Journal of Critical Reviews

ISSN- 2394-5125 Vol 7, Issue 8, 2020

DETECTION OF PHISHING WEBSITES USING DEEP LEARNING AND MACHINE LEARNING

T. Sujithra[1], Naveen Dwivedi[2], Anuwaya Utakarsha[3]

[1]Assistant Professor, [2,3]Undergraduate Student
[1,2,3]Computer Science and Engineering, SRM Institute of Science and Technology ,
Kattankulathur Tamilnadu-603203, India
E-mail: [1]sujithrt@gmail.com, [2]dwivedi.naveen272208@gmail.com, [3]anuwayautkarsh@gmail.com

Received: 08.05.2020 *Revised: 06.06.2020* *Accepted: 30.06.2020*

Abstract
Phishing can be described as a way by which someone may try to steal some personal and important information. By appearing as a trusted body. Many websites, which look perfectly legitimate to us, can be phishing and could well be the reason for various online frauds. These phishing websites may try to obtain our important information through many ways, for example: phone, calls, messages, and pop up windows. When a user opens a fake webpage and enters the username and protected password, the credentials of the user are acquired by the attacker which can be used for malicious purposes . Phishing websites look very similar in appearance to their corresponding legitimate websites to attract large number of Internet users.

Keywords--- machine learning(ML) , phishing , deep learning(DL), Algorithms , Support vector machine(SV)

DOI: http://dx.doi.org/10.31838/jcr.07.08.215

INTRODUCTION

In this cyber world, most of the people communicate with each other either through a computer or a digital device connected over the Internet. The number of people using e-banking, online shopping and other online services has been increasing due to the availability of convenience, comfort, and assistance. An attacker takes this situation as an opportunity to gain money or fame and steals sensitive information needed to access the online service websites.

Phishing is one of the ways to steal sensitive information from the users. It is carried out with a mimicked page of a legitimate site, directing online user into providing sensitive information. The term phishing is derived from the concept of 'fishing' for victims sensitive information.

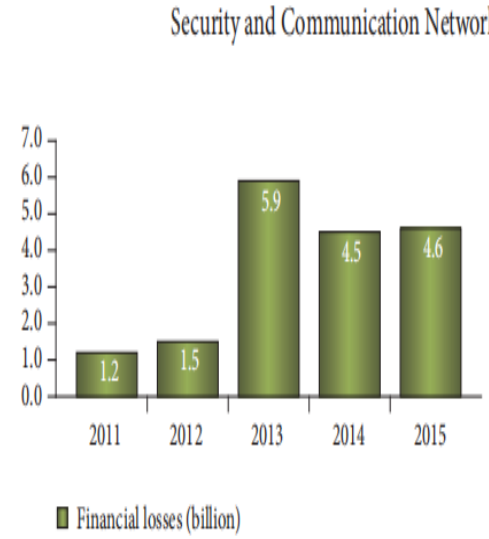

Figure 1. worldwide financial losses(in billion)due to phising attack

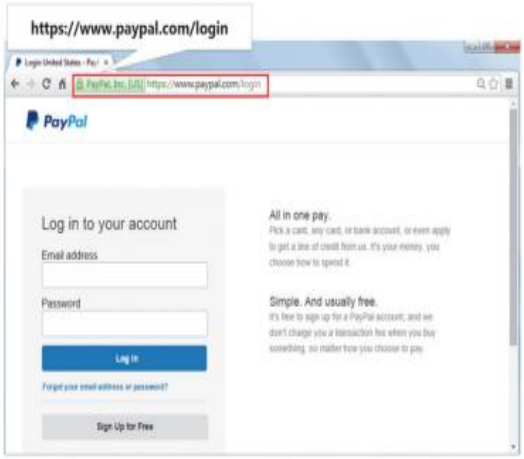

Figure 2. Legitimate Paypal Webpage

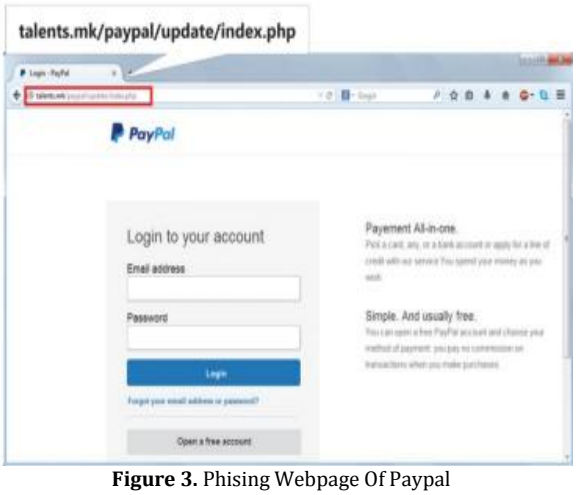

Figure 3. Phising Webpage Of Paypal

LITERATURE SURVEY

[1] This paper explains a novel method to find phishing websites using machine learning algorithms. Comparison of accuracy of five machine learning algorithms Decision Tree (DT), Random Forest (RF), Gradient Boosting (GBM), Generalized Linear Model (GLM) and Generalized Additive Model (GAM), Accuracy, Precision and Recall evaluation approach were calculated for each algorithm and dealt according to need. With the help of Python and open source programming language RWebsite attributes (30) are found. Algorithms namely Decision Tree, Random Forest and GBM performance were compared in table. Tables of accuracy, recall and performance, it is shown that Random Forest algorithm has given highest 98.4% accuracy, 98.59% recall and 97.70% precision.

[2] This paper suggest a classification mode in order to distinguish the phishing attacks. This model consist of feature extraction from sites and classification of website. In feature extraction, 30 features has been taken and phishing feature extraction rules has been clearly outlined. for classification of these features, Support Vector Machine (SVM), Naïve Bayes (NB) and Extreme Learning Machine (ELM)were used. In Extreme Learning Machine (ELM), six activation functions were used‘ for process and achieved 95.34% accuracy than SVM and NB. The results were obtained with the help of MATLAB.

[3] Authors presents a way to find phishing email attacks using natural language processing and machine learning. This is used to run the semantic analysis of the text to find malicious intent. A natural Language Processing (NLP) technique is used to parse each sentence and finds the semantic jobs of words in the sentence in connection to the predicate. In light of the job of each word in the sentence, this strategy find whether the sentence is an inquiry or an order. Supervised machine learning is used to generate the blacklist of malicious pairs.

Authors made algorithm SEAHound for detecting phishing emails and Netcraft Anti-Phishing Toolbar is used to check the validity of a URL. This algorithm is implemented with Python scripts and dataset Nazario phishing email set is used. Results of Netcraft and SEAHound are compared and obtained precision 98% and 95% respectively.

[4] Another method by authors proposes feature selection algorithms to decrease the components of dataset to get higher order execution . It also compared with other data mining classification algorithms and results obtained. Dataset for phishing websites was taken from UCI machine learning repository.

From the result, it is seen that some classification strategies increment the execution; some of them decline the execution with decreased component. Bayesian Network, Stochastic Gradient Descent (SGD), Lazy.K.Star, Randomizable Filtered Classifier, Logistic model tree (LMT) and ID3 (Iterative Dichotomiser) are useful for reduce phishing dataset and Multilayer Perception, JRip, PART, J48, Random Forest and Random Tree algorithms are not valuable for the diminished phishing dataset. Lazy.K.Star obtained 97.58% accuracy with 27 reduced features. This study is obtained with the help of WEKA software.

[5] Proposed model with answer for recognize phishing sites by using URL identification strategy utilizing Random Forest algorithm. Show has three stages, namely Parsing, Heuristic Classification of data, Performance Analysis Parsing is used to analyze feature set. Dataset found from Phishtank. Out of 31 features only 8 features are taken for parsing. Random forest method obtained accuracy level of 95%.

[6] It is proposed a flexible filtering decision module to extract features automatically without any particular expert knowledge of the URL domain using neural network model.

In this method authors used all the characters included in the URL strings and count byte values. They not only count byte values and also overlap parts of neighbouring characters by shifting 4-bits. They embed combination information of two characters appearing simultaneously and counts how many times each value appears in the original URL string and achieves a 512 dimension vector. Neural network model tested with three optimizers Adam, AdaDelta and SGD. Adam was the best optimizer with accuracy 94.18% than others. Authors also conclude that this model accuracy is higher than the previously stated complex neural network topology.

[7] Alicious URL with classical machine learning technique – logistic regression using bigram, deep learning techniques like convolution neural network (CNN) and CNN long short-term memory (CNN-LSTM) as architecture. The dataset received from Phishtank, OpenPhish for phishing URLs and dataset Malware Domainlist, Malware Domains were collected for malicious URLs. As a result of comparison, CNN-LSTM obtained 98% accuracy. In this paper authors used TensorFlow in conjuction with Keras for deep learning architecture.

[8] Authors in this paper also proposed reduced feature selection model to detect phishing websites. They used Logistic Regression and Support Vector Machine (SVM)as classification methods to validate the feature selection method. 19 features reduced from 30 site features have been selected and used for phishing detection. The LR and SVM calculations performance was surveyed dependent on precision, recall, f-measure and accuracy. Study shows that SVM algorithm achieved best performance over LR algorithm.

[9] In this paper authors proposed a phishing detection model to find the phishing performance effectively by using mining the semantic features of word embedding, semantic feature and multi-scale statistical features in Chinese web pages.

Eleven features were extracted and categorized into five classes to acquire statistical features of web pages. AdaBoost, Bagging, Random Forest and SMO are used to implement learning and testing the model. Legitimate URLs Dataset received from Direct Industry web guides and phishing data was received from Anti-Phishing Alliance of China. According to study, only semantic features well identified the phishing sites with high detection efficiency and fusion model achieved the best detection performance. This model is unique to Chinese web pages and it has dependency in certain language.

[10] This paper states an efficient method to check phishing URL websites by using c4.5 decision tree approach. This technique extracts features from the sites and calculates heuristic values. These values were given to the c4.5 decision tree algorithm to determine whether the site is phishing or not. Dataset is collected from PhishTank and Google. This process includes two phases namely pre-processing phase and detection phase. In which features are extracted based on rules in pre-processing phase and the features and their respected values were given to the c4.5 algorithm and obtained 89.40% accuracy

[11] Authors in this paper created an extension to Google Chrome to detect phishing websites content with the help of machine learning algorithms. Dataset UCI-Machine Learning Repository used and 22 features were extracted for this dataset.

Algorithms kNN, SVM and Random Forest were chosen for precision, recall,f1-score and accuracy comparison. Random Forest obtained a best score and HTML, JavaScript, CSS used for

implementing chrome extension along with python. This extension is having a drawback of declared malicious site list which is surging up every day.

[12] This paper takes a framework to extract features flexible and simple with new strategies. Data is collected from PhishTank and legitimate URLs from Google. To receive the text properties C# programming and R programming were used. 133 features were obtained from the dataset and third party service providers. CFS subset based and Consistency subset based feature selection methods used for feature selection and analyzed with WEKA tool. Naïve Bayes and Sequential Minimal Optimization (SMO) algorithms were compared for performance evaluation and SMO is preferred by the author for phishing detection than NB.

[13] Another heuristic features detection method by authors explains about the feature of URL such as PrimaryDomain, SubDomain, PathDomain and ranking of website such as PageRank, AlexaRank, AlexReputation to identify the phishing websites. Dataset used from PhishTank and experimental is splitted into 6 phases through MYSQL, PHP with 10 testing datasets.

The stated model contains two phases. In Phase I site features were found and in Phase II six values of heuristic are calculated. According to authors, if heuristic value is nearest to one, the site is regarded as legitimate and if it is nearest to zero then the site is doubted as phishing site. Root Mean Square Error (RMSE) is used to calculate accuracy and obtained 97% accuracy.

[14] In this paper author introduces a phishing URL detection system depends on URL lexical analysis named PhishScore. This approach is based on intra-URL relatedness. This relatedness reflects the relationship into part of the URL Right around 12 site highlights removed from a solitary URL are utilized to include machine learning algorithms to identify phishing URLs. This experiment results accuracy of 94.91%.

[15] The focus on detecting phishing website URLs with domain name features. Web spoofing attack categories content-based, heuristic-based and blacklist-based approaches are said and the proposed model PhishChecker is developed with the help of Microsoft Visual Studio Express 2013 and C# language Dataset used from Phishtank and Yahoo directory set and obtained an accuracy of 96%. This paper checks only the validity of URLs.

[16] Phishing is a an important security threat to the Internet; it is an electronic online identity theft in which the attackers use spoofing techniques like fake websites that mimic legal websites to trick users into giving their private information.

Many of successful phishing attacks do exist and subsequently a considerable number of anti-phishing methods have been given out. However, they vary in terms of their accuracy and error rate. The author proposed an algorithm for phishing websites detection using data mining classification model. It is executed and experimented using a dataset of 20 different webpage features and 1,000 instances. The experimental results showed that the stated algorithm outperforms the original one in terms of the number of classification rules, accuracy (87%) and less error rate (0.1 %).

[17] With the ever surging use of Internet by different stake holders in various fields, information on web browsers and servers is highly susceptible to different security attacks. Though high security measures and enhanced techniques are used to protect the information on the web browsers and servers, they are still prone to a number of attacks. Phishing is one such type of attack in which users are tricked by the phishers using social engineering methods to steal their personal or confidential information. Detection of phishing attack with high accuracy is a challenging research issue. Users are duped by the phishers to enter their confidential information into websites created by them and thereby are steal the vital user's credentials. Phishing sites are normally detected by using blacklist based approach but this approach fails as white listed phishing sites cannot be detected using this approach.

This research work aims to use data mining algorithms to analyze E-mails and also helps in protection from phishing attacks. The author stated an architectural model to differentiate between the fake E-mail and real E-mail with a high accuracy and use naive Bayesian classification for the said purpose. The proposed algorithm works in various stages for fake E-mail detection and hence tries to protect the users from leaking their confidential information.

[18] Phishing attacks are common online, which have resulted in financial losses through using either malware or social engineering. Thus, phishing email detection with high accuracy has been an issue of great interest. Machine learning-based detection methods, particularly Support Vector Machine (SVM), have been proved to be effective.

However, the parameters of kernel method, whose default is that class numbers reciprocals in general, affect the classification accuracy of SVM. In order to improve the classification accuracy, this paper states a model, called Cuckoo Search SVM(CS-SVM). The CS-SVM extracts 23 features, which are used to construct the hybrid classifier. In the hybrid classifier, Cuckoo Search (CS) is integrated with SVM to optimize parameter selection of Radial Basis Function(RBF). Experiments are performed on a dataset consisting of 1,384 phishing emails and 20,071 non-phishing emails.

[19] Phishing attacks are common online, which have resulted in financial losses through using either malware or social engineering. Thus, phishing email detection with high accuracy has been an issue of great concern. Machine learning-based detection methods, particularly Support Vector Machine (SVM), have been proved to be effective.

However, the parameters of kernel method, whose default is that class numbers reciprocals in general, affect the classification accuracy of SVM. In order to improve the classification accuracy, this paper proposes a model, called Cuckoo Search SVM(CS-SVM). The CS-SVM extracts 23 features, which are used to construct the hybrid classifier. In the hybrid classifier, Cuckoo Search (CS) is integrated with SVM to optimize parameter selection of Radial Basis Function(RBF). Experiments are performed on a dataset consisting of 1,384 phishing emails and 20,071 non-phishing emails.

[20] Social engineering has emerged as a serious hurdle in virtual communities and is an important means to attack information systems. The services used by today's knowledge workers prepare the base for complicated social engineering attacks. Phishing is a kind of technically generated social engineering attack and is the type of identity theft that uses the social engineering techniques and complex attack vectors to harvest financial information from unsuspecting consumers.

It is a kind of attack in which phishers use spoofed emails and fraudulent web sites to trick people into giving personal information. Victims perceive these emails as trusted, while in reality they are the work of phishers interested in identity theft. Therefore, there is an urgent need for anti-phishing solutions and hereabout have been identified, a number of solutions to mitigate phishing attacks have been suggested. In this work, a phishing detection method is proposed by using machine learning and data mining techniques. Success rate of %89 has been achieved against phishing attacks coming from email messages. [6]

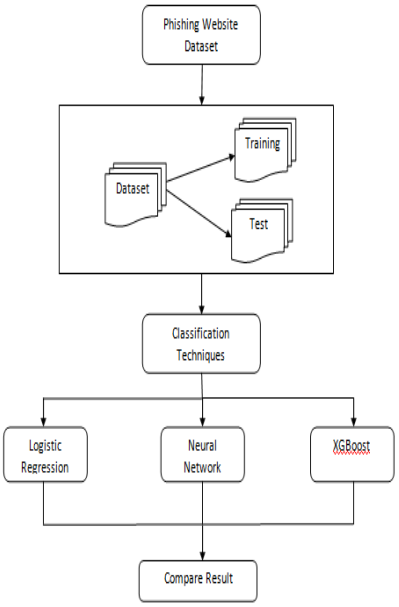

Figure 4. Methodology

PROPOSED WORK

We are using Machine learning algorithms to reduce the false positives in detecting new phishing sites. We are also making an attempt to identify the best machine learning algorithm to detect phishing sites with high accuracy than the existing techniques. We are Using machine learning algorithms (Logistic regression (LR), Neural Network and XG Boost to classify the websites as legitimate and phishing. Based on the experimental observations, XG Boost Outperformed the others. The choice of considering these machine learning algorithms is based on the classifiers used in the recent literature.

Figure 5. Interface Of Model

ALGORITHM

Logistic Regression

In simple, linear regression, predict scores on one variable from the scores on a second variable. The variable that predicted is called the criterion variable and is referred to as Y. The variable base for predictions on is called the predictor variable and is referred to as X. When there is only one predictor variable, the prediction method is called simple regression. In simple linear regression, the topic of this section, the predictions of Y when plotted as a function of X form a straight line.

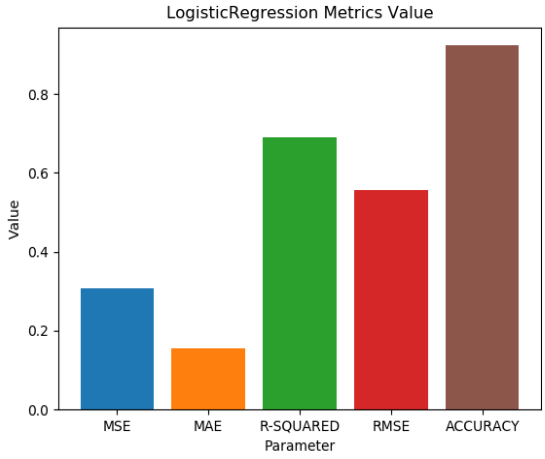

Figure 6. Logistic regression Metrics Value

Neural Network

We used the deep learning model from keras neural networks. It is a form of machine learning, which mimics same learning process as human brain. An artificial neural network has three layers such as input, hidden, and output layers. This requires high data and processing power. The method of learning can be supervised, semi-supervised or unsupervised

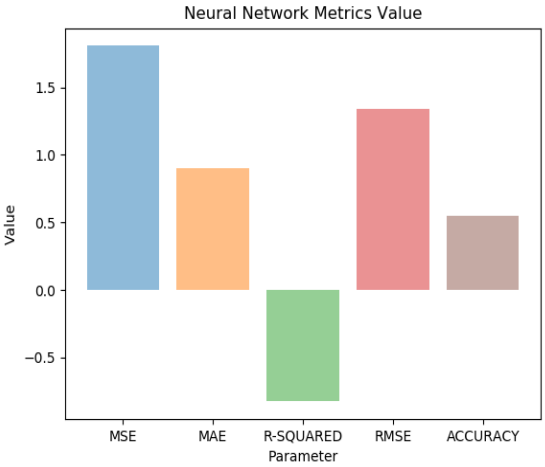

Figure 7. Neural Network Metrics Value Table

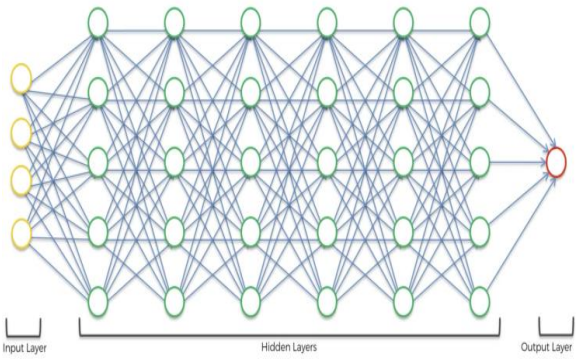

Figure 8. Neural Network Figure

XGBOOST model

XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. Rather than training all of the models in isolation of one another, boosting trains models in succession, with each new model being trained

to correct the errors made by the previous ones. Models are added sequentially until no further improvements can be made.

The advantage of this iterative approach is that the new models being added are focused on correcting the mistakes which were caused by other mode.

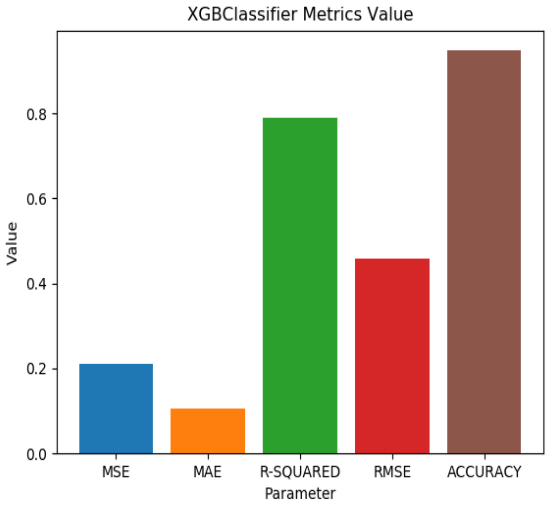

Figure 9. XGB classifier Metrics Value

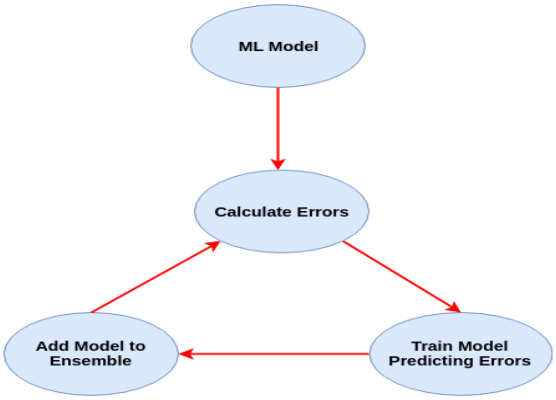

Figure 10. Evaluation model

EVALUATION METHOD

Researchers and developers calculate Precision, Recall, accuracy. These are the standard metrics to judge any phishing detection system.

Accuracy is defined as the ratio of correct predictions to the total predictions (both correct and incorrect).

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

Precision can be defined as the ratio of correctly predicted positive observations to the total predicted positive observations.

$$Precision = \frac{TP}{TP+FP}$$

Recall is the ratio of correctly predicted positive observations to the all observations in actual class.

$$Recall = \frac{TP}{TP+FN}$$

Where TP is number of cases which were positive and were also predicted positive, TN is number of cases which were negative and were also predicted negative, FP is number of cases which were negative but predicted positive and FN is the number of cases which were positive but predicted negative.

EXPERIMENTAL RESULTS (ATTRIBUTES OF URL)

There are 30 attributes of a website that were considered for detection purpose. Out of those 30 attributes only 8 are chosen in detection of phising data.

A. *has_ip*
B. *long_url*
C. *short_service*
D. *has_at*
E. *slash_redirect*
F. *pref_suf*
G. *has_sub_domain*
H. *ssl_state*
I. *long_domain*

RESULT DISCUSSION

Implemented three learning algorithm on the given dataset for phishing website detection shows that XG Boost model outperforms other models. The accuracy of XG boost is high compared to other machine learning algorithms.

Algorithm	Accuracy
Logistic Regression	92.29
Neutral Network	54.81
XG Boost	94.7

Figure 11. Accuracy Table Of Algorithms

CONCLUSION

Phishing is a cyber crime procedure utilizing both social building and specialized deception to take individual sensitive data. Besides, Phishing is considered as another extensive type of fraud. Experimentations against recent dependable phishing data sets utilizing different classification algorithm have been performed which received different learning methods. The base of the experiments is accuracy measure. The aim of this research work is to predict whether a given URL is phishing website or not. It turns out in the given experiment that XG Boost based classifiers are the best classifier with great classification accuracy of 94% for the given dataset of phishing site. As a future work we might use this model to other Phishing dataset with larger size then now and then testing the performance of those classification algorithm's in terms of classification accuracy.

REFERENCES

1. J. Shad and S. Sharma, "A Novel Machine Learning Approach to Detect Phishing Websites Jaypee Institute of Information Technology," pp. 425–430, 2018.
2. Y. Sönmez, T. Tuncer, H. Gökal, and E. Avci, "Phishing web sites features classification based on extreme learning machine," 6th Int. Symp. Digit.Forensic Secur. ISDFS 2018 - Proceeding, vol. 2018–Janua, pp. 1–5, 2018.
3. T. Peng, I. Harris, and Y. Sawa, "Detecting Phishing Attacks Using Natural Language Processing and Machine Learning," Proc. - 12th IEEE Int. Conf. Semant.Comput. ICSC 2018, vol. 2018–Janua, pp. 300–301, 2018.
4. M. Karabatak and T. Mustafa, "Performance comparison of classifiers on reduced phishing website dataset," 6th Int. Symp.Digit.Forensic Secur. ISDFS 2018 - Proceeding, vol. 2018–Janua, pp. 1–5, 2018.
5. S. Parekh, D. Parikh, S. Kotak, and P. S. Sankhe, "A New Method for Detection of Phishing Websites: URL Detection," in 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), 2018, vol. 0, no.Icicct, pp. 949–952.

6. K. Shima et al., "Classification of URL bitstreams using bag of bytes," in 2018 21st Conference on Innovation in Clouds, Internet and Networks and Workshops (ICIN), 2018, vol. 91, pp. 1–5.
7. A. Vazhayil, R. Vinayakumar, and K. Soman, "Comparative Study of the Detection of Malicious URLs Using Shallow and Deep Networks," in 2018 9th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2018, 2018, pp. 1– 6.
8. W. Fadheel, M. Abusharkh, and I. Abdel-Qader, "On Feature Selection for the Prediction of Phishing Websites," 2017 IEEE 15th Intl Conf Dependable, Auton. Secur. Comput.15th Intl Conf Pervasive Intell.Comput. 3rd Intl Conf Big Data Intell.Comput. Cyber Sci. Technol. Congr., pp. 871–876, 2017.
9. X. Zhang, Y. Zeng, X. Jin, Z. Yan, and G. Geng, "Boosting the Phishing Detection Performance by Semantic Analysis," 2017.
10. L. MacHado and J. Gadge, "Phishing Sites Detection Based on C4.5 Decision Tree Algorithm," in 2017 International Conference on Computing, Communication, Control and Automation, ICCUBEA 2017, 2018, pp. 1–5.
11. A. Desai, J. Jatakia, R. Naik, and N. Raul, "Malicious web content detection using machine leaning," RTEICT 2017 - 2nd IEEE Int. Conf.Recent Trends Electron. Inf. Commun. Technol. Proc., vol. 2018–Janua, pp. 1432–1436, 2018.
12. M. Aydin and N. Baykal, "Feature extraction and classification phishing websites based on URL," 2015 IEEE Conf. Commun. NetworkSecurity, CNS 2015, pp.769–770, 2015.
13. L. A. T. Nguyen, B. L. To, H. K. Nguyen, and M. H.Nguyen, "A novel approach for phishing detection using URL-based heuristic," 2014 Int. Conf. Compu t.Manag. Telecommun. Com ManTel 2014, pp. 298–303, 2014.
14. S. Marchal, J. Francois, R. State, and T. Engel, "Phish Score: Hacking phishers' minds," Proc. 10th Int. Conf. Netw. Serv. Manag. CNSM 2014, pp. 46–54, 2015.
15. A. Ahmed and N. A. Abdullah, "Real time detection of phishing websites," 7th IEEE Annu. Inf. Technol. Electron. Mob. Commun .Conf. IEEE IEMCON 2016, 2016.
16. Phishing Websites Detection Using Data Mining Classification Model Jabri, Riad& Ibrahim, Boran.(2015). Phishing Websites Detection Using Data Mining Classification Model. Transactions on Machine Learning and Artificial Intelligence.
17. Data mining a way to solve Phishing Attacks P. K. Sahoo, "Data mining a way to solve Phishing Attacks," 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT), Coimbatore, 2018, pp. 1-5.
18. Phishing Emails Detection Using CS-SVM W. Niu, X. Zhang, G. Yang, Z. Ma and Z. Zhuo, "Phishing Emails Detection Using CS-SVM," 2017 IEEE International Symposium on Parallel and Distributed Processing with Applications and 2017 IEEE International Conference on Ubiquitous Computing and Communications (ISPA/IUCC), Guangzhou, 2017, pp. 1054-1059.
19. Detection of phishing attacks Baykara and Z. Z. Gürel, "Detection of phishing attacks," 2018 6th International Symposium on Digital Forensic and Security (ISDFS), Antalya, 2018, pp. 1-5.
20. Email phishing detection and prevention by using data mining techniques Ş. Şentürk, E. Yerli and İ. Soğukpınar, "Email phishing detection and prevention by using data mining techniques," 2017 International Conference on Computer Science and Engineering (UBMK), Antalya, 2017, pp. 707-712.