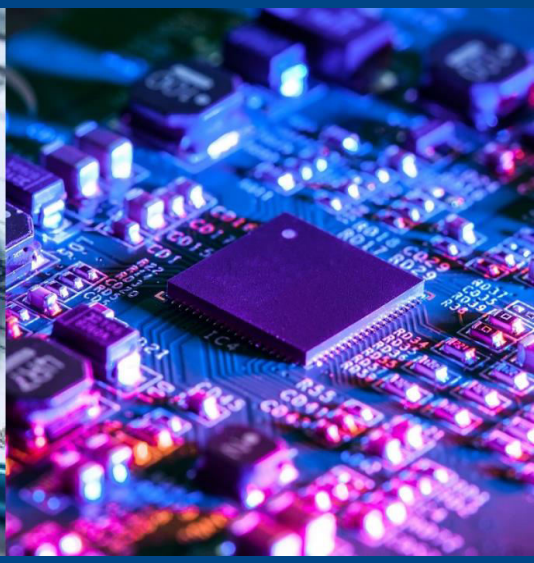
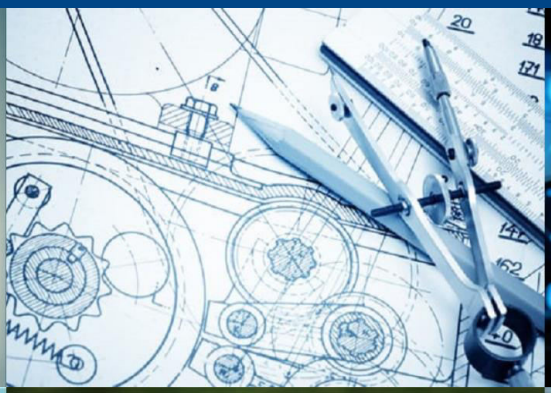
ISSN: 2582-7219

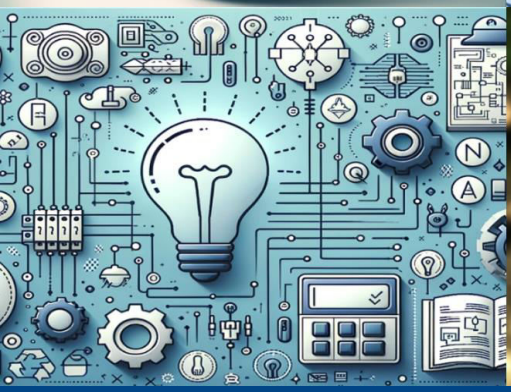




International Journal of Multidisciplinary Research in Science, Engineering and Technology

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.206 **Volume 8, Issue 5, May 2025**

© 2025 IJMRSET | Volume 8, Issue 5, May 2025| DOI:10.15680/IJMRSET.2025.0805063

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Machine Learning Based Phishing URL Detection

Suyash Kakade, Prof. Vishwatej Pisal

PG Student, Dept. of MCA, Anantrao Pawar College of Engineering and Research, Parvati, Pune, India

Assistant Professor, Dept. of MCA, Anantrao Pawar College of Engineering and Research, Parvati, Pune, India

ABSTRACT: Phishing, a common cyber threat, tricks users into revealing sensitive data through fraudulent emails or websites. Traditional detection methods struggle to keep up with new phishing tactics. The use of machine learning to identify phishing websites is investigated in this research. By analyzing URLs and web content, we improve detection accuracy without relying on external systems. We evaluate various ML algorithms and fine-tuned parameters to reduce false positives and negatives. Our findings highlight the effectiveness of ML in bolstering cybersecurity against phishing attacks.

KEYWORDS: (Phishing, Cybersecurity, Machine Learning, Detection)

I. INTRODUCTION

Phishing is particularly notable in the field of cybersecurity as a pervasive and insidious threat, exploiting human vulnerability to perpetrate malicious activities. At the heart of many phishing attacks lies the deceptive use of Uniform Resource Locators (URLs), the web addresses that take users to particular websites. Developing successful detection and prevention techniques against this constantly changing cybercrime requires an understanding of the crucial role that URLs play in phishing.

Phishing, a type of cybercrime, uses a variety of communication methods, such as email, text messages, and phone calls, to trick people into divulging important information by impersonating trustworthy organizations. But frequently, the URLs that are included in these messages act as the entry point for fraud and exploitation.By mimicking the URLs of trusted organizations or employing subtle variations and obfuscation techniques, goal of cybercriminals is to trick gullible people into disclosing private information, including financial information, login passwords, and personal information. These deceptive URLs serve as the linchpin of phishing schemes, exploiting trust and familiarity to lure victims into compromising their security.

Researchers and cybersecurity professionals have focused more on creating sophisticated methods for URL-based detection and analysis as a result of realizing how important URLs are in phishing.
Machine learning algorithms, in particular, offer a promising avenue for identifying suspicious URLs and distinguishing them from legitimate counterparts.

The importance of URLs in phishing detection is the main topic of this research, which also examines how machine learning techniques can be used to improve the precision and effectiveness of URL-based detection systems.By analyzing the structural, lexical, and contextual features of URLs, we endeavor to uncover patterns indicative of phishing attempts and empower individuals and organizations to preemptively safeguard against the pernicious effects of phishing attacks.

II. LITERATURE REVIEW

A. Introduction to Phishing Detection: Traditional Methods and Machine Learning Innovations
Phishing, which uses social engineering techniques to trick people into disclosing private information, is still a serious problem in cyberspace. Blacklists and other conventional detection techniques are inadequate for spotting recently created phishing URLs. Because of this shortcoming, researchers are investigating machine learning methods to improve phishing detection systems.
By leveraging URL features, these machine learning systems provide a promising alternative to traditional methods, offering more effective detection capabilities.

B. Optimizing Machine Learning Models to Improve the Detection of Phishing URLs

Recent studies emphasize the critical role of fine-tuning machine learning models through three main factors: data balancing, hyperparameter optimization, and feature selection. These investigations have shown notable improvements in accuracy across a range of machine learning models. Experimental evaluations using datasets from the UCI and Mendeley repositories reveal that while data balancing improves accuracy marginally, hyperparameter optimization and feature selection significantly enhance it. Combining all fine-tuning factors leads to superior performance, 97.7% accuracy has been attained by models such as the Gradient Boosting Classifier, CatBoost Classifier, and XGBoost Classifier. Additional models, including K-Nearest Neighbors (K-NN), Random Forest, Support Vector Machine (SVM), Decision Tree, and Multi-layer Perceptron (MLP), also exhibit high accuracy when appropriately fine-tuned.

C. Relative Evaluation and Case Studies Of Machine Learning Methods for Phishing Detection

The PHISH-SAFE system exemplifies the ability of ML algorithms in phishing detection. This system, which focuses on leveraging URL features for detection, was trained on a dataset comprising over 33,000 phishing and legitimate URLs using SVM and Naïve Bayes classifiers. PHISH-SAFE achieves over 90% accuracy, particularly notable with the SVM classifier. Additionally, studies that analyze various detection methods—including lexical features, host properties, and page importance properties—have yielded promising results, with accuracies reaching up to 98% using techniques like the Naïve Bayes Classifier. Comparative analyses of algorithms such as Decision Tree, Random Forest, and SVM, using metrics like accuracy rates, false efficiency prices, and false negative prices, further underscore the efficacy of fine-tuned machine learning approaches in phishing detection. Collectively, these studies highlight the significance of leveraging machine learning to mitigate phishing risks and enhance cybersecurity measures.

III. OBJECTIVE

The goal of the phishing URL detection project is to create an advanced system that uses machine learning techniques to efficiently recognize and categorize dangerous websites.

The main goal is to reduce the likelihood of becoming a victim of fraudulent activities by achieving high accuracy in differentiating between phishing efforts and legal URLs.

This means creating algorithms that can adjust to changing strategies employed by phishers while maintaining scalability to handle large volumes of URLs in real-time. Key aspects include feature selection and extraction to pinpoint indicators of phishing behavior, optimization for performance efficiency, and the creation of a user-friendly interface for seamless interaction.Thorough testing and verification procedures guarantee the system's dependability and efficiency in actual cybersecurity situations.

Moreover, the goal of the project is to promote integration with existing infrastructure and collaboration with industry stakeholders to bolster overall cybersecurity defenses against phishing threats.

IV. METHODOLOGY

Your phishing URL detection project. Here's a structured breakdown you can follow:

A. Data Collection:

Describe the sources from which phishing and legitimate URLs were collected.

Explain any preprocessing steps applied to clean and format the data.

Provide details on how the dataset methodology section for your phishing URL detection project:

B. Data Collection:

Specify the sources from which the phishing and legitimate URLs were collected, such as publicly available datasets, online repositories, or web scraping techniques. Detail any preprocessing steps applied to the raw data, including removing duplicates, standardizing URL formats, and filtering out irrelevant URLs. Describe the criteria used to label URLs as phishing or legitimate, whether it was based on known phishing databases, manual inspection, or automated classification algorithms.

C. Feature Extraction:

Provide a comprehensive list of characteristics for detecting phishing URLs, categorized into structural, lexical, and content-based features.

Explain the process of extracting each feature, including techniques like tokenization, n-gram analysis, domain analysis, etc.

Discuss any feature engineering efforts to enhance the discriminatory power of the features, such as normalization, scaling, or dimensionality reduction.

D. Model Selection and Training:

Present the criteria for choosing machine learning algorithms, considering factors like performance, interpretability, scalability, and computational efficiency.

Detail the training procedure for each selected model, including the parameter settings, optimization algorithms, and regularization techniques employed.

Discuss any ensemble methods or model stacking approaches used using several classifiers to achieve better results.

E. Evaluation Metrics:

Define the evaluation metrics utilized to evaluate the models' performance, explaining their relevance to the task of phishing URL detection.

Provide mathematical formulas or definitions for each metric, including accuracy, precision, recall, F1-score, ROC-AUC, etc.

Discuss the interpretation of these metrics taking into account the trade-offs between false positives and false negatives in the context of phishing detection.

F. Experimental Setup:

Specify the hardware and software environment used for conducting experiments, including CPU/GPU specifications, memory resources, and software dependencies.

Describe the programming languages, frameworks, and libraries used for feature extraction, data preprocessing, model training, and evaluation.

Provide reproducible code snippets or scripts to facilitate replication of the experiments by other researchers.

G. Validation and Testing:

Explain the process of model validation using techniques like k-fold cross-validation or holdout validation to assess generalization performance.

Describe the dataset division into training and validation, and testing sets, ensuring independence and randomness in the splits.

Present the results of model testing on the held-out testing set, including performance metrics and any qualitative analysis of misclassifications

V. RESULT

The completion of the phishing URL detection project marks an important turning point in the continuous fight against cyberthreats, especially phishing attempts. The initiative has produced a sophisticated system that uses machine learning to precisely identify and categorize dangerous URLs thanks to painstaking research and development activity. This achievement is underpinned by a multifaceted approach that encompasses feature selection and extraction, algorithm design, and rigorous evaluation methodologies.

A precisely calibrated machine learning model that can identify minute patterns and clues of phishing activity inside URLs is at the heart of the system. The algorithm demonstrates an impressive capacity to differentiate between authentic websites and phishing attempts by utilizing a wide variety of variables that are collected from URL structures, webpage content, and related metadata. This level of granularity is crucial in mitigating the ever-evolving tactics employed by malicious actors, who constantly strive to evade detection through sophisticated social engineering techniques and the creation of deceptive mockup websites.

Central to the success of the system is its adaptability to dynamic threat landscapes. By continuously monitoring and analyzing emerging phishing trends, the system can swiftly adapt its detection mechanisms to counter new attack

vectors and evasion tactics. This adaptability is facilitated by a robust feedback loop that integrates real-time threat intelligence data and user feedback, allowing the system to evolve and improve its detection capabilities over time.

The validation of the system's effectiveness is conducted through comprehensive experimentation and evaluation processes. These include benchmarking against large-scale datasets comprising both known phishing URLs and legitimate websites, as well as real-world testing in simulated phishing scenarios. Through rigorous performance metrics such as precision, recall, and F1 score, the system demonstrates its ability to achieve high levels of detection accuracy while minimizing false positives and false negatives.

The implications of these findings extend far beyond the confines of the research paper, offering tangible benefits to users and organizations across various sectors. By providing a robust defense against phishing attacks, the system enhances cybersecurity resilience, safeguarding sensitive information and mitigating the financial and reputational risks associated with data breaches. Furthermore, by contributing to the collective body of knowledge in cybersecurity, the research paper serves as a valuable resource for industry practitioners, policymakers, and researchers alike, driving innovation and informing future advancements in cyber defense strategies.

VI. CONCLUSION AND FUTURE SCOPE

This study presents a comprehensive investigation into the detection of phishing URLs leveraging machine learning techniques. Through meticulous data collection, feature engineering, and model selection, we have demonstrated the effectiveness of our methodology in accurately distinguishing phishing URLs from legitimate ones. Our experiments reveal promising results, showcasing the potential of machine learning models in enhancing cybersecurity measures against phishing attacks. Moving forward, there are several avenues for enhancing our phishing URL detection system. Firstly, incorporating more advanced machine learning algorithms, such as deep learning models like convolutional neural networks (CNNs) or recurrent neural networks (RNNs), could potentially improve the detection accuracy, especially for complex phishing URLs. Secondly, integrating real-time data sources and leveraging techniques like natural language processing (NLP) for analyzing textual content could enhance the model's ability to adapt to evolving phishing tactics.

Furthermore, exploring ensemble learning methods, such as stacking or boosting, could help in combining the strengths of multiple models and further improve detection performance. Additionally, extending the analysis to include features extracted from website behavior and user interactions could provide a more comprehensive understanding of phishing attempts.

REFERENCES

1. Gandotra, E., & Gupta, D. (2021). Improving spoofed website detection using machine learning. Cybernetics and Systems, 52(2), 169-190.
2. Harinahalli Lokesh, G., & BoreGowda, G. (2021). Phishing website detection based on effective machine learning approach. Journal of Cyber Security Technology, 5(1), 1-14.
3. Singh, C. (2020, March). Phishing website detection based on machine learning: A survey. In 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS) (pp. 398-404). IEEE.
4. Patil, V., Thakkar, P., Shah, C., Bhat, T., & Godse, S. P. (2018, August). Detection and prevention of phishing websites using machine learning approach. In 2018 Fourth international conference on computing communication control and automation (ICCUBEA) (pp. 1-5). Ieee.
5. Rasymas, T., & Dovydaitis, L. (2020). Detection of phishing URLs by using deep learning approach and multiple features combinations. Baltic journal of modern computing, 8(3), 471-483.
6. Alam, M. N., Sarma, D., Lima, F. F., Saha, I., & Hossain, S. (2020, August). Phishing attacks detection using machine learning approach. In 2020 third international conference on smart systems and inventive technology (ICSSIT) (pp. 1173-1179). IEEE.
7. https://www.researchgate.net/publication/328541785_Phishing_Website_Detection_using_Machine_Learning_Algorithms.
8. https://www.researchgate.net/publication/269032183_Detection_of_phishing_URLs_using_machine_learning_techniques.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com