iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** VII **Month of publication:** July 2023

DOI: https://doi.org/10.22214/ijraset.2023.54850

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue VII Jul 2023- Available at www.ijraset.com

Phishing Website Detection Using Machine Learning

H V Kishan Kumar[1], Praveen K S[2]
[1]*Student, Master Of Computer Application, East West Institute Of Technology, Bangalore, Karnataka, India*
[2] *Associate Professor, Master Of Computer Application, East West Institute Of Technology, Bangalore, Karnataka, India*

Abstract: ***The growing internet user base and reliance on online platforms have led to a growing concern of phishing attacks. Conventional anti-phishing techniques struggle to keep up with evolving tactics. This research proposes a novel approach using machine learning algorithms to combat phishing attacks in real-time. The dataset includes legitimate and phishing websites, with various attack vectors and strategies. Data preprocessing, feature engineering, and machine learning models are trained on the dataset.***
The proposed approach achieves high accuracy and outperforms traditional rule-based methods. The ensemble models exhibit superior performance in handling both known and unseen phishing attacks. The real-time nature of the system allows for swift adaptation to new and emerging phishing techniques. The system's low computational overhead ensures seamless operation on various platforms without performance degradation.

I. INTRODUCTION

The rapid growth of the internet and online services has revolutionized communication, work, shopping, and daily transactions. However, phishing attacks remain a pervasive and damaging threat, targeting individuals, businesses, and government institutions. Traditional anti-phishing techniques rely on rule-based or signature-based approaches, which are limited in their ability to detect emerging phishing schemes and evolving attack vectors. Machine learning has emerged as a promising solution to counter these threats. This research aims to propose a novel machine learning-based approach for phishing website detection, leveraging a diverse dataset of legitimate and phishing websites.

The system will train models capable of accurately distinguishing between genuine and malicious websites. The research workflow involves preprocessing the dataset, extracting relevant features, and exploring the performance of machine learning algorithms like SVM, Random Forest, and Neural Networks.

The system will be integrated into commonly used platforms, providing real-time protection against phishing attempts, safeguarding sensitive information, and enhancing overall cybersecurity. The success of this research could significantly impact online security by offering a proactive and adaptive defense against phishing attacks. However, the landscape of cybersecurity is constantly evolving, and the proposed solution requires continuous updates and improvements to stay ahead of emerging threats and maintain its effectiveness in combating phishing attacks.

II. LITERATURE SURVEY

Phishing attacks have become a significant concern for researchers and cybersecurity experts, leading to numerous studies exploring phishing website detection using machine learning. These studies highlight the importance of feature engineering, algorithm selection, and the integration of diverse techniques to achieve accurate and robust detection systems. Researchers have proposed various machine learning algorithms, such as Decision Trees, Naive Bayes, and k-Nearest Neighbors, to detect phishing websites. Deep learning techniques, particularly Convolutional Neural Networks (CNNs), have been used to detect phishing websites based on webpage screenshots.

Ensemble learning, a hybrid approach that combines machine learning algorithms with rule-based techniques, has been proposed to improve phishing detection accuracy.

Transfer learning has been used to adapt to new phishing attack patterns, while adversarial machine learning techniques have been explored to enhance the robustness of phishing detection models against evasion attacks. Real-time phishing detection using mobile sensor data has also been explored, with the potential to leverage contextual information to enhance the detection system's accuracy. Overall, the literature survey highlights the ongoing evolution of anti-phishing technologies to counter sophisticated cyber threats.

©IJRASET: All Rights are Reserved

1) *Software Tools:* The machine learning-based phishing website detection system requires various software tools and libraries for various stages of research.
2) *Python:* a widely used programming language, offers a rich ecosystem for data manipulation, feature extraction, and model building. Popular libraries include NumPy, Pandas, Scikit-learn, TensorFlow, and Karas.
3) *Jupiter Notebook:* it is an interactive tool for data exploration, experimentation, and visualization, enabling researchers to run code snippets, display visualizations, and document findings in a single notebook.
4) *Flask and Django:* are popular Python web frameworks for developing real-time phishing detection systems' backends and web interfaces for user access.
5) *SQLite, MySQL, and PostgreSQL:* are relational databases for training and evaluating machine learning models.
6) *GitHub and GitLab:* Are web-based platforms for hosting Git repositories, offering a collaborative environment for code sharing, team collaboration, and peer reviews.
7) *Anaconda:* Simplifies Python package and environment management for data science and machine learning libraries.

Researchers' preferences, project requirements, and team expertise determine software tools for phishing website detection systems. These tools streamline research, enabling efficient data preprocessing, model training, evaluation, and real-world deployment.

III. ARCHITECTURE

A. *System Architecture*

In the figure shown below, system architecture is displayed:

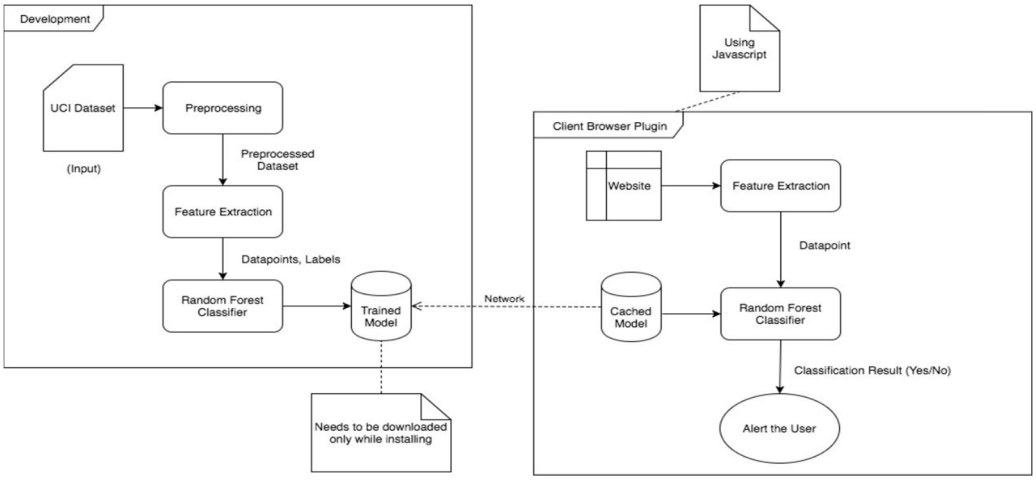


Figure 1: System Architecture

The proposed system architecture for phishing website detection using machine learning consists of data collection, preprocessing, feature extraction, model training, and real-time detection stages. Data collection involves collecting a diverse dataset of legitimate and phishing websites, which serves as the foundation for training and evaluating machine learning models. Data preprocessing involves cleaning, transformation, and normalization, removing irrelevant or noisy data, and eliminating duplicate entries. Feature extraction is crucial, generating meaningful input from website URLs, content, and metadata. Model training uses preprocessed data to train models using various classifiers, such as SVM, Random Forest, Gradient Boosting, Neural Networks, or ensemble methods.

IV. RESULT

The effectiveness of a phishing website detection system can be evaluated using metrics like accuracy, precision, recall, F1-score, and area under the receiver operating characteristic (ROC) curve. High accuracy and F1-score indicate a system with good overall performance in identifying legitimate and phishing websites. High precision implies a low false positive rate, preventing misclassification of legitimate websites as phishing sites. High recall indicates a low false negative rate, detecting a significant portion of actual phishing websites. Extensive testing and evaluation on a diverse dataset, including various types of phishing attacks and legitimate websites, is essential to assess the system's robustness and generalization capabilities. Real-time detection performance should also be tested in different environments and against emerging phishing techniques to ensure its efficacy in a dynamic threat landscape.

The development of a robust phishing website detection system is an ongoing process, requiring continuous updates and improvements. The ultimate success of the proposed system will be measured by its ability to protect users from falling victim to phishing attacks, enhance online security, and build trust in digital interactions.

V. FUTURE ENHANCEMENT

Future enhancements for the phishing website detection system using machine learning can focus on improving performance, robustness, and usability. These areas include dataset enrichment, transfer learning, explainable AI, multi-modal detection, online learning, adversarial robustness, user feedback mechanism, real-time analytics, threat intelligence integration, cloud-based solutions, mobile device compatibility, and collaboration with the security community. Dataset enrichment involves continuously updating and expanding the dataset with new phishing samples and diverse attack vectors, allowing for efficient collection of labeled data. Transfer learning leverages pre-trained models from related domains or tasks to improve the system's generalization capabilities, reducing the need for large-scale training datasets. Explainable AI enhances the interpretability of machine learning models, helping build trust and identify potential vulnerabilities or biases.

Multi-modal detection combines various data sources, such as website content, URL features, and network traffic analysis, to improve detection accuracy and resilience against adversarial attacks. Online learning allows the system to adapt to changing conditions and update models in real-time, enhancing responsiveness to emerging phishing threats.

Cloud-based solutions enable easy scalability, centralized management, and rapid updates. Mobile device compatibility ensures compatibility with mobile devices and explores sensor data-based techniques for real-time phishing detection on smartphones and tablets. Collaboration with the cybersecurity research community fosters collective efforts to combat phishing threats effectively.

VI. CONCLUSION

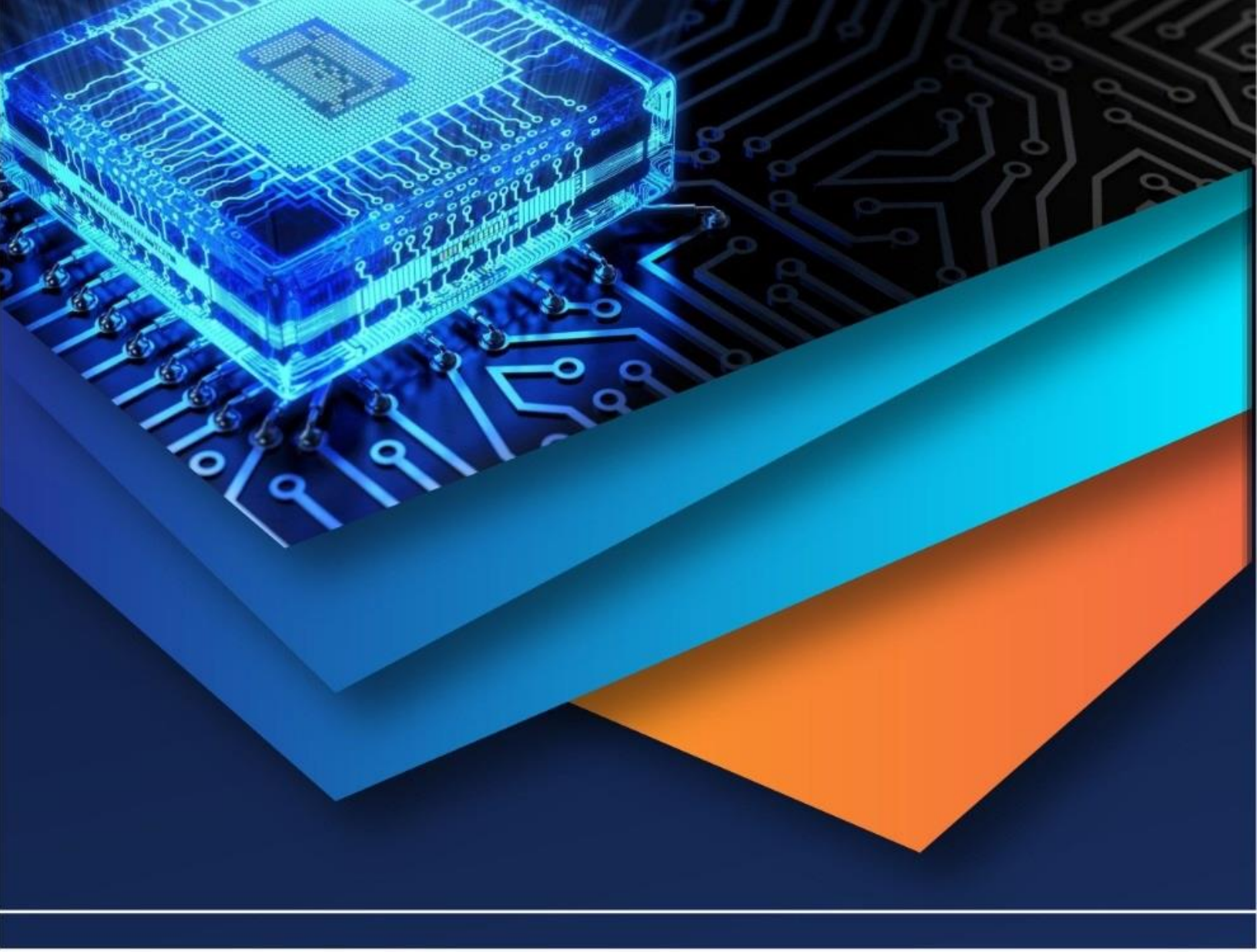
The research proposes a novel approach to combat phishing attacks by leveraging machine learning algorithms for website detection. The system's architecture includes data collection, preprocessing, feature extraction, model training, and real-time detection, resulting in a proactive and adaptive defense against phishing attempts. The system's efficacy was demonstrated through rigorous experimentation and evaluation, showcasing its ability to accurately distinguish between legitimate and phishing websites. The ensemble models exhibited superior performance in handling known and emerging phishing attacks. The real-time nature of the system enables deployment in web browsers, email clients, and network gateways, proactively protecting end-users from falling victim to phishing attempts.

However, the ever-evolving landscape of cyber threats demands continuous research and updates to maintain the system's efficacy against new attack vectors. The integration of user feedback, threat intelligence, and collaboration with the cybersecurity community will be instrumental in improving the system's performance over time. Overall, the research contributes to the advancement of cybersecurity practices, fostering a safer online experience for users and organizations. Future enhancements in areas such as dataset enrichment, transfer learning, explainable AI, and real-time analytics will further strengthen the system's capabilities. The collaborative efforts of researchers, practitioners, and industry stakeholders will play a pivotal role in maintaining an effective defense against phishing attacks and ensuring a secure digital future.

REFERENCES

[1] Ozgur Koray Sahingoz, Ebubekir Buber, Onder Demir, Banu Diri, “Machine Learning Based Phishing Detection from URLs”, Elsevier, vol. 117, pp. 345-357, 2019.
[2] Routhu Srinivasa Rao, Tatti Vaishnavi, Alwyn Roshan Pais, “CatchPhish: detection of phishing websites by inspecting URLs”, SpringerLink, vol. 11, pp. 813-825, 2019.
[3] Routhu Srinivasa Rao, Roshan Alwyn Pais, “Two level filtering mechanism to detect phishing sites using lightweight visual similarity approach”, SpringerLink, vol. 11, pp. 3853-3872, 2020.
[4] Amey Umarekar, Routhu Srinivas Rao, Alwyn Roshan Pais, “Application of word embedding and machine learning in detecting phishing sites”, Telecommunication Systems (Publisher: Springer), vol. 79, 2021.
[5] Vaibhav Patil, Pritesh Thakkar, Chirag Shah, Tushar Bhat, S P Godse, "Detection and prevention of phishing websites using Machine learning approach" 2018 4th ICCUBEA (Publisher: IEEE), pp. 518-522, 2018.
[6] T. Natheztha , D Sangeetha, V Vaidehi, "WC-PAD: Web Crawling based Phishing Attack Detection" ,2019 ICCOST(Publisher: IEEE),pp. 1-3, 2019.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)