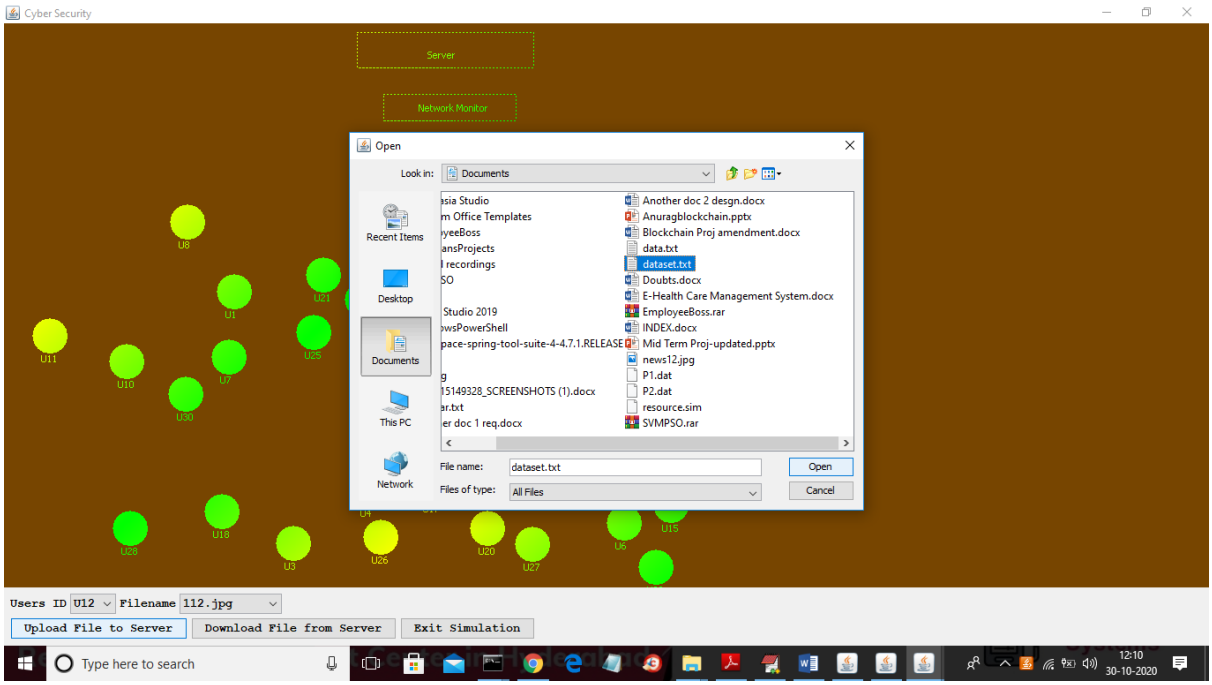


Fig:9.5 Upload file to server

In above screen I selected U12 and then click on upload button and then select file as ‘dataset.txt’ and now click open button to upload file to server

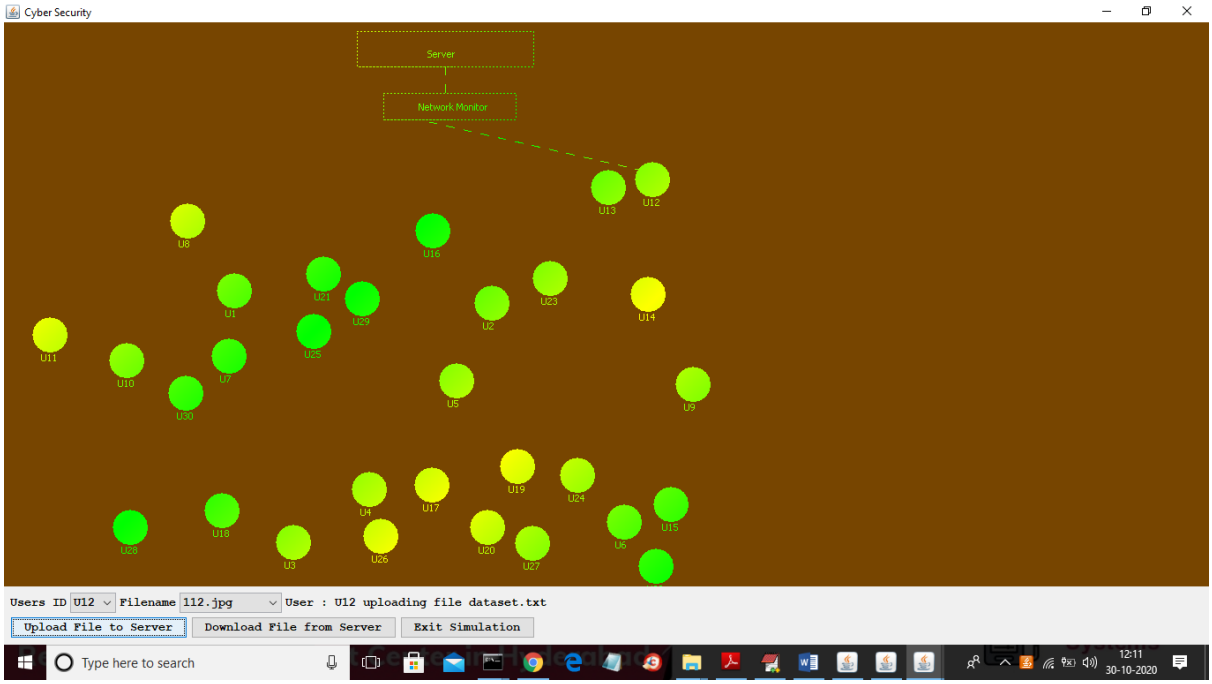


Fig: 9.6uploaded to server

In above screen we can see U12 is sending data to server and its analysing by network monitor and below is the result

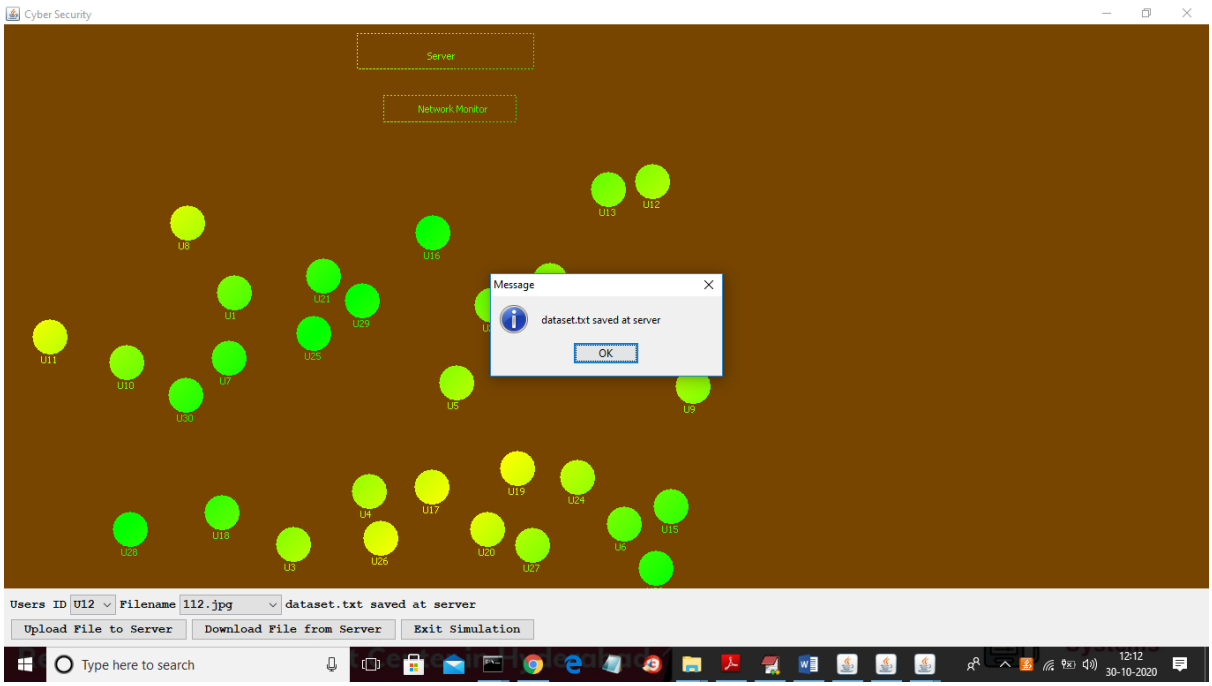


Fig: 9.7 saved to server

In above screen we can see file is in limit so network monitor allow it saved on server and the uploaded dataset.txt file we can see in drop down box

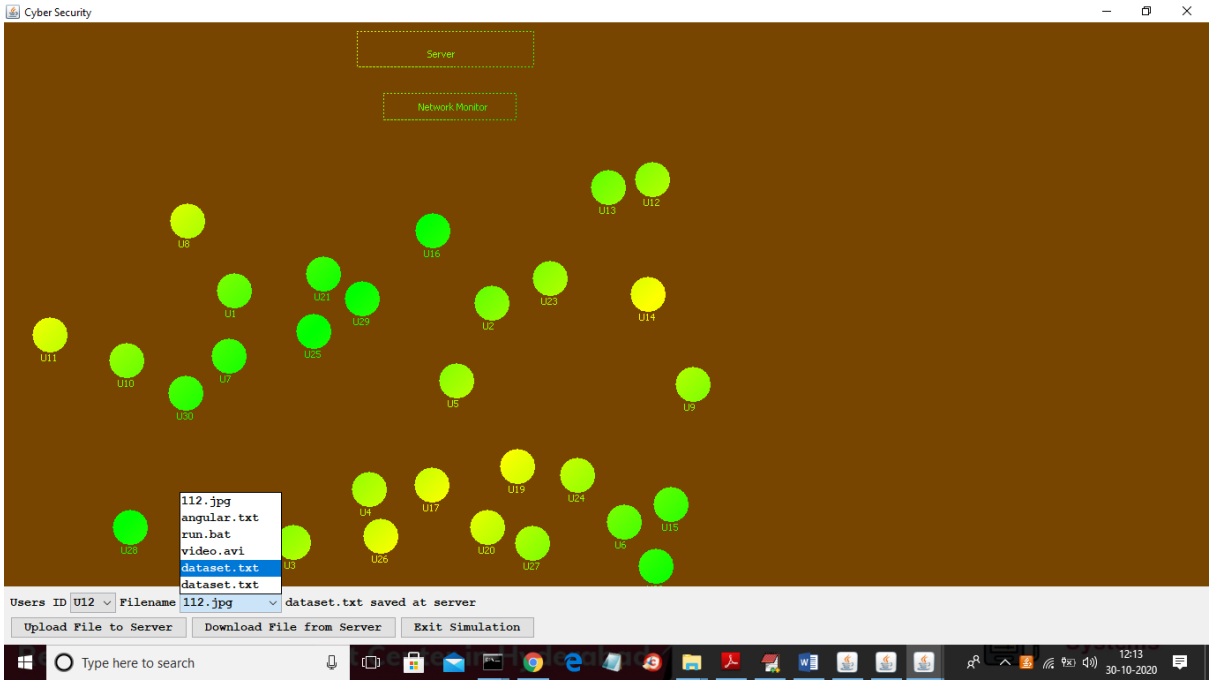


Fig: 9.8 File from server

Now in below screen we can see network monitor status

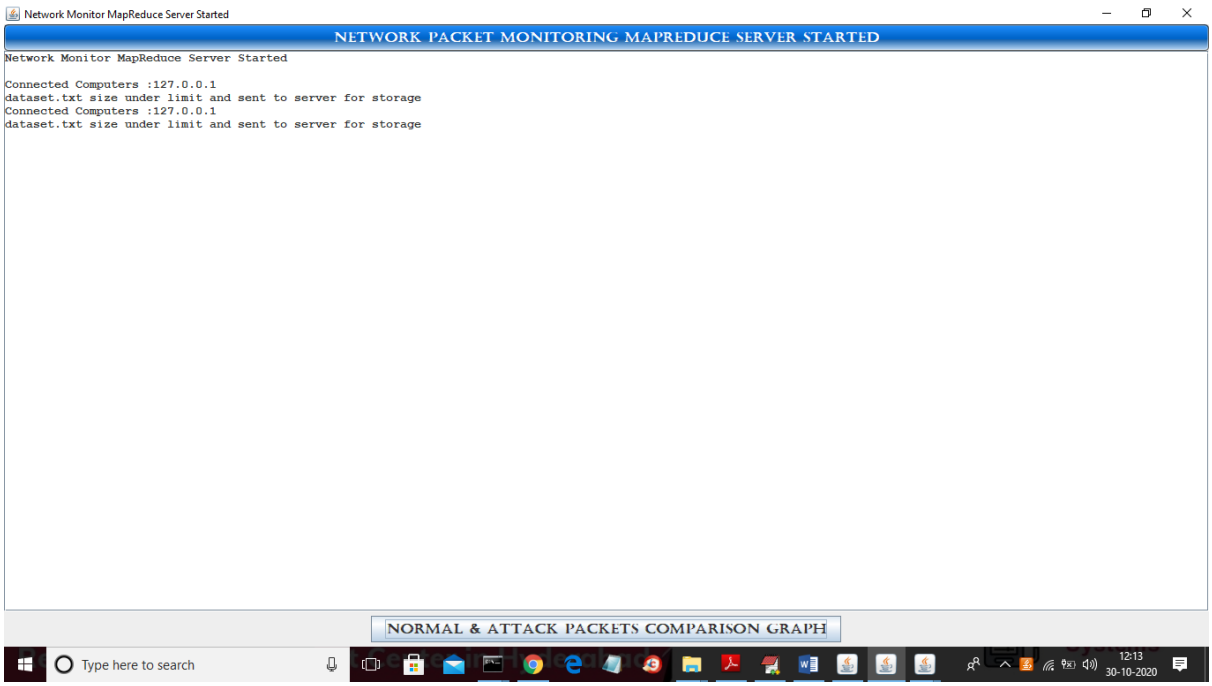


Fig: 9.9 Connected computers

Now in below screen we can map reduce processing at network monitor console

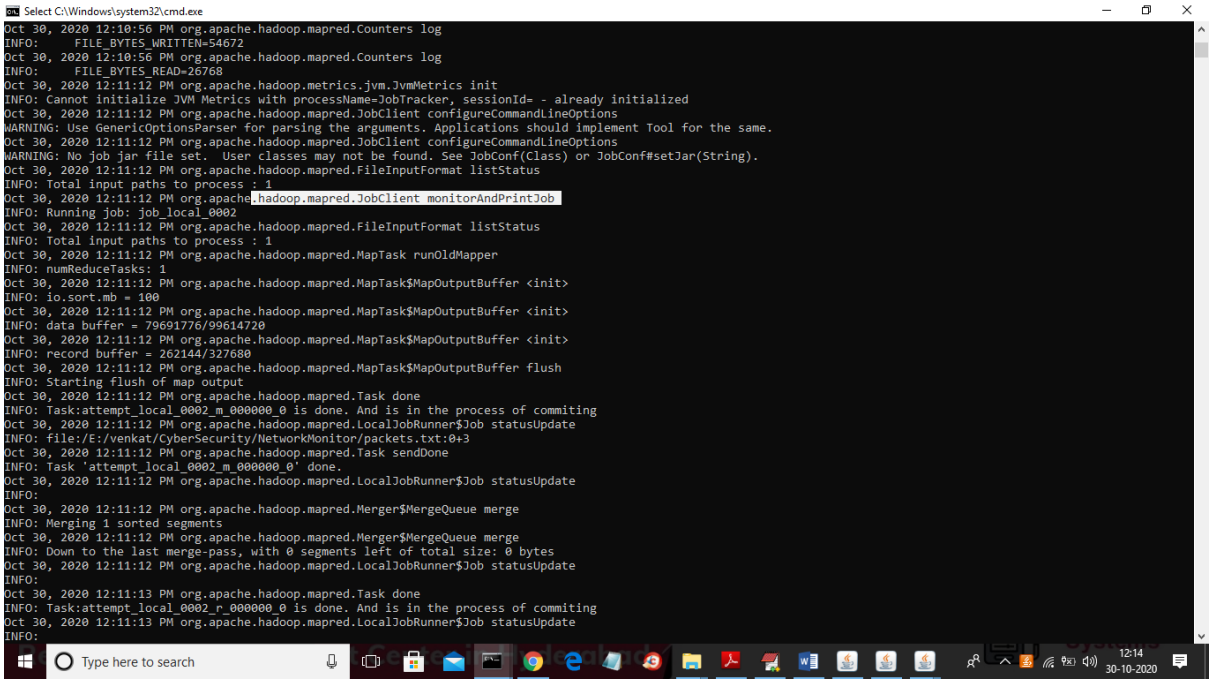


Fig: 9.10 hadoop map reduce monitoring job

In above screen we can see Hadoop map reduce monitoring job and now we will upload big file and see result

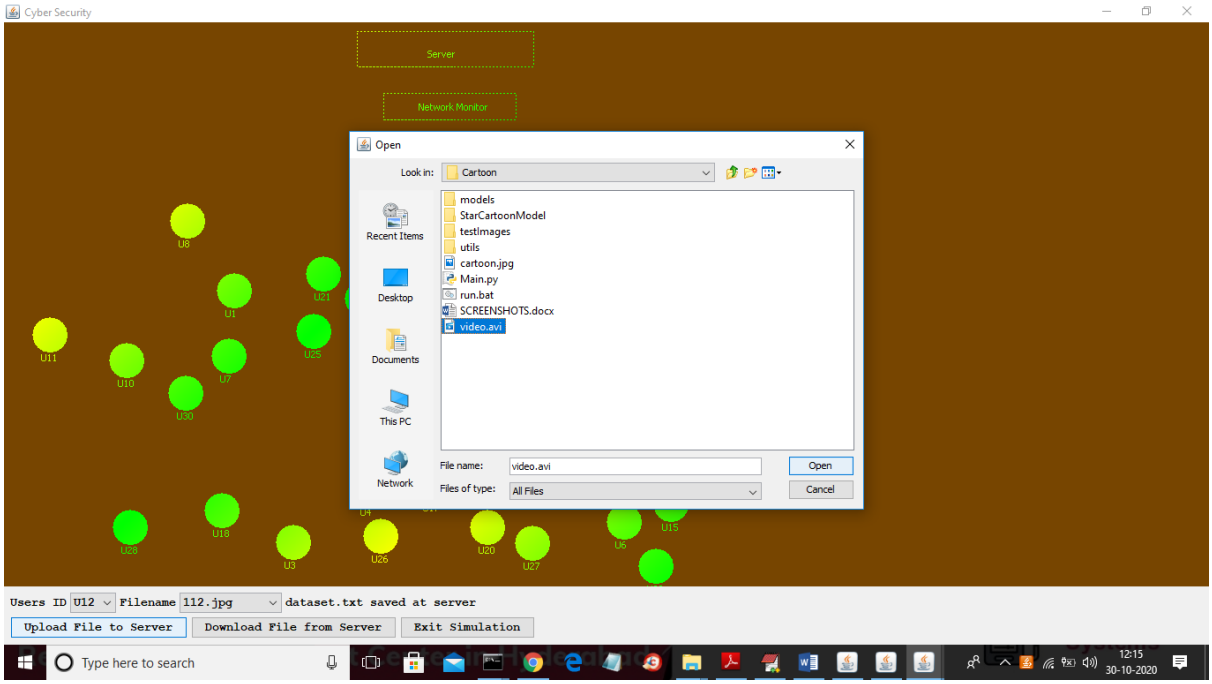


Fig:9.11 file upload

In above screen I am uploading video file and below is the result from network monitor

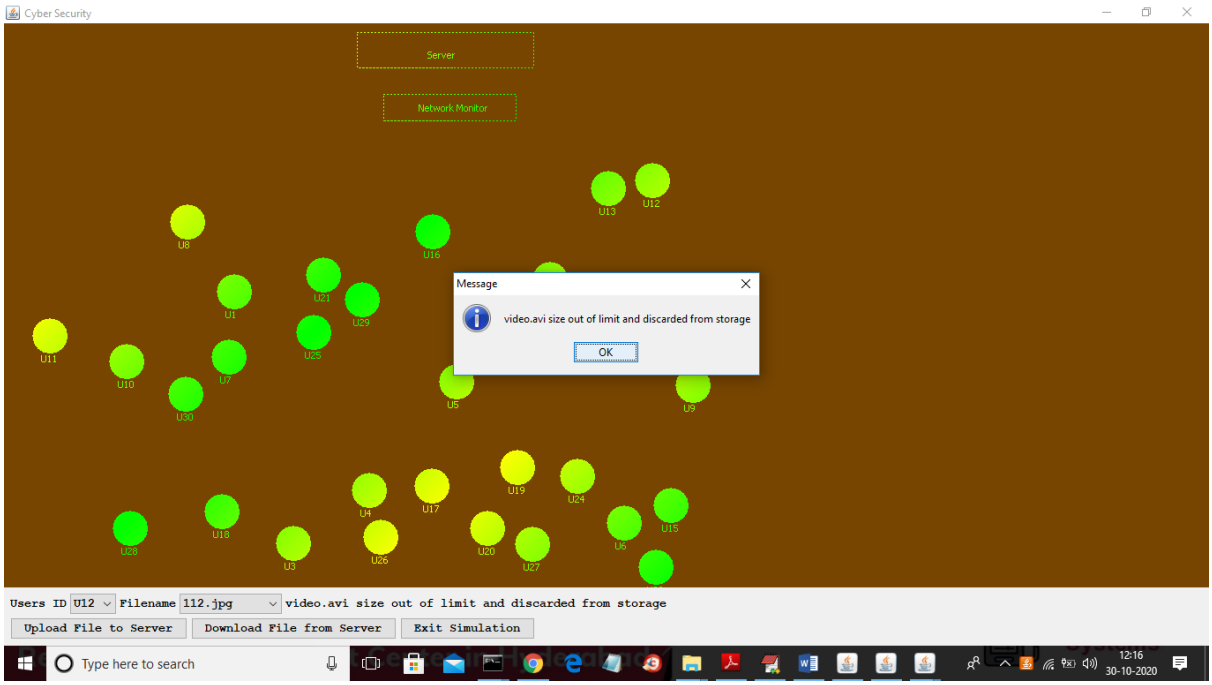


Fig: 9.12 Network monitor

In above screen we can see network monitor discarding video file as its size is out of server limit and now click on 'Download File From Server' button by selecting user and file name from drop down box

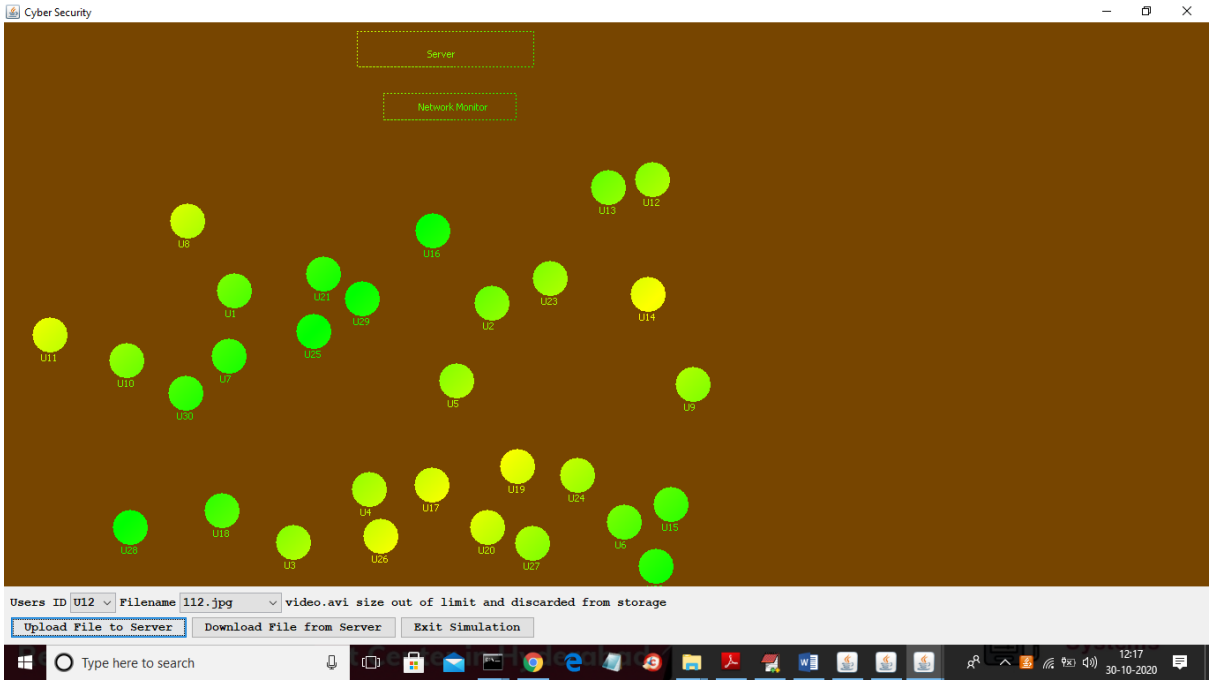


Fig:9.13 users screen

In above screen I selected U12 and 112.jpg file from drop down box and now click on 'Download File' button to get below result

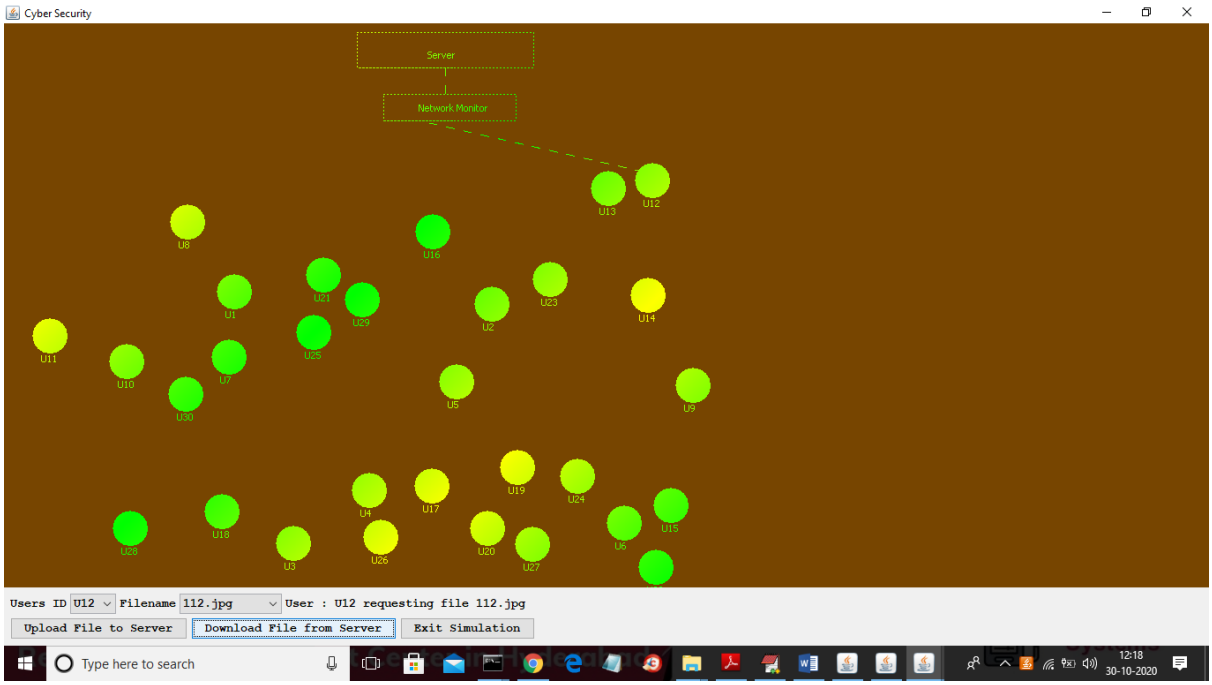


Fig: 9.14 file downloaded after file downloaded the output screen

In above screen file is downloading and after download will get below screen

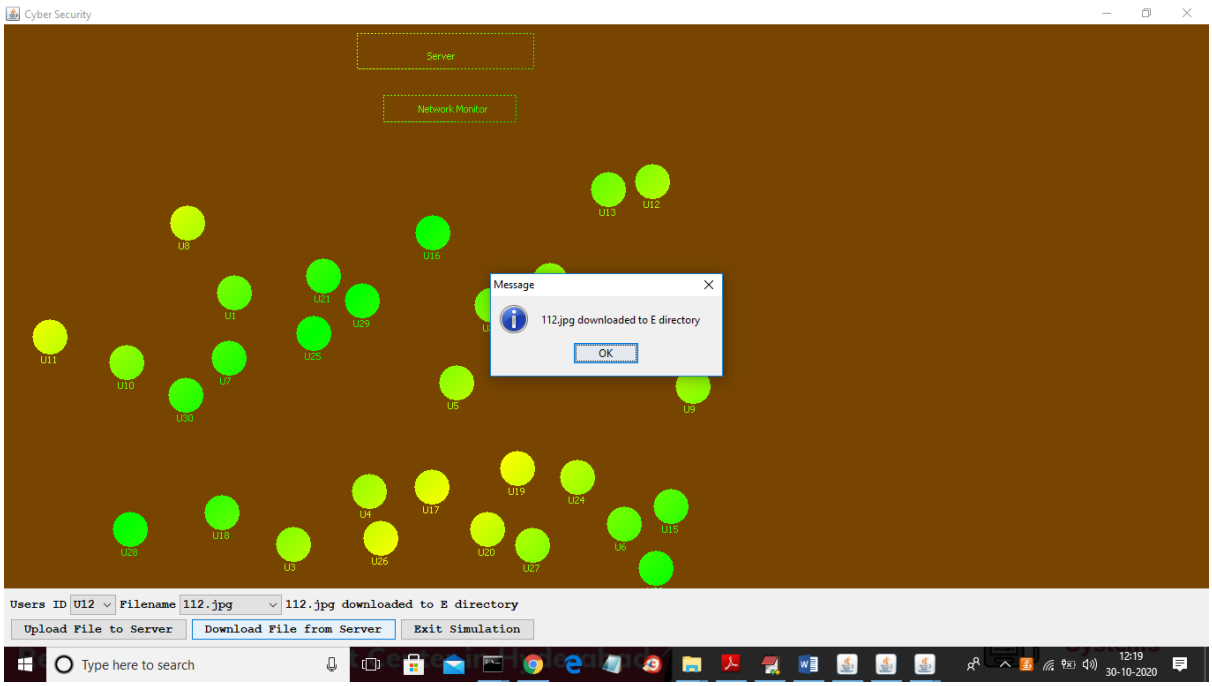


Fig: 9.15 file is downloaded at E directory

In above screen file is downloaded at E directory and now go to Network Monitor screen and click on button to get below graph

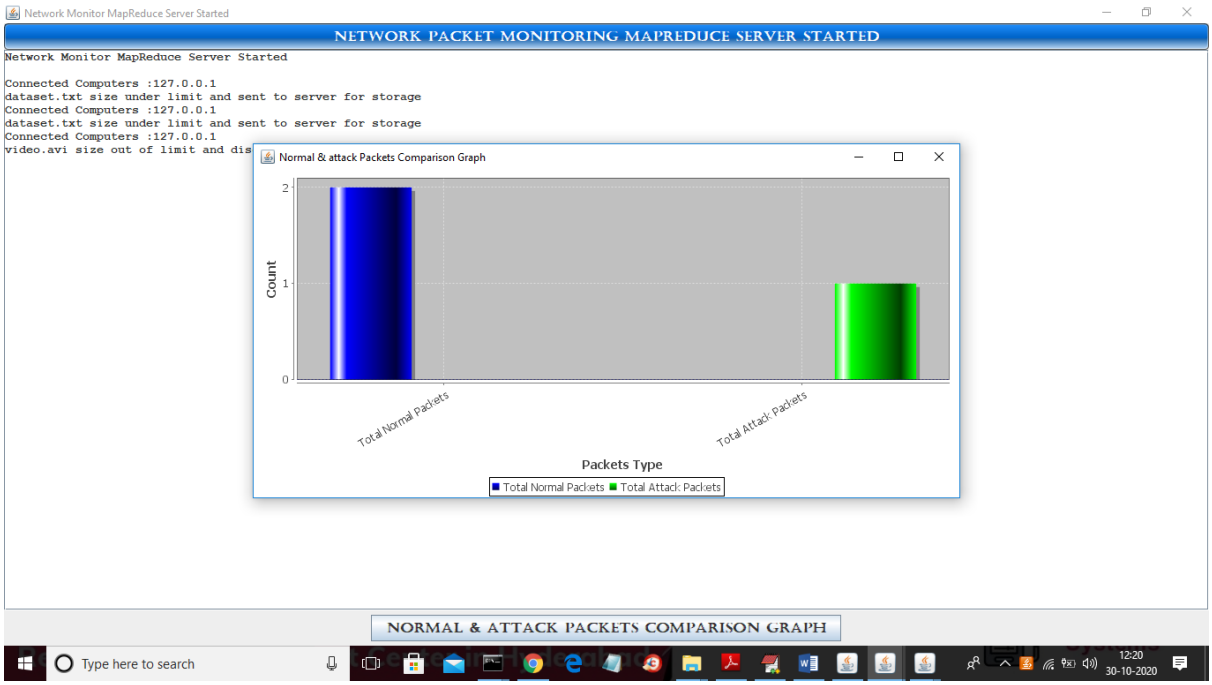


Fig: 9.16 showing request monitor has recevied

In above normal and attack graph we can see how many request monitor has received and out of that how many are normal and how many are attack request. In above graph x-axis represents packet type and y-axis represents count of those packets.

Below is the server screen of all processed requests

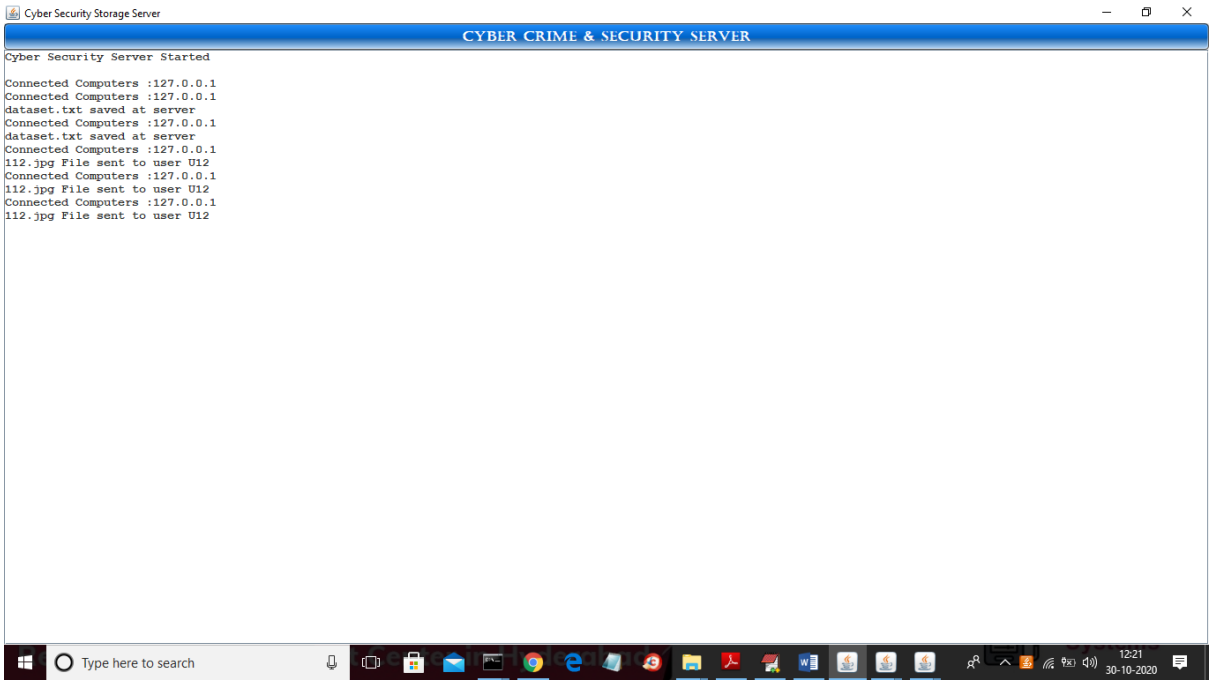


Fig: 9.17 the processed requests

10.CONCLUSIONS AND FUTURE ENHANCEMENT

In this modern era of technology, the role and usage of internet is increasing worldwide rapidly, therefore it becomes easy for cyber criminals to access any data and information with the help of their knowledge and their expertise. Cyber crime is an unlawful act or a menace that needs to be tackled firmly and effectively. There is a need to create more awareness among the people and basically users of internet about cyber space, diverse forms of cyber crime and some preventive measures as “Prevention is always better than cure”, so it is seriously advised to take some previous precautions while operating the internet. Security nowadays is becoming a prominent and major concern. In the following paper, some security issues are introduced, threats, Trojans, and attacks over internet. Computer security becomes critical in many of the technology-driven industries which operate on the computer systems. Computer security is nothing more than computer safety. Countless vulnerabilities and computer or network based issues are acts as an integral part of maintaining an operational industry.

11. REFERENCES

[1] Pooja Aggarwal , Neha, Piyush Arora , Poonam , "REVIEW ON CYBER CRIME AND SECURITY", IJREAS, Vol. 02, Issue 01, Jan 2014.

[2] Ammar Yassir and Smitha Nayak, "Cybercrime: A threat to Network Security", IJCSNS International Journal of Computer Science and Network Security, VOL.12 No.2, February 2012.

[3] Atul M. Tonge, Suraj S. Kasture, Surbhi R. Chaudhari, "Cyber security: challenges for society- literature review", IOSR Journal of Computer Engineering (IOSR-JCE) , Volume 12, Issue 2 (May. - Jun. 2013), PP 67-75.

[4] C. Catlett (ed.), "A Scientific Research and Development Approach to Cyber Security", Report submitted to the U.S. Department of Energy, December 2008.

[5] Seema Vijay Rane & Pankaj Anil Choudhary, April 2012-September 2012, "Cyber Crime and Cyber Law in India", Cyber Times International Journal of Technology and Management, Vol. 5 Issue 2.

[6] Casey, E. Digital Evidence and Computer Crime: Forensic Science, Computers and the Internet. London: Academic Press, 2011: Pp. 5-19.

[7] Richards, James. Transnational Criminal Organizations, Cybercrime, and Money Laundering: A Handbook for Law Enforcement Officers, Auditors, and Financial Investigators. Boca Raton, FL: CRC Press, 1999: Pp. 21-54.

[8] Ravi Sharma, Study of Latest Emerging Trends on Cyber Security and its challenges to Society, International Journal of Scientific & Engineering Research, Volume 3, Issue 6, June-2012 1 ISSN 2229-5518 IJSER © 2012.

[9] BinaKotiyal, R H Goudar, and Senior Member, A Cyber Era Approach for Building Awareness in Cyber Security for Educational System in India PritiSaxena, IACSIT International Journal of Information and Education Technology, Vol. 2, No. 2, April 2012.

[10] B.T. Wang and H. Schulzrinne, "An IP traceback mechanism for reflective DoS attacks", Canadian Conference on Electrical and Computer Engineering, Vol. 2, 2-5 May 2004, pp. 901 – 904.

[11] Y. C. Hu, A. Perrig, and D.B. Johnson, "Packet leashes: A defense against wormhole attacks in wireless networks", in Proceedings of the 22 nd Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM "03), vol. 3, San Francisco, CA, Mar. 2003, pp. 1976-1986.

[12] Shio Kumar Singh, M P Singh, and D K Singh, “A Survey on Network Security and Attack Defense Mechanism For Wireless Sensor Networks”, International Journal of Computer Trends and Technology- May to June Issue 2011.

[13] M. Cagalj, S. Capkun, and J.P. Hubaux, “Wormhole-based Anti-Jamming Techniques in Sensor Networks” from http://lcawww.epfl.ch/Publications/Cagalj/CagaljCH05-worm.pdf.

[14] A. T. Zia, “A Security Framework for Wireless Sensor Networks”. 2008, http://ses.library.usyd.edu.au/bitstream/2123/ 2258/4/02whole.pdf.

[15] en.wikipedia.org/wiki/Cyber_security_standards.

[16] en.wikipedia.org/wiki/Cyber_crime.

12. PUBLICATIONS

CONFERENCE:

International conference on “Innovations in computers networks ,computational intelligence

And iot(ICICCI - 21)

Paper ID: ICICCI – 21 – 0156

13.STUDENT PROFILE

Paripally Shilpa is currently pursuing her Bachelors of Technology in the stream of Computer Science and Engineering at St.Martin's Engineering College. She Completed her 12th standard from Sri Chaitanya Junior College and 10th from Viveakanda vidhyanikethan High School. She is well trained in C, Java and have basic knowledge on Python. She was the part of Employability skill development program conducted by Zensor. She is also a student in the Smart Interviews. Her participations include: Women Online Five Days Workshop on "Women in Cyber Security and Privacy in 2020" which was conducted on 6th to 10th July, a webinar on "Digital Transformation in Education Sector Post-Covid era" conducted on 11th June 2021 conducted collegedunia Two days National Level Seminar on "Recent Trends in Cloud Computing and Fog, Edge Computing" in online mode during the period from 18th June to 19 June, 2021conducted by St.Martin's Engineering College. She completed a summer internship program conducted by Goal Street On Machine Learning using python from June 2020 to August 2020 and also done a capstone project-"To preict whether the cance is Benign or Malignant".She completed few certification courses from various platforms such as Coursera, Udemy, CursaApp and Solo Learn.

Kanki Subhashini is currently pursuing her Bachelors of Technology in the stream of Computer Science and Engineering at St.Martin's Engineering College. She Completed her 12th standard from Narayana Junior College and 10th from MNR Scottsdale group of High School. She is well trained in C, Java and have basic knowledge on Python. She was the part of Employability skill development program conducted by Zensor. She is also a student in the Smart Interviews. Her participations include: Women Online Five Days Workshop on "Women in Cyber Security and Privacy in 2020" which was conducted on 6th to 10th July, a webinar on "Digital Transformation in Education Sector Post-Covid era" conducted on 11th June 2021 conducted collegedunia Two days National Level Seminar on "Recent Trends in Cloud Computing and Fog, Edge Computing" in online mode during the period from 18th June to 19 June, 2021conducted by St.Martin's Engineering College. She completed a summer internship program conducted by Goal Street On Machine Learning using python from June 2020 to August 2020 and also done a capstone project-"To preict whether the cance is Benign or Malignant".She completed few certification courses from various platforms such as Coursera, Udemy, CursaApp and Solo Learn.

Kokkonda Sumanth is currently pursuing her Bachelors of Technology in the stream of Computer Science and Engineering at St.Martin's Engineering College. He Completed her 12th standard from Sri chaitanya Junior College and 10th from Sri Chaitanya techno High School. He is well trained in C, Java and have basic knowledge on Python. He was the part of Employability skill development program conducted by Zensor. He is also a student in the Smart Interviews. His participations include: a webinar on "Digital Transformation in Education Sector Post-Covid era" conducted on 11th June 2021 conducted collegedunia Two days National Level Seminar on "Recent Trends in Cloud Computing and Fog, Edge Computing" in online mode during the period from 18th June to 19 June, 2021conducted by St.Martin's Engineering College. He completed a summer internship program conducted by Goal Street On Machine Learning using python from June 2020 to August 2020 and also done a capstone project-"To preict whether the cance is Benign or Malignant".She completed few certification courses from various platforms such as Coursera, Udemy, CursaApp and Solo Learn.

Rangoli Jagan mohan goud is currently pursuing her Bachelors of Technology in the stream of Computer Science and Engineering at St.Martin's Engineering College. He Completed her 12th standard from S R junior College and 10th from English Union School. He is well trained in C, Java and have basic knowledge on Python. He was the part of Employability skill development program conducted by Zensor. He is also a student in the Smart Interviews. His participations include: a webinar on "Digital Transformation in Education Sector Post-Covid era" conducted on 11th June 2021 conducted collegedunia Two days National Level Seminar on "Recent Trends in Cloud Computing and Fog, Edge Computing" in online mode during the period from 18th June to 19 June, 2021conducted by St.Martin's Engineering College. He completed a summer internship program conducted by Goal Street On Machine Learning using python from June 2020 to August 2020 and also done a capstone project-"To preict whether the cance is Benign or Malignant".She completed few certification courses from various platforms such as Coursera, Udemy, CursaApp and Solo Learn.

14.APPENDICES

Since September 11, 2001, many cybersecurity activities have been undertaken by the federal government,[1] the research community, and private industry. This appendix reviews these activities, providing a snapshot of the efforts undertaken to address cybersecurity concerns over the past several years. Specifically, federal cybersecurity policy activity since 2001 is reviewed. A number of federal government reports that detail cybersecurity risks and challenges that need to be overcome are summarized. Also summarized are best practices and procedures, as well as options for making progress, as identified in these reports. Efforts for improving public-private collaboration and coordination are identified. Reports aimed at elaborating the necessary elements of a research agenda are also reviewed. The final section reviews the current federal research and development (R&D) landscape and describes the particular focus and the types of support being provided at various federal agencies with cybersecurity responsibilities.

Several general impressions about the state of cybersecurity and some common themes about the type of actions required to improve it can be drawn from the various activities summarized here. First, there are

[1] The Congressional Research Service issued the report *Computer Security: A Summary of Selected Federal Laws, Executive Orders, and Presidential Directives* on April 16, 2004; the report outlines the major roles and responsibilities assigned various federal agencies in the area of computer security. See http://www.fas.org/irp/crs/RL32357.pdf.

A

PROJECTREPORT

On

VEHICLE DETECTION AND SPEED DECTETION

Submittedby

A.SAICHARAN REDDY	17601A0502
B.RAJITH	17K81A05K1
G.RETHIK REDDY	17K81A05K9
MANOOKANTH RATHI	17K81A05P8

inpartialfulfillmentfortheawardofthedegree of

BACHELOR OF TECHNOLOGY

IN

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Under the Guidance of

MR.J.SUDHAKAR

ASSOCIATE PROFESSOR

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ST.MARTIN'S ENGINEERING COLLEGE

An Autonomous Institute

Dhulapally, Secunderabad – 500 100

JUNE 2021

BONAFIDE CERTIFICATE

This is to certify that the project entitled **VEHICLE DETECTION AND SPEED DECTETION**, is being submitted by **A.SAICHARAN REDDY 17601A0502, B.RAJITH 17K81A05K1, G.RETHIK REDDY 17K81A05K9, MANOOKANTH RATHI 17K81A05P8** in partial fulfillment of the requirement for the award of the degree of **BACHELOR OFTECHNOLOGY INCOMPUTER SCIENCE** is recorded of bonafide work carried out by them. The result embodied in this report have been verified and found satisfactory.

J SUDHAKAR
Associate professor
Department of CSE

Head of the Department
Dr.M.NARAYANAN
Department of CSE

Internal Examiner External Examiner

Place:

Date:

DECLARATION

We, the student of **Bachelor of Technology** in Department of Computer Science and Engineering', session: 2017 – 2021, St. Martin's Engineering College, Dhulapally, Kompally, Secunderabad, hereby declare that work presented in this Project Work entitled **VEHICLE DETECTION AND SPEED DETECTION** is the outcome of our own bonafide work and is correct to the best of our knowledge and this work has been undertaken taking care of Engineering Ethics. This result embodied in this project report has not been submitted in any university for award of any degree.

A.SAICHARAN REDDY 17601A0502

B.RAJITH 17K81A05K1

G.RETHIK REDDY 17K81A05K9

MANOOKANTH RATHI 17K81A05P8

ABSTRACT

In recent times, there has been a drastic change in people's lifestyles and with an increase in incomes and lower cost of automobiles there is a huge increment in the number of cars on the roads which has led to traffic and commotion. The manual efforts to keep people from breaking traffic rules such as the speed limit are not enough. There is not enough police and man force available to track the traffic and vehicles on roads and check them for speed control. Hence, we require technologically advanced speed calculators installed that effectively detect cars on the road and calculate their speeds.

To implement the above idea two basic requirements, need to be met which are the effective detection of the cars on roads and their velocity measurement. For this purpose, we can use OpenCV software which uses the Haar cascade to train our machine to detect the object, in this case the car.

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible and whose encouragements and guidance have crowded effects with success.

We extended our deep sense of gratitude to Principal**, Dr. P. SANTOSH KUMARPATRA**, St. Martin's Engineering College, Dhulapally, for permitting us to undertake this project.

We are also thankful to **Dr. M.NARAYANAN**, Head of the Department, Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

We would like to thank **Dr. T. POONGOTHAI,** Department of Computer Science and Engineering (AI & ML) for her encouragement and insightful comments and as well as our project coordinators **Dr. B.RAJALINGAM**, Associate Professor and **Mr. J.SUDHAKAR**, Associate Professor, in Department of Computer Science and Engineering for their valuable support.

We would like to express our sincere gratitude and indebtedness to our project supervisor **Mr.J.SUDHAKAR**Associate professor Computer Science and Engineering, St. Martin's Engineering College, Dhulapally, for his support and guidance throughout our project.

Finally, we express thanks to all those who have helped us successfully to completing this project. Furthermore, we would like to thank our family and friends for their moral support and encouragement.

We express thanks to all those who have helped us in successfully completing the project.

A.SAICHARAN REDDY 17601A0502

B.RAJITH 17K81A05K1

G.RETHIK REDDY 17K81A05K9

MANOOKANTH RATHI 17K81A05P8

TABLEOFCONTENTS

LISTOFTABLES

LIST OF FIGURES

LISTOFACRONYMS

<AVI>	AudioVideoInterlace
<BMP>	Bitmap
<CPU>	CentralProcessingUnit
<GB>	GigaBytes
<GUI>	GraphicalUserInterface

INTRODUCTION

We have developed a Haar cascade to detect cars on the roads, whose velocities are then measured using a python script. The real-time application of this project proves to be much useful as it is easy to implement, fast to process and efficient with low cost development. Also, the tool might be useful to apply in simulation tools to measure velocities of cars. This can be further developed to identify all kinds of vehicles as well as to check anyone who breaks a traffic light.
The improvements in the project can be done by creating a bigger haar cascade since bigger the haar cascade developed, more the number of vehicles that can be detected on the roads. Better search algorithms can allow a faster search and better detection of these vehicles for better efficiency.

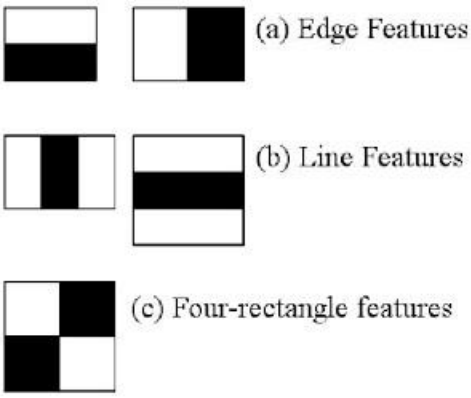
This paper is to develop an algorithm to calculate the speed of the object(vehicle) detected. We have implemented the algorithm using
Python Script.

The complete implementation uses two basic processes: -
1. Car detection using Haar cascades in OpenCV
2. Measurement of velocity of detected cars using python script.

Car Detection:

Object Location utilizing Haar highlight based course classifiers is a compelling item discovery strategy that uses a machine learning based approach where a course capacity is prepared from a considerable measure of positive and negative pictures. It is then used to recognize protests in different pictures.

• Initially, the calculation needs a considerable measure of positive (pictures of autos) and negative (pictures without autos) to prepare the classifier. At that point, we have to concentrate highlights from it. For this, haar highlights appeared in beneath picture are utilized. They are much the same as our convolutional part. Each component is a solitary esteem acquired by subtracting total of pixels under white rectangle from aggregate of pixels under dark rectangle.



Now every single conceivable size and areas of every part is utilized to ascertain a lot of components. (Simply envision what amount of calculation it needs? Indeed, even a 24x24 window comes about more than 160000 components). For each component computation, we have to discover whole of pixels under white and dark rectangles. To tackle this, they presented the necessary pictures.

• Now, we apply each component on all the preparation pictures. For each component, it finds the best limit which will characterize the countenances to positive and negative. Be that as it may, clearly, there will be blunders or misclassifications. We select the elements with least mistake rate, which implies they are the elements that best orders the auto and non-auto pictures.

• So now you take a picture. Take each 24x24 window. Apply 6000 elements to it. Check on the off chance that it is auto or not.

Speed Calculation

• Once a car is detected, using the cascadeClassifier() function on the haar cascade developed.

• Now the time is started which was initialized to 0.

• Using the ratio in the image for each cm travelled by the detected image and real-time distance in meters, the actual distance covered by the car is calculated.

• As soon as the car reaches the center of the detection window whose distance is already known to us the time is stopped.

• Now the actual distance calculated is divided by the time calculated and velocity is obtained.

• This velocity and the distance of the camera in feet from the car (i.e. the height of camera above the car) is printed on the output screen.

For this use multiple object detection algorithms could have been used but the algorithm of developing the Haar cascade and its implementation proves to be the best since it is the least time consuming, most efficient and highly reliable.

CHAPTER -2

LITERATURE SURVEY

2.1 SURVEY ON BACKGROUND:

A number of techniques are available for vehicle detection from a captured video. Detection of moving vehicle accurately from the video is a challenging task. The moving vehicle detection provides a classification of the pixels in the video frames either for foreground or background data. The approaches for vehicle detection are Frame differencing, background subtraction algorithm [1]. Frame differencing method can be used to eliminate the still objects i.e. which are not moving, and will consider the objects which are under motion. The main drawback of this method is slow moving objects are not detected, so improvisation is needed for frame differencing technique to detect slow motion objects too.

Background subtraction means that the background or the static scene is extracted from the video image. Because the camera is fixed, each pixel in the image has a corresponding value which is basically fixed over a period of time. The purpose of background subtraction is to find the background value of each point of the image. We have used Mixture of Gaussian technique to detect the moving vehicle since it overcomes the limitations of the other methods. The best approach for removing the background and considering the vehicle from the video is through background subtraction were each frames of the video is compared with reference or a background model, and the pixels were deviation is observed from the background are considered to be the vehicles.

2.2 CONCLUSION ON SURVEY:

Once the moving vehicle is detected from the image frames, then the next step is to track the vehicle from thevideo sequence, by tracking the vehicle we will be able to determine the velocity, acceleration and position of vehicle at different appetence pattern and occlusions between the object-to-scene. In general, tracking can be defined as the problem of determining the trajectory motion of an object as it moves around in a scene. Here the object tracker assigns a label to all the tracked vehicles In other words; a tracker can also provide centric-information, orientation and shape of a vehicle.

TITILE:A vision-based vehicle identification system," in Pattern Recognition.

.

AUTHOR: H. Chung-Lin and L. Wen-Chieh.

This work presents a vision-based vehicle identification system which consists of object extraction, object tracking, occlusion detection and segmentation, and vehicle classification. Since the vehicles on the freeway may occlude each other, their trajectories may merge or split. To separate the occluded objects, we develop three processed: occlusion detection, motion vector calibration, and motion field clustering. Finally, the segmented objects are classified into seven different categorized vehicles.

TITLE: Real-Time Incremental Segmentation and Tracking of Vehicles at Low Camera Angles Using Stable Features.

AUTHOR:N. K. Kanhere and S. T. Birchfield.

We present a method for segmenting and tracking vehicles on highways using a camera that is relatively low to the ground. At such low angles, 3-D perspective effects cause significant changes in appearance over time, as well as severe occlusions by vehicles in neighboring lanes. Traditional approaches to occlusion reasoning assume that the vehicles initially appear well separated in the image; however, in our sequences, it is not uncommon for vehicles to enter the scene partially occluded and remain so throughout. By utilizing a 3-D perspective mapping from the scene to the image, along with a plumb line projection, we are able to distinguish a subset of features whose 3-D coordinates can be accurately estimated.

TITLE: Vehicle Tracking using Computer Vision Technique for Car-Following Model.

AUTHOR: Elysia ,Gunawan FE.

Detect, classify and keep track, in real-time, on different kinds of objects or vehicles that are moving on a road is crucial for traffic managements systems, among other research areas. In this paper, a vision based system to detect, track, count and classify moving vehicles, on any kind of road, is shown. The data acquisition system consists of a HD-RGB camera placed on the road, while the information processing is performed by clustering and classification algorithms. The system obtained an efficiency score over the 95 percent in test cases, as well, the correct classification of 85 percent of the test objects. Also, the system achieves 30 fps in image processing with a resolution of 1280×720.

TITLE: Speed Detection Camera System using Image Processing. International Journal of Computer and Electrical Engineering.

AUTHOR: Ibrahim O, ElGendy H, Ahmed M. ElShafee.

This paper, presents a new Speed Detection Camera System (SDCS) that is applicable as a radar alternative. SDCS uses several image processing techniques on video stream in online -captured from single camera- or offline mode, which makes SDCS capable of calculating the speed of moving objects avoiding the traditional radars' problems. SDCS offers an en-expensive alternative to traditional radars with the same accuracy or even better.

CHAPTER-3

SOFTWARE AND HARDWARE REQUIRNMENTS

3.1SOFTWARE REQUIREMENTS:

- Operating System : Windows 7
- Technology : Python and Django
- Python Version : Python 3.9.0

3.2 HARDWARE REQUIREMENTS:

- Hardware : Pentium IV
- RAM : 256MB
- Hard Disk : 512mb

CHAPTER- 4

SOFTWARE DEVELOPMENT ANALYSIS

4.1 OVERVIEW OF PROBLEM:

Now-a-days traffic in two-tier cities has increased lot, so surveillance system is needed to monitor the traffic and avoid unwanted delays and accidents. Vehicles speed estimation can be done by analyzing the video captured. The conventional or the vehicles is through using RADAR device and other models like constant g-factor ,sequence method and motion median. The limitations of this device or models are accuracy is bit less and costly and does not capture long distance images. Different techniques have been used for speed estimation; here in this paper we propose a mixture of Gaussian technique to estimate the speed.

4.2 DEFINE PROBLEM:

Vehicle tracking is the process of locating a moving vehicle using a camera. Capture vehicle in video sequence from surveillance camera is demanding application to improve tracking performance. This technology is increasing the number of applications such as traffic control, traffic monitoring, traffic flow, security etc. The estimated cost using this technology will be very less. Video and image processing has been used for traffic surveillance, analysis and monitoring of traffic conditions in many cities and urban areas. Various methods for speed estimation are proposed in recent years.

4.3 MODULES:

Moving Vehicle Detection:

It also known as Foreground Detection, it is a technique in the field of image processing where the foreground of image is extracted. A general rule determines which intensities are of the background and that of foreground. The pixels that do not match to these are known as foreground pixels. Foreground pixels are grouped using 8 connectivity connected component analysis.

Feature Extraction

Corner detection is a widely used approach to extract certain kinds of features. It is mainly used to deduce the contents of an image. The Harris corner detector has some corner selection criteria. A score is calculated for each pixel in the image, and if the score is above a threshold value, then the pixel is marked as a corner otherwise it is not.

Speed Estimation

Distance is calculated using the centroids of previous and next frames. For distance calculation, Euclidean distance is used. Time is calculated as vehicle enters the ROI. Initially the speed is calculated using the formula If (x1, y1) is centroid of vehicle in first frame and (x2, y2) is centroid of second frame then Distance travelled= $(x2-x1)^2+ (y2-y1)^2$ Speed=Distance travelled/Time taken.

Vehicle tracking

Tracking of vehicles from a video where vehicles are in motion is a challenging task. Many difficulties may arise in vehicle tracking such as no rigid vehicle structure, changes in background subtraction. Different techniques are available but the most efficient one is Gaussian mixture Model (GMM).

4.4SYSTEM STUDY

FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are

- ECONOMICAL FEASIBILITY
- TECHNICAL FEASIBILITY
- SOCIAL FEASIBILITY

ECONOMICAL FEASIBILITY

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

TECHNICAL FEASIBILITY

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

SOCIAL FEASIBILITY

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

IMPLIMENTATION:

1.Upload Image:

We apply each component on all the preparation pictures. For each component, it finds the best limit which will characterize the countenances to positive and negative. Be that as it may, clearly, there will be blunders or misclassifications. We select the elements with least mistake rate, which implies they are the elements that best orders the auto and non-auto pictures.

• So now you take a picture. Take each 24x24 window. Apply 6000 elements to it. Check on the off chance that it is auto or not.

2.Train Dataset:

Now every single conceivable size and areas of every part is utilized to ascertain a lot of components. (Simply envision what amount of calculation it needs? Indeed, even a 24x24 window comes about more than 160000

3.Upload Test & Classify:

This velocity and the distance of the camera in feet from the car (i.e. the height of camera above the car) is printed on the output screen.

For this use multiple object detection algorithms could have been used but the algorithm of developing the Haar cascade and its implementation proves to be the best since it is the least time consuming, most efficient and highly reliable.

The complete implementation uses two basic processes: -

1. Car detection using Haar cascades in OpenCV

2. Measurement of velocity of detected cars using python script. components). For each component computation, we have to discover whole of pixels under white and dark rectangles. To tackle this, they presented the necessary pictures.

• Now, we apply each component on all the preparation pictures. For each component, it finds the best limit which will characterize the countenances to positive and negative. Be that as it may, clearly, there will be blunders or misclassifications. We select the elements with least mistake rate, which implies they are the elements that best orders the auto and non-auto pictures.

• So now you take a picture. Take each 24x24 window. Apply 6000 elements to it. Check on the off chance that it is auto or not.

CHAPTER-5

PROJECT SYSTEM DESIGN

5.1 SYSTEM DESIGN

This project is to develop an algorithm to calculate the speed of the object(vehicle) detected. We have implemented the algorithm using Python Script. In urban areas the traffic monitoring cameras are stationary. They will be located above the ground level to get a clear view of vehicles moving on road. The vehicle detection from stationary camera will be much easier as compared to cameras which capture the dynamic change of the vehicles and its surrounding environment. In this paper we are analyzing the videos captured by stationary cameras, and to detect multiple vehicles optical-flow based technique is used. Using this technique multiple images are recognized at different times, optical-flow based technique indirectly detects obstacles by analyzing the velocity field. Object detection techniques can be classified depending on the background subtraction and frame differencing.

5.2 UML DIAGRAMS

Any complex system is best understood by making some kind of diagrams or pictures. These diagrams have a better impact on our understanding. If we look around, we will realize that the diagrams are not a new concept but it is used widely in different forms in different industries.

We prepare UML diagrams to understand the system in a better and simple way. A single diagram is not enough to cover all the aspects of the system. UML defines various kinds of diagrams to cover most of the aspects of a system. You can also create your own set of diagrams to meet your requirements. Diagrams are generally made in an incremental and iterative way. There are two broad categories of diagrams and they are again divided into

subcategories –

- Structural Diagrams
- Behavioral Diagrams

Structural Diagrams:

The structural diagrams represent the static aspect of the system. These static aspects represent those parts of a diagram, which forms the main structure and are therefore stable.

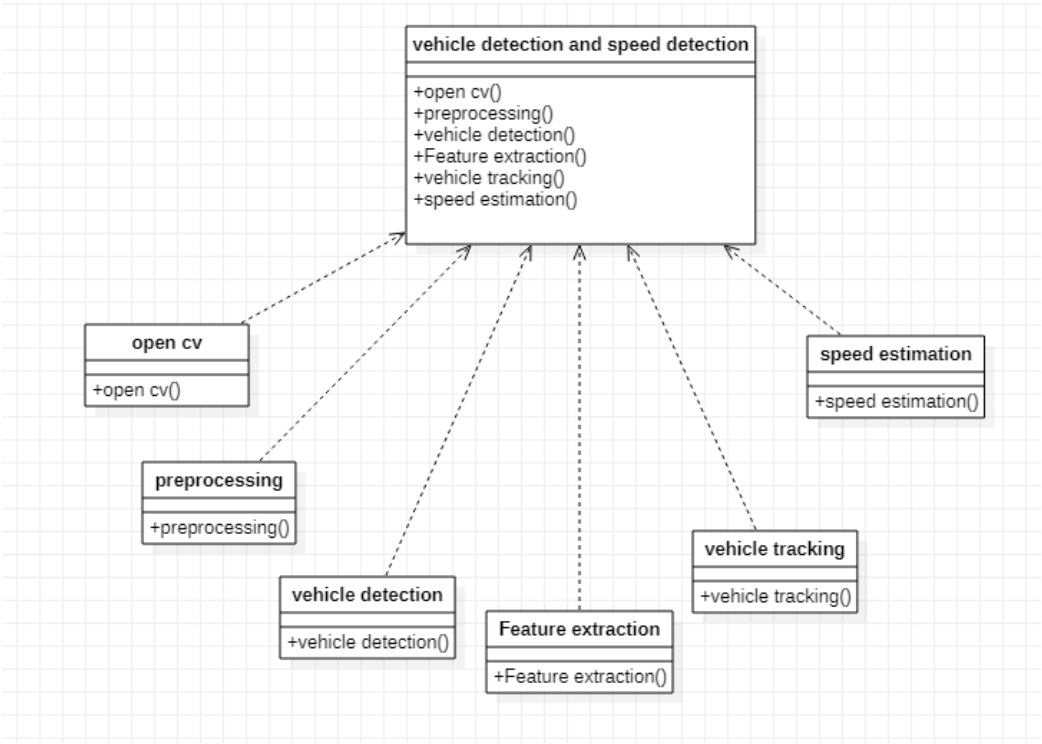
These static parts are represented by classes, interfaces, objects, components, and nodes. The four structural diagrams are

- Class diagram
- Object diagram
- Component diagram
- Deployment diagram

Class Diagram

Class diagrams are the most common diagrams used in UML. Class diagram consists of classes, interfaces, associations, and collaboration. Class diagrams basically represent the object-oriented view of a system, which is static in nature. Active class is used in a class diagram to represent the concurrency of the system.

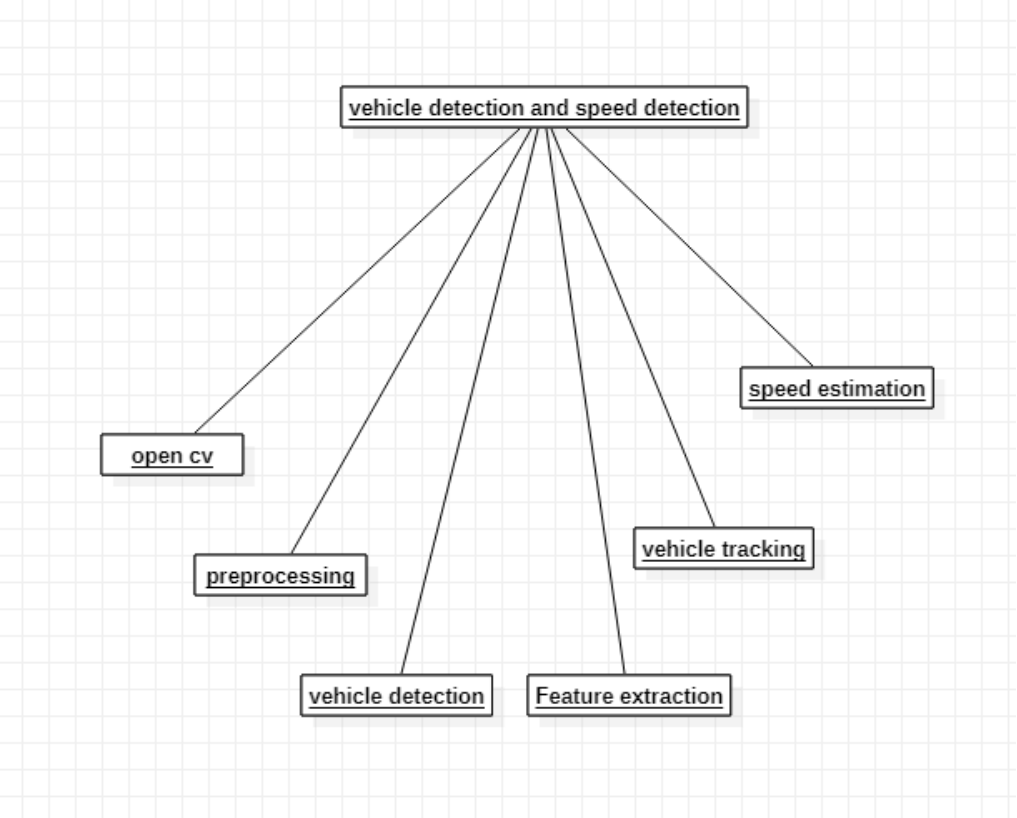
Class diagram represents the object orientation of a system. Hence, it is generally used for development purpose. it is the most widely used diagram at the time of system construction.



Object Diagram:

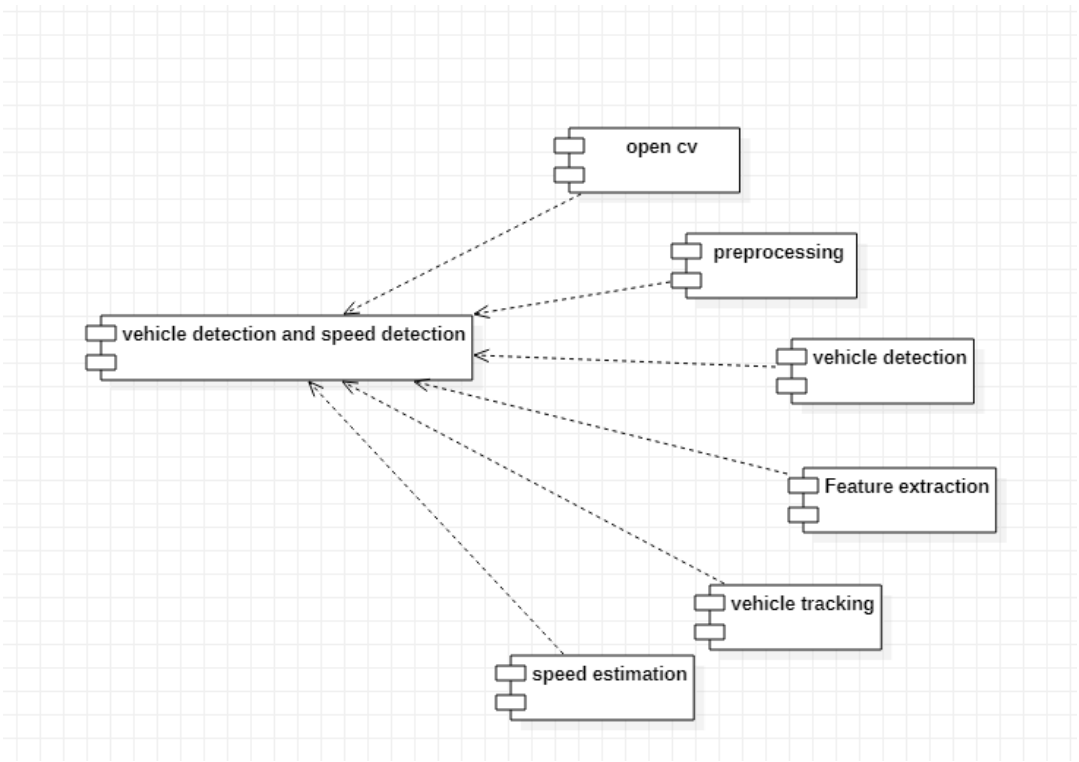
Object diagrams can be described as an instance of class diagram. Thus, these diagrams are more close to real-life scenarios where we implement a system. Object diagrams are a set of objects and their relationship is just like class diagrams. They also represent the static view of the system.

The usage of object diagrams is similar to class diagrams but they are used to build prototype of a system from a practical perspective.



Component Diagram:

Component diagrams represent a set of components and their relationships. These components consist of classes, interfaces, or collaborations. Component diagrams represent the implementation view of a system. During the design phase, software artifacts (classes, interfaces, etc.) of a system are arranged in different groups depending upon their relationship. Now, these groups are known as components. Finally, it can be said component diagrams are used to visualize the implementation.



Deployment Diagram:

Deployment diagrams are a set of nodes and their relationships. These nodes are physical entities where the components are deployed. Deployment diagrams are used for visualizing the deployment view of a system. This is generally used by the deployment team. Note − If the above descriptions and usages are observed carefully then it is very clear that all the diagrams have some relationship with one another. Component diagrams are dependent upon the classes, interfaces, etc. which are part of class/object diagram. Again, the deployment diagram is dependent upon the components, which are used to make component diagrams.

Behavioral Diagrams:

Any system can have two aspects, static and dynamic. So, a model is considered as complete when both the aspects are fully covered.

Behavioral diagrams basically capture the dynamic aspect of a system. Dynamic aspect can be further described as the changing/moving parts of a system.

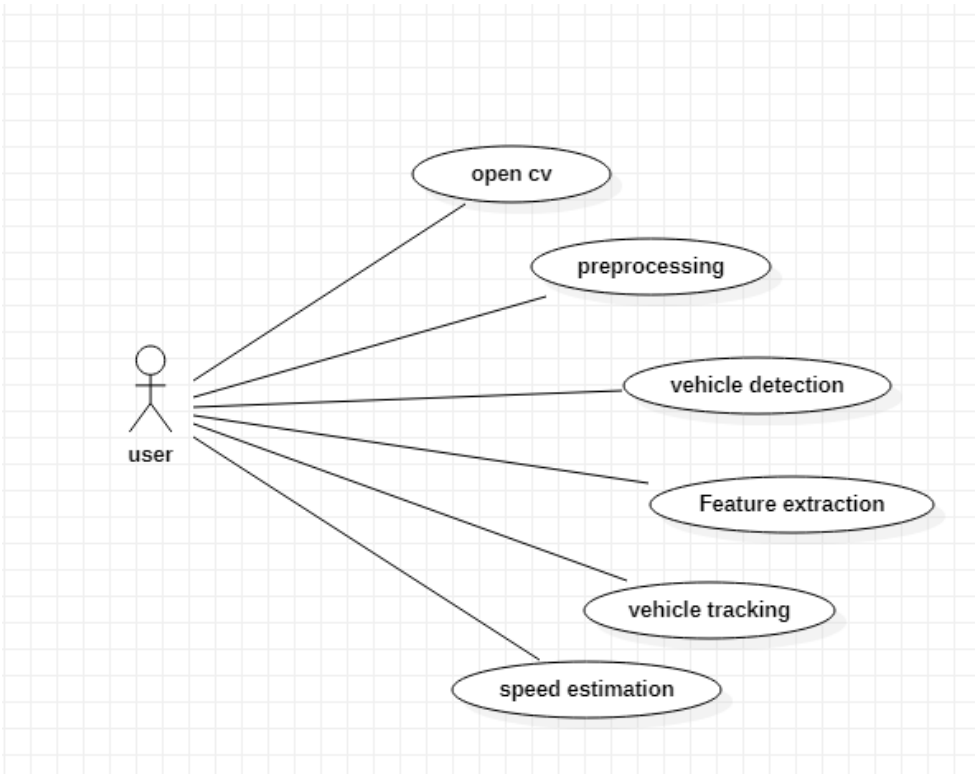
UML has the following five types of behavioral diagrams −

- Use case diagram
- Sequence diagram
- Collaboration diagram

- Statechart diagram

5.3 Use Case Diagram:

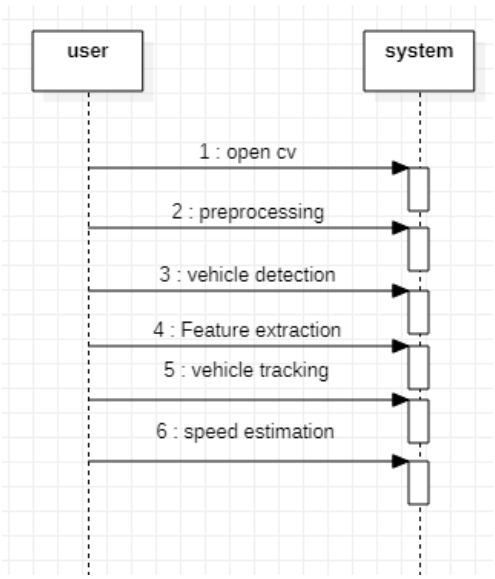
Use case diagrams are a set of use cases, actors, and their relationships. They represent the use case view of a system. A use case represents a particular functionality of a system. Hence, use case diagram is used to describe the relationships among the functionalities and their internal/external controllers. These controllers are known as actors.



Sequence Diagram:

A sequence diagram is an interaction diagram. From the name, it is clear that the diagram deals with some sequences, which are the sequence of messages flowing from one object to another.

Interaction among the components of a system is very important from implementation and execution perspective. Sequence diagram is used to visualize the sequence of calls in a system to perform a specific functionality.

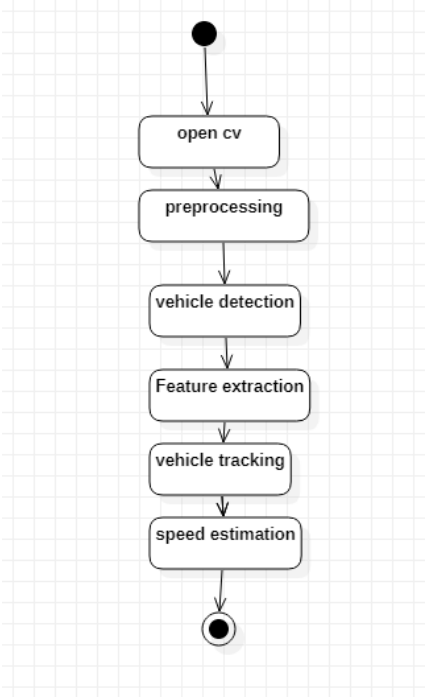


Collaboration Diagram:

Collaboration diagram is another form of interaction diagram. It represents the structural organization of a system and the messages sent/received. Structural organization consists of objects and links. The purpose of collaboration diagram is similar to sequence diagram. However, the specific purpose of collaboration diagram is to visualize the organization of objects and their interaction.

Statechart Diagram:

Any real-time system is expected to be reacted by some kind of internal/external events. These events are responsible for state change of the system. Statechart diagram is used to represent the event driven state change of a system. It basically describes the state change of a class, interface, etc. State chart diagram is used to visualize the reaction of a system by internal/external factors.



Activity Diagram:

Activity diagram describes the flow of control in a system. It consists of activities and links. The flow can be sequential, concurrent, or branched. Activities are nothing but the functions of a system. Numbers of activity diagrams are prepared to capture the entire flow in a system. Activity diagrams are used to visualize the flow of controls in a system. This is prepared to have an idea of how the system will work when executed.

Note − Dynamic nature of a system is very difficult to capture. UML has provided features to capture the dynamics of a system from different angles. Sequence diagrams and collaboration diagrams are isomorphic, hence they can be converted from one another without losing any information. This is also true for Statechart and activity diagram

CHAPTER-6

PROJECT CODING

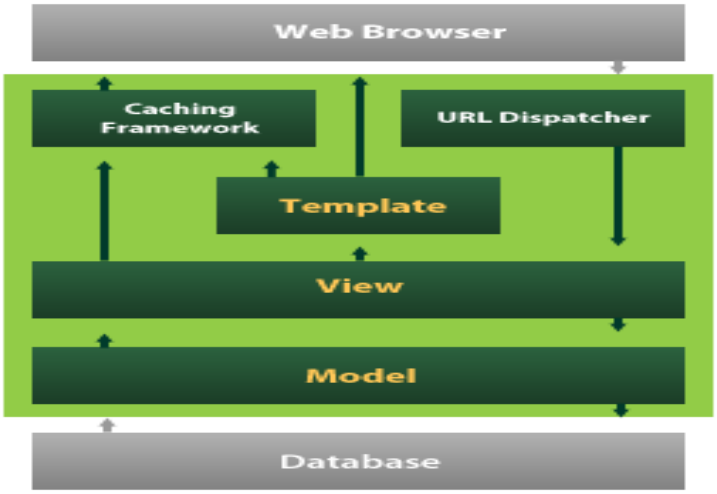
6.1 PYTHON:

Python is a general-purpose interpreted, interactive, object-oriented, and high-level programming language. An interpreted language, Python has a design philosophy that emphasizes code readability (notably using whitespace indentation to delimit code blocks rather than curly brackets or keywords), and a syntax that allows programmers to express concepts in fewer lines of code than might be used in languages such as C++or Java. It provides constructs that enable clear programming on both small and large scales. Python interpreters are available for many operating systems. CPython, the reference implementation of Python, is open source software and has a community-based development model, as do nearly all of its variant implementations. CPython is managed by the non-profit Python Software Foundation. Python features a dynamic type system and automatic memory management. It supports multiple programming paradigms, including object-oriented, imperative, functional and procedural, and has a large and comprehensive standard library

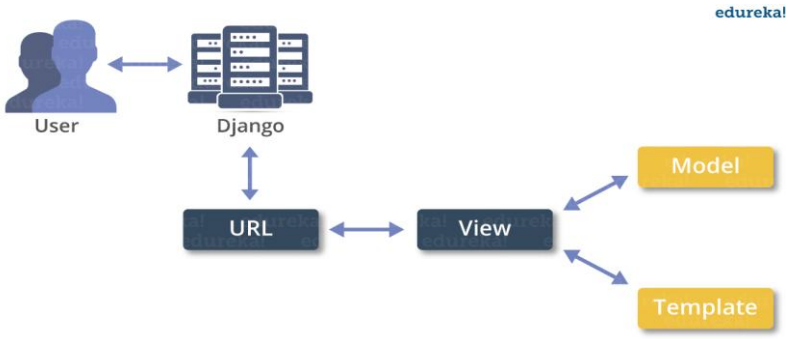
DJANGO:

Django is a high-level Python Web framework that encourages rapid development and clean, pragmatic design. Built by experienced developers, it takes care of much of the hassle of Web development, so you can focus on writing your app without needing to reinvent the wheel. It’s free and open source.

Django's primary goal is to ease the creation of complex, database-driven websites. Django emphasizes reusabilityand "pluggability" of components, rapid development, and the principle of don't repeat yourself. Python is used throughout, even for settings files and data models.



Django also provides an optional administrative create, read, update and delete interface that is generated dynamically through introspection and configured via admin models



SOFTWARE ENVIRONMENT:

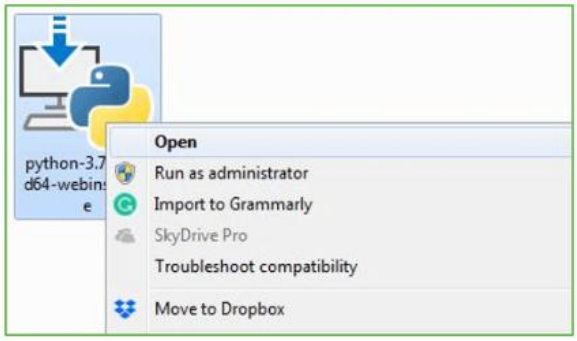
Python is currently the most widely used multi-purpose, high-level programming language.

Python language is being used by almost all tech-giant companies like – Google, Amazon, Facebook, Instagram, Dropbox, Uber… etc.

The biggest strength of Python is huge collection of standard library which can be used for the following –

- Machine Learning
- GUI Applications (like Kivy, Tkinter, PyQt etc.)
- Web frameworks like Django (used by YouTube, Instagram, Dropbox)
- Image processing (like Opencv, Pillow)
- Web scraping (like Scrapy, BeautifulSoup, Selenium)
- Test frameworks
- Multimedia

Step 1: Go to Download and Open the downloaded python version to carry out the installation process.



Step 2: Before you click on Install Now, Make sure to put a tick on Add Python 3.7 to PATH.

Step 3: Click on Install NOW After the installation is successful. Click on Close.

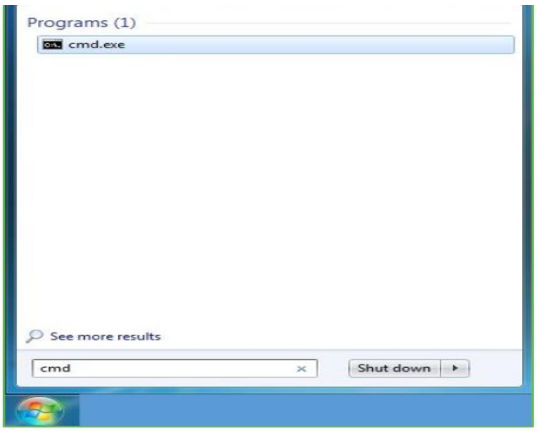
With these above three steps on python installation, you have successfully and correctly installed Python. Now is the time to verify the installation.

Note: The installation process might take a couple of minutes.

Verify the Python Installation

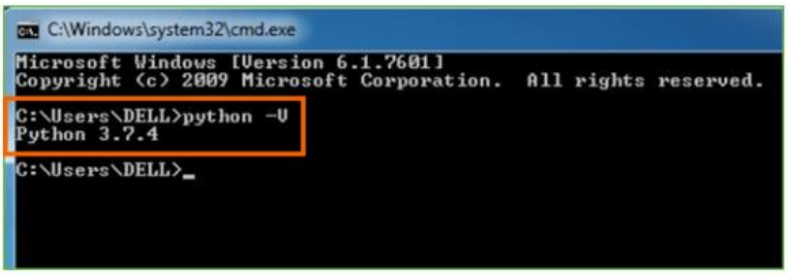
Step 1: Click on Start

Step 2: In the Windows Run Command, type “cmd”.



Step 3: Open the Command prompt option.

Step 4: Let us test whether the python is correctly installed. Type **python –V** and press Enter.



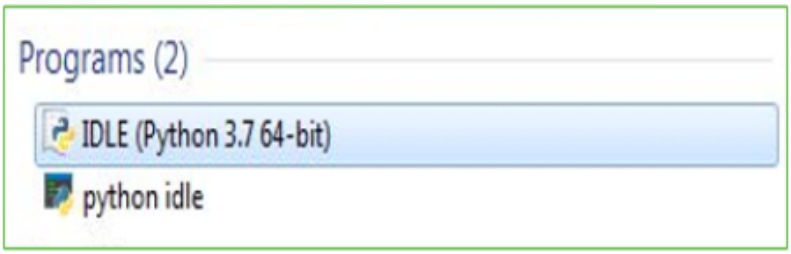
Step 5: You will get the answer as 3.7.4

Note: If you have any of the earlier versions of Python already installed. You must first uninstall the earlier version and then install the new one.

Check how the Python IDLE works

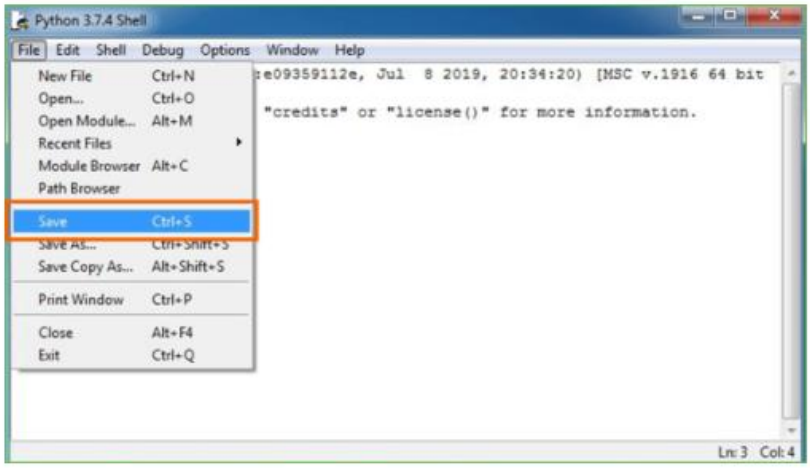
Step 1: Click on Start

Step 2: In the Windows Run command, type “python idle”.



Step 3: Click on IDLE (Python 3.7 64-bit) and launch the program

Step 4: To go ahead with working in IDLE you must first save the file. **Click on File > Click on Save**



Step 5: Name the file and save as type should be Python files. Click on SAVE. Here I have named the files as Hey World.

Step 6: Now for e.g. **enter print**

PROJECT CODE:

```
importnumpy as np
import cv2
import time
car_cascade = cv2.CascadeClassifier('hand.xml')
cap = cv2.VideoCapture('car.mp4')
wide=0.1   #depends upon size of car(~2.5)
flag=True
start=end=0
time_diff=0
while(cap.isOpened()):
ret, img = cap.read()
height,width,chan=img.shape
   #print(height,width,chan)
gray = cv2.cvtColor(img, cv2.COLOR_BGR2GRAY)
cars = car_cascade.detectMultiScale(gray, 1.3, 5)
   #crp=gray[0:480,0:int(width/2)+10]
for(x,y,w,h) in cars:
```

```
cv2.rectangle(img, (x,y), (x+w,y+h), (0,255,0),2)
center_x=(2*x+w)/2
center_y=(2*y+h)/2
      #print(center_x,center_y)
      dist1=((wide*668.748634441)/w)
      #print("Distance from car:",round(dist1,2),"m")
roi_gray = gray[y:y+h,x:x+w]
roi_color = img[y:y+h,x:x+w]
      dist0=((wide*668.748634441)/w)
actual_dist=dist0*(width/2)/668.748634441
      #print("Actual Distance:",actual_dist)
if flag is True and int(round(center_x)) in (range(0,80) or range(400,480)):
start=time.time()
flag=False
         #print("Start:",start)
      if flag is False and int(round(center_x)) in range(int(round(width/2))-
10,int(round(width/2))+10):
end=time.time()
time_diff=end-start
         #print("End:",end)
flag=True
#print("Time Difference:",time_diff)
iftime_diff>0 and s_flag==True:
velocity=actual_dist/time_diff
      #print(round(start),round(end))
vel_kmph=round(velocity*3.6,2)
print("Speed:",vel_kmph,"kmph")
print("Distance from car:",round(dist1,2),"m")
s_flag=False
   cv2.line(img,(int(width/2),0),(int(width/2),height),(255,0,0),2)
cv2.imshow('frame',img)
if cv2.waitKey(1) & 0xFF == ord('q'):
break
cap.release()
cv2.destroyAllWindows()
```

CHAPTER-7

PROJECT TESTING

7.1 VARIOUS TEST CASES:

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

TYPES OF TESTS:

Unit testing:

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

Integration testing:

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

Functional test:

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items:

Valid Input : identified classes of valid input must be accepted.

Invalid Input : identified classes of invalid input must be rejected.

Functions : identified functions must be exercised.

Output : identified classes of application outputs must be exercised.

Systems/Procedures : interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

System Test:

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

7.2Black Box Testing:

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box .you cannot “see” into it. The test provides inputs and responds to outputs without considering how the software works.

7.3White Box Testing:

White Box Testing is a testing in which in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is purpose. It is used to test areas that cannot be reached from a black box level.

Unit Testing:

Unit testing is usually conducted as part of a combined code and unit test phase of the software lifecycle, although it is not uncommon for coding and unit testing to be conducted as two distinct phases.

Test strategy and approach

Field testing will be performed manually and functional tests will be written in detail.

Test objectives

- All field entries must work properly.
- Pages must be activated from the identified link.
- The entry screen, messages and responses must not be delayed.

Features to be tested

- Verify that the entries are of the correct format
- No duplicate entries should be allowed
- All links should take the user to the correct page.

Integration Testing

Software integration testing is the incremental integration testing of two or more integrated software components on a single platform to produce failures caused by interface defects.

INPUT DESIGN

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system.

The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the following things:

- What data should be given as input?
- How the data should be arranged or coded?
- The dialog to guide the operating personnel in providing input.
- Methods for preparing input validations and steps to follow when error occur.

OBJECTIVES

1.Input Design is the process of converting a user-oriented description of the input into a computer-based system. This design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerized system.

2.It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The goal of designing input is to make data entry easier and to be free from errors. The data entry screen is designed in such a way that all the data manipulates can be performed. It also provides record viewing facilities.

3.When the data is entered it will check for its validity. Data can be entered with the help of screens. Appropriate messages are provided as when needed so that the user will not be in maize of instant. Thus the objective of input design is to create an input layout that is easy to follow

OUTPUT DESIGN

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making.

1. Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each

output element is designed so that people will find the system can use easily and effectively. When analysis design computer output, they should Identify the specific output that is needed to meet the requirements.

2.Select methods for presenting information.

3.Create document, report, or other formats that contain information produced by the system.

The output form of an information system should accomplish one or more of the following objectives.

- Convey information about past activities, current status or projections of the
- Future.
- Signal important events, opportunities, problems, or warnings.
- Trigger an action.
- Confirm an action.

CHAPTER-8

OUTPUT SCREENS

8.1 USER INTERFACES:

Figure. Noise Removal of vehicle

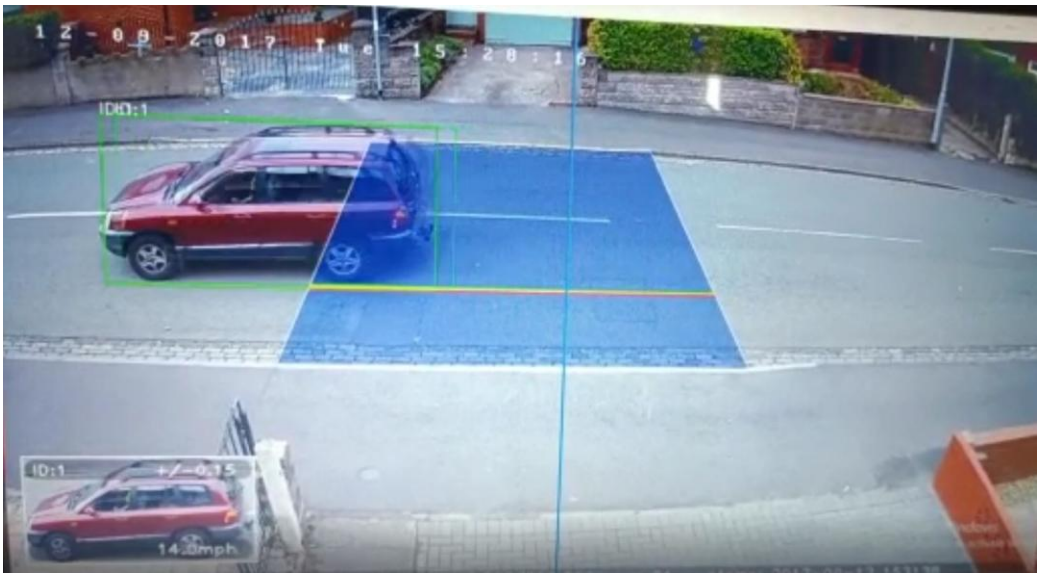
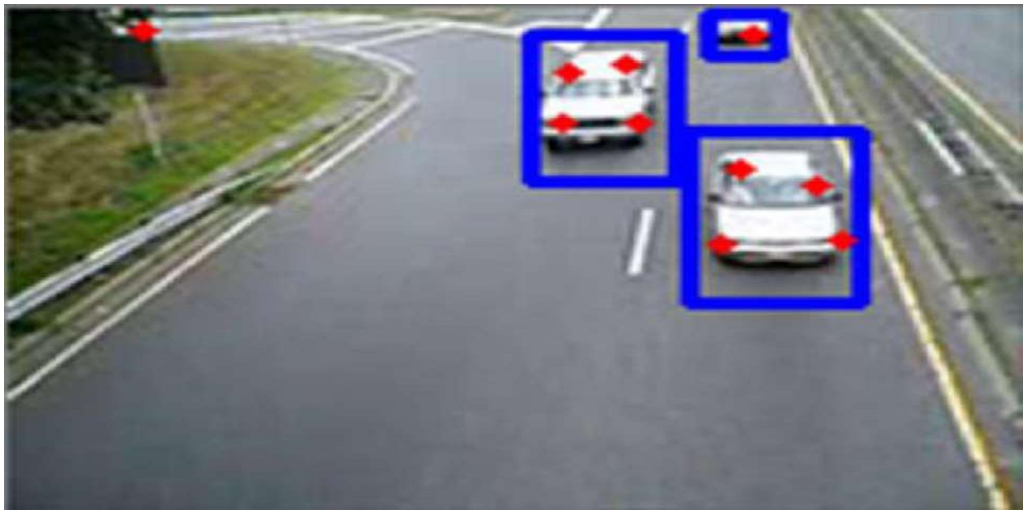


Figure: Detection of vehicle and estimating the speed

8.2 OUTPUT SCREENS:



F

igure:Identifying the corners of vehicle

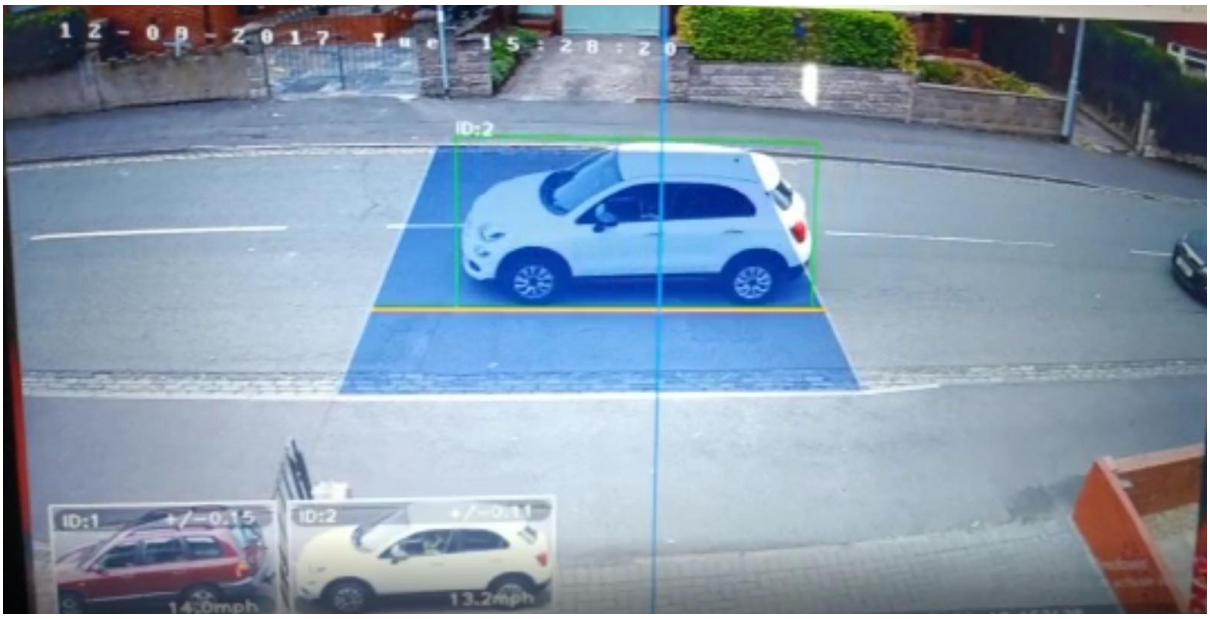
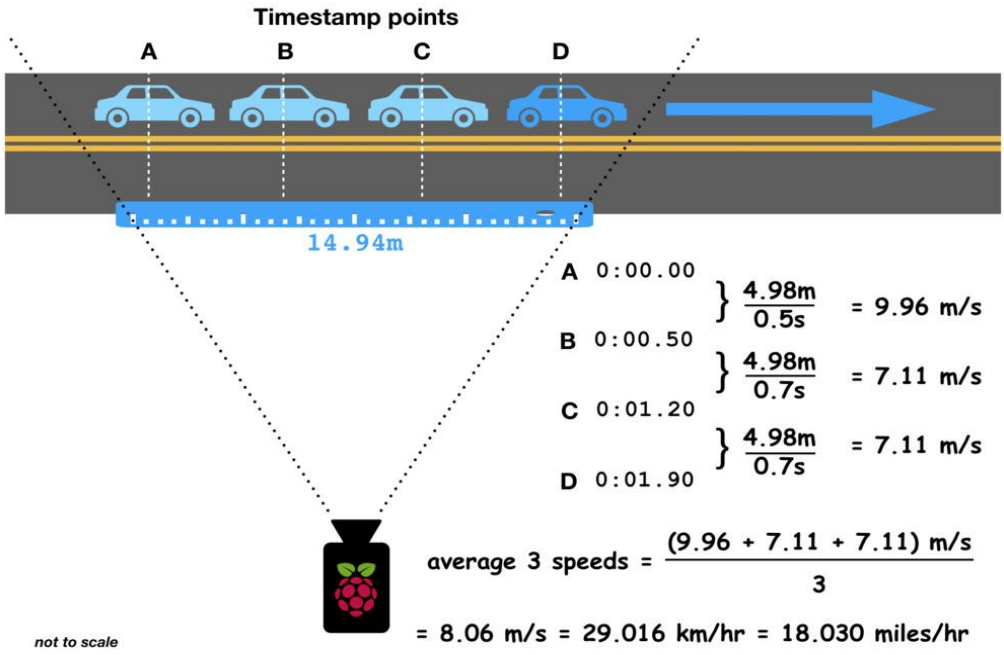


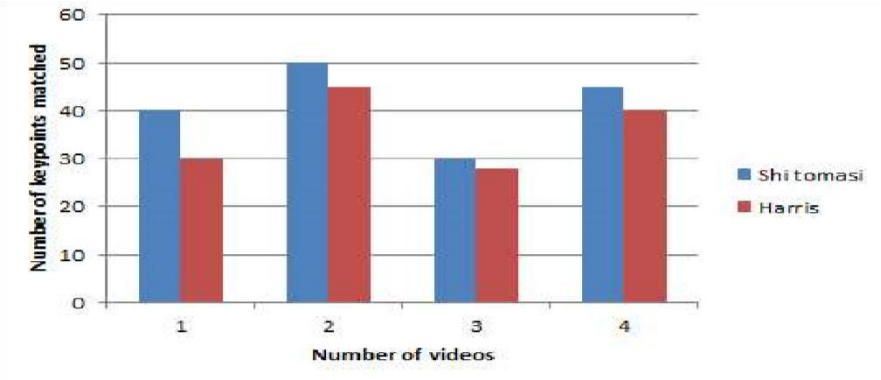
Figure: Detection of vehicle

9.EXPERIMENTAL RESULTS



Analysing Experimental Data:

The table 1 shows the experimental results for speed estimation of different cars.



Comparison between Shi-Tomasi and Harris corner detection techniques

Table 1: Speed Evaluation

Sl.no.		Distance covered(m)	Actual speed(km/hr)	Estimated speed(km/hr)	Error rate
Car1		25	43	42.6	-0.4
Car2		21	39	39.4	+0.4
Car3		18	36	36.2	+0.2

From the above table we can infer that results have minimum error rate which can be further improved.

10. CONCLUSION AND FUTURE ENHANCEMENTS

CONCLUSION:

A real time system for detection, tracking and classification of vehicles. Pattern recognition helps eliminates false alarms caused by shadows and headlight reflections. The system not only lessens the work load on the traffic police. For the concerned educational institute's authority whose workload is tremendously reduced and for the government at large, whose services are protected from exploitation.If over speed is detected it sends alert massage to driver, Repeated violations results to increase in penalty amount which will help in reduction of violations by the vehicle user.Can easily incorporate additional knowledge to improve calibration accuracy, quick setup for short term data collection applications.

FUTURE ENHANCEMENTS:

VEHICLE SECURITY: The lost out vehicle cases are increasing day by day the stolen vehicle can be easily compared with registered entry of stolen vehicles.

PARKING: The vehicles can be registered using automatic system with this system in parking lounge or for similar purposes

11.REFERENCES

[1]H. Chung-Lin and L. Wen-Chieh, "A vision-based vehicle identification system," in Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on, 2004,pp.364-367Vol.4.

[2] Z. Wei, et al., "Multilevel Framework to Detect and Handle Vehicle Occlusion," Intelligent Transportation Systems, IEEE Transactions on, vol. 9, pp. 161-174, 2008.

[3] N. K. Kanhere and S. T. Birchfield, "Real-Time Incremental Segmentation and Tracking of Vehicles at Low Camera Angles Using Stable Features," Intelligent Transportation Systems, IEEE Transactions On,vol9,pp.148-160,2008.

[4] N. K. Kanhere, "Vision-based detection, tracking and classification of vehicles using stable features with automatic camera calibration," ed, 2008, p. 105.

[5] A. H. S. Lai, et al., "Vehicle type classification from visual-based dimension estimation," in Intelligent Transportation Systems, 2001. Proceedings. 2001 IEEE, 2001, pp. 201-206.

[6] Z. Zhigang, et al., "A real-time vision system for automatic traffic monitoring based on 2D spatio System based on artificial intelligence.

[7]Ibrahim O, ElGendy H, Ahmed M. ElShafee. Speed Detection Camera System using Image Processing. International Journal of Computer and Electrical Engineering. 2011 December ; 3(6): p. 1-8. [11] P. Nyoni and M. Velempini, “Privacy and user awareness on facebook,” South African Journal of Science, vol. 114, no. 5-6, pp. 27–31, 2018.

[8] Elysia , Gunawan FE, Vehicle Tracking using Computer Vision Technique for Car-Following Model, 2015.

[9] Burger W, Burge MJ. Principles of Digital Image Processing: Core Algorithms: Springer; 2009.

[10] Wang G, Li J, Zhang P, Zhang X, Song H. Pedestrian Speed Estimation Based on direct Linear in International Conference on Audio, Language and Image

Processing;2014; Shanghai. p. 1-5.

[11] Kim H. Vehicle Detection and Speed Estimation for Automated. TehnickiVjesnik.
2019; 26(1): p. 87-94.

[12]Drews P, Williams G, Goldfain B, Theodorou EA, Rehg JM. Vision-Based High-Speed Driving With a Deep Dynamic Observer. IEEE Robotics and Automation Letters.2019; 4(2)

[13] . Gandhi SA, Kulkarni CV. MSE Vs SSIM. International Journal of Scientific & Engineering Research.2013 July 7; 4(7): p. 930.

[14]KaewTraKulPong P, Bowden R. An Improved Adaptive Background Mixture Model for Real-time Tracking with Shadow Detection. In Remagnino P, Jones GA, ParagiosN,Regazzoni CS, editors. Video-Based Surveillance Systems: Kluwer Academic; 2002.

[15]Firdaus A. Kopo Toll Road Video. [Online].; 2016. Available from: https://youtu.be/OcKtb6reBl0

12. PUBLICATIONS

International Conference on “Vehicle Detection and Speed Detection” (ICICCI-21)

Paper ID: ICICCI-21-0140

A.SAICHARAN REDDY is currently pursuing his Bachelor of Technology in the stream of Computer Science and Engineering at St. Martin's Engineering College. He completed his intermediate from Sri Chaitanya Junior College and 10th class from Sri Chaitanya School. He is one of the volunteer in street cause hyderabad. His technical skills include C, Python, HTML. He also has a basic understanding of C++.He also participated in Online Two Day National Level Seminar on "Recent Trends in Cloud Computing Fog and Edge Computing" from 18th to 19th June 2021.He participated in National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020. His areas of interest are Python, Artificial Intelligence, Machine Learning. He completed few certification courses from online platforms like Coursera, CursaApp and SoloLearn.

G. RETHIK REDDY is currently pursuing his Bachelor of Technology in the stream of Computer Science & Engineering at St. Martin's Engineering College. He completed his intermediate from Urbane Junior College and schooling from ShantiniketanVidyalaya . He was a volunteer in the student run NGO, Street Cause, during the year 2017-2018. His technical skills include C, HTML ,CSS and Python. His participations include:, Workshop on "Arduino/Robotics" which was conducted in the college on 12th February 2019 and 13th February 2019, Workshop on "Ethical Hacking" which was conducted in the college on 31st January 2020 and 1st February 2020, National Level Three Day Online Workshop on "AI & ML in Speech and Audio Processing" which was conducted from 10th to 12th December 2020.He was also a student organizing member during two days "National Level Hackathon-2020" held on 7th and 8th February 2020 at the college. He spends his free time taking online certification courses related to his field of study as well as personal interests from platform such as Coursera, Cursa and EdX. His areas of interest are Python, Artificial Intelligence, Machine Learning and Deep Learning.

14.APPENDICES

Feedback page code:

```
importnumpy as np
import cv2
import time
car_cascade = cv2.CascadeClassifier('hand.xml')
cap = cv2.VideoCapture('car.mp4')
wide=0.1   #depends upon size of car(~2.5)
flag=True
start=end=0
time_diff=0
while(cap.isOpened()):
ret, img = cap.read()
height,width,chan=img.shape
   #print(height,width,chan)
gray = cv2.cvtColor(img, cv2.COLOR_BGR2GRAY)
cars = car_cascade.detectMultiScale(gray, 1.3, 5)
   #crp=gray[0:480,0:int(width/2)+10]
for(x,y,w,h) in cars:
cv2.rectangle(img, (x,y), (x+w,y+h), (0,255,0),2)
center_x=(2*x+w)/2
center_y=(2*y+h)/2
      #print(center_x,center_y)
      dist1=((wide*668.748634441)/w)
      #print("Distance from car:",round(dist1,2),"m")
roi_gray = gray[y:y+h,x:x+w]
roi_color = img[y:y+h,x:x+w]
      dist0=((wide*668.748634441)/w)
actual_dist=dist0*(width/2)/668.748634441
      #print("Actual Distance:",actual_dist)
if flag is True and int(round(center_x)) in (range(0,80) or range(400,480)):
start=time.time()
flag=False
```

```
        #print("Start:",start)
    if flag is False and int(round(center_x)) in range(int(round(width/2))-10,int(round(width/2))+10):
end=time.time()
time_diff=end-start
        #print("End:",end)
flag=True
#print("Time Difference:",time_diff)
iftime_diff>0 and s_flag==True:
velocity=actual_dist/time_diff
    #print(round(start),round(end))
vel_kmph=round(velocity*3.6,2)
print("Speed:",vel_kmph,"kmph")
print("Distance from car:",round(dist1,2),"m")
s_flag=False
  cv2.line(img,(int(width/2),0),(int(width/2),height),(255,0,0),2)
cv2.imshow('frame',img)
if cv2.waitKey(1) & 0xFF == ord('q'):
break
cap.release()
cv2.destroyAllWindows()
```