Proceedings of the Sixth International Conference on Inventive Computation Technologies [ICICT 2021]
IEEE Xplore Part Number: CFP21F70-ART; ISBN: 978-1-7281-8501-9

Detection of Phishing Websites from URLs by using Classification Techniques on WEKA

Buket Geyik
Istanbul Kultur University
Computer Engineering Department
Istanbul, Turkey
bukettgk@hotmail.com

Kübra Erensoy
Istanbul Kultur University
Computer Engineering Department
Istanbul, Turkey
erensoykubra13@gmail.com

Emre Kocyigit
Yildiz Technical University
Department of Computer Engineering
Istanbul, Turkey
kocyigit.emre.30@gmail.com

2021 6th International Conference on Inventive Computation Technologies (ICICT) | 978-1-7281-8501-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICICT50816.2021.9358642

***Abstract*—The Internet is getting stronger day by day and it makes our lives easier with many applications that are executed on cyberworld. However, with the development of the internet, cyber-attacks have increased gradually and identity thefts have emerged. It is a type of fraud committed by intruders by using fake web pages to access people's private information such as userid, password, credit card number and bank account numbers, etc. These scammers can also send e-mail from many important institutions and organizations by using phishing attacks which imitate these web pages and acts as if they are original. Traditional security mechanisms can not prevent these attacks because they directly target the weakest part of connection : end-users. Machine learning technology has been used to detect and prevent this type of intrusions. The anti-phishing method has been developed by detecting the attacks made with the technologies used. In this paper, we combined the websites used by phishing attacks into a dataset, then we obtained some results using 4 classification algorithms with this dataset. The experimental results showed that the proposed systems give very good accuracy levels for the detection of these attacks.**

***Index Terms*—phishing attacks, machine learning, classification algorithms, phishing detection, cybersecurity**

I. Introduction

In the developing world, we use the internet very actively to provide communication and reach information, and the user base has recently increased with internet applications. For this reason, thanks to the internet, communication and information transfer with social networks such as banking, e-commerce, e-mail, and social media applications like Instagram have drastically ascended, and it has a huge positive influence in our lives [1]. On the other hand, security measures are not sufficiently organized and capable of preventing a wide variety of cyber-attack threats or protecting computer users. This is a vital security problem for even experienced and educated users and any cyber-threat like phishing attack can cause crucial losses.

Phishing is an online attack by fraudsters and it is sent to user accounts to collect sensitive, personal, and financial information. Phishing attacks seek access to especially financial information using emails, official websites, credit card companies. While doing these, there is a URL link for the user to point to another website. The website that the user is connected to is a fake website with an innocent appearance and the information that the user share with this website is transmitted to the phisher.

The numbers of phishing websites detected in the first, second, and third quarters of 2020 were in the order of 165,772 146,994 and 571,764. Totally 884,530 unique phishing websites were detected for the first third quarter of this year. If we look for 2019 in the same order, the detected websites were 180,768 182,465 and 266,378 totally 629,611. This means an increase of approximately 40% in phishing websites in a year [2].

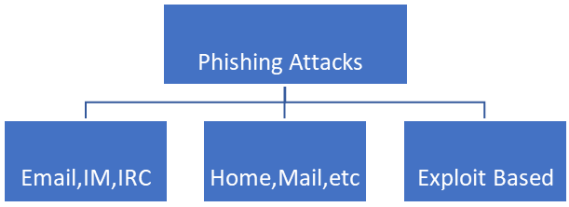

Fig. 1. Phishing attacks.

In Phishing attacks, phisher mostly designs a fake web page. This web page appears similar with the original web page and has a different but deceptive URL. By this way, they can access the private information of the users. A careful user may notice that the URL is malicious and belongs to phishing. However, phishers take advantage of human vulnerabilities and social engineering techniques to hide their scam.

E-mails, which are sent by phishers, have the appearance of official e-mail account of institutions and organizations as a part of deceptive process of phishing. When the user clicks on these e-mails, it leads to a malicious website. This website uses the credentials entered by the user. This information is saved on a different server. The phishing uses them to commit a cybercrime. Various phishing techniques and methods have been developed with the advancement of technology in order to acquire confidential data of users. Phishing techniques are shown in the table below [3].

There are also anti-phishing techniques and measures in order to get rid of spam messages and protect vital information of users. Recently, they have been developing with different approaches. These techniques are:

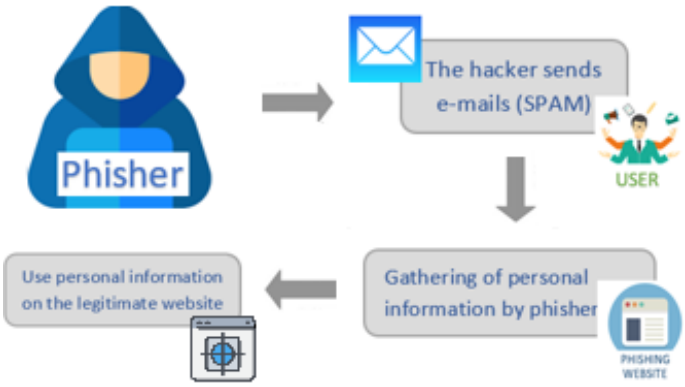

Fig. 2. Phishing process.

Phishing Techniques	Definition
Spear Phishing	Spear phishing, hackers don't target random people or organizations here. So phisher does specific research to launch an attack and organize personal attacks to trap the target.
Session Hijacking	Phisher uses this method to steal information from users through its web session control mechanism. The identity hunter accesses the web server illegally with the help of listeners.
Email/Spam	E-mail, one of the most common techniques, asks to access user information by e-mail sent to millions of people. These messages sends a form to the users to fill in their personal account information to access their accounts.
Content Injection	Content Injection, is connected to a different web page to access the personal information of audiophiles varying portions of their content in a trusted website.
Web Based Delivery	Phisher observes our actions on the website, transmitting our information to the phishing site without the user's knowledge.
Phishing through Search Engines	It is a technique that users carry out by stealing the credit card information of the products they research on search engines.
Link Manipulation	It is a method of connecting to a malicious website when you click on the link sent by the phishing.
Smishing (SMS Phishing)	It is the method made by SMS sent to our phones. They can access our personal information via the link in the message section.
Malware	Phishing scammers, as soon as we click on the link they send to our e-mail, the malware will start running on our computer.
Trojan	It is a type of maliciously written software to access credentials.
Ransomware	Here it is denied access to a file and device to get a ransom from the user. When the user clicks on a link or is tricked by a malicious ad, the malware gets installed on their computer.

1) Spam filters have been designed to detect and prevent detrimental and phishing e-mails.
2) Web browsers such as Google Chrome, Internet Explorer, Mozilla Firefox have taken browser measures that warn us of phishing on the websites we enter.
3) Using different password entries. For example, some banks have added a password by adding images to be selected by the users other than a certain password. This has increased the password entries.
4) To prevent phishing scams, some organizations analyze websites to remove them.
5) Users of multi-factor authentication methods application for entry to specific pages.

To be aware and detect phishing we should take a look at the techniques used by phishers. Some of these techniques are [4]:

- Using an IP address in the URL
- Using long URL to hide doubtful part
- Using tiny URL to hide a long URL
- Having @ symbol in the URL
- Adding known prefix or suffix to domain name
- Using "https" word in domain name

The rest of this paper is organized as follows: in Section II, the related works about phishing detection are reviewed. Section III focuses on proposed system and gives details about the dataset, tools and machine learning methods that are used in this work then gives information about data preprocessing part. Results are depicted in Section IV. Section V discusses about the work and about how to improve this work in the future. Finally, conclusion is presented in Section VI.

II. Related Work

Phishing is an old problem in the internet history. Hackers consistently and insidiously try to obtain and abuse people's information. Users must be quite careful to avoid from these kind of attacks and effective and well-organized strategies should be generated. This study aims to carry out accurate predictions of phishing websites by several algorithms. Marchal, François, State and Engel (2014), build a phishing dataset by downloading the daily PhishTank blacklist data between October 11th and November 10th, 2012 with 53,089 unique URLs [5]. After a selection they had 48,009 extended phishing URLs. Then for a balanced dataset they get same amount of malicious URLs from DMOZ. This study uses supervised classification techniques. They build a feature vector matrix from the dataset, each one is composed of 12 elements. Predicted variable is 0 for legitimates and 1 for phishings. Using Weka they have tested seven classifiers. With Random Forest Classifier they achieve 94.91% accuracy with 1.44% false positive rate.

Jain and Gupta [6], extracts nineteen features from client side only, URL and source code of the websites. The data is collected mostly from Phishtank, Openphish for verified URLs and Alexa for the legitimate ones, which includes 4059 websites with 2141 phishing and 1918 legitimate sites in training dataset. They get 99.39% of TPR and 1.25 of FPR. They implemented intuitive methods to produce the feature vector and generate a singular feature vector for each website sample to build labelled dataset. They have evaluated the dataset with 10-fold cross-validation. The study has reached 99.09% accuracy with random forest, 96.16% with SVM, 98.05% with neural networks, 98.25% with Logistic Regression (LR), and 97.59% with Bayes by using WEKA.

Weedon, Tsaptsinos and Denholm-Price [7], get the dataset from Phishtank which is completely verified and DMOZ websites. Their study contained 4000 URLs for training process

and quarter of the training data belongs to malicious dataset and remaining data belongs to the opposite one. Their testing set consists of 7000 URLs and nearly 57% of them belong malicious dataset. The study used a literal only dataset in order to assess the accuracy of Random Forest algorithm and gets 86.9%. With other algorithms they get 83.9% accuracy with j48, 64.6% with Bayes, 81.5% with LR.

Sahingoz, et. al. [8], [14], provide the phishing URLs mostly from Phishtank by writing a script. Over 70000 URLs were available in their dataset and roughly half of them were legitimate websites and other half of the dataset were phishing websites. They extract each word from these URLs to use in analyses. Then they implemented a Random Word Detection Module and all words, which had over seven characters, were examined via Word Decomposer Module(WDM) to separate their subwords. For the words are not compound, they obtained only the original ones by WDM. After that their Maliciousness Analysis Module examined and processed the output words of WDM and the words that were up to seven characters. Then a couple of auxilary features were extracted from these words. Random Forest was the most successul algorithm with 97.98% accuracy. Natural Language Processing based features increased the performance and had better scores than word vectors in all algorithms but Naive Bayes.

Liu, Wang, Lang and Zhou [9], uses WEKA library to execute Random Forest, J48, LR, SVM, MLP and Bayes algorithms. 29000 URLs were available in their dataset and approximately 12500 malicious of them were obtained from Phishtank. Then they were combined with 16.516 legitimate URLs taken from digg58 website. They identified 41 features in their study and adjusted the Random Forest algorithm. After their implementations, considerable scores were acquired by Random Forest classification. The algorithm's precision was 99.7% and FPR, which is important factor in this kind of problems, was less than 0.4%.

Rakesh et al. [10], also use Weka tool. They collected the legitimate URL set from DMOZ and fake samples from widespread source, Phishtank. The dataset consists of balanced 2000 URLs. This study had 9 features that were extracted using a java program. They generated 6 particular subsets in variable rates to observe the difference of accuracy rate by dataset size. In their project, they classified the URLs by C4.5 algorithm in WEKA. As a result higher accuracy scores belonged to C4.5 and AdaBoost algorithms.

Aydın et. al. [11] point out the most attacked websites and their devious URLs from the Phishtank. The study analyzed totally 8,538 URLs. Nearly 40% of them were legitimate and remaining ones were fraudulent. Their program got textual properties and "whois" records. Additionally, they obtained some data manually. Dataset has 133 separate features related to URLs. Gain Ratio Attribute (GRA) and ReliefF Attribute used for the feature selection and analyzed by WEKA. The SMO and J48 algorithms achieve their best results by using ReliefF attribute-based(58) selection technique and got 96.42%, 98.47% accuracies. Naive Bayes reaches better result using Gain Ratio Attribute(36) with 87.08% accuracy.

İbrahim and Hadi [12], use WEKA tool for implementing the classifiers on public NASA repository dataset. The dataset has 30 attributes with 11055 instances. They categorized their dataset into four parts such as address bar, abnormal, website content and domain based features. They used K-fold cross validation first and k value was 10. Random Forest get the highest accuracy 95.2% with and without feature selection algorithms and Bayes get the least. All classifiers get better accuracy results than using feature selection.

James et al. [13], analyzed some algorithms using WEKA and MATLAB. First, they extract the features. Then they choose a classifier to implement in MATLAB. They collect URLs of benign websites from Alexa, DMOZ websites and web browser past. They collected 37000 URLs and 45% of them were phishing samples from Phishtank. They collect WHOIS information of some websites. By using only the lexical features, they generate a successful classification rates as 93.2% for test section of 60% and 93.78% for test section of 90%. They used Regression Tree by MATLAB and the accuracy rate was 91.08% in 60% of dataset although accuracy rate was 85.63% in 90% of dataset.

Priya and Meenakshi [15], analysis C4.5 (J48) algorithm using WEKA tool. Phishing and legitimate websites are collected from PhishTank. Thirty-two features are extracted from the websites. Two training datasets are created with 750 URLs and 2000 URLs to train the J48 algorithm. The test dataset has 300 URLs. The size of tree is 45 nodes out of 28 nodes are leaves. But if 2000 URLs used then 75 nodes are created, out of 43 are leaf nodes. The algorithm has 82.6% accuracy rate.

III. Proposed System

A. Dataset Description

Phishers try to click the URL of the site their victims enter in their attacks. Identity hunters use some differences to change the appearance of the URL structure in various ways. These differences used change the URLs and look different from the legitimate site. What we are going to do here is that by doing extensive research on the malicious URL, we use some properties to classify the web page. And we analyze the URLs we detected. Some features are defined below for the malicious URL [10].

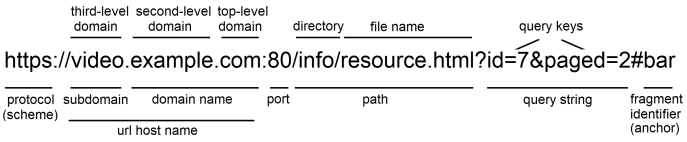

Fig. 3. URL components.

The CatchPhish_D3.csv dataset had 126,077 rows and 2 columns which is full of site names and phishing value if it is a phishing site 1 if not 0 is given. We get the dataset from [21]. In this dataset, legitimate sites are collected from common-crawl and Alexa while phishing sites are collected from Phishtank. After some pre-processing steps 122,055 values 85,220 with 0 and 36,835 with 1 values stayed and those

values are in order 0's comes first. After extracting values from address bar values we had 15 columns for example protocol, site length, host name, file name, path, path length, fragment, number of query keys, port, number of delimiters, number of reserved characters etc.

TABLE I
DESCRIPTION OF RELEVANT NOMINAL ATTRIBUTES

Description	Values
site	full-length sites in the dataset
site_len	length of sites
protocol	http (0) https (1)
url_host_name	combine of subdomain and domain name
url_host_name_len	length of host name of a URL
file_name	name of a file
file_name_len	length of a file name
file_name_without_ext	file name length without extension
file_name_len_without_ext	length of file name without extension
path	location of a file
path_len	length of path of a URL
query	pass data to the server
query_len	length of a query string
num_of_query_keys	count of query keys of a URL
fragment	internal page reference
fragment_len	length of fragment identifier
port	by default 80 for HTTP, 443 for HTTPS
num_of_reserved_chars	reserved chars: ; / ? : @ & = + $
num_of_delimiters	delimiters: < > # % ‘
num_of_unreserved_chars	unreserved chars: - _ . ! * ~ ’ ( )
num_of_unwise_chars	unwise chars: { } | \ ^ [ ]
phish	not phish (0) phish (1)

B. Tools

1) Python: Python is created by Guido Van Rossum in early 1990s [19]. It has a simple structure so that a nice choice for educational purposes as a student or a beginner. Since being one of the preferred languages contributes more to open source projects related to this language, it leads to a rapid development cycle.

2) Jupyter Notebook: Jupyter notebook is an open source web program to lets developers to execude codes in pieces and make visualizations. In this way, it allows users to view code blocks and their results.

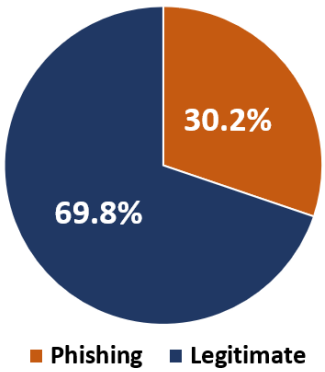

Fig. 4. Distribution of dataset according to phishing sites.

3) Weka: Waikato Environment for Knowledge Analysis (WEKA) is a Java based open-source tool which is developed by the University of Waikato. It is used for data mining. Weka includes machine learning algorithms and it is easy to use. Besides, visualization and data preprocessing tools are also included [16].

C. Methods

Data Mining (DM) is the business of accessing information or mining among big data. Our job here is to predict precisely what will come from the large chunks of data. The computer program we use for this is WEKA. The use of WEKA provides a great convenience here because WEKA obtains a very quick result in machine utilization algorithms. Thanks to the results we obtained, we can compare the algorithms we use. We used the CatchPhish dataset in our program. We obtained some features by breaking URLs in the dataset. This became a multivariate dataset. Classification is a data mining function that assigns items in a dataset to target categories or classes. The purpose of classification is to make accurate predictions for each value in the data. In the classification model, we looked at functional algorithms such as Random Forest, Decision Tree. Finally, we looked at Naive Bayes, where we used multiple algorithms, because you have the advantage of comparing the extracted information in each algorithm. As a result, the multiple classification system wants to make the best use of the data in the data set.

1) Random Forest Classifier: Random Forest is one of the most used and most popular machine learning algorithms in Classification. In this algorithm, it creates a forest made up of words as its name and combines this data to make a random prediction [17]. Here, our algorithm gets faster results than other algorithms. Additionally, it works even better.

2) Decision Tree: Tree-based learning algorithm is one of the most used algorithms for data mining classification. The algorithm, which has a tree-like model, makes some decisions to reach the desired results. But it contains some conditional statements to arrive at this conclusion [1]. This algorithm can be used in all decision trees, classification and regression problems, and this algorithm gives the best result to achieve the goal. A decision tree is used to divide a data set into even smaller sets by applying certain rules. In other words, even easier steps are taken in the data that is divided into small pieces. In addition, a decision tree that can be visualized is much easier to understand.

3) Logistic Regression: Logistic regression is the categorization problem of dependent variables used in the linear classification problem. The purpose here is to obtain an analysis of the independent variables that we use in our data set. Like the phenomenon of all regression algorithms, this algorithm is a prediction analysis. The result we will get here is taken from a binary variable. These variables are 0 (false) and 1 (true). Events in logistic regression must be independent from each other. There is no linear relationship between dependent and independent variables.

4) Naive Bayes: This classification is the simplest network model with the family of "probabilistic classifiers". The purpose is to use a vector with multiple properties. Then training is created from the information provided, and It is received at the end of this training the new data classified correctly.

D. Dataset Preprocessing

Data preprocessing is a step to get qualified data because the dataset can have incomplete, inconsistent and outdated data in it [17]. Our dataset had 2 columns in the beginning which is full site names and phishing detail. Firstly, we transform the dataset csv to excel. Then some rows in phishing column was missing so we fill them with NaN value and some rows had different numbers so we found and delete them. We drop the empty rows. After that we extract the needed features from the site column and had 15 features. We normalize some rows like port which means we give a number to each value. Then we made it available for use on Weka and made predictions J48, Bayes, Logistic Regression and Decision Tree. But before doing that we applied 5-fold cross validation which divides data into subsets and leaves the last part as test data.

TABLE II
CONFUSION MATRIX TERMS

	Predicted Positive Value	**Predicted Negative Value**
Real Positive Value	True Positive (TP)	False Negative (FN)
Real Negative Value	False Positive (FP)	True Negative (TN)

True Positive Rate (TPR): Here the classifier calculates how accurately it predicts true positive values. The higher the better.
TPR = TP / (TP+FN)
False Positive Rate (FPR): Here the classifier calculates how accurately it predicts true negative values.
FPR = FP / (FP+TN) Accuracy: Here is how often the classifier gets the correct predictions.
Accuracy = (TP+TN) / (TP+FP+TN+FN)
Precision: It is a measure of accuracy in all estimated classes. It is preferable to be high.
Precision = TP/ (TP+FP)

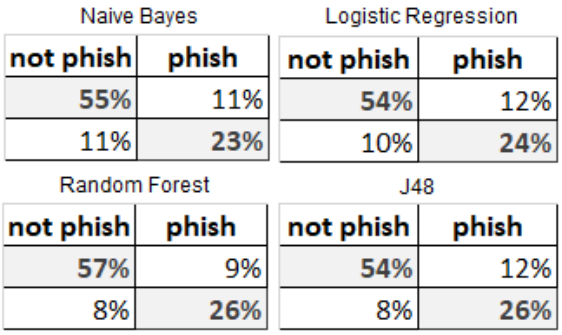

Fig. 5. Confusion matrix for each algorithm.

The confusion matrix is shown above. The diagonal values, which is TP and FN, of the matrices shows the estimated correct values. By looking at this, we can say that Random Forest predicted the highest correct values.

IV. EXPERIMENTAL RESULTS

Recall: The proportion of positive samples is calculated according to the total number of positive samples in the correct classification used.
Recall = TP / (TP+FN)
F1-Score: It is the harmonic mean of Recall and Precision values. The purpose here is to measure the performance value shown by the classifiers. It is mostly used to compare classifiers.
F1-Score = 2 * Precision * Recall/ (Precision + Recall)
ROC Curve: Here are the graphs used by calculating the performance for all values consisting of classifiers. The ROC curve creates a Sensitivity / Specificity report. The area under the ROC curve is called AUC. It uses this field as an evaluation criterion. AUC is a measure of how well a parameter can be distinguished between two classes.

TABLE III
PERFORMANCE METRICS

Algorithm	Class	Precision	Recall	F1-Score
Random Forest	Not Phish	0,877	0,864	0,870
	Phish	0,743	0,765	0,754
Decision Tree	Not Phish	0,871	0,818	0,844
	Phish	0,684	0,765	0,722
Naive Bayes	Not Phish	0,833	0,833	0,833
	Phish	0,676	0,676	0,676
Logistic Regression	Not Phish	0,844	0,818	0,831
	Phish	0,667	0,706	0,686

Random Forest Classifier's results are shown below. We can say that this algorithm's execution time takes a little longer than the other algorithms that we use except logistic regression but still fast. Here, the accuracy value of Random Forest Classifier is calculated as 83%. Logistic Regression and Naive Bayes algorithms get 78% accuracy but Naive Bayes was faster. J48 was also fast and get 80% accuracy percentage. As a result, we can say that the Random Forest Classifier gave us the best result.

TABLE IV
ALGORITHMS

Algorithm	**Test results**	
Names	***Accuracy***	***Run time***
Random Forest	83%	0.06sec
Decision Tree (J48)	80%	0.03sec
Naive Bayes	78%	0.01sec
Logistic Regression	78%	0.07sec

Algorithm accuracy and runtimes.

V. DISCUSSION

In this study, the accurate prediction of phishing websites by different classification techniques is the ultimate goal and

we divided our dataset into two main parts as training and test in the first phase. Initially, we tried to extract some effective features from the URL dataset that we can use to detect phishing. Then we make data preprocessing to clear and prepare the data. Then we apply Random Forest, Decision Tree, Naive Bayes and Logistic Regression algorithms to reach the most qualified result and compare each one's scores. We observed that addition of phishing samples increased the accuracy score of algorithms. Balanced and enhanced dataset can create better solutions in this case. In addition to all, some deep learning models which showed the efficiency in [14] can be adopted to the proposed model in the future work.

VI. Conclusion

In this paper, we have executed a phishing detection system on WEKA and tested its efficiency by using a public dataset as CatchPhish_D3 by using different classification techniques. The dataset has 2 columns and we extracted some features and create a new dataset with a structured format. To make this, it is needed to make some preprocessing steps to use the dataset in Weka system. For detection of phishing sites URL of the web pages are mainly used. By using this data some features are produced and these features are used for detection of whether the web page is phishing or not. To predict this, four different machine learning models are used as random forest, naive Bayes, logistic regression and decision tree algorithms.

As a conclusion of this work, we found that the Random Forest algorithm works better than the others with relatively high accuracy rates. The models can be enhanced by using new features in the system as in [4], [18]. Additionally, apart from the URL based features, some content-based features [20] can also be used here. Finally, we can also get help from some third party organization/web pages as Alexa and Whois to identify whether the page is phishing or not.

References

[1] M. Korkmaz, O. K. Sahingoz and B. Diri, "Detection of Phishing Websites by Using Machine Learning-Based URL Analysis," 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kharagpur, India, 2020, pp. 1-7, doi: 10.1109/ICCCNT49239.2020.9225561.

[2] Phishing Activity Trends Report, Summary – 3rd Quarter 2020. (2020). [Online]. Available: https://apwg.org/trendsreports/

[3] A. Das, S. Baki, A. El Aassal, R. Verma and A. Dunbar, "SoK: A Comprehensive Reexamination of Phishing Research From the Security Perspective," in IEEE Communications Surveys Tutorials, vol. 22, no. 1, pp. 671-708, First quarter 2020, doi: 10.1109/COMST.2019.2957750.

[4] M. Korkmaz, O. K. Sahingoz and B. Diri, "Feature Selections for the Classification of Webpages to Detect Phishing Attacks: A Survey," 2020 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), Ankara, Turkey, 2020, pp. 1-9, doi: 10.1109/HORA49412.2020.9152934.

[5] S. Marchal, J. François, R. State and T. Engel, "PhishScore: Hacking phishers' minds," 10th International Conference on Network and Service Management (CNSM) and Workshop, Rio de Janeiro, 2014, pp. 46-54, doi: 10.1109/CNSM.2014.7014140.

[6] A. K. Jain and B. B. Gupta, "Towards detection of phishing websites on client-side using machine learning based approach," Telecommunication Systems, vol. 68, 2018, pp. 687-700, doi: 10.1007/s11235-017-0414-0.

[7] M. Weedon, D. Tsaptsinos and J. Denholm-Price, "Random forest explorations for URL classification," 2017 International Conference On Cyber Situational Awareness, Data Analytics And Assessment (Cyber SA), London, 2017, pp. 1-4, doi: 10.1109/CyberSA.2017.8073403.

[8] E. Buber, B. Dırı and O. K. Sahingoz, "Detecting phishing attacks from URL by using NLP techniques," 2017 International Conference on Computer Science and Engineering (UBMK), Antalya, 2017, pp. 337-342, doi: 10.1109/UBMK.2017.8093406.

[9] C. Liu, L. Wang, B. Lang and Y. Zhou, "Finding effective classifier for malicious URL detection," Proceedings of the 2018 2nd International Conference on Management Engineering, Software Engineering and Service Sciences, 2018, pp. 240-244, doi: 10.1145/3180374.3181352.

[10] Rakesh R, Kannan A, Muthurajkumar S, Pandiyaraju V and SaiRamesh L, "Enhancing the precision of phishing classification accuracy using reduced feature set and boosting algorithm," 2014 Sixth International Conference on Advanced Computing (ICoAC), Chennai, 2014, pp. 86-90, doi: 10.1109/ICoAC.2014.7229752.

[11] M. Aydin, I. Butun, K. Bicakci and N. Baykal, "Using Attribute-based Feature Selection Approaches and Machine Learning Algorithms for Detecting Fraudulent Website URLs," 2020 10th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 2020, pp. 0774-0779, doi: 10.1109/CCWC47524.2020.9031125.

[12] D. R. Ibrahim and A. H. Hadi, "Phishing Websites Prediction Using Classification Techniques," 2017 International Conference on New Trends in Computing Sciences (ICTCS), Amman, 2017, pp. 133-137, doi: 10.1109/ICTCS.2017.38.

[13] J. James, Sandhya L. and C. Thomas, "Detection of phishing URLs using machine learning techniques," 2013 International Conference on Control Communication and Computing (ICCC), Thiruvananthapuram, 2013, pp. 304-309, doi: 10.1109/ICCC.2013.6731669.

[14] C.B. Cebi, F.S. Bulut, H. Firat, O.K. Sahingoz and G. Karatas, "Deep Learning Based Security Management of Information Systems: A Comparative Study", Journal of Advances in Information Technology Vol 11 (3), 2020.

[15] A. Priya and E. Meenakshi, "Detection of phishing websites using C4.5 data mining algorithm," 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), Bangalore, 2017, pp. 1468-1472, doi: 10.1109/RTEICT.2017.8256841.

[16] K. P. S. Attwal and A. S. Dhiman, "Exploring data mining tool-Weka and using Weka to build and evaluate predictive models," Advances and Applications in Mathematical Sciences, vol. 19, 2020, pp. 451-469.

[17] B. Geyik and M. Kara, "Severity Prediction with Machine Learning Methods," 2020 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), Ankara, Turkey, 2020, pp. 1-7, doi: 10.1109/HORA49412.2020.9152601.

[18] E. Buber, Ö. Demir and O. K. Sahingoz, "Feature selections for the machine learning based detection of phishing websites," 2017 International Artificial Intelligence and Data Processing Symposium (IDAP), Malatya, 2017, pp. 1-5, doi: 10.1109/IDAP.2017.8090317.

[19] A. Rawat, "A Review on Python Programming," International Journal of Research in Engineering, Science and Management, vol. 3, 2020, pp. 8-11.

[20] U. Ozker and O. K. Sahingoz, "Content Based Phishing Detection with Machine Learning," 2020 International Conference on Electrical Engineering (ICEE), Istanbul, Turkey, 2020, pp. 1-6, doi: 10.1109/ICEE49691.2020.9249892.

[21] R.S. Rao, T. Vaishnavi and A.R. Pais "CatchPhish: detection of phishing websites by inspecting URLs", Journal of Ambient Intelligence and Humanized Computing 11, 2020, pp. 813–825, doi: 10.1007/s12652-019-01311-4