A Survey of Phishing Website Detection Systems

Rohith R [1] ,Srivanth S [2] ,Dhanush G [3] ,Tarun MK [4]

Department of Information Technology, St.Joseph's College of Engineering, Chennai, Tamil Nadu, India

----------***----------

Abstract –

Phishing URL is a widely used and common technique for cybersecurity attacks. Phishing is a cybercrime that tries to trick the targeted users into exposing their private and sensitive information to the attacker. The motive of the attacker is to gain access to personal information such as usernames, login credentials, passwords, financial account details, social networking data, and personal addresses. These private credentials are then often used for malicious activities such as identity theft, notoriety, financial gain, reputation damage, and many more illegal activities. This paper aims to provide a comprehensive and comparative study of various existing free service systems and research based systems used for phishing website detection. The systems in this survey range from different detection techniques and tools used by many researchers. The approach included in these researched papers ranges from Blacklist and Heuristic features to visual and content-based features. The studies presented here use advanced machine learning and deep learning algorithms to achieve better precision and higher accuracy while categorizing websites as phishing or benign. This article would provide a better understanding of the current trends and existing systems in the phishing detection domain.

***Key Words*:** Phishing, Phishing Websites, Detection, Machine Learning.

1. INTRODUCTION

The advancement of internet is resulting in attracting more and more users into this huge Internet Sea. There are a lot of perks of using internet, one can buy stuff online, way of learning and gaining knowledge has improved, etc. On the contrary, possible threats comes hand in hand. One of them is Phishing Attack. Phishing is an attack where a legitimate user is deceived to disclose sensitive information and assets with economic value. Loss of such sensitive information might cause potential economic or reputational harm an organization. Phishing basically uses social engineering techniques to trick users such as creating fake websites which clones with same attributes and design of the existing legitimate one. In a classic phishing attack a phisher send a link enclosed in a message to the user. The link redirects the user to the cloned malicious page which looks similar to the original webpage but is not and is intended to steal user's sensitive data. Such phishing attacks have proven to cause a lot of financial loss to various organizations. Thus, phishing attacks can be prevented by exterminating such harmful websites with the aid of a "Phishing

Detection" tool. Machine learning is one of the powerful techniques which can make the detection of phishing websites a lot simpler. A machine learning based tool will easily filter out Phishing and Non Phishing websites with the help of algorithms.

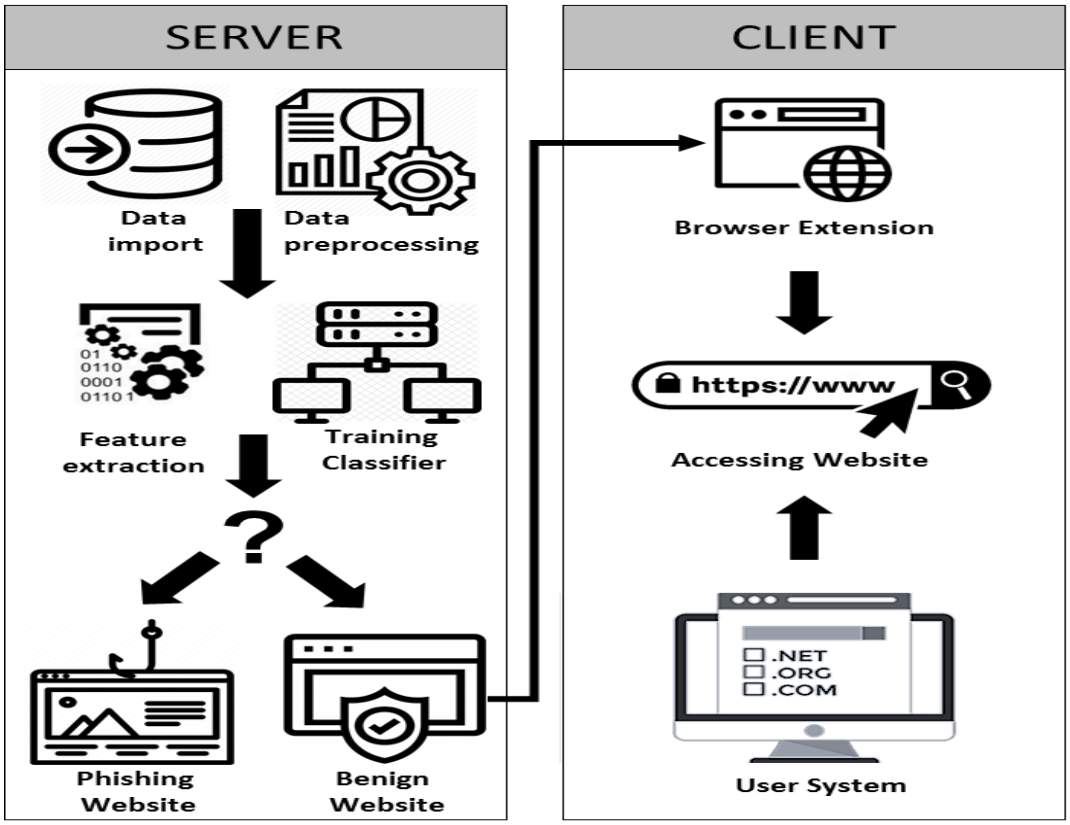

Fig-1: System Architecture

Website URLs are categorized into the following three classes:

a. Benign: These are Safe websites that provide normal services to people.
b. Malware: These websites which are created by attackers look like normal websites can make use of sensitive contents of people.
c. Spam: These websites flood the user's system with advertisements, fake surveys, etc.

In this survey, we review the phishing website detection systems which use advanced tools and techniques to provide promising results in this domain. We specifically focused on the work which presented the feature representation model with an advanced machine learning algorithm for development.

2.LITERATURE SURVEY

In this paper, the authors proposed a system with a collection or set of Hybrid features to classify websites based on machine learning algorithms. The main feature set is extracted using the cumulative distribution gradient technique, while the data perturbation ensemble technique is used to extract the secondary feature set.

The algorithm used for training the classifier is Random Forest in association with ensemble learner identifies the phishing websites with a precision of 94.6 percent. [1]

The authors made a relative study to detect phishing website URLs with machine learning and deep learning algorithms. Convolution Neural Network (CNN) and CNN Long Short-Term Memory (CNN-LSTM) with Logistic Regression formed the architecture of the classification model. The system was designed using tools like TensorFlow along with Keras for machine learning and deep learning model. The dataset was imported from multiple sources to provide better scalability. The phishing website URL dataset was obtained from OpenPhish and Phishtank, while the malicious or spam website URLs were imported from MalwareDomains. [2]

The proposed system detected phishing websites using a machine learning algorithm. The feature set included six features based on the website structure and was chosen after a comparative study by the authors. The classifier was trained using Support Vector Machine which worked effectively to classify websites whether legitimate or phishing. The model presented obtained an accuracy of 84 percent for the classification of websites. [3]

In this paper, the authors designed a browser extension to detect phishing websites. The system used multiple machine learning algorithms which included Random Forest, Support Vector Machine (SVM), and k-Nearest Neighbor (kNN) to train the classifier to achieve higher precision by doing a comparative study. The feature set included a content-based approach for extracting the JavaScript and HTML features of the websites. The dataset was imported from UCI-Machine Learning Repository and boasted a 22 feature classification technique to detect phishing websites. [4]

Authors made a comparative study of various machine learning algorithms such as Random Forests (RF), Support Vector Machines (SVM), Logistic Regression (LR), Bayesian Additive Regression Trees (BART), and Neural Networks to implement an efficient phishing website detection system. The dataset imported included a list of 2889 websites which were termed as phishing and a set of true blue messages. In total 43 features were extracted from the acquired dataset and were used extensively to train the classifier using the machine learning algorithms to obtain higher precision and accuracy. [5]

This paper proposes a phishing website detection method using reduces feature classification. The extracted features were analyzed using Support Vector Machine (SVM) and Logistic Regression algorithms. Out of the total 30 features identified, 19 features were selected and used for classification. The model was implemented using Big Data and the Dataset was obtained from the UCI Irvine machine learning repository. Between the

two algorithms used, Support Vector Machine (SVM) showed better performance and accuracy of 95.62%. [6]

Authors designed a system with a detection technique involving a fresh approach for phishing website detection named PhishLimiter. The proposed system used Deep Packet Inspection (DPI) along with Software-Defined Networking (SDN) through web communications and emails for identifying malicious activities. The real-time DPI and phishing signature classification based on SDN programmability provided PhishLimiter, the flexibility to address phishing attacks in real-time. This also helped in better network traffic management and evaluated attacks in real-world environments proving an effective solution to identify phishing attacks. [7]

The authors in this paper proposed a phishing detection system with a feature classification methodology. The phishing and legitimate website URL dataset was imported from Google and Phishtank. In total 133 features were extracted from the obtained dataset using the consistency subset based feature selection methods and the WEKA tool. Algorithms like Sequential Minimal Optimization (SMO) and Naïve Bayes (NB) were used to train the classifier to detect the phishing website URL. After analyzing the algorithms and doing a comparative study, the authors concluded that Sequential Minimal Optimization (SMO) achieved better performance than Naïve Bayes in terms of detecting the websites. [8]

In this paper, the authors used artificial intelligence techniques like neural networks for detecting phishing websites. The obtained data set from third-party service providers was divide into two parts, each for a specific purpose. The training module imported 80 percent of the dataset while the remaining 20 percent was used for the Testing phase. The Neural network model utilized the input of 17 neurons to compare with 17 characteristics in the imported dataset. The system determined whether the website is legitimate or phishing based on one hidden layer level of processing and output of two neurons. The proposed system showed an accuracy of 92.48 percent. [9]

The authors in this paper propose a model to classify websites as legitimate or phishing. The model was implemented in MATLAB and the Data Set was imported from the UCI Irvine machine learning repository. The system comprises of extraction of features from websites using Extreme Learning Machine (ELM), Naïve Bayes (NB), and Support Vector Machine (SVM). Among the algorithms used, the Extreme Learning Machine (ELM) obtained an accuracy of
95.34%. The model was implemented in MATLAB and the Data Set was imported from the UCI Irvine machine learning repository. [10]

In this paper, the authors designed a phishing website detection tool using the technique of linear classifier. They followed the content-based approach for the detection of websites and categorizing them into phishing or benign. The features extracted from the dataset included HTML, JavaScript features, and the website domain names for training the classifier using the linear classifier methodology. A total of 10 features were extracted and used in the training module. The authors were able to achieve 89 percent accuracy for the presented model. [11]

The authors proposed a system to detect phishing using heuristic-based methods and feature extraction. The c4.5 decision tree algorithm was used for analysis and computing the heuristic values to determine whether a website is legitimate or phishing. The Dataset

was imported from Phishtank and Google, proposed using the trained classifier, and used for detection purposes. The model achieved an accuracy of 89.40%. [12]

This paper proposed a system that determines phishing mails using two existing systems, Machine Learning Anti- Phishing System (MLAPT) and Phishzoo. The Phishzoo system uses the visually based approach for phishing detection while the Machine Learning Anti- Phishing System (MLAPT) helps in determining the mails present on the system into a phishing or benign category. The presented model proved effective to manage personal sensitive information on social networking websites. [13]

The authors in this paper presented a feature selection methodology to detect phishing website detection. The dataset was obtained from the UCI Irvine machine learning repository. Various algorithms were implemented by the authors for training purposes and after a comparative study, the authors finally concluded that different classification methods and strategies showed different results. Based on classification strategies and the data mining techniques the execution outcome results are incremented or decremented. [14]

In this paper, the authors proposed a multidimensional feature classification technique to detect phishing websites. The model used a deep learning methodology for faster detection of websites. The classifier was trained using the dataset obtained from third-party service providers. The URL features like visual and lexical features were extracted from the imported dataset and used for training the classifier using deep learning techniques. The presented model showed effective results by achieving an accuracy of 98 percent. [15]

The paper proposes an efficient way to detect phishing websites using a URL identification strategy utilizing the approach of the Random Forest algorithm. Phishtank was used to gather the required dataset. Out of the total 30 features listed, only 8 features were used for parsing to analyze the feature classification. The system model was partitioned into three stages which consisted of classification, parsing, and analysis. The model achieved 95% accuracy for the Random forest algorithm implemented using Rstudio. [16]

In this paper, the authors proposed a system for phishing website detection using machine learning algorithms. The system used a domain name based approach for determining the phishing website URL. The Random Forest algorithm was used to train the classifier by importing the dataset and extracting features for classification purposes. In total 10 URL features including the host-based and lexical features were extracted from the obtained dataset. The testing phase of this model achieved an accuracy of 96 percent for the Random Forest classifier using the labeled dataset. [17]

The paper proposed a flexible decision filter to extract and classify features from the inputted website URL. The system was implemented using a neural network model and other optimizers included AdaDelta, Adam, and Stochastic Gradient Descent (SGD). The Dataset was imported from Phishtank and Chainer was used to develop and implement the model. Among the three optimizers used, Adam obtained an accuracy of 94.18%. [18]

Authors in this paper proposed a system named BaitAlarm for determining phishing websites. The presented model used the visual content-based approach for feature representation and classification purposes. The visual features included HTML, JavaScript, and CSS features extracted from the list of websites obtained from the

imported dataset. The system relies on the visual content and layout based features of the website to detect whether the website is phishing or legitimate. [19]

In this paper, the authors proposed a novel approach using machine learning algorithms to detect phishing websites. The model utilized multiple machine learning algorithms for training and testing purposes. In total 30 features were extracted from the imported dataset obtained from various repositories and third-party service providers. The classifier was trained using Random Forest (RF), Decision Tree (DT), Generalized Linear Model (GLM), Gradient Boosting (GBM), and Generalized Additive Model (GAM). Out of all the machine learning algorithms, the Random forest classifier achieved an accuracy of 98.4 percent in the testing phase. [20]

The author presented a system for phishing website detection using different machine learning algorithms and a bag of words technique. The dataset was imported and the classifier was trained using the features extracted by a content-based approach. The training classifier included machine learning algorithms like Support Vector Machine (SVM) and Naive Bayes (NB) for training and testing purposes. Out of the two algorithms implemented, Support Vector Machine (SVM) showed higher accuracy achieving 95 percent. [21]

3.CONCLUSION

Phishing URL detection plays a pivotal role for many cybersecurity software and applications. In this paper, we researched and reviewed works based on the advanced machine learning techniques and approaches that promise a fresh approach in this domain. This article includes summary of the reviewed works after a systematic and comprehensive study on Phishing Website Detection systems. We believe that the presented survey would help researchers and developers with the insight of the progress achieved in the past years. Despite the tremendous progress in the field of cybersecurity, phishing website detection still pose a challenging problem with the ever evolving technology and techniques.

REFERENCES

[1] Kang Leng Chiew, Choon Lin Tan, KokSheik Wong, Kelvin SC Yong, and Wei King Tiong, "A new hybrid ensemble feature selection framework for machine learning-based phishing detection system," Information Sciences, vol. 484, pp. 153-166, 2019

[2] A. Vazhayil, R. Vinayakumar, and K. Soman, "Comparative Study of the Detection of Malicious URLs Using Shallow and Deep Networks," in 2018 9th International Conference on Computing, Communication and Networking Technologies, ICCCNT, 2018, pp. 1–6.

[3] Pan, Ying, and Xuhua Ding. —Anomaly based web phishing page detection.‖ In Computer Security Applications Conference, 2006. ACSAC'06. 22nd Annual, pp. 381-392. IEEE, 2006.

[4] A. Desai, J. Jatakia, R. Naik, and N. Raul, “Malicious web content detection using machine leaning,” RTEICT 2017 - 2nd IEEE Int. Conf. Recent Trends Electron. Inf. Commun. Technol. Proc., vol. 2018–Janua, pp. 1432– 1436, 2018.

[5] Abu-Nimeh, Saeed, Dario Nappa, Xinlei Wang, and Suku Nair. "An examination of machine learning systems for phishing recognition." In Proceedings of the counter phishing working gatherings second yearly eCrime specialists summit,ACM, pp. 60-69, 2007.

[6] W. Fadheel, M. Abusharkh, and I. Abdel-Qader, “On Feature Selection for the Prediction of Phishing Websites,” 2017 IEEE 15th Intl Conf Dependable, Auton. Secur. Comput. 15th Intl Conf Pervasive Intell. Comput. 3rd Intl Conf Big Data Intell. Comput. Cyber Sci. Technol. Congr., pp. 871–876, 2017.

[7] Tommy Chin, Kaiqi Xiong and Chengbin Hu, “PhishLimiter: A Phishing Detectionand Mitigation Approach Using Software-Defined Networking”, IEEE Access, 2018.

[8] M. Aydin and N. Baykal, “Feature extraction and classification phishing websites based on URL,” 2015 IEEE Conf. Commun. NetworkSecurity, CNS 2015, pp. 769–770, 2015.

[9] Rami M Mohammad, Fadi Thabtah, and Lee McCluskey, "Predicting phishing websites based on self-structuring neural network," Neural Computing and Applications, vol. 25, pp. 443-458, 2014.

[10] Y. Sönmez, T. Tuncer, H. Gökal, and E. Avci, “Phishing web sites features classification based on extreme learning machine,” 6th Int. Symp. Digit. Forensic Secur. ISDFS 2018 - Proceeding, vol. 2018, pp. 1–5, 2018.

[11] Xiang, Guang, and Jason I. Hong. —A hybrid phish detection approach by identity discovery and keywords retrieval. In Proceedings of the 18th international conference on World Wide Web, pp. 571-580. ACM, 2009.

[12] L. MacHado and J. Gadge, “Phishing Sites Detection
Based on C4.5 Decision Tree Algorithm,” in 2017 International Conference on Computing, Communication, Control and Automation, ICCUBEA 2017, 2018, pp.

[13] Meena, p., m. kavitha, s. jeyanthi, and cpnijithamahalakshmi. "Phishing prevention using datamining techniques." International Journal of Pure and Applied Mathematics 119, no. 10 117-123, 2018.

[14] M. Karabatak and T. Mustafa, “Performance comparison of classifiers on reduced phishing website dataset,” 6th Int. Symp. Digit. Forensic Secur. ISDFS 2018 - Proceeding, vol. 2018–Janua, pp. 1–5, 2018.

[15] Peng Yang, Guangzhen Zhao, Peng Zeng, “Phishing Website Detection based on Multidimensional Features driven by Deep Learning”, IEEE Access, 2018.

[16] S. Parekh, D. Parikh, S. Kotak, and P. S. Sankhe, “A New Method for Detection of Phishing Websites: URL Detection,” in 2018 Second International Conference on Inventive Communication and Computational
Technologies (ICICCT), 2018, vol. 0, no., pp. 949–952.

[17] Solomon Ogbomon Uwagbole, William J Buchanan, and Lu Fan, "Applied machine learning predictive analytics to SQL injection attack detection and prevention," in 2017 IFIP/IEEE Symposium on Integrated Network and Service Management (IM), 2017, pp. 1087- 1090.

[18] K. Shima et al., “Classification of URL bitstreams using bag of bytes,” in 2018 21st Conference on Innovation in Clouds, Internet and Networks and Workshops (ICIN), 2018, vol. 91, pp. 1–5.

[19] Mao, Jian, Pei Li, Kun Li, Tao Wei, and Zhenkai Liang. —BaitAlarm: detecting phishing sites using similarity in fundamental visual features.‖ In Intelligent Networking and Collaborative Systems (INCoS), 2013 5th
International Conference on, pp. 790-795. IEEE, 2013.

[20] J. Shad and S. Sharma, “A Novel Machine Learning Approach to Detect Phishing Websites Jaypee Institute of Information Technology,” pp. 425–430, 2018.

[21] Azad, B. Recognizing Phishing Attacks.