e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:05/May-2024 Impact Factor- 7.868 www.irjmets.com

PHISH-SHIELD: IDENTIFYING DECEPTIVE DOMAINS

Jaya Tripathi*1, Mahak Srivastava*2, Radhey Shyam*3, Ajay Srivastava*4

*1,2UG Student Of Department Of Information Technology Shri Ramswaroop Memorial Collage Of Engineering & Management, Lucknow, Uttar Pradesh, India.

*3,4Professor Of Department Of Information Technology Shri Ramswaroop Memorial Collage Of Engineering & Management, Lucknow, Uttar Pradesh, India.

DOI : https://www.doi.org/10.56726/IRJMETS55348

ABSTRACT

With advancements in technology, with each passing day, a large number of users are joining and surfing the web. Along with this, there has also been a rise in cybercrimes. Phishing website creation is one of the most prevalent one which traps innocent users by making them to access or click on the URLs (Uniform Resource Locators). Detecting phishing websites is vital in the fight against online misinformation and fraud. This research proposes a comprehensive approach, utilizing machine learning (ML) technique, to identify fraudulent websites. By scrutinizing a range of features, including content, design, and user interactions, our method employs ML algorithms to categorize websites as genuine or fake. We introduce a multi-stage detection framework that integrates feature extraction, classification, and validation steps to improve accuracy. Moreover, we explore the integration of external data sources such as social media signals to enhance detection performance. Through extensive experimentation across various datasets, our approach demonstrates both robustness and effectiveness in discerning between authentic and fake websites. This study contributes to the advancement of techniques aimed at combating online deception and protecting users from malicious online entities.

Keywords: Phishing, Phishing Website Detection, Machine Learning, Cybersecurity, Cybercrime.

I. INTRODUCTION

Anti-phishing strategies encompass both educating internet users and employing technical defenses. This paper primarily focuses on reviewing recent advancements in technical defense methodologies. A key aspect involves the identification of phishing websites, which serves as an effective means within the broader process of deceiving user information. A key aspect involves the identification of phishing websites, which serves as an effective means within the broader process of deceiving user information. Cybercriminals obtain this information through various illegal means and forge these users to carry out illegal activities on the Internet. In the early days of the invention of the Internet, network security issues have already appeared. With the development of the Internet, network attack techniques have also changed rapidly, which has brought many challenges to network security. According to the methods and forms of network attacks, cybersecurity issues are mainly divided into the denial-of-service attack (DoS), man-in-the middle (MitM), SQL injection, zero-day exploit, DNS tunnelling, phishing, and other categories.

Nowadays, the risk of cyber frauds is increasing rapidly along with increase in the level of danger. Most of the cyber criminals try to use end-to-end technology to exploit human vulnerabilities. These methods include social engineering, phishing, farming etc.

It is seen that that in most cyber-attacks the attackers try to deceive users by making them to click on phishing URLs (Uniform Resource Locators). This makes the detection of phishing websites a topic of great importance for anyone on the web. A large number of research studies being done are using various methods based on machine learning and deep learning techniques.

In our project we propose a phishing URL detection system running on machine learning algorithms by using URL behaviour and attributes in dataset to predict the output.

II. LITERATURE SURVEY

A brief survey of the existing phishing website detection methodologies is essential before describing the proposed system:

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:05/May-2024 Impact Factor- 7.868 www.irjmets.com

Empirical Study on Phishing Website Detection Using Machine Learning [2018] by Ripon Patgiri (B), Hemanth Katari(B), Ronit Kumar(B), and Dheeraj Sharma suggests that malicious web sites are the basis of most of the criminal activities over the internet. The dangers that arise due to the malicious sites are enormous and the end-users must be prohibited from visiting such sites. The users should prohibit themselves from clicking on such URLs.

Detection of URL based Phishing Attacks using Machine Learning [Nov -2019] by Ms. Sophiya Shikalgar Department of Computer Engineering Datta Meghe College of Engineering, Airoli, Navi Mumbai, INDIA Dr. S. D. Sawarkar Department of Computer Engineering Datta Meghe College of Engineering, Airoli, Navi Mumbai, INDIA Mrs. Swati Narwane Department of Computer Engineering Datta Meghe College of Engineering, Airoli, Navi Mumbai, INDIA addresses the widespread cybersecurity concern where threat actors bypass security defenses and use URLs to launch various forms of malicious attacks on unsuspecting individuals. In order to prevent such attacks, the paper proposes the use of machine learning algorithms to detect Phishing Websites. The proposed MuD (Phishing Website Detection) model is trained using an existing dataset which contains URLs, each with unique features, and is applied to three different machine learning classifiers—support vector machine, logistic regression and Naïve Bayes. After training and testing the algorithms, it is observed that Naïve Bayes classifier recorded the highest accuracy.

Phishing Website Detection Based on Associative Classification [2020] by Sandra Kumi 1 , ChaeHo Lim 2 and Sang-Gon Lee 1 suggests that Cybercriminals have invented sophisticated ways such as injecting malicious code into websites to disseminate malware in an attempt to infect target systems. Associative classification approaches to detect Phishing Websites mainly focus on phishing websites. Regarding this, we present an approach based on classification based on association (CBA) algorithm to detect malicious URLs comprising phishing, malware, and drive-by-download websites.

Using Deep Learning to Detect Malicious URLs [2019] by Yuchen Liang Shady Side Academy Pittsburgh, PA, United States and Xiaodan Yan Beijing University of Posts and Telecommunications Beijing, China suggests This paper presents different approaches to detect DGA generated domains based on the features of URLs. The result proves that the DBLSTM algorithm is superior to other conventional machine learning methods. The source code is posted on GitHub for other groups to use or to reproduce the same result (https://github.com/liangy2019/UsingDeep Learning-to-Detect-MaliciousURLs). The deep learning technique presented in the paper can be widely utilized in the realm of cybersecurity, especially for energy network security, to detect attacks initiated by different domain generation algorithms.

III. METHODOLOGY

Our proposed system extracts the features from the training data categorized into lexical features, network-based features and host-based features. The collected dataset was then divided into two subsets in the ratio of 80:20 for training purposes and testing purposes. The dataset consists of a huge number of URL's which consists of both malicious and non-phishing websites.

The completion of this project involves several key steps. Firstly, we need to collect a dataset containing both phishing and legitimate websites from open-source platforms. Following this, we'll develop code to extract the necessary features from the URL database. Once we have the dataset, we'll analyze and preprocess it using exploratory data analysis (EDA) techniques. After preprocessing, the dataset will be divided into training and testing sets. Then, we'll apply various machine learning and deep neural network algorithms such as Support Vector Machines (SVM), Random Forest, and Autoencoder to the dataset. Subsequently, we'll write code to display the evaluation results, considering accuracy metrics. Finally, we'll compare the results obtained from the trained models and specify which model performs better based on the evaluation metrics.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:05/May-2024 Impact Factor- 7.868 www.irjmets.com

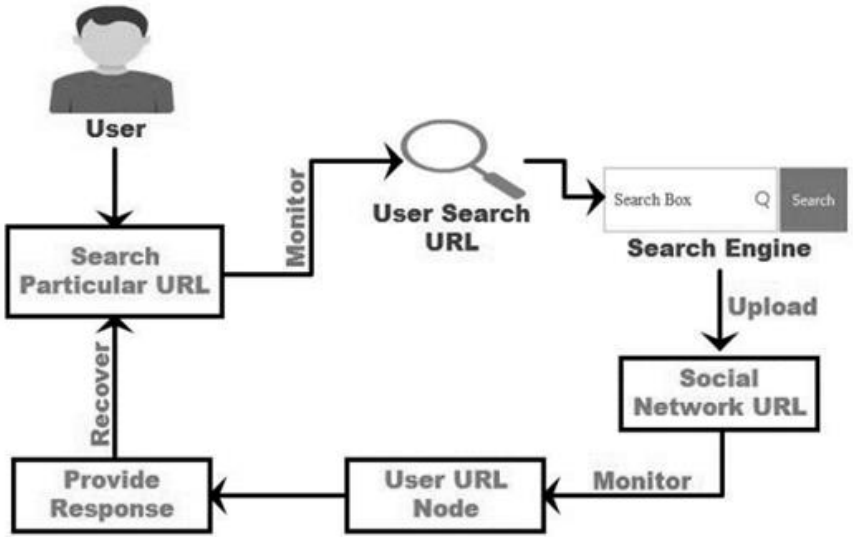

Figure 1: System architecture of proposed system.

Anti-phishing

There exist five sequential stages preceding an attacker's successful theft from a user's account or exploitation of acquired information for subsequent attacks. Consequently, interrupting any of these stages holds the potential to thwart a phishing attempt.

Web scraping

While it's challenging to entirely prevent perpetrators from crafting deceptive web pages, certain strategies can elevate their operational costs. Attackers often employ scripts to automate the extraction of content from legitimate web pages, subsequently repurposing this information for phishing endeavors. Hence, legitimate websites can hinder web scraping through various techniques, such as employing obfuscation methods like CSS sprites to present crucial data and substituting text with images.

Detecting Fake website

When individuals land on a phishing webpage masquerading as a genuine site, they often struggle to recall the authentic domain name, especially for lesser-known companies or startups. Consequently, users may fail to discern the phishing site solely based on the URL. To address this challenge, certain web browsers incorporate security features designed to identify phishing or malware-infested sites. For instance, Chrome presents warning messages when users attempt to access potentially hazardous web pages.

Heuristic strategies

A phishing detection method known as PhishWHO, is structured around three key phases. Initially, it acquires identity keywords through a weighted URL token system and employs an ensemble N-gram model based on the page's HTML. Subsequently, it utilizes these keywords to query mainstream search engines, aiming to identify the legitimate website and its authorized domain.

IV. MODELING AND ANALYSIS

Machine learning based method

Proposed are machine learning-driven countermeasures aimed at combating dynamic phishing attacks, boasting superior accuracy performance and reduced false positive rates compared to alternative methods [4]. This machine learning approach comprises six key components: data collection, feature extraction, model training, model testing, and prediction.

Data collection and feature extraction

Data serves as the cornerstone of each approach and significantly influences performance outcomes. Two primary methods are employed for data collection: loading existing datasets from publications and directly retrieving URLs from the internet. Table 1 highlights several prominent data sources. Within these three published datasets.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:05/May-2024 Impact Factor- 7.868 www.irjmets.com

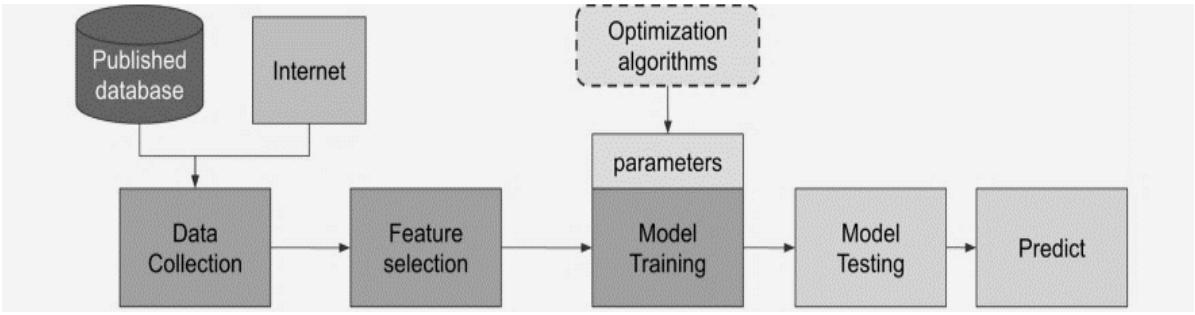

Figure 2: Machine learning based prediction.

Deep learning

A subset of machine learning, utilizes intricate structured architectures. Among the commonly employed deep learning algorithms are convolutional neural networks (CNNs), recurrent neural networks (RNNs), and long short-term memory (LSTM) networks. As natural language processing (NLP) and deep learning continue to advance, numerous deep learning-based solutions have emerged for phishing detection.

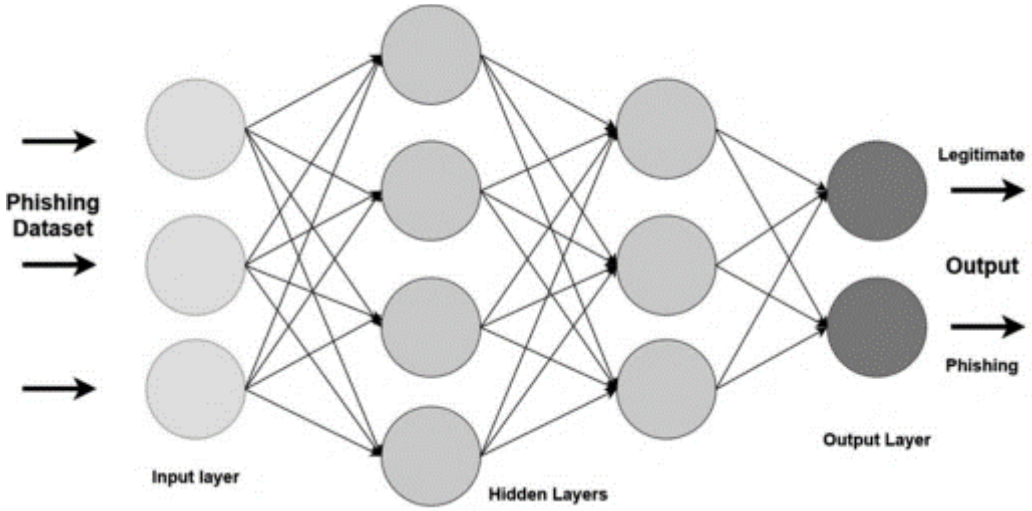

Figure 3: Internal working of deep learning model.

V. CONCLUSION

Detecting phishing attacks is crucial due to the potential leakage of highly sensitive user information through phishing websites. It's imperative to address this issue to safeguard users' personal data. Utilizing machine learning algorithms with a classifier offers a promising solution to this problem. The proposed technique not only identifies both new and previously known phishing sites but also ensures enhanced security. The plan involves deploying the model online by integrating it as a web browser plug-in capable of providing real-time warnings to users about potential phishing websites. This plug-in will analyze URLs clicked or typed by users, assessing their features to determine potential malicious intent.

If a URL is flagged as malicious or suspected to be so, a pop-up notification will alert the user about the potential threat, and access to the URL will be temporarily blocked unless the user chooses to proceed. Future endeavours will include evaluating the proposed model against more recent and diverse datasets and exploring the integration of additional classifiers such as decision trees and random forests.

VI. REFERENCES

[1] Vanhoenshoven, F., N´apoles, G., Falcon, R., Vanhoof, K., K¨oppen, M.: Detecting Phishing Websites using machine learning techniques. In: 2016 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 1–8. IEEE (2016)

[2] F. Vanhoenshoven, G. Nápoles, R. Falcon, K. Vanhoof, M. Köppen, Detecting Phishing Websites using machine learning techniques, in 2016 IEEE SymposiumSeries on Computational Intelligence (SSCI) (IEEE, 2016), pp. 1–8

[3] A. Singh, N. Goyal, A comparison of machine learning attributes for detecting malicious websites, in 2019 11th International Conference on Communication Systems & Networks (COMSNETS) (IEEE, 2019), pp. 352–358

[4] A.S. Manjeri, R. Kaushik, M. Ajay, P.C. Nair, A machine learning approach for detecting malicious websites using URL features, in 2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA) (IEEE, 2019), pp. 555–561

[5] Internet Security Threat Report (ISTR) 2019–Symantec. https://www.symantec.com/content/dam/symantec/docs/reports/istr-24-2019-en.pdf

[6] Kim, K., Hwang, S., & Kim, Y. (2020). Machine learning-based phishing website detection using URL and HTML features. Computers & Security, 92, 101740

[7] Zhang, Z., Zhang, X., Wang, H., Wu, S., & Wang, S. (2019). A survey on phishing website detection techniques. Computers & Security, 84, 63-7[6] Kim, K., Hwang, S., & Kim, Y. (2020). Machine learning-based phishing website detection using URL and HTML features. 6.

[8] Almomani, A., Gupta, B. B., & Gupta, N. (2018). A novel model for detection of phishing websites using fuzzy logic. Computers & Security, 77, 1-11.

[9] Kim, K., Hwang, S., & Kim, Y. (2020). Machine learning-based phishing website detection using URL and HTML features. Computers & Security, 92, 101740.

[10] Hong, J., Yang, S., Yang, D., & Hong, J. (2007). Detecting phishing web sites by analysing HTML content. In Proceedings of the 16th international conference on World Wide Web (pp. 581-590).

[11] Ramzan, Z., & Rubin, A. D. (2007). Phish and HIPs: Human interactive proofs to detect phishing attacks. In Proceedings of the 16th international conference on World Wide Web (pp. 411-420).

[12] Kumar, S., & Reddy, P. K. (2012). A survey on phishing detection techniques. International Journal of Computer Applications, 55(9), 24-29.

[13] Ma, J., Saul, L. K., Savage, S., & Voelker, G. M. (2009). Beyond blacklists: Learning to detect malicious web sites from suspicious URLs. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1245-1254).

[14] Ramachandran, A., & Feamster, N. (2006). Understanding the network-level behaviour of spammers. ACM SIGCOMM Computer Communication Review, 36(4), 291-302.

[15] Li, Y., Li, J., Zheng, X., & Gao, D. (2020). Phishing detection based on machine learning algorithms: a survey. Journal of Cyber Security Technology, 4(2), 127-144.

[16] Vartika Srivastava, Radhey Shyam base on machine learning (2021) . View article (google.com)