Phishing detection system using machine learning classifiers

Nur Sholihah Zaini[1], Deris Stiawan[2], Mohd Faizal Ab Razak[3], Ahmad Firdaus[4], Wan Isni Sofiah Wan Din[5], Shahreen Kasim[6], Tole Sutikno[7]
[1,3,4,5]Faculty of Computer Systems & Software Engineering, University Malaysia Pahang, Malaysia
[2]Department of Computer Engineering, Universitas Sriwijaya, Indonesia
[6]Faculty of Computer Science & Information Technology, Universiti Tun Hussein Onn Malaysia, Malaysia
[7]Department of Electrical and Computer Engineering, Universitas Ahmad Dahlan, Indonesia

Article Info

Article history:

Received Jul 29, 2019
Revised Sep 20, 2019
Accepted Oct 11, 2019

Keywords:

Intrusion detection
Machine learning
Malware
Phishing
Website

ABSTRACT

The increasing development of the Internet, more and more applications are put into websites can be directly accessed through the network. This development has attracted an attacker with phishing websites to compromise computer systems. Several solutions have been proposed to detect a phishing attack. However, there still room for improvement to tackle this phishing threat. This paper aims to investigate and evaluate the effectiveness of machine learning approach in the classification of phishing attack. This paper applied a heuristic approach with machine learning classifier to identify phishing attacks noted in the web site applications. The study compares with five classifiers to find the best machine learning classifiers in detecting phishing attacks. In identifying the phishing attacks, it demonstrates that random forest is able to achieve high detection accuracy with true positive rate value of 94.79% using website features. The results indicate that random forest is effective classifiers for detecting phishing attacks.

Corresponding Author:

Ahmad Firdaus,
Faculty of Computer Systems & Software Engineering,
University Malaysia Pahang,
Lebuhraya Tun Razak, 26300 Gambang, Kuantan, Pahang, Malaysia.
Email: firdausza@ump.edu.my

1. INTRODUCTION

The Phishing defined as a way of attempting to acquire information such as usernames, passwords, and credit card details by masquerading as a trustworthy entity in electronic communication. It is a tool used by cyber criminals to steal personal information from the user. The criminals will create a fake website that looks the same as the real websites. The user will get fraud by entering their confidential information such as password, bank details and account credentials into the fake websites [1-3]. The criminal will then use the information provided to access the account to buy stuff, transfer money, or other damaging activities [3, 4]. For example, in 2016 the phishing attack up to 65% worldwide which costs about $1.6 million [5]. The number of phishing attacks has increased significantly in recent years, where 2.3 million sites create in May 2017 [6]. Approximately nearly 1.5 million phishing sites created each month [6]. Over the years, phishing attacks have increased globally. The total number of phishes detected was 263,538 in first quartile 2018. This increased by 46 percent compared to the 180,577 observed in fourth quartile 2017. It was also considerably more than in third quartile 2017 in 190,942 [7]. Figure 1 illustrates the statistic of phishing attacks.

Figure 1 demonstrates the increasing of phishing websites from the year 2017 until March 2018. The increasing of websites also because there a lot of phishing toolkits such as Rock Phish and Super Phisher that make easy for attackers to create fraudulent websites [8]. This fraudulent website able to steal the source code normal websites [8]. Therefore, there is a need for an effective anti-phishing solution for detecting phishing websites and control this internet threat. There are several anti-phishing detections that has been developed by the previous researcher such as using heuristic [9], blacklist [10], and content-based approach [11]. Even though these anti-phishing solutions have been solving phishing attacks, but the users still prone to new phishing attacks. This happens because attackers are not static in their activities; attackers always change their mode activities as often as possible to stay undetected [12, 13]. This motivates this paper into seeking a new solution to solve known and unknown phishing websites.

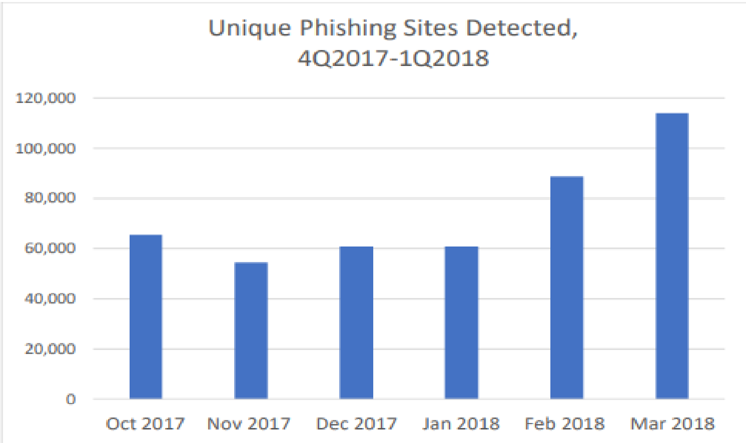

Figure 1. Statistic of phishing attack

Although they are many existing systems for detecting phishing website, however, there are still have room for improvement in detecting phishing websites. Even though properly applied technology, along with security awareness able to reduce the phishing attacks, but it difficult to apply in daily life [14]. For example, email has their own protection approach from a phishing attack, but unable to protect from emerging phishing attack patterns [11, 15]. This is because email used existing phishing patterns, thus making users prone to new phishing attacks. This led to the discovery of machine learning classifiers to detect phishing attacks.

Machine learning is a part of artificial intelligence (AI) that apply data mining approach to discover unknown or existing features from the dataset [16, 17]. Then the features will be used with a classification algorithm to classify either phishing or normal. This paper proposes a phishing detection system which is used to identify phishing attacks as well as to examine the presence of dangerous in websites. The proposed study applies a heuristic based approach and using features from the websites. Hence, the focus of this paper is to detect a phishing attack on the website, the main contributions of this paper are the following:

a) The evaluation study applied phishing website features for a malicious and benign sample from PhishTank dataset.
b) The proposed particle swarm optimization has improved the optimization of phishing website features using tenfold cross-validation.
c) The proposed naïve random forest has increased the accuracy in classifying the phishing attacks on websites applications.

The rest of the paper is organized as follows. In Section 2 discusses related works of the research. Section 3 describes the methodology which includes features optimization and general architecture. Section 4 evaluates the effectiveness of phishing detection system. Lastly, Section 5 conclusion of this paper.

2. RELATED WORK

There are currently various types of phishing attacks. It has been categorized into three different types which are deceptive phishing, malware based phishing and content injection phishing. Deceptive phishing is the messages required to verify account information, requesting that users re-enter their information, bogus account charges, unwanted account changes, new free services requiring immediate action, and many other malicious sites are sent to many recipients in the hope that the unsuspecting person will react by clicking on a link to or signing on a fake site [15]. Malware-based phishing refers to attacks that lead to the installation and execution of malicious software on computers of users [18, 19]. Malware is generally introduced as an email attachment that can be downloaded. Malware commonly installed in phishing attacks includes keyloggers and screen grabbers, spyware that captures and logs input keyboards or display the screen and sends information to the phisher. In other cases, the target of the attack is to control the computer of the victim [20]. The injection of content is a technique in which the phisher changes a part of the content on a reliable website page. This is done in order to mislead the user to go to a page outside the legitimate website where personal information is to be entered [15, 21].

Three types of approaches are used to phishing attacks which are a blacklist based approach, content-based approach, and heuristic-based approach. A blacklist is a list of malicious URLs [17]. Blacklist is obtained using a number of methods, such as heuristics from web crawlers, manual voting, and honeypots. When a website is visited, the browser refers it to the blacklist to check whether the current URL is included in the list [21]. The drawback of this approach is that blacklists cannot normally cover all phishing websites because a newly created fraudulent website takes a considerable amount of time before it is added [16]. Content-based technique for the detection of phishing websites using the term- frequency-inverse document-frequency (TF–IDF) measurements [22]. Heuristic-based approaches collect features from the website to identify them as either phishing or legitimate [20]. Unlike the blacklist method, a heuristic solution can identify in real time newly created phishing websites. The efficiency of the heuristic methods depends on the selection of a set of discriminative features that could help to distinguish the website type.

3. RESEARCH METHOD

Explaining The phishing detection system consists of five components that are collect data, define phishing features, create a model, testing and finally, the result will be compared. Figure 2 shows the component of the phishing detection system.

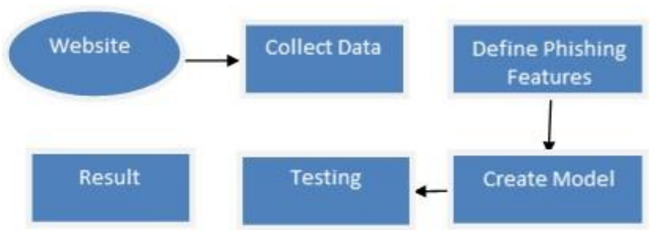

Figure 2. Component of phishing detection system

3.1. Data Collection

The first part of the implementation is to collect dataset. The dataset phase is important for maintaining result accuracy. The dataset will give more understanding and explanation of phishing and legit activities. For further examination, the dataset is then analyzed and the results are used to foresee or predict the future events in phishing.

All the features were collected from (Mohammad, McCluskey, & Thabtah, 2012). There are a total of 30 phishing website features that have been collected. This dataset collected mainly from a well-known phishing database, PhishTank archive, MillerSmiles archive and Google search operators. The collected dataset holds categorical values those are “Legitimate”, ”Suspicious” and “Phishy”, these values have been transformed to numerical values by replacing the values “1”, “0” and “-1” instead of “Legitimate”, “Suspicious” and “Phishy” respectively [19, 20].

3.2. Machine Learning Approach

Machine learning approach is used to ensure that website users are able to optimize the phishing features through the feature optimization approach. This approach provides shorter training and testing time thus it simplifies the phishing detection system. This study applied WEKA tools (Waikato Environment for Knowledge Analysis). It is a popular machine learning software program developed in Java at Waikato University, New Zealand [23, 24]. WEKA supports a number of standard data mining tasks, including data pre-processing, clustering, classification, regression, visualization and selection of features [4, 19, 25].

Feature selection methods were used to identify and remove irrelevant and redundant attributes from data that do not contribute to the accuracy of a predictive model [26].The features of the phishing website were first trained and then classified by using significant features. In order to choose the significant features for effective phishing website detection, this study applies the feature selection approach. Hence, the number of phishing features was reduced from 30 features to 15 features only. This is to ensure that there is a unique pattern appearing between the normal and phishing websites. Then this features trained and testing using machine learning classifier. The result from the machine learning classifiers will become the phishing detection model. Lastly, this model used for testing the phishing websites. Table 1 presents the list of phishing website features used by the study.

Table 1. Phishing Website Features

Phishing Features	Description
SSLFinal_State	SSL Certificates are small data files that digitally bind a cryptographic key to an organization's details. When installed on a web server, it activates the padlock and the https protocol and allows secure connections from a web server to a browser.
URL_of_Anchor	An anchor is an element defined by the <a> tag. This feature is treated exactly as "Request URL".
Website Traffic	This feature measures the popularity of the website by determining the number of visitors and the number of pages they visit.
Prefix_Suffix	The dash symbol is rarely used in legitimate URLs. Phishers tend to add prefixes or suffixes separated by (-) to the domain name so that users feel that they are dealing with a legitimate webpage.
Page_Rank	PageRank is a value ranging from "0" to "1". PageRank aims to measure how important a webpage is on the Internet. The greater the PageRank value the more important the webpage.
Having_Sub_domain	A subdomain is a domain that is a part of a larger domain under the Domain Name System (DNS) hierarchy. It is used as an easy way to create a more memorable Web address for specific or unique content with a website.
Age_of_domain	This feature can be extracted from WHOIS database (Whois 2005). Most phishing websites live for a short period of time.
Domain_registration_length	Based on the fact that a phishing website lives for a short period of time, it was believed that trustworthy domains are regularly paid for several years in advance.
Request_URL	Request URL examines whether the external objects contained within a webpage such as images, videos and sounds are loaded from another domain. In legitimate webpages, the webpage address and most of the objects embedded within the webpage are sharing the same domain.
Links_in_tags	It is common for legitimate websites to use <Meta> tags to offer metadata about the HTML document; <Script> tags to create a client-side script; and <Link> tags to retrieve other web resources. It is expected that these tags are linked to the same domain as the webpage.
DNSRecord	DNS records are basically mapping files that tell the DNS server which IP address each domain is associated with, and how to handle requests sent to each domain
Google_Index	This feature examines whether a website is in Google's index or not.
Links_pointing_to_page	The number of links pointing to the webpage indicates its legitimacy level, even if some links are of the same domain (Dean, 2014).
SFH	Server Form Handler (SFH) that contain an empty string or "about: blank" is considered doubtful because action should be taken upon the submitted information. In addition, if the domain name in SFHs is different from the domain name of the webpage, this reveals that the webpage is suspicious because the submitted information is rarely handled by external domains.
URL_Length	To ensure the accuracy of our study, it has been calculated the length of URLs in the dataset and produced an average URL length. The results showed that if the length of the URL is greater than or equal 54 characters then the URL classified as phishing. By reviewing our dataset we were able to find 1220 URLs lengths equals to 54 or more which constitute 48.8% of the total dataset size.

4. RESULTS AND ANALYSIS

This research applies the supervised machine learning approach since the sample data set have labels (phishing and normal). In addition, supervised machine learning offers good results through the reduction of errors. This study implements four classifiers in order to observe the distinctive results noted in the various machine learning classifiers. The four classifiers are Random Forest (RF), J48, Multi-Layer Perceptron (MLP) and K-Nearest Neighbors (KNN). This study used the parameters for evaluation such as accuracy, FPR, precision, recall, and f-measure to investigate the different measurements. Table 2 shows the results achieved after training and testing the dataset using four classifiers.

Table 2. Performance of Each Classifiers

Classifiers	Accuracy (%)	FPR	Precision (%)	Recall (%)	F-measure (%)
Random Forest	94.79	5.3	94.8	94.8	94.8
J48	93.93	6.0	94.0	93.9	93.9
MLP	93.28	7.0	93.3	93.3	93.3
KNN	93.08	6.8	93.1	93.1	93.1

The results indicated that random forest classifiers had achieved the highest accuracy result of 94.79 percent when compared to KNN which achieved only 93.08 percent. This outcome shows that the random forest classifiers are more effective than other selected classifiers in detecting phishing website. It also shows that feature selection plays a crucial role in determining the effectiveness of phishing website detection. The high precision rate shows that the classifier produced more relevant results and producing accurate results.

4.1. Confusion Matrix

A confusion matrix is a technique for summarizing the performance of a classification model. The table shows two possible classes' prediction, normal and phishing. For example, if a model predicts the presence of phishing activities, the result will show "phishing" and vice versa. Table 3 shows the performance of the four classifiers.

The Table 3 shows that the study produced corrected and magnificent results by predicting the unknown phishing with 1033 for the J48 classifiers. In the incorrectly predicted perspective, the J48 shows the most minimal value. Hence, the outcomes show that J48 classifiers able to predict unknown phishing more accurately.

Table 3. Confusion Matrix of Classifiers

Classifiers	Actual	Predicted	
		Predicted normal	Predicted phishing
RF	Actual normal	1032	62
	Actual phishing	66	1296
J48	Actual normal	1033	61
	Actual phishing	88	1274
MLP	Actual normal	1005	89
	Actual phishing	76	1286
KNN	Actual normal	1023	71
	Actual phishing	99	1263

4.2. Receiver Operating Characteristics Curve (ROC)

In this study, based on the phishing website features, the processes were classified as normal and phishing. Aside from using the performance matrix, this study also calculated the receiver operating characteristics (ROC) curve for each of the machine learning classifiers. In this phase, the TPR was regarded as the detection rate which will correctly predict the phishing process and the FPR was selected as the detection rate which incorrectly predicted normal as phishing. Figure 3 demonstrates ROC curve.

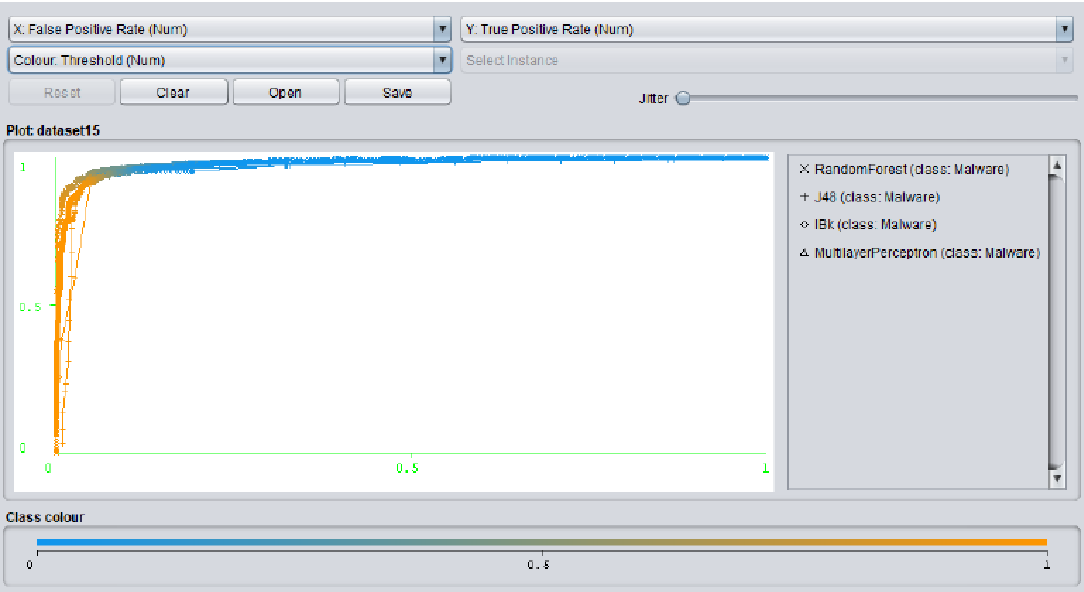

Figure 3. ROC curve

The horizontal axis in the above figure indicates the error detection rate meanwhile the vertical axis indicates the detection rate. Four lines represent the individual ROC curve of the machine learning classifiers. The ROC curve is difficult to compare because it seems to be similar under the same conditions. Hence, the area under the curve (AUC) was used to measure detection accuracy. The AUC results identified were able to measure whether the detection approach was good or bad. An area of 1 shows perfect prediction while an area of 0.5 shows a bad prediction.

Table 4 shows that the random forest and MLP classifiers provide the best AUC values, with over 0.97. This signifies perfect prediction. Overall, the curve and the AUC values confirmed that the most recent phishing experiments had provided compelling accurate results in the phishing website applications detection.

Table 4. AUC Results

Classifier	AUC	Indicator
Random Forest	0.985	Perfect prediction
KNN	0.961	Perfect prediction
MLP	0.978	Perfect prediction
J48	0.957	Perfect prediction

5. CONCLUSION

This paper has presented the performance of the proposed approach in detecting phishing attacks. The proposed approach that implements the machine learning classifier and has correctly classified phishing by using relevant features. In the experiments, this paper considers applied real phishing and benign samples application dataset. The experiment results show that the proposed approach recorded high accuracy in classifying the phishing.

ACKNOWLEDGEMENTS

This work was supported by Universiti Malaysia Pahang, under the Grant Faculty of Computer Systems and Software Engineering (FSK1000), RDU1803163.

REFERENCES

[1] A. Firdaus, N. B. Anuar, M. F. A. Razak, and A. K. Sangaiah, “Bio-inspired computational paradigm for feature investigation and malware detection: interactive analytics,” *Multimed. Tools Appl.*, 2017.
[2] Muhammad Taseer Suleman and Shahid Mahmood Awan, “Optimization of URL-Based Phishing Websites Detection through Genetic Algorithms,” *Autom. Control Comput. Sci.*, vol. 53, no. 4, pp. 333–341, 2019.
[3] A. Kulkarni and L. L., “Phishing Websites Detection using Machine Learning,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 7, 2019.
[4] M. Hazim, N. B. Anuar, M. F. Ab Razak, and N. A. Abdullah, “Detecting opinion spams through supervised boosting approach,” *PLoS One*, vol. 13, no. 6, pp. 1–23, 2018.
[5] PhishMe, “Analysis of Susceptibility, Resiliency and Defense Against Simulate and Real Phishing Attacks,” 2017.
[6] W. S. Cybersecurity, “Nearly 1.5 Million New Phishing Sites Created Each Month,” *Webroot Smarter Cybersecurity*, 2017. .
[7] APWG, “APWG Phishing Attack Trends Reports,” *APWG Unifying Global Response to Cybercrime*, 2018. .
[8] R. Gowtham and I. Krishnamurthi, “A comprehensive and efficacious architecture for detecting phishing webpages,” *Comput. Secur.*, vol. 40, pp. 23–37, 2014.
[9] S. G. Selvaganapathy, M. Nivaashini, and H. P. Natarajan, “Deep belief network based detection and categorization of malicious URLs,” *Inf. Secur. J.*, vol. 27, no. 3, pp. 145–161, 2018.
[10] L. McCluskey, F. Thabtah, and R. M. Mohammad, “Intelligent rule-based phishing websites classification,” *IET Inf. Secur.*, vol. 8, no. 3, pp. 153–160, 2014.
[11] A. A. Akinyelu and A. O. Adewumi, “Classification of phishing email using random forest machine learning technique,” *J. Appl. Math.*, vol. 2014, 2014.
[12] M. F. A. Razak, N. B. Anuar, R. Salleh, A. Firdaus, M. Faiz, and H. S. Alamri, “‘Less Give More’: Evaluate and zoning Android applications,” *Meas. J. Int. Meas. Confed.*, vol. 133, pp. 396–411, 2019.
[13] M. Akiyama, T. Yagi, T. Yada, T. Mori, and Y. Kadobayashi, “Analyzing the ecosystem of malicious URL redirection through longitudinal observation from honeypots,” *Comput. Secur.*, vol. 69, pp. 155–173, 2017.
[14] B. Li, G. Yuan, L. Shen, R. Zhang, and Y. Yao, “Incorporating URL embedding into ensemble clustering to detect web anomalies,” *Futur. Gener. Comput. Syst.*, vol. 96, pp. 176–184, 2019.
[15] S. Nisha and A. N. Madheswari, “*Secured authentication for internet voting in corporate companies to prevent phishing attacks,*” vol. 22, no. 1, pp. 45–49, 2016.

[16] H. B. Kazemian and S. Ahmed, "Comparisons of machine learning techniques for detecting malicious webpages," *Expert Syst. Appl.*, vol. 42, no. 3, pp. 1166–1177, 2015.
[17] K. Thomas, C. Grier, J. Ma, V. Paxson, and D. Song, "Design and Evaluation of a Real-Time URL Spam Filtering Service," *2011 IEEE Symp. Secur. Priv.*, pp. 447–462, 2011.
[18] A. Firdaus, N. B. Anuar, M. F. A. Razak, I. A. T. Hashem, S. Bachok, and A. K. Sangaiah, "Root Exploit Detection and Features Optimization: Mobile Device and Blockchain Based Medical Data Management," *J. Med. Syst.*, vol. 42, no. 6, 2018.
[19] M. F. A. Razak, N. B. Anuar, F. Othman, A. Firdaus, F. Afifi, and R. Salleh, "Bio-inspired for Features Optimization and Malware Detection," *Arab. J. Sci. Eng.*, 2018.
[20] J. A. Chaudhry, S. A. Chaudhry, and R. G. Rittenhouse, "Phishing attacks and defenses," *Int. J. Secur. its Appl.*, vol. 10, no. 1, pp. 247–256, 2016.
[21] R. Gowtham and I. Krishnamurthi, "A comprehensive and efficacious architecture for detecting phishing webpages," *Comput. Secur.*, vol. 40, pp. 23–37, 2014.
[22] G. Xiang, J. Hong, C. P. Rose, and L. Cranor, "Cantina+," *ACM Trans. Inf. Syst. Secur.*, vol. 14, no. 2, pp. 1–28, 2011.
[23] R. Thakur, *"Preprocessing and Classification of Data Analysis in Institutional System using Weka,"* vol. 112, no. 6, pp. 9–11, 2015.
[24] F. Afifi, N. B. Anuar, S. Shamshirband, and K.-K. R. Choo, "DyHAP: Dynamic Hybrid ANFIS-PSO Approach for Predicting Mobile Malware," *PLoS One*, vol. 11, no. 9, p. e0162627, 2016.
[25] C. Science, *"Comparative evaluation of the different data mining techniques,"* vol. 10, no. 3, pp. 233–238, 2016.
[26] T. Chou and J. Pickard, *"Machine Learning based IP Network Traffic Classification using Feature Significance Analysis,"* vol. 16, no. 3, pp. 9–12, 2018.