Prediction Of Phishing Websites And Analysis Of Various Classification Techniques

Selvan K

Abstract: Phishing is a mendacious technique used over the Internet. It was invented in the 1996. The phishing attacks are created to swindle the users with the intention of extracting their personal and valuable information. The attackers will make use of the intentionally acquired sensitive information to gain access over the users financial resources such as bank accounts or personal web accounts. Data mining algorithms are most commonly used to identify the phishing websites. The accuracy of the identification process is used to assess the technique employed. This research work is focused on the classification and analyzing the phishing attacks. Four classification algorithms are used to classify the phishing attacks and their performance are evaluated using different performance metrics.

Index Terms: Accuracy, Attack, Classification, K-Nearest Neighbour, Phishing, Precision, Random Forest

1 INTRODUCTION

Internet made human life more sophisticated. Digital technology enables us to perform many operations through online and at any time. On the other side, novice users are being losing their valuable information [11]. One of the prime attack is phishing. Though the financial organizations are the primary targets, attacks over government institutions, automation industries, social media have being increased. Many of us are aware of phishing [15] and careful at the online transactions. But many people are tricked to provide their sensitive information. Phishing emails and spams are due to the absence of authentication mechanisms in the Simple Mail Transfer Protocol (SMTP) [12]. The reports of Anti-Phishing Working Group (APWG) [10], shows that the economy and phishing are closely associated. Financial losses created through phishing attacks are increasing considerably. The economy of the country is affected by the phishing attacks. As per the reports of [14], the total number of phishing attacks detected in first quarter of 2018 was 263,538 which is 46% increase when compared to that of the previous quarter (4Q 2017). Incidents like these pose a serious issue to detect the phishing attacks. The objective of this research work is to classify the phishing attacks either as legitimate or normal. The results will be then analyzed using difference performance measures. This research paper is organized as follows. The review of the literature is elaborated in the section II. The phases of the classification process are explained in the section III. The results and discussion are presented in section IV. The research work is concluded in section V.

2 BACKGROUND STUDY

In the research work [1], Aksu et al. employed deep learning to assess whether the websites are legitimate or phishing websites. The authors used different classification methods such as neural networks and Support Vector Machine (SVM), decision tree and stacked autoencoders. They produced a success rate of 86% using the stacked autoencoders. The Proposed algorithm described in [6] is Machine Learning based automated real-time system. This will automatically detect the phishing attacks. The URLs of the phishing websites usually have connections between the parts. This can be identified with the help of features. The extracted features will aid to identify the phishing websites using Machine Learning classifications. In [13], the authors concluded that no single technique is suitable of detecting the phishing attacks. The authors of [7] insisted on the need for an automated classifier to classify the phishing attacks. They created a classification model using Bayesian statistical classification, J48 Decision tree, Random Forest, K-Nearest Neighbour (KNN) and SVM. Here the classifiers are constructed automatically from the pre-classified sample phishing data set. Among the techniques Random Forest offers better results in terms of accuracy and time efficiency. In [8], the authors recorded that associate classifiers will yield greater classification accuracy than the traditional models. They also identified there is a significant relation between the URL & Domain identity and Security & Encryption criteria of phishing websites. Namasivayam [9], extracted 59 attributes from the website URL, URL redirection, hosting domain and popularity. He created a learning model using decision tree classifiers and neural networks. The model developed is able to classify the attacks without bias towards any features. It can also predict new type of phishing attacks with shortened URLs or newly compromised websites. In [16], the authors introduced a multilayered authentication system using mathematical expression, One Time Password, Session key, Picture Selection, Animated Image Selection, and also usual Username and Password. The Synchronized Feature Vector (SFV) is used as an input dataset to the classifier for distinguishing Phishing web pages from Ordinary web pages, which uses sixteen attributes. The following section will elaborate the classification techniques used for this research work.

3 PROPOSED FRAMEWORK

The research work consists of three phases namely, feature selection, classification and analysis as shown in the figure 1. These phases are explained in the following section.

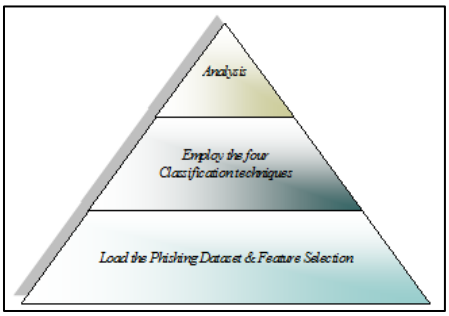

Fig.1 *Phases of the proposed work*

3.1 Feature Selection Phase

The dataset consists of more number of features. Some of them really contribute for the classification process while the others are not. The features used to classify the phishing attacks should be selected from the dataset. Here the features are selected using feature importance. The features influencing the result are identified and listed. The top contributed features are selected for the classification phase. By doing so, the features are reduced.

3.2 Classification Phase

The features are selected using the first phase. The techniques such as Random forest, Decision tree, K-Nearest Neighbour and Support Vector Machine are used to perform the classification. The performance of the classification techniques are then analysed using the third phase. The metrics used for analysis are described in the next section.

3.3 Analysis Phase

The classification techniques are implemented using Java Servlets. Instead of Java, tools like Rapid Miner [5] can also be used to perform the implementation. This is one of the best tools for data mining applications. Based on the outcomes obtained from the implementations, the results are analyzed as described in the following sections.

4 EMPIRICAL STUDY

This section describes the experimental study and provides a detailed discussion about the results obtained. The performance analysis is also presented to visualize the impact of the classification techniques employed to classify the phishing attacks.

4.1 Dataset Description

The dataset used for this research work is downloaded from the kaggle repository [2]. This dataset is a public dataset. The phishing data set consists of 11055 records. There are totally 32 attributes in this dataset. The attributes in the .csv files are listed below. This dataset is uploaded into the repository for public usage in the year 2018.

TABLE I. *ATTRIBUTES OF THE PHISHING DATASET.CSV FILE*

S. No	Attribute Name
1	Index
2	having_IPhaving_IP_Address
3	URLURL_Length
4	Shortining_Service
5	having_At_Symbol
6	double_slash_redirecting
7	Prefix_Suffix
8	having_Sub_Domain
9	SSLfinal_State
10	Domain_registeration_length
11	Favicon
12	Port
13	HTTPS_token
14	Request_URL
15	URL_of_Anchor
16	Links_in_tags
17	SFH
18	Submitting_to_email
19	Abnormal_URL
20	Redirect
21	on_mouseover
22	RightClick
23	popUpWidnow
24	Iframe
25	age_of_domain
26	DNSRecord
27	web_traffic
28	Page_Rank
29	Google_Index
30	Links_pointing_to_page
31	Statistical_report
32	Result

4.2 Advanced Classification Metrics

The outcome is assessed using the confusion matrix. The confusion matrix is given below.

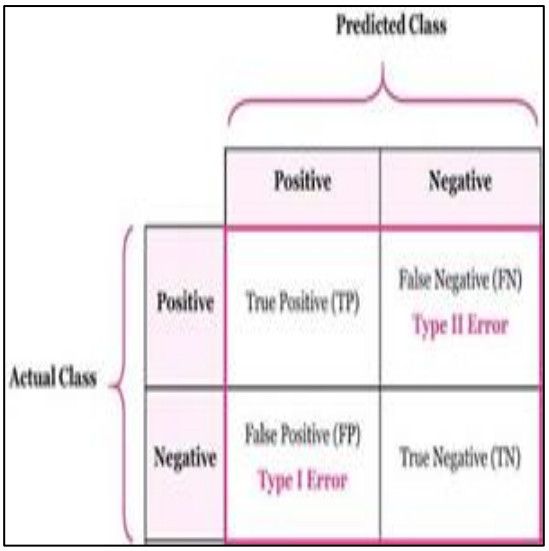

Fig.2 *Confusion Matrix with advanced classification metrics [4]*

The performance of the classification techniques used to classify the phishing dataset are analyzed in terms of the metrics [3] given below.

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

$$Misclassification\ Rate = \frac{(FP + FN)}{(TP + TN + FP + FN)}$$

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

$$F1\ Score = \frac{2}{(\frac{1}{Recall} + \frac{1}{Precision})}$$ Where,

True positive (TP): Number of samples that are correctly classified as phishing websites.
True negative (TN): Number of normal samples that are correctly classified as legitimate websites.
False positive (FP): Number of normal samples that are incorrectly classified as phishing websites.
False negative (FN): Number of attack samples that are incorrectly classified as legitimate websites.

In addition to the above mentioned metrics, the classification is also analyzed in terms of False Positive Rate (FPR), False Discovery Rate (FDR), False Omission Rate (FOR), number of Type I and Type II Errors.

4.3 Results and Discussions

The phishing dataset is classified using the above mentioned four techniques. The accuracy of the four techniques and misclassification rates are represented in the following figures (Fig. 3-6).

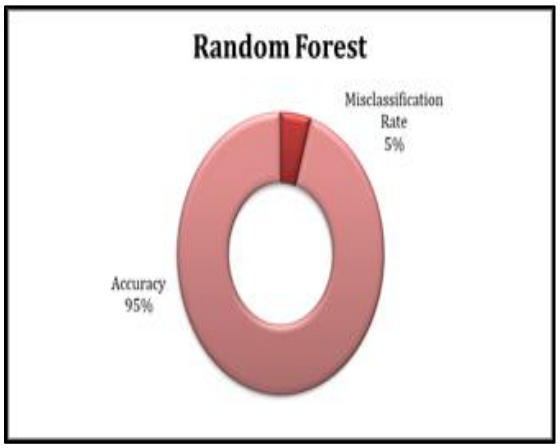

Fig. 3. *Classification Accuracy and Misclassification Rate of Random Forest Technique*

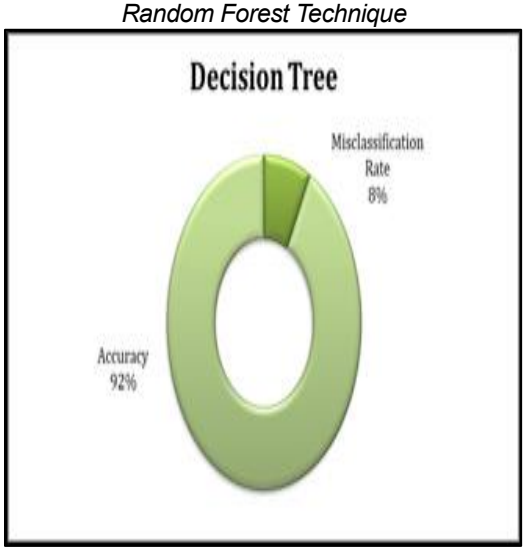

Fig. 4. *Classification Accuracy and Misclassification Rate of Decision Tree Technique*

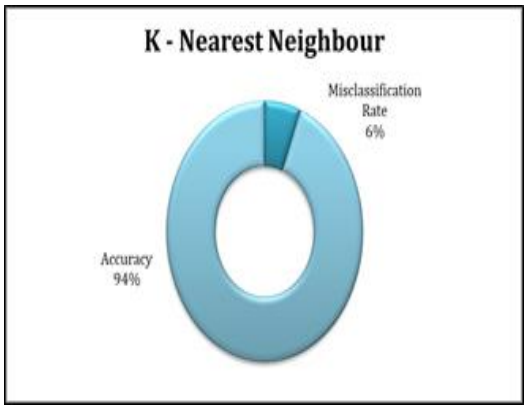

Fig. 5. *Classification Accuracy and Misclassification Rate of K-Nearest Neighbour Technique*

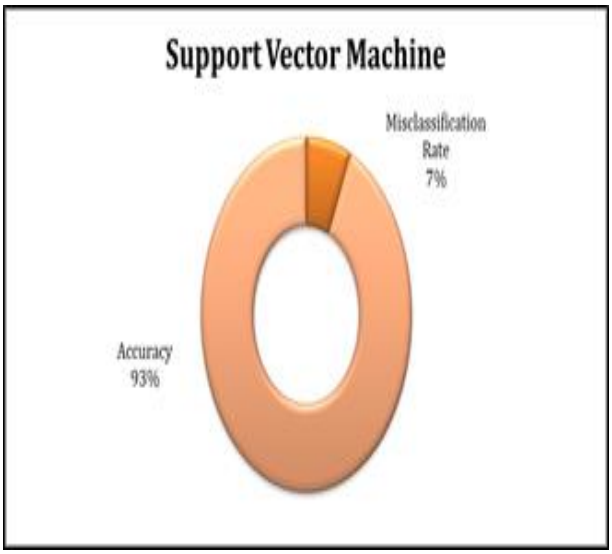

Fig. 6. *Classification Accuracy and Misclassification Rate of Support Vector Machine Technique*

Among the four classification techniques, Random forest is having better classification accuracy and lower misclassification rate. The following table represents the comparison of the classification techniques employed in terms of False Positive Rate (FPR), False Discovery Rate (FDR), False Omission Rate (FOR), Specificity, number of Type I and Type II Errors.

Table II. *Comparison in terms of FPR, FDR, FOR, Sencitivity, Type I and Type II errors*

Metric	*Random Forest*	*Decision Tree*	*KNN*	*SVM*
FPR	*0.03*	*0.10*	*0.05*	*0.04*
FDRD	*0.04*	*0.14*	*0.06*	*0.04*
FOR	*0.06*	*0.03*	*0.06*	*0.09*
Specificity	*0.97*	*0.90*	*0.95*	*0.96*
Type I Error	*177*	*679*	*290*	*209*
Type II Error	*348*	*190*	*368*	*552*

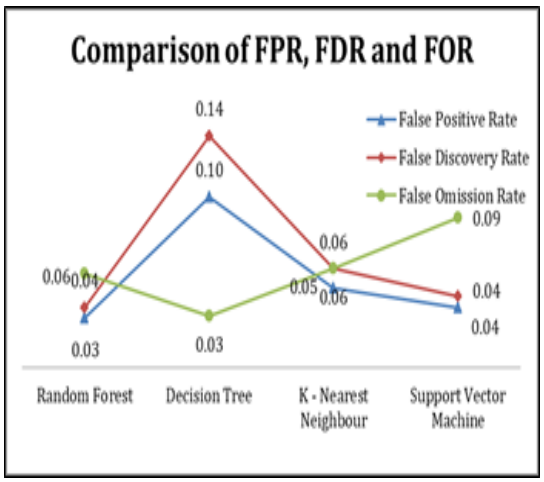

Fig.7 *Comparison in terms of FPR, FDR and FOR*

From the above chart, it is clear that the False Positive Rate (FPR), False Discovery Rate (FDR) is high for the Decision Tree technique. Similarly the False Omission Rate (FOR) of the Support Vector Machine technique is higher when compared to the other three methods used. The specificity of the classification techniques is shown in the following figure 8.

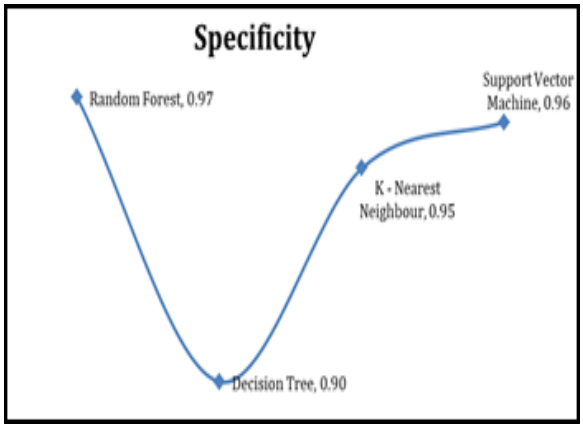

Fig.8 *Comparison in terms of FPR, FDR and FOR*

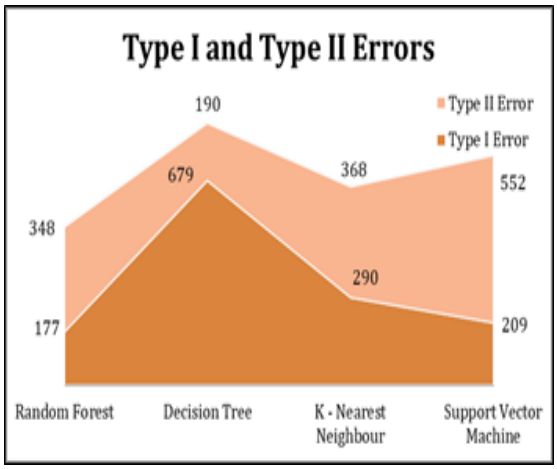

Fig. 9 *Comparison in terms of Type I and II errors*

Figure 9 clearly shows that the number of Type I errors is less when compared to that of the Type II errors for the classification techniques used except the Decision tree technique. Also when compare to the other two methods the Random forest and SVM are having few Type I error than others.

4.4 Performance Analysis

The classification accuracy is the prime metric that supports the justification of using specific classification technique. The performance analysis of the methods are analyzed using the classification accuracy and is depicted in the following figure. Among the four techniques, the Random Forest is having higher accuracy of 95.25%.

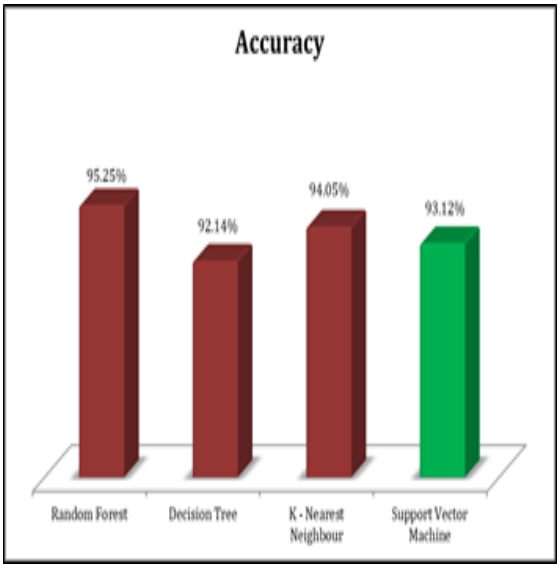

Fig. 10 *Performance Analysis – Accuracy*

The misclassification is inversely related to the classification accuracy. The misclassification rate of the techniques is shown in the following figure. Obliviously Random forest is having minimum misclassification rate (4.75%) when compared to others.

Fig *11Performance Analysis – Misclassification Rate*

The following table represents the Precision, Recall and F1 Score of the four classification techniques. Also, the Precision, Recall and F1 Score of the four classification techniques are depicted in the following figures (Fig 12-14).

Table III. *Precision, Recall and F1 Score of the classification techniques used*

Metric	*Random Forest*	*Decision Tree*	*KNN*	*SVM*
Precision	*0.96*	*0.86*	*0.94*	*0.96*
Recall	*0.93*	*0.96*	*0.93*	*0.89*
F1 Score	*0.95*	*0.91*	*0.93*	*0.92*

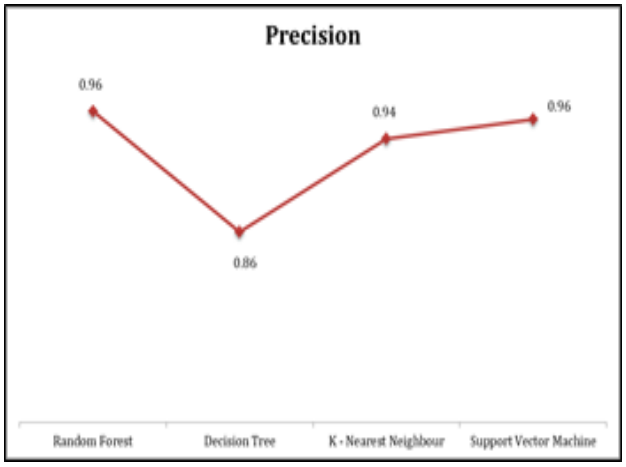

Fig. 12 *Performance Analysis – Precision*

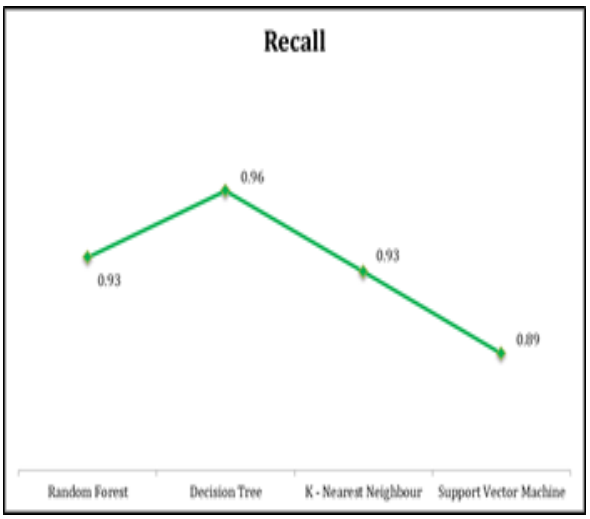

Fig. 13 *Performance Analysis – Recall*

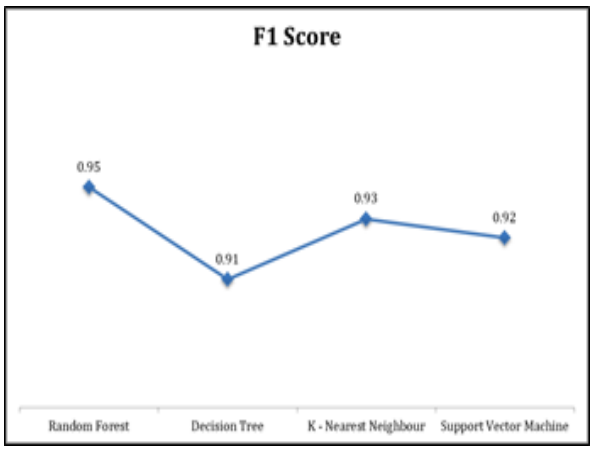

Fig. 14 *Performance Analysis – F1 Score*

5 CONCLUSION

The research work highlights the need for an automatic phishing detection mechanism to overcome the impacts caused due to phishing attacks. The existing works are clearly described in the section 2. The proposed work classifies the phishing attacks of a synthesized dataset using four classification techniques. The accuracy and misclassification rate of the techniques are analyzed. Among them the Random forest algorithm is producing better accuracy when compared to the Decision Tree, K-Nearest Neighbour and Support Vector Machine. Similarly the misclassification rate of Random forest is lesser than the others. The F1 score of the Random Forest and K-Nearest neighbour are closer. This work can be further extended to identify the correlation between the features used to identify phishing.

6 REFERENCES

[1] Aksu D., Turgut Z., Üstebay S., Aydin M.A., "Phishing Analysis of Websites Using Classification Techniques", In: Boyaci A., Ekti A., Aydin M., Yarkan S. (eds) International Telecommunications Conference, Lecture Notes in Electrical Engineering,Springer,Singapore, vol 504, 2019.

[2] Dataset : https://www.kaggle.com/akashkr/phishing-website-dataset accessed on Nov 30,2019.

[3] Cortez, Paulo & Morais, A. " A Data Mining Approach to Predict Forest Fires using Meteorological Data",In: 2007.

[4] https://manisha-sirsat.blogspot.com/2019/04/confusion-matrix.html Last accessed on Nov 30,2019.

[5] https://rapidminer.com/ Last accessed on Nov 15, 2019.

[6] Sneha Mande1, D.S.Thosar, "Detection of Phishing Web Sites Based On Extreme Machine Learning", IJARIIE, Vol-4 Issue-6, 2018, pp. 405-410, ISSN(O)-2395-4396.

[7] Nandhini.S 1, Dr.V.Vasanthi, "Extraction of Features and Classification on Phishing Websites using Web Mining Techniques", International Journal of Engineering Development and Research, Volume 5, Issue 4, 2017. ISSN: 2321-9939.

[8] Aburrous, M., Hossain, M. A., Dahal, K., & Thabtah, F., "Predicting phishing websites using classification mining techniques with experimental case studies", In 2010 Seventh IEEE International Conference on Information Technology: New Generations, April 2010. (pp. 176-181).

[9] Namasivayam, B., "Categorization of Phishing Detection Features" (Doctoral dissertation, Arizona State University), 2017.

[10] APWG. Phishing Activity Trends: Technical report, Anti Phishing Working Group, [online], http://www.antiphishing.org/reports/apwg_trends, 2013.

[11] Jabri, R., & Ibrahim, B., "Phishing websites detection using data mining classification model", Transactions on Machine Learning and Artificial Intelligence, 3(4),2015.

[12] Medvet, E., Kirda, E., and Kruegel, C., "Visual-Similarity-Based Phishing Detection", In Proceedings of the 4th international conference on Security and privacy in communication networks, ACM 22, Istanbul, Turkey, 22 – 25, September, 2008.

[13] Pujara, P., & Chaudhari, M. B. (2018). Phishing Website Detection using Machine Learning: A Review.

[14] APWG report 2, http://docs.apwg.org/reports/apwg_trends_report_q1_2018.pdf

[15] Gaurav Varshney et al, "A survey and classification of web phishing detection schemes", SECURITY AND

COMMUNICATION NETWORKS, Security Comm. Networks 2016; 9:6266–6284, John Wiley & Sons, Ltd.

[16] Selvan. K, Vanitha Muthuraman, “Detection of Phishing web pages based on Features Vector and Prevention using Multi Layered Authentication”, International Journal of Pure and Applied Mathematics, Vol 119, No.: 15, January 2018. ISSN : 1314-3395. PP. 565-573