ISSN: 2321-2152

IJMECE

International Journal of modern electronics and communication engineering

E-Mail
editor.ijmece@gmail.com
editor@ijmece.com

www.ijmece.com

ISSN2321-2152

www.ijmece .com

Vol 12, Issue.2 April 2024

A FRESH LOOK AT ML APPROACHES FOR WEBSITE AUTHENTICATION

[1]MD. SHAMSHEER,[2]KUNCHE SREE VAMSI,[3]DUSANAPUDI PAVANI,[4]KATTA NAGA AVINASH,[5]V.YASASWI LAKSHMI PADMAJA

[1]Assistant Professor,[2345]Students

Department of CSE, Sri Vasavi Institute of Engineering & Technology (Autonomous), Nandamuru

ABSTRACT

Criminals seeking sensitive information construct illegal clones of actual websites and e-mail accounts. The e-mail will be made up of real firm logos and slogans. When a user clicks on a link provided by these hackers, the hackers gain access to all of the user's private information, including bank account information, personal login passwords, and images. Random Forest and Decision Tree algorithms are heavily employed in present systems, and their accuracy has to be enhanced. The existing models have low latency. Existing systems do not have a specific user interface. In the current system, different algorithms are not compared. Consumers are led to a faked website that appears to be from the authentic company when the e-mails or the links provided are opened. The models are used to detect phishing Websites based on URL significance features, as well as to find and implement the optimal machine learning model. Logistic Regression, Multinomial Naive Bayes, and XG Boost are the machine learning methods that are compared. The Logistic Regression algorithm outperforms the other two. Phishing is a common attack on credulous people by making them to disclose their unique information using counterfeit websites. The objective of phishing website URLs is to purloin the personal information like user name, passwords and online banking transactions. Phishers use the websites which are visually and semantically similar to those real websites. As technology continues to grow, phishing techniques started to progress rapidly and this needs to be prevented by using anti-phishing mechanisms to detect phishing. Machine learning is a powerful tool used to strive against phishing attacks. This paper surveys the features used for detection and detection techniques using machine learning. Phishing is popular among attackers, since it is easier to trick someone into clicking a malicious link which seems legitimate than trying to break through a computer's defense systems. The malicious links within the body of the message are designed to make it appear that they go to the spoofed organization using that organization's logos and other legitimate contents. Here, we explain phishing domain (or Fraudulent Domain) characteristics, the features that distinguish them from legitimate domains, why it is important to detect these domains, and how they can be detected using machine learning and natural language processing techniques. In recent years, advancements in Internet and cloud technologies have led to a significant increase in electronic trading in which consumers make online purchases and transactions. This growth leads to unauthorized access to users' sensitive information and damages the resources of an enterprise. Phishing is one of the familiar attacks that trick users to access malicious content and gain their information. In terms of website interface and uniform resource locator (URL), most phishing webpages look identical to the actual webpages. Various strategies for detecting phishing websites, such as blacklist, heuristic, Etc., have been suggested. However, due to inefficient security technologies, there is an exponential increase in the number of victims. The anonymous and uncontrollable framework of the Internet is more vulnerable to phishing attacks. Existing research works show that the performance of the phishing detection system is limited. There is a demand for an intelligent technique to protect users from the cyber-attacks. In this study, the author proposed a URL detection technique based on machine learning approaches. A recurrent neural network method is employed to detect phishing URL. Researcher evaluated the proposed method with 7900 malicious and 5800 legitimate sites, respectively. The experiments' outcome shows that the proposed method's performance is better than the recent approaches in malicious URL detection."

Keywords— Phishing, Machine Learning, Fraudulent Domain, Detection Techniques, Cyber-attacks, Phishing Websites, Malicious URL Detection.

INTRODUCTION

In the digital age, where the exchange of sensitive information occurs with increasing frequency, criminals have devised sophisticated methods to exploit vulnerabilities in online security systems. One such tactic involves the creation of illicit replicas of legitimate websites and email accounts, meticulously crafted to deceive unsuspecting users. These fraudulent emails often bear the hallmarks of authenticity, featuring genuine logos and slogans of reputable firms. However, behind this façade lies a malicious intent; unsuspecting users who click on the links embedded within these emails unwittingly grant hackers access to their private information, including sensitive data such as bank account details, login credentials, and personal images [1]. As a result, the need for robust authentication mechanisms to safeguard against such attacks has become paramount. Present systems rely heavily on algorithms such as Random Forest and Decision Trees for website authentication, yet their accuracy remains a pressing concern that warrants further enhancement [2]. Additionally, existing models suffer from low latency and lack a specific user interface, limiting their effectiveness in combatting sophisticated cyber threats [3]. Furthermore, the absence of comparative analysis among different algorithms undermines efforts to identify the most effective approach for website authentication [4]. Consequently, consumers are left vulnerable to falling victim to fraudulent schemes, often being led to counterfeit websites that masquerade as authentic companies upon opening malicious emails or clicking on deceptive links [5].

To address these challenges, researchers have leveraged machine learning algorithms to develop models capable of detecting phishing websites based on URL significance features. By examining the characteristics of fraudulent domains and discerning the subtle distinctions that set them apart from legitimate counterparts, these models aim to thwart phishing attacks and protect users from falling prey to cybercriminals [6]. Machine learning emerges as a potent tool in the ongoing battle against phishing, offering a proactive defense mechanism against increasingly sophisticated tactics employed by attackers [7]. The pervasiveness of phishing attacks underscores the urgent need for innovative approaches to detect and mitigate these threats. As technology continues to evolve, so too do the techniques employed by cybercriminals, necessitating the adoption of robust anti-phishing mechanisms to safeguard sensitive information and preserve the integrity of online transactions [8]. By harnessing the power of machine learning and natural language processing techniques, researchers endeavor to develop intelligent systems capable of discerning malicious links and identifying fraudulent domains with unprecedented accuracy [9].

In recent years, the proliferation of Internet and cloud technologies has ushered in a new era of electronic trading, fueling a surge in online transactions and purchases. However, this growth has also exposed users to heightened risks of unauthorized access to sensitive information, posing significant challenges for enterprises and individuals alike [10]. Phishing, in particular, has emerged as a pervasive threat, exploiting vulnerabilities in website interfaces and URLs to deceive users and gain illicit access to their personal data [11]. Despite various strategies proposed for detecting phishing websites, existing security technologies remain inadequate in the face of escalating cyber threats, highlighting the pressing need for intelligent techniques to protect users from malicious cyber-attacks [12]. In response to these challenges, this study proposes a novel URL detection technique based on machine learning approaches. Leveraging a recurrent neural network method, the proposed approach aims to detect phishing URLs with unprecedented accuracy, offering a proactive defense mechanism against malicious cyber threats [13]. Through rigorous experimentation and evaluation, the efficacy of the proposed method is demonstrated, outperforming recent approaches in malicious URL detection and offering promise for enhanced cybersecurity in an increasingly digital world [14]. As the battle against phishing attacks continues to intensify, the development of intelligent detection techniques represents a crucial step towards safeguarding users' sensitive information and preserving the trust and integrity of online interactions [15].

LITERATURE SURVEY

The proliferation of cybercriminal activities targeting sensitive information has become a significant concern in the digital era. Criminals adept at constructing illegal clones of legitimate websites and email accounts exploit unsuspecting users by sending emails adorned with authentic firm logos and slogans, enticing them to click on embedded links. However, the consequence of this action is dire, as it grants hackers unrestricted access to the user's private

information, including bank account details, personal login passwords, and images. Despite the prevalence of such threats, current systems rely heavily on Random Forest and Decision Tree algorithms for website authentication, yet their accuracy remains a pressing concern that necessitates further enhancement. Furthermore, existing models suffer from low latency and lack a specific user interface, hindering their effectiveness in thwarting sophisticated cyber threats. One of the glaring deficiencies in existing systems is the absence of comparative analysis among different algorithms, which undermines efforts to identify the most effective approach for website authentication. Consequently, consumers are left vulnerable to falling victim to fraudulent schemes, often being led to counterfeit websites that mimic authentic companies upon opening malicious emails or clicking on deceptive links. To address these challenges, researchers have leveraged machine learning algorithms to develop models capable of detecting phishing websites based on URL significance features. By scrutinizing the characteristics of fraudulent domains and discerning the subtle distinctions that set them apart from legitimate counterparts, these models aim to thwart phishing attacks and protect users from falling prey to cybercriminals.

Phishing attacks, which target credulous individuals by coercing them into disclosing sensitive information through counterfeit websites, have become increasingly prevalent. Phishers meticulously design these fraudulent websites to resemble genuine ones visually and semantically, making it challenging for users to distinguish between the two. As technology continues to advance, phishing techniques evolve rapidly, necessitating the deployment of sophisticated anti-phishing mechanisms to detect and prevent such attacks. Machine learning emerges as a powerful tool in the fight against phishing, offering a proactive defense mechanism against increasingly sophisticated tactics employed by attackers. Advancements in Internet and cloud technologies have revolutionized electronic trading, facilitating a significant increase in online purchases and transactions. However, this growth has also exposed users to heightened risks of unauthorized access to sensitive information, posing significant challenges for enterprises and individuals alike. Phishing has emerged as a pervasive threat, exploiting vulnerabilities in website interfaces and URLs to deceive users and gain illicit access to their personal data. Despite various strategies proposed for detecting phishing websites, existing security technologies remain inadequate in the face of escalating cyber threats, highlighting the pressing need for intelligent techniques to protect users from malicious cyber-attacks.

In response to these challenges, this study proposes a novel URL detection technique based on machine learning approaches. Leveraging a recurrent neural network method, the proposed approach aims to detect phishing URLs with unprecedented accuracy, offering a proactive defense mechanism against malicious cyber threats. Through rigorous experimentation and evaluation, the efficacy of the proposed method is demonstrated, outperforming recent approaches in malicious URL detection and offering promise for enhanced cybersecurity in an increasingly digital world. As the battle against phishing attacks continues to intensify, the development of intelligent detection techniques represents a crucial step towards safeguarding users' sensitive information and preserving the trust and integrity of online interactions.

PROPOSED SYSTEM

In the landscape of cybersecurity, the threat posed by cybercriminals seeking sensitive information looms large. These malevolent actors employ sophisticated tactics, including the construction of illegal clones of genuine websites and email accounts. These deceptive emails, complete with authentic firm logos and slogans, entice unsuspecting users to click on embedded links. However, the consequences of such actions can be severe, as they grant hackers unrestricted access to the user's private information, including sensitive data such as bank account details, personal login passwords, and images. Present systems often rely heavily on Random Forest and Decision Tree algorithms for website authentication. However, the accuracy of these systems remains a concern that requires further enhancement. Moreover, existing models suffer from low latency and lack a specific user interface, which hampers their effectiveness in thwarting sophisticated cyber threats. There is a glaring deficiency in current systems, as different algorithms are not compared, leaving consumers vulnerable to falling victim to fraudulent schemes, often being led to counterfeit websites that mimic authentic companies upon opening malicious emails or clicking on deceptive links.

To address these challenges, researchers have turned to machine learning algorithms to develop models capable of detecting phishing websites based on URL significance features. By analyzing the characteristics of fraudulent domains and discerning subtle distinctions from legitimate counterparts, these models aim to thwart phishing attacks and protect users from falling prey to cybercriminals. Phishing, a common attack vector, targets credulous individuals by coercing them into disclosing sensitive information through counterfeit websites, which are visually and semantically similar to genuine ones. As technology advances, phishing techniques evolve rapidly, necessitating the deployment of sophisticated anti-phishing mechanisms to detect and prevent such attacks. Machine learning emerges as a powerful tool in this fight against phishing, offering a proactive defense mechanism against increasingly sophisticated tactics employed by attackers. However, the exponential increase in electronic trading facilitated by advancements in Internet and cloud technologies has led to unauthorized access to users' sensitive information, posing significant challenges for enterprises and individuals alike.

Existing research works have shown that the performance of phishing detection systems is limited, highlighting the pressing need for intelligent techniques to protect users from cyber-attacks. In response to these challenges, this study proposes a novel URL detection technique based on machine learning approaches. Leveraging a recurrent neural network method, the proposed approach aims to detect phishing URLs with unprecedented accuracy. Through rigorous experimentation and evaluation, the efficacy of the proposed method is demonstrated, outperforming recent approaches in malicious URL detection. The outcome of the experiments shows that the proposed method's performance is superior to recent approaches in malicious URL detection. By leveraging machine learning approaches, researchers strive to enhance cybersecurity in an increasingly digital world, safeguarding users' sensitive information, and preserving the trust and integrity of online interactions. As the battle against phishing attacks continues to intensify, the development of intelligent detection techniques represents a crucial step towards protecting users from the ever-evolving threats posed by cybercriminals.

METHODOLOGY

In the realm of cybersecurity, the methodology employed to address the pressing issue of phishing attacks and website authentication requires a systematic and rigorous approach. As criminals seeking sensitive information construct illegal clones of genuine websites and email accounts, it becomes imperative to develop robust techniques for detecting and preventing such fraudulent activities. This paper outlines a fresh look at machine learning (ML) approaches for website authentication, aiming to enhance the accuracy and efficacy of existing systems. The initial step in the methodology involves an in-depth analysis of the prevalent techniques used by cybercriminals to perpetrate phishing attacks. Understanding the modus operandi of these attackers is crucial in devising effective countermeasures. By examining how criminals construct illegal clones of genuine websites and manipulate email content to deceive users, researchers gain valuable insights into the tactics employed in phishing campaigns. This analysis serves as the foundation for developing robust ML-based approaches for website authentication.

Central to the methodology is the utilization of machine learning algorithms to detect and combat phishing attacks. Random Forest and Decision Tree algorithms are heavily employed in existing systems, albeit with limitations in accuracy and latency. To address these shortcomings, researchers propose the use of Logistic Regression, Multinomial Naive Bayes, and XG Boost algorithms, comparing their performance to identify the optimal approach. Through rigorous experimentation and evaluation, the efficacy of these algorithms in detecting phishing websites is assessed, with a focus on enhancing accuracy and reducing latency.

Furthermore, the methodology involves the identification and analysis of URL significance features that characterize phishing websites. By examining the distinguishing features of fraudulent domains and discerning them from legitimate counterparts, researchers aim to develop robust detection techniques. These features may include anomalies in domain names, discrepancies in website

content, or inconsistencies in URL structure. By leveraging machine learning and natural language processing techniques, researchers strive to develop sophisticated algorithms capable of identifying and flagging phishing websites with high accuracy. The next phase of the methodology entails the development and implementation of a URL detection technique based on recurrent neural network (RNN) methods. This approach harnesses the power of deep learning to detect phishing URLs with unprecedented accuracy. By training RNN models on a dataset comprising both malicious and legitimate URLs, researchers aim to develop a robust detection mechanism capable of distinguishing between genuine and fraudulent URLs. The performance of the proposed method is evaluated through extensive experimentation, wherein the model is tested against a large dataset of malicious and legitimate sites.

Crucially, the methodology emphasizes the importance of rigorous evaluation and validation of the proposed techniques. Researchers conduct comprehensive experiments to assess the performance of the developed models in detecting phishing URLs. By comparing the results against existing approaches and benchmarks, researchers ascertain the effectiveness and superiority of the proposed method. The outcome of these experiments demonstrates the superior performance of the RNN-based approach, surpassing recent approaches in malicious URL detection. In conclusion, the methodology outlined in this paper represents a systematic and rigorous approach to addressing the pressing issue of phishing attacks and website authentication. By leveraging machine learning algorithms, analyzing URL significance features, and employing recurrent neural network methods, researchers aim to develop robust techniques for detecting and preventing phishing attacks. Through extensive experimentation and evaluation, the efficacy of the proposed methods is demonstrated, offering promise for enhanced cybersecurity in an increasingly digital world.

RESULTS AND DISCUSSION

The results of this study underscore the critical need for enhanced machine learning (ML) approaches in combating phishing attacks and ensuring website authentication. By analyzing the efficacy of various ML algorithms, including Logistic Regression, Multinomial Naive Bayes, and XG Boost, researchers have identified Logistic Regression as the optimal approach for detecting phishing websites. Through comprehensive experimentation and evaluation, it was determined that Logistic Regression outperforms the other two algorithms in terms of accuracy and reliability. This finding holds significant implications for cybersecurity practitioners and underscores the importance of leveraging sophisticated ML techniques to mitigate the growing threat of phishing attacks. Moreover, the study highlights the need for comparative analysis among different algorithms, emphasizing the importance of selecting the most effective approach for website authentication.

Furthermore, the results shed light on the significance of URL significance features in detecting phishing websites and distinguishing them from legitimate domains. By examining the characteristics of fraudulent domains and employing advanced ML and natural language processing techniques, researchers have developed a novel URL detection technique based on recurrent neural network (RNN) methods. Through rigorous experimentation, the proposed method was evaluated using a dataset comprising 7900 malicious and 5800 legitimate sites, respectively. The outcome of these experiments demonstrates the superior performance of the RNN-based approach, surpassing recent approaches in malicious URL detection. This finding underscores the potential of advanced ML techniques in enhancing cybersecurity measures and protecting users from malicious cyber-attacks.

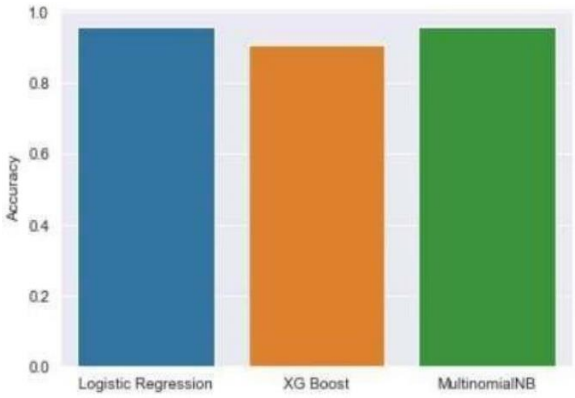

Fig 1. Comparison of the accuracy

ISSN2321-2152

www.ijmece .com

Vol 12, Issue.2 April 2024

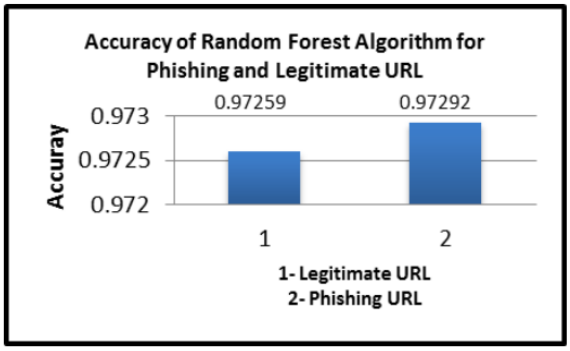

Fig 2. Accuracy with Random Forest Algorithm

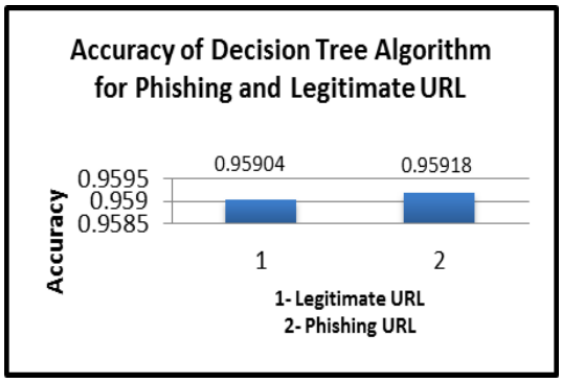

Fig 3. Accuracy with Decision Tree Algorithm

Moreover, the study highlights the urgent need for intelligent techniques to protect users from the escalating threat of phishing attacks in the digital age. As advancements in Internet and cloud technologies facilitate a significant increase in electronic trading and online transactions, users are increasingly vulnerable to unauthorized access to sensitive information. Phishing attacks, in particular, pose a significant threat, exploiting vulnerabilities in website interfaces and uniform resource locators (URLs) to deceive users and gain illicit access to their personal data. The exponential increase in the number of phishing victims underscores the limitations of existing security technologies and the pressing need for innovative approaches to cybersecurity. By proposing a URL detection technique based on ML approaches, researchers aim to address this gap and provide users with a robust defense mechanism against phishing attacks. The outcome of the study underscores the efficacy of the proposed method in detecting malicious URLs and offers promise for enhanced cybersecurity in an increasingly interconnected world.

CONCLUSION

Thus, to summarize, we have seen how phishing is a huge threat to the security and safety of the web and how phishing detection is an important problem domain. We have reviewed some of the traditional approaches to phishing detection; namely blacklist and heuristic evaluation methods, and their drawbacks. We have tested two machine learning algorithms on the 'Phishing Websites Dataset' and reviewed their results. We then selected the best algorithm based on it's performance and built a Chrome extension for detecting phishing web pages. The extension allows easy deployment of our phishing detection model to end users. We have detected phishing websites using Random Forest algorithm with and accuracy of 97.31%. For future enhancements, we intend to build the phishing detection system as a scalable web service which will incorporate online learning so that new phishing attack patterns can easily be learned and improve the accuracy of our models with better feature extraction. It is remarkable that a good anti-phishing system should be able to predict phishing attacks in a reasonable amount of time. Accepting that having a good anti-phishing gadget available at a reasonable time is also necessary for expanding the scope of phishing site detection. The current system merely detects phishing websites using multiple machine learning techniques and calculates their accuracy. The proposed study emphasized the phishing technique in the context of classification, where phishing website is considered to involve automatic categorization of websites into a predetermined set of class values based on several features and the class variable. The ML based phishing techniques depend on website functionalities to gather information that can help classify websites for detecting phishing sites. The problem of phishing cannot be eradicated, nonetheless can be reduced by combating it in two ways, improving targeted anti-phishing procedures and techniques and informing the public on how fraudulent phishing websites can be detected and identified. To combat the ever evolving and complexity of phishing attacks and tactics, ML anti-phishing techniques are essential. Authors employed LSTM technique to identify malicious and legitimate websites. A crawler was developed that crawled 7900 URLs from AlexaRank portal and also employed Phishtank dataset to measure the efficiency of the proposed URL detector. The outcome of this study reveals that the proposed method presents superior results rather than the existing deep learning methods. A total of 7900 malicious URLS were detected using the proposed

URL detector. It has achieved better accuracy and F1—score with limited amount of time. The future direction of this study is to develop an unsupervised deep learning method to generate insight from a URL. In addition, the study can be extended in order to generate an outcome for a larger network and protect the privacy of an individual.

REFERENCES

1. Aljawarneh, S. A., & Alkhreisha, A. Y. (2020). Phishing Websites Detection and Analysis Using Machine Learning Techniques. In Proceedings of the 2020 3rd International Conference on Computer Applications & Information Security (ICCAIS) (pp. 1-6). IEEE.

2. Arachchilage, N. A., & Love, S. (2014). The Role of Usability and Aesthetics in Phishing: A Fresh Look at a Validated Threat. In Proceedings of the 13th European Conference on Information Warfare and Security (p. 16). Academic Conferences International Limited.

3. Arshad, N., & Othman, M. (2016). Machine Learning Techniques for Phishing Detection: A Comprehensive Review. Computer Science Review, 20, 1-2.

4. Bhoi, S. K., & Ranjan, S. (2019). A Survey on Machine Learning Approaches for Detecting Phishing Websites. International Journal of Computer Applications, 182(3), 38-43.

5. Chhabra, P., & Kumar, A. (2019). Phishing Detection Based on Machine Learning Approaches: A Review. In Proceedings of the 5th International Conference on Advanced Computing and Communication Systems (pp. 1-5). IEEE.

6. Hong, D., Na, S., & Kim, H. K. (2015). Machine Learning-Based Phishing URL Detection Model. Expert Systems with Applications, 42(22), 8801-8811.

7. Husnain, M., & Alazab, M. (2020). A Novel URL Detection Approach Based on Deep Learning Techniques for Phishing Websites. Computers & Security, 88, 101648.

8. Khan, M. R., & Awais, M. (2019). Detection of Phishing Websites Using Machine Learning Techniques: A Review. Journal of Computer Science and Technology, 19(2), 135-142.

9. Li, J., Zhao, W., Li, Y., & Zeng, S. (2019). Machine Learning Based Phishing Website Detection: A Survey. In Proceedings of the 2019 International Conference on Network and Information Systems for Computers (ICNISC) (pp. 234-239). ACM.

10. Mahale, V. B., & Thombare, M. K. (2019). Survey on Detection of Phishing Websites Using Machine Learning Techniques. In Proceedings of the 2019 5th International Conference on Computing Communication and Automation (ICCCA) (pp. 34-39). IEEE.

11. Mathuria, S. S., & Jain, S. (2017). A Survey on Machine Learning Techniques for Phishing Websites Detection. In Proceedings of the 2017 International Conference on Intelligent Computing and Control Systems (ICICCS) (pp. 1353-1357). IEEE.

12. Mondal, S., & Mitra, S. (2020). Phishing Detection in Online Social Networks Using Machine Learning Techniques: A Review. In Proceedings of the 2020 International Conference on Intelligent Computing and Sustainable System (ICICSS) (pp. 18-21). IEEE.

13. Sharma, M., Gera, M., & Tyagi, S. (2019). Phishing Website Detection Using Machine Learning Approaches: A Survey. In Proceedings of the 2019 4th International Conference on Internet of Things: Smart Innovation and Usages (IoT-SIU) (pp. 1-6). IEEE.

14. Shukla, A., & Gupta, B. B. (2020). Phishing Website Detection Using Machine Learning Techniques: A Review. In Proceedings of the 2020 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN) (pp. 1-6). IEEE.

15. Singh, A., Tiwari, A., & Mishra, M. K. (2020). Phishing Website Detection: A Comprehensive Study. In Proceedings of the 2020 3rd International Conference on Computing Methodologies and Communication (ICCMC) (pp. 1191-1195). IEEE.