Phishing Detection using Machine Learning based URL Analysis: A Survey

Arathi Krishna V*, Anusree A, Blessy Jose, Karthika Anilkumar, Ojus Thomas Lee#
Department of Computer Science and Engineering,
College of Engineering Kidangoor
Kottayam, India

***Abstract*—As we have moved most of our financial, work related and other daily activities to the internet, we are exposed to greater risks in the form of cybercrimes. URL based phishing attacks are one of the most common threats to the internet users. In this type of attack, the attacker exploits the human vulnerability rather than software flaws. It targets both individuals and organizations, induces them to click on URLs that look secure, and steal confidential information or inject malware on our system. Different machine learning algorithms are being used for the detection of phishing URLs, that is, to classify a URL as phishing or legitimate. Researchers are constantly trying to improve the performance of existing models and increase their accuracy. In this work we aim to review various machine learning methods used for this purpose, along with datasets and URL features used to train the machine learning models. The performance of different machine learning algorithms and the methods used to increase their accuracy measures are discussed and analysed. The goal is to create a survey resource for researchers to learn the current developments in the field and contribute in making phishing detection models that yield more accurate results.**

Keywords—Phishing; URL features; machine learning; phishing detection

I. INTRODUCTION

The year 2020 saw peoples' life being completely dependent on technology due to the global pandemic. Since digitalization became significant in this scenario, cyber criminals went on an internet crime spree. Recent reports and researches point to an increased number of security breaches that costs the victims a huge sum of money or disclosure of confidential data. Phishing is a cybercrime that employs both social engineering and technical subterfuge in order to steal personal identity data or financial account credentials of victims[1]. In phishing, attackers counterfeit trusted websites and misdirect people to these websites, where they are tricked into sharing usernames, passwords, banking or credit card details and other sensitive credentials. These phishing URLs may be sent to the consumers through email, instant message or text message. According to the FBI crime report 2020, phishing was the most common type of cyber attack in 2020 and phishing incidents nearly doubled from 114,702 in 2019 to 241,342 in 2020[2]. The Verizon 2020 Data Breach Investigation Report states that 22% of data breaches in 2020 involved phishing[3].

The number of phishing attacks as observed by the Anti-Phishing Work Group (APWG) grew through 2020, doubling

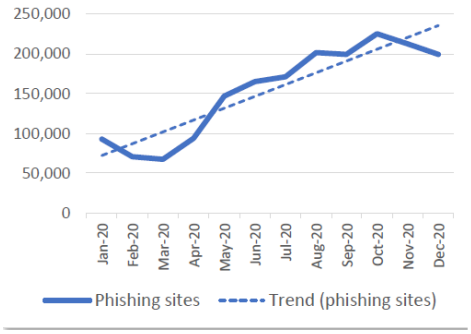

Fig. 1. Phishing Activity – 2020 [1]

over the course of the year. In the 4th quarter of 2020, it was found that phishing attacks against financial institutions were the most prevalent. Phishing attacks against SaaS and Webmail sites were down and attacks against E-commerce sites escalated, while attacks against media companies decreased slightly from 12.6% to 11.8%[1]. In light of the prevailing pandemic situation, there have been many phishing attacks that exploit the global focus on Covid-19. According to WHO, many hackers and cyber scammers are sending fraudulent emails and WhatsApp messages to people, taking advantage of the coronavirus disease[4]. These attacks are coming in the form of fake job offers, fabricated messages from health organizations, covid vaccine themed phishing and brand impersonation.

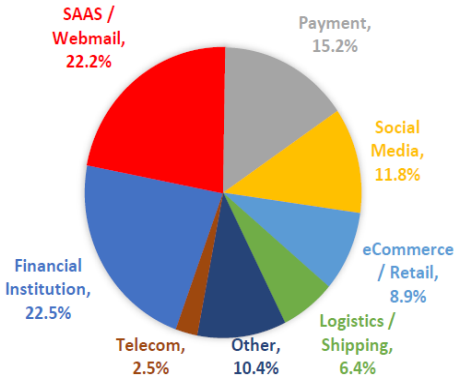

Fig. 2. Most targeted industries, 4Q 2020 [1]

In the next section, various phishing detection approaches are analysed. The most common machine learning algorithms used in case of machine learning based approach are discussed.

II. BACKGROUND

A. Phishing Detection

A URL based phishing attack is carried out by sending malicious links, that seems legitimate to the users, and tricking them into clicking on it. In phishing detection, an incoming URL is identified as phishing or not by analysing the different features of the URL and is classified accordingly. Different machine learning algorithms are trained on various datasets of URL features to classify a given URL as phishing or legitimate.

B. Phishing Detection Approaches

In List Based approach, there are two lists, called whitelist and blacklist to classify legitimate and phishing URLs respectively. In [5], access to websites takes place only if the URL is in the whitelist. In [6] blacklist is used. In Heuristic Based approach, the structure of a phishing URL is analysed. A pattern of URLs that were previously classified as phishing is created. URLs are classified according to their compliance with this pattern. The methods used to process the features of the URL plays a significant role in classifying websites accurately [7].

Visual similarity Based approach works by comparing the visual similarity of the website pages. Websites are classified as phishing or not by taking a server side view of them as in [8]. These two data are then compared with image processing techniques. Fake web pages are designed very close to the original ones and it is easier to notice minor differences with image processing techniques, as users cannot notice them easily.

Content Based approach analyses the pages' content. This method extracts features from page contents and third-party services like search engines and DNS servers. In [9] authors proposed a detection method by specifying weights to the words that draw out from URLs and HTML contents. The words might include brand names that attackers use in the URL to make it look like a real one. Weights are specified according to their presence at different positions in URLs. The most probable words are chosen and then sent to Yahoo search to return the domain name with the highest frequency between the top 30 outcomes. The owners of the domain name are compared to decide if the website is phishing or not. In [10], they utilized a logo image to find the identity of web pages by matching real and fake web pages.

Fuzzy Rule based approach allows processing of ambiguous variables, then integrates human experts to classify those variables and relations between them. It is used to classify web pages based on the level of phishing that appears in the pages by employing a specific group of metrics and predefined rules[11]. From the experimental results in the paper, for fuzzy logic systems, lower number of features leads to higher accuracy. If a fuzzy logic algorithm is affected by irrelevant features, the effectiveness of the classifier will decrease and vice-versa.

In Machine Learning based approach, machine learning models are created to classify a given URL as phishing or not using supervised learning algorithms. Different algorithms are trained on a dataset and then tested to learn the performance of each model. Any variations in the training data directly affects the performance of the model. This approach provides efficient techniques with high-performance for detecting phishing. This is a significant field of research and there are many papers that discuss machine learning based phishing detection.

C. Machine Learning Algorithms

There are several machine learning algorithms such as Naive Bayes, Decision Tree, Random Forest, Support Vector Machine, Logistic Regression and K-Nearest-Neighbor for detecting phishing websites. This is a very popular approach that has proved to be very efficient and accurate compared to other methods.

III. LITERATURE REVIEW

In this section, few of the research works that deploy the above mentioned algorithms are reviewed and their results are summarized.

In the paper [12], the authors Rishikesh Mahajan and Irfan Siddavatam chose three algorithms for classification– Decision Tree, Random Forest and Support Vector Machine. Their dataset contained 17,058 benign URLs and 19,653 phishing URLs collected from Alexa website and PhishTank respectively, with 16 features each. The dataset was divided into training and testing set in the ratios 50:50, 70:30 and 90:10 respectively. The accuracy score, false negative rate and false positive rate were considered as performance evaluation metrics. They achieved 97.14% accuracy for Random Forest algorithm with the lowest false negative rate. The paper concluded that accuracy increases when more data is used for training.

The study conducted by Jitendra Kumar et al. in [13] trained different classifiers like Logistic Regression, Naive Bayes Classifier, Random Forest, Decision Tree and K-Nearest Neighbor based on the features extracted from the lexical structure of the URL. They created the dataset of URLs in such a way that it solved the issues of data imbalance, biased training, variance and overfitting. The dataset contained an equal number of labeled phishing and legitimate URLs, and was further split in the ratio 7:3 for training and testing. All the classifiers had almost the same AUC (area under ROC curve), but the Naive Bayes Classifier turned out to be more suitable as it had the highest AUC value. Naive Bayes achieved the highest accuracy of 98% with a precision=1, recall=0.95 and F1-score=0.97.

Mehmet Korkmaz et al. proposed in [14] a machine-learning based phishing detection system by using 8 different algorithms on three different datasets. The algorithms used were Logistic Regression (LR), K-Nearest Neighbor(KNN), Support Vector Machine (SVM), Decision Tree (DT), Naive Bayes (NB), XGBoost, Random Forest (RF) and Artificial Neural Network (ANN). It was observed that the models using LR, SVM and NB have low accuracy rate. In terms of training time, NB, DT, LR and ANN algorithms gave better results. They concluded that RF algorithm or ANN algorithm may be used because of less training time along with a high accuracy rate.

Mohammad Nazmul Alam et al. [15] proposed a system to detect phishing attacks using Random Forest and Decision Tree. The Kaggle dataset with 32 features was used along with

feature selection algorithms like principal component analysis (PCA). Feature selection reduces redundancy of data that is irrelevant or unnecessary in the dataset. The proposed model used REF, Relief-F, IG and GR algorithm for feature selection before applying PCA. Random Forest achieved an accuracy of 97%. It had less variance, and it could handle the over-fitting problem.

Abdulhamit Subasi et al. in [16] presented an intelligent phishing detection system using UCI dataset. Different machine learning tools namely, Artificial Neural Networks (ANN), K-Nearest Neighbor (K-NN), Support Vector Machine (SVM), C4.5 Decision Tree, Random Forest (RF), and Rotation Forest (RoF) were used as classifiers for detection of phishing websites. The performance of proposed RF classifier was higher than the others in terms of accuracy, F-measure and AUC. RF was faster, robust and more accurate than the other classifiers.

The rest of the paper is organized as follows: In section IV, the characteristics of the datasets that are used for training machine learning models are discussed. Section V explains the feature extraction process. The parameters used for performance evaluation of algorithms are discussed in Section VI. The observations obtained from the survey are pointed out in Section VII. Section VIII concludes the paper.

IV. DATASETS

Usually, the phishing website data is collected from Phish Tank or OpenPhish. PhishTank.com is a website where phishing URLs are detected and can be accessed via API call. Their data is used by companies like McAfee, Kaspersky, Mozilla and APWG. Since it does not store the content of webpages, it is a good source for URL based analysis[14]. The legitimate sites are generally collected from Alexa's top-ranking websites database or from common-crawl. There are publicly available datasets like the UCI machine learning repository dataset used in [16] which contains 11,055 records, each record having 31 features and the Kaggle phishing dataset used in [15].

V. FEATURE EXTRACTION

URLs have certain characteristics and patterns that can be considered as its features. The Fig. 3 shows the relevant parts of a typical URL.

In case of URL based analysis for designing machine learning models, we need to extract these features in order to form a dataset that can be used for training and testing. There are four categories of features that are most commonly considered for feature extraction as in [18]. They are as follows:

1) Address Bar based features

2) Abnormal based features

3) HTML and JavaScript based features

4) Domain based features

TABLE I. SUMMARY OF LITERATURE REVIEW

Paper	Approach	Conclusion	Accuracy
[14]	8 different algorithms are applied on three different datasets making use of 48 features.	RF has the highest accuracy, on all three datasets. ANN is also preferred.	Dataset 1 : 94.59% Dataset 2 : 90.5% Dataset 3 : 91.26%
[12]	Dataset is split into training and testing set in 50:50, 70:30 and 90:10 ratios respectively. DT, RF and SVM classifiers are applied.	RF has better accuracy with least false negative rate. Accuracy increases when more data is used for training.	50:50 split ratio : 96.72% 70:30 split ratio : 96.84% 90:10 split ratio : 97.14%
[13]	A balanced dataset is used to train LR, NB RF, DT, k-NN classifiers based on features extracted from the lexical structure of a URL.	The RF and NB classifiers have better accuracies among all classifiers. In terms of AUC, Gaussian Naive Bayes has a slightly higher value of 0.991	Random Forest :98.03% Gaussian Naive Bayes : 97.18%
[16]	Accuracy, F-measure and AUC are used to evaluate performance of classifiers ANN, k-NN SVM, C4.5 DT, RF and RoF on UCI dataset.	RF produces reliable results in terms of Accuracy, F-measure and AUC. It is faster, robust and more accurate.	Random Forest :97.36%
[15]	Uses feature selection algorithms like PCA before applying classifiers RF and DT.	RF has less variance and it could handle the problem of overfitting.	Random Forest :97%

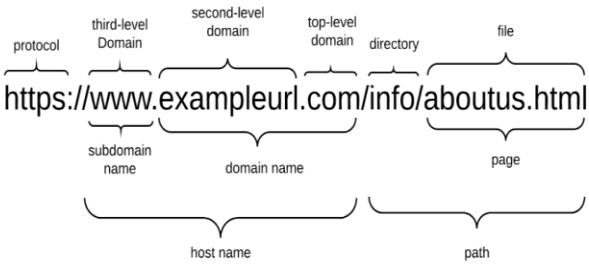

Fig. 3. Structure of a URL[6]

The features used in UCI dataset with their column names are given in Table II below. They are represented with binary values that indicate if the property is present or absent. Some features have 3 values that represent its strength ranging from low, medium and high. At the end, there is a result which

TABLE II. FEATURES OF URL -SET I

ID		Name of Features
1	Address Bar Based Features	Having_ip_Address
2		URL_Length
3		Shortening_Service
4		Having_At_Symbol
5		Double_slash_redirecting
6		Prefix_suffix
7		Having_sub_Domain
8		SSLfinal_State
9		Domain_registration_length
10		Favicon
11		Port
12		HTTPS_token
13	Abnormal Based Features	Request_URL
14		URL_of_Anchor
15		Links_in_tags
16		SFH
17		Submitting_to_email
18		Abnormal_URL
19	HTML & JavaScript Based Features	Redirect
20		On_Mouseover
21		Right_Click
22		popUpWidnow
23		Iframe
24	Domain Based Features	Age_of_domain
25		DNS_Record
26		Web_traffic
27		Page_Rank
28		Google_Index
29		Links_poiniting_to_page
30		Statistical_report
31		Result

identifies the true nature of the URL, -1 if it is a phishing site and 1 if it is a legitimate site.

The paper [14] examined URL features into hostname, domain and path sections. The author used the best 48 features out of the 58 they obtained, in order to perform URL classification in a short time without content analysis and without using third party services. The features used for that study are listed in Table III.

TABLE III. FEATURES OF URL -SET II [14]

#		Name		#	Name	#		Name
1	NUMBER OF	Words	NUMBER OF	17	Underscore	33	LENGTH OF	Path
2		Url Paths		18	Dots in Host	34		Sub domain
3		Digits		19	Dots in Path	35		Url
4		Ampersand		20	Hyphen in Host	36		Domain Name
5		Sensitive Words	HAS	21	Url without www	37		Longest Word
6		“?”		22	Query	38		Parameters
7		Special Chars		23	Character Repetition	39		Average Word
8		Punctuation		24	Https Protocol	40		Shortest Word
9		Dots in Sub Domain		25	Digits in Domain name	41		Longest Word in Host name
10		TId in Paths		26	Ip Address	42		Host
11		Subdomain		27	subdomain	43	RATIO OF	Url/Path
12		Digits in Host		28	“www” or “com”	44		Vowel/Consonant
13		Dots		29	“@”	45		Digit/ Letter
14		Words in Host Name		30	Hyphen in Url	46		Longest/ Shortest Word Length
15		Hyphen in Path		31	Suffix	47	-	STD of Words Length
16		“=”		32	Redirected	48	-	Port Number

VI. PERFORMANCE EVALUATION METRICS

To evaluate the efficiency of a system, we use certain parameters. For each machine learning model, we calculate the Accuracy, Precision, Recall, F1 Score and ROC curve to determine its performance. Each of these metrics is calculated based on True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN).

In the case of URL classification, True Positive (TP) is the number of phishing URLs that are correctly classified as phishing. True Negative (TN) is the number of legitimate URLs that are correctly classified as legitimate. False Positive (FP) is the number of legitimate URLs that are classified as phishing. False Negative (FN) is the number of phishing URLs that are classified as legitimate. These values are summarized in Table IV called Confusion Matrix.

TABLE IV. CONFUSION MATRIX FOR PHISHING DETECTION

	Predicted Phishing	Predicted Legitimate
Actual Phishing	TP	FN
Actual Legitimate	FP	TN

Precision is the number of URLs that are actually phishing out of all the URLs predicted as phishing. It measures the classifier's exactness. The formula to calculate precision is given by Equation (1) below.

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

Recall is the number of URLs that the classifier identified as phishing out of all the URLs that are actually phishing. It is also called sensitivity or true positive rate. It is an important measure and should be as high as possible. The formula to calculate recall is given by Equation (2) below.

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

F1-Score is the weighted average of precision and recall. It is used to measure precision and recall at the same time. The formula to calculate F1-Score is given by Equation (3) below.

$$F1Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (3)$$

Accuracy is the number of instances that were correctly classified out of all the instances in the test data. The formula to calculate accuracy is given by Equation (4) below.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

Receiver Operating Characteristic (ROC) curve is an important evaluation metric for binary classification models. The curve is plotted with True Positive Rate (TPR) on the y-axis and False Positive Rate (FPR) on the x-axis. The Area Under the ROC curve shows how well a classifier is able to distinguish between phishing and legitimate URLs. The formula to calculate FPR and TPR are given by Equation (5) and Equation (6) respectively.

$$FPR = \frac{FP}{FP + TN} \quad (5)$$

$$TPR = \frac{TP}{TP + FN} \quad (6)$$

VII. OBSERVATIONS

Phishing attacks are constantly evolving and the cyber world is hit by new types of attacks often. Hence a particular detection approach or algorithm cannot be tagged as the best one giving exact results. Through the literature survey, it is evidently visible that Random Forest gives better results in most scenarios. But then the performance of each algorithm varies depending on the dataset used, train-test split ratio, feature selection techniques applied etc. Researchers prefer to create machine learning models that perform phishing detection with best value for evaluation parameters and least training time. Therefore, the future works should focus on these aspects of phishing detection.

VIII. CONCLUSION

Phishing detection is now an area of great interest among the researchers due to its significance in protecting privacy and providing security. There are many methods that perform phishing detection by classification of websites using trained machine learning models. URL based analysis increases the speed of detection. Furthermore, by applying feature selection algorithms and dimensionality reduction techniques, we can reduce the number of features and remove irrelevant data. There are many machine learning algorithms that perform classification with good performance measures. In this paper, we have done a study of the process of phishing detection and the phishing detection schemes in the recent research literature. This will serve as a guide for new researchers to understand the process and to develop more accurate phishing detection systems.

REFERENCES

[1] Anti-phishing Working Group (APWG) Phishing Activity Trends Report 4th quarter 2020, https://docs.apwg.org/reports/apwg trends report q4 2020.pdf

[2] FBI Internet Crime Report 2020, https://www.ic3.gov/Media/PDF/AnnualReport/2020 IC3Report.pdf

[3] Verizon 2020 Data Breach Investigation Report, https://enterprise.verizon.com/resources/reports/2020-data-breachinvestigations- report.pdf

[4] World Health Organization, Communicating for Health, Cyber Security, https://www.who.int/about/communications/cyber-security

[5] Ye Cao, Weili Han, and Yueran Le, "Anti-phishing based on automated individual white-list," Proceedings of the 4th ACM workshop on Digital identity management-DIM 08, pp. 51-60, 2008

[6] M. Sharifi, and S. H. Siadati, "A phishing sites blacklist generator," 2008 IEEE/ACS International Conference on Computer Systems and Applications, pp. 840-843, 2008

[7] N. Abdelhamid, A. Ayesh, and F. Thabtah, "Phishing detection based associative classification data mining," Expert Systems with Applications, vol. 41, no.13, pp. 5948-5959, 2014

[8] L. Wenyin, G. Huang, L. Xiaoyue, Z. Min, and X. Deng, "Detection of phishing webpages based on visual similarity," Special interest tracks and posters of the 14th international conference on World Wide Web-WWW 05, pp. 1060-1061, 2005

[9] C. L. Tan, K. L. Chiew et al., "Phishing website detection using url assisted brand name weighting system," 2014 International Symposium on Intelligent Signal Processing and Communication Systems(ISPACS), IEEE, pp. 054-059, 2014

[10] K. L. Chiew, E. H. Chang, W. K. Tiong et al., "Utilisation of website logo for phishing detection," Computers & Security, vol. 54, pp. 16-26, 2015

[11] K. M. kumar, K. Alekhya, "Detecting phishing websites using fuzzy logic," International Journal of Advanced Research in Computer Engineering Technology(IJARCET), vol. 5, no. 10, 2016

[12] Rishikesh Mahajan, and Irfan Siddavatam, "Phishing website detection using machine learning algorithms," International Journal of Computer Applications(0975-8887), vol. 181, no. 23, 2018

[13] Jitendra Kumar, A. Santhanavijayan, B. Janet, Balaji Rajendran, and Bindhumadhava BS, "Phishing website classification and detection using machine learning," International Conference on Computer Communication and Informatics(ICCCI), 2020

[14] Mehmet Korkmaz, Ozgur Koray Sahingoz, Banu Diri, "Detection of phishing websites by using machine learning-based URL analysis," 11nth International Conference on Computing, Communication and Networking Technologies(ICCCNT), 2020

[15] Mohammad Nazmul Alam, Dhiman Sarma et al., "Phishing attacks detection using machine learning approach," 3rd International Conference on Smart Systems and Inventive Technology(ICSSIT), 2020

[16] Abdulhamit Subasi, Esraa Molah, Fatin Almkallawi, Touseef J. Chaudhery, "Intelligent phishing website detection using Random Forest classifier," International Conference on Electrical and Computing Technologies and Applications(ICECTA), 2017

[17] Structure of a URL – image, https://towardsdatascience.com/phishingdomain-detection-with-ml-5be9c99293e5

[18] Rami M. Mohammad, Fadi Thabtah, Lee McCluskey, "Phishing websites features"