A Comprehensive Survey on Identification and Analysis of Phishing Website based on Machine Learning Methods

Mohammed Hazim Alkawaz
Faculty of Information Sciences & Engineering
Management and Science University
Shah Alam, Selangor, Malaysia
mohammed_hazim@msu.edu.my

Stephanie Joanne Steven
School of Graduates Studies
Management and Science University
Shah Alam, Selangor, Malaysia
stephaniejoanne_my@yahoo.com

Asif Iqbal Hajamydeen
Faculty of Information Sciences & Engineering
Management and Science University
Shah Alam, Selangor, Malaysia
asif@msu.edu.my

Rusyaizila Ramli
Faculty of Information Sciences & Engineering
Management and Science University
Shah Alam, Selangor, Malaysia
rusyaizilaramli@msu.edu.my

2021 IEEE 11th IEEE Symposium on Computer Applications &amp; Industrial Electronics (ISCAIE) | 978-1-6654-0338-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/ISCAIE51753.2021.9431794

Abstract*— *Phishing is a cybercrime which is carried out by imitating a legal website to trick users to steal their personal data, including usernames, passwords, account numbers, national insurance numbers, etc. Phishing frauds may be the most widespread cybercrime used today. Machine learning focuses on computer algorithms which improves automatically through experience. Machine learning methods were utilized to detect phishing URLs that typically evaluates an URL based on a feature or set of features extracted from it. This paper presents an approach to identify phishing websites using trained machine learning models. It also delivers a detailed analysis of phishing attacks with a comparison on machine learning approaches used for analysis and classification of phishing and legitimate websites.

Keywords: *Security, phishing, detection, Machine Learning, Visual Similarity, Hybrid, Random Forests, Resource Description Framework (RDF)*

I. INTRODUCTION

Phishing impersonated the overlook of email and appears to be as the legitimate source [1]. This tricks users to visit phished sites through links given in phishing email. The phishers will trick users to fill in their personal information by sending alert messages asking to validate their account. This is done so that user will think that it is a mandatory action needed from their end [2]. Anti-Phishing Working Group has defined phishing as a crime of stealing user's credential data from financial accounts and identity information [3]. In first quarter of 2018, 263538 number of phish were reported in APWG 1Q report. Referring to the fourth quarter of 2017, this is forty-six percent of from 180577[4]. In addition, 233040 phish were detected in second quarter of 2018 compared to the first quarter which was 263538[5]. This was reported in APWG 2Q. This numbers were higher compared to fourth and third quarter of 2017 which were only 180577 and 190942. With an increase of twenty-one percent of phishing attacks. There were rises in the SAAS/webmail targeted market. The payment industry continues as the most appealing phishing target.

Today many security researchers are depending on machine learning techniques to overcome weakness of existing approaches [6]. This technique only requires previous data to analyze or predict future data hence it comprises a wide range or algorithms. Machine learning establishes analytical models by using complex intervention during supervised leaning [7]. This technique appears to be appropriate for the detection of phishing pages, as this tricky can be transformed into a classification mission. Machine learning techniques were used to progress measures to perceive phishing actions founded on the classification of existing web sites, and these replicas can be incorporated hooked on the browser [8]. Machine Learning models instantaneously detect the valid site and then advancing the production to the user at the other end. The key accomplishment is the structures of the website in the input data set and the accessibility of satisfactory sites for the structure of reliable logical replicas for the expansion of Machine Learning models for computerized anti-phishing identification [9]. The main aim of this survey is to study the comprehensive analysis and relevant literature review that supports the research questions and propose an approach in analyzing and classifying phishing and legitimate websites. Phishing is one of cyber-world's major problems and leads to both industry and individual financial losses. High accuracy phishing attack detection has always been a challenge [10]. More than 80 percent of users believe the green lock icon on the browser URL bar means the webpage is legitimate and secure, according to a report by Phish Labs [11].

The confusing "Safe" label given to verifiable HTTPS websites by modern browsers, even if it is a fake one, just aggravates the situations, as shown in Figure 1. To compare, Figure 2 gives a screenshot of the legitimate PayPal login page, visually almost the same as the fake one. This was the limitation of phishing website which gave machine learning to penetrate through it making it possible to identify if it is possible to identify whether the website is fake or legitimate [12].

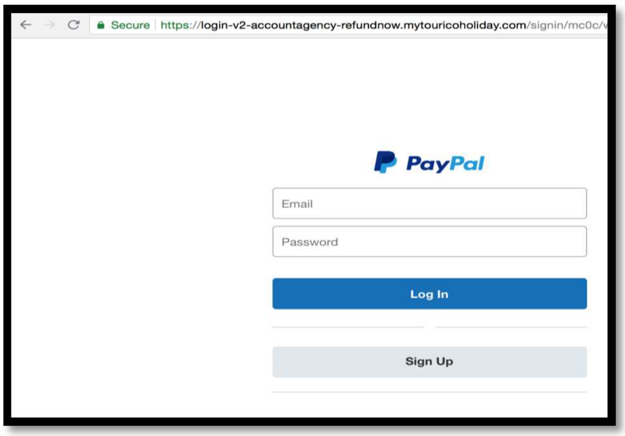
Secure
https://login-v2-accountagency-refundnow.mytouricoholiday.com/signin/mc0c/v
PayPal
Email
Password
Log In
Sign Up

Figure 1: Secure Label Shown in Browser Address Bar for a Fake PayPal Page

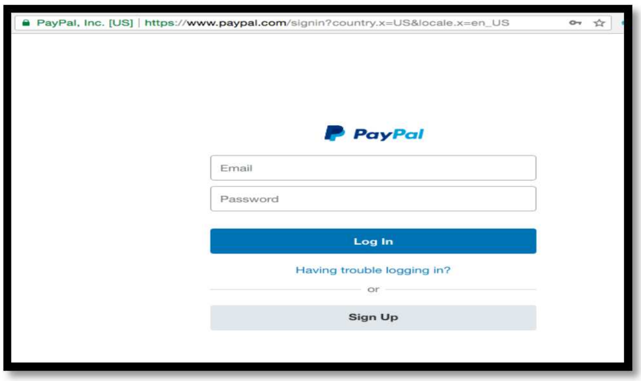
PayPal, Inc. [US]
https://www.paypal.com/signin?country.x=US&locale.x=en_US
PayPal
Email
Password
Log In
Having trouble logging in?
or
Sign Up

Figure 2: Secure Label Shown in Browser Address Bar for the True PayPal Page

II. LITERATURE REVIEW

This section delivers a complete and relevant literature review that supports the research questions and their possible solutions. Different kinds of phishing have grown over the last several years, during which assailants find new ways and means of using innovative ideas by scrutinizing the users and upgrading themselves with the latest technology to make the pages appear stronger and more comfortable than ever [13]. This chapter also discuss on the related proposed method and the overview on phishing is taken place.

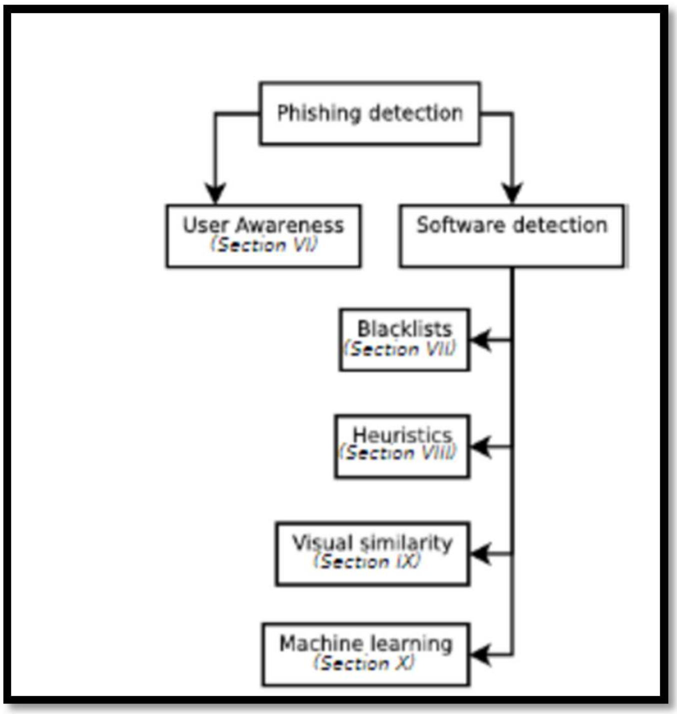

Figure 3: An overview of phishing detection approaches.

Figure 3 illustrates the overview and approaches that have been introduced during phishing detection approach that suffer from low detection accuracy and high false alarm. In addition, the blacklist-based method, the most common technique used, is ineffective in responding to phishing attacks as it has become easier to setup a new domain, and a detailed blacklist can hardly ensure a perfect up-to-date approach [14].

The hybrid approach proposed by Bhawna Sharma and Parvinder Singh [15] includes a grouping of content similarity whitelists, style similarity and heuristics which is called PhishAlert. As compared to existing algorithms namely CANTINA and CANTINA+, PhishAlert executed more effectively on experimental dataset that includes 500 phishing sites and 500 valid sites [15]. The legal URLs have been retrieved from the stuffgate server [17] while the phishing URL from phishtank [18] have been obtained. It has been found that PhishAlert is 98% precise as shown in Figure 4. The performance of the PhishAlert model decreases with an increase in dataset size. In order to achieve a better detection efficiency, authors proposed that more functions should be included within the future PhishAlert algorithms.

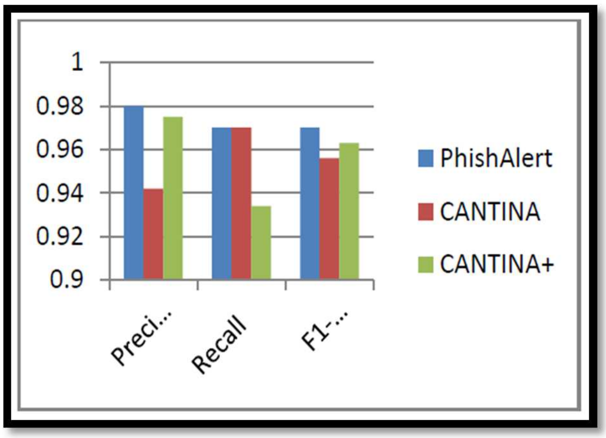

Figure 4: An Evaluation of PhishAlert with CANTINA and CANTINA+ based on precision, recall and f1-measure.

Vamsee Muppavarapu, Archanaa Rajendran, and Shriram Vasudevan proposed [19] an approach on phishing detection using Resource Description Framework (RDF) and Random Forests. This paper aims to identify phishing websites and to provide possible targeted domain. There is 2 level of process. In the first level is founded on RDD model of webpages and second is founded on machine learning technique. Both levels collaborate to reduce the number of false positives to improvise the accuracy and precision of system.

Hypertext Markup Language (HTML) to RDF model generation is carried out once extraction is done of collecting feature from suspicious web page [20]. Twenty-one features were selected where there are no similarity webpages that will carry the same element sent. Other vocabularies have been added such as Extensible Hypertext Markup Language (XHTML) [21], HTTP [22] and Dublin Core [23][24]. Among the machine learning algorithms, Random Forest has shown best performed in classification where it has good accuracy even with less sensitive outliers and missing values in parameter choices [25].Three metrics were used to evaluate the performance of system as shown in Figure 5 where 98.8% achieved from True Positive Rate (TPR), whereas 1.5% for False Positive Rate (FPR) and 98.68% for Accuracy (ACC).

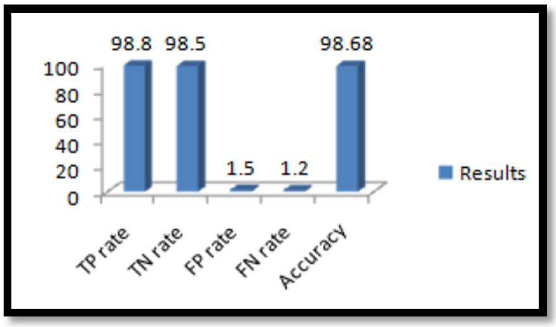

Figure 5: Comparison Result

Ankit Kumar Jain, B. B. Gupta introduces a visual similarity method to phishing detection [26].This study involves a thorough review of phishing occurrences, their trick, some of the greatest current visual similarity methods to phishing detection, and its survey - based. The survey presented a deeper considerate of the situation, the current explanation space, and the context of future research to effectively transaction with phishing occurrences using visual similarity-based strategies. A comparison was made to show which Visual similarity-based techniques was able to detect number of attacks experimented. As a result, Hybrid approach obtained high accuracy in detecting most of phishing attacks compared to other techniques as shown in Table 1. However, Authors have stated that no solitary method is enough for the use of phishing.

Table 1: Analysis of hybrid approach of various attacks

Approach	Zero-hour protection	Embedded object	Language independence	Partially copied webpage	DNS attack
Bayesian model	×	✓	×	×	×
Hybrid features	×	✓	×	×	×
Using phishing target	✓	✓	×	✓	×
Website logo for phishing detection	✓	✓	✓	✓	×

Rishikesh Mahajan and Irfan Siddavatam **Error! Reference source not found.** [28]present an approach on phishing detection using machine learning. This paper aims to perceive phish URL's by looking into the rate of accuracy, false positive and false negative of machine learning algorithms which are Decision Tree (DT), Random Forest (RF) and Support Vector Machine (SVM). Dataset is segregated into training and testing set to evaluate the performance of classifiers. As in Table 2, Random Forest achieved accuracy of 97.14 with ninety percent of data used as training dataset.

Table 2: Classifier's performance

Dataset Split Ratio	Classifier	Accuracy Score	False Negative Rate	False Positive Rate
50:50	DT	96.71	3.69	2.93
	RF	96.72	3.69	2.91
	SVM	96.40	5.26	2.08
70:30	DT	96.80	3.43	2.99
	RF	96.84	3.35	2.98
	SVM	96.40	5.13	2.17
90:10	DT	97.11	3.18	2.66
	RF	97.14	3.14	2.61
	SVM	96.51	4.73	2.34

Kahksha and Sameena Naaz [25]present a machine learning approach in phishing detection website [29]. The aim of this research was to build a model to protect users from phishing attacks. Decision Tree, Linear Model, Random Forest and Neural Network algorithms were used on a phishing dataset in this paper. The dataset was gathered from MillerSmiles archive, PhishTank archive and Google searching operators. The data set consists of 2456 instances and 30 features. Value of attributes is in the form of integer -1, 0, and 1, -1 represents phishing, 0 denotes suspicious and 1 denotes legitimate First the data set has been processed in order to get mature data in the desired format, then it is split into two parts, seventy percent of training and thirty percent of testing. Four machine learning algorithms were used in this specifically Decision tree, Random forest, Neural Network and Linear model [30]. In this work 2456 websites having 30 attributes have been investigated from the data taken from UCI Machine learning repository data. Rstudio has been used here for the implementation. In terms of accuracy, the rate of true positive rate, the rate of true negative , accuracy, the F measure and the rate of false positive, the decision tree, the random forest, the neural network, and the linear model were implemented on the above dataset and the effects. These values were determined using performance indicator formulas and using the confusion matrix in the test data collection. As shown in Figure 6 The Random Forest algorithm works best than other algorithms in term of different parameters. At 95.70 percent, it achieved the highest precision of all, while other algorithms function with 90.4 percent (DT), 90.7 percent (NN) and 92.1 percent accuracy (LM).

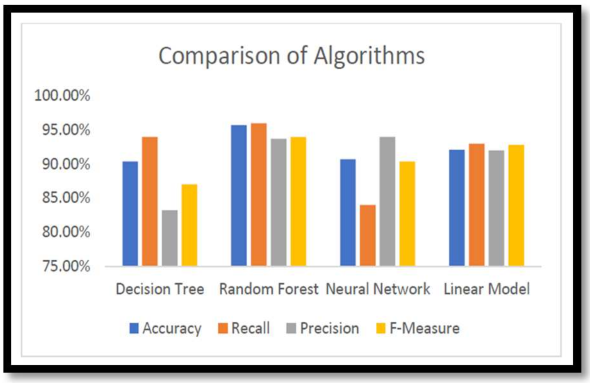

Figure 6: Comparison of Existing Algorithms

Many methods have been proposed for detection using machine learning for intrusion detection [31] [32] [33] [34], concentrating on logs form heterogenous sources. This demonstrates the capability of machine learning algorithms in detection of anomalies and its utilization in numerous zones of computer and network security.

III. DISCUSSION

This section explains the overall concept of this study where it provides a whole vision of the study and to clarify the details and related information that is needed to accomplish the objectives of this research. This survey presented various algorithms and approaches to classify phishing websites by several researchers in anti-phishing techniques. On reviewing the papers, most of the work done is successfully achieved using machine learning algorithms. Hence proving that the use of machine learning is the best technique to analyze and classify based on the latest dataset and the implementation of machine learning provides more advantages comparing to the related works.

Machine learning algorithm is a computation innovative and a subgroup of artificial intelligence extensively recognized as Intelligent Machine Learning. This method works proficiently in a big number of data where it reduces the disadvantage of the new approach and the possibilities for a zero-day attack. Machine Learning based classifiers achieved accuracy more than 99% which proves to be most effective method. Performance be contingent on the scope of the training data, the set of features and the type of classifier. Figure 7 shows an architecture of proposed work presented by Kahksha and Sameena Naaz in previous research.

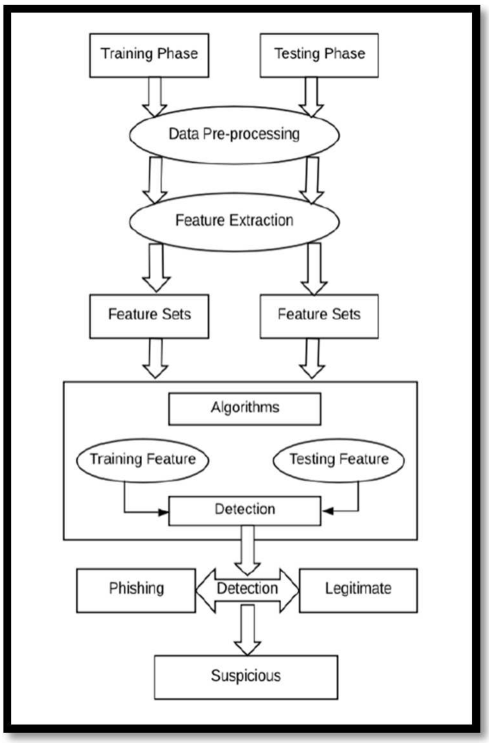

Figure 7 Architecture of proposed work using Machine Learning

IV. CONCLUSION

In a nutshell, phishing detection tools are crucial in ensuring safe online experience to users, to ensure that online users do not become targets to online fraud, to distribute confidential information to an assailant and to effectively use phishing as an attacking device. Unfortunately, a significant number of existing phishing detector devices has error classification caused by a delay in updating the blacklist due to human intervention in classification. Such weaknesses are often caused by the flaws in the precision of detection. These critical problems led many researchers to work on different approaches to improve phishing attacks detection accuracy and to minimize false alarm rates. The inconsistent nature and constantly changing URL patterns of attacks requires the reference model to be updated in due time. The goal of this survey paper is to study and analyze the previous works in order to identify the suitable approach in classifying phishing and legitimate websites. Based on the literature review, Machine Learning can be concluded and has been proven to be the most effective approach to achieve the aim of the paper.

ACKNOWLEDGEMENT

Authors are grateful to the School of Graduate Studies and the Faculty of Information Sciences and Engineering, Management and Science University, Malaysia for their support.

REFERENCES

[1] Narendra. M. Shekokar, C. S. (2015). An Ideal Approach for Detection and Prevention of Phishing Attacks. *Proceedings of 4th International Conference on Advances in Computing, Communication and Control, Elsevier,, Vol. 49.*

[2] Er Purvi Pujara, M. B. (2019). Phishing Website Detection using Machine Learning : A Review.

[3] Stu Sjouwerman. (2019). Phishing Activity Trends Report, 3rd Quarter 2019.

[4] Greg Aaron (2018). Phishing Activity Trends Report, 1st Quarter

[5] Greg Aaron (2018). Phishing Activity Trends Report, 2st Quarter

[6] Routhu Srinivasa Rao, Tatti Vaishnavi, Alwyn Roshan Pais, "CatchPhish.(2019) Detection of Phishing Websites by Inspecting URLs", Journal of Ambient Intelligence and Humanized Computing, Springer, Vol. 10.

[7] Routhu Srinivasa Rao, Alwyn Roshan Pais.(2018) "Detection of Phishing Websites Using an Efficient Feature-Based Machine Learning Framework", Neural Computing and Applications, Springer.

[8] Alkawaz, M. H., Steven, S. J., & Hajamydeen, A. I. (2020). Detecting Phishing Website Using Machine Learning. In 2020 16th IEEE International Colloquium on Signal Processing & Its Applications (CSPA) (pp. 111-114). IEEE.

[9] Gandotra, E., & Gupta, D. (2021). An Efficient Approach for Phishing Detection using Machine Learning. In Multimedia Security (pp. 239-253). Springer, Singapore.

[10] Makkar, A., Kumar, N., Sama, L., Mishra, S., & Samdani, Y. (2021). An Intelligent Phishing Detection Scheme Using Machine Learning. In Proceedings of the Sixth International Conference on Mathematics and Computing (pp. 151-165). Springer, Singapore.

[11] Crane Hassold. (2017). A Quarter of Phishing Attacks are Now Hosted on HTTPS Domains:

[12] Yang, Y. (2019). Effective Phishing Detection usingMachine Learning Approach.

[13] Abbasi, A., Dobolyi, D. G., Vance, A., & Zahedi, F. M. (2021). The Phishing Funnel Model: A Design Artifact to Predict User Susceptibility to Phishing Websites. Information Systems Research.

[14] Mahmoud Khonji, Y. I. (2013). Phishing Detection: A Literature Survey.

[15] Bhawna Sharma, P. S. (2019). PhishAlert: An Efficient Phishing URL Detection via Hybrid Methodology. *International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8*

[16] Harikrishnan NB, Vinayakumar and Soman KP(2018) "A machine learning approach towards Phishing email detection; 2018.

[17] Gururaj Harinahalli Lokesh & Goutham BoreGowda (2020): Phishing website detection based on effective machine learning approach, Journal of Cyber Security Technology

[18] Ushamary Sharma, B. S. (2015). Phishing-An Analysis on the Types, Causes, Preventive Measuresand Case Studies in the Current Situation. National Conference on Advances in Engineering, Technology & Management (AETM'15), *Volume: IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 2278-8727, PP 01-08.*

[19] Vamsee Muppavarapu, A. R. (2018). Phishing Detection using RDF and Random Forests. *The International Arab Journal of Information Technology, Vol. 15, No. 5.*

[20] Kremic E. and Subasi A.,(2015) "Performance of Random Forest and SVM in Face Recognition," *The International Arab Journal of Information Technology*, vol. 13, no. 2, pp. 287-293.

[21] Ali, W. (2017). Phishing Website Detection based on Supervised Machine Learning with Wrapper Features Selection. International Journal of Advanced Computer Science and Applications, 8(9), 72-78.

[22] Subasi, A., Molah, E., Almkallawi, F., & Chaudhery, T. J. (2017). Intelligent phishing website detection using random forest classifier. In Electrical and Computing Technologies and Applications (ICECTA), 2017 International Conference on (pp. 1-5). IEEE.

[23] Samar Muslah Albladi, George R. S. Weir (2018)"User Characteristics that Influence Judgment of Social Engineering Attacks in Social Networks", Human-centric Computing and Information Sciences, SpringerOpen, Vol. 8, No. 1, pp. 1-24.

[24] 32. Tore Pedersen, Christian Johansen, Audun Josang, (2018)"Behavioural Computer Science: An Agenda for Combining Modelling of Human and System Behaviours",

Human-centric Computing and Information Sciences, SpringerOpen, Vol. 8, No. 1, pp. 1-20.

[25] Abdelhamid N, Thabtah F, Abdel-jaber H(2017). Phishing detection: a recent intelligent machine learning comparison based on models content and features. In Beijing, China: IEEE; 2017

[26] Ankit Kumar Jain, B. B. (2017). *Phishing Detection: Analysis of Visual Similarity Based Approaches.* In Springer Science+Business Media, LLC, part of Springer Nature.

[27] Rishikesh Mahajan, I. S. (2018). Phishing Website Detection using Machine Learning Algorithms. *International Journal of Computer Applications(0975-8887)*, Volume 181-No.23.

[28] Ankit Kumar Jain, B. B. (2016). A novel approach to protect against phishing attacks at client side using auto-updated white-list. EURASIP Journal on Information Security, Springer Open,, Vol. 2016, No. 1, pp. 1-11.

[29] Kahksha, S. N. (2019). Detection of Phishing Websites using Machine Learning Approach. *International Conference on Sustainable Computing in Science, Technology & Management (SUSCOM-2019).*

[30] Wenbin Yao, Yuanhao Ding, Xiaoyong Li.(2018).“Deep Learning for Phishing Detection”, 2018 IEEE International Conference on Parallel & Distributed Processing with Applications, Ubiquitous Computing & Communications, Big Data & Cloud Computing, Social Computing & Networking,

[31] Asif-Iqbal, H., Udzir, N. I., Mahmod, R., & Ghani, A. A. A. (2011). Filtering events using clustering in heterogeneous security logs. Information Technology Journal, 10(4), 798-806.

[32] Hajamydeen, A. I., Udzir, N. I., Mahmod, R., & GHANI, A. A. A. (2016). An unsupervised heterogeneous log-based framework for anomaly detection. Turkish Journal of Electrical Engineering & Computer Sciences, 24(3), 1117-1134.

[33] Hajamydeen, A. I., & Udzir, N. I. (2016). A refined filter for UHAD to improve anomaly detection. Security and Communication Networks, 9(14), 2434-2447.

[34] Hajamydeen, A. I., & Udzir, N. I. (2019). A detailed description on unsupervised heterogeneous anomaly based intrusion detection framework. Scalable Computing: Practice and Experience, 20(1), 113-160.