International Journal of Aquatic Science
ISSN: 2008-8019
Vol 12, Issue 03, 2021

Survival Study On Website Phishing Attack Detection

R.Sakunthala jenni[1],S.Shankar[2]

[1]Research Scholar, Department of Computer Science and Engineering, Hindusthan College of Engineering and Technology, Coimbatore, India.
[2]Professor, Department of Computer Science and Engineering, Hindusthan College of Engineering and Technology, Coimbatore, India.

Email: [1]kalaivanijenni@gmail.com,[1]shankarhicet@hindusthan.net

Abstract. In World Wide Web, cybercriminalsutilize the opportunities to hack the personal information like username, password, account number and national insurance numbers called Web phishing attack (WPA). WPA is performed viasending link to emails. Victims receive email to update information. When link is clicked by victims, Web browser sends phishing website that appears likeoriginal website.Phishing website is identifiedthrough the characteristics such as URL and domain identity. The data mining techniquesis employed to identifywebsite is phishing website or not. However, the WPAdetection (WPAD) waschallenging task. Our main objective is to improve the WPAD performance through studying the existing problems.
Keywords: Web browser, Web phishing attack,Cybercriminal, Victims, Attack detection, Domain identity

1. INTRODUCTION

Phishing is fraudulent one to get sensitive information via hiding as trust worthy entity in electronic communication. Via email spoofing or instant messaging, Phishing was carried ou tand directed users provide personal information to fake website and identical to legitimate site.Phishing is a type of cyber attack that everyone protects themselves. Phishing is afake e-mail designed to attract the victim. When the attacker is deceiving victim, it is encouraged to presentthe confidential information in fraud website. Phishing e-mails are transmitted to retrieve the login details of employees to utilize for advanced attack against particular company.

This paper is structured as below: Section 2 describes various WPAD review in cloud environment, Section 3 elucidates study and analysis of existing WPAD, Section 4 depicts the comparison of existing WPAD techniques. In Section 5, the discussion and issues of existing WPAD techniques are portrayed and Section 6 concludes the paper.

2. LITERATURE REVIEW

XSS attack detection method was designed in [1] using ensemble learning approach. However, the designed technique failed to consider the inside weakness like vulnerability. In

[2], a phishing website detection technique was introducedwith meta-heuristic-based nonlinear regression (NR) and feature selection method. The runtime was not reduced as it failed to have parallel memory for HS. The reliability of HS was not improved.

To identify online phishing attacks, A novel phishing email detection system (PEDS) was presented in [3] to integrate neural network (NN) with reinforcement learning. The designed framework not classified the spam email, phishing and ham email. The designed framework not increased the richness of designed model. Different approaches were designed in [4] to detect spammers on Twitter through finding the similarities between the spam accounts. A number of features were introduced to enhance the classification algorithm performance. But, the scalability was not enhanced without reducing the accuracy.

For both spam message and account identification process, A unified framework was presented in [5]. In designed framework, four datasets were employed. A novel lightweight phishing detection approach was designed in [6] based on the uniform resource locator (URL). The designed system enhanced the recognition rate. However, the designed approach failed to analyze the system constantly on the gigantic phishing websites database to enhance it when it was mandatory.

In [7], the aspects of Cyber kill chain depended taxonomy of banking Trojans werepresented. However, the taxonomy did not hide other malware families through the defense using evolutionary computational intelligence. In [8],a new feature selection method with semantic ontology was presented to gather words into topics to build feature vectors. Though the feature selection accuracy was enhanced, the time consumption was not reduced.

In [9], anovel spam filter combined N-gram tf.idf feature selection, varied distribution-based balancing and regularized deep multi-layer perceptron NN with rectified linear units (DBB-RDNN-ReL). But, DBB-RDNN-ReL has high computational cost and it was difficult to address the concept drift problem. To identify phishing attacks, a two-level authentication approach was designed in [10]. But, the designed system failed to identify the non-HTML websites with higher accuracy.

A new approach was designed in [11] to identify the phishing attack. However, designed system was not employed to identify non-HTML websites. The phishing websites detection in mobile environment remained an open issue. In [12], URL and web traffic features were presented to discover phishing websites. But, phishing attack detection time was not minimized using anti-phishing model.

3. WEBSITE PHISHING ATTACK DETECTION

Phishing attack is the severe Internet security threats. In WPA, user gives his/her secret credential to fake website that resembles like genuine one[13]. The WPA affects the online payment services, e-commerce, and social networks. Via considering visual resemblance merits, a phishing attack is performed. Attacker creates the webpage that resemble same aslegitimate webpage. The phishing webpage link is distributed to the large number of Internet users via emails and communication website[14][15]. The fake email content describesfear sense, significance and request user to perform urgent action. Fake email is pressuringuser to renew PIN and evade debit/credit card suspension. Cyber criminals gather user details, when the user wrongly updates confidential credentials. Phishing attack contains more cyber fraud thatinfluences Internet users.

3.1 An ensemble learning approach for XSS attack detection with domain knowledgeand threat intelligence

International Journal of Aquatic Science
ISSN: 2008-8019
Vol 12, Issue 03, 2021

Based on ensemble learning approach, Cross-site scripting (XSS) attack detection method was presented to utilize BN. Via the domain knowledge and threat intelligence, each BN was created. An analysis method was sort nodes in BN consistent with influences on outcome node.

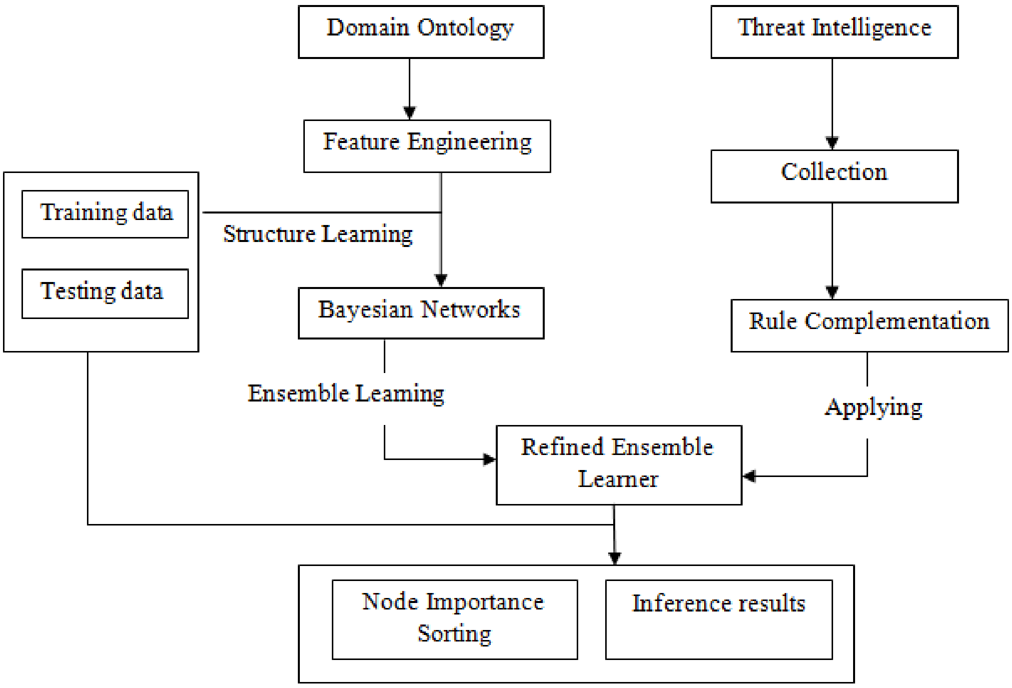

Fig. 1. XSS Attack Detection method

Fig. 1 explains theXSS Attack Detection method. Initially,the ontology wasconstructed to create features which indicate XSS attackfeatures. The features were set asnodes in BN and values are obtained. Learning algorithm employs scoring and searching learning. Each BN wasan individual learner, a voting method groupindividual model to produceensemble learner. Then, threat intelligence wasdiscovered to enhance results. To face concealed XSS attack, gathered intelligence was employed to generate complement rules.

By utilizing an ensemble learner and complement rules, new data input was identified. Depends on ensemble BN learner, the node importance sorting was performed to discovernodes influences in detection outcome. BN was white-box model where outcomewasunderstandable and complement rules discoverhidden attack.To handle the incidents, the BN and threat intelligence rules were utilized. XSS payload was not similarto normal requests or inputs, like abnormal input length, sensitive words, sensitive characters and redirection link.

Due to malicious codes, XSS payload was longer than normal one. Input length was attained as one feature. To discover XSS, Sensitive words and characters wereessential. For one payload, diverse words and characters are exist. In machine learning (ML) model, the words and characters are employed to generate one malicious payload and appearances wereutilized. To conceal their original form, an XSS payload utilizes redirection link. For redirecting current page to another page, the designed payload was employed in one payload. The appearance time of protocols was counted and redirection address wasattained for analysis.

3.2 Heuristic nonlinear regression strategy for detecting phishing websites

A phishing website detection approach was introduced with two feature selection methodsto pick best feature subset. Then, two meta-heuristic algorithms wereemployedtodiscoverfraudulent websites. Harmony search (HS) was employed with NR method and support vector machine (SVM). The NRcategorizes the websiteswhere regression model metrics were achievedwith HS. HS algorithm employsdynamic pitch adjustment rate and generate new one.

Decision Tree and wrapper techniques were used to attain clear dissemination of feature set and eradicate noisy features. DT was employed in initial phase. If nodes removal in sub-tree not affected root, then feature in root was consideredas fundamentalfeature. The significant feature wasfound, iteradicated from DT list and next significant feature wasrestored. The wrapper process with genetic algorithm (GA) was pickbest feature subset. The classificationalgorithms in wrapper method were taken as black box. The classification techniqueswere assumed for identify the optimal subsets for classification methods.

In wrapper method, the features were embedded to discover optimal feature subset with greater accuracy. NR with HS discovers phishing websites via the extracted feature. NRtried to discover functional relationship among inputs and outcome. The coefficients of NRwerecalculatedthrough modified HS (MHS). MHS lessen mean-square-error (MSE) amongforecasted and target outcomes. NRwasperformingregression analysis with independent variables combination address the nonlinear issues. HSwasestimating the best weights for NR. HSmethod was used for optimization issues. A solution vector was same as harmony inmusic. Solution vector searching wassimilar to process employedin orchestra.

3.3 Detection of Online Phishing Email using Dynamic Evolving Neural Network Based on Reinforcement Learning

To detect WPA in online mode, a new PEDS framework was presented which combinedNN with reinforcementlearning. The designed system performance was improved via adopting the reinforcement learning. The designed model addressed the limiteddataset issues.

Depends on supervised and unsupervised ML methods, PEDS frameworkperforms online phishing email detection. The supervised MLtechnique employed training dataset to builddetection model while unsupervised MLadapted detection model bynovel delivered email tosystem. The designedframework determine the new phishing behaviors in four stages, such as pre-processing, FEaR, DENNuRL and RL-Agent.

The pre-processing includes two steps. The feature from every email text and header are extracted in first phase. Thefeatures are described in diverse properties of each email. Thesecond step comprised selection ofefficient features to speed-up adaptation of classification model. The features were chosen from email headers and email content. A new algorithm was designed to discover new behavior and rank selected features list. In online phishing email detection field, the essential feature was varying one. The designed algorithm alteredessential features and obtained from next email. NN was core of classification model. Dynamic Evolving NN algorithm with Reinforcement learning (DENNuRL) permitted NN vary dynamically and build best NN to resolve desired issue. The reinforcement learning approach studied the optimal behavior depending on trial-and-error interaction. RL-agent observed PEDS outcome in online mode.

4. PERFORMANCE ANALYSIS OF WEBSITE PHISHING ATTACK DETECTION TECHNIQUES

In order to compare differentWPADtechniques, number of website data and features wereobtained from Phishing Websites Data Set from UCI MLRepositoryfor experimental. Various parameters are used for website phishing attack detection.

4.1 Feature Selection Time (FST)

FSTis measured as time consumedto select relevant features to perform WPAD. It is variation of starting time and ending time of feature selection for WPAD. It is calculated in milliseconds (ms) and given by,

$$FeatureSelectionTime = Endingtime - Startingtimeoffeatureselection (1)$$

From (1), the feature selection is calculated.

Number of Features (Number)	Feature Selection Time (ms)		
	XSS attack detection method	Phishing website detection approach	PEDS framework
3	25	37	48
6	27	39	50
9	29	42	53
12	26	40	51
15	24	38	49
18	22	36	46
21	23	39	48
24	27	43	52
27	30	46	55
30	33	49	58

Table 1. Tabulation for Feature Selection Time

FST is illustrated in Table 1 with number of features ranging from 3 to 30. FSTcomparison takes place on existing XSS attack detection method, Phishing website detection approach andPEDS framework.

International Journal of Aquatic Science
ISSN: 2008-8019
Vol 12, Issue 03, 2021

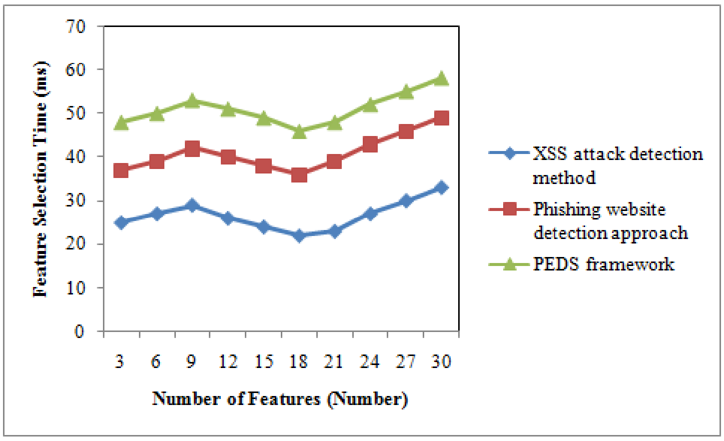

Fig. 2. Measure of Feature Selection Time

FST is portrayed in Fig.2 with number of features. From Fig.2, it is clear thatFSTusingXSS attack detection methodis lesser when compared to phishing website detection approach and PEDS framework. This is because, this methodutilizesensemble learning approach and BN. Scoring and searching learning algorithm was used in ensemble learning approach. FSTof XSS attack detection method is 35% lesser than phishing website detection approach and 48% lesser than PEDS framework.

4.2. Phishing Attack Detection Accuracy(PADA)

PADA is calculated as ratio of number of website data which are correctly classified as phishing attack to total number of website data. It is computed in percentage (%) and given by,

$$PADA = \frac{Numberofwebsitedatacorrectlyclassifiedasphishingattack}{Totalnumberofwebsitedata} \quad (2)$$

From (2), the PADAis determined.

Table 2. Tabulation for Phishing Attack Detection Accuracy

Number of Website data (Number)	Phishing Attack Detection Accuracy (%)		
	XSS attack detection method	Phishing website detection approach	PEDS framework
50	72	85	78
100	75	87	81
150	78	89	84
200	76	86	82
250	73	84	80
300	77	87	83
350	81	90	86
400	79	88	84
450	76	85	82
500	80	89	85

PADA is explained in Table 2 with number of website data. PADA compared with XSS attack detection method, Phishing website detection approach and PEDS framework.
PADA of three methods is portrayed in Fig. 3 with number of website data. From Fig. 3, PADA using phishing website detection approach is higher when compared to XSS attack detection method and PEDS framework. As a result, PADA of phishing website detection approach is 13% higher than XSS attack detection methodand 5% higher than PEDS framework.

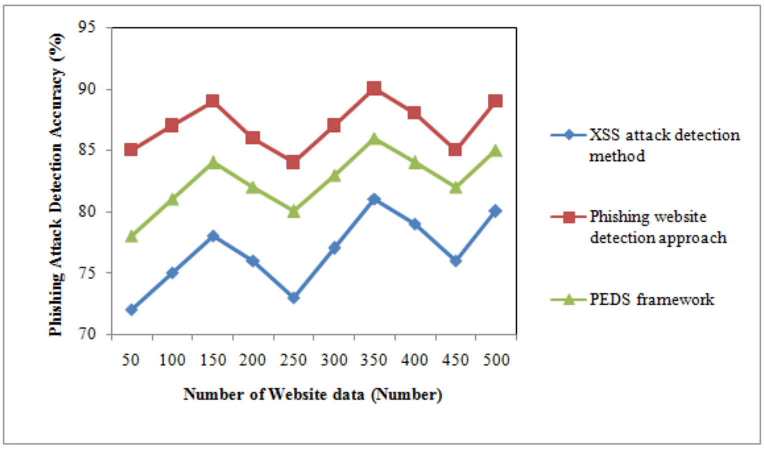

Fig. 3. Measure of Phishing attack Detection Accuracy

4.3 False Positive Rate (FPR)

FPR is calculated as ratio of number of website data which are incorrectly detected as phishing attack to total number of website data. It is measured in percentage (%) and formulated as,

$$FPR = \frac{Numberofwebsitedataincorrectlyclassifiedasphishingattack}{Totalnumberofwebsitedata} \quad (3)$$

From (3), the FPRis measured.

Number of Website data (Number)	False Positive Rate (%)		
	XSS attack detection method	Phishing website detection approach	PEDS framework
50	28	34	12
100	26	32	10
150	23	30	9
200	27	33	11
250	30	36	13
300	33	39	17
350	31	37	15
400	29	34	12
450	33	38	18
500	37	42	22

Table 3. Tabulation for False Positive Rate

FPR comparison of three methods is explained in Table 3with number of website data in the range of 50 to 500.

Fig. 4 described the FPR with number of website data. From Fig. 4, FPR of PEDS framework is minimal than the other conventional methods. This is because designed model used the NN, reinforcement learning and data mining associative classification methods to detect the phishing attacks. FEaR identified the new behavior and ranked the features list. DENNuRL allowed NN to vary dynamically and constructed the NN for addressing the existing problem. RL-agent examined PEDS output in online mode. As a result, the FPR of PEDS framework is 54% and 61% lesser than XSS attack detection methodand phishing website detection approach.

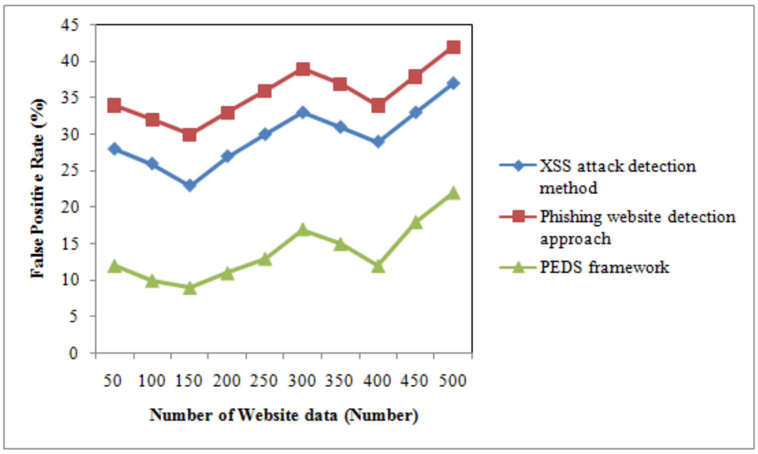

Fig. 4. Measure of False Positive Rate

5. DISCUSSION AND LIMITATION ON WEBSITE PHISHING ATTACK DETECTION TECHNIQUES

XSS attack detection method was introducedwithBN. The collected threat intelligence enhanced the learning accuracy. A model explanation method determined node importance. BNs identified theessential factors for the attacks detection. Designed method failed to assume outside web attacks and inside weakness like vulnerability. The designed method and their outputs were not employed in web security risk assessment system.
Phishing website detection was introduced with feature selection approach. NRcomputed thefunctional relationship betweeninputs and outputs. MHSlessen theMSEamongforecasted and target result. But, the runtime was notminimizedas it failed to haveparallel memory for HS. The reliability of HS was minimal.A novelPEDS frameworkwas developing the best NN to discovernovel behavior. The designed model adapted to producePEDS which reflectsvariations with newly explored behaviors. But, designed framework failed to categorize the spam email, phishing and ham email. The designed framework failed to improve the model richness.

5.1 Future Direction

The forthcoming direction of WPADcan be performedwithML and deep learning (DL) techniques to enhancePADA and lessen theFPR.

6. CONCLUSION

A different conventionalWPADtechniquescomparison is studied. From survival study, the conventional method does not enhance the WPADaccuracy. In addition, the reliability was not increased.XSS attack detection method failed to consider the outside web attacks like XSS and inside weakness like vulnerability. In existing PEDS framework, it failed to classify the spam email, phishing and ham email. The experiment on conventional methods portrays the performance of WPADtechniques with its issues. To conclude that, the research work can be performedwith MLand DL techniques for improving the performance of WPAD.

7. REFERENCES

[1] Zhou, Y., Wang, P.: An ensemble learning approach for XSS attack detection with domain knowledge and threat intelligence. Computers and security 82, 261–269 (2019).
[2] Babagoli, M., Aghababa, M., VahidSolouk.: Heuristic nonlinear regression strategy for detecting phishing websites. Soft Computing 1–13 (2018).
[3] Smadi, S., Aslam, N., Zhang, L.: Detection of Online Phishing Email using Dynamic Evolving Neural Network Based on Reinforcement Learning. Decision Support Systems 107, 88-102 (2018).
[4] Adewole, K., Han, T., Wu, W., Song, H., Sangaiah, A.: Twitter spam account detection based on clustering and classification methods. The Journal of Supercomputing, 1–36 (2018).
[5] Adewole, K., Han, T., Wu, W.,Song, H., Sangaiah, A.: SMSAD: a framework for spam message and spam account detection. Multimedia Tools and Applications, 1–36 (2017).

[6] Zouina, M., Outtaj, B.: A novel lightweight URL phishing detection system using SVM and similarity index. Human-centric Computing and Information Sciences 7, 1-13 (2017).

[7] Kiwia, D,. Dehghantanha, A., Choo, K., Slaughtera. J.: A cyber kill chain-based taxonomy of banking Trojans for evolutionary computational intelligence. Journal of Computational Science 27, 394-409 (2018).

[8] Mendez, J.R., Cotos-Yanez, T.R., Ruano-Ordas, D.: A new semantic-based feature selection method for spam filtering. Applied Soft Computing 76, 89-104 (2017).

[9] Barushka, A., Hajek, P.: Spam filtering using integrated distribution-based balancing approach and regularized deep neural networks. Applied Intelligence 48, 3538–3556 (2018).

[10] Jain, A. K., Gupta, B. B.: Two-level authentication approach to protect from phishing attacks in real time.Journal of Ambient Intelligence and Humanized Computing 9(6), 1783–1796 (2018).

[11] Jain, A. K., Gupta, B. B.: A machine learning based approach for phishing detection using hyperlinks information. Journal of Ambient Intelligence and Humanized Computing 1–14, (2018).

[12] Pham, C., Nguyenz, L. A. T., Tran, N. H., Huh, E., Hong, C.: Phishing-Aware: A Neuro-Fuzzy Approach for Anti-Phishing on Fog Networks, IEEE Transactions on Network and Service Management1 5(3), 1076 – 1089 (2018).

[13] Sujatha, K & Shalini Punithavathani, D 2016, 'Fuzzy Based Weight Estimation and Sub band Architecture in Image Fusion for Multi Exposure Images', Asian Journal of Information Technology (AJIT), ISSN:1682-3915, Vol. 15, No.3, pp.384-392

[14] Viji, C., Rajkumar, N., Suganthi, S.T. et al. An improved approach for automatic spine canal segmentation using probabilistic boosting tree (PBT) with fuzzy support vector machine. J Ambient Intell Human Comput (2020).

[15] K. Venkatachalam, A. Devipriya, J. Maniraj, M. Sivaram, A. Ambikapathy, and S. A. Iraj, "A novel method of motor imagery classification using eeg signal," *Artificial intelligence in medicine*, vol. 103, p. 101787, 2020.