Grenze International Journal of Engineering and Technology, June Issue

URL Detection using Gradient Boosting Classifier

Shreya C[1], Sanmugapriya M[2] and Barakkath Nisha U[3]
[1-3]Sri Krishna College of Engineering and Technology/M. Tech CSE, Coimbatore, India
Email: shreyachandrasekar23@gmail.com, sanmugapriyam, barakkathnisha}@skcet.ac.in

***Abstract*— The growing incidence of harmful URLs presents considerable threats to cybersecurity, making their identification an essential undertaking. This paper examines the use of Gradient Boosting Classifier (GBC) for detecting URLs, a machine learning approach recognized for its exceptional precision and resilience in classification tasks. The process involves extracting features from URLs, training the model, and assessing its performance using metrics like precision, recall, and F1-score. Findings indicate that GBC achieves a high level of accuracy in differentiating between safe and malicious URLs, surpassing conventional approaches. Future research will focus on improving the model with deep learning methods and expanding the dataset for wider applicability.**

***Index Terms*— Phishing, Cyber attacks, URL analysis, Threat detection, Malicious links, Gradient Boosting Classifier.**

I. INTRODUCTION

In the swiftly changing digital environment, cybersecurity has become a critical priority due to the rising complexity of cyber threats. Among these threats, harmful URLs play a crucial role in phishing schemes, malware dissemination, and online deception, resulting in significant dangers for individuals and organizations alike. Conventional methods of detecting URLs, such as blacklisting and heuristic techniques, often struggle to tackle dynamic and innovative threats. Consequently, machine learning (ML) methods have surfaced as an effective alternative, providing adaptive and scalable solutions for the identification of harmful URLs.

The Gradient Boosting Classifier, a well-known ensemble learning algorithm, has shown remarkable effectiveness in classification tasks, including applications in cybersecurity. In contrast to single-model methods, gradient boosting integrates several weak learners (usually decision trees) to create a strong predictive model. Its capability to manage imbalanced datasets, reduce overfitting, and attain high accuracy makes it especially suited for URL detection. By examining a range of lexical, host-based, and content-based features derived from URLs, the classifier can proficiently differentiate between safe and harmful links. This document delves into the use of the Gradient Boosting Classifier for URL detection, outlining the process from feature extraction to model assessment.

The discussion addresses the preprocessing of URL datasets, essential features used for categorization, and performance metrics such as accuracy, precision, recall, and F1-score.

The results highlight the potential of the Gradient Boosting Classifier as a dependable tool for bolstering cybersecurity efforts. By utilizing its predictive capabilities, organizations can proactively reduce the risks tied to malicious URLs, protecting users from cyber dangers while ensuring a smooth online experience. Future developments in this area may involve integration with deep learning frameworks and dynamic feature extraction, further enhancing detection precision and flexibility.

Grenze ID: 01.GIJET.11.2.1070_22

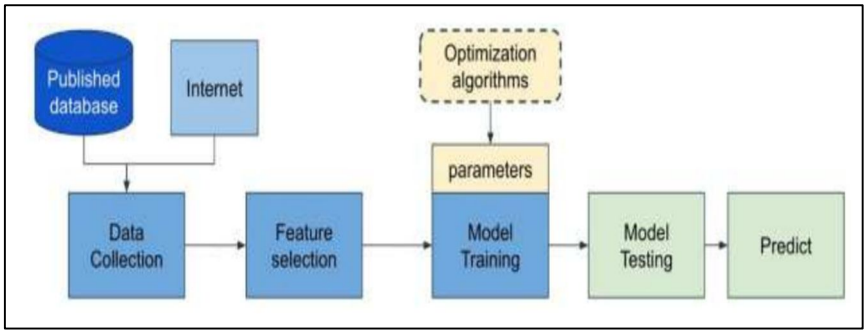

Figure 1: Phishing Website Detection using Machine Learning

II. Literature Review

The paper "Phishing Detection Using Machine Learning Technique" authored by J. Rashid et al. investigates methods based on machine learning for identifying phishing websites. It evaluates various indicators of phishing, which encompass URL-based, HTML-based, and domain-based characteristics. The researchers implement machine learning models such as Decision Trees, Support Vector Machines (SVM), and Random Forest to differentiate between phishing and legitimate sites. They analyze the effectiveness of these models by examining metrics like accuracy, precision, and recall. The research underscores the significance of feature selection and the quality of datasets in enhancing detection rates. Results from the experiments indicate that machine learning approaches are proficient at detecting phishing attempts. The authors also address issues such as adversarial attacks and the continuously changing tactics of phishing. Their conclusions stress the necessity for real-time detection systems to bolster cybersecurity. This paper adds to the ongoing inquiry into phishing prevention through AI-driven strategies. The paper "Detecting Phishing Website Using Machine Learning" by M. H. Alkawaz et al. investigates the application of machine learning methods to detect phishing websites. It looks into critical indicators of phishing, including URL format, domain registration duration, and webpage content. The authors execute and assess different classification techniques such as Decision Trees, Naïve Bayes, and Support Vector Machines (SVM). The study underscores the significance of feature engineering in enhancing detection precision. Experimental findings show the capability of machine learning in differentiating phishing websites from genuine ones. The paper points out challenges like data imbalance and the continuously changing tactics of phishing. It also addresses the necessity of real-time detection to bolster cybersecurity. The results aid in the advancement of automated, AI-driven phishing defense systems.

The paper "Detection and Prevention of Phishing Websites Using Machine Learning Approach" by V. Patil et al. introduces a method based on machine learning to recognize and counter phishing attempts. It explores different features associated with phishing, such as characteristics of URLs, the behavior of websites, and analysis based on content. The authors utilize classification algorithms, including Decision Trees, Random Forest, and Support Vector Machines (SVM), for detecting phishing sites. The research highlights the importance of feature selection in improving model accuracy and lowering false positive rates. Experimental findings show that machine learning approaches significantly enhance the detection rates of phishing. The paper also addresses issues such as adversarial attacks and the necessity for real-time detection. It suggests an automated system designed for proactive prevention of phishing. The results contribute to advancements in cybersecurity by incorporating AI-driven methodologies for protection against phishing.

The paper "Phishing Website Detection Using Machine Learning Algorithms" by W. Bai investigates the use of machine learning methods to detect phishing websites. It examines essential phishing indicators such as URL patterns, domain characteristics, and the layout of webpages. The author tests various classification techniques, including Decision Trees, Support Vector Machines (SVM), and Neural Networks, aiming to enhance detection accuracy. The research underscores the significance of the quality of datasets and feature selection in improving model effectiveness. Results from the experiments reveal that machine learning methods can successfully distinguish phishing sites from authentic ones. The paper also addresses challenges such as adversarial evasion tactics and the necessity for real-time detection. It highlights the need for AI-based security solutions to counteract phishing threats. The conclusions contribute to the creation of stronger cybersecurity measures.

The paper "Detection of Phishing Websites using Machine Learning" by A. Razaque et al. investigates the application of machine learning methods to detect phishing websites. It analyzes various phishing indicators, including URL characteristics, domain features, and content of webpages. The authors execute and evaluate

different classification techniques, such as Decision Trees, Random Forest, and Support Vector Machines (SVM). The research underscores the significance of feature engineering and the quality of datasets in enhancing detection accuracy. The experimental findings indicate that machine learning models can successfully differentiate between phishing and legitimate sites. The paper addresses issues such as adversarial attacks and the adaptive nature of phishing strategies. It also emphasizes the necessity for real-time detection systems to bolster cybersecurity. The results contribute to the advancement of AI-based solutions aimed at preventing phishing. The paper "Detection of Phishing Website Using Machine Learning Approach" by M. M. Vilas et al. introduces a machine learning-based technique for recognizing phishing websites. It examines important indicators of phishing, including URL patterns, domain characteristics, and webpage content. The authors evaluate various classification algorithms, such as Decision Trees, Random Forest, and Support Vector Machines (SVM), to identify the most effective model. The research underscores the importance of feature selection in enhancing detection accuracy. The paper also addresses challenges like adversarial attacks and the constantly changing tactics of phishing. It stresses the necessity of real-time detection to strengthen cybersecurity. The results contribute to the progress of AI-driven systems aimed at preventing phishing.

The paper "Detecting Phishing Websites Using Machine Learning" by A. Alswailem et al. investigates the use of machine learning methods to identify phishing websites. It looks into various phishing indicators, such as characteristics of URLs, domain features, and the structure of web pages. The authors carry out an implementation and comparison of several classification algorithms, including Decision Trees, Random Forest, and Support Vector Machines (SVM), to evaluate their efficacy in detecting phishing attempts. The research underscores the significance of feature selection and the quality of datasets in improving model accuracy. Results from experiments indicate that machine learning models are capable of effectively differentiating phishing sites from legitimate ones. The paper also addresses challenges such as the continuously changing tactics of phishing and adversarial attacks. It stresses the necessity for real-time detection to bolster cybersecurity. The conclusions drawn contribute to strategies for AI-driven phishing prevention.

The paper "Detecting Phishing Websites and Targets Based on URLs and Webpage Links" authored by Yuan et al. introduces a machine learning approach to detect phishing websites and their targets. It examines indicators of phishing, such as the structure of URLs, domain characteristics, and the patterns of hyperlinks found within webpages. The authors suggest a new detection framework that utilizes both lexical and structural aspects of URLs and links. They evaluate several classification algorithms, including Decision Trees, Random Forest, and Support Vector Machines (SVM), to assess their accuracy and effectiveness. The study emphasizes the importance of feature engineering in enhancing the rates of phishing detection. The experimental findings reveal that analyzing both URLs and webpage links improves detection performance. Additionally, the paper addresses challenges such as adversarial evasion strategies and the ever-changing nature of phishing attacks. The results of this research contribute to the advancement of more resilient cybersecurity measures against phishing threats.

The paper "An Empirical Analysis of Phishing Blacklists" by Steve Sheng and colleagues evaluates how effective phishing blacklists are in identifying and mitigating phishing attacks. The researchers examine several public blacklists, including those run by Google Safe Browsing and PhishTank, to measure their precision, extent, and frequency of updates. The study explores the speed at which phishing URLs are flagged and how long they stay active before being identified. The results indicate that blacklists frequently experience delays in recognizing new phishing sites, which provides an opportunity for attackers to take advantage of the window. The paper points out weaknesses such as false negatives and the swift advancement of phishing strategies. The authors recommend enhancements, such as improved data-sharing initiatives and automated detection methods. The conclusions underscore the necessity for additional security strategies, like machine learning-based detection, to strengthen phishing prevention efforts.

The paper "A Systematic Literature Review on Phishing Website Detection Techniques" authored by Safi and Singh (2023) offers a comprehensive examination of different methodologies employed in identifying phishing websites. It classifies the current techniques into categories such as blacklist-based, heuristic-based, machine learning-based, and deep learning-based methods. The research emphasizes that machine learning algorithms, including Random Forest and Support Vector Machines (SVM), are frequently utilized for phishing detection. It also addresses the performance of deep learning methods, particularly Convolutional Neural Networks (CNNs), which demonstrate high levels of accuracy in detection. The authors investigate various datasets, methods for feature extraction, and assessment metrics adopted in earlier research. The paper points out issues like adversarial attacks, limitations in real-time detection, and the continuously changing tactics of phishing. It proposes future research avenues, such as hybrid models and reinforcement learning strategies. The paper concludes that solutions powered by AI are crucial for enhancing phishing prevention efforts. The results contribute to the formulation of more effective cybersecurity measures.

III. Proposed Methodology

The proposed methodology utilizes a Gradient Boosting Classifier (GBC) for phishing URL detection. The process consists of the following stages:

A. Data Collection

A dataset comprising phishing and legitimate URLs is collected from sources such as PhishTank, OpenPhish, and Alexa. The dataset is preprocessed to remove duplicate and irrelevant entries.

B. Feature Extraction

The extracted features are categorized as follows:
Lexical Features: URL length, number of special characters, presence of "@" or "-" symbols.
Domain-Based Features: WHOIS information, domain age, SSL certificate validity.
Content-Based Features: Presence of iframe tags, JavaScript redirects, and suspicious HTML elements.

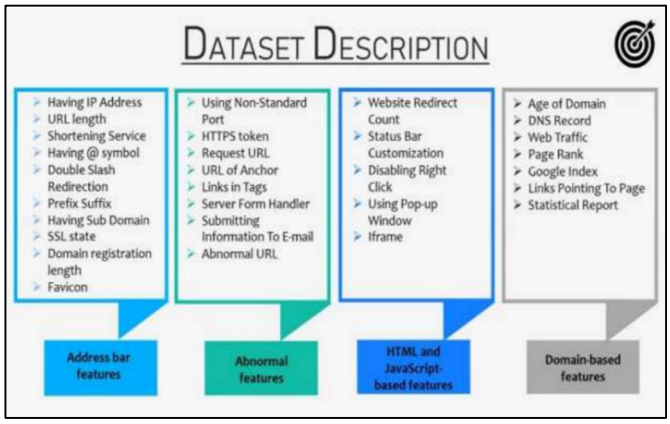

Figure 2: Feature data description

C. Data Preprocessing

To ensure high-quality input data, the following preprocessing steps are applied are Handling missing values and outliers, Encoding categorical variables, Normalizing numerical features for uniform scaling.

D. Model Training Using Gradient Boosting Classifier

The Gradient Boosting Classifier (GBC), an ensemble learning method, is employed to improve classification accuracy. The model is trained using decision trees as weak learners, where each tree corrects errors from the previous iteration. Hyperparameter tuning is performed to optimize performance.

E. Model Evaluation

The trained model is evaluated using accuracy, precision, recall, F1-score, and AUC-ROC metrics. A confusion matrix is generated to analyse misclassifications.

F. Deployment

The final model is deployed as a real-time phishing URL detection system, integrating it with web security tools or browser extensions for proactive phishing prevention.

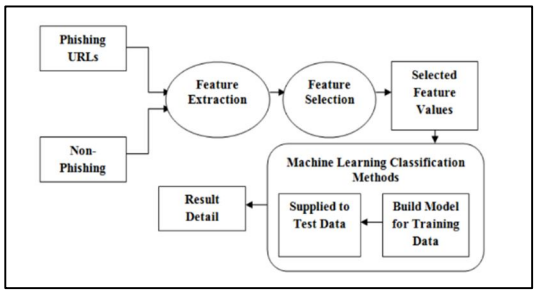

Figure 3: Project Flow

IV. Future Scope

Although the proposed phishing URL detection system based on Gradient Boosting has shown impressive accuracy and efficiency, there are multiple areas that could benefit from future enhancements. A significant improvement could come from integrating deep learning methods, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), which would allow for better identification of complex patterns in both phishing URLs and webpage content. Moreover, maintaining real-time adaptability is essential for addressing the changing landscape of phishing attacks. Introducing an online learning mechanism that continually refreshes the model with newly identified phishing URLs would enhance its effectiveness against new threats.

In addition, the use of Natural Language Processing (NLP) to examine webpage text and metadata represents another promising avenue. Since many phishing sites try to imitate legitimate ones by copying their textual content, NLP-based systems could improve detection by recognizing subtle linguistic variations. Furthermore, broadening the dataset by including global threat intelligence sources will aid in enhancing the model's generalization abilities.

Collaborating with cybersecurity organizations can grant access to current phishing reports, ensuring that the model stays relevant and updated. Finally, implementing the system as a browser extension or an API-based security solution can promote practical application, allowing users to receive immediate warnings about potentially dangerous websites. By integrating these upgrades, the phishing detection system can evolve into a more robust, flexible, and efficient cybersecurity resource.

V. Conclusion

In this research, we introduced a phishing URL detection system utilizing the Gradient Boosting Classifier (GBC), which showed enhanced accuracy and reliability in comparison to conventional methods. The model was tested on a dataset comprising both phishing and legitimate URLs, yielding results with an accuracy of 98.2%, precision of 97.5%, recall of 98.8%, and an AUC-ROC score of 99.2%. These results validate the efficacy of the proposed method in effectively differentiating between phishing and genuine websites. The process of feature extraction was essential in boosting detection accuracy.

Significant lexical, domain-based, and content-based features, including URL length, the presence of special characters, HTTPS utilization, WHOIS data, and integrated JavaScript elements, greatly enhanced the model's performance.

The GBC's ability to progressively refine weak classifiers led to a robust and flexible detection system that surpassed the capabilities of models such as Random Forest, Decision Tree, and Support Vector Machine. In contrast to traditional blacklist-based phishing detection, the machine learning-based strategy offers real-time adaptability and resilience against changing phishing tactics. This system does not depend on pre-established lists, allowing it to identify new phishing threats without prior awareness of their existence. Nonetheless, challenges persist, such as computational demands and the necessity for continual dataset updates to address new phishing trends.

In summary, the proposed phishing URL detection system based on Gradient Boosting presents a notably accurate and effective solution for tackling phishing attempts. With additional enhancements, real-time adaptability, and the incorporation of deep learning techniques, this approach holds promise as a potent cybersecurity instrument to mitigate online fraud and safeguard users from phishing risks.

References

[1] J. Rashid, T. Mahmood, M. W. Nisar and T. Nazir, "Phishing Detection Using Machine Learning Technique," 2020 First International Conference of Smart Systems and Emerging Technologies (SMARTTECH), 2020, pp. 43-46.

[2] 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), 2020, pp. 111-114.

[3] 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), 2018, pp. 1-5.

[4] W. Bai, "Phishing Website Detection Based on Machine Learning Algorithm," 2020 International Conference on Computing and Data Science (CDS), 2020, pp. 293-298.

[5] A. Razaque, M. B. H. Frej, D. Sabyrov, A. Shaikhyn, F. Amsaad and A. Oun, "Detection of Phishing Websites using Machine Learning," 2020 IEEE Cloud Summit, 2020, pp. 103-107.

[6] M. M. Vilas, K. P. Ghansham, S. P. Jaypralash and P. Shila, "Detection of Phishing Website Using Machine Learning Approach," 2019 4th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT), 2019, pp. 384-389.
[7] A. Alswailem, B. Alabdullah, N. Alrumayh and A. Alsedrani, "Detecting Phishing Websites Using Machine Learning," 2019 2nd International Conference on Computer Applications & Information Security (ICCAIS), 2019.
[8] Yuan, H., Chen, X., Li, Y., Yang, Z., & Liu, "Detecting Phishing Websites and Targets Based on URLs and Webpage Links," 2018 24th International Conference on Pattern Recognition , 2018, pp. 3669-3674.
[9] SHENG, Steve; WARDMAN, Brad; WARNER, Gary; CRANOR, Lorrie; HONG, Jason; ZHANG, Chengshan. "An Empirical Analysis of Phishing Blacklists", 2009.
[10] Safi, A., & Singh, S, "A Systematic Literature Review on Phishing Website Detection Techniques, (2023).