Phishing Website Detection

Joshma K J, Vineetha Sankar P

***Abstract:** Phishing websites have emerged as a serious security risk. Phishing is the starting point for many cyberattacks that compromise the confidentiality, integrity, and availability of customer and business data. Decades of effort have gone into developing novel methods for automatically identifying phishing websites. Modern systems aren't very adept at spotting new phishing threats and require a lot of manual feature engineering, even though they can produce better outcomes. Thus, an open problem in this discipline is to identify tactics that can swiftly handle zero-day phishing attempts and automatically recognize phishing websites. The web page that the URL hosts has a plethora of information that can be utilized to assess the maliciousness of the web server. One useful technique for spotting phishing emails is machine learning. Additionally, it does away with the drawbacks of the earlier approach. After a careful analysis of the literature, we proposed a novel approach that combines a machine learning algorithm with feature extraction to identify phishing websites. Using the gathered dataset, this study aims to train deep neural networks and machine learning models to detect phishing websites.*

***Keywords:** Deep Neural Networks, Machine Learning, Phishing Websites, Cybersecurity, Feature Extraction, and Zero-day Attacks.*

I. INTRODUCTION

The widespread expansion of the internet has completely changed how we communicate, work, and live. But there is a dark side to the digital age, and that is the threat posed by cyber threats. Phishing assaults are one particularly sneaky hazard that has emerged. Phishing is the use of fraudulent techniques to fool someone into disclosing private information, such as bank account information or personal credentials. Traditional detection techniques are unable to keep up with the increasing sophistication of these attacks, thus novel strategies to strengthen our digital defences are required. This project takes on the express goal of identifying phishing websites by delving into the field of machine learning in response to this urgent situation. The vast number of worldwide phishing events that have been documented, impacting individuals, businesses, and organizations equally, highlights the seriousness of the matter. As cybercriminals utilize more advanced methods, it is critical to have flexible and effective solutions. The primary dataset used in this investigation comes from Kaggle and includes a wide range of website parameters. Cybersecurity and machine learning come together to build a powerful coalition. This project, which aims to create strong models for phishing detection and provide insightful information to the continuing efforts to strengthen our digital infrastructure, is situated at the intersection of these domains. Come along on this investigation into the core of cybersecurity, where necessity and ingenuity collide in the unwavering quest for a secure digital future. Our study utilizes a comprehensive dataset from Kaggle, encompassing a wide array of website parameters. By integrating cybersecurity principles with machine learning, we strive to build powerful models for phishing detection. This project seeks to enhance our understanding and capability to combat phishing, contributing to the ongoing efforts to secure our digital infrastructure. Join us in exploring the forefront of cybersecurity, where innovation meets necessity in the relentless pursuit of a safer digital future.

II. LITERATURE REVIEW

[1] M. Bahaghigat, M. Ghasemi, and F. Ozen, 2023. "A high accuracy machine learning phishing website detection." The study published in the Journal of Applications and Information Security explores advanced machine learning techniques to enhance the accuracy of phishing website detection. The authors focus on the development and evaluation of a robust model capable of distinguishing between legitimate and fraudulent websites with high precision [11].

[2] H. Zhang, D. Gao, and D. Wang, 2019. "An Empirical Study on Phishing Detection with Machine Learning." Published in IEEE Transactions on Information Forensics and Security, this empirical study investigates the application of various machine learning algorithms in phishing detection. The researchers assess the performance and effectiveness of these algorithms, providing valuable insights into their practical utility in real-world scenarios.

[3] B. Soh, A. A. Pirzada, and M. Alsaleh, 2017. "A Survey of Phishing Detection Techniques." In this survey published in Computers & Security, the authors review a range of detection methods, the survey categorizes these techniques and evaluates their strengths and weaknesses, offering a comprehensive overview of the current state of phishing detection technologies.

[4] C. Jones and H. J. C. Ellis, 2011. "An Evaluation of Machine Learning Techniques for Phishing Detection." Presented at the 6th Annual Symposium on Information Assurance (ASIA '11), this paper evaluates different machine learning techniques used for detecting phishing websites. The authors compare various algorithms such as support vector machines and decision trees, highlighting their performance and suitability for phishing detection [7][8][9][10].

Manuscript received on 05 May 2024 | Revised Manuscript received on 13 May 2024 | Manuscript Accepted on 15 May 2024 | Manuscript published on 30 May 2024.
**Correspondence Author(s)*

Joshma K J*, Department of Computer Science, St. Albert's College, Kochi (Kerala), India. Email: joshmakj2001@gmail.com , ORCID ID: 0009-0002-2597-0399

Vineetha Sankar P, Department of Computer Science, St. Albert's College, Kochi (Kerala), India. Email: vineethasankarp@alberts.edu.in

[5] O. K. Sahingoz, E. Buber, O. Demir, and B. Diri, 2019. "Machine learning based phishing detection from URLs." This study, published in Expert Systems with Applications, vol. 117, pp. 345-357, demonstrates the use of machine learning for phishing detection by analyzing URLs. The authors detail the feature extraction process and the application of machine learning models to achieve high detection accuracy.

[6] M. N. Seghir, A. B. Hamida, and F. Labidi, 2019. "Machine Learning Techniques in Phishing URL Detection." Presented at ICOASE 2019, the International Conference on Advanced Science and Engineering, this paper discusses the implementation of various machine learning techniques for detecting phishing URLs. The authors evaluate the performance of these techniques in practical applications, identifying key challenges and proposing solutions for effective phishing detection.

III. METHODOLOGY

Phishing website detection is the broad term for a variety of methods used to detect and stop fraudulent websites that try to trick people into disclosing personal information, credit card numbers, or passwords. These techniques make use of a variety of strategies, such as network-based detection mechanisms, machine learning algorithms, and heuristic analysis. Heuristic analysis is the process of locating possible phishing websites by applying pre-established rules or patterns. One of these guidelines could be to look for oddities in the URL structure, including misspellings, additional characters, or dubious domain names. Heuristic analysis can also be used to examine web page content for common phishing indicators, like demands for private information through email or pop-up windows. Heuristic analysis is a rapid and effective way to identify some phishing attempts, however it might not be adequate to recognize complex phishing attempts that use cutting-edge obfuscation methods. Because machine learning algorithms can examine enormous datasets and find intricate patterns, they have become effective tools for phishing website identification. Support vector machines (SVM), logistic regression, decision trees, random forests, and neural networks are a few examples of supervised learning algorithms that can be trained on labelled datasets that contain attributes that have been taken from both authentic and fraudulent websites. These features could be behavioural traits, HTML content, SSL certificate details, and URL properties. Machine learning algorithms are able to reliably classify websites as either legitimate or phishing based on attributes and can generalize to new, unseen instances by learning from labelled ones. Gradient boosting and bagging are examples of ensemble techniques that combine several weak learners into a robust classifier, hence improving detection accuracy. Although machine learning techniques yield encouraging outcomes, They need a substantial amount of labelled data for training and ongoing updates in order to adjust to changing phishing strategies.

Techniques for behavioural analysis evaluate how users behave when interacting with websites in order to spot indicators of phishing activities. Behavioural analysis systems can detect abnormalities in user behaviour that might point to phishing efforts by tracking mouse movements, keystrokes, and click patterns. Abrupt alterations in surfing activity, including clicking or typing quickly, could indicate automated credential stuffing or phishing attempts. Furthermore, abnormalities in session length, page navigation, and form submission habits can be found using behavioural analysis, which enables the early identification of phishing campaigns. Although behavioural analysis has the potential to identify subtle phishing attempts, user authorization is required for data collection and analysis, which may pose privacy concerns.

To sum up, a wide range of techniques are used in phishing website detection, such as behavioural analysis methods, machine learning algorithms, network-based detection mechanisms, and heuristic analysis. Employing machine learning and artificial intelligence can enhance detection accuracy by analyzing vast datasets and identifying subtle patterns indicative of phishing. Additionally, integrating heuristic analysis helps in quickly flagging known phishing tactics, while behavioral analysis can detect anomalies in user interactions that may suggest phishing activity. Continuous updating and retraining of these models are crucial to adapt to the ever-evolving techniques used by cybercriminals. Collaboration between cybersecurity researchers, industry experts, and organizations can foster the development of more robust and comprehensive anti-phishing solutions.

IV. RESULT AND DISCUSSION

A.Results

The Gradient Boosting algorithm was shown to be the most effective of the five algorithms tested for phishing website identification, with an astounding accuracy score of 97.4%. Gradient Boosting showed the best accuracy among the competitors, including Multilayer Perceptron, Random Forest, Support Vector Machine (SVM), and Logistic Regression. Even with a decent accuracy of 93.4%, Logistic Regression was not as successful as the other techniques. Both Random Forest and SVM had comparable accuracy of 96.4%, demonstrating their efficacy in detecting phishing websites. Furthermore, the Multilayer Perceptron algorithm proved to be a formidable rival to Random Forest and SVM, with a somewhat elevated accuracy of 96.5%. But the Gradient Boosting algorithm really shone out, demonstrating its strength and ability to identify phishing websites with remarkable precision. These findings highlight how well ensemble techniques, such as Gradient Boosting, handle challenging datasets and pick up on subtle patterns in the information. Gradient Boosting's increased accuracy is a measure of its capacity to reduce misclassifications and enhance detection performance in general. As such, companies and cybersecurity professionals might consider the Gradient Boosting algorithm to be an invaluable tool in their toolbox for detecting and reducing possible risks associated with phishing websites. They can strengthen their defences and increase their resistance to changing cyberthreats in the digital sphere by utilizing the power of gradient boosting.

Table 1. Accuracy for Machine Learning Models

Model Name	Accuracy
Logistic Regression	93.4%
Random Forest	96.8%
Support Vector Machine	96.4%
Multilayer Perceptron	96.6%
Gradient Boosting	97.4%

B. Discussion

Implications: The phishing detection system's high accuracy has important ramifications for cybersecurity. Given that phishing assaults continue to be a concern to both individuals and companies, protection measures can be revolutionized by the system's ability to recognize rogue websites. Based on website features, the system quickly identifies and flags phishing attempts, offering a strong defence against online attacks. This feature promotes trust and confidence in online interactions while protecting sensitive data and financial assets. Moreover, the technology makes the digital world safer and more secure by giving consumers better awareness and defence against phishing frauds.

Limitations: The phishing detection system has several restrictions and difficulties even with its high level of effectiveness. The detection capabilities of the system may be challenged by the ever-changing landscape of attack routes, advanced evasion techniques used by cybercriminals, and variability in phishing approaches. Furthermore, there may be false positives and false negatives as a result of the dynamic nature of web content and the growth of trustworthy websites that share characteristics with phishing sites. Furthermore, the system can have trouble identifying extremely focused or unique phishing assaults that diverge from known patterns, highlighting the necessity for constant improvement and adaption to new threats.

Future Enhancement: Continued research and development activities are essential to overcome these issues and enhance the capabilities of phishing detection systems. In order to improve detection accuracy and flexibility, future initiatives can entail investigating sophisticated machine learning algorithms, anomaly detection methods, and behavioural analysis strategies. In addition, the incorporation of contextual data, user feedback mechanisms, and threat intelligence feeds can improve the system's ability to make decisions and its resistance to new threats. Furthermore, in order to effectively tackle phishing assaults, researchers, cybersecurity experts, and industry stakeholders must collaborate in order to share best practices, threat intelligence, and insights. This collaborative ecosystem is fostered by this.

V. CONCLUSION

Detecting phishing websites is an essential part of cybersecurity since online attacks are becoming more and more sophisticated. Using cutting-edge machine learning methods for predictive modelling improves the capacity to recognize and reduce any hazards related to phishing attempts. Pre-processing is a crucial step that addresses problems such as unbalanced datasets, missing data, and superfluous features. Methods like data scaling and feature engineering are essential for maximizing the predictive power of the model. A complete defensive plan is comprised of several algorithms, such as k-nearest neighbours, logistic regression, random forest, support vector machines, and others. However, meticulous data preparation, ongoing model validation, and iterative improvements are necessary for these models to be effective. Best practices include keeping representative, clean, and normalized datasets; prioritizing the selection of pertinent features; and releasing updates often to keep up with changing risks. Metrics for evaluating models, such as recall, accuracy, precision, and F1 score, offer performance information that help firms adjust their tactics. By combining these predictive models with strong cybersecurity procedures, it is possible to protect sensitive data, prevent phishing attacks, and preserve the integrity of online spaces. Strengthening the protection mechanisms against phishing assaults will require constant study, collaboration, and adaption of cutting-edge technologies as the threat landscape changes.

DECLARATION STATEMENT

Funding	No, I did not receive.
Conflicts of Interest	No conflicts of interest to the best of our knowledge.
Ethical Approval and Consent to Participate	No, the article does not require ethical approval and consent to participate with evidence.
Availability of Data and Material	Yes, The dataset for the project is taken from the Kaggle Website.
Authors Contributions	Each author has made an independent contribution to the article. The individual contributions of each author are presented below for clarity and transparency. Joshma K J is the main contributor and Ms. Vineetha Sankar P is the project guide.

REFERENCES

1. M. Bahaghigat, M. Ghasemi, and F. Ozen, "A high accuracy machine learning phishing website detection," Journal of Applications and Information Security, 2023.
2. H. Zhang, D. Gao, and D. Wang, "An Empirical Study on Phishing Detection with Machine Learning," IEEE Transactions on Information Forensics and Security, 2019.
3. B. Soh, A. A. Pirzada, and M. Alsaleh, "A Survey of Phishing Detection Techniques," Computers & Security, 2017.
4. C. Jones and H. J. C. Ellis, "An Evaluation of Machine Learning Techniques for Phishing Detection," in Proceedings of the 6th Annual Symposium on Information Assurance (ASIA '11), 2011.
5. O. K. Sahingoz, E. Buber, O. Demir, and B. Diri, "Machine learning based phishing detection from URLs," Expert Systems with Applications, vol. 117, pp. 345-357, 2019.
6. M. N. Seghir, A. B. Hamida, and F. Labidi, “Machine Learning Techniques in Phishing URL Detection," presented at ICOASE 2019, the International Conference on Advanced Science and Engineering, 2019.
7. Rajeev, H., & Chakkaravarty, Dr. M. (2023). Prediction of Cybercrime using the Avinashak Algorithm. In Indian Journal of Artificial Intelligence and Neural Networking (Vol. 4, Issue 1, pp. 5–10). https://doi.org/10.54105/ijainn.a1078.124123
8. Meenu, & godara, S. (2019). Phishing Detection using Machine Learning Techniques. In International Journal of Engineering and Advanced Technology (Vol. 9, Issue 2, pp. 3820–3829). https://doi.org/10.35940/ijeat.b4095.129219
9. Priya Darshini, Smt. V., Srilatha, P., & Neelima, P. (2019). Detecting Phishing Website with Machine Learning. In International Journal of Recent Technology and Engineering (IJRTE) (Vol. 8, Issue 3, pp. 5626–5629).

https://doi.org/10.35940/ijrte.k1439.098319

10. Phishing Detection using Machine Learning Techniques. (2019). In International Journal of Innovative Technology and Exploring Engineering (Vol. 8, Issue 12S2, pp. 73–78). https://doi.org/10.35940/ijitee.l1014.10812s219
11. Sharma, D., & Sharma, Dr. P. (2021). Design and Implementation of Rainfall Prediction Model using Supervised Machine Learning Data Mining Techniques. In Indian Journal of Data Mining (Vol. 1, Issue 2, pp. 20–26). https://doi.org/10.54105/ijdm.b1615.111221

AUTHORS PROFILE

Joshma K J, currently pursuing Master of Science in Computer Science from the prestigious St. Albert's College (Autonomous), Ernakulam. Prior to this she had completed her Bachelor of Science degree in Computer Science from SN Arts and Science College, Kedamangalam. Her area of interests includes prominent fields like IoT, Networking. She had given attention to details as well as she is able to think outside the box, she loves to solve problems and had been keenly observing the latest technology. When she is not studying or working on new project. She enjoys to listen music, reading. She is an active member of the Computer Science community and coordinates in various events conducted.

Ms. Vineetha Sankar P, is a distinguished academic and educator at St. Albert's College (Autonomous) in Ernakulam, contributing significantly through teaching, research, and mentorship. She holds a Master of Computer Applications (MCA) from the College of Applied Science, Palakkad, and an MPhil in Computer Science. Additionally, she has a degree in Physics from St. Teresa's College, Ernakulam. Currently, she serves as the Class Tutor for MSc Computer Science, Department Exam Coordinator, and Department Placement Coordinator, enhancing the academic and professional development of her students.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of the Lattice Science Publication (LSP)/ journal and/ or the editor(s). The Lattice Science Publication (LSP)/ journal and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

.