# Detection Analysis Of URL Attack In SDR Systems For Network Data Security Using Machine Learning Methods

**[1]Parvathapuram Pavan Kumar,**

*Research Scholar, Department of ECE, VISTAS, India*

**[2]Dr. T.Jaya,**

*Assistant Professor, Department of ECE, VISTAS, India*

**[3]Dr. V.Rajendran,**

*Director, Department of ECE, VISTAS, India*

## *Abstract*

*During day-to-day life, most of the people using internet for net banking, shopping, etc which provides risk to the user from internet. Spoofing (Phishing) means the appearance of treachery wherein the aggressor endeavor to discover sympathetic information such as users access code or account details through conveyance it to either electronic mail or some more communication lines as a trustworthy individual or someone. In general, the affected party got statement which emerges to be mailed by recognized person or association. This statement consists of either malevolent software focused the user's desktop or having associations of personal endure to malicious websites with the intention of scam them into revealing private information namely account ID, credit card, password details. Now, lack of security in the data network caused if aspiration to analyze with big data, Artificial Intelligence and Internet of Things. In such type of situation, URL phishing attacks in the website have to be found through statistical analysis in the proposed study. Such type of phishing URL has to be classifying either as phishing URL or non-phishing URL via statistical analysis in machine learning techniques. Such type of classification methods can be done through data pre-processing, feature extraction, training, testing and validation phase of machine learning techniques. This proposed manuscript deals with data mining classification methods like decision tree classifier, K- Nearest Neighbor classifier and ensemble gradient boosting classifier also to classify the phishing URL as malicious or normal for various features of URL and websites. Thereby, the enhancement of overall model performance can be evaluated by finding metrics as accuracy, precision and recall.*

***Keywords**: Phishing, Uniform Resource Locator (URL), Decision Tree Classifier, K-Nearest Neighbor Classifier, ensemble Gradient Boosting Classifier, accuracy.*

## 1. Introduction

Most of the users unintentionally select fraudulent domains each and every hour day-by-day. The aggressors has been attacking both the organizations and the users. In accordance with the third Microsoft Computing Safer Index Report (MCSIR) liberated on the year 2014 feb, the overall international influence of phishing may perhaps as high as five billion dollar. In a nutshell, phishing URL begins with phishing communication through text message, electronic mail, and also internet community. Pratima Sharma et. al [2] exposed how URL attacks give risk to the internet and also how to handle the risk during attack in the network. This study mainly focused on giving better perceptive of URL harasses, URL manipulation attack along with malicious machinery. Doyen Sahoo et. al [4] surveyed malicious URL attack using machine learning techniques based on some process mainly focused on spam detection, web

ISSN: 2233-7857 IJFGCN

page classification, feature extraction and also scheming novel learning algorithms for detecting malicious URL attack in the website.

The overall configuration of URL along with its significant parts can be formulated as below.

**Table 1. Configuration of Unifrom Resource Locator with its significant parts**

| |
|---|
| https://www.exemplaryurl.com/information/aboutus.html |

| | |
|---|---|
| https:// | Protocol |
| exemplaryurl.com | Domain name |
| www. | Subdomain name |
| Information | Directory |
| information/aboutus.html | Path |
| aboutus.html | Page/ file |
| www.exemplaryurl.com | Name of the host |

Ravi kumar G et. al [3] demonstrates various classification methods explicitly KNN, SVM, Naïve Bayes, as well as Random Forest. Among all algorithms, Support Vector Machine generates better accuracy based upon predicting phishing or non-phishing category of URL in the website and regression technique also used for uninterrupted prophecy of homogeneous data for better improvement of the model.

## 2. Existing Work

Hyunsang Choi et. al [5] composed different real dataset from various resource to detect benign URL from open directory, spam URL from Spam spy, phishing URL from phish tank community furthermore malware URL obtained from DNS-BH also. Using that, performance of the model can be evaluated for both malicious detection and attack identification as well.

Harshal Tupsamudre et. al [6] suggested the performance improvement of URL based detection technique for feature extraction technique based on segmentation of words, Phish-list, numerical features. The efficacy of the model evaluated through logistic regression classifier during training phase includes 100,000 URL. The accuracy can be evaluated founded on word segmentation, phishing list, and other numerical features also.

Immadisetti et. al [7] exposed some classification technique such as Black listing, Heuristic classifier for detecting malevolent URL attack in the website that classifies either phishing or legitimate.

Dharmaraj et. al [9] recommended binary dataset and multi-class dataset has constructed using 49935 malevolent and compassionate URL. One-vs-one (OVO) SVM, One-vs-all (OVA) SVM, and Online Confidence multi-class weighted learning for evaluating the efficiency and performance.

Joby James et. al [10] anticipated many classification algorithm namely Naïve Bayes, J48, IBK, SVM, focused on splitting phase as 60% and 90% for finding success rate, error rate, accuracy and confusion matrix. This novel work separates phishing and non-phishing URL achieves the better model performance depends upon the tuning of hyper parameters.

Cho Do Xuan et. al [11] suggested many machine learning techniques like Random Forest, SVM, Naïve Bayes, Regression, Clustering, collaborative filtering. The experimental outcome reveals that among all classifiers, random forest works very fast and accurate on a very large datasets.

Shraddha Parekh et. al [12] focused primarily on Random forest classification algorithm only to identify phishing website by categorize into three phases namely parsing, Heuristic classification of data, and performance analysis of dataset.

Vaibhav Patel et. al [13] conferred on three approaches behind this work as initially scrutinized different parameters in URL, second approach is to examine the legitimate website where it is being hosted, who is handling the website, finally, analyzing perception appearance for guiding authenticity of website.

## 3. Outline of proposed work

### 3.1 Survey of our work

The proposed method consists of three phases 1: Data Collection 2: Supervised learning algorithm 3: URL detection & Identify type of attack

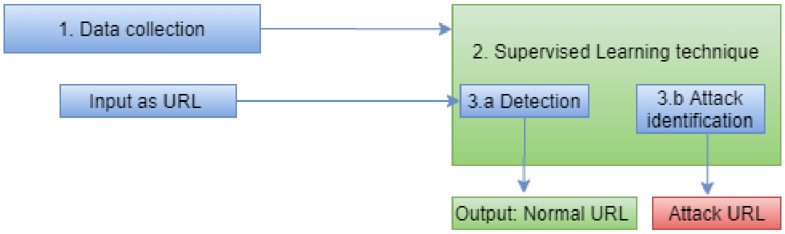

**Figure 1. Outline of the proposed work**

These phases can manage consecutively same as in batched learning else in interleaving manner, classification methods for detecting the outcome and also identification of attack URL or normal.

### 3.2 Features used for Phishing domain detection

Mainly, there are four various kinds of features in machine learning method for phishing domain detection process. They are explained as follows

1. URL Based Features 2. Domain Based Features

3. Content Based Features 4. Page Based Features

1. URL Based Features: URL based features primarily used to investigate the website to make a decision just phishing URL or any other attack website. Some URL based features used in this study as digit count, total length, total number of sub domains in URL.

2. Domain Based Features: The domain based features intention is to detect phishing domain names such as IP address in blacklists, domain entropy, long domain token length etc.

3. Content Based Features: These kinds of features needs energetic examine of target domain which is used for phishing or legitimate. A quantity of features used in the proposed work as title of the page, Meta tags, hidden text, body text, Images etc.

4. Page Based Features: These kinds of features provide information about client commotion on target site includes category of domain, total number of times domain visit.

### 3.3 Workflow for the proposed method

Suppose if we wish to pipeline with big data analysis, Internet of Things and Artificial Intelligence, lack of security in the network happens. To overcome the issues in the network, finding URL phishing attack in the websites have to be found. Thus, the flow of work for proposed model shows how the URL website classified as phishing URL or non-phishing URL.

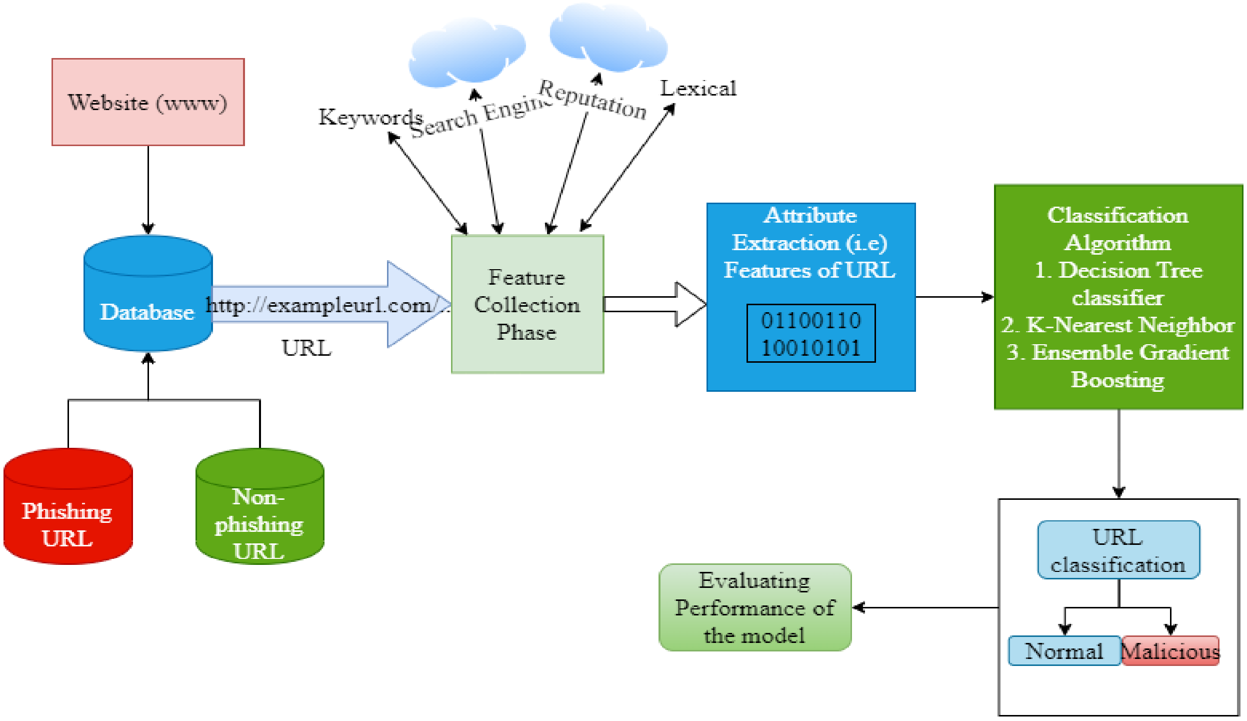

**Figure 2. Workflow for the proposed method**

The database consists of Uniform Resource Locator website as http://exampleurl.com, the feature collection phase includes keywords, search engines, reputation, as well as lexical words goes behind pre-processing stage which complete feature extraction as 01100110, Min Max scalar normalization, $x_{new} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$ leads to feature transformation moves to further stage as algorithm classification such as decision tree, KNN, ensemble gradient boosting. By Usage of classification technique, the website categorized into phishing or legitimate, the model performance may be evaluated with good quality of metrics estimation. After encoding, number of features described as 1389 were established.

### 3.4 Usage of algorithm for proposed model:

The proposed algorithm mainly used for categorizing the phishing Uniform Resource Locator (URL) detection as normal or malicious. Here are the procedures summarizing as follows:

*i) Decision Tree algorithm:* Decision tree algorithm is a data mining technique which is used to construct either classification or regression models. In this study, the scams emails can be identified in the Uniform

Resource Locator could be classified either as Phishing URL or Non-phishing URL through decision tree classifier model. The illustration for the proposed study as follows

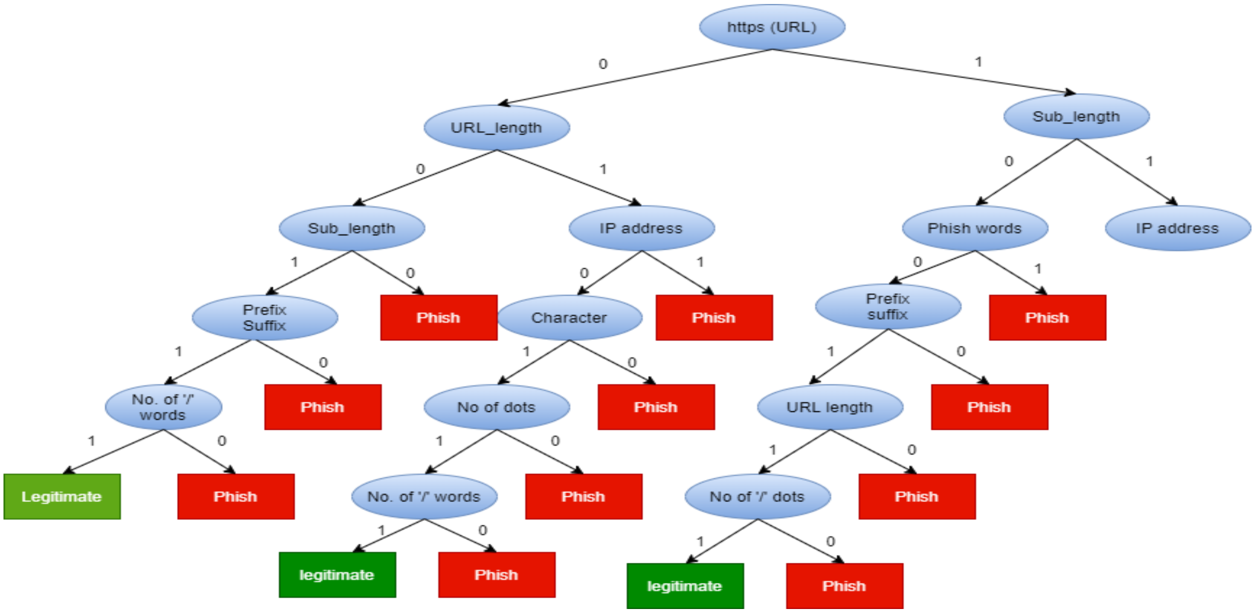

**Figure 3. Decision tree for identifying phish or legitimate**

The above diagram explains how to classify the dataset into legitimate (normal) as well as Phishing (attack). The URL (https) can be classified into length, sub length, and then prefix suffix words, characters, finally classify it as legitimate and phishing using decision tree classification technique.

### 3.5 Detecting email Phishing decision tree

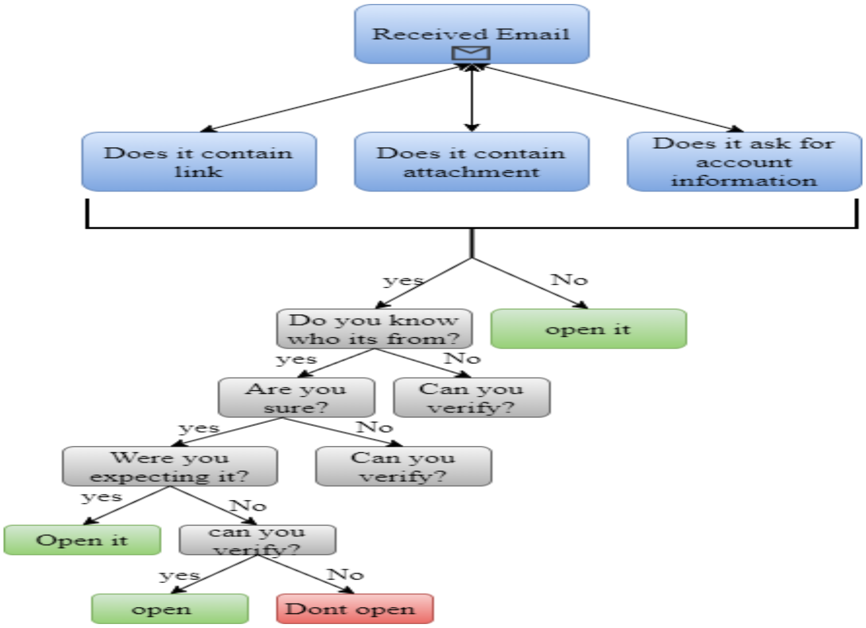

**Figure 4. Decision Tree for Detecting email phishing**

The above decision tree shows the detection of email phishing using data mining technique. Here, if any email received by the user, they must check whether it contain link or it consists of any attachment or it may ask any account details such as customer id, credit card information etc.. If is it true, then the user should check where the mail comes from? Or else open it. Confirm whether the user should completely know the person whom it was sent by or else can you verify if yes open it otherwise doesn't open. This method explains how to detect the email phishing through decision tree algorithm in data mining technique.

*ii) K-Nearest Neighbor:* K-Nearest Neighbor is the simplest machine learning technique mainly focused on supervised learning algorithm which solves both classification and regression issues. KNN algorithm used at the training stage of the dataset in which the dataset categorized into new form by classifying as class A and class B for further processes. To establish which of the k instances in the training dataset are further related to a new input, a space determine is used.

*iii) Ensemble Gradient Boosting*: Gradient Boosting is prevailing ensemble machine learning technique that mainly utilizes decision trees for further classification which categorize either attack or normal. The generalization method of Ada boosting is primarily Gradient Boosting to enhance the model performance however, initializing thoughts from bootstrap aggregation such as randomly selecting features along with samples for promote enhancement in the model. Through ensemble gradient boosting, classification and regression issues could be solved. Tie Li et. al [1] proposed arrangement of both linear and non-linear space transformation methods for classifying malicious URL attack and normal. Supervised learning method namely KNN, Neural Networks and SVM classifiers for categorize the URL attack from overall website. The results of each classifiers reveals better performance of the model with KNN predict accuracy as 86%, SVM calculate accuracy as 81% and MLP accuracy forecasting as 82%. In this model, ensemble gradient boosting achieves accuracy around **92, 97 and 98 percent** placed in 1%, 10% and 100 % of training samples.

## 4. Dataset representation

In this work, the dataset has been collected from Kaggle data source which has 15367 samples with 78 features each. The dataset has split into two phases in the 80:20 ratios with training set as 80% with 12293 samples, testing set as 20% with 3074 samples. Now, the percentage of attacks can be found as 49.37 %. To discover the attacks in the network, the novel method projected with classification method using statistical analysis in ML approach. Ammar Yahya Daeef et. al [8] demonstrates N-gram method for independently built the host, path and query via phishing and non-phishing datasets. Herein, three classifiers namely J48, Logistic Regression, Support Vector Machine have been carried out for phishing detection area. The datasets collected as 4,65,461 URL from phish tank as well as 4641 URL features from Open phish for detecting phishing.

**Table 2. Dataset description**

| Total number of records | Number of attacks | Number of normal records | Percentage of attacks |
|---|---|---|---|
| **15367** | 7586 | 7781 | 49.37% |

Some of the features used in this novel study specifically Query length, domain token count, path token count, average domain token length, long domain token length, average path token length, token id, char comp vowels, char comp ace, Id1 url, Symbol count file name, symbol count extension, symbol count after path, entropy domain, Entropy filename, Entropy extension, Entropy_afterpath, Attack type. Seventy-one significant features can be identified with an alpha level as 0.05.

### 4.1 Metrics Evaluation:

The overall performance of the model can be estimated via metrics calculation such as accuracy, precision and recall. Though, predicting accuracy need to be found, confusion matrix has to be calculated as shown

**Table 3. Confusion matrix**

| | **Class 1<br>(Predicted value)** | **Class 2<br>(Predicted value)** |
|---|---|---|
| **Class 1<br>(Actual value)** | TP (True Positive) | FP (False Positive) |
| **Class 2<br>(Actual value)** | FN (False Negative) | TN (True Negative) |

The confusion matrix (error matrix) is defined as explicit outline table that allows prophecy performance of an algorithm especially supervised algorithm in machine learning classification issues during statistical analysis.

*Accuracy:* Accuracy is the measure of selecting the number of corrected predictions from total length of the attacks. Based upon accuracy calculation metrics, the performance of the model will be guessed whether the model is good or bad.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

$$Accuracy\,Naive = \frac{float\,(TP + TN)}{float\,(length(attacks))}$$

*Recall*: Recall is defined as the fraction of predicted positive observations acceptably to total actual class observations.

$$True\,Positive = \frac{True\,Positive}{\Pr edicted \operatorname{Re} sults} \tag{2}$$

(or)

$$True\,Positive = \frac{float\,(True\,Positive)}{float\,(True\,Positive + False\,Negative)}$$

*Precision*: Precision refers the ratio of true positive to the actual results which means estimate the total number of positive class predictions that fit in to the positive class.

$$\Pr ecision = \frac{float\,(True\,Positive)}{float\,(Actual \operatorname{Re} sults)} \tag{3}$$

*F-Score:* F-score offers a single score that together both precision and recall of the concerns in one number. F-score is the reciprocal mean of arithmetic mean of recall as well as precision

$$F - Score = \frac{2}{\frac{1}{\Pr ecision} + \frac{1}{\operatorname{Re} call}} \tag{4}$$

On the other hand, F-Score can be represented as

$$F - Score = 2 * \frac{\Pr ecision * \operatorname{Re} call}{\Pr ecision + \operatorname{Re} call}$$

*FNR (False Negative Rate):* FNR is calculated using the formula

$$FNR = \frac{FN}{FN + FP} \tag{5}$$

### 4.2 Finding Mean, Median, Mode using Statistical Analysis

Herein, the mean, median, mode values for both phishing URL as well as non- Phishing URL can be evaluated for certain specified features using statistical analysis.

*Mean:* Mean is defined as the average of all numbers containing in the list. Sometimes, Mean is called as Arithmetic Mean (AM). The mean value can be intended using the following formula.

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{6}$$

*Median:* To calculate median, first arrange the given list of numbers in an ascending order. After that, the middle number should be chosen from the ordered list.

*Mode:* Mode is defined as a number occur repeatedly contained in set of numbers.

**Table 4. Central tendency measures**

| Central Tendency Measures | | |
|---|---|---|
| Appraise | **Modus operandi** | **Depiction** |
| **Mean** | Σx/n | Average of list of given numbers. |
| **Median** | n+1/2 position | Step 1: Arrange the list in ascending order<br>Step 2: Choose middle value |
| **Mode** | - | Most repeated value |

For any model, this can be achieved by merely importing an inbuilt library 'statistics' in Python 3 using inbuilt function namely mean (), median (), mode ().

Additional formula for standardization, SD, and proportion value is shown below

| Standardization formula | Standard Deviation | Proportion value |
|---|---|---|
| $z = \frac{x - \mu}{\sigma}$ | $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$ | $True\ precentage = 100 * \frac{float(True)}{float(True + False)}$ |

**Measures for central tendency for both attack and normal (some specific features)**

**Table 5. Measures for central tendency (Mean median mode)**

| Features | Attack | | | Normal | | |
|---|---|---|---|---|---|---|
| | Mode | Median | Mean | Mode | Median | Mean |
| Querylength | 0.000000 | 0.000000 | 2.818613 | 0.0 | 0.000000 | 4.057705 |
| domain_token_count | 3.000000 | 3.000000 | 3.016346 | 2.0 | 2.000000 | 2.082894 |
| path_token_count | 4.000000 | 5.000000 | 6.343396 | 8.0 | 10.000000 | 10.557255 |
| avgdomaintokenlen | 6.000000 | 6.000000 | 6.448547 | 5.5 | 5.500000 | 5.270317 |
| longdomaintokenlen | 9.000000 | 10.000000 | 12.219879 | 8.0 | 8.000000 | 7.889988 |
| ... | ... | ... | ... | ... | ... | ... |
| Entropy_Domain | 0.916850 | 0.817188 | 0.823926 | 1.0 | 0.884870 | 0.883778 |
| Entropy_DirectoryName | 0.871049 | 0.777077 | 0.628993 | 0.0 | 0.750563 | 0.491593 |
| Entropy_Filename | 1.000000 | 0.887436 | 0.780120 | -1.0 | 0.743766 | 0.571433 |
| Entropy_Extension | 0.000000 | 0.579380 | 0.430392 | 0.0 | 0.000000 | 0.199649 |
| Entropy_Afterpath | -1.000000 | -1.000000 | -0.808380 | -1.0 | -1.000000 | -0.641047 |

### 4.3 Experimental analysis

The statistical analysis for some of the features in URL attack showed how the phishing attacks classified as normal (Blue color) or attack (red color). The features namely query length, domain token count, average domain token length, long domain token length, entropy domain, entropy directory name, entropy filename, entropy extension, entropy after path were shown the classification of attacks.

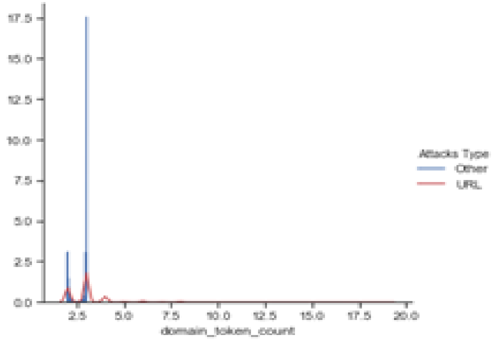

Figure 5. Domain token count

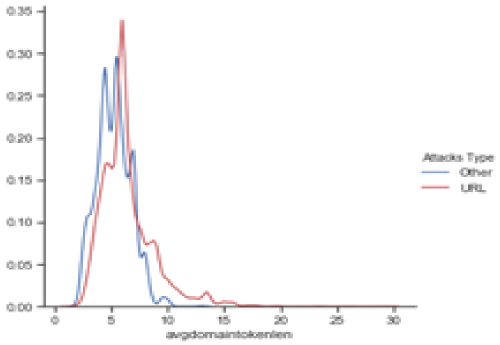

Figure 6. length of average domain token

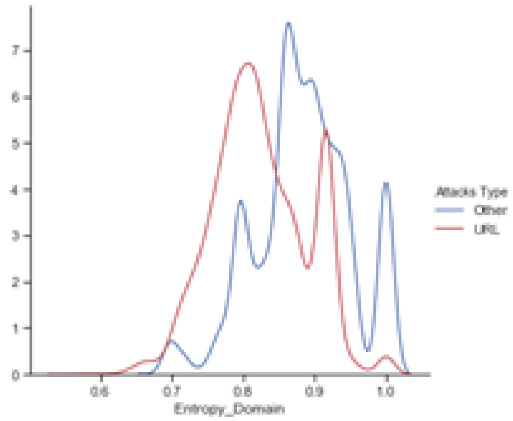

Figure 7. Entropy domain

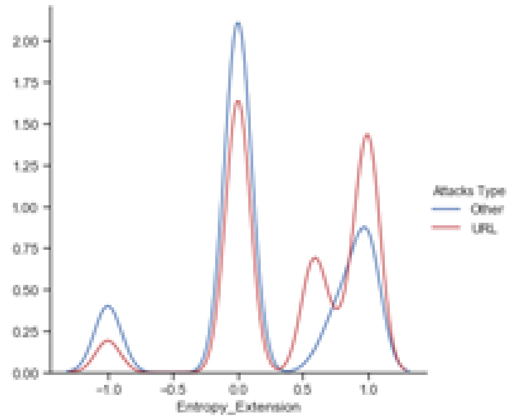

Figure 10. Entropy extension

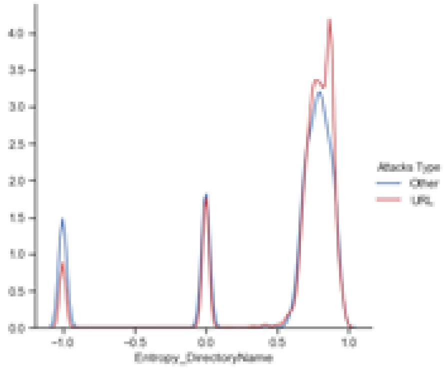

Figure 8. Entropy for directory name

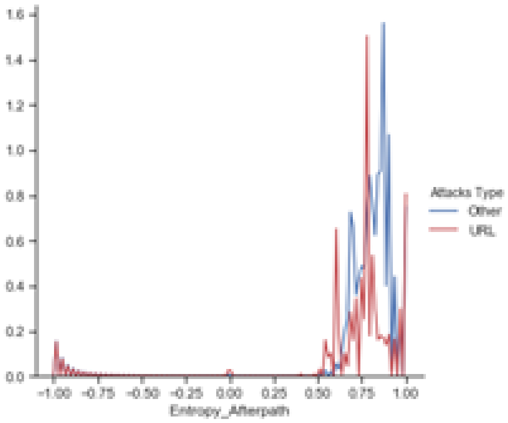

Figure 11. Entropy afterpath

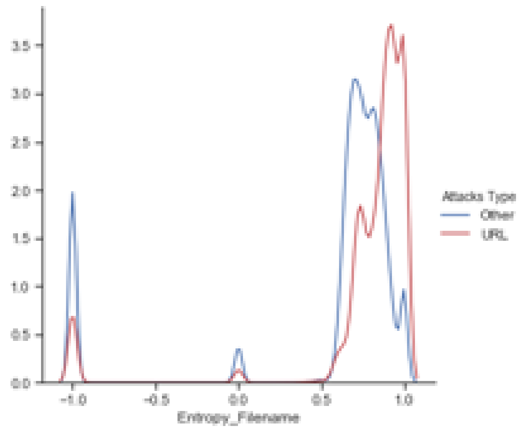

Figure 9. Entropy filename

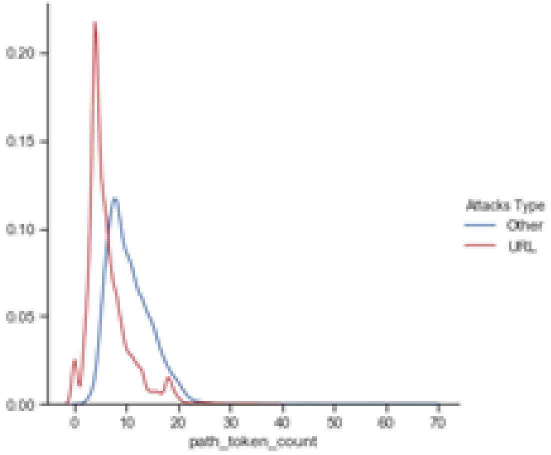

Figure 12. path token count

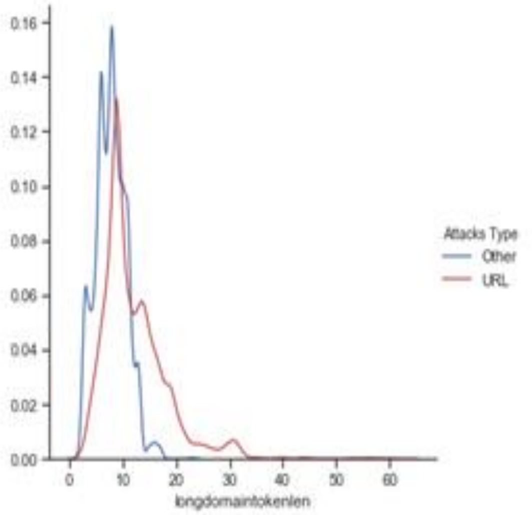

Figure 13, Long domain token length

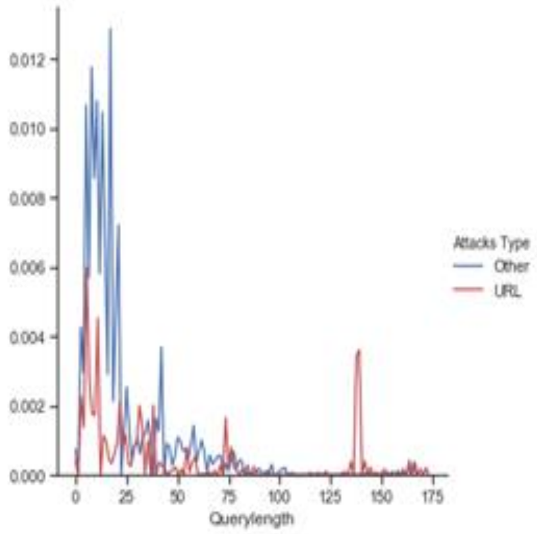

Figure 14, Query length

### 4.4 Comparative Analysis among different existing work with proposed model

The comparative analysis reveals the comparison between existing models and proposed model with metrics, algorithm used in machine learning techniques

**Table 6. Comparative analysis for existing and proposed model**

| Existing work | Features | Machine Learning Technique | Metrics | | |
|---|---|---|---|---|---|
| | | | **FNR** | **0/1 (URL attack/normal)** | **Accuracy** |
| **Tie et. al [1]** | 3,31,622 | SVD, SVDR, NYS, NYS-DML, DML-NYS | ✓ | ✓ | 86.4% |
| **Pratima et. al [2]** | Real data | URL manipulation attack | - | ✓ | - |
| **G. Ravi kumar et. al [3]** | 21 fixed features | SVM, NLP | - | ✓ | - |
| **Doyen sahoo et. al [4]** | | SVM, Logistic regression, Naïve Bayes, Decision tree | | | |
| **Hyunsang et. al [5]** | 40, 000 Benign URL<br>32,000 malicious URL | SVM, multi-label classification | - | ✓ | 93% |

| | | | | | |
|---|---|---|---|---|---|
| **Harshal et. al [6]** | 100,000 | Logistic Regression | ✓ | ✓ | - |
| **Immadisetti et. al [7]** | 18 features | Black listing<br>Heuristic classification | - | ✓ | - |
| **Ammar et. al [8]** | 46,5461 from phish tank<br>4647 from open phish | J48 classifier | ✓ | ✓ | 93% |
| **Dharmaraj et. al [9]** | 26041 benign URL | One-vs-one SVM, One-vs-All SVM, Multi class online weighted learning | ✓ | ✓ | ✓ |
| **Joby James et. al [10]** | 17000 Phishing URL<br>20,000 benign URL | Naïve Bayes,<br>J48<br>SVM<br>K-NN | ✓ | ✓ | 91.08% |
| **Cho Do et. al [11]** | 10,000 malicious URLs<br>4,70,000 normal URLs | SVM<br>RF | ✓ | ✓ | 93.39%<br>96.28% |
| **Shraddha et. al [12]** | 31 features | Random Forest | ✓ | ✓ | 95% |
| **Valibhav et. al [14]** | 9076 test websites | Logistic Regression<br>Decision Tree<br>Random Forest | ✓ | ✓ | 96.23%<br>96.23%<br>96.58% |

### 4.5 Proposed work Accuracy estimation

The proposed work established classification algorithm for finding accuracy in both training and testing to enhance the performance.

**Table 7. Accuracy estimation**

| **Proposed Algorithm** | **Training Accuracy** | | | **Testing Accuracy** | | |
|---|---|---|---|---|---|---|
| | 1% | 10% | 100% | 1% | 10% | 100% |
| Gradient Boosting | 0.933 | 0.997 | 0.98 | 0.923 | 0.968 | 0.979 |
| Decision Tree Classifier | 0.937 | 1.00 | 1.00 | 0.906 | 0.951 | 0.975 |
| KNN classifier | 0.883 | 0.957 | 0.98 | 0.863 | 0.94 | 0.960 |

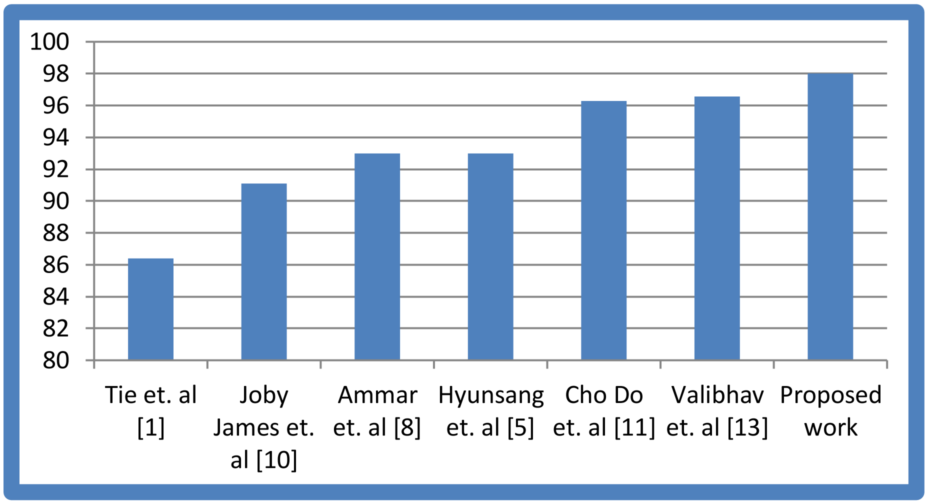

**Figure15. Accuracy calculation for existing and proposed**

## 5. Conclusion

In this proposed study, various data mining techniques has been proposed to recognize phishing (spoofing) websites through classification algorithm namely decision tree, KNN, ensemble gradient boosting using python programming from anaconda software. Herein, we experimentally confirmed that the proposed features like length of the query, domain token count including 78 features are most appropriate for finding phishing websites in URL as well.  The metrics achievement along with related work also manifested the level of accuracy of ensemble gradient boosting around 98% among all three algorithms. The detection of phishing attack can be identified via metrics specifically False Negative Rate, Accuracy, Precision, Recall, F-Score for scrutinizing purpose thus generates the performance of the model. The future scope is to detect phishing URL attack through some other related machine learning classification technique for higher performance rate.

## 7. References:

1. T. Li, G. Kou, Y. peng, “Improving malicious URLs detection via feature engineering: Linear and nonlinear space transformation methods”, Information systems 91 (2020).
2. Pratima Sharma, Bharti Nagpal “A study on URL manipulation attack methods and their countermeasures”, International Journal of Emerging Technology in Computer Science and Electronics (IJETCSE)Pg: 116-119, 2015.
3. G. Ravi Kumar, s. Gunasekaran, Nivetha R, Sangeetha Prabha. K, Shanthini. G, Vignesh A. S, “URL phishing data analysis and detecting phishing attacks using machine learning in NLP”, International Journal of  Engineering Applied Sciences and Technology (IJEAST) Pg: 26-31, 2019 (Journal).
4. DOYEN SAHOO, CHENGHAO LIU, STEVEN C.H. HOI “Malicious URL Detection using MachineLearning: A Survey”, Pg: 1-37, 2019 (Article).
5. Hyunsang Choi, Bin B. Zhu, Heejo Lee “Detecting Malicious Web Links and Identifying Their Attack Types”, Pg: 1-12, (Article)
6. Harshal Tupsamudre, Ajeet Kumar Singh, and Sachin Lodha “Everything is in the Name − A URL based Approach for Phishing Detection”, research gate.net Pg: 1-19, 2019 (Article)

7. Immadisetti, Naga Venkata, Durga Naveen, Manamohana K, Rohit Verma "Detection of Malicious URLs using Machine Learning Techniques", International Journal of Innovative Technology and Exploring Engineering (IJITEE), Pg: 389-393, 2019(Journal).
8. Ammar Yahya Daeef, R. Badlishah Ahmad, Yasmin Yacob, Ng Yen Phing "Wide Scope and Fast Websites Phishing Detection Using URLs Lexical Features", International conference on electronic design, Pg: 410-416, 2016 (conference)
9. Dharmaraj R. Patil, Jayantrao B. Patil "Feature-based Malicious URL and Attack Type Detection Using Multi-class Classification", The ISC International Journal of Information security, PP: 141-162, 2018 (Journal).
10. Joby James, Sandhya. L, Ciza Thomas "Detection of phishing URLs using machine learning techniques", International conference on control communication and computing (ICCC), pp: 304-309, 2013 (conference).
11. Cho Do Xuan1, Hoa Dinh Nguyen, Tisenko Victor Nikolaevich "Malicious URL detection based on machine learning", International Journal of Advanced Computer Science and Applications (IJACSA) pp: 148-153, 2020 (Journal).
12. Shraddha Parekh, Dhwanil Parikh, Srushti Kotak, Prof. Smita Sankhe "A new method for Detection of Phishing Websites: URL Detection", International Conference on Inventive Communication and Computational Technologies (ICICCT 2018), pp: 549-552 (conference)
13. Vaibhav PatilPritesh Takkar, Chirag Shah, Tushar Bhat, S. P. Godse "Detection and Prevention of Phishing Websites using Machine Learning Approach", International Conference on Computing Communication Control and Automation (ICCUBEA), 2018 (conference).
14. https://www.kaggle.com