International Journal of Scientific Research and Engineering Development-– Volume 9 Issue 2, Mar-Apr 2026
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Comparative study of URL based Phishing Detection Using Machine Learning

A. Beulah*, Mrs. M. Saranya, M.Sc., M.Phil.,**
*(Department of Computer Science, Sri Kaliswari College(Autonomous), Sivakasi
Email: a14pcs001@kaliswaricollege.edu.in)
** (Department of Computer Science, Sri Kaliswari College(Autonomous), Sivakasi
Email: msaranya.skc@gmail.com)

----------**********----------

Abstract:

Phishing detection is crucial for protecting users from cyber threats, especially as online transactions and digital communications continue to grow rapidly. Fraudulent websites are designed to mimic legitimate platforms, making it difficult for users to distinguish between genuine and malicious URLs. Machine learning techniques, particularly ensemble learning models such as Random Forest and XGBoost (Extreme Gradient Boosting), have demonstrated significant potential in improving phishing detection accuracy. A comparative analysis of cost-sensitive Random Forest and cost-sensitive XGBoost is conducted to evaluate their effectiveness in detecting phishing URLs. URL-based features, including length, entropy, presence of suspicious keywords, number of special characters, domain characteristics, and HTTPS usage, are extracted and used as input for training and testing both models. Since phishing datasets are typically imbalanced, cost-sensitive learning is incorporated to assign higher misclassification penalties to phishing instances. Results indicate that while both models enhance detection performance compared to conventional approaches, XGBoost slightly outperforms Random Forest in terms of recall, F1-score, and ROC-AUC, whereas Random Forest provides stable and interpretable results. The findings contribute toward developing reliable and scalable phishing detection mechanisms for strengthening cybersecurity systems.

Keywords — **Phishing Detection, Cost-Sensitive Learning, Random Forest, XGBoost, Class Imbalance, Cybersecurity, Machine Learning.**

----------**********----------

I. INTRODUCTION

Phishing detection plays a critical role in safeguarding digital communication systems, online banking platforms, and e-commerce services, particularly as cyber attacks continue to increase in frequency and sophistication. Fraudulent websites are carefully designed to imitate legitimate platforms, deceiving users into revealing sensitive information such as login credentials, credit card numbers, and personal data. The consequences of successful phishing attacks include financial losses, identity theft, and reputational damage to organizations. As internet usage expands globally, the need for reliable and intelligent phishing detection mechanisms has become increasingly important. Traditional phishing detection methods, such as blacklist-based filtering and rule-based systems, are often insufficient in identifying newly generated or zero-day phishing URLs. Attackers frequently create dynamic domains and obfuscated URLs to evade detection systems. The growing complexity and evolving nature of phishing techniques demand advanced methodologies capable of capturing subtle structural and lexical patterns embedded within malicious URLs.

Machine learning (ML) has emerged as a powerful alternative for phishing detection, demonstrating superior capability in handling large-scale datasets and identifying nonlinear relationships among features. Unlike conventional approaches, ML models learn from historical data patterns and generalize to detect previously unseen phishing websites. By analyzing URL-based features such as length, domain structure, entropy, suspicious keywords, and security indicators, ML algorithms can effectively distinguish between legitimate and malicious URLs. Among various ML techniques, ensemble learning models such as Random Forest and Extreme Gradient Boosting (XGBoost) have gained significant attention due to their robustness and high predictive performance. These algorithms combine multiple weak learners to construct a strong classifier, reducing variance and improving generalization. Random Forest operates through bootstrap aggregation of decision trees, while XGBoost utilizes gradient boosting with regularization to enhance predictive accuracy. However, phishing datasets typically exhibit class imbalance, where legitimate URLs significantly outnumber phishing URLs. This imbalance often leads to biased models that favor the majority class, resulting in higher false negative rates. In cybersecurity applications, misclassifying a phishing URL as legitimate can have severe consequences. Therefore, cost-sensitive learning techniques are incorporated to assign higher penalties to phishing misclassification, improving detection reliability. In this context, a comparative evaluation of cost-sensitive Random Forest and cost-sensitive XGBoost provides valuable insights into their effectiveness in reducing phishing risk. URL datasets collected from public repositories serve as the foundation for training and evaluation. Feature extraction, preprocessing, and model optimization techniques are applied to ensure optimal performance. Performance assessment is conducted using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. Emphasis is placed on recall and false negative reduction, as these metrics are critical in cybersecurity systems. Additionally, feature importance analysis provides interpretability regarding which URL characteristics contribute most significantly to phishing detection. Enhancing phishing detection capability has significant implications for cybersecurity infrastructure, financial institutions, and online service providers. More accurate detection mechanisms enable proactive threat mitigation, protecting users from fraudulent activities. The integration of cost-sensitive ensemble learning approaches contributes to building adaptive and scalable security systems capable of responding to evolving cyber threats. With the increasing availability of large-scale cybersecurity datasets and improvements in computational power, phishing detection models can be further refined. Future research may explore hybrid ensemble techniques, deep learning integration, and real-time detection systems to strengthen protection against sophisticated phishing attacks. As cyber threats continue to evolve, continuous advancement in intelligent detection frameworks remains essential for ensuring secure digital environments.

II. LITERATURE SURVEY

Phishing detection has become a major research focus in cybersecurity due to the continuous growth of malicious online activities. Traditional approaches such as blacklist filtering and rule-based detection systems have been widely used to identify phishing websites. However, these techniques are limited in detecting newly generated phishing URLs and sophisticated attacks that frequently change their structure. As a result, researchers have increasingly adopted machine learning and artificial intelligence techniques to improve phishing detection accuracy and reliability.

Alqarni et al. (2024) proposed a feature-based machine learning framework for detecting phishing URLs by extracting structural and lexical characteristics from URLs. Their approach analyzes

various URL features such as length, special characters, and domain patterns to distinguish phishing websites from legitimate ones. The study demonstrates that feature engineering combined with machine learning classifiers can significantly enhance web security by accurately identifying malicious URLs.

Sharma and Gupta (2023) introduced an AI-driven phishing detection model using reinforcement learning techniques to enhance cybersecurity defenses. Their system continuously learns from new phishing patterns and adapts its decision-making process to improve detection accuracy. The research highlights the potential of reinforcement learning for developing adaptive phishing detection systems capable of responding to evolving cyber threats.

Kumar et al. (2024) explored the use of machine learning algorithms for phishing URL detection through data-driven analysis. Their study applied various classification algorithms to analyze patterns within URL structures and identify phishing characteristics. Experimental results indicate that machine learning models can effectively classify phishing URLs by learning from historical data and extracting meaningful patterns from URL attributes.

Zhang and Wang (2025) investigated multiple machine learning techniques for phishing website detection. Their study compared several classifiers and found that ensemble-based models achieved higher detection accuracy due to their ability to combine multiple decision trees and reduce overfitting. The results demonstrate that machine learning approaches provide reliable solutions for identifying phishing websites in large datasets.

Ahmed et al. (2025) proposed an ensemble learning-based model called EnLeM for phishing website detection. This model integrates multiple classifiers to improve prediction accuracy and reduce classification errors. By combining the strengths of different algorithms, the ensemble approach enhances the overall reliability of phishing detection systems and minimizes the risk of false predictions.

Silva and Costa (2024) focused on real-time phishing attack detection using machine learning techniques. Their research emphasizes the importance of developing systems capable of identifying phishing attacks in real-time environments. The study demonstrates that machine learning models can effectively detect phishing websites while maintaining efficient computational performance for practical deployment.

Chen et al. (2025) introduced the EXPLICATE framework, which enhances phishing detection using explainable artificial intelligence (XAI) and large language model (LLM) interpretability. Their approach focuses on improving transparency in machine learning models by providing explanations for classification decisions. This method helps cybersecurity analysts better understand the reasoning behind phishing detection outcomes and increases trust in automated security systems.

Patel and Rao (2025) proposed a dual-path phishing detection framework that integrates transformer-based natural language processing techniques with structural URL analysis. Their model combines textual analysis of URLs with structural feature extraction to capture complex phishing patterns. The study demonstrates that hybrid approaches combining deep learning and feature-based analysis can significantly improve phishing detection accuracy. Although numerous studies have explored various machine learning and deep learning techniques for phishing detection, limited research focuses on comparative analysis of cost-sensitive ensemble models for URL-based detection. Therefore, this study aims to evaluate and compare Random Forest and XGBoost models under a cost-sensitive learning framework to improve phishing detection performance while minimizing critical misclassification errors.

III. METHODOLOGY

To enhance phishing URL detection using cost-sensitive Random Forest and XGBoost, the proposed methodology follows a systematic and structured machine learning pipeline designed to improve accuracy, recall, F1-score, and overall detection reliability. The process begins with dataset acquisition from publicly available phishing repositories and legitimate URL sources. The dataset consists of labeled URLs categorized as phishing and legitimate instances. Since raw URL datasets may contain missing entries, duplicate records, and inconsistent labeling, data preprocessing is performed to ensure dataset integrity. Missing values in URL and class label fields are removed, and duplicate URLs are eliminated to avoid model bias. Class labels are converted into binary format for supervised classification. Stratified sampling is applied during train-test splitting to preserve class distribution and prevent imbalance distortion. Feature engineering plays a critical role in the proposed framework. Instead of relying on deep learning-based automatic feature extraction, handcrafted lexical, structural, and statistical features are derived directly from URL strings. Extracted features include URL length, path length, query length, number of dots, hyphens, digits, special characters, uppercase characters, subdomain count, domain length, HTTPS presence, IP address detection, suspicious keyword indicators, shortening service detection, suspicious top-level domains, and URL entropy. Entropy is computed using Shannon's entropy formula to quantify the randomness of characters within the URL, as phishing URLs often exhibit higher entropy due to obfuscation techniques. Random Forest and XGBoost, both ensemble learning algorithms capable of modeling nonlinear relationships, are employed for classification. Random Forest operates through bootstrap aggregation of decision trees, while XGBoost utilizes gradient boosting with regularization to enhance predictive performance. Since phishing detection is inherently an imbalanced classification problem, cost-sensitive learning techniques are integrated to reduce false negatives. In Random Forest, class weights are assigned to penalize phishing misclassification more heavily. In XGBoost, the scale_pos_weight parameter is calculated based on the ratio of legitimate to phishing samples to balance gradient updates during training. Model performance is evaluated using stratified train-test validation. Key evaluation metrics such as accuracy, precision, recall, F1-score, confusion matrix analysis, ROC-AUC, and Precision-Recall curves are analyzed to assess detection effectiveness. Particular emphasis is placed on recall and false negative reduction, as misclassifying phishing URLs as legitimate poses significant security risks. Feature importance analysis is conducted using built-in importance scores from Random Forest and XGBoost to interpret model decisions and identify the most influential URL characteristics contributing to phishing detection.

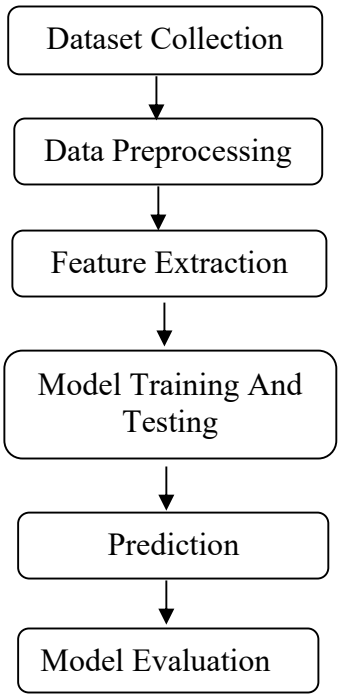

Fig 1. Workflow Diagram

IV. IMPLEMENTATION

To improve phishing URL detection accuracy, this project proposes a machine learning–based classification framework using two ensemble learning algorithms.

Random Forest – A bagging-based ensemble algorithm that builds multiple decision trees using bootstrap sampling and aggregates their predictions to improve stability and reduce overfitting. **XGBoost (Extreme Gradient Boosting)** – A boosting-based algorithm known for its gradient optimization, regularization techniques, and ability to handle nonlinear feature interactions efficiently.

Data Collection & Pre-processing: The dataset used in this study is the **PhishLegitURLs dataset** obtained from Mendeley Data (DOI: 10.17632/j43jtv3zzc.1). It contains approximately 1,000 labeled URLs categorized as phishing (0) and legitimate (1). Phishing URLs were sourced from URLHaus, while legitimate URLs were collected from trusted platforms such as Wikipedia and Stack Overflow.

The following preprocessing steps were performed:

- Removal of missing values in URL and ClassLabel columns
- Elimination of duplicate URLs to prevent bias
- Conversion of class labels into binary format

- Splitting the dataset into 80% training and 20% testing sets using stratified sampling

Feature Extraction: Since the dataset contains only raw URL strings, handcrafted feature extraction techniques are applied. The extracted features include lexical, structural, and statistical characteristics derived directly from the URL.

Key features generated include:

- URL length, path length, and query length
- Number of dots, hyphens, digits, and special characters
- Number of uppercase letters
- Subdomain count and domain length
- Presence of HTTPS protocol
- Detection of IP address in URL
- Identification of URL shortening services
- Suspicious keyword detection (login, verify, secure, bank, update)
- Suspicious top-level domain detection
- Double slash occurrence
- URL entropy to measure randomness

These features help the model distinguish between legitimate and malicious URL patterns.

Model Training & Cost-Sensitive Learning

Random Forest and XGBoost models are implemented using Scikit-learn and XGBoost libraries.Since phishing detection involves class imbalance, cost-sensitive learning techniques are applied:

- In Random Forest, class weights are assigned to penalize phishing misclassification more heavily.
- In XGBoost, the scale_pos_weight parameter is calculated using the ratio of legitimate to phishing samples to balance gradient updates.

The models are trained on the extracted feature set and evaluated using unseen test data.

Feature Importance Analysis

Feature importance scores are obtained from both Random Forest and XGBoost models. The most influential features contributing to phishing detection are identified and analyzed to interpret model behavior.

Prediction: The trained models classify URLs as phishing or legitimate based on extracted features.

Probability outputs are also generated to evaluate prediction confidence.

Evaluation: Model performance is assessed using the metrics are Accuracy, Precision, Recall, F1-score, Confusion Matrix, ROC-AUC Curve. Special emphasis is given to recall to minimize false negatives, as undetected phishing URLs pose serious cybersecurity risks.

V. RESULT

A phishing URL detection model was developed using cost-sensitive Random Forest and XGBoost, focusing on improving accuracy, precision, recall, and F1-score. After comprehensive data preprocessing, handcrafted feature extraction, and implementation of cost-sensitive learning techniques, both models were evaluated based on their classification performance. The results showed that the Random Forest model achieved an accuracy of 98.89%, demonstrating strong predictive capability in distinguishing phishing and legitimate URLs. The model achieved a phishing recall of 99%, successfully identifying the majority of malicious URLs while maintaining high precision. Additionally, the model exhibited zero misclassification of legitimate URLs in the test dataset, indicating strong generalization performance. The XGBoost model achieved an accuracy of 97.79%, with similarly high phishing detection performance. While XGBoost maintained strong precision and recall values, it showed slightly higher misclassification compared to Random Forest, particularly in legitimate URL prediction. Overall, both ensemble models demonstrated excellent performance in URL-based phishing detection. However, Random Forest slightly outperformed XGBoost in terms of overall accuracy and stability on the PhishLegitURLs dataset. The integration of cost-sensitive learning significantly reduced false negatives, which is critical in cybersecurity applications where undetected phishing attacks can cause severe security risks. These findings highlight the effectiveness of ensemble learning combined with cost-aware optimization in enhancing phishing detection systems.

Prediction performance

Algorithm	Accuracy	precision	Recall	F1-score
Random Forest	98.89%	0.98	0.99	0.98
XGBoost	97.79%	0.97	0.97	0.97

Table 1: Prediction Report

VI. CONCLUSION

The development and evaluation of a cost-sensitive phishing URL detection model using Random Forest and XGBoost demonstrated the effectiveness of ensemble learning techniques in cybersecurity applications. By leveraging handcrafted URL-based feature engineering and cost-aware optimization, the proposed framework achieved strong classification performance in distinguishing phishing URLs from legitimate ones.Among the two models, Random Forest achieved superior performance with an accuracy of 98.89% and high precision, recall, and F1-score values, making it a more reliable choice for phishing detection on the PhishLegitURLs dataset. The model showed excellent capability in minimizing misclassification, particularly reducing false negatives, which is critical in cybersecurity systems. XGBoost also demonstrated strong predictive performance with an accuracy of 97.79%, though it exhibited slightly higher misclassification compared to Random Forest.The integration of cost-sensitive learning significantly improved phishing detection by assigning higher penalty to malicious URL misclassification. These findings highlight the potential of ensemble

machine learning techniques in strengthening web security systems. Future work may focus on incorporating larger datasets, real-time detection frameworks, cross-validation strategies, and advanced feature selection techniques to further enhance model robustness and generalization performance.

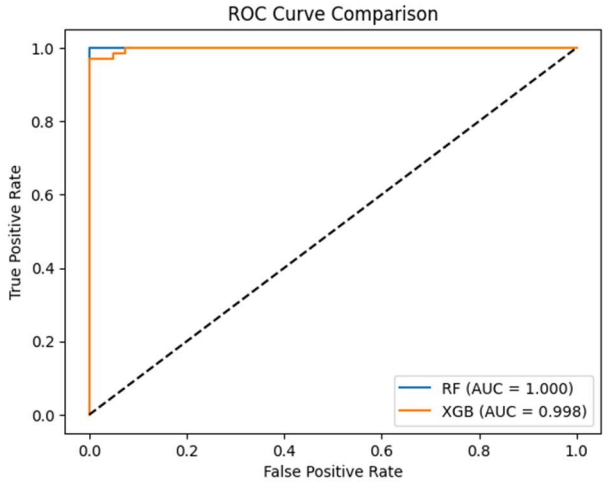

Fig 2. ROC Curve Comparison

REFERENCES

[1] M. A. Alqarni, A. S. Alfakeeh, and M. Alshahrani, "Improving Web Security through Machine Learning: A Feature-Based Methodology for Detecting Phishing URLs," *Engineering, Technology & Applied Science Research*, vol. 14, no. 2, 2024.

[2] A. Sharma and R. Gupta, "AI-Driven Phishing Detection: Enhancing Cybersecurity with Reinforcement Learning," *Cybersecurity*, vol. 5, no. 2, Art. no. 26, 2023.

[3] S. Kumar, P. Singh, and R. Mehta, "Phishing URL Detection Using Machine Learning: Harnessing Data Analysis to Strengthen Cyber Security," in *Proceedings of the International Conference on Intelligent Systems and Applications*, Springer, 2024.

[4] H. Zhang and L. Wang, "Phishing Website Detection Using Machine Learning Techniques," *International Journal of Information Engineering and Electronic Business*, vol. 17, no. 2, pp. 95–110, 2025.

[5] R. Ahmed, K. Rahman, and M. Islam, "EnLeM: Ensemble Learning-Based Model to Detect Phishing Websites," *EURASIP Journal on Information Security*, vol. 2025, Art. no. 19, 2025.

[6] J. Silva and M. Costa, "Application of Machine Learning for Real-Time Phishing Attack Detection," *Journal of Engineering Technology and Industrial Applications (JETIA)*, vol. xx, no. xx, pp. xx–xx, 2024.

[7] Y. Chen, L. Zhao, and T. Wang, "EXPLICATE: Enhancing Phishing Detection through Explainable AI and LLM-Powered Interpretability," arXiv preprint arXiv:2503.20796, 2025.

[8] D. Patel and S. Rao, "Dual-Path Phishing Detection: Integrating Transformer-Based NLP with Structural URL Analysis," arXiv preprint arXiv:2509.20972, 2025.

[9] A. A. Albishri and M. M. Dessouky, "A Comparative Analysis of Machine Learning Techniques for URL Phishing Detection," *Engineering, Technology & Applied Science Research*, vol. 14, no. 6, pp. 18495–18501, Dec. 2024.

[10] A. Al-qasmi, A. Al-anazi, L. AL-shehri, S. A-lshaman, and W. Al-atawi, "Machine Learning-Based Phishing Detection System," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 12, no. 4, pp. 4367–4372, 2024.

[11] S. Patil, M. Patil, and K. Chinnaiah, "Machine Learning and Deep Learning for Phishing Page Detection," *Research Reports on Computer Science*, vol. 2, no. 3, pp. 45–54, May 2023.

[12] N. F. Almujahid, M. A. Haq, and M. Alshehri, "Comparative Evaluation of Machine Learning Algorithms for Phishing Site Detection," *PeerJ Computer Science*, vol. cs-2131, pp. 1–19, Jun. 2024.

[13] S. Mundodagi, N. Patel, and Z. Choudhari, "Detection of Phished URL's Using Machine Learning," *Journal of Web Engineering & Technology*, vol. 11, no. 03, 2024.

[14] M. A. Mukunthan, P. S. V. Reddy, M. Trinath Reddy, and P. Chakradhar Reddy, "Comparative Analysis of Machine Learning Algorithms for Phishing Detection Using URL Features," in *Proc. 4th Int. Conf. Info. Tech., Civil Innovation, Science, and Management (ICITSM 2025)*, Tiruchengode, India, 2025, pp. 1–8.

[15] J. Lukasz Wilk-Jakubowski, L. Pawlik, G. Wilk-Jakubowski, and A. Sikora, "Machine Learning and Neural Networks for Phishing Detection: A Systematic Review (2017–2024)," *Electronics*, vol. 14, no. 18, Art. 3744, Sep. 2025.

[16] B. Bezerra, I. Pereira, and M. Â. Rebelo, "A Case Study on Phishing Detection with a Machine Learning Net," *Int. J. Data Sci. Anal.*, vol. 20, pp. 2001–2020, 2025.

[17] S. Joshi and S. M. Joshi, "Phishing Urls Detection Using Machine Learning Techniques," *Int. J. Comput. Eng. Res. Trends*, vol. 6, no. 6, pp. 326–333, 2019.

