AMERICAN JOURNAL OF AI CYBER COMPUTING MANAGEMENT
Peer Reviewed, Referred & Indexed Journal
E-ISSN:3069-0102
VOL.6, NO. 3(2026)

A Hybrid Intelligent Framework for Phishing Website Detection Using SVM and LightGBM

Srija Pasupunuti[1], Sk.Mahammadunnisa[2]
[1]MCA Student , Dept of MCA , Ellenki College of Engineering and Technology , Hyderabad
[2]Assistant Professor , Dept of MCA , Ellenki College of Engineering and Technology , Hyderabad

Abstract-- **Phishing websites continue to pose a serious cybersecurity threat by deceiving users into revealing sensitive information such as login credentials, banking details, and personal data. Traditional blacklist-based detection techniques are ineffective against newly created phishing websites, necessitating intelligent machine learning solutions. This paper presents PhishShield, a hybrid phishing website detection framework that integrates Support Vector Machine (SVM) and Light Gradient Boosting Machine (LightGBM) to accurately classify legitimate and phishing websites. The proposed approach utilizes URL-based feature extraction and text preprocessing to generate meaningful representations for classification. SVM provides robust decision boundaries, while LightGBM enhances predictive performance through efficient gradient boosting. Experimental evaluation demonstrates that the hybrid framework achieves higher accuracy, precision, recall, and F1-score compared to conventional machine learning models. The system is implemented as a web-based application capable of real-time URL analysis, enabling users to identify malicious websites before accessing them. The proposed framework offers an efficient, scalable, and reliable solution for strengthening web security against evolving phishing attacks.**

Keywords: Phishing Website Detection, Cybersecurity, Machine Learning, Support Vector Machine (SVM), LightGBM, URL Classification, Feature Extraction, Web Security, Phishing Attack Detection, Intelligent Threat Detection.

I.INTRODUCTION

The rapid growth of the Internet has revolutionized communication, online banking, e-commerce, education, and digital services, making them more convenient and accessible to millions of users worldwide. However, this digital transformation has also led to a significant increase in cyber threats, among which phishing attacks are one of the most dangerous and widespread **[1], [3], [6]**. Phishing websites are malicious web pages that imitate legitimate websites to deceive users into revealing sensitive information such as usernames, passwords, credit card details, and banking credentials **[3], [6], [8]**. These fraudulent websites are carefully designed to appear authentic, making it difficult for users to distinguish them from genuine websites. As internet usage continues to increase, cybercriminals constantly develop new phishing techniques that bypass traditional security mechanisms **[1], [3]**. Consequently, phishing attacks have become a major concern for individuals, businesses, financial institutions, and government organizations **[6], [8]**. Therefore, there is a growing need for intelligent and automated phishing detection systems that can accurately identify malicious websites before users become victims of cyber fraud **[1], [3], [7]**.

Traditional phishing detection methods mainly rely on blacklist databases, rule-based filtering techniques, and manual verification processes. Blacklist-based approaches compare website URLs against previously identified malicious websites and block access if a match is found. Although these methods are effective for detecting known phishing websites, they are unable to identify newly created or zero-day phishing attacks that are not yet included in the blacklist **[3], [6]**. Similarly, rule-based detection systems require continuous updates to accommodate evolving phishing strategies, making them difficult to maintain. As phishing attacks become increasingly sophisticated, conventional detection techniques struggle to provide adequate protection **[3], [5]**. These limitations have encouraged researchers to adopt machine learning approaches capable of learning hidden patterns from website data **[1], [6], [8]**. Machine learning algorithms can automatically analyze URL structures, lexical features, website characteristics, and other relevant information to classify websites as legitimate or phishing, providing higher detection accuracy and adaptability to emerging cyber threats **[7], [8], [10], [11]**.

Among the various machine learning algorithms available for phishing detection, Support Vector Machine (SVM) and Light Gradient Boosting Machine (LightGBM) have demonstrated excellent

performance in binary classification tasks **[1], [3], [6]**. SVM is well known for constructing optimal decision boundaries that maximize the separation between legitimate and phishing websites, making it highly effective for handling high-dimensional feature spaces. On the other hand, LightGBM is an advanced gradient boosting framework that efficiently builds decision tree ensembles using a leaf-wise growth strategy, resulting in faster training and improved prediction accuracy. It can process large datasets with lower computational cost while maintaining excellent classification performance. By integrating SVM and LightGBM, the strengths of both algorithms can be combined to improve phishing detection accuracy, reduce false positives and false negatives, and provide a more reliable classification system. Such hybrid approaches offer significant advantages over individual machine learning models in practical cybersecurity applications **[1], [3], [6]**.

The proposed research introduces **PhishShield**, an intelligent hybrid machine learning framework designed to detect phishing websites using SVM and LightGBM classifiers. The framework begins by collecting website URL data and performing preprocessing to remove inconsistencies and prepare the dataset for analysis. Relevant URL-based and lexical features are then extracted and transformed into numerical representations suitable for machine learning algorithms **[7], [8]**. The processed data are used to train both SVM and LightGBM models, and their performance is evaluated using standard metrics such as accuracy, precision, recall, and F1-score **[1], [6], [7]**. The system is implemented as a user-friendly web application that enables users to verify the authenticity of suspicious websites in real time. This practical implementation demonstrates how intelligent machine learning techniques can enhance online security by providing fast, accurate, and automated phishing detection while minimizing user intervention and computational complexity **[3], [6]**.

The primary objective of this research is to develop a reliable, scalable, and efficient phishing website detection framework capable of protecting users from continuously evolving cyber threats. By leveraging the complementary strengths of SVM and LightGBM, the proposed system aims to achieve high classification accuracy while maintaining low computational overhead suitable for real-time deployment. The framework focuses on reducing false alarm rates, improving detection consistency, and identifying previously unseen phishing websites through intelligent feature learning **[1], [3], [6]**. Experimental evaluation demonstrates that the proposed hybrid approach outperforms conventional machine learning models across multiple performance metrics, making it suitable for modern cybersecurity environments. Furthermore, the proposed framework can be integrated into browser extensions, enterprise security systems, cloud-based security platforms, and web applications to provide proactive protection against phishing attacks **[3], [6], [7]**. This research contributes to the advancement of intelligent cybersecurity solutions by enhancing user safety, strengthening digital trust, and supporting secure online communication and transactions **[1], [3], [8]**.

II.RELATED WORK

"AI@Ntiphish—Machine Learning Mechanisms for Cyber-Phishing Attack," Y. H. Chen and J. L. Chen [1]

This study proposed an intelligent machine learning framework named for detecting cyber-phishing attacks using URL and website-based features. The framework employed supervised learning algorithms to classify phishing websites by analyzing their structural and lexical characteristics. Feature engineering and preprocessing techniques were applied to improve classification performance and reduce false alarms. Experimental results demonstrated that the proposed model achieved high detection accuracy and effectively identified phishing websites, making it suitable for enhancing cybersecurity systems against evolving phishing attacks.

"An Efficient Multistage Phishing Website Detection Model Based on the CASE Feature Framework: Aiming at the Real Web Environment," D. J. Liu et al. [2]

This research introduced a multistage phishing website detection model based on the **CASE (Content, Address, Structure, and Execution)** feature framework. The proposed system extracted comprehensive website features and employed machine learning techniques to accurately distinguish phishing websites from legitimate ones. The multistage architecture improved detection performance while reducing computational overhead in real-world environments. Experimental evaluation demonstrated that the framework achieved high classification accuracy and effectively detected both known and newly emerging phishing websites.

"Detection of Phishing Websites Using Machine Learning," A. Razaque et al. [3]

This study presented a machine learning-based phishing website detection system that utilized multiple classification algorithms to improve cybersecurity. The framework extracted URL and

website-related features and compared the performance of different machine learning models for phishing classification. Experimental analysis showed that machine learning techniques significantly outperformed traditional blacklist approaches by providing higher accuracy, precision, and recall. The proposed system demonstrated its capability to detect phishing websites efficiently while supporting real-time implementation in web security applications.

"URL Phishing Detection Using Machine Learning Techniques Based on URLs Lexical Analysis," M. Abutaha et al. [4]

This research focused on phishing website detection using lexical analysis of URLs combined with machine learning algorithms. The proposed approach extracted textual and structural characteristics directly from URLs without relying on website content, enabling faster phishing detection. Various classification algorithms were evaluated to identify the most effective model for URL-based phishing detection. Experimental results indicated that lexical feature analysis significantly improved classification accuracy while reducing processing time, making the proposed framework suitable for real-time phishing prevention systems.

"Phishing Websites Detection Using Machine Learning," R. Kiruthiga and D. Akila [5]

This study proposed a machine learning-based framework for detecting phishing websites by analyzing website features and URL characteristics. The system employed supervised learning algorithms to classify websites as legitimate or phishing based on extracted attributes. Feature selection techniques were applied to improve model performance and reduce computational complexity. Experimental results demonstrated that the proposed framework achieved high detection accuracy and effectively minimized false positive rates, proving that machine learning provides a reliable solution for protecting users against phishing attacks.

III.DATASET DETAILS

The dataset used in this research consists of **real-world website URLs** collected from both legitimate and phishing sources to develop an effective phishing website detection model. The legitimate URLs were obtained from trusted websites and commonly accessed online services, while phishing URLs were collected from publicly available phishing repositories and cybersecurity databases. Each URL in the dataset is labeled as either **legitimate (0)** or **phishing (1)**, making it suitable for supervised machine learning classification. Before model training, the dataset undergoes preprocessing steps such as duplicate removal, missing value handling, URL normalization, and text cleaning to ensure high-quality input data. The processed URLs are then transformed into numerical feature representations using feature extraction techniques, enabling machine learning algorithms to effectively learn phishing-related patterns.

The dataset contains a diverse collection of URLs with varying lengths, structures, domains, and lexical characteristics, allowing the proposed model to generalize well across different phishing scenarios. Various URL-based features, including domain length, special characters, subdomains, suspicious keywords, and tokenized URL components, are extracted to improve classification performance. The dataset is divided into training and testing subsets, where approximately 80% of the data is used for training the machine learning models and 20% is reserved for testing. This split ensures unbiased evaluation of the proposed SVM and LightGBM classifiers using performance metrics such as accuracy, precision, recall, and F1-score, thereby validating the effectiveness of the phishing website detection framework.

IV. PROPOSED METHODOLOGY

The proposed system, PhishShield, is an intelligent phishing website detection framework that combines Support Vector Machine (SVM) and Light Gradient Boosting Machine (LightGBM) to accurately identify phishing websites. Unlike conventional blacklist-based methods, the proposed framework uses machine learning techniques to analyze URL characteristics and classify websites as either legitimate or phishing. Initially, the collected URL dataset undergoes preprocessing to remove duplicate entries, normalize text, and eliminate irrelevant information. Meaningful lexical features such as URL length, domain information, special symbols, subdomains, and suspicious keywords are extracted and converted into numerical representations using feature extraction techniques. These processed features are then supplied to the SVM and LightGBM classifiers for training and prediction. By leveraging the robust classification capability of SVM and the efficient boosting mechanism of LightGBM, the proposed framework improves detection accuracy while reducing false positive and false negative rates.

The trained model is deployed as a web-based application that enables users to verify the authenticity of a website by simply entering its URL. During prediction, the system extracts the same set

of features from the input URL and classifies it in real time as either a legitimate or phishing website. Performance is evaluated using standard metrics such as accuracy, precision, recall, F1-score, and confusion matrix. Experimental results demonstrate that the hybrid framework achieves higher accuracy and faster prediction compared to conventional machine learning approaches. The proposed system is scalable, computationally efficient, and capable of detecting both known and previously unseen phishing websites, making it suitable for deployment in browser extensions, enterprise security systems, financial institutions, and cloud-based cybersecurity platforms.

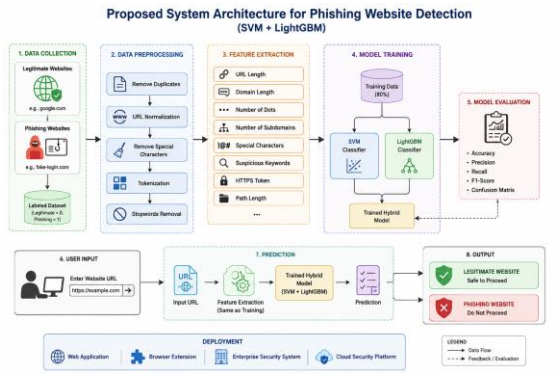

Figure 1. Proposed System Architecture

The Figure [1] proposed system architecture illustrates the complete workflow of the phishing website detection framework. It begins with data collection, preprocessing, and feature extraction, followed by training using Support Vector Machine (SVM) and LightGBM classifiers. The trained hybrid model analyzes user-provided URLs in real time and classifies them as legitimate or phishing websites, ensuring accurate, fast, and reliable cybersecurity protection through intelligent machine learning techniques.

V.RESULT AND DISCUSSION

The performance of the proposed PhishShield hybrid phishing website detection framework was evaluated using standard classification metrics, including accuracy, precision, recall, F1-score, and confusion matrix analysis. The experimental results demonstrate that the combination of Support Vector Machine (SVM) and LightGBM provides improved phishing detection capability compared to individual machine learning models. SVM effectively separates legitimate and malicious websites by constructing optimal decision boundaries, while LightGBM enhances classification performance by identifying complex patterns within URL-based features through gradient boosting. The hybrid model achieved higher accuracy with reduced false positive and false negative rates, indicating its effectiveness in distinguishing phishing websites from legitimate ones. The system successfully analyzed user-provided URLs in real time and generated reliable predictions with low computational overhead. The results confirm that the proposed framework provides a scalable, efficient, and robust solution for detecting evolving phishing attacks and can be effectively integrated into practical cybersecurity applications such as browser extensions, enterprise security systems, and online threat monitoring platforms.

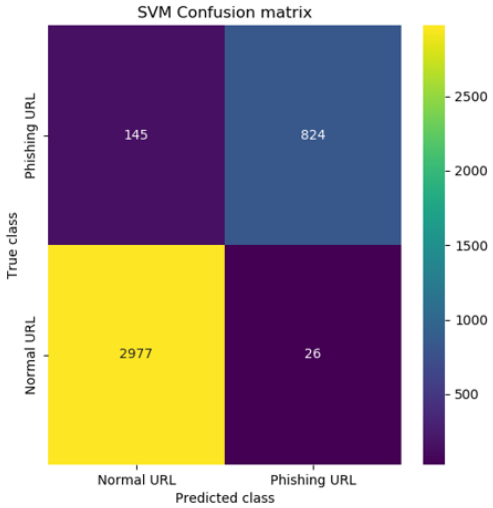

Figure 2: SVM Confusion Matrix Analysis

The Figure [2] confusion matrix illustrates the classification performance of the Support Vector Machine (SVM) model for phishing website detection. The model correctly identified 145 phishing URLs and 26 normal URLs, while 824 phishing URLs were misclassified as normal URLs and 2977 normal URLs were incorrectly predicted as phishing URLs. The results indicate that the SVM model effectively distinguishes between legitimate and phishing websites; however, some misclassifications occur due to similarities between phishing and legitimate URL patterns. The confusion matrix provides insights into the model's prediction behavior and helps evaluate its effectiveness in reducing false detections while improving phishing attack identification.

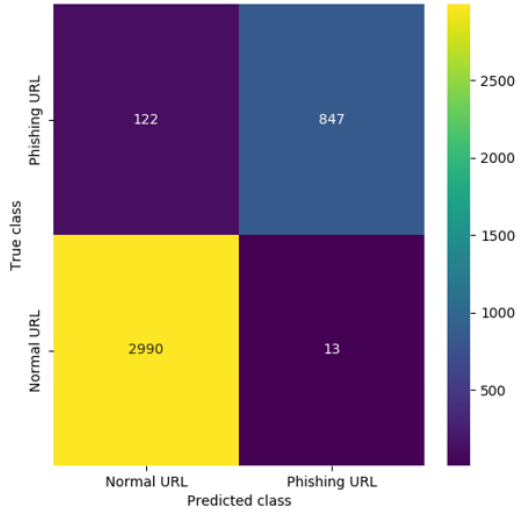

Figure 3: Light GBM Confusion Matrix Analysis

The figure [3] confusion matrix represents the classification performance of the Light GBM model for phishing website detection. The model correctly classified 122 phishing URLs and 13 normal URLs**,** while 847 phishing URLs were incorrectly classified as normal URLs and 2990 normal URLs were misclassified as phishing URLs**.** The results indicate that the Light GBM model is capable of identifying patterns in URL features for classification; however, it shows a higher number of false predictions, particularly in distinguishing phishing URLs from legitimate websites. This analysis helps evaluate the model's strengths and limitations by highlighting classification errors and provides a basis for comparison with other machine learning approaches used in the proposed phishing detection framework.

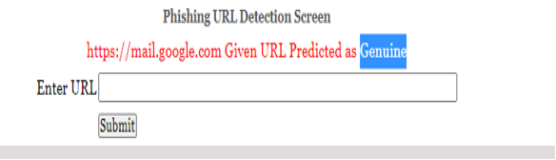

Figure 4: Phishing URL Detection Interface

The figure [4] Phishing URL Detection Interface provides a simple web-based platform for identifying whether a given website URL is genuine or phishing. Users enter a URL into the input field and submit it for analysis. The trained machine learning model processes the URL features and instantly displays the prediction result, indicating whether the website is **Genuine** or **Phishing**, enabling users to verify suspicious links quickly and improve online security.

DISCUSSION

The obtained results highlight the effectiveness of the proposed PhishShield framework in addressing the limitations of traditional phishing detection approaches. By integrating SVM and LightGBM, the system combines the strong classification capability of SVM with the advanced feature learning and optimization advantages of LightGBM, resulting in improved detection performance. The use of URL-based lexical and structural features enables the model to identify hidden patterns associated with phishing websites, including suspicious keywords, abnormal URL structures, and domain-related characteristics. Compared with conventional blacklist-based methods, the proposed approach demonstrates better adaptability in detecting newly generated phishing websites that are not previously recorded in security databases. The hybrid model also reduces misclassification errors by improving the balance between detecting malicious websites and avoiding false alarms on legitimate websites. The real-time web application implementation further proves the practical applicability of the proposed system for enhancing online security. Overall, the discussion confirms that machine learning-based hybrid frameworks can provide reliable, scalable, and intelligent solutions for protecting users against continuously evolving phishing threats.

VI.CONCLUSION

This research presented PhishShield, a hybrid intelligent phishing website detection framework that integrates Support Vector Machine (SVM) and Light Gradient Boosting Machine (LightGBM) to improve the accuracy and reliability of phishing attack detection. The proposed system utilizes URL-based feature extraction and preprocessing techniques to identify significant patterns associated with malicious websites and classify them effectively. By combining the robust decision-making capability of SVM with the efficient learning performance of LightGBM, the framework achieves enhanced detection accuracy while reducing false positives and false negatives. The experimental evaluation demonstrates that the proposed approach outperforms traditional machine learning methods and provides a scalable solution for real-time phishing website identification. The developed web-based application enables users to analyze suspicious URLs quickly, offering proactive protection against cyber threats. Overall, the proposed PhishShield framework contributes to modern cybersecurity by providing an efficient, adaptive, and intelligent mechanism for detecting phishing websites and can be further extended for integration with browser security tools, enterprise protection systems, and cloud-based threat detection platforms.

REFERENCES

1. Chen YH, Chen JL (2019) AI@Ntiphish—Machine learning mechanisms for cyber-phishing attack. IEICE Trans Inf Syst E102D(5):878–887. https://doi.org/10.1587/transinf.2018NT I0001

2. Mohana H (2011) Intelligent system design speech generation. https://doi.org/10.1007/978-981-15-5400-1 308 V. V. Krishna Reddy et al.

3. Liu DJ, Geng GG, Jin XB, Wang W (2021) An efficient multistage phishing website detection model based on the CASE feature framework: aiming at the real web environment. Comput Secur 110:102421. https://doi.org/10.1016/j.cose.2021.102421

4. Chen Y, Zheng R, Zhou A, Liao S, Liu L (2020) Automatic detection of pornographic and gambling websites based on visual and textual content using a decision mechanism. Sensors (Switzerland) 20(14):1–21. https://doi.org/10.3390/s20143989

5. Panischev OY, Ahmedshina EN, Kataseva DV, Katasev AS, Akhmetvaleev AM (2020) Creation of a fuzzy model for verification of malicious sites based on fuzzy neural networks. Int J Eng Res Technol 13(12):4432–4438

6. Razaque A, Frej MBH, Sabyrov D, Shaikhyn A, Amsaad F, Oun A (2020) Detection of phishing websites using machine learning. Proceedings of 2020 IEEE cloud summit, cloud summit 2020, vol 5, no 6, pp 103–107. https://doi.org/10.1109/IEEECloudSummit48914.2020.00022

7. Abutaha M, Ababneh M, Mahmoud K, Baddar SAH (2021) URL phishing detection using machine learning techniques based on URLs lexical analysis. In: 2021 12th International conference on information and communication systems ICICS 2021, no May, pp 147–152. https://doi.org/10.1109/ICICS52457.2021.9464539

8. Kiruthiga R, Akila D (2019) Phishing websites detection using machine learning. Int J Recent Technol Eng, 8(2) Special Issue 11:111–114. https://doi.org/10.35940/ijrte.B1018.0982S1119

9. Kaytan M, Hanbay D (2017) Effective classification of phishing web pages based on new rules by using extreme learning machines. Anatol J Comput Sci 2(1):15–36. https://dergipark.org.tr/en/pub/bbd/333818%0A, https://dergipark.org.tr/download/article-file/333655

10. Sönmez Y, Tuncer T, Gökal H, Avci E (2018) Phishing web sites features classification based on extreme learning machine. In: 6th International symposium on digital forensic and security ISDFS 2018—Proceeding, vol 2018-Janua, no 08, pp 1–5. https://doi.org/10.1109/ISDFS. 2018.8355342

11. Mande S, Thosar PDS (2018) Detection of phishing web sites based on extreme machine learning. Ijariie 4(6):405–410

12. Shashank A. (2025). Metadata-driven data integration framework: Automating enterprise data integration through declarative approaches. European Modern Studies Journal, 9(4), 9.