JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

INTELLIGENT METHODS FOR ACCURATELY DETECTING PHISHING WEBSITES

[1] Dr. R. S. Khule, [2]Shraddha Jadhav, [3]Nikita Sanap, [4]Shital Satote, [5]Sakshi Ugale

[1]Professor, Information Technology Department, Matoshri College of Engineering and Research Centre, Nashik
[2,3,4,5]Student, Information Technology Department, Matoshri College of Engineering and Research Centre, Nashik

Abstract: Phishing attacks pose a severe threat to online security, making the accurate detection of phishing URLs a critical endeavor. This abstract introduces a novel approach that leverages intelligent methods based on machine learning algorithms to enhance the precision of phishing URL detection. In an era where cybercriminals continuously adapt and evolve their tactics, this research aims to develop a robust and proactive defense mechanism. By harnessing the power of machine learning, this study seeks to train models capable of identifying subtle patterns and characteristics in URLs that are indicative of phishing attempts. Through a comprehensive analysis of various features, including domain structure, textual content, and lexical attributes, the proposed methods aim to provide a multi-faceted and dynamic approach to detect phishing URLs, ultimately bolstering cybersecurity efforts. The proposed work focuses on the design and evaluation of a range of machine learning algorithms, including decision trees, random forests, and naive bays, to uncover the most effective techniques for phishing URL detection. By harnessing a diverse set of features and utilizing labeled datasets for training, the models are expected to learn and adapt to emerging phishing techniques. As phishing threats continue to evolve in complexity and scale, the integration of intelligent methods based on machine learning is poised to be a critical step forward in the ongoing battle to protect digital identities and sensitive information.

Index Terms **- Phishing URL Detection, Cybersecurity, Feature Analysis, Decision Trees, Random Forests, Pattern Recognition.**

I. INTRODUCTION

Intelligent methods for accurately detecting phishing URLs have become a crucial aspect of cybersecurity, given the growing sophistication of phishing attacks. Machine learning algorithms have proven to be valuable tools in this endeavor. These algorithms use a combination of feature engineering and training on labeled datasets to differentiate between legitimate and malicious URLs. Features like URL length, domain reputation, presence of suspicious keywords, and SSL certificate status are commonly used to train models. Machine learning models, such as decision trees, random forests, support vector machines, and deep neural networks, are then employed to classify URLs as either benign or phishing. By learning from historical data, these algorithms can adapt to new and evolving phishing techniques, enhancing their accuracy.

One of the key advantages of using machine learning for phishing URL detection is its ability to generalize patterns and identify subtle characteristics of malicious URLs that may be difficult to detect using traditional rule-based systems. Machine learning models can analyze vast amounts of data quickly, making them effective in real-time detection and scalable for large-scale web monitoring. Additionally, these methods can be continually updated with new data, allowing them to stay current with emerging phishing threats. However, it's essential to ensure that the training data is representative and up to date, as well as to employ ensemble techniques and anomaly detection to minimize false positives and negatives.

The challenge of accurately detecting phishing URLs using machine learning lies in the constantly evolving tactics employed by cybercriminals. Phishers frequently change their techniques to bypass detection systems, making it crucial to keep machine learning models up-to-date and adapt them to emerging threats. Collaboration between cybersecurity experts, data scientists, and machine learning engineers is essential to stay ahead of the ever-changing phishing landscape. As machine learning technology continues to advance, it holds great promise in the ongoing battle against phishing attacks, offering more sophisticated and accurate methods for protecting individuals and organizations from falling victim to these fraudulent schemes.

II. Literature Survey

This paper [1] serves as a valuable reference, focusing on the crucial issue of phishing website detection. Phishing websites, by mimicking legitimate ones, pose a significant threat by attempting to illicitly obtain sensitive information from unsuspecting users. The study provides an in-depth exploration of a novel approach that employs joint features for the detection of these deceptive websites. The research draws its original data from Phish Tank, a reputable source in the field, and utilizes a hybrid training approach, combining Support Vector Machine (SVM) and Bayesian methods to train the classifier effectively. Additionally, the implementation incorporates the use of Wireshark for packet flow detection, enhancing the system's ability to discern potential threats. Impressively, after the training phase, the classifier demonstrates a remarkable detection speed of 1000 websites per second, highlighting its efficiency and scalability.

As the internet continues to grow in popularity, traditional shopping habits are shifting towards electronic commerce. However, this transition has also brought forth new challenges in the form of cybercrime, with criminals employing anonymous internet frameworks and sophisticated tactics like phishing to deceive and gather sensitive user information. To address this issue, this study [2] utilizes machine learning techniques to detect phishing websites by analyzing the HTML code structure within hyperlinks. Two machine learning algorithms, Random Forest, and Support Vector Machine were evaluated, with Random Forest achieving the highest accuracy at 99.98 percent, making it the preferred choice for the system's implementation in detecting phishing websites.

The rise of e-commerce transactions has exposed users and industries to significant cybersecurity threats, especially from phishers and cybercriminals. Phishing attempts, where users are deceived into divulging sensitive information, pose a substantial risk to online safety and the global economy. To combat this, the implementation of effective phishing detection algorithms is imperative. Phishing involves deceptive practices like mimicking legitimate websites to trick users into sharing personal credentials. This study [3] focuses on developing a system utilizing various machine learning methods, including Logistic Regression and ensemble algorithms like Adaboost and Gradient Boost, while introducing a hybrid stacking model. The research findings demonstrate that the proposed Stacking Classifier achieves an impressive 85.6% accuracy rate, showcasing the effectiveness of the hybrid approach in enhancing phishing website prediction accuracy compared to individual classifiers.

III. Methodology

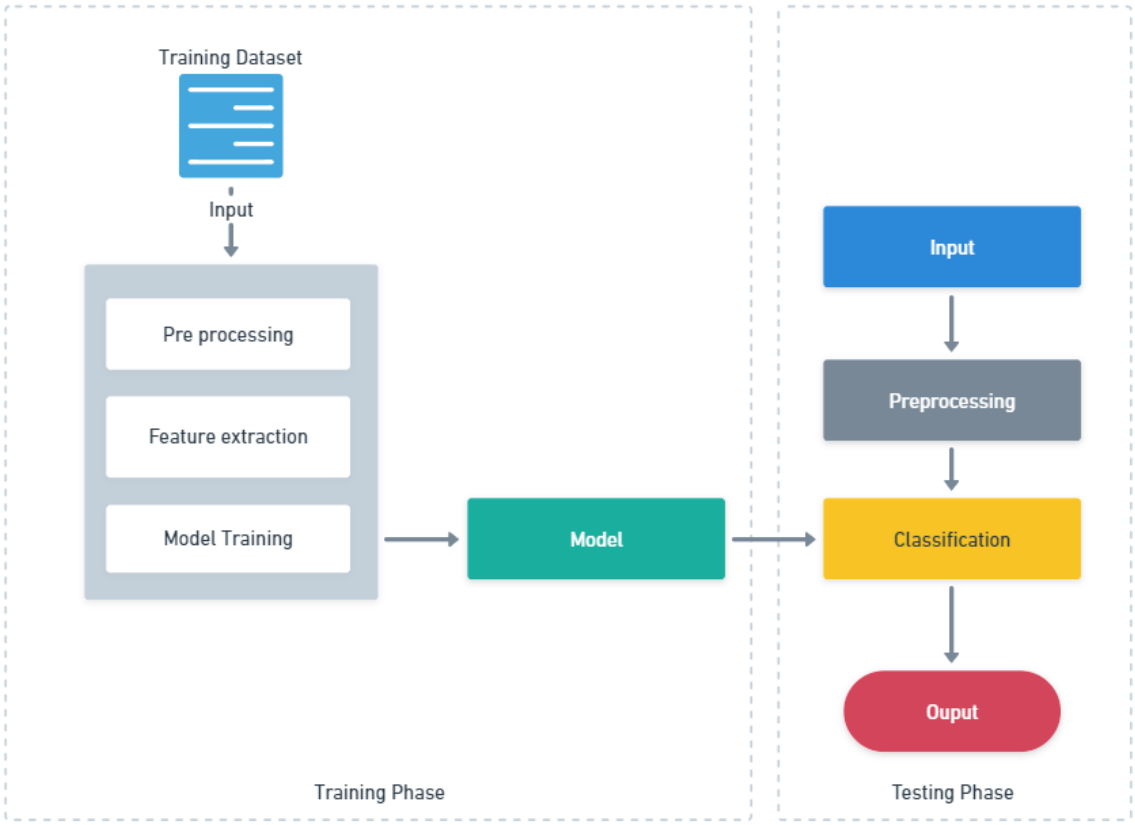

Figure 1 System Architecture

Detecting phishing URLs accurately is a critical challenge in cybersecurity, and machine learning algorithms offer a promising approach to tackle this problem. The methodology for developing intelligent methods to accurately detect phishing URLs typically involves several key steps. Firstly, data collection is vital, as a substantial dataset of both legitimate and phishing URLs is required for training and testing the machine learning models. Features extraction plays a pivotal role, where relevant attributes such as URL structure, domain reputation, and content are identified to characterize each URL. Feature engineering is essential to create informative input vectors for the machine learning models.

Next, model selection and training are crucial steps. Various machine learning algorithms, such as decision trees, random forests, support vector machines, and deep neural networks, can be employed. The selected algorithms are trained on the feature-rich dataset, and techniques like cross-validation are used to assess their performance. Hyperparameter tuning may also be necessary to optimize the models. To evaluate the models, metrics like precision, recall, F1-score, and ROC-AUC are employed, considering the trade-off between false positives and false negatives. Finally, the chosen model is deployed in a real-world environment, integrated into security systems to proactively identify, and mitigate phishing threats, contributing to enhanced online security. Regular updates and retraining are vital to keep the model effective against evolving phishing techniques.

VI. TRADITIONAL APPROACHES VS MACHINE LEARNING APPROACH

Traditional approaches to phishing URL detection primarily rely on static blacklisting and heuristic-based methods. These methods involve maintaining databases of known malicious URLs and using predefined rules to flag suspicious ones. While these methods are simple and efficient, they often fall short in dealing with novel or rapidly evolving phishing attacks. Phishers can easily circumvent these techniques by constantly changing their URLs or using obfuscation techniques. Moreover, the reliance on static databases can result in false negatives, as it may take some time for a malicious URL to be added to the blacklist. Overall, traditional approaches offer a basic level of protection but are not well-suited to combat the constantly changing landscape of phishing attacks. On the other hand, machine learning approaches leverage the power of artificial intelligence to dynamically adapt to new and emerging threats. These approaches use supervised learning algorithms to analyze various features of URLs, learning to distinguish between legitimate and phishing URLs based on historical data. This adaptability allows them to detect previously unseen phishing URLs and identify evolving attack patterns. Machine learning models can consider a wide range of features, including lexical properties, domain age, and content analysis, making them more robust and accurate. However, machine learning approaches require substantial amounts of labelled data for training, and their performance heavily depends on the quality and representativeness of the training dataset. Additionally, they may be susceptible to adversarial attacks designed to fool the model. Traditional approaches are simple but limited in their ability to tackle evolving phishing threats, while machine learning approaches offer a more adaptive and dynamic defence, but they come with their own set of challenges, such as the need for high-quality training data and potential vulnerability to adversarial attacks. The choice between these approaches often depends on the specific security requirements and the organization's capacity to implement and maintain sophisticated machine learning systems.

V. CONCLUSION

Phishing is a way to deceive via fake e-mails and websites to steal people’s private information. Phishing prevents individuals from carrying out their activities via the internet. Phishing website detection is crucial for the internet community since it has a big impression on online transactions performed. Random Forest is an intelligent machine learning method that was recently paid attention to by researchers due to its speed and high classification accuracy. The phishing website problem has been investigated in this work in which we are developing a machine learning model to determine correlations between the features and yield them from simple and effective rules. In this study, we adopted a classifier model that is used for detecting phishing websites in an intelligent and automated way by using publicly available datasets.

REFERENCES

[1] Z. Fan, "Detecting and Classifying Phishing Websites by Machine Learning," 2021 3rd International Conference on Applied Machine Learning (ICAML), Changsha, China, 2021, pp. 48-51, doi: 10.1109/ICAML54311.2021.00018.
[2] N. Binti Md Noh and M. N. Bin M. Basri, "Phishing Website Detection Using Random Forest and Support Vector Machine: A Comparison," 2021 2nd International Conference on Artificial Intelligence and Data Sciences (AiDAS), IPOH, Malaysia, 2021, pp. 1-5, doi: 10.1109/AiDAS53897.2021.9574282.
[3] A. Lakshmanarao, P. S. P. Rao and M. M. B. Krishna, "Phishing website detection using novel machine learning fusion approach," 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), Coimbatore, India, 2021, pp. 1164-1169, doi: 10.1109/ICAIS50930.2021.9395810.
[4] A. Mandadi, S. Boppana, V. Ravella and R. Kavitha, "Phishing Website Detection Using Machine Learning," 2022 IEEE 7th International conference for Convergence in Technology (I2CT), Mumbai, India, 2022, pp. 1-4, doi: 10.1109/I2CT54291.2022.9824801.
[5] C. Gu, "A Lightweight Phishing Website Detection Algorithm by Machine Learning," 2021 International Conference on Signal Processing and Machine Learning (CONF-SPML), Stanford, CA, USA, 2021, pp. 245-249, doi: 10.1109/CONF-SPML54095.2021.00054.
[6] Y. Sönmez, T. Tuncer, H. Gökal and E. Avcı, "Phishing web sites features classification based on extreme learning machine," 2018 6th International Symposium on Digital Forensic and Security (ISDFS), Antalya, Turkey, 2018, pp. 1-5, doi: 10.1109/ISDFS.2018.8355342.
[7] M. H. Alkawaz, S. J. Steven and A. I. Hajamydeen, "Detecting Phishing Website Using Machine Learning," 2020 16th IEEE International Colloquium on Signal Processing & Its Applications (CSPA), Langkawi, Malaysia, 2020, pp. 111-114, doi: 10.1109/CSPA48992.2020.9068728.
[8] M. D. Bhagwat, P. H. Patil and T. S. Vishawanath, "A Methodical Overview on Detection, Identification and Proactive Prevention of Phishing Websites," 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), Tirunelveli, India, 2021, pp. 1505-1508, doi: 10.1109/ICICV50876.2021.9388441.
[9] Z. Fan, "Detecting and Classifying Phishing Websites by Machine Learning," 2021 3rd International Conference on Applied Machine Learning (ICAML), Changsha, China, 2021, pp. 48-51, doi: 10.1109/ICAML54311.2021.00018.

[10] V. Patil, P. Thakkar, C. Shah, T. Bhat and S. P. Godse, "Detection and Prevention of Phishing Websites Using Machine Learning Approach," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 2018, pp. 1-5, doi: 10.1109/ICCUBEA.2018.8697412.(4), 754-764.