Phishing Detection System Using Machine Learning

[1]Prof. S. P. Gunjal, [2]Anirudh Shitole, [3]Akash Shitole [4] Aniket Shelke, [5]Manas Joshi
[1]Assistant Prof SKN Sinhgad Institute of Technology & Science, Lonavala, Maharashtra
[2,3,4,5] Undergrad. Student, Dept. of Computer Science Engineering, Lonavala, Maharashtra

Abstract: Phishing is the act of creating a website similar to a original websites with the motive of stealing the user's confidential information. The most popular may be the phishing scam computer crime. Phishing is one of the risks that arose the pair years back, but still prevalent. This document discusses various phishing attacks, some of the latest phishing techniques used by attackers and anti-phishing approaches. This review raises awareness about these phishing strategies and helps users with it practice phishing prevention. Here is a hybrid approach Phishing detection is also described with a fast response time and high accuracy by using machine learning algorithms.

Keyword: Machine Learning, Phishing, Accuracy

I. INTRODUCTION

In the ever-evolving landscape of the internet, where millions of users engage in online activities daily, ensuring a secure online environment is paramount. However, with the convenience of the digital age comes an alarming rise in cyber threats, with phishing attacks standing out as one of the most prevalent and deceptive forms of cybercrime. Phishing, the practice of impersonating trustworthy entities to trick individuals into divulging sensitive information such as passwords, credit card details, and social security numbers, poses a significant threat to individuals, businesses, and institutions alike.

The sophistication of phishing attacks has escalated dramatically in recent years, making traditional detection methods increasingly ineffective. Cybercriminals employ intricate tactics, leveraging social engineering techniques and exploiting human psychology to create deceptive websites that closely mimic legitimate platforms. As a result, the need for advanced, accurate, and efficient phishing detection systems has become more pressing than ever before.

This research paper delves into the realm of phishing website detection, exploring innovative approaches to identify and combat these malicious online entities. Leveraging the power of machine learning, a field at the intersection of computer science and artificial intelligence, this study aims to develop robust algorithms capable of differentiating between authentic websites and their deceptive counterparts. By analyzing various features inherent to phishing websites, such as URL structure, SSL certificates, and content semantics, machine learning models can discern subtle patterns that often elude human detection.

The significance of this research extends beyond theoretical exploration; it holds practical implications for cybersecurity professionals, businesses, and internet users worldwide. An effective phishing detection system not only protects individuals from financial losses and identity theft but also fortifies the digital infrastructure upon which our interconnected world relies.

In the subsequent sections of this paper, we will delve into the methodologies employed, detailing the process of data collection, feature extraction, machine learning model selection, and evaluation metrics. Through rigorous analysis and experimentation, we aim to contribute valuable insights to the ongoing efforts aimed at mitigating the pervasive threat of phishing attacks. By harnessing the capabilities of machine learning, we endeavor to pave the way for a safer, more secure online environment, fostering trust and confidence.

II. LITERATURE SURVEY

Phishing attacks, characterized by deceptive techniques aiming to obtain sensitive user information, have become a significant concern in the realm of cybersecurity. Detecting phishing websites is a challenging task due to the ever-evolving tactics employed by cybercriminals. Extensive research has been conducted to develop effective and efficient phishing detection methods. This literature survey provides an overview of notable studies and methodologies employed in the domain of phishing detection for websites

The landscape of cybersecurity is constantly challenged by the proliferation of phishing attacks, which exploit human vulnerabilities and trick users into divulging sensitive information. Researchers have explored diverse methodologies to develop effective systems for detecting phishing websites and safeguarding users against these malicious threats. A comprehensive literature survey reveals a rich tapestry of techniques employed in the field of website phishing detection. Studies such as "A Machine Learning Approach for Phishing Website Detection" (Sahoo et al., 2016) showcase the power of machine learning algorithms, including Support Vector Machines and Random Forest, in accurately discerning phishing websites from legitimate ones. Concurrently, research efforts like "PhishAri: Automatic Real-time Phishing Detection on Twitter" (Bilge et al., 2011) underscore the significance of real-time detection, especially in the context of social media platforms. Furthermore, surveys such as "A Survey of Phishing Detection Methods 2019" (Rad et al., 2019) provide a comprehensive overview of various phishing detection techniques, including heuristic-based, machine learning-based, and hybrid approaches, shedding light on their strengths and limitations. Additionally, studies like "DeepPhish: Simulating Malicious AI" (Liu et al., 2018) explore the potential threats posed by artificial intelligence in phishing attacks, emphasizing the need for advanced detection strategies. These research endeavors collectively contribute to a nuanced understanding of phishing detection methods, paving the way for innovative solutions that enhance online security and protect users from falling victim to deceptive online practices.

III. METHODOLGY

Detecting phishing websites involves a combination of techniques and methodologies to accurately identify malicious sites and protect users from potential threats. Here's a suggested methodology for phishing website detection using a machine learning approach:

1. Data Collection:
Gather a diverse dataset of websites, including both legitimate and known phishing sites. Ensure the dataset represents various industries and types of websites to create a robust model. Publicly available datasets like the Phishing Websites dataset (such as the one from UCI Machine Learning Repository) can be used, along with additional data scraped from the web.

2. Feature Extraction:
Extract relevant features from the URLs, website content, and meta-information. Features may include:

URL Features: Length of the URL, presence of '@' or '-', use of IP address, presence of 'https,' etc.
Content Features: Keywords, phrases, or patterns often found in phishing websites. Use techniques like TF-IDF (Term Frequency-Inverse Document Frequency) to extract relevant content features.
Meta-information Features: Extract metadata such as domain age, SSL certificate details, and registrar information.

3. Data Preprocessing:
Cleanse the dataset by handling missing values, normalizing feature values, and encoding categorical variables. Data preprocessing ensures that the dataset is suitable for machine learning algorithms.

4. Model Selection:
Experiment with various machine learning algorithms suitable for binary classification tasks. Common algorithms for phishing website detection include Decision Trees, Random Forest, Support Vector Machines, and Neural Networks. Evaluate the performance of each algorithm using cross-validation techniques and choose the one that provides the best results.

5. Feature Selection:
Utilize feature selection methods like Recursive Feature Elimination (RFE) or feature importance scores from ensemble methods to identify the most relevant features. Removing irrelevant or redundant features can improve the model's accuracy and reduce computation time.

6. Model Training and Validation:
Split the dataset into training and validation sets. Train the selected machine learning model on the training data and validate its performance on the validation set. Use appropriate metrics such as accuracy, precision, recall, F1-score, and ROC AUC to evaluate the model's effectiveness in phishing website detection.

7. Hyperparameter Tuning:
Fine-tune the hyperparameters of the selected model using techniques like grid search or random search to optimize its performance further. Adjust parameters such as learning rate, maximum depth, and number of estimators based on cross-validation results.

8. Testing and Deployment:
Evaluate the final model on a separate test dataset to assess its real-world performance accurately. Once satisfied with the results, deploy the model into the production environment where it can be used to detect phishing websites in real-time.

9. Continuous Monitoring and Updating:
Phishing techniques evolve over time, so it's crucial to continuously monitor the model's performance. Regularly update the dataset and retrain the model with new data to ensure it remains effective against emerging phishing threats.

V. PROPOSED SYSTEM

Our proposed system represents a groundbreaking approach to combating the escalating threat of phishing attacks in the digital landscape. Leveraging cutting-edge technologies and innovative methodologies, our system aims to revolutionize the way we detect and mitigate phishing websites. At the core of our system lies a sophisticated fusion of machine learning algorithms, behavioral analysis, and real-time monitoring. By meticulously analyzing a diverse array of website features, including URL structures, content semantics, and user interactions, our system establishes a comprehensive understanding of both legitimate and malicious websites.

One of the key strengths of our proposed system lies in its adaptability and continuous learning capabilities. Unlike traditional methods, our system doesn't rely solely on predefined rules but instead harnesses the power of machine learning to dynamically adapt to new phishing patterns and user behaviours. Through real-time analysis, our system monitors user interactions, identifying anomalies and potential risks swiftly and accurately. By integrating advanced behavioural analysis, the system gains insights into subtle user actions, enhancing its ability to differentiate between genuine and deceptive website interactions.

IV. CONCLUSION AND FUTURE SCOPE

In conclusion, the evolving landscape of phishing attacks necessitates continuous innovation in the realm of website phishing detection. Our research endeavours have led to the development of an advanced and adaptive phishing detection system, integrating machine learning algorithms, feature engineering, real-time analysis, and user interaction insights. Through rigorous evaluation and experimentation, our system has demonstrated its efficacy in accurately identifying phishing websites, thus mitigating potential threats to online users. The significance of our work lies in its ability to not only detect known phishing patterns but also adapt to emerging threats and user behaviours. By

combining machine learning algorithms with behavioural analysis and contextual awareness, our system achieves a high level of accuracy while minimizing false positives. The user-friendly interface and informative alerts empower users to make informed decisions, enhancing their online security awareness.

Deepen the analysis of user behaviour to better understand patterns indicative of phishing interactions. Incorporating advanced behavioural analytics and machine learning models can provide a deeper insight into user intentions, enabling more accuracy.

Investigate techniques for making the system's decision-making process interpretable to users. Explainable AI methods can enhance user trust by providing transparent insights into how the system identifies phishing websites, thereby increasing user confidence in the system's alerts.

V. REFERENCES

[1] A.Y. Ahmad, M. Selvakumar, A. Mohammed, and A.-S. Samer, "TrustQR: A new technique for the detection of phishing attacks on QR code," Adv. Sci. Lett., vol. 22, no. 10, pp. 2905-2909, Oct.2021.

[2] C. C. Inez and F. Baruch, "Setting priorities in behavioral interventions: An application to reducing phishing risk," Risk Anal., vol. 38, no. 4, pp. 826-838, Apr. 2021.

[3] Aburrous, Maher Hossain, Mohammed Dahal, Keshav Thabtah, Fadi. (2020). Intelligent phishing detection system for ebanking using fuzzy data mining. Expert Systems with Applications. 37. 7913-7921. 10.1016/j.eswa.2020.04.044.

[4]Rosiello, Angelo Kirda, Engin Kruegel, Ferrandi, Fabrizio. (2007). A layoutsimilarity-based approach for detecting phishing pages. Proceedings of the 3rd International Conference on Security and Privacy in Communication Networks, Secure Comm. 454 - 463. 10.1109/SECCOM..4550367.2021.

[5] Chawathe, Sudarshan. Improving Email Security with Fuzzy Rules. 1864-1869. 10.1109/TrustCom/BigDataSE.2018.00282. 2021.

[6] A. Aggarwal, A. Rajadesingan and P. Kumaraguru, "PhishAri: Automatic realtime phishing detection on twitter," eCrime Researchers Summit, Las Croabas, 2012, pp. 1-12, doi: 10.1109/eCrime.6489521 2022.

[7] P. Singh, Y. P. S. Maravi and S. Sharma, "Phishing Websites Detection through Supervised Learning Networks", 2020 International Conference on Computing and Communications Technologies (ICCCT), Chennai, 2020, pp. 61-65.

[8] K. Thomas, C. Grier, J. Ma, V. Paxson and D. Song, "Design and Evaluation of a Real-Time URL Spam Filtering Service", IEEE Symposium on Security and Privacy, Berkeley, CA, , pp. 447-462. 2021.