Detection and prevention of Phishing attacks based on classification models

Juwairiyyah
Department of Computer Science and Engineering
Chaitanya Bharathi Institute of technology
Hyderabad, India
ugs19129_cse.juwairiyyah@cbit.org.in

G.Kiran Kumar
Associate Professor
Department of Computer Science and Engineering
Chaitanya Bharathi Institute of technology
Hyderabad, India
ganipalli.kiran@gmail.com

***Abstract*— There has been a shift in the last few years that saw a rapid increase in eservices that need the transmission of sensitive and private information over the internet. Along with the number of people utilizing the internet, crime-attacks are expanding quickly. Malicious attacks using malicious URLs are becoming the simplest way to compromise the security chain's weakest link. Machine learning is an extremely well-liked method for identifying malicious or fake URLs. Tree based classification ML models may be utilized to detect non-benign URLs by training them by utilizing a dataset that contains common malicious URLs. This paper explains the lexical features in a URL that may be utilized to train the machine learning models. It also calculates the accuracy rate of different models and selects one for prediction based on the calculated accuracies.**

Keywords—Malicious attacks, non-benign URLs, Machine Learning, lexical features

I. Introduction

Phishing is one of the most common social engineering cyber-attacks which often targets at collecting sensitive data from internet users, such as personal credentials and bank account information, through posing as a reliable entity of online contacts. The online industry is the one that is most frequently targeted by phishing. Phishing is done by establishing a different URL, either by making a brand-new one or by slightly changing an existing valid one so that the internet user cannot realize the difference between the two as depicted in Fig. 1. Phishing attacks can occur in many different places, such as online payment systems, websites, bank institutions, and many more.

As of 2023, 5.16 billion people used the internet, representing 64.4 percent of the world's population. From 2020 to 2023, we also saw a huge increase in internet usage across all sectors of the economy. Internet usage has increased among people ever since they got habituated to it because of the pandemic. To put it in a different way, when internet usage rises as mentioned in [1], so does the frequency of cyberattacks. 67% of the data thefts were caused by malicious assaults, according to the Verizon Data Breach Investigations Report 2020 [2].

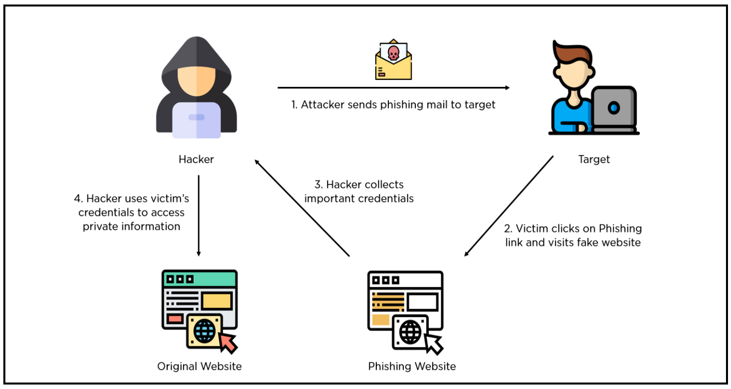

Hacker
1. Attacker sends phishing mail to target
Target
4. Hacker uses victim's credentials to access private information
3. Hacker collects important credentials
2. Victim clicks on Phishing link and visits fake website
Original Website
Phishing Website

Fig. 1. Working of a phishing attack using a fake URL.

There is legitimate cause for concern given the increased prevalence of phishing assaults among all cyber-crimes. As total of 182,465 phishing sites have been discovered in 2019 based on a report from the Anti-Phishing Working Group (APWG Q2 2019). The industries under this that are most directly targeted are webmail and SaaS.

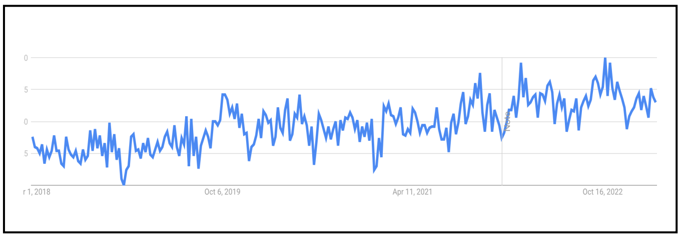


Fig.2. Phishing attacks research trend over 5 years

According to the Phishlab Phishing Report 2019, the targets of 84% of malicious attempts include financial, online data storage, shipping, and bank services. Fig.2 shows the growing interest for the said study over a period of five years. The most appealing market for phishing is the payments industry. Many studies have been conducted recently to identify phishing attack prevention and detection methods.

Machine learning techniques are more widely used and are effective than others for phishing attack detection. As phishing attempts get more sophisticated, machines must be able to anticipate, spot, and recognize them more quickly. Many techniques and models are available in machine learning to automate and quickly identify phishing attacks. Yet, hackers also improve their hacking techniques by using the same models. ML is very successful in predicting and analysing purposes. In this paper we describe the lexical features needed to train a classification based ML model for the identification of malicious URLs. We also exhibit our experiment of the detection strategies utilising different classification models of ML.

A. Machine Learning technique

In phishing detection, an incoming URL is analysed to determine whether it is phishing or not, and then it is categorised accordingly. To determine whether a particular URL is real or phishing, several machine learning algorithms are trained on diverse datasets of URL attributes. These Machine Learning models are classification models that classify the give input into two categories. This is done by using decision trees. The decision tree model is enhanced to make more effiecient models like Random Forest and XGBoost. All these algorithms or models are tree-based and give different accuracies for different datasets.

Machine learning algorithms use a calculation called Ginni Index (1). The Gini Index is a potent indicator of the randomness, impurity, or entropy in a dataset's values. It determines the likelihood that a certain characteristic will be erroneously classified if it is randomly chosen. The formula for calculation of Gini Index is shown below.

$$\begin{aligned} Gini\ Index &= 1 - \sum_{i=1}^{n} (P_i)^2 \\ &= 1 - [(P_+)^2 + (P_-)^2] \qquad (1) \end{aligned}$$

B. URL-Lexical Features

Lexical characteristics are retrieved out of raw URLs for the purpose of training a machine learning model and employed as the model's input characteristics. Some of which are extracted from the existing systems in [3]–[4]. The following 22 features are used:

- Presence of IP address: Typically, IP addresses are used in place of domain-names by online criminals for concealing the identity of a website.
- Existence of hostname: Identity is often included in the URL of a trustworthy site.
- Google indexing: Using this, it can be checked if Google-Search Console has indexed a particular URL.
- Number of dots: More than two sub-domains are typically included in the URL of a non-benign site. A dot(.) separates every domain. Any URL with more than three dot characters (.) raises the risk of a malicious website.
- Count of www: Most secure websites often just have one www in their URL. If the URL has no www or more than one www, this feature aids in the detection of fraudulent websites.
- Number of @: When an '@' is encountered everything prior to it is overlooked.
- Number of directories: Malicious URLs typically include many directories.
- Number of embedded domains: The count of embedded domains can aid in the identification of non-benign URLs. Presence of character "//" indicates this.
- Suspicious terms: Terms like PayPal, login, sign in, banks, accounts, updates, bonuses, services, ebayisapi, tokens, etc. are frequently seen in malicious Websites.
- URL shortening: The purpose of this feature is to show whether a URL has been shortened using a service, such as Rebrandly, BL.INK, etc.
- Number of https: Non-benign URLs typically avoid using these protocols because of their security.
- Count of HTTP: Often, benign URLs include a single HTTP.
- Number of %: Normal URL encoding substitutes the symbol (%) for spaces. Secure websites typically have less spaces in their URLs.
- Presence of (?): The string, which includes information that is provided to the server, is indicated by the symbol (?) in the URL.
- Number of Hyphens: To make a URL look legitimate, phishers put dashes (-) before or after the brand name.
- Count of (=): The equals sign (=) in a URL denotes that variables are being sent from one form page to another. As anyone may alter the values in a URL to change the page, it is regarded as being riskier.
- Length of URL: Attackers generally use URLs of greater length in order to not reveal the domain name.
- Length of hostname: This feature is also crucial to determine if the URL is malicious or not.
- First directory length: Counting the number of characters after the first '/' encountered gives this length.
- top-level domains: A TLD is, to put it simply, everything that comes after the last dot in a domain name.

- Number of digits: In general, URLs that contain numbers are considered suspicious.
- Number of letters: Attackers add more letters and numbers to the URL to lengthen it and conceal the domain name.

II. Related work

[5] is a research study that investigates various approaches for detecting phishing attacks. The paper begins with an overview of phishing attacks and their potential impact on individuals and organizations. The authors discuss the various techniques that phishers use to trick their victims, including social engineering, URL obfuscation, and the use of fake websites. They also discuss the different categories of phishing attacks, like spear phishing and whaling-attacks. Next, the authors review the existing approaches for detecting phishing attacks. They classify these approaches into the following types: signature-oriented, behaviour-oriented, and a combination technique. The authors then suggest a technique to identify phishing attacks which uses BayesNet which is also discussed in [6] and [7]. Their approach is based on analysing the features of phishing websites and identifying patterns that are common to phishing attacks. They use utilize the dataset containing malicious websites and authentic sites for training their model as well as evaluate its performance. The results of their experiments suggest that this approach demonstrates superior accuracy and false positive percentage compared to current techniques.

[8] is a research study that suggest an innovative technique to identify malicious attacks. This paper begins with an introduction to phishing attacks, their impact on users, and the various techniques used by phishers to deceive their victims. The authors then review the existing approaches for detecting phishing attacks, including signature, heuristic and ML based methods. The authors suggest a novel approach to identify malicious attacks which uses ML algorithms to analyse the features of phishing websites. They extract lexical-features such as URL length, age of domain, and the existence of particular key-words, and use these to train the model. Then the model is utilized for classifying new websites as malicious or authentic. Malicious websites are illegitimate and the output shows the presence of malicious content in the URL provided. Authentic URLs are legitimate sites that already exist with proper identity and are shown similarly. The authors evaluate this technique by utilizing a dataset containing non-benign and benign websites and compare its performance with other algorithms.

In [9], the authors provide a comprehensive review of the present state-of-the-art techniques for malicious attack identification and propose new approach that uses machine learning algorithms. The paper begins with an introduction to phishing attacks. They propose a ML-based technique to identify malicious-attacks that utilizes three different algorithms: SVM, Random Forest, and Neural Network with Backpropagation The output of their proceedings reflect that RF and SVM outperform the Neural Network with Backpropagation regarding precision , accuracy and recall. [10] has a similar approach where Support Vector Machine and Random Forest machine learning models are talked about. The UCI Machine Learning Repository is where the datasets are obtained. The initial dataset has 30 lexical features, whereas the next dataset has 10 lexical features. 95.11% accuracy was obtained when the RF technique was used. The Support vector machine model showd an accuracy of 92.6%. Similarly The Random Forest Classifier was used by [11] to identify phishing webpages. In their research, they discovered 26 feature combinations that together produced a training dataset with an accuracy of 98.8%. They only used the Random Forest Classifier algorithm. They made no mention of whether or not using a different model would result in a greater level of accuracy for these 26 features.

[12] discusses Artificial Neural Network and Deep Neural Network, two classification techniques. There are 27 features in all, including the URLs length and count of subdomains. The original dataset was divided in 10 sects, and cross validation was performed ten times. A dataset was utilised for testing, whereas nine datasets were used for training. There was just one hidden layer structure utilised in artificial neural networks. This categorization system has a 91% accuracy rate. There were two hidden layers in the Deep Neural Network. Tensorflow was utilised to create this classification technique. This approach yields a 96% accuracy rate.

[13] described a project that explored surface-level URL features for developing a strong-weighted ML model. The goal is to prevent the vulnerability of extracting host-based information by limiting the source of potential characteristics to the URL's character-string. Each URL is displayed as a binary characteristic vector. These are the vectors that are fed into the online model, which, at the time of testing, maps the binary feature vector to URLs that haven't yet been found. The model goes down the vector and produces the final result, which may or may not be good. In contrast they described a method according lexical and host-specific traits in [14]. The suggested model demonstrated an accuracy ranging from 93% to 98% and detected a significant amount of servers that were engaged in phishing.

[15]-[16] proposes an approach for identifying and classifying phishing URLs by employing ML algorithms. This proposed approach involves the extraction of various features from the particular URL, like the URL’s length and the use of unique characters. These features are then fed into different machine learning techniques, including RF, SVM and logistic regression, for classifying URLs as either authentic or phishing. The features are calculated based on previous work done in papers and identifying the most salient characteristics which can efficiently help in training these Machine Learning models. The paper concludes that ML techniques can be effectively utilized for identifying phishing URLs and the fact that feature engineering is crucial for achieving accurate classification results.

III. Proposed approach

Recent years have seen a sharp rise in cybersecurity attacks on many websites around the world, including ransomware, phishing, malware injection, etc. As a result, numerous financial institutions, e-commerce businesses, and people suffered significant financial losses. Since new attack types are being developed daily, controlling a cyber-security attack in such a situation is a significant problem for cyber security specialists.

This approach is aimed at identifying non-benign URLs by utilizing the features extracted through decision tree based models . We tackle the multi-class classification problem of malicious URL detection. In this paper, we classify the unprocessed URLs into a number of groups, such as safe or innocuous, malware, phishing, and defacement. We generate lexical numeric features from input URLs since machine learning techniques only accept numeric inputs. Therefore, the numeric-lexical features will be the input to ML models instead of actual raw URLs.

Four common ML model classifiers: Extreme Gradient Boosting, Random Forest, Lite Gradient-Boosting machine, Logistic Regression and Random Forest, as suggested in [17]-[18] and Logistic Regression after referring to [19] are utilised. Eventually, in order to determine which features are crucial for predicting dangerous URLs, additionally we will assess how they performed and display the average feature-importance.

A. *Dataset Description*

We use a dataset from Kaggle shown in Table 1. Among the 6,51,191 URLs in the Malicious URLs dataset, 4,28,103 are benign or non-malicious URLs, 96,457 are defacement URLs, 94,111 are phishing URLs, and 32,520 are malware URLs.

TABLE I. MALICIOUS-URLS-DATASET

	URL	*TYPE*
0	br-icloud.com.br	phishing
1	bopsecrets.org/rexroth/cr/1.html	benign
2	mp3raid.com/music/krizz_kaliko.html	benign
3	http://www.garage-pirenne.be/index.php?	defacement
4	http://adventure-nicaragua.net/index.php	defacement

B. Model Building and Evaluation

The Lexical features are utilised as input to train the classification models. Exploratory analysis on different features is done as displayed in Fig. 3. It can be seen how the feature ip address of the URL is only relevant to the Malware URLs. All the other URLs do not contain IP address. In this way an EDA graph of every lexical feature can be plotted to show the Exploratory analysis done on them.

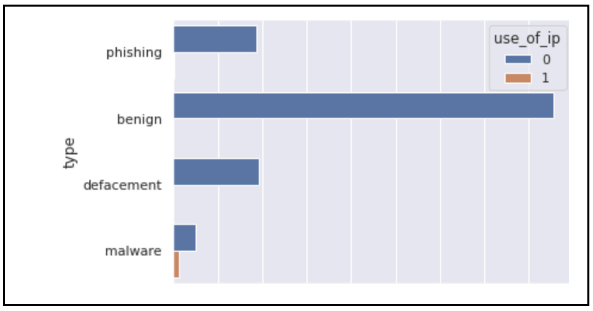
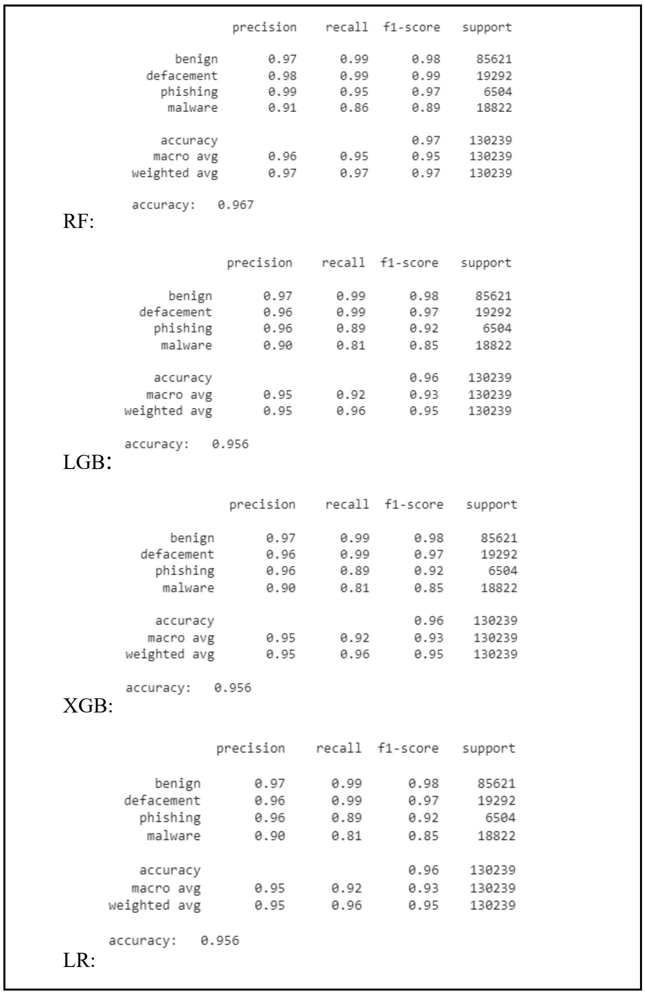


Fig.3. EDA graph.

A split of 80:20 was utilised to divide the dataset, in which the data accounting to 80% was utilized to train the classification algorithms and rest being utilized to test these models. There is evident asymmetry in the dataset.



```
              precision    recall  f1-score   support

      benign       0.97      0.99      0.98     85621
  defacement       0.98      0.99      0.99     19292
    phishing       0.99      0.95      0.97      6504
     malware       0.91      0.86      0.89     18822

    accuracy                           0.97    130239
   macro avg       0.96      0.95      0.95    130239
weighted avg       0.97      0.97      0.97    130239

accuracy:   0.967
```

RF:

```
              precision    recall  f1-score   support

      benign       0.97      0.99      0.98     85621
  defacement       0.96      0.99      0.97     19292
    phishing       0.96      0.89      0.92      6504
     malware       0.90      0.81      0.85     18822

    accuracy                           0.96    130239
   macro avg       0.95      0.92      0.93    130239
weighted avg       0.95      0.96      0.95    130239

accuracy:   0.956
```

LGB:

```
              precision    recall  f1-score   support

      benign       0.97      0.99      0.98     85621
  defacement       0.96      0.99      0.97     19292
    phishing       0.96      0.89      0.92      6504
     malware       0.90      0.81      0.85     18822

    accuracy                           0.96    130239
   macro avg       0.95      0.92      0.93    130239
weighted avg       0.95      0.96      0.95    130239

accuracy:   0.956
```

XGB:

```
              precision    recall  f1-score   support

      benign       0.97      0.99      0.98     85621
  defacement       0.96      0.99      0.97     19292
    phishing       0.96      0.89      0.92      6504
     malware       0.90      0.81      0.85     18822

    accuracy                           0.96    130239
   macro avg       0.95      0.92      0.93    130239
weighted avg       0.95      0.96      0.95    130239

accuracy:   0.956
```

LR:

Fig.4. Accuracies of different classification models used

The data contains benign URLs in about 66% of cases, malware in 5%, phishing in 14%, and defacement URLs in 15% of cases. So it's feasible that the distribution of different types changed after the dataset was arbitrarily divided into train-set and test-set, that would have a major effect on the performance of the classification model. Thus, stratification is required to keep the desired variable's proportion the same.

The split created by this stratify parameter ensures that there is equality between the proportion of rates created in the sample and the proportion of rates supplied to the stratify parameter. We build four classification models i.e.,Random forest, Light Gradient-Boosting Machine , Extreme Gradient Boosting and Logistic Regression. We make predictions on the test set after fitting the model. Fig. 4 shows a performance comparison between Random Forest, Lite GBM, XGBoost, and Logistic Regression. The findings above show that Random Forest performs best when it comes to test accuracy, reaching the highest accuracy of 96.7% and having a greater identification percentage for malware, phishing, and innocuous content.

A confusion matrix for each model is also plotted to evaluate the performance of the ML model. It contrasts projected values with the actual goal values by the ML model. This model's confusion matrix uses a RF algorithm is depicted in Fig. 5.

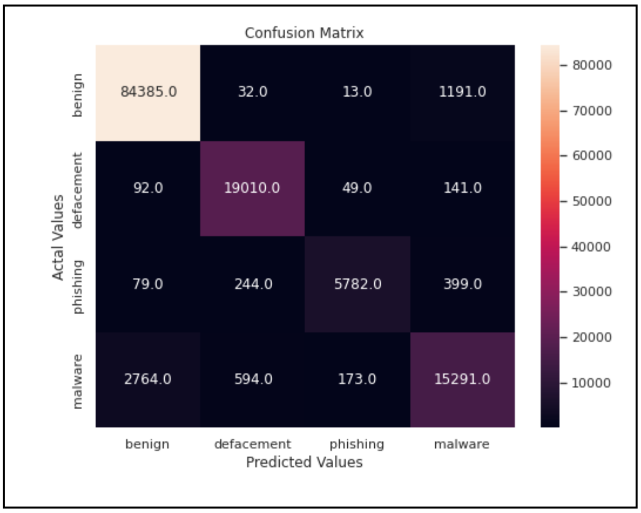


Fig.5. Confusion matrix of Random Forest model

The highlighted diagonal squares show the true positives found for each URL type. The highest being 84385 for benign URLs. After selecting the model i.e., Random Forest, due to its highest accuracy finding, we look for the highly contributing features. This is done by plotting a feature importance graph as shown in Fig.6. The techniques for rating each input feature for a particular model are referred to as "feature importance"; The ratings only indicate the significance of each feature. A higher rating means the specific characteristic will have a bigger affect on the forecasting model for that specific variable. From the plot shown it can be seen how five lexical features, count of www, count of directories, hostname size, tld length, and count-http are the most significant for identifying non-benign URLs by visualizing the feature importance of RF model.

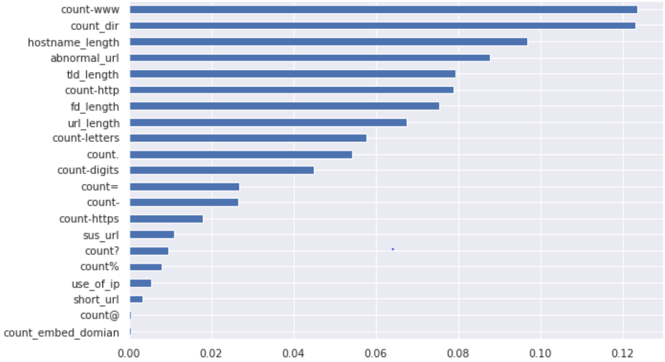


Fig.6. Feature importance graph for RF

IV. CONCLUSION

In this paper, we have shown how to identify malicious URLs using machine learning. Using raw URLs, 22 lexical characteristics were extracted and trained the Random Forest, Light Gradient-Boosting Machine, Extreme Gradient Boosting and Logistic Regression classification-based learning models. In addition, we evaluated the effectiveness of the three classification models and found that Random Forest model has exceeded in terms of performance with the highest accuracy of 96.6% as shown in Fig.7. Finally, RF model, the prediction algorithm for categorizing any unprocessed URL, has been coded.

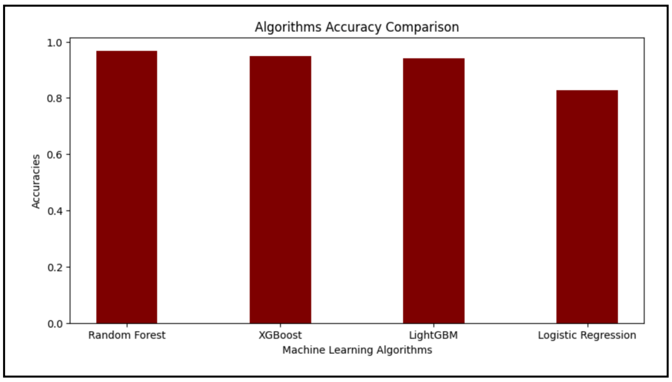


Fig.7. Algorithm Accuracy Comparison graph

REFERENCES

[1] Alsharnouby, Mohamed, F. Alaca, and S. Chiasson. ”Why phishing still works: User strategies for combating phishing attacks.” International Journal of Human-Computer Studies 82 (2015): 69-82.

[2] Ali, Waleed. “Phishing website detection based on supervised machine learning with wrapper features selection.” International Journal of Advanced Computer Science and Applications 8.9 (2017): 72-78.B. Rieder, *Engines of Order: A Mechanology of Algorithmic Techniques*. Amsterdam, Netherlands: Amsterdam Univ. Press, 2020.

[3] Jain,AnkitKumar,andBrijB.Gupta,"Towardsdetectionofphishing websites on client-side using machine learning based approach." Telecommunication Systems 68.4 (2018): 687-700.

[4] Jain, Ankit Kumar, and Brij B. Gupta, "A machine learning based approach for phishing detection using hyperlinks information." Journal of Ambient Intelligence and Humanized Computing 10.5(2019): 2015-2028.

[5] M. Baykara and Z. Z. Gürel, "Detection of phishing attacks," *2018 6th International Symposium on Digital Forensic and Security (ISDFS)*, Antalya, Turkey, 2018, pp. 1-5, doi: 10.1109/ISDFS.2018.8355389.

[6] V. Muralidharan, and V. Sugumaran. "A comparative study of Naïve Bayes classifier and Bayes net classifier for fault diagnosis of monoblock centrifugal pump using wavelet analysis." Applied Soft Computing 10.1016/j.asoc.2012.03.021.

[7] Bouckaert, Remco R. "Bayesian network classifiers in weka." Hamilton: Department of Computer Science, University of Waikato, 2007.

[8] A. A.A. and P. K., "Towards the Detection of Phishing Attacks," 2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184), Tirunelveli, India, 2020, pp. 337-343, doi: 10.1109/ICOEI48184.2020.9142967.

[9] S. Sindhu, S. P. Patil, A. Sreevalsan, F. Rahman and M. S. A. N., "Phishing Detection using Random Forest, SVM and Neural Network with Backpropagation," 2020 International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE), Bengaluru, India, 2020, pp. 391-394, doi: 10.1109/ICSTCEE49637.2020.9277256.

[10] S. Jagadeesan, Anchit Chaturvedi, Shashank Kumar, "URL Phishing Analysis using Random Forest", International Journal of Pure and Applied Mathematics, 2018

[11] A. Alswailem, B. Alabdullah, N. Alrumayh and A. Alsedrani, "Detecting Phishing Websites Using Machine Learning," 2019 2nd International Conference on Computer Applications & Information Security (ICCAIS), Riyadh, Saudi Arabia, 2019, pp. 1-6, doi: 10.1109/CAIS.2019.8769571.

[12] Ozgur Koray Sahingoz, Saide Işilay Baykal and Deniz Bulut, " Phishing Detection from URLs by using Neural Networks", International Conference on Computer Science engineering and Applications, 2018

[13] Aaron Blum, Brad Wardman, Thamar Solorio, Gary Warner "Lexical Feature Based Phishing URL Detection Using Online Learning", Proceedings of the 3rd ACM Workshop on Security and Artificial Intelligence, AISec 2010, Chicago, Illinois, USA, October 8, 2010

[14] Feroz, M. Nazim, and S. Mengel. "Phishing URL detection using URL ranking." In 2015 ieee international congress on big data, pp. 635-638. IEEE, 2015.

[15] F. Vanhoenshoven, G. Nápoles, R. Falcon, K. Vanhoof and M. Köppen, "Detecting malicious URLs using machine learning techniques," 2016 IEEE Symposium Series on Computational Intelligence (SSCI), Athens, Greece, 2016, pp. 1-8, doi: 10.1109/SSCI.2016.7850079.

[16] A. Crişan, G. Florea, L. Halasz, C. Lemnaru and C. Oprisa, "Detecting Malicious URLs Based on Machine Learning Algorithms and Word Embeddings," 2020 IEEE 16th International Conference on Intelligent Computer Communication and Processing (ICCP), Cluj-Napoca, Romania, 2020, pp. 187-193, doi: 10.1109/ICCP51029.2020.9266139.

[17] Chen, Tianqi, C. Guestrin. "Xgboost: A scalable tree boosting system." arXiv:1603.02754 [cs.LG], pp. 785-794. 2016

[18] Tianqi Chen, Tong He, M. Benesty, V. Khotilovich, Y. Tang, and H. Cho. "Xgboost: extreme gradient boosting." R package version 0.4-2 1, no. 4 (2015): 1-4.

[19] R. Chiramdasu, G. Srivastava, S. Bhattacharya, P. K. Reddy and T. Reddy Gadekallu, "Malicious URL Detection using Logistic Regression," 2021 IEEE International Conference on Omni-Layer Intelligent Systems (COINS), Barcelona, Spain, 2021, pp. 1-6, doi: 10.1109/COINS51742.2021.9524269.