IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
Special Issue || March. 2022 || PP 01-07
National Conference on Innovations and Research in Science & Technology (NCIRST-22)

Detection of Phishing Web as an Attack: A Comprehensive Analysis of Machine Learning Algorithms on Phishing Dataset

Mr. Swapnil S. Chaudhari[1], Dr. Satish N. Gujar[2,] Dr. Farhat Jummani[3]
[1]*(Research Scholar in Computer Engineering, JJT University, India)*
[2, 3]*(Research Guide in Computer Engineering, JJT University, India)*

Abstract: *In this day's era, the Internet and the web users are increasing day-by-day along with electronic trading, making bank payments online, purchases and transactions on daily life products and services. Due to this growth, it may lead to another's excess Like stealing the user's information or logging on behalf of users in their systems and damaging the resources. Phishing is one of the techniques which may lead to diverting the users into malicious content websites and stealing all the information. The objective of the Phishing mechanism Is to steal or take a user's credentials like Username, passwords, Credentials for banking transactions, etc. As technology continues to grow the Phishing mechanism starts to progress so we have to prevent it somewhere by using the anti-phishing mechanisms to detect the Phishing at the source. Machine learning is a powerful Tool against Phishing attacks, due to machine learning algorithms we can classify all content and we can detect if the Phishing is there or not. In this paper as an Experimental setup, we have taken the dataset of a phishing website named 'phising.csv' having 10887 Rows and 31 Columns. We have checked cross-validation as the correlation between features. By using ExtraTreesClassifier we have encountered the feature importances. Finally, we have tested XGBoost Classifier, Random Forest Classifier, Decision Tree Classifier, KNN Model, SVM Classifier, Logistic Regression Model, AdaBoost Classifier algorithms for better accuracy and we found out XGBoost Classifier & Random Forest Classifier has Better accuracy. Again we have applied SMOTE and PCA Techniques on Dataset for accuracy deviations on XGBoost Classifier & Random Forest Classifier but we got the same results as in normal state as XGBoost Classifier accuracy is 96.8312% & Random Forest Classifier accuracy is 96.7853%*

Keywords*: Phishing Attack, Detection Phishing Website, Machine Learning Algorithms, Classification, Random Forest Classifier, XGBoost Classifier.*

I. Introduction

To steal personal and important information of the users, cyber criminals mostly uses phishing mechanism. in this reaction mechanism the cybercriminal are going to use the phish websites or the phishy content in the form of email or a phone call who is pretending to be the existing someone, but they are not in many cases such as your bank. So on the behalf of your bank they will provide you the email, they will provide you a website and according to that if you login on that particular website your information will be disclosed and further you will redirect to the main page as a legitimate. Cyber criminals use phishing emails because they are very easy and effective. The email addresses are easy to maintain or obtain, the emails are free to send so that's why they encode all the things to your email addresses and send email addresses to you on behalf of your legitimate or original sources. Machine learning algorithms that identify phishing URLs typically calculate a URL based on some feature or set of features taken from it. There are two typical types of features that can be taken from URLs, namely host-based features and token based/lexical features. Host based features tell characteristics of the website, such as where it is located, who manages it , and when was the site installed. Alternatively, lexical features describe textual properties of the URL. URLs are only text strings that can be divided into categories including the protocol, hostname, and path, a system can assess a site's legitimacy based on any combination of those components.

There are different type of Machine Learning concepts and formats. One of them is classification technique. Basically, classification is about identifying in which set of categories a certain observation belongs in the system datasets. Classifications are normally belonging to supervised learning techniques in the field of Machine Learning. A typical classification is Spam detection in e-mails in Gateway – the two possible classifications in this case are either "spam" or "no spam". The two most common classification algorithms are the Naive Bayes classification, the random forest classification, Decision Tree Classifier & Linear SVC Classifier In this study we have concentrated on these four classification techniques and took observations on two different datasets from Malware Detection.

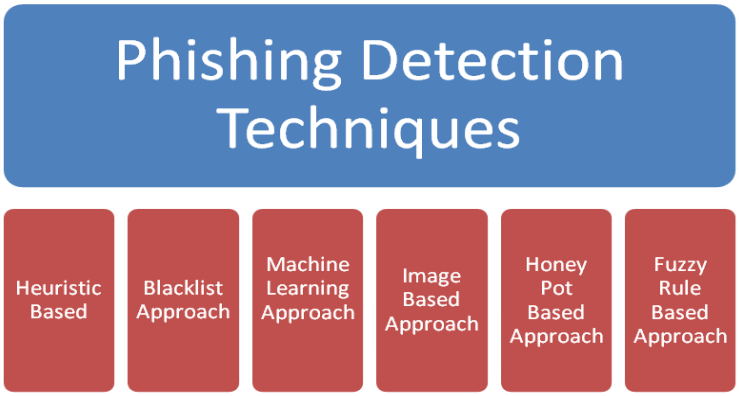

Fig 1: Phishing detection techniques

II. Related Work and Terminologies

Jain A.K., Gupta B.B, at el [15], In this paper, Authors proposed a machine learning based anti-phishing system for computers (named as PHISH-SAFE) based on Uniform Resource Locator (URL) features. To evaluate the performance of their proposed system, they have taken 14 features from the URL to detect a website as phishing or non-phishing. The proposed system is trained using more than 33,000 phishing and good URLs with SVM and Naive Bayes classifiers used. their experiment results show more than 90% accuracy in detecting phishing websites using the SVM classifier. Purbay M., Kumar D, at el [19], This article deals with methods to detect phishing URLs by monitoring different components of URLs using machine learning and deep learning techniques. they have proposed different supervised learning methods used for phishing URL detection based on lexical feature analysis, WHOIS properties, Page Rank, Traffic Rank details and page importance properties. Another, they have analyzed how different volumes of training data affect the classification accuracy. The algorithms being analyzed are support vector machine (SVM), K-NN, decision tree classification (DTC), random forest classification (RFC), and artificial neural network (ANN) etc. Gandotra E., Gupta D, at el [5] , They deal with the role of feature selection methods in detecting phishing WebPages efficiently and effectively in systems. A comparative analysis of machine learning algorithms is proposed on the basis of their performance without and with feature selection methods. Experiments are conducted on a phishing dataset with 30 features containing 4898 phishing and 6157 good web pages. Several machine learning algorithms are used for gaining the best results. Another, a feature selection method is introduced to improve the efficiency of the models. The best accuracy monitored by random forest both before and after feature selection with a significant improvement in model building time. The experiments demonstrate that employing a feature selection method along with machine learning algorithms can improve the build time of classification models for phishing detection without compromising their accuracy. Y. Sonmez, T. Tuncer, H. Gokal, and E. Avci, at el [31] In this article authors propose a classification mode in order to separate the phishing possibilities. This model comprises feature extraction from sites and classification of websites. In feature extraction, 30 features have been taken from UCI Irvine machine learning set repository data set and phishing feature extraction rules have been straightly defined. In order to classify these features, Support Vector Machine (SVM), Naïve Bayes (NB) and Extreme Learning Machine (ELM)were used. In the Extreme Learning Machine (ELM), six active functions were used and reached 95.34% accuracy than SVM and NB. The results were obtained with the help of MATLAB Tool.

2.1 Uniform Resource Locator

URL imply Uniform Resource Locator. A URL is the address of a novel resource on the online. Each valid address points to a novel resource in computer network. These resources can be an HTML page, an image, etc. Each URL has a definite structure.

http://www.Yourworld.com:80/Download?key1=value1&key2=value2#otherContent

Scheme:http

Domain Name: www.Yourworld.com

Port: 80

Path to file: Download

Parameters: Anchor key1=value1&key2=value2

Other Content

• **Scheme**-Scheme is the first element of URL. It express the protocol the browser must use (a protocol is a set of rules for switching or transferring data in a computer network). Most normally used protocol is HTTPS or HTTP but other than that it can have FTP or SMTP, etc.
• **Authority**-The next fraction is authority which additionally consists of two parts- Domain Name and other is the port. Domain Name points out which server to request for and is usually the registered name for IP addresses. Port points out the path used to access the resource on the web server. For example Http protocol uses port 8080 like that
• **Path to the resource**- It points to indicate the place of the file requested on the web server.
• Parameters- Parameters or the arguments are the added information supplied to the web server. They are in the form of a key-value couple and are separated using the '&' symbol.
• **Anchor**-An anchor gives guidelines to the browser to find the content located at that "bookmarked" speck. The anchor is written after the '#' symbol.

2.2 Malicious URL
Attackers make certain changes to legitimate URLs such that it appears a real website but the user will be redirected to a fake website. Attackers usually change the subdomain name and path of the URL.
http://YourBank.com-webpageuser123.Passive.com/webpage12345
Protocol: http://
Domain Name: Passive.com
Path : webpage12345
Subdomain item1: Com-webpageuser123
Subdomain item2: YourBank
Attackers use Cyber squatting and Typo squatting techniques to tempt users.

Cyber squatting entails buying website URLs of previously established businesses that do not have a related website. Typo squatting entails buying a look-alike website URL that appears alike to the genuine URL of an recognized organization but actually contains a misprint. Inventory or using a domain name for phishing purposes is called Cyber squatting. For example, if the real website domain of a company is Mypepeer.com then phishers will register a domain such as Mypepppers.com or Mypeppers.in.

Typo squatting depends on the typographical faults which are being completed by users. The URL appears to be a good one but isn't. For example, google.com is typo squat as goggle.com to attract users, some other examples such as microhoft.com likewise.

2.3 Types of Phishing Attacks

a) **Malware:** By clicking link Malware Downloaded into system and get executed
b) **Smishing:** Using short message services, attackers send malicious sites to users.
c) **Vishing:** By using Voice changing software, attackers are used to call users and take the information
d) Spear phishing: these email messages are sent to specific people within an organization, usually high-level priority holders.
e) **Link manipulation:** Email or SMS messages contain a link to a malicious site that looks like the official business site or official bank site
f) **CEO/Principal fraud:** these messages are sent mainly to organizational people to trick them into believing that the CEO or Principal of college or other executive is asking them to transfer money.
g) **Content injection:** an attacker who can inject malicious content into an official site will trick users into accessing the site to provide them a malicious popup or redirect them to a phishing website.
h) **Wi-Fi:** spoofing free Wi-Fi, attackers trick users into connecting to a malicious hotspot so that they can perform a man-in-the-middle attack and see all the conversation.
i) **Clone phishing:** Clone phishing is a type of phishing in which previously delivered emails are taken with the same attachment and create the same clone of that email. Sometimes The attachment or link within the email is replaced with a malicious link and then sent from a spoofed email address to look like it came from the original sender.
j) **Spear phishing:** It targets a specific person, enterprise or organization, as opposed to random application users for their organization. It requires special knowledge about an organization, including its power structure.

III. Results

We have taken the dataset of a phishing website named 'phising.csv' having 10887 Rows and 31 Columns. We have checked cross-validation as the correlation between features. By using ExtraTreesClassifier we have encountered the feature importances. Finally, we have tested XGBoost Classifier, Random Forest Classifier, Decision Tree Classifier, KNN Model, SVM Classifier, Logistic Regression Model, AdaBoost Classifier algorithms for better accuracy and we found out XGBoost Classifier & Random Forest Classifier has Better accuracy. Following figure shows the ROC curves for above said Classifiers individually.

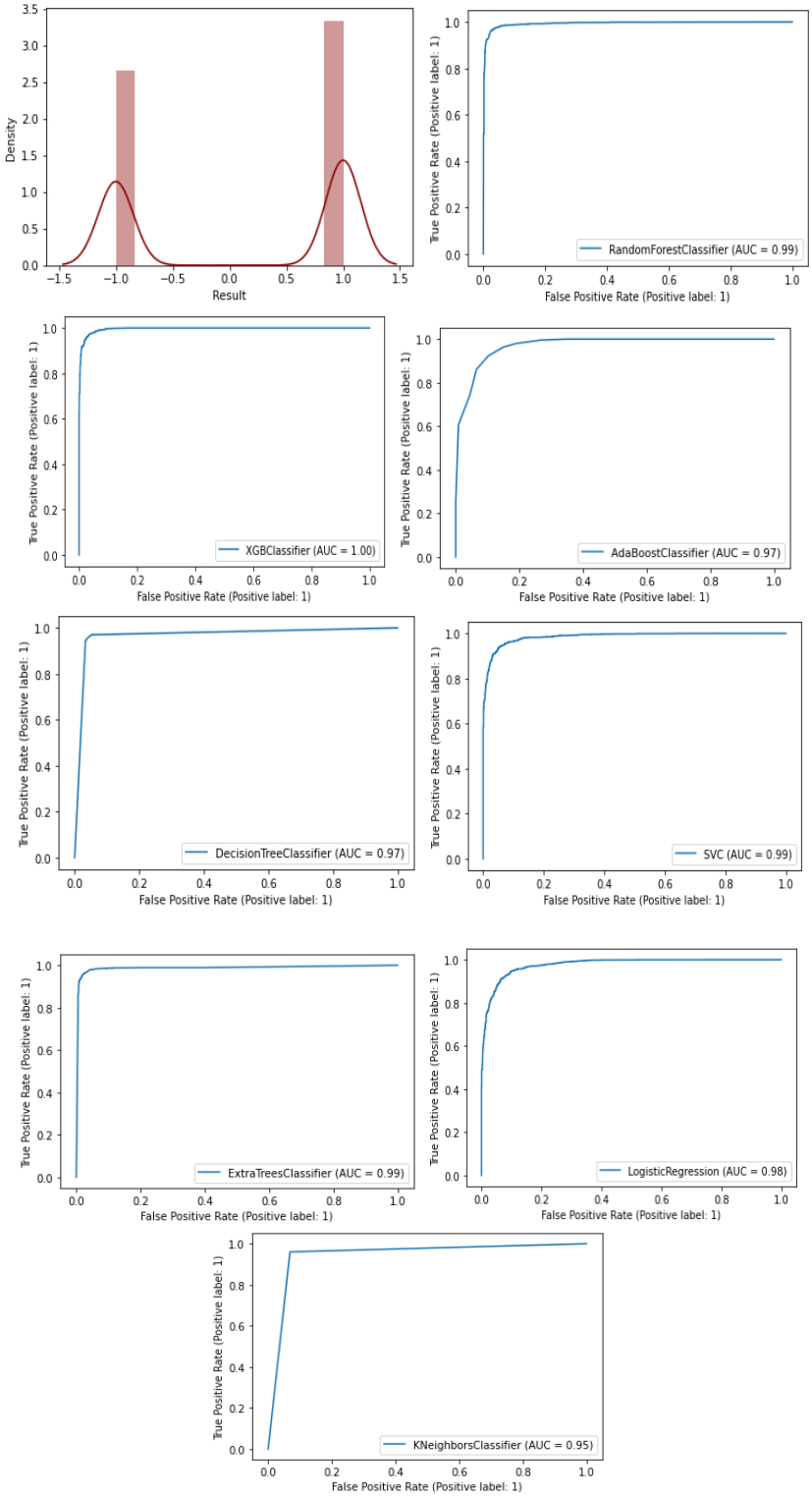

Fig. 2 : ROC curves for above said Classifiers individually

Table no 1 Shows Phishing website data classification score as Ada Boost Classifier=91.13662457, Logistic Regression Model=92.30769231, SVM Classifier=94.1216992, KNN Model=94.71871412, Decision Tree Classifier=95.66016073, Extra Trees Classifier=96.57864524, Random Forest Classifier=96.78530425, XGBoost Classifier=96.83122847.

Table no 1: Shows Phishing website data Machine Learning classification score

Classifier Model	Score
Ada Boost Classifier	91.13662457
Logistic Regression Model	92.30769231
SVM Classifier	94.1216992
KNN Model	94.71871412
Decision Tree Classifier	95.66016073
Extra Trees Classifier	96.57864524
Random Forest Classifier	96.78530425

3.1 Comprehensive Analysis of all Machine Learning Classifiers (Above Experimental Classifiers)

Here we have combines all ROC curves as X axis: False Positive Rates, Y axis: True Positive Rates, taken from Machine Learning Classifiers like XGBoost Classifier, Random Forest Classifier, Decision Tree Classifier, KNN Model, SVM Classifier, Logistic Regression Model, AdaBoost Classifier algorithms for better accuracy and we found out XGBoost Classifier & Random Forest Classifier has Better accuracy. Following figure shows the Combines ROC comprehensive analysis for above said classifiers & Accuracy Scores.

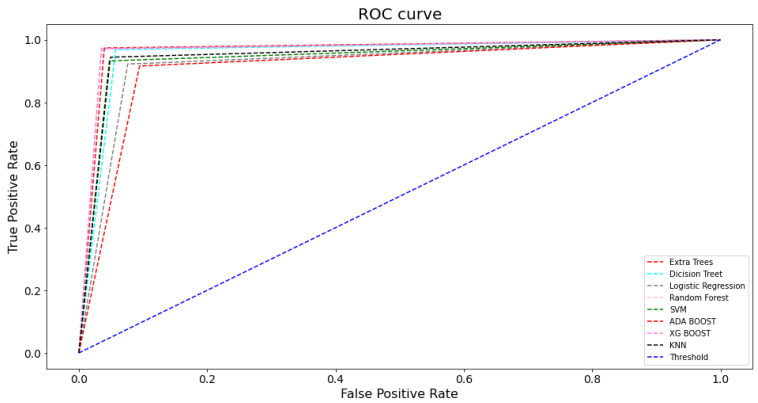

Fig. 3: Comprehensive ROC curves for Classifiers

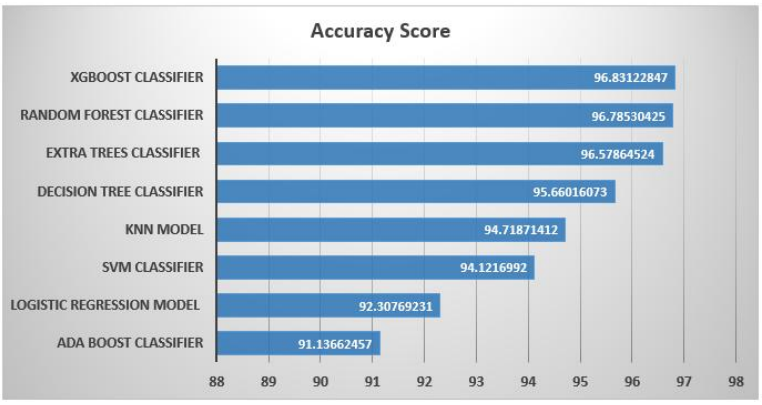

Fig. 4: Accuracy Score for Different classifiers

IV. Conclusion

we have taken the dataset of a phishing website named 'phising.csv' having 10887 Rows and 31 Columns. We have checked cross-validation as the correlation between features. By using ExtraTreesClassifier we have encountered the feature importance. Finally, we have tested XGBoost Classifier, Random Forest Classifier, Decision Tree Classifier, KNN Model, SVM Classifier, Logistic Regression Model, AdaBoost Classifier algorithms for better accuracy and we found out XGBoost Classifier & Random Forest Classifier has

Better accuracy. Again we have applied SMOTE and PCA techniques on Dataset for accuracy deviations on XGBoost Classifier & Random Forest Classifier but we got the same results as in normal state as Random Forest Classifier accuracy is 96.7853% & XGBoost Classifier accuracy is 96.8312%

Data and Code

To facilitate to other researchers for future work or to obtain highest accuracy of the research in this paper, all codes and data are shared at this repository: https://github.com/swapnlc39/Phishing-detection-classification-using-ML

Acknowledgement

I express our sincere thanks to Dr. Farhat Jummani for her kind cooperation and Valuable Guidance. I also express our sincere thanks to Dr. Satish N. Gujar for her kind Cooperation in research work.

References

[1]. 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.

[2]. AlEroud A, Karabatis G. Bypassing detection of URL-based phishing attacks using generative adversarial deep neural networks. In: Proceedings of the Sixth International Workshop on Security and Privacy Analytics 2020 Mar 16 (pp. 53–60).

[3]. Aljofey A, Jiang Q, Qu Q, Huang M, Niyigena JP. An effective phishing detection model based on character level convolutional neural network from URL. Electronics. 2020 Sep; 9(9):1514.

[4]. Chiew KL, Chang EH, Tiong WK,"Utilisation of website logo for phishing detection", Computer Security, pp.16–26, 2015.

[5]. Gandotra E., Gupta D, "An Efficient Approach for Phishing Detection using Machine Learning", Algorithms for Intelligent Systems, Springer, Singapore, 2021, https://doi.org/10.1007/978-981-15-8711-5_ 12.

[6]. Gandotra E., Gupta D, "An Efficient Approach for Phishing Detection using Machine Learning", Algorithms for Intelligent Systems, Springer, Singapore, 2021, https://doi.org/10.1007/978-981-15-8711-5_12.

[7]. Gupta D, Rani R, "Improving malware detection using big data and ensemble learning", Computer Electronic Engineering, vol. 86, no.106729, 2020.

[8]. Hassan Y.A. and Abdelfettah B, "Using case- based reasoning for phishing detection", Procedia Computer Science, vol. 109, 2017, pp. 281–288.

[9]. Hong J., Kim T., Liu J., Park N., Kim SW, "Phishing URL Detection with Lexical Features and Blacklisted Domains", Autonomous Secure Cyber Systems. Springer, https://doi.org/10.1007/978-3-030-33432-1_12.

[10]. Hung Le, Quang Pham, Doyen Sahoo, and Steven C.H. Hoi, "URLNet: Learning a URL Representation with Deep Learning for Malicious URL Detection", Conference'17, Washington, DC, USA, arXiv:1802.03162, July 2017.

[11]. J. Anirudha and P. Tanuja,"Phishing Attack Detection using Feature Selection Techniques ", Proceedings of International Conference on Communication and Information Processing (ICCIP), 2019, http://dx.doi.org/10.2139/ssrn.3418542

[12]. J. Kumar, A. Santhanavijayan, B. Janet, B. Rajendran and B. S. Bindhumadhava, "Phishing Website Classification and Detection Using Machine Learning," 2020 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2020, pp. 1–6, 10.1109/ICCCI48352.2020.9104161.

[13]. J. Shad and S. Sharma, "A Novel Machine Learning Approach to Detect Phishing Websites Jaypee Institute of Information Technology," pp. 425–430, 2018.

[14]. Jain A.K., Gupta B.B. "PHISH-SAFE: URL Features-Based Phishing Detection System Using Machine Learning", Cyber Security. Advances in Intelligent Systems and Computing, vol. 729, 2018, https://doi. org/10.1007/978-981-10-8536-9_44

[15]. Jain A.K., Gupta B.B. "PHISH-SAFE: URL Features-Based Phishing Detection System Using Machine Learning", Cyber Security. Advances in Intelligent Systems and Computing, vol. 729, 2018, https://doi. org/10.1007/978-981-10-8536-9_44

[16]. 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.

[17]. 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.

[18]. 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.

[19]. Purbay M., Kumar D, "Split Behavior of Supervised Machine Learning Algorithms for Phishing URL Detection", Lecture Notes in Electrical Engineering, vol. 683, 2021, https://doi.org/10.1007/978-981- 15-6840-4_40

[20]. Purbay M., Kumar D, "Split Behavior of Supervised Machine Learning Algorithms for Phishing URL Detection", Lecture Notes in Electrical Engineering, vol. 683, 2021, https://doi.org/10.1007/978-981-15-6840-4_40

[21]. Rao RS, Pais AR. Jail-Phish: An improved search engine based phishing detection system. Computers & Security. 2019 Jun 1; 83:246–67.

[22]. 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.

[23]. Sahingoz OK, Buber E, Demir O, Diri B, "Machine learning based phishing detection from URLs", Expert System Application, vol. 117, pp. 345–357, 2019.

[24]. Srinivasa Rao R, Pais AR, "Detecting phishing websites using automation of human behavior", In: Proceedings of the 3rd ACM workshop on cyber-physical system security, ACM, pp 33–42, 2017.

[25]. Swapnil S. Chaudhari, Dr. Satish N. Gujar, Dr. Farhat Jummani, "Comparative Analysis Of Malware Detection Datasets Using Different Machine Learning Classifiers", International Journal of Emerging Technologies and Innovative Research (www.jetir.org), ISSN:2349-5162, Vol.9, Issue 1, page no.d698-d703, January-2022, Available :http://www.jetir.org/papers/JETIR2201390.pdf

[26]. 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.

[27]. 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.

[28]. Wu CY, Kuo CC, Yang CS," A phishing detection system based on machine learning" In: 2019 International Conference on Intelligent Computing and its Emerging Applications (ICEA), pp 28–32, 2019.
[29]. X. Zhang, Y. Zeng, X. Jin, Z. Yan, and G. Geng, "Boosting the Phishing Detection Performance by Semantic Analysis," 2017.
[30]. 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– January, pp. 1–5, 2018.
[31]. 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