Review on the Utilization of Machine Learning Techniques for Detecting Phishing Websites

Rashmi R M[1], Dr. Mohammed Rafi[2]

[1]PG Student, UBDTCE, Davangere, Karnataka, Email id: rmrashmikanth@gmail.com

[2]Professor, UBDTCE, Davangere, Karnataka, Email id: mdrafi2km@yahoo.com

ABSTRACT:

Rapid growth of the Internet made life much easier and unsafe. Usage of social media platform such as Facebook, Instagram, YouTube, Twitter, LinkedIn and many more for communication with each other is increasing and providing the required entertainment. Enchasing this dependency on Internet, attacker steals the sensitive information of users and use the credentials details for their profit. Phishing is one such cyber-attack that uses email as weapon. The goal is to trick the email recipient into believing that the message is genuine and promotes to share their information. Phishing website looks legitimate and traps the innocent users and gets information such as password, username, credit card details, phone number and many more details from the users. This paper reviews machine learning methods used for detecting the phishing website. Algorithms like Random forest, Decision tree, Support Vector Machine(SVM), Naive Baye's are used in detecting. Some basic general concepts of cyber security and phishing are also included.

Index/Keywords: Cyber Security, Phishing website, Machine Learning, Supervised learning.

I. INTRODUCTION:

Cyber Security is a practice of defending any devices which are connected to Internet against malicious attacks. Pillar of cyber security is to provide CIA where C-Confidentiality, I-Integrity, A-Availability. Cyber security relates to information security, data security, protection against risk. Under many types of attack in cyber security is the Phishing attack. Now a days Phishing becomes a main area of concern for security researchers because it is not difficult to create the fake website which looks so close to legitimate website. Experts can identify fake websites but not all the users can identify the fake website and such users become the victim of phishing attack. Main aim of the attacker is to steal banks account credentials. Phishing attacks are becoming successful because lack of user awareness [1]. Phishing is that the fraudulent plan to obtain sensitive information like username, password, and credit card details, often malicious purposes, by disguising as a trustworthy entity in an electronic communication. 'Phishing' recorded on 2nd January, 1996 according to Internet records. Social media phishing is when attackers use social networking sites like Facebook, Twitter, and Instagram rather than email to obtain your sensitive personal information or click. In this attack phishers use fake websites and emails to expose a user's sensitive private information. They plan to create a uniform false copy of an ingenious website. The rapid growth of Information Technology indeed created many connivances to us, but on the other hand it also resulted and increased security challenges to us to protect our information securely especially from social engineering

attack now a day. Phishing is the major security threats faced by the cyber-world and could lead to financial losses for both industries and individuals. In this attack, Phisher makes a fake web page by copying contents of the legitimate page, so that a user cannot differentiate between phishing and legitimate sites. Social engineering schemes prey on unwary victims by fooling them into believing they are dealing with a trusted, legitimate party, such as by using deceptive email addresses and email messages [2].

II. BACKGROUND THEORY: PHISHING LIFE CYCLE

The number of phishing attacks has been growing considerably in recent years and is considered as one of the most dangerous modern internet crimes, which may lead individuals to lose confidence in e-commerce. Consequently, it has a tremendous negative effect on online commerce, marketing efforts, organization income, relationships, customers, and overall business operations. In order to steal the user identities and credentials, the phisher usually develops a fake replica of the original website, which is similar in appearance to the original website. Subsequently, the phisher sends a forged email to victims in order to criminally perform fraudulent financial transactions on behalf of the web users. Basically, the phisher constantly sends emails to many Web users including hyperlinks to the forged website in as attempt to deceive Web users. As most of Web users are not specialists in Internet security, they follow the link in the phishing email and log in to the fake website. Thus, they would simply fall into the phishing website trap and credentials information such as account information, passwords, and credit card numbers would fall under the control of the phisher [3].

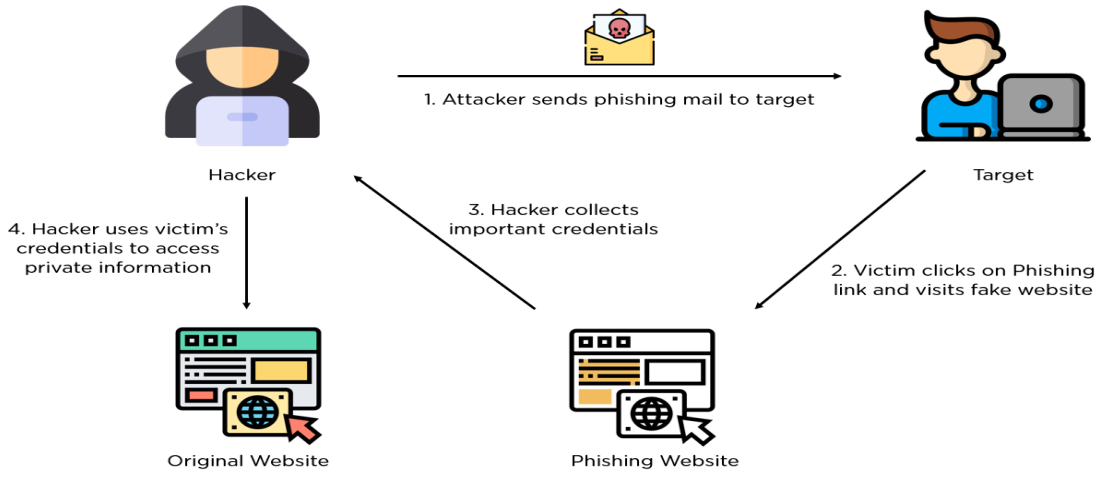

Fig 1: Steps of Web Phishing Process [11]

Types of Phishing Scams:

Deceptive Phishing: Deceptive phishing refers to any attack by fraudsters impersonate a legitimate company and plan to steal people's personal information or login credentials.

Spear Phishing: It is method of sending a Phishing messages to a particular organization to gain organizational information for more targeted social engineering. For example, Social media sites like LinkedIn.

CEO Fraud: Phishers use an email address similar to that of an authority to request payments or data from others within in the company. For example, Use CEO ID.

Pharming: In this attack phisher hijack a website's domain name and use it to redirect visitors to a fake site. Pharmer targets a DNS server and changes the IP address.

Dropbox Phishing: Many people use Dropbox every day to backup, access and share their files. The attackers would attempt to maximize the platforms popularity by targeting users with phishing email.

Google Docs Phishing: Phishers could choose to target Google Drive similar to the way they might prey upon Dropbox users. Specially, as Google Drive supports documents, spreadsheets, presentations, photos and even entire website, phishers can misuse the service to create a web page that mimics the Google account log-in screen [2].

The components for detection and classification of phishing websites are as follows: [1]

1. Address Bar based Features
2. Abnormal Based Features
3. HTML and JavaScript Based Features
4. Domain Based Features

III. RELATED WORK:

Asadullah Safi et al. [4] carried out a comprehensive review methodology including constructing research questions, identifying the list of electronic databases to be explored, data collection, data analysis, discussion on findings, and a comparison study of final selected research articles. The procedure includes searching primary and secondary databases, implementing inclusion-exclusion criteria, analyzing results, and discussions. To identify and prevent phishing attacks, various anti-phishing methods are available. As illustrated, it is classified into five groups in this work.

Fig 2. Phishing Classification

Atharva Deshpande et al. [1] have reviewed the traditional approaches to phishing detection namely blacklist and heuristic evaluation methods. They have tested two machine learning algorithms on the 'Phishing Websites Dataset' to check whether a URL is legitimate or fraudulent. First algorithm used is the Random forest algorithm which creates the forest with number of decision trees and the higher number of tree gives higher detection accuracy. Bootstrap method is used for creation of tress in which features and samples of dataset are randomly selected to construct a single tree. Second algorithm used is the Decision tree which begins its work by choosing best splitter from the available attributes for classification in turn treated as a root of the tree, works until it finds the leaf node. Detecting phishing websites using Random Forest algorithm gave accuracy of 97.31%.

Mr. B. Ravi Raju et al. [5] addressed methodology focusing on detecting phishing attacks using the properties of phishing websites, the Blacklist, and the WHOIS database. For distinguish between real and faked web pages -URLs, domain identification, security & encryption, source code, page style and contents, web address bar, and social human component characteristics need to be checked. IP addresses, long URL addresses, adding a prefix or suffix, redirecting with the sign "//," and URLs with the symbol "@" are among the features of URLs and domain names that are verified. As Atharva Deshpande et al. [1] highlighted the algorithms of Random search and Decision tree, a Support Vector Machine(SVM) has been added by the author in this to detect phishing websites. In SVM each input item is displayed as a point in n-dimensional space and the algorithm creates a separating line for classification of two classes, which is known as a hyperplane. The SVM looks for the closest points, which are called support vectors, and then constructs a line linking then creates a separation line that is perpendicular to and bisects the connecting line. The margin should be as large as possible in order to accurately classify data. The margin is the distance between the hyperplane and support vectors in this case. Because it is impossible to separate complicated and nonlinear data in real life, the support vector machine employs a kernel approach that converts lower dimension space to higher dimensional space to tackle this difficulty.

Rishikesh Mahajan et al. [6] conducted a systematic review to detect phishing URLs and finding best machine learning algorithm by comparing accuracy rate, false positive and false negative rate of each algorithm. Implemented python program to extract features from URL. The features that extracted for

detection of phishing URLs are like- Presence of IP address in URL, Presence of @ symbol in URL, Number of dots in Hostname, Prefix or Suffix separated by (-) to domain, URL redirection, HTTPS token in URL, Information submission to email, URL Shortening Services "TinyURL", Length of host name, Presence of sensitive words in URL, Number of slash in URL, Presence of Unicode in URL, Age of SSL certificate, URL of anchor, IFRAME, Website rank. The author used Scikit-learn tool to import Machine learning algorithms. The training set and testing set are divided in 50:50, 70:30 and 90:10 ratios respectively. Each classifier is used to train the training dataset and test datasets are used to measure the performances. 97.14% of accuracy has been found using random forest algorithm with lowest false positive rate.

Arun Kulkarni et al. [7] aimed to contribute to the goal of evaluating the performance of the commonly used machine learning algorithms on the data set. The author worked on SVM, Naïve Bayes' classifier, decision tree, and neural network testing. Implemented four classifiers using MATLAB scripts. General Neural networks are non-parametric classifiers and a powerful alternative to statistical classifiers. Neural networks can learn with a training set data and make decisions. MATLAB script is used for neural networks. The three layers are input layer, hidden layer and output layer. The number of units in the input layer is equal to number of features, and number of units in the output layer is equal to number of classes. The backpropagation algorithm is a well-known algorithm. The classifiers were used to detect phishing URLs. In detecting phishing URLs, there are two steps. The first step is to extract features from the URLs, and the second step is to classify URLs using the model that has been developed with the help of the training set data. Over fitting is the main concerns in the decision tree classifiers.

Waleed Ali [3] proposed two main categories used for features evaluation like wrapper-based evaluation and filter-based evaluation. In the filter-based evaluation techniques, the high dependency on target class and less inter-correlation are used to select the important features in order to be utilized later in a classification or a regression model. Information gain (IG) is one of the most common filter-based techniques. In the wrapper-based evaluation, an inductive classifier and search algorithm is used to evaluate the significance of the features subset along with to search through the space of possible features and evaluate each subset by running a model on the subset. Popular machine learning techniques such as Back-Propagation Neural Network (BPNN), Radial Basis Function Network (RBFN), Support Vector Machine (SVM), Naïve Bayes classifier (NB), Decision Tree (C4.5), Random Forest (RF), and k-Nearest Neighbor (kNN) are focused by the author. The experimental results showed that BPNN, kNN and RF achieved the best CCR (Correct Classification Rate) while RBFN and NB achieved the worst CCR for detecting the phishing websites.

Er Purvi Pujara et al. [8] provided a comprehensive literature review using various anti phishing solutions such as Blacklist, heuristic, visual similarity, machine learning. Machine Learning based classifiers are efficient classifiers which achieved accuracy more than 99%. Performance depends on size of training data,

feature set, and type of classifier. Limitation of this is it fails to detect when attacker use compromised domain for hosting their site. At end the author has concluded that tree-based classifiers in machine learning approach is best suitable than other.

P. Amba Bhavani et al. [9] provided a brief overview of various methodologies such as Logistic regression, XGBoost algorithm, CNN-LSTM algorithm, CNN BI-LSTM algorithm and Dataset. A regression model where the dependent variable (DV) is categorical is defined as a Logistic regression. It is used to predict a result with two values, such as 0 or 1, pass or fail, yes or no, and so on. Instead of a linear function, this cost function is called the 'Sigmoid function' or 'logistic function'. XG Boost stands for eXtreme Gradient Boosting. It incorporates additional techniques to correct faults in previously presented models. CNN and LSTM integration is a typical notion for integrating benefits due to the accessibility of CNN and LSTM. The notion for a novel deep learning scheme was proposed in this work by integrating CNN and LSTM. The CNN-LSTM URL, a web page code, a text function, and a rapid grading result are integrated to generate multidimensional features. A recurrent NN is a kind of Bidirectional Long Short-Term Memory. It consists of two hidden layers that process data in both forward and backward directions. From Kaggle.com the datasets are extracted to use in training the model. The Logistic regression and XGBoost algorithm model achieve a good accuracy of 92% in detecting the phishing URL's.

IV. RELATED WORK SUMMARY:

In the paper of Atharva et al. some of the traditional approaches to phishing detection were used; namely blacklist and heuristic evaluation methods, and their drawbacks. They detected phishing websites using Random Forest algorithm with an accuracy of 97.31%. From the paper of Asadullah Safi et al. with the evolution of Convolution Neural Network (CNN), the Accuracy of the CNN algorithm is the best, i.e., 99.98%. Mr. B. Ravi Raju et al. paper suggested methodology focuses on detecting phishing attacks using the properties of phishing websites, the Blacklist, and the WHOIS database. Few criteria can be utilized to distinguish between real and faked web pages, according to experts. URLs, domain identification, security & encryption, source code, page and contents, web address bar, and social human component are only a few of the features that have been chosen. Also Rishikesh Mahajan et al. paper aims to enhance detection method to detect phishing websites using machine learning technology. They achieved 97.14% detection accuracy using random forest algorithm with lowest false positive rate. Also result shows that classifiers give better performance when we used more data as training data. In the work of Arun Kulkarni et al, they implemented four classifiers using MATLAB scripts, which are the decision tree, Naïve Bayes' classifier, Support Vector Machine (SVM), and the Neural Network. The pruned decision tree provided the highest classification accuracy 90.39 percent. with more features in the data set it may be possible to obtain higher accuracy. In Waleed Ali paper, the wrapper-based features selection method was used for selecting the most significant features to be utilized in predicting the phishing websites accurately. The experimental results showed that

BPNN, kNN and RF achieved the best CCR while RBFN and NB achieved the worst CCR for detecting the phishing websites. More significantly, the machine learning classifiers using wrapper-based features selection outperformed the machine learning classifiers with PCA and IG features selection methods. In the paper of P. Amba Bhavani et al, the issue of phishing attacks are considered and thus proposed a constructive model using CNN LSTM , CNN-Bi-LSTM, Logistic regression and XGBoost algorithms which combined machine learning mechanism and deep neural networks in data science to detect and classify the illegitimate URL's. Analysis results show the adequacy of the model, and results into 92% accuracy.

V. CONCLUSION

Today's Internet services tremendously changed the lives of people. In the busy schedule people are highly dependent on the online services where user need to register and give their credentials for using the services. These credentials information are protected by network security technology. However, cybercriminals use different techniques to attack and steal the sensitive details. Phishing stands in the top attacks of cyber. The work done in this review paper involves the systematic literature survey of those studies which analyzed the performances of phishing website detection techniques. This paper describes detailed highlights of various literature survey about phishing website detection using Machine Learning. It is observed from various papers that Random Forest algorithm has an accuracy of 97.31% and the accuracy of the CNN algorithm is the best, i.e., 99.98% among all the studies included in this survey. According to this, Machine Learning is the efficient technique to detect phishing website.

REFERENCES:

[1]. Atharva Deshpande, Omkar Pedamkar, Nachiket Chaudhary, Dr. Swapna Borde "Detection of Phishing Websites using Machine Learning", International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181 IJERTV10IS050235, Vol. 10 Issue 05, May-2021

[2]. Oza Pranali P, Deepak Upadhyay "Review on Phishing Sites Detection Techniques", International Journal of Engineering Research & Technology, Vol. 9 Issue 04, April-2020

[3]. Waleed Ali "Phishing Website Detection based on Supervised Machine Learning with Wrapper Features Selection", International Journal of Advassnced Computer Science and Applications, Vol. 8, No. 9, January 2017.

[4]. Safi, A., Singh, S., "A Systematic Literature Review on Phishing Website Detection Techniques", Journal of King Saud University - Computer and Information Sciences, 2023.

[5]. Mr. B. Ravi Raju, S. Sai Likhitha, N. Deepa, S. Sushma "Survey on Phishing Websites Detection using Machine Learning", International Journal for Research in Applied Science & Engineering Technology, ISSN: 2321-9653; IC Value: 45.98.

[6]. Rishikesh Mahajan, Irfan Siddavatam "Phishing Website Detection using Machine Learning Algorithms", International Journal of Computer Applications, Volume 181 – No. 23, October 2018

[7]. Arun Kulkarni, Leonard L. Brown "Phishing Websites Detection using Machine Learning", International Journal of Advanced Computer Science and Applications, Volume. 10, No. 7, 2019

[8]. Purvi Pujara, M. B.Chaudhari "Phishing Website Detection using Machine Learning : A Review", International Journal of Scientific Research in Computer Science, Engineering and Information Technology Volume 3, Issue 7.

[9]. P. Amba Bhavani, Chalamala Madhumitha, Pinnam Sree Likhitha, "Phishing Website Detection using Machine Learning".

[10]. Jain, Ankit Kumar, and B. B. Gupta." Comparative analysis of features-based machine learning approaches for phishing detection." Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on.IEEE, 2016.

[11]. https://www.simplilearn.com/tutorials/cryptography-tutorial/what-is-phishing-attack