International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 05 (May) 2026
Received: 10-3-2026 Accepted: 21-4-2026 Published: 07-5-2026

HYBRID SVM AND LIGHTGBM FRAMEWORK FOR PHISHING URL DETECTION

[1] B. RAMA DEEPTHI, [2] NAGIREDDY SRI LALITHA, [3] SYED NAZIYA, [4] BORRA NANDINI, [5] POTTI NAGA LAKSHMI KUMARI, [6] GONTLA SAI BHARGAVI PUSHPALATHA

ASST., PROESSOR, DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING, KRISHNA CHAITANYA INSTITUTE OF TECHNOLOGY & SCIENCES, DEVARAJUGATTU, MARKAPUR

[2,3,4,5,6] STUDENT, DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING, KRISHNA CHAITANYA INSTITUTE OF TECHNOLOGY & SCIENCES, DEVARAJUGATTU, MARKAPUR.

ABSTRACT

Phishing websites pose a significant threat to online users by attempting to steal sensitive information such as login credentials, banking details, and personal data through deceptive techniques. With the rapid growth of digital services and e-commerce platforms, detecting and preventing phishing attacks has become a critical cybersecurity challenge. Traditional rule-based detection systems often fail to adapt to evolving phishing strategies, making machine learning approaches more effective for dynamic threat detection.

This project presents a robust phishing website detection system using advanced machine learning algorithms, specifically Support Vector Machine (SVM) and Light Gradient Boosting Machine (LightGBM). The proposed system analyzes various features extracted from URLs and webpage content, such as domain age, URL length, presence of special characters, HTTPS usage, redirection behavior, and abnormal request patterns. These features are preprocessed and fed into the models to classify websites as either legitimate or phishing.

Keywords: Phishing Detection, Cybersecurity, Machine Learning, Support Vector Machine (SVM), LightGBM, URL Analysis, Feature Extraction, Classification, Data Mining, Fraud Detection, Web Security, Malicious Websites, Ensemble Learning, Predictive Modeling

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 05 (May) 2026
Received: 10-3-2026 Accepted: 21-4-2026 Published: 07-5-2026

I. INTRODUCTION

With the rapid expansion of the internet and digital technologies, online services such as banking, e-commerce, social media, and cloud platforms have become an integral part of daily life. However, this growth has also led to a significant rise in cyber threats, among which phishing attacks are one of the most common and dangerous. Phishing websites are malicious web pages designed to imitate legitimate websites in order to deceive users into revealing sensitive information such as usernames, passwords, credit card details, and personal data. These attacks not only result in financial losses but also compromise user privacy and organizational security.

Traditional approaches to phishing detection, such as blacklist-based and rule-based systems, have several limitations. Blacklist methods rely on previously reported malicious URLs, making them ineffective against newly created phishing websites. Similarly, rule-based systems depend on predefined patterns, which can be easily bypassed by attackers using advanced obfuscation techniques. As phishing tactics continue to evolve rapidly, there is a growing need for intelligent and adaptive detection mechanisms that can identify previously unseen threats.

In recent years, machine learning techniques have emerged as powerful tools for detecting phishing websites. These methods analyze various characteristics of websites, including URL structure, domain-related features, HTML content, and user behavior patterns, to classify them as legitimate or malicious.

II. LITERATURE REVIEW

Phishing website detection has been an active area of research in the field of **Cybersecurity** and **Machine Learning**, with various techniques proposed to improve accuracy and adaptability. Early studies primarily relied on blacklist-based approaches, where known phishing URLs were stored in databases. However, these methods were limited in detecting newly generated phishing sites, leading researchers to explore more intelligent solutions.

A study by **Abdelhamid et al. (2014)** introduced a phishing detection model using rule-based classification combined with machine learning techniques. The system utilized URL-based and HTML-based features to identify malicious websites. Although the approach showed improved detection rates, it struggled with scalability and adaptability to new phishing patterns.

Ma et al. (2009) proposed one of the foundational works in phishing detection using machine learning algorithms such as Logistic Regression and Support Vector Machines. Their research focused on lexical features of

2250-3676 ijesatj@gmail.com https://www.ijesat.com/

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 05 (May) 2026
Received: 10-3-2026 Accepted: 21-4-2026 Published: 07-5-2026

URLs and demonstrated that SVM could effectively classify phishing websites with high accuracy. However, the model required extensive feature engineering and was computationally expensive for large datasets.

EXISTING SYSTEM

The existing systems for phishing website detection primarily rely on traditional approaches such as blacklist-based detection, heuristic methods, and rule-based techniques. These systems were developed to identify malicious websites by comparing URLs against known phishing databases or by applying predefined rules to detect suspicious patterns. While these methods have been effective to some extent, they suffer from several limitations when dealing with modern and rapidly evolving phishing attacks.

One of the most commonly used approaches is the blacklist-based system, where URLs of known phishing websites are stored in a centralized database. When a user attempts to access a website, the system checks whether the URL exists in the blacklist. If found, the website is blocked. Although this method is simple and fast, it is highly ineffective against newly created (zero-day) phishing websites that are not yet listed in the database. Attackers can easily generate new URLs, making blacklist systems outdated quickly.

Another approach is the heuristic-based detection system, which uses predefined rules and patterns such as URL length, presence of special characters, abnormal domain names, and suspicious page behaviors. These systems attempt to detect phishing websites based on known characteristics. However, cyber attackers continuously modify their techniques to bypass these rules, reducing the effectiveness of heuristic methods over time.

PROPOSED SYSTEM

The proposed system aims to develop an intelligent and efficient phishing website detection framework using advanced machine learning techniques, specifically **Support Vector Machine (SVM)** and **Light Gradient Boosting Machine (LightGBM)**. Unlike traditional systems, this approach is designed to identify both known and unknown phishing websites by learning patterns from data rather than relying solely on predefined rules or blacklists.

The system begins with **data collection**, where a dataset consisting of legitimate and phishing website URLs is gathered from reliable sources. These URLs are then processed through a **feature extraction module**, which derives important characteristics such as URL length, domain

2250-3676 ijesatj@gmail.com https://www.ijesat.com/

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 05 (May) 2026
Received: 10-3-2026 Accepted: 21-4-2026 Published: 07-5-2026

age, use of HTTPS, presence of special characters, number of subdomains, redirection behavior, and abnormal request patterns. These features play a crucial role in distinguishing phishing websites from legitimate ones.

After feature extraction, the data undergoes **preprocessing**, including handling missing values, normalization, and encoding of categorical variables. The processed data is then split into training and testing sets for model development and evaluation.

METHODOLOGY

The proposed phishing website detection system follows a structured machine learning methodology to accurately classify websites as legitimate or malicious. The process begins with data collection, where a labeled dataset of phishing and legitimate URLs is obtained from publicly available sources such as security repositories and phishing databases. This dataset forms the foundation for training and evaluating the models.

Next, a feature extraction phase is performed to derive meaningful attributes from the URLs and associated webpage data. These features include lexical properties (such as URL length, presence of special characters, number of subdomains), host-based features (domain age, DNS records), and content-based indicators (use of HTTPS, redirection behavior, iframe usage). These extracted features are crucial for identifying patterns commonly associated with phishing websites.

VI. SYSTEM MODEL

System Architecture

III. RESULTS AND DISCUSSIONS

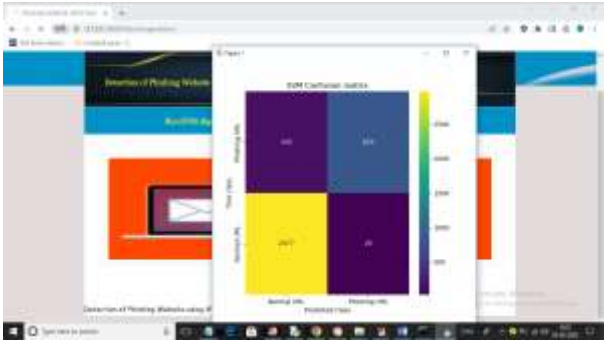
In above screen click on 'Admin Login Here' link to get below login screen

In above screen enter username and password as 'admin' and 'admin' and then press button to get below output

2250-3676 ijesatj@gmail.com https://www.ijesat.com/

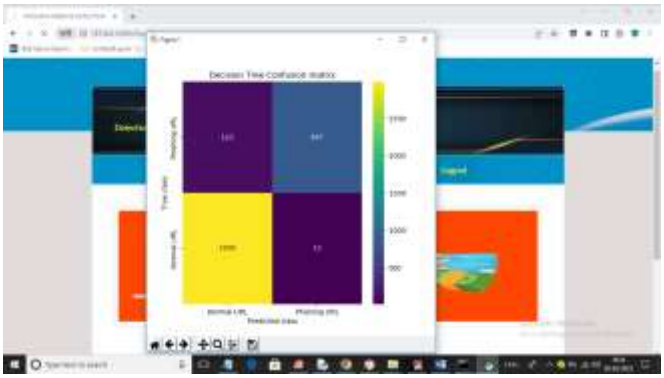
International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 05 (May) 2026
Received: 10-3-2026 Accepted: 21-4-2026 Published: 07-5-2026

In above screen click on 'Run SVM Algorithm' link to train SVM algorithm and get below output



In above screen we can see SVM confusion matrix where x-axis represents predicted class and y-axis represents TRUE class and we can see SVM predict 2977 records correctly as NORMAL and only 145 are incorrect prediction and it predict 824 records as PHISHING URL and only 26 are incorrect prediction and now close above graph to get below output

In above screen with SVM we got 95% accuracy and now click on 'Run Light GBM Algorithm' link to get below output



In above screen we can see Decision Tree confusion matrix graph and now close above graph to get below output

In above screen with Light GBM also we got 96% accuracy and now click on 'Test Your URL' link to get below screen

In above screen enter any URL and then press button and then Light GBM will predict whether that URL IS normal or phishing

2250-3676 ijesatj@gmail.com https://www.ijesat.com/

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 05 (May) 2026
Received: 10-3-2026 Accepted: 21-4-2026 Published: 07-5-2026

In above screen I entered URL as https://mail.google.com and then press button to get below output

In above screen in blue colour text we can see given URL predicted as GENUINE (normal) and now test other URL. Similarly now I will enter Google.com in below screen

In above screen I gave URL as Google.com and below is the output

In above screen Google.com also predicted as Genuine. Now in below screen from internet I am taking one phishing URL and then input to my application to get prediction

In above screen blue colour URL is the phishing URL and I will input that to my application in below screen and below is the phishing URL from internet

'https://in.xero.com/3LQDhRwfvoQfeDtlDMqkk1JWSqC4CMJt4VVJRsGN'

In above screen I entered same URL and press button to get below output

In above screen in blue colour text we can see application detected PHISHING in given URL and similarly you can enter any URL and detect it as NORMAL or phishing

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 05 (May) 2026
Received: 10-3-2026 Accepted: 21-4-2026 Published: 07-5-2026

VIII. CONCLUSION

In this project, an effective and intelligent system for detecting phishing websites has been developed using advanced machine learning algorithms, namely **Support Vector Machine (SVM)** and **Light Gradient Boosting Machine (LightGBM)**. The system focuses on analyzing various URL-based and content-based features to accurately classify websites as legitimate or phishing, thereby addressing the growing threat of cyber fraud in the digital world.

The implementation of SVM provides strong classification capabilities, especially in handling high-dimensional data, while LightGBM enhances the system with faster processing speed and improved accuracy through gradient boosting techniques. The comparative analysis of both models demonstrates that the proposed approach achieves high detection accuracy with reduced false positives, making it more reliable than traditional rule-based and blacklist methods.

IX. FUTURE WORK: Future work for this

Although the proposed phishing detection system using SVM and LightGBM achieves high accuracy and reliability, there are several areas where further improvements and enhancements can be made to increase its effectiveness and real-world applicability.

In future work, the system can be enhanced by integrating deep learning techniques such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) to analyze webpage content, screenshots, and sequential URL patterns more effectively. This would help in capturing complex relationships and visual similarities between phishing and legitimate websites.

Another important improvement is the incorporation of real-time threat intelligence feeds and continuous learning mechanisms. By updating the model with newly detected phishing URLs and patterns, the system can remain adaptive to emerging threats and reduce the risk of zero-day attacks.

XI. REFERENCES

➤ Abdelhamid, N., Ayesh, A., & Thabtah, F. (2014). *Phishing detection based associative classification data mining*. Expert Systems with Applications.

➤ Ma, J., Saul, L. K., Savage, S., & Voelker, G. M. (2009). *Beyond blacklists: Learning to detect malicious web sites from suspicious URLs*. ACM SIGKDD.

➤ Zhang, Y., Hong, J. I., & Cranor, L. F. (2017). *Cantina+: A feature-rich machine learning framework for detecting phishing*

web sites. ACM Transactions on Information and System Security.

➤ Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. Y. (2017). *LightGBM: A highly efficient gradient boosting decision tree*. Advances in Neural Information Processing Systems (NeurIPS).

➤ Sahingoz, O. K., Buber, E., Demir, O., & Diri, B. (2019). *Machine learning based phishing detection from URLs*. Expert Systems with Applications.

➤ Rao, R. S., & Pais, A. R. (2019). *Detection of phishing websites using an efficient feature-based machine learning framework*. Neural Computing and Applications.

➤ Basnet, R., Mukkamala, S., & Sung, A. H. (2014). *Detection of phishing attacks: A machine learning approach*. Soft Computing Journal.

➤ Mohammad, R. M., Thabtah, F., & McCluskey, L. (2015). *Phishing detection: A recent intelligent machine learning comparison based on models content and features*. IEEE Conference.

➤ Chiew, K. L., Tan, C. L., Wong, K., Yong, K. S. C., & Tiong, W. K. (2018). *A new hybrid ensemble feature selection framework for machine learning-based phishing detection system*. Information Sciences.

➤ Verma, R., & Das, A. (2017). *What's in a URL: Fast feature extraction and malicious URL detection*. ACM Workshop on Artificial Intelligence and Security.

➢ Jajam Venkata Anil Kumar, Dr. G. Charles Babu, "Automating Content Utilizing Big Data Innovations", *Journal of Advances and Scholarly Researches in Allied EducationVol. 15, Issue No. 9, October-2018, ISSN 2230-7540,* IIFS : 1.6 (2014), INDEX COPERNICUS : 49060 (2018), IJINDEX : 3.46 (2018), pp.635-639, 2018.

➢ Jajam Venkata Anil Kumar, Dr. G. Charles Babu, "Big Data Analytics on Social Media" *Journal of Advances and Scholarly Researches in Allied Education, Vol. XII, Issue No. 23, October-2016, ISSN 2230-7540,* IIFS : 1.6 (2014), INDEX COPERNICUS : 49060 (2018), IJINDEX : 3.46 (2018), pp. 389-393,2016.