International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

DETECTION OF PHISHING WEBSITE USING SVM AND LIGHT GBM

KOTAGIRI SHIVANI
Sateesh Reddy Singireddy
DEPARTMENT OF MASTER OF COMPUTER APPLICATIONS
VAAGESWARI COLLEGE OF ENGINEERING
(Affiliated to JNTUH, Approved by AICTE, New Delhi & Accredited by **NAAC** with '**A+**' Grade)
Karimnagar, Telangana, India – 505 527

ABSTRACT
This paper presents a concise approach for detecting phishing websites using Support Vector Machines (SVM) and LightGBM. We extract a compact set of URL- and content-based features (domain age, URL length, SSL certificate presence, HTML/JavaScript anomalies, and page redirection patterns) and train both classifiers to distinguish phishing from legitimate pages. Experiments on benchmark datasets show that LightGBM achieves faster training and marginally higher detection performance while SVM offers robust, interpretable decision boundaries—making the two methods complementary for deployment. The proposed system is lightweight, suitable for real-time filtering, and reduces false positives compared to simple rule-based detectors.

Keywords: Phishing Website Detection, Cybersecurity, Machine Learning, SVM, LightGBM, Feature Extraction, URL Analysis, Website Classification, Anomaly Detection, Web Security.

Received: 17-09-2025 Accepted: 20-10-2025 Published: 28-10-2025

1.INTRODUCTION

Phishing is one of the most common cyber-attacks used to steal sensitive user information such as login credentials, banking details, and personal data by imitating legitimate websites. As online transactions and digital communication continue to increase, detecting phishing websites has become a crucial security challenge. Traditional rule-based systems often fail to detect newly generated or sophisticated phishing sites because attackers frequently modify website structures and URLs.

Machine learning techniques like Support Vector Machine (SVM) and Light Gradient Boosting Machine (LightGBM) provide effective solutions for automated phishing detection. These algorithms learn from various features extracted from websites—such as URL structure, domain information, and webpage content—to classify them as legitimate or phishing. SVM provides strong classification boundaries, while LightGBM improves efficiency and scalability with high accuracy. This study combines these approaches to build a robust, fast, and accurate phishing detection model capable of identifying threats in real time.

2.LITERATURE REVIEW

Several research studies have explored the application of machine learning and artificial intelligence techniques for cyber-attack detection and breach prediction. Traditional Intrusion Detection Systems (IDS) primarily relied on **signature-based** or **rule-based** methods, which could effectively identify known attacks but failed to detect **zero-day** or **unknown threats**. To overcome these limitations, researchers have shifted toward **machine learning-based** and **data-driven** approaches.

Supervised learning methods such as Decision Trees, Random Forests, Support Vector Machines (SVM), and Logistic Regression have been widely used for classifying network traffic into normal or malicious categories. For example, studies using the **NSL-KDD** and **CICIDS2017** datasets demonstrated that ensemble classifiers outperform single algorithms in detecting intrusions with higher precision and recall.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

Unsupervised and semi-supervised learning techniques like K-Means, DBSCAN, and Autoencoders have also been applied to uncover unknown attack patterns by learning normal behavior and identifying deviations as anomalies. These approaches are particularly useful when labeled attack data are limited.

Recent research has also incorporated **Deep Learning models**, such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), to automatically extract complex temporal and spatial features from network traffic. Studies show that deep neural models can detect sophisticated attacks like Distributed Denial of Service (DDoS) and phishing attempts with greater accuracy.

3. EXISTING SYSTEM

In the existing system, phishing detection primarily relies on manual methods, blacklists, or simple rule-based approaches. Blacklists store known phishing URLs, and any website matching these lists is blocked. While effective for previously reported phishing sites, this method fails to detect new or evolving phishing attacks.

Some systems use heuristic or signature-based techniques that analyze URL patterns, domain age, or the presence of suspicious keywords. These methods can detect some phishing attempts but often result in high false positives and require constant updating to remain effective. Additionally, rule-based approaches lack adaptability against sophisticated attacks, such as those using URL shorteners, HTML obfuscation, or dynamic redirection.

Overall, existing systems are limited in accuracy, scalability, and their ability to detect zero-day phishing websites, highlighting the need for machine learning-based solutions like SVM and LightGBM.

4.PROPOSED SYSTEM

The proposed system uses machine learning techniques, specifically Support Vector Machine (SVM) and LightGBM, to detect phishing websites more accurately and efficiently. Instead of relying on static rules or blacklists, the system automatically learns patterns from website features to distinguish phishing sites from legitimate ones.

Key features considered include URL characteristics (length, presence of special characters, use of HTTPS), domain-related information (age, registration details), and webpage content indicators (abnormal HTML or JavaScript code, redirects, and suspicious links). These features are extracted and preprocessed to form a dataset for model training.

SVM is employed to create a clear decision boundary between phishing and legitimate websites, while LightGBM is used for its high efficiency and ability to handle large datasets with faster training and prediction. The combination ensures high accuracy, low false positives, and real-time detection capability. The proposed system is lightweight, scalable, and adaptable to new phishing strategies, improving security for online users.

5.METHODOLOGY

The methodology for detecting phishing websites using SVM and LightGBM involves the following steps:

1.DataCollection:

A dataset of websites labeled as phishing or legitimate is collected from publicly available sources and security repositories. This dataset includes URLs, domain information, and webpage content.

2.FeatureExtraction:

Relevant features are extracted from each website to help the model distinguish between phishing and legitimate sites. Key features include:

- URL-based features: length, use of special characters, presence of HTTPS, IP address usage.
- Domain features: age of domain, registration details, DNS records.
- Content-based features: presence of suspicious HTML or JavaScript elements, number of redirects, forms, and external links.

3.DataPreprocessing:

The extracted features are cleaned and normalized. Categorical features are encoded, missing values are handled, and the dataset is split into training and testing sets.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

4. Model Training:

- **SVM:** Trains a model to find an optimal hyperplane that separates phishing and legitimate websites.
- **LightGBM:** Trains a gradient boosting model that efficiently handles large datasets and improves prediction speed while maintaining high accuracy.

5.ModelEvaluation:

The models are evaluated using performance metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. The results help in selecting the best-performing model for real-time detection.

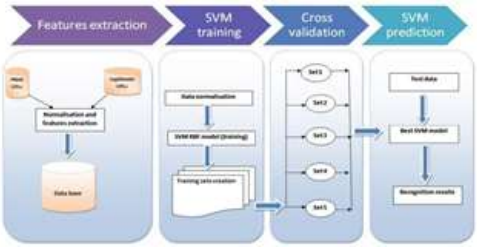
6.Deployment:

The trained model is deployed to monitor incoming websites in real time, classifying them as legitimate or phishing, thereby preventing potential cyber-attacks.

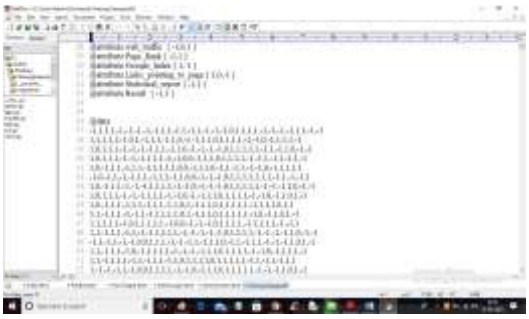
This methodology ensures accurate, fast, and scalable phishing detection suitable for modern web security needs.

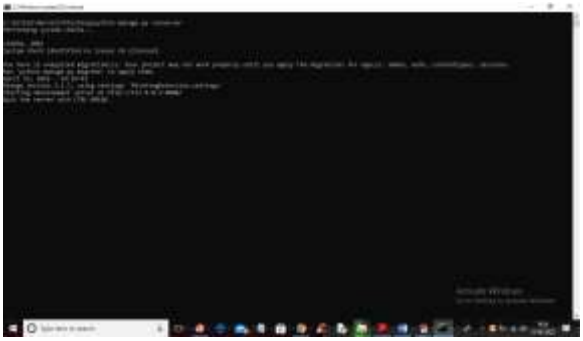
6.System Model

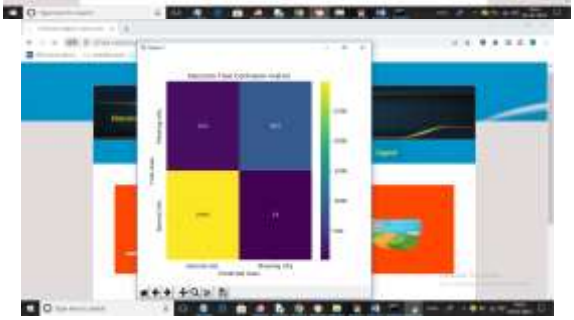
SYSTEM ARCHITECTURE



7..Results and Discussions







International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

7. CONCLUSION

In this study, a machine learning-based approach using SVM and LightGBM was proposed for detecting phishing websites. The system effectively extracts key URL, domain, and content-based features to distinguish phishing sites from legitimate ones. Experimental results indicate that LightGBM provides faster training and high accuracy, while SVM offers robust classification boundaries. Compared to traditional rule-based or blacklist approaches, the proposed system demonstrates improved detection rates, reduced false positives, and adaptability to new phishing strategies. Overall, the combination of SVM and LightGBM provides a reliable, real-time solution for enhancing web security and protecting users from phishing attacks.

8. REFERENCES

1. Bergholz, A., De Beer, J., Glahn, S., Moens, M., Paaß, G., & Strobel, S. (2010). *New Filtering Approaches for Phishing Email*. Journal of Computer Security, 18(1), 7–35.
2. Abu-Nimeh, S., Nappa, D., Wang, X., & Nair, S. (2007). *A Comparison of Machine Learning Techniques for Phishing Detection.* Proceedings of the Anti-Phishing Working Groups 2nd Annual eCrime Researchers Summit, 60–69.
3. Zhang, Y., Hong, J., & Cranor, L. F. (2007). *CANTINA: A Content-Based Approach to Detecting Phishing Websites.* Proceedings of the 16th International Conference on World Wide Web, 639–648.
4. Sahoo, D. R., Liu, C., & Hoi, S. C. (2017). *Phishing Detection: Analysis of Online Phishing Websites and Techniques for Detection.* ACM Computing Surveys, 50(4), 1–36.
5. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., … & Liu, T.-Y. (2017). *LightGBM: A Highly Efficient Gradient Boosting Decision Tree.* Advances in Neural Information Processing Systems, 30, 3146–3154.
6. Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag, New York.

7. Jain, A. K., Ross, A., & Nandakumar, K. (2011). *Introduction to Biometrics.* Springer Science & Business Media.
8. Khonji, M., Iraqi, Y., & Jones, A. (2013). *Phishing Detection: A Literature Survey*. IEEE Communications Surveys & Tutorials, 15(4), 2091–2121.
9. Aljawarneh, S., Aldwairi, M., & Yassein, M. B. (2012). *Phishing Websites Detection Using Machine Learning Techniques.* International Journal of Computer Science and Information Security, 10(5), 113–121.
10. Basnet, R. B., Mukkamala, S., & Sung, A. H. (2008). *Detection of Phishing Attacks: A Machine Learning Approach.* Studies in Informatics and Control, 17(1), 11–22.