Webpage Classification for Detecting Phishing Attack

[1] Omejevwe Efe-odenema, [2] Dr. Jitendra Jaiswal

[1]Student, [2]Associate Professor,

1Department of Computer Science and technology, Jain (Deemed to be University), Bengaluru, India.

Abstract: Despite numerous research efforts, phishing attacks remain prevalent and highly effective in luring unsuspecting users to reveal sensitive information, including account credentials and social security numbers. In this paper, we propose the use of three machine algorithm to help in the detection of phishing attacks. Machine learning algorithm has been popular over the years for implementing and solving different problems. Different features were observed and approximately 112 features were used from 88,648 dataset, gotten from Vrbancic UC Machine Learning Repository database. Through the use the algorithms, high accuracy were gotten especially after the application of PCA feature selection.

***IndexTerms* - principal component analysis, Machine Learning Framework.**

I. Introduction

Recent advances in technology, which have to enable online users to have a better experience while carrying out their day to day activities with ease, this has also created an avenue for criminals to carry out their illegal activities. One of the most recent of them is the phishing attack. This is done by stealing private information both personal identity data and financial account credentials such as credit card details, password to carry out fraudulent activities. The attacker usually does this by using a phishing URL and an email to deceive users. Phishing employs the use of social engineering to fool users that they are dealing with a legitimate source. They lead the user into phish websites into divulging financial details.[1]

There has been recent trend according to Phishing Activity Trends Report, during the first quarter of 2020, during the time of COVID-19, cybercriminals launched numerous attacks relating to COVID-19 as they launched phishing and malware attacks on health workers, healthcare facilities and recently unemployed as they were vulnerable, having just lost their means of livelihood. There was a slight increase in phishing attacks. [2]Recorded that 165,772 attacks took place as against 162,155 as of the end of the 4th quarter of 2019.

A new trend was discovered as most phishing website now uses an SSL certificate for security, this makes a phishing attack more cumbersome to curb. Phishing targeting webmail and software as a service user continued as the biggest category of phishing. Zoom technology, an online platform that is used by business owners for meetings was used by the Cybercriminal to carry out crime during the pandemic by creating fake video-conferencing meetings, leading the user into a fake website where their credential was stolen. The figure below shows the statistic and trends of phishing attack in the first quarter of 2020.

	January	**February**	**March**
Number of unique phishing Web sites detected	54,926	49,560	60,286
Number of unique phishing e-mail reports (campaigns) received by APWG from consumers	52,407	43,270	44,008
Number of brands targeted by phishing campaigns	374	331	344

fig. 1.1 statistical highlights for 1st quarter 2020

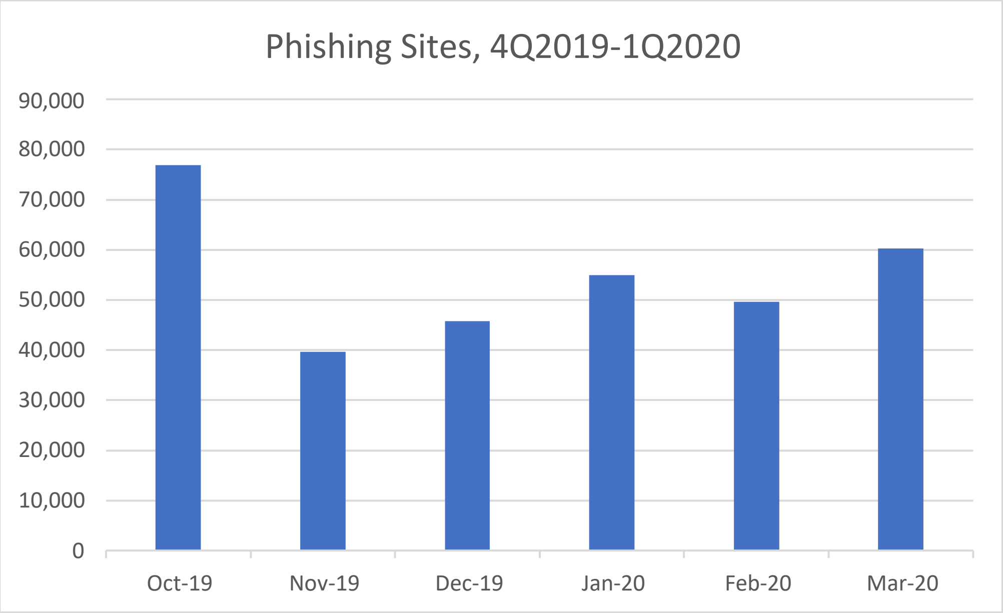

fig. 1.2 phishing activity trends report 1st quarter 2020

1.1 Types of phishing

1.1.1 Spear phishing

Spear phishing as against phishing is targeted at a specific user, as the emails are crafted to suit that user. The email messages are specifically personalized to a particular victim, which purportedly comes with familiar personal information in order to gain the trust of the user. The message is well thought out and usually requires some time to achieve. The aim of this type of phishing attack is to gain the trust of the user to increase their chances of deceiving the user. Due to how personal this type of email is, it is very difficult to combat.

1.1.2 Whaling

Attacker utilizes spear phishing to target high profile employees in an organization. In order to carry out this attack successfully, attackers compel the user with some sense of urgency. The goal is to capture some sensitive information that the high profile user has access to. [3]

1.1.3 Clone phishing

Replicating a legitimate email that was sent earlier to the user, but uses a spoof email address. The link is replaced with malicious links, which oftentimes contain malware or put their personal data at the risk of being stolen. The user most often would not notice the difference because the email is very familiar and think its true and valid. This is very risky [4]

Most phish URL has similar features with a legitimate URL; this paper proposes the use of URL based phishing detection system. This study looks into the various features contained in both phishing and legitimate URL and also links that are found on the webpage and use a machine-learning algorithm to train the detection model for automatic detection.

In this study, we use the three machine learning algorithms; Support Vector Machine, Random Forest algorithm, and Naïve Bayes algorithm to help classify the detection model. Before recent times, most researchers use the old method like a blacklist for classification. This method is not only time consuming but also has to be done manually as blacklist involves the use of a signature database where features of phish URL are stored and verified. The stored features are then mapped to see if there is a match.

One of the drawbacks is that it cannot detect zero-day attacks due to delay in verification. In order to reduce the time and cost of the blacklist, a machine learning algorithm was used which resulted in better results. Various URL features were put into consideration especially the top features which gave better results. In the following section, we take a look at the literature review, methodology, result, and analysis then conclusion.

II. LITERATURE REVIEW

2.1 Content-based

CANTINA[5], the technique used by this author, he used TF-IDF technology to compare whether the current domain appears in the top 30 domains of the search result, some are legitimate sites, if not some are not. The technique takes the first top 5 nouns of the TF-IDF of the browsed webpage. The following features were used together with Google assisted TF-IDF; the age of the domain, known images, suspicious URL, suspicious links, and forms.

2.2 Phishing Detection by Blacklists

Blacklist is frequently updated from a previous phishing attack. Before being stored on the database, they are verified by 5 or more persons. One of the drawbacks of the blacklist is that they usually do not provide for zero-day phishing attacks. According to [6] blacklist has only detected only 12% of zero-day attacks. The study also shows that 47% to 83% of phishing URL was blacklisted after 12 hours, meanwhile phishing campaigns end within 2hours. This delay causes lots of drawbacks to this technique.

Due to the limitations of blacklist [7] proposes a PhishNet method, which involves the processing of the parent blacklisted URL and producing multiple variations to produce a new URL(which s the children element). Each parent URL will produce 3,210 variations each with a different top-level domain from the parent element.

2.3 Phishing Detection by Heuristics

The work from [8] is based on authentication which focuses on non-piped phishing attempts. PhishGuard follows the following steps to test a suspected page. When the user visits a page, if the visited page sends an authentication request, and if the user submitted the authentication form, then PhishGuard starts its testing procedures. PhishGuard would send the same user ID, followed by a random password that does not match the real password, for random n times.

If the page responded with HTTP 200 OK message, then it would mean the page is a phishing site and is simply returning fake authentication success messages. If the page responded with HTTP 401 Unauthorized message, then it could possibly mean: The site is a phishing site that blindly responds with failure authentication messages. The site is a legitimate site.

2.4 Phishing Detection By Visual Similarity

Using classification with discriminative keypoint features[9] makes use of content presentation instead of the actual code. For example, a phishing website that mimics a legitimate website by displaying similar content using IMG HTML tags instead of using HTML, will be able to bypass anti-phishing mechanisms, because image contents are invisible to them. [9] The Proposed solution for this is for the web browser to take snapshots of every suspected site, then RGB channels are then converted to grayscale channels by taking the mean of red, green,and blue values.

The resultant Grayscale image is analyzed to find key features or salient points. The mechanism used to detect salient points is through detection of corners. This proposal uses the Harris-Laplace algorithm to detect corners in an image. The advantage of Harris-Laplace algorithm is its accuracy in conditions of different resolutions and rotations.

2.5 Machine Learning-Based Phishing Detection

Due to high labor cost and time consumption, the use of artificial intelligence has been made popular over the years, and there has been an increase in the use of machine learning algorithms in detection phishing attacks.

One of the many methods of machine learning is hybrid features, which is a combination of natural language processes and word vector features. In [10] results, the random forest algorithm with hybrid feature set had an accuracy of 96.36%. [10] uses hyperlinks found in HTML source as data features that were extracted from the client-side. This was done to achieve to detect and provide real-time solutions.

In [11] three supervised learning algorithms were used, namely, Adaline network, backpropagation network, support vector machine. About 15 features of these websites were extracted. After the application of the various algorithms, the Adaline network with the support vector machine got the highest accuracy. According to [12], they used features from X.509 public-key certificates to detect phishing websites. the features used are, NotBefore, NotAfter, Date- Downloaded,Issuer, Subject and Domain Name, etc The features were collected from PhishTank entries, HTTPS inclusion was not considered. The machine learning algorithm used are; Decision Trees, Random Forest, Naïve Bayes, and logistic regression. Their best prediction came from the random forest with accuracy of 95.5%.

They [13] proposed a method that makes a prediction based on the features of the URL and the ranking of the site. Alexa Reputation was calculated using the URL. Root mean squared error was also calculated to find the accuracy of different values. The accuracy rate of this technique is 97.16% with a threshold of 0.4. Their shortcoming were the important features like status bar customization, submission of information, website forwarding that were neglected which may lead to the wrong prediction.

III. MATERIAL AND METHODOLOGY

3.1 The Phishing Dataset

To carry out this research, a large dataset, approximately 88,648 containing the 112 features was collected from the websites in Vrbancic UC Machine Learning Repository database. Following are some the features considered for machine learning based phishing detection; IP address, long URL length, shortened URL, having @ symbol, double slash redirecting, prefix suffix, Iframe tag, anchor tag, disabling right click, age of domain, record, HTTPS with SSL, domain registration length, website traffic, statistical based report feature, using non standard port, abnormal URL, sub domain and multi sub domains, request URL, server form handle, submitting information to email, website forwarding, pageRank, using pop-up window, Google index, number of links pointing to page and many more.

3.2 Machine Learning Algorithms

The main contribution of this paper is to propose an adaptive detection system which can detect phishing websites from URL using a set of the outstanding features. Since we design our system in python, the machine learning kit we use is scikit-learn, a common used machine learning library in python.

Classification is used for this problem since it is used to determine the class to which each data sample of the methods belongs, which methods are used when the outputs of input data are qualitative. The purpose is to divide the whole problem space into a certain number of classes. As mentioned in literature studies, the aim of classification is to assign the new samples to classes by using the pre-labeled samples. In our experiments three classification algorithms were tested. This was done using the open source programming language python.

3.2.1 Support vector machine

A support vector machine (SVM) is a supervised machine learning model that uses classification algorithms for two-group classification problems.[14]

3.2.2 Naïve bayes

A Naive Bayes classifier is a probabilistic machine learning model that's used for classification task. The crux of the classifier is based on the Bayes theorem. Using Bayes theorem, we can find the probability of **A** happening, given that **B** has occurred. [15]

3.2.3 Random Forest

Random Forest adds randomness to the generation of decision trees. Instead of relying on one single decision tree to cover the entire dataset and features, this approach selects features and training data randomly from the given sets and constructs a series of decision trees based on these randomly selected inputs.[16]

3.3 Model performance evaluation

Split data into training and testing dataset, which uses 80% for training and 20% for testing. Train and test all possible combination of 112 features dataset to get the strongest features that comes from the accuracy of detection. Execute the final classifier. Feature selection like PCA was applied to combine highly correlated variables by using suitable combinations. Hence it creates new variables in dataset removing highly correlated variables. This newly formed dataset gave the best prediction for random forest algorithm and naïve bayes algorithm, a slight decrease was observed for support vector algorithm after the application of PCA.

IV. RESULT AND DISCUSSION

Table 1 Percentage of The Result Of The Various Algorithm

Algorithm	Accuracy	Error Rate	After PCA
Random Forest	0.9697687535250987	0.030231246474901274	0.9953186689227298
Support Vector Machine	0.9319796954314721	0.06802030456852792	0.863282571912013 5
Naïve Bayes	0.8045121263395375	0.1954878736604625	0.8631697687535251
K-Nearest-Neighbors (KNN)	0.9547659334461365	0.04523406655386353	0.9239706711787931

The accuracy of the classifiers was evaluated when it is trained with different percentage of the dataset. Three machine learning algorithms were applied on the dataset. The results obtained for the three algorithms are mentioned in Table I.
The accuracy of Random Forest Algorithm (RFA) was 0.96%, that of Support Vector Machine (SVM) was 0.93% and Naïve Bayes Algorithm(NBA) was 0.80%. Random forest algorithm gave the best accuracy, this maybe because they perform better with large dataset.

In other to get better performance, principal component analysis was implemented. PCA was used because of how good it is for large pool of datasets, as it helps to reduce number of variables in the data by extracting important features from the data and also reducing the dimension of the data with the aim of retaining as much information as possible. This it does by dividing the features unto two component parts. After applying Principal Component Analysis (PCA), the accuracy increased by a decent margin in Radom forest algorithm. The accuracy for random forest algorithm after applying PCA was 0.99%, that of Support Vector Machine (SVM) however decreased to 0.86% and Naïve Bayes Algorithm(NBA) increased with an accuracy was 0.86%.

Like accuracy, the precision rates too were highest for random forest. The algorithms were first implemented independently, and then after applying PCA. K-Nearest-Neighbors (KNN) was also introduced into the model, the accuracy gotten was 0.95%, and after the application of PCA, it decreased to 0.92%. from this result it is therefore save to say that oly RF and NB had an increase after the application of PCA.

The graph below shows principal component, pc1 and pc2, x and y axis respectively. This is the different classification region. Green color and red color represent the phish and non-phishing classification along the x and y axis. The NB algorithm and SVM similar graphs which represent their outcome after the PCA was applied; in which both of the result was 0.86%. the KNN and RF algorithm gave sharp distinction between both axis.

When compared with [17], better accuracy was gotten in RF and SVM.

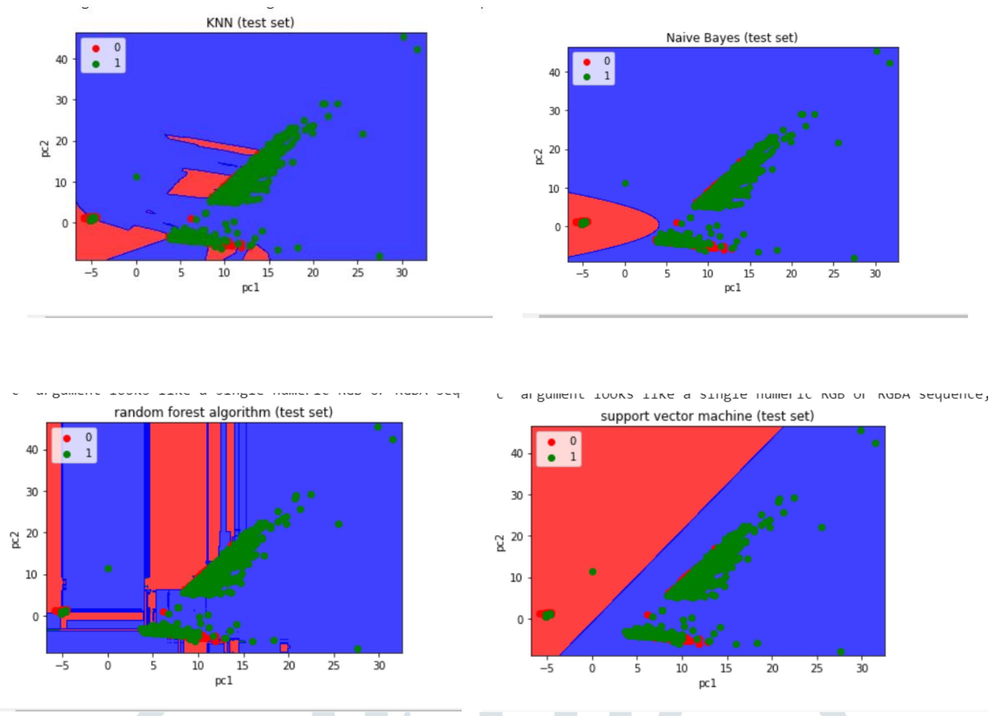

fig 4.1. pca of all four machine learning algorithms

V. Conclusion

This study focused on the classification of web URL into phishing and non-phishing groups, using machine learning algorithms to improve its accuracy efficiency. The use of three machine learning algorithm was considered in hopes of finding a better accuracy from previous study. The accuracy gotten from the random forest was higher compared to naïve bayes and support vector algorithm. Better accuracy was also gotten compared to previous study. The application of PCA was applied for improvement of result. In the future, additional adjustment of datasets features would be considered in other to get more relevant features to get better result especially for naïve bayes algorithm and support vector machine. More study will be focused on other types of phishing methods.

VI. Reference

[1] Che-Yu Wu, Tainan, Taiwan, Cheng-Chung Kuo, Chu-Sing Yang "A Phishing Detection System based on Machine Learning"019 International Conference on Intelligent Computing and its Emerging Applications (ICEA).

[2] Anti-Phishing Working Group Phishing Attack Trends Report 1Q2020, https://docs.apwg.org/reports/apwg_trends_report_q1_2020.pdf

[3] https://www.rapid7.com/fundamentals/whaling-phishing-attacks/

[4] https://www.pentestpeople.com/clone-phishing/

[5] Y. Zhang, J. I. Hong, and L. F. Cranor, "Cantina: a content-based approach to detecting phishing web sites," in Proceedings of the 16th international conference on World Wide Web, ser. WWW '07. New York, NY, USA: ACM, 2007, pp. 639–648.S. Sheng, B. Wardman

[6] G. Warner, L. F. Cranor, J. Hong, and C. Zhang, "An empirical analysis of phishing blacklists," in Proceedings of the 6th Conference in Email and Anti-Spam, ser. CEAS'09, Mountain view, CA, July 2009.

[7] P. Prakash, M. Kumar, R. R. Kompella, and M. Gupta, "Phishnet: predictive blacklisting to detect phishing attacks," in INFOCOM'10: Proceedings of the 29th conference on Information communications. Piscataway, NJ, USA: IEEE Press, 2010, pp. 346–350.

[8] P. Likarish, D. Dunbar, and T. E. Hansen, "Phishguard: A browser plug-in for protection from phishing," in 2 nd International Conference on Internet Multimedia Services Architecture and Applications, 2008. IMSAA 2008, 2008, pp. 1 – 6

[9] K.-T. Chen, J.-Y. Chen, C.-R. Huang, and C.-S. Chen, "Fighting phishing with discriminative keypoint features," Internet Computing, IEEE, vol. 13, no. 3, pp. 56 –63, may-june 2009

[10] K. Jain and B. Gupta, "A machine learning based approach for phishing detection using hyperlinks information," Journal of Ambient Intelligence and Humanized Computing, pp. 1-14, 2018.

[11] Mustafa Aydin, Nazife Baykal, "Feature Extraction and Classification Phishing Websites Based on URL", in IEEE International Conference on Communications and Network Security (CNS), pp.769 – 770, 2015.

[12] Rami M. Mohammad, FadiTabah, Lee McCluskey, "Phishing Website Features" Unpublished. Available via http://eprints.hud.ac.uk/24330/6/RamiPhishing_Websites _Features.pdf.

[13] R. M. Mohammad, F. Thabtah, and L. McCluskey, "Predicting phishing websites based on self-structuring neural network," Neural Comput &Applic, vol. 25, no. 2, pp. 443 458, Aug. 2014.

[14] https://monkeylearn.com/blog/introduction-to-support-vector-machines-svm/

[15] https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c

[16] https://towardsdatascience.com/random-forest-classification-and-its-implementation-d5d840dbead0

[17] Hieu Nguyen and Thai Nguyen "Machine Learning Based Phishing Web Sites Detection"