DETECTION OF PHISHING ATTACKS USING IN WEB ENVIRONMENT USING UNSUPERVISED MACHINE LEARNING

Kamalakanta Shaw[1], Satya Sobhan Panigrahi,[1]Archana Panda[1]

[1]Assistant Professor,[1] Dept. of CSE

[1]Gandhi Institute for Technology, Bhubaneshwar, India

ABSTRACT

Phishing has no fixed complete solution. Phishing means getting information from a person without their knowledge such as user's id, passwords, credentials, etc.,Phishing attacks occur in over every day-to-day life, where one person who does not know the outside world threats give their personal information through emails send by some attackers andthrough SMS send by spammers etc., There are different types of phishing they are vishing(means through phone calls), whaling(attacking a group of targeted people), search engine phishing(using the web for searching online), etc.In this research we perform Phishing attacks detection using machine algorithms include Support vector Machine(SVM) and we constructed dataset from UCI machine Repository of phishing sites and dataset is trained and tested and then apply these classification and regression algorithms to find accuracy of detecting phishing sites. And tools such as Phish Tank are used whether the site is phished or not. MX Toolbox is used to blacklist the phished sites and WHOIS is used to know the details of phished sites. Counter Phishing technique is used to know the unauthorized loins of that websites. This results shows that the Unauthorized logins and by applying various machine learning algorithms the SVM shows highest accuracy in detecting the phished sites.

Keywords:logistic regression, machine learning, Spam detection and E-mail.

1.INTRODUCTION

Phishing is fraudulent activity which involves the use of counterfeit websites by attackers to steal personal and sensitive user details. These may involve email login credentials, one-time password for transactions, bank account username and password, credit & debit card pin numbers and so on. In Phishing, the attacker appears to be a reputable entity and tricks the user into sharing sensitive details. Phishing involves tricking the user to share details with the attacker which makes it a simpler way of breaking into a computer's defence system in comparison to hacking. Phishing attacks are often carried out through e-mails containing spoofed logos with malicious links which appear to be legitimate to an unsuspecting user. Based on data from [1], a new phishing website is created every 20 seconds on the internet. Also, recipients open 70% of the phishing attempts they receive. From [2], 0.47% of bank account holders become targets of phishing attacks each year leading to $2.4M to $9.4M losses per million clients. These statistics reveal the ease with which attackers can target unsuspecting users and the need to have a robust phishing attack detection mechanism.

The process of carrying out a phishing attack is as follows. The attacker mimics the login page of a popular website and registers it with a URL which looks very similar to a legitimate website. An email is then sent to the user with the link of the phishing website. The body of the e-mail is disguised to make it seem legitimate to the person reading it. The user then clicks on the link and enters the login credentials. The login page of the cloned website has a script running at the backend which extracts the credentials entered by the unsuspecting user and makes it available to the attacker. The attacker can then utilize these credentials in the legitimate website and exploit the user. This process is illustrated in Fig. 1.

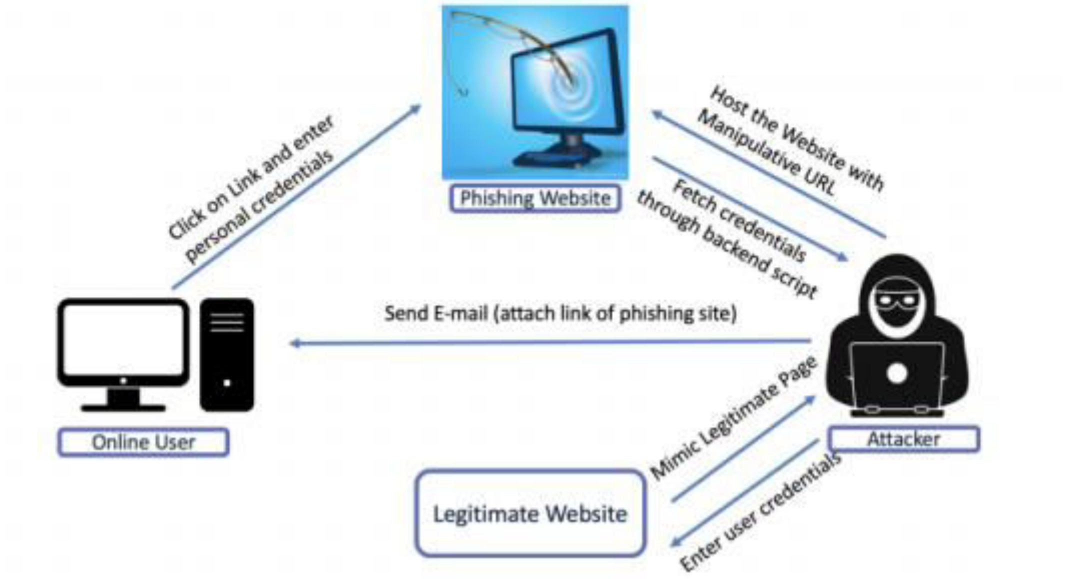
Phishing Website
Host the Website with Manipulative URL
Fetch credentials through backend script
Click on Link and enter personal credentials
Send E-mail (attach link of phishing site)
Online User
Attacker
Mimic Legitimate Page
Enter user credentials
Legitimate Website

Fig. 1: Phishing Mechanism

The commonality of all phishing attacks is the disguise of the website URL. The victims of phishing attacks are most often tricked by the URL of the phishing website. There are 2 methods used by attackers – Cybersquatting and Typosquatting. Cybersquatting is the process of URL hijacking. The attacker buys the domain name of an already established company which does not have a website related to the domain name. Typosquatting refers to buying a website URL similar to a legitimate website but containing a typographical error. An example of this is google.com and goggle.com. Often, internet users make typing errors while entering the website URL which is exploited by attackers. Besides these, the attacker may also choose to manipulate the URL by altering the sub domain names, query lengths, adding redirect requests or making the URL excessively long. Since phishing data is easily available in phishing databases such as Phishtank and OpenPhish, once a website is suspected of being related to phishing, the attacker can easily modify the website URL by altering the sub-domain names to make a new website. Therefore, there is a need for an intelligent method for identifying phishing URLs and reduce phishing attacks. Data mining techniques can help in classification of website URLs into phishing and legitimate URLs.

The major contributions of the paper as follows:

- In this paper, present a general thought of phishing assaults and different conceivable shield plans using SVM classification.
- To develop a method with low computational cost and to analyses its performance under different scenarios.
- To improve the performance and accuracy in terms of the classification and prediction of phishing e-mail in the future.
- To optimize memory consuming in the classifier process and to reduce the time needed for classify the email with unlimited learning, while the characteristics of phishing e-mail features have change.
- To evaluate the proposed framework against using approaches for the purpose of phishing email detection

Rest of the paper is organized as follows; section 2 deals with the various literatures with their drawbacks respectively. Section 3 deals with the detailed analysis of the proposed method with its operation. Section 4 deals with the analysis of the results with the comparison analysis. Section 5 concludes the paper with possible future enhancements.

2. REALTED WORK

Authors of [3] detected phishing websites by using various machine learning algorithms and then compared the accuracy of the different algorithms. Their experimental results indicated at Random Forest algorithm having the highest accuracy, recall and precision. Their work also illustrates the important features that are used for phishing detection. A classification model is proposed in [4] to classify the phishing attacks. Feature extraction was done from various sites based on the UCI Irvine ML repository. The study was carried out using MATLAB and it was found that Extreme Learning algorithm gave the highest accuracy of 95.34%. Authors of [5] demonstrated a method of detecting phishing email attacks using NLP and ML. They performed semantic analysis of text for detecting any kind of malicious activity. NLP was used to parse sentences. This work was mainly carried out using Python and results indicate 95% precision on phishing website classification. The authors of [6] have proposed algorithms for feature selection for phishing detection to improve the quality of the dataset. They compared their algorithm with other commonly used algorithms for classification on basis of accuracy. Their study showed that Tree algorithms didn't work well on diminished datasets. Lazy K Star algorithm showed the best results. The complete study was done using Weka. In [7], a model was created using Random Forest algorithm to classify phishing URLs. The URLs were parsedin order to analyze the feature set. Their algorithm showed 95% accuracy on the test dataset. The authors of [8] used neural networks to extract features from the URL without any specific knowledge about the URL. An accuracy of 94.18% was obtained using Adam optimizer. The authors of [9] did a comparative study between logistic regression techniques with bigrams and deep learning techniques like CNN and CNN-LSTM architectures. The study was done using Tensorflow and Keras with the dataset being drawn up from OpenPhish and Phishtank. The authors reported an accuracy of 98% using CNN-LSTM architecture. In [10], the authors made a proposal of phishing detection model in Chinese Websites. The performance of the model was studied by mining the semantic features of words in Chinese web pages. Different machine learning algorithms like Random Forest, Adaboost and Bagging were compared on the dataset. In [11], the authors have developed an extension to Google Chrome for phishing website detection using machine learning algorithms. The UCI ML Repository has been used for this purpose. The drawback of this extension is that the number of malicious sites is increasing daily with new sites coming up each day and the training set for the study is too small. The authors of [12] have explained about the different URL features such as primary domain, sub domain, and ranking of websites for phishing detection. The authors of [13] developed a tool called PhishScore which does lexical analysis on the URL to detect phishing. They used the relatedness of the URLs in their study for developing the tool. In [14], the authors have explained about web spoofing attacks categories and tried to use domain name features to determine phishing URLs. The authors of [15] have studied the accuracy of different classifiers for prediction of spam emails. Their dataset comprised of 2289 phishing emails comprising of legitimate and phishing websites. They identified a bag of words from the body of the email by text mining which was subsequently utilized for classification using common ML algorithms like random forests, decision trees, SVM, BART and neural networks. Random Forest algorithm was found to be the most accurate among these classifiers. In [16], the authors have gathered 2456 websites which may be classified as legitimate, suspicious and phishing. They have used data mining techniques like Random Forest, Neural Network, Decision Tree and Neural Networks to classify the website into one of the abovementioned categories. The results of the work have been compared with other works on both similar and dissimilar datasets. The limitation of this work is that there is no description of the phishing features taken into consideration for machine learning purpose.

3. MACHINE LEARNING IN E-MAIL CLASSIFICATION

Determining whether a given website URL is phishing or legitimate is a binary classification problem which can be solved with the help of labelled data on which supervised learning can be applied. Data collection for this problem requires recent website URLs belonging to both classes - phishing and legitimate. This is followed by preparation of the dataset by extracting relevant features which helps in distinguishing phishing websites from legitimate websites. The features need to be processed in order to give as input to the machine learning algorithm. Then the model is trained using the training set and its accuracy is determined on the testing set. The flow chart depicting the methodology is summarized in the Fig. 2.

After collecting the dataset, each feature has a threshold value, by using it we can compare the values .If the value is in range then it is indicated as “1”. Or else “0” in binary vector and this is further used in testing. There are six parts

- In this the data is transformed in to tree structure format.
- Rule generation is applied it means applying the “if then else” rule to the tree
- Bayes classifier is used to remove the rules and get the remaining generated rules to next stage.
- Test dataset contains both the legitimate and phishing URLs and calculate their performance
- Here classification is done it declares accuracy of the URL through rules and class label for each URL.
- According to accuracy and performance whether the sites is phished or not is declared

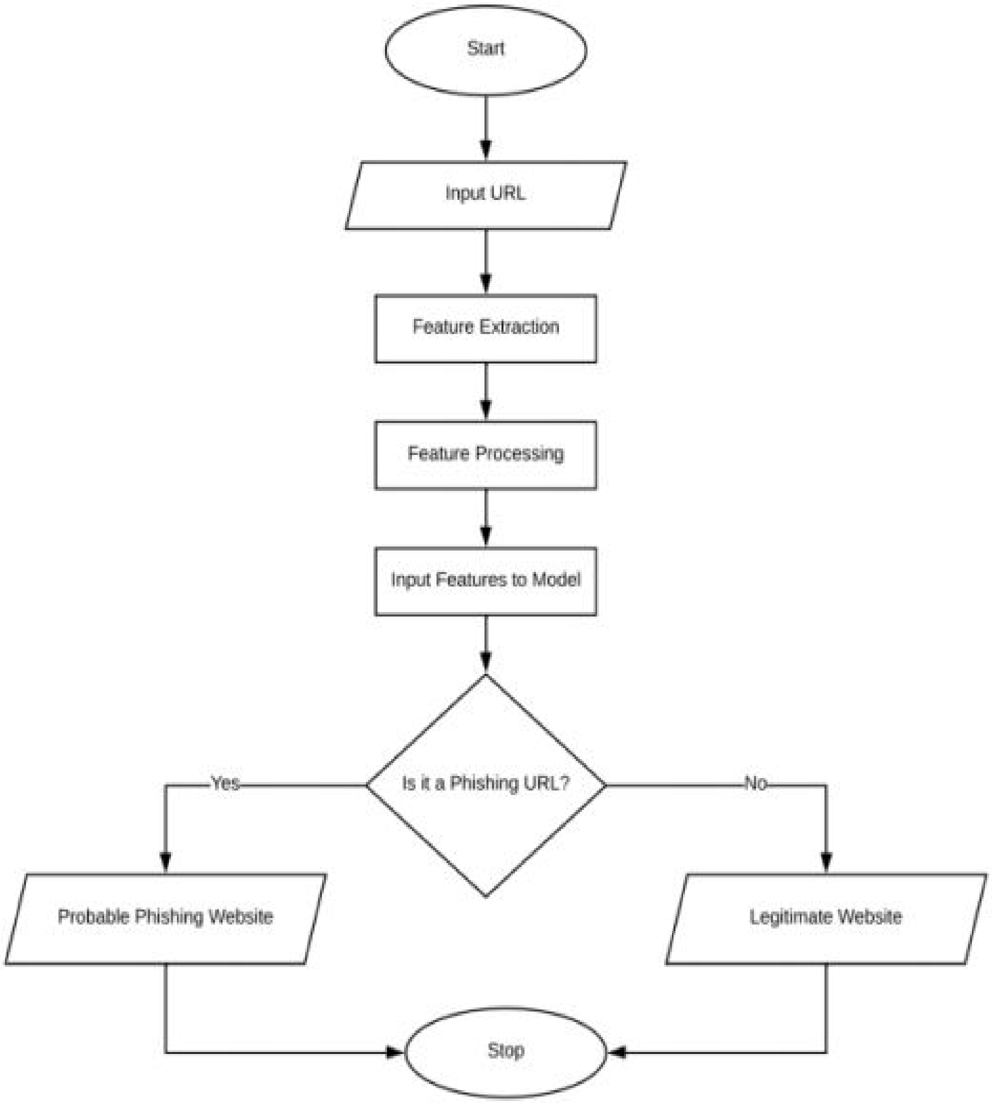

Fig. 2. Methodology

3.1 Feature extraction

A. URL Based Features:

- IP Address: The use of IP address or hexadecimal characters in the domain of the URL instead of a textual domain name can be a probable phishing website. An example of IP address in URL is http://102.24.134.12/page.html. In 46.66% of cases, the use of IP address or hexadecimal characters have been linked to phishing websites and suspicious activity according to a study in [17].
- @ symbol in URL: The use of '@' in the URL causes the browser to disregard all the contents prior to the symbol. Often, the phishing website address follows the '@' symbol. This is often utilized as a means to exploit phishing since users do not often read the full URL. The appearance of '@' has been found in 20% of the phishing websites in the study done in [17].

- HTTPS in the middle of URL: The presence of “https” in the domain of the URL is used by phishers for tricking people. https.paypal.com-account-update.e3d3idw3k4securityalert.cenksen.com.tr/paypal.com/ is an example of how users are deceived by http in URL domain
- Dashes in URL: The presence of ‘-’ in the website URL is a sign of phishing because legitimate web page URL’s don’t use dashes. This is often utilised by phishers to mislead the users by making they feel as though they are dealing with authentic websites. In our study, the average number of dashes in legitimate websites was only 0.06 whereas in phishing websites, this number was 0.49 suggesting the use of dashes predominantly in phishing websites
- Long URL: An abnormally long URL is usually a sign of phishing which is used by phishers to disguise the suspicious part of the URL. In our dataset the average character count of all the URLs was 69.3. The average length of phishing website URL was significantly longer at 92.7 whereas the length of legitimate website URLs was found to be 36.63. This difference in length is a significant feature that differentiates legitimate websites from phishing websites and hence has been taken into consideration in this study. In the study done in [17], the author has found that 73.33% of the phishing URLs have an abnormally long URL.
- Delimiter Characters: The presence of unusual number of delimiter characters such as “#, ~, _, %, &” have been taken into consideration because delimiter characters are predominantly found in suspicious URLs.
- Query Length: The query length portion of phishing websites is significantly longer in comparison to standard legitimate websites which is another feature used in our study for classifying phishing and non-phishing websites.
- Dot Count: The count of dot symbols in the domain of the URL can be used as a marker for identifying suspicious phishing websites. Usually the domain portion of a phishing website is characterized by more dots than a legitimate website. The Urllib library of python along with Pandas has been utilized to count the dots in the URL of the website. The protocol and TLD portions of the URL are omitted while determining the number of dots in the URL. In our study, the average count of dots in the domain of the URL was 0.55. However, in phishing websites the average number of dots was 0.87.

- Sub-domains in URL: Likewise, phishing websites are often characterized by the number of sub-domains in the URL. An unusually high number of sub-domains is a probable phishing website. The average number of subdomains in our study was 1.29. However, in phishing websites, this number came to be close to 1.5.
- Redirect Request: The existence of '//' in the URL website is often a sign of a redirect request. An illustration of this can be seen in the following: "http://www.legitimate.com//http://www.phishing.com". Here, the redirect request follows the double slash. This feature serves as a valuable marker for identifying phishing websites

B. Domain Based Features:

- Page Rank of the Website: The importance of a website is often marked by its page rank. Alexa maintains a database of websites is used to determine the page rank of a given website. A phishing website is usually unranked or very lowly ranked in comparison to legitimate websites. This is again an important differentiating factor between phishing and non-phishing pages.
- Age of the Domain: A phishing website has a very short lifetime in comparison to legitimate websites which are usually up for a very long duration of time. The Whois API helps to determine the age of the domain which is a useful metric for differentiating between phishing and non-phishing pages.
- Validity of the Website: The Google Whois API helps to identify whether the website is still operational or not. Most of the phishing sites have a short lifetime and are pulled down once detected of suspicious activity. The validity is an important marker that separates legitimate website from phishing websites.

3.2 Classification

Support Vector Machine: SVM is very similar to the decision tree i.e., it is supervised machine leaning algorithm. It is also used for both the classification and regression. In this we are using SVM for classification purpose. For this we plot the dataset in the dimensional space format with each value to the feature being in particular. Then we perform classification and find the differences in 2 layers.In this, we consider the dataset in 2 parts: training dataset and testing dataset. In training state, we give 33 featured values and are organised in particular format in SVM. At the testing state, we compare the values in two layers. The output values indicate whether the URLs are phished or not. If the result is equal to "-1" then it is phished site or if the result is equal to "1" then it is non-phished site.

3.3 Description of Tools:

The tools that are used in this paper are WHOIS, MX Toolbox, Phish Tank. In this paper, the main aim of tools is to tell which tool is better to use at which time. WHOIS: whois is a query and response protocol tool. It is used mostely to know whteher a webiste is legal or illegal. As in whois.com, many users register their website or their internet resouruces in it. If we want to know any domain information about a legal website like the IP address etc., It provides the information in fraction of seconds in human readable format. The main advantgae of the whois, it tells whether a website is legall or not and gives information about a legal webiste.

MX Toolbox: MX Toolbox is a phishing tool that provides information likewise HOSI, but it also provides protection. It checks if any website is under blacklist, if true or having any problem with email delivery and not getting any help. MX Toolbox provides a solution by offering blacklist protection. In this we can add blacklists that are newly added or created. Phish Tank: Phish Tank is an anti-phishing site. It is based on community verification i.e., in this site we upload a website and others users in the site vote whether it is a phished site or not. If maximum people vote that it is a phished site then it comes under phished site or else the site is nonphished site.

3.4 Technical Approach:

In phishing, there are different types of phishing attacks without the knowledge of the people by just one click. There are different ways of obtaining personal information from users. As technology is increasing day-by-day cybercrime is also increasing drastically. To prevent from being phished people should have the knowledge of phishing techniques in present world and how they work. There are different types of phishing techniques some of them are Spear-phishing, email spam, etc., in this paper, counter phishing technique is demonstrated on how personal information is gone to attackers.In counter phishing, a duplicate link is created and sent to a lot of random people through mails and without knowledge users input their personal details like email Id, passwords. This is a lot of information to the attackers to attack.

4. EXPERIMENTAL ANALYSIS

4.1 Dataset

The dataset is been taken from the website i.e., UCI machine learning repository. The file contains Dataset of phishing websites which is available in .arff format and needs to be converted to .csv format for the classification. In the dataset there are 32 columns and 11,056 rows. Each row represents an attribute that can be a part of each phished websites. And the values 1 refers to the success state, 0 refers to the not detected state , 1 refers to the failure state. Each attribute represents its classification taken from the phished sites. Some of the attributes are

- Using Ip address- In case if an IP address is used as an alternative for a website URL such as, **https://192.16.10.564.html** means a phished website that can be used for stealing the user information.
- Tiny URL- these are used in some of the sites such as https://bit.ly.3dfe6sd.html which seems to be redirecting to the original site but can lead to phished site.
- HTTPs token- phishers may add the https token in the URL domain which seems to be normal website to trick the users.
- Domain registration length-based on the present scenarios, each phished site can only be livedfor a very short time, so secure domains pay a high amount to be used for a long time periods.
- Abnormal URL-host name which needs to be added to each website to gain the trust worthiness. So, if not added can be considered as a part of the malicious site.
- Website forwarding-few websites in the internet makes user to redirect from page to page when selects a link, but redirects to another website which can use some issue to the system and can access users information.
- Popup window-it is unusual to ask user credentials in a popup window. On the other hand, this feature has been used to gain.

4.2 Performance classification

In this study, seven machine learning algorithms have been tested on the dataset. The results of the different algorithms have been tabulated below. The accuracy of different algorithms was determined for the original feature set as well as for PCA applied feature set as depicted in Table I. The accuracy of Random Forest algorithm was identified as the best algorithm with 95.82% after PCA. Logistic regression was the worst performing algorithm amongst all of models. Decision Tree, Gradient Boosting, Fuzzy Pattern Classifier and Adaboost gave accuracies which were close to Random Forest Algorithm.

Table 1: Performance comparison

Algorithm	Accuracy	Sensitivity	Specificity	Precision	Recall	F1
SVM	95.33	95.82	95.00	96.00	95.00	96.00
Decision Tree	94.09	94.26	94.00	94.00	94.00	94.00
GradientBoosting	92.19	92.22	92.00	92.00	92.00	92.00

Fuzzy PatternTree	91.22	92.23	91.00	92.00	91.00	91.00
Adaboost	91.00	90.64	91.00	91.00	91.00	91.00
Gaussian NB	83.28	85.17	83.00	85.00	84.00	86.00
Logistic Regression	73.78	82.89	74.00	83.00	74.00	84.00

The recall and precision of the different algorithms was determined in the same manner as accuracy for the original feature set and PCA applied dataset respectively. The precision and recall for Random Forest algorithm stood at 96% which is the best among all the testedalgorithms. Application of PCA on the dataset improved the precision and recall by a slight margin of 1% over the base precision and recall.Random Forest algorithm showed the highest F1 score among all the algorithms under study. The overall conclusion based on the results is that Random Forest is reasonably good model for classification of website URLs into phishing and legitimate websites.

CONCLUSIONS

This paper explains the existing security problems in today's digital world with respect to phishing and the process through which phishing is carried out. Phishing is a serious security concern which may lead to loss of sensitive personal information due to clever disguising of phishing mails by attackers. This work mainly focuses on identifying features useful for detecting phishing websites based on solely the URL of the website and applying machine learning algorithms to classify websites into legitimate and phishing. The study involves comparison of results of 7 machine learning algorithms with Random Forest algorithm emerging as the most accurate and hence, most suited algorithm for this binary classification.The accuracy of the model can be further improved by considering more features such as favicons, meta tags, pop-up windows, redirect links on the web page and so on. Also, the number of training examples can be increased to make the model more robust. Phishing will remain a prevalent issue and a web browser extension to block phishing websites can prove to be very helpful to users.

REFERENCES

[1]. Yahaya, Saudi, Madihah ,Abdullah, Ismail. "A Review and Proof of Concept for Phishing Scam Detection and Response using Apoptosis". International Journal of Advanced Computer Science and Applications (IJACSA) ,Volume 8,Issue 6.

[2]. S .Laidlaw and M .Hillick, "Profiling cyber threats detected in a target environment and automatically generating one or more rule bases for an expert system usable to

profile cyber threats detected in a target environment“. U.S. Patent 9,503,472. Cyberlytic Limited, 2016.

[3]. Routhu Srinivasa Rao∗ and Syed Taqi Ali, “PhishShield: A Desktop Application to Detect Phishing Webpages through Heuristic Approach”, Eleventh International Multi-Conference on Information Processing-2015 (IMCIP-2015). Volume 54, Pages 147-156

[4]. Namrata Singh, Nihar Ranjan Roy, “A Hybrid Approach to Detect Zero Day Phishing Websites”, International Journal of Information & Computation Technology. ISSN 0974-2239 Volume 4, Number 17 (2014), pp. 1761-1770

[5]. Ratinder Kaur and Maninder Singh, “A Hybrid Real-time Zero-day Attack Detection and Analysis System”, I. J. Computer Network and Information Security, 2015, Volume 9, 19-31.

[6]. N. M. Shekokar, C. Shah, M. Mahajan, S. Rachh, “An Ideal Approach for Detection and Prevention of Phishing Attacks”, Procedia Computer Science Volumn 49 (2015) page no. 82 – 91.

[7]. Mouna Jouini, Latifa Ben Arfa Rabai, Anis Ben Aissa, “Classification of security threats in information systems”, 5th International Conference on Ambient Systems, Networks and Technologies (ANT-2014).Volumn 32,page no. 489-496

[8]. A. Mishr and B. B. Gupta, “Hybrid Solution to Detect and Filter Zero-day Phishing Attacks”, Emerging research in computing, information, communication and applications, ERCICA 2014. page no.373-379

[9]. Neda Abdelhamid, “Multi-label rules for phishing classification”, Applied Computing and Informatics (2015) volume 11, page no.29–46.

[10]. MAAWG (2011). Messaging Anti-Abuse Working Group (MAAWG) Email Metrics Program. 15. third Quarter.

[11]. Singh, D. K., & Ashraf, M. (2019). Detect the phishing websites in the contex of internet security by using machine learning approach. International Journal of Advanced Science and Technology, 27(1), 104-111.

[12]. APWG (2010). “Phishing Activity Trends Report“.From http://www.antiphishing.org/reports/apwg_report_Q1_2010.pdf.

[13]. GARTNER (2007). “Gartner Survey Shows Phishing Attacks Escalated in 2007; More than $3 Billion Lost to These Attacks.” Retrieved December 17,from http://www.gartner.com/it/page.jsp?id=565125.

[14]. Bimal Parmar, F. (2012). “Protecting against spear-phishing.” Computer Fraud & Security 2012(1): page no 8-11.

[15]. Steve Sheng,1 Mandy ,Holbrook, Ponnurangam Kumaraguru, Lorrie Cranor,Julie Downs, “Who Falls for Phish? A Demographic Analysis of Phishing Susceptibility and Effectiveness of Interventions”, Copyright 2010 ACM.

[16]. Christy Jeba Malar, A., Kanmani, R., Vijayavarman, R., PraveenKumar, R., & Poorna Bharathi, G. (2020). Implementation of phishing detection using SVM. Test Engineering and Management, 83, 3287-3295.

[17]. Ali Darwish, Ahmed El Zarka and Fadi Aloul,” Towards Understanding Phishing Victims' Profile”,2013 IEEE.

[18]. https://blog.returnpath.com/10-tips-on-how-to-identifya-phishing-or-spoofing- email

[19]. http://www.techrepublic.com/blog/10-things/10-tipsfor-spotting-a-phishing-email

[20]. Dhamdhere V. , P. Joeg ,” To Study of Phishing Attacks and User Behavior”, International Conference on Inventive Computation Technologies (ICICT 2017)

[21]. Dhamdhere V., P. Joeg ,” A Study User Behavior Using Phishing Education and Training”, International Journal of Engineering Research in Computer Science and Engineering,2017,pp. 50- 55.

[22]. Dhamdhere V, S. Vanjale,” A novel approach for phishing email real time classification using kmean algorithm, International Journal of Electrical and Computer Engineering,2018,pp.5326- 5332.

[23]. Dhamdhere V, S. Vanjale,”PHISH SAFE GUARD-Phishing Detection: Enhance Anti-Phishing System Using Machine Learning Algorithm”, International Journal of Engineering and Advanced Technology, 2019, pp. 1668-1671.

[24]. Dhamdhere V, S. Vanjale,”To Enhance Phishing Emails Classification using Machine Learning Algorithm.”, International Journal of Recent Technology and Engineering, 2019, pp.2240-2242.