WEB PHISHING DETECTION

Neethu Mol (19104117), Nidhin R (19104118), Nihal Abdul Gafoor (19104119), Manikandan (19104092), Manojkumar R (19104094)

1.Introduction

There are a number of users who purchase products online and make payments through e-banking. There are e-banking websites that ask users to provide sensitive data such as username, password & credit card details, etc., often for malicious reasons. This type of e-banking website is known as a phishing website. Web service is one of the key communications software services for the Internet. Web phishing is one of many security threats to web services on the Internet.

Common threats of web phishing:

- Web phishing aims to steal private information, such as usernames, passwords, and credit card details, by way of impersonating a legitimate entity.
- It will lead to information disclosure and property damage.
- Large organizations may get trapped in different kinds of scams.

This Guided Project mainly focuses on applying a machine-learning algorithm to detect Phishing websites.

In order to detect and predict e-banking phishing websites, we proposed an intelligent, flexible and effective system that is based on using classification algorithms. We implemented classification algorithms and techniques to extract the phishing datasets criteria to classify their legitimacy. The e-banking phishing website can be detected based on some important characteristics like URL and domain identity, and security and encryption criteria in the final phishing detection rate. Once a user makes a transaction online when he makes payment through an e-banking website our system will use a data mining algorithm to detect whether the e-banking website is a phishing website or not.

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

3.**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

learningbased 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. 381392. 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. 6069, 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.