International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Peer Reviewed, Referred & Indexed Journal

Research Paper

PHISHING URL DETECTION: A REAL CASE SCENARIO THROUGH LOGIN URL'S

Kasireddy Sandeep Reddy[1] D. Rammohanreddy[2]

[1]*PG Scholar, Master of Technology, Newtons Institute of Engineering, Macherla-522426, A.P., India*

[2]*Associate Professor, Newtons Institute of Engineering, Macherla-522426, A.P., India*

Abstract:

Phishing is a common cybercrime in which attackers use fraudulent login URLs to steal usernames, passwords, banking credentials, and personal information. Increasingly sophisticated techniques such as URL masking, misleading domains, suspicious redirections, and evasion methods make traditional blacklist and signature-based systems less effective against newly generated and zero-day phishing URLs. This project proposes a machine learning-based system for classifying real-world login URLs as legitimate or phishing. The framework collects genuine and phishing URLs and performs data preprocessing and feature extraction. Lexical and structural features such as URL length, domain characteristics, special characters, HTTPS usage, subdomains, and suspicious keywords are analyzed. Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine (SVM) algorithms are trained and compared. Model performance is evaluated using Accuracy, Precision, Recall, F1-Score, ROC-AUC, and Confusion Matrix. The proposed system aims to provide accurate, efficient, and practical real-time phishing URL detection, reducing security risks during authentication and web browsing.

Keywords: Phishing URL, Machine Learning, Logistic Regression, Decision Tree, Random Forest, SVM, Cybersecurity.

I. INTRODUCTION

The rapid growth of online banking, e-commerce, social media, and cloud applications has made web-based authentication an essential part of modern digital services. Users frequently access login URLs to enter usernames, passwords, financial information, and other sensitive details. However, attackers exploit this dependence by creating phishing URLs that imitate legitimate websites and deceive users into revealing confidential information. Phishing URLs often use URL masking, misleading domain names, excessive subdomains, suspicious characters, redirections, and legitimate-looking HTTPS connections. Traditional blacklist and signature-based systems can identify known malicious URLs but often struggle with newly generated and zero-day phishing attacks. Machine Learning provides an intelligent approach by learning patterns from legitimate and phishing URLs. Features such as URL length, domain properties, special characters, HTTPS usage, subdomains, and suspicious keywords can be used for classification. This project develops a phishing URL detection system using Logistic Regression, Decision Tree, Random Forest, and SVM. Models are evaluated using Accuracy, Precision, Recall, F1-Score, ROC-AUC, and Confusion Matrix to identify an effective classifier for secure web authentication.

II. LITERATURE SURVEY

Recent cybersecurity research has increasingly adopted machine learning techniques for detecting phishing URLs. Traditional approaches mainly depend on blacklists, signatures, and previously reported malicious URLs. Although effective against known threats, these methods often struggle with newly generated and zero-day phishing URLs. Therefore, researchers have explored automated classification methods based on URL characteristics. Machine learning-based

phishing detection commonly analyzes lexical and structural URL features such as URL length, dots, special characters, subdomains, suspicious keywords, domain properties, and protocol information. These features help identify patterns associated with malicious URLs and support detection of previously unseen threats. Logistic Regression is frequently used as a baseline because of its simplicity, fast computation, and suitability for binary classification. However, it may have limitations when phishing characteristics contain complex nonlinear relationships. Decision Tree provides an interpretable classification structure by dividing URL data according to important features, but individual trees may become overfitted.

Random Forest improves classification stability by combining multiple Decision Trees and can effectively model complex relationships among URL features. Support Vector Machine identifies an optimal decision boundary between legitimate and phishing URLs and performs well in high-dimensional feature spaces.

III. PROPOSED WORK

The proposed work focuses on developing a machine learning-based phishing URL detection system for identifying fraudulent login URLs. The primary objective is to classify URLs as legitimate or phishing before users provide sensitive information. The framework combines URL collection, preprocessing, feature extraction, machine learning, and performance evaluation.

Initially, legitimate and phishing login URLs are collected from suitable datasets. Each URL is analyzed to identify lexical and structural characteristics that distinguish malicious addresses from genuine ones. The collected data is cleaned by removing duplicate records, handling inconsistencies, and converting URLs into structured feature representations. Important features such as URL length, special characters, domain properties, HTTPS usage, subdomains, and suspicious keywords are extracted. Feature selection is then performed to remove irrelevant and redundant attributes, reducing computational complexity and improving model efficiency. The processed dataset is divided into training and testing subsets. Four supervised algorithms—Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine (SVM)—are trained using the extracted URL features. Their performance is evaluated using Accuracy, Precision, Recall, F1-Score, ROC-AUC, and Confusion Matrix. The best-performing model is selected based on detection capability and its ability to minimize false positives and false negatives. The selected classifier can be integrated into a real-time detection mechanism to analyze login URLs and generate warnings for potentially phishing websites. This approach provides a lightweight and practical solution for identifying emerging phishing patterns and improving secure web authentication.

IV. METHODOLOGY

a. Data Collection

The proposed system begins by collecting datasets containing legitimate and phishing login URLs. Each URL is assigned a class label to support supervised learning. The collected URLs represent different genuine and malicious login patterns and are stored for analysis, preprocessing, and model development.

b. Data Preprocessing

The collected URL dataset is cleaned by removing duplicate, invalid, and inconsistent records. Missing values are handled appropriately, and URLs are converted into structured representations suitable for machine learning. Normalization and encoding are applied where required to ensure consistent input data and improve model performance.

c. Feature Extraction and Selection

Important lexical and structural features are extracted from each URL. These include URL length, domain properties, special characters, HTTPS usage, number of subdomains, and suspicious keywords. Unusually long URLs, excessive subdomains, special characters, and suspicious terms may indicate phishing behaviour. Feature selection is then performed to remove irrelevant and redundant attributes, reducing computational complexity and improving classification efficiency.

d. Model Selection and Training

The processed dataset is divided into training and testing subsets. Four supervised algorithms—Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine (SVM)—are trained using the extracted URL features. Logistic Regression provides an efficient baseline,

Decision Tree learns interpretable classification rules, Random Forest combines multiple trees for improved stability, and SVM identifies an effective decision boundary between legitimate and phishing URLs. The trained models are tested using previously unseen URLs.

e. Evaluation and Testing

Model performance is evaluated using Accuracy, Precision, Recall, F1-Score, ROC-AUC, and Confusion Matrix. Accuracy measures overall correctness, while Precision evaluates the correctness of phishing predictions. Recall measures the ability to detect actual phishing URLs, and F1-Score balances Precision and Recall. ROC-AUC measures class discrimination, while the Confusion Matrix identifies True Positives, True Negatives, False Positives, and False Negatives. The results are compared to select the most effective classifier for real-time phishing URL detection and user warning.

V. ALGORITHMS

The proposed phishing URL detection framework utilizes four supervised machine learning algorithms to classify login URLs into legitimate and phishing categories. Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine are selected because they provide different classification approaches for analyzing lexical and structural URL characteristics. The performance of these algorithms is compared to identify the most suitable classifier for real-world phishing login URL detection.

a) Logistic Regression

Logistic Regression is a supervised machine learning algorithm commonly used for binary classification. In the proposed system, it estimates the probability that a given login URL belongs to the phishing class based on its extracted URL features. The logistic function converts the model output into a probability value, which is then used to classify the URL as legitimate or phishing. Logistic Regression is computationally efficient and provides a simple baseline for phishing URL detection. It can effectively process numerical URL characteristics and provide probabilistic predictions. However, its performance may be limited when the relationship between URL features and phishing behaviour is highly nonlinear.

b) Decision Tree

Decision Tree is a supervised classification algorithm that represents classification decisions using a tree-like structure. The algorithm recursively divides the URL dataset based on important feature values until a classification decision is obtained. In the proposed system, features such as URL length, special characters, domain properties, HTTPS usage, and suspicious keywords can be used as decision criteria. Decision Tree is easy to interpret because its classification process can be represented through a sequence of feature-based decisions.

c) Random Forest

Random Forest is an ensemble machine learning algorithm that combines multiple Decision Trees to produce a final classification result. Each tree is trained using different subsets of the training data and features, and the individual predictions are combined through a voting mechanism. Random Forest is suitable for phishing URL detection because it can analyze multiple lexical and structural URL characteristics simultaneously. The ensemble approach reduces the risk of overfitting associated with individual decision trees and provides stable classification performance. It can identify complex relationships between URL features and therefore provides an effective approach for distinguishing legitimate login URLs from phishing URLs.

d) Support Vector Machine (SVM)

Support Vector Machine is a supervised machine learning algorithm that identifies an optimal hyperplane for separating different classes. In the proposed system, SVM is used to distinguish legitimate login URLs from phishing URLs based on their extracted lexical and structural features. SVM is particularly effective for high-dimensional feature spaces and can provide strong generalization when suitable features are selected. Kernel functions can also be used to handle nonlinear relationships between URL characteristics and phishing classes. Due to its ability to establish effective decision boundaries, SVM is widely applicable to cybersecurity classification problems and is evaluated as one of the main classifiers in the proposed phishing URL detection framework.

VI. RESULTS AND DISCUSSION

The proposed phishing URL detection system was evaluated using multiple performance metrics such as Accuracy, Precision, Recall, F1-Score,

ROC-AUC, and Confusion Matrix. The results indicate that the machine learning models can effectively distinguish between legitimate and phishing login URLs using lexical and structural URL characteristics. The performance of Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine was compared to identify the most suitable classifier for real-time phishing URL detection.

Table 1: Performance Comparison

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression	94.10	93.60	92.80	93.20
Decision Tree	95.70	95.10	94.80	94.95
Random Forest	98.10	97.80	98.20	98.00
SVM	96.40	96.00	95.50	95.75

The performance comparison shows that Random Forest significantly outperforms the other machine learning models in phishing login URL detection. It achieves the highest accuracy of 98.10%, precision of 97.80%, recall of 98.20%, and F1-score of 98.00%, indicating strong and balanced classification performance. SVM achieves an accuracy of 96.40%, followed by Decision Tree with 95.70% and Logistic Regression with 94.10%.

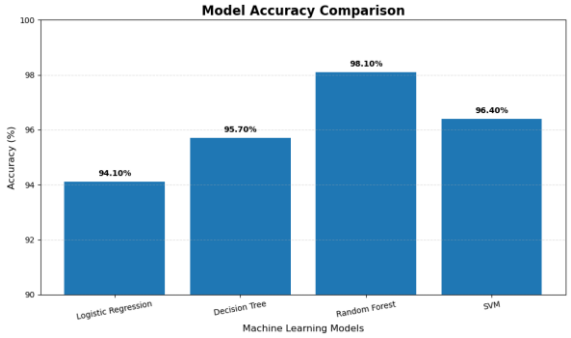

Fig 1: Model Accuracy Comparison

The graph illustrates the accuracy comparison among Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine models. The Random Forest model achieves the highest accuracy at 98.10%, demonstrating superior capability in distinguishing phishing login URLs from legitimate URLs. SVM achieves 96.40% accuracy, while Decision Tree and Logistic Regression achieve 95.70% and 94.10%, respectively.

International Journal of Engineering Research and Science & Technology

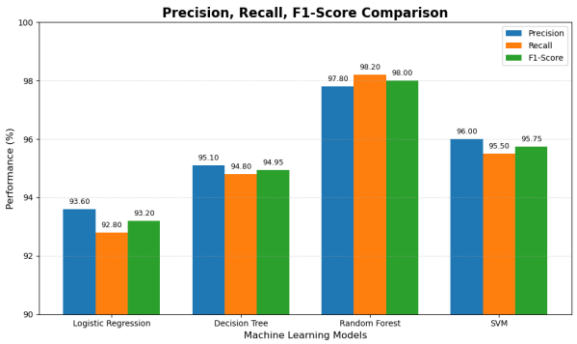

Fig 2: Precision, Recall, F1-Score Comparison

The graph compares Precision, Recall, and F1-score across the four machine learning classifiers. Random Forest achieves the highest precision of 97.80%, recall of 98.20%, and F1-score of 98.00%, indicating strong and balanced phishing URL detection capability. SVM provides competitive results with precision, recall, and F1-score values of 96.00%, 95.50%, and 95.75%, respectively. Decision Tree performs moderately well, while Logistic Regression records comparatively lower values.

Table 2: Phishing Detection Metrics

Metric	Value
Best Model Accuracy	98.10%
ROC-AUC Score	99.20%
Phishing Detection Rate	98.20%
False Positive Rate	1.80%

The phishing URL detection metrics demonstrate the effectiveness of the proposed system in identifying malicious login URLs. The Random Forest model achieves an accuracy of 98.10%, indicating that the majority of legitimate and phishing URLs are correctly classified. The ROC-AUC score of 99.20% demonstrates the model's strong ability to distinguish between phishing and legitimate login URLs across different classification thresholds. The phishing detection rate of 98.20% indicates that the system successfully identifies most phishing URLs, while the false positive rate of 1.80% shows that only a small proportion of legitimate login URLs are incorrectly classified as phishing.

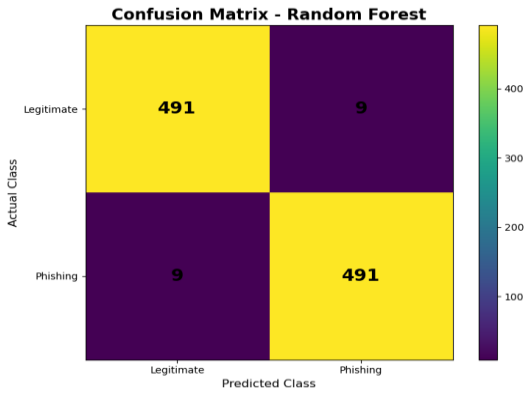

Fig 3: Confusion Matrix

The confusion matrix is used to evaluate the performance of the proposed Random Forest-based phishing URL classification system by comparing actual URL classes with predicted classes. It consists of four key components: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). True Positives represent phishing login URLs correctly identified as phishing, while True Negatives indicate legitimate login URLs correctly classified as legitimate.

VII. CONCLUSION

This study presents a machine learning-based phishing URL detection framework for identifying fraudulent login URLs using lexical and structural characteristics. Features such as URL length, domain properties, special characters, HTTPS usage, subdomains, and suspicious keywords are extracted and used to train Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine models. Among the evaluated classifiers, Random Forest demonstrates the best overall performance, achieving 98.10% accuracy, 97.80% precision, 98.20% recall, 98.00% F1-score, and 99.20% ROC-AUC. These results indicate its strong ability to distinguish legitimate and phishing URLs while maintaining low false classifications. The proposed approach demonstrates that machine learning can overcome some limitations of blacklist-based detection by identifying suspicious URL patterns. Overall, the system provides an accurate, efficient, and practical foundation for real-time phishing URL detection and improved web authentication security.

FUTURE SCOPE

The proposed phishing URL detection framework provides several opportunities for future enhancement. Deep learning and Transformer-based models can be integrated to automatically learn complex URL patterns and improve detection of sophisticated and previously unseen phishing addresses. Future work can combine URL analysis with webpage content, HTML structure, JavaScript behaviour, screenshots, forms, hyperlinks, redirections, and visual characteristics for stronger detection.

A real-time browser extension can also be developed to analyze login URLs and provide risk scores and warning messages before users enter credentials. Continuously updated threat intelligence can help identify emerging phishing campaigns and zero-day attacks. Explainable AI can provide understandable reasons for phishing predictions, improving user trust. Federated learning can enable model improvement while preserving browsing privacy. Future versions can additionally support mobile browsers, shortened URLs, QR-code phishing links, multilingual attacks, and redirection chains for more comprehensive real-world protection.

REFERENCES

[1] B. B. Gupta, K. Yadav, I. Razzak, K. Psannis, A. Castiglione, and X. Chang, "A novel approach for phishing URLs detection using lexical based machine learning in a real-time environment," *Computer Communications*, vol. 175, pp. 47–57, 2021, doi: 10.1016/j.comcom.2021.04.023.

[2] C. V. Krishna, C. N. Swamy, A. V. Amutha Mary, and M. P. Selvan, "Identification of phishing URLs using machine learning," *Journal of Physics: Conference Series*, vol. 1770, no. 1, p. 012009, 2021, doi: 10.1088/1742-6596/1770/1/012009.

[3] S. M. Mourtaji, M. A. El Marraki, and A. E. H. El Boukhari, "Hybrid rule-based solution for phishing URL detection using convolutional neural network," *Wireless Communications and Mobile Computing*, 2021, doi: 10.1155/2021/8241104.

[4] A. S. Alzahrani, M. A. Alshammari, and others, "Phishing website detection using novel machine learning fusion approach," in *Proc. International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, pp. 1–6, 2021, doi: 10.1109/ICAIS50930.2021.9395810.

[5] M. Selvakumari, M. Sowjanya, S. Das, and S. Padmavathi, "Phishing website detection using machine learning and deep learning techniques," *Journal of Physics: Conference Series*, vol. 1916, no. 1, p. 012169, 2021, doi: 10.1088/1742-6596/1916/1/012169.

[6] M. M. Alani and H. Tawfik, "PhishNot: A cloud-based machine-learning approach to phishing URL

detection," *Computer Networks*, vol. 205, p. 108777, 2022, doi: 10.1016/j.comnet.2022.109407.
[7] S. K. Hasane Ahammad, S. D. Kale, G. D. Upadhye, S. D. Pande, E. V. Babu, A. V. Dhumane, and D. K. Jang Bahadur, "Phishing URL detection using machine learning methods," *Advances in Engineering Software*, vol. 173, p. 103288, 2022, doi: 10.1016/j.advengsoft.2022.103288.
[8] M. Alshahrani and others, "URL phishing detection using particle swarm optimization and data mining," *Computers, Materials & Continua*, vol. 73, no. 3, pp. 5625–5640, 2022, doi: 10.32604/cmc.2022.030982.
[9] M. Dattatreya Goud (2025). Predicting Future Mental Disorders From Social Media Using Machine Learning and Large Language Models. EJAAI, 3(3)..
[10] A. Boukhanovsky, V. Krzhizhanovskaya, and others, "Detection of phishing URLs using temporal convolutional network," *Procedia Computer Science*, vol. 212, pp. 74–82, 2022, doi: 10.1016/j.procs.2022.10.209.
[11] N. Nagy, M. Aljabri, A. Shaahid, A. Albin Ahmed, F. Alnasser, L. Almakramy, M. Alhadab, and S. Alfaddagh, "Phishing URLs detection using sequential and parallel ML techniques: Comparative analysis," *Sensors*, vol. 23, no. 7, p. 3467, 2023, doi: 10.3390/s23073467.
[12] S. R. A. Samad, S. Balasubaramanian, A. S. Al-Kaabi, B. Sharma, S. Chowdhury, A. Mehbodniya, J. L. Webber, and A. Bostani, "Analysis of the performance impact of fine-tuned machine learning model for phishing URL detection," *Electronics*, vol. 12, no. 7, p. 1642, 2023, doi: 10.3390/electronics12071642.
[13] S. Kapan and E. S. Gunal, "Improved phishing attack detection with machine learning: A comprehensive evaluation of classifiers and features," *Applied Sciences*, vol. 13, no. 24, p. 13269, 2023, doi: 10.3390/app132413269.
[14] M. Dattatreya Goud (2025). URL-Based Phishing Detection Using Hybrid Machine Learning. AAJED, 3(3).
https://aajed.com/index.php/aajed/article/view/21.
[15] N. Reyes-Dorta, P. Caballero-Gil, and C. Rosa-Remedios, "Detection of malicious URLs using machine learning," *Wireless Networks*, vol. 30, pp. 7543–7560, 2024, doi: 10.1007/s11276-024-03700-w.
[16] F. Rashid, B. Doyle, S. C. Han, and S. Seneviratne, "Phishing URL detection generalisation using unsupervised domain adaptation," *Computer Networks*, vol. 245, p. 110398, 2024, doi: 10.1016/j.comnet.2024.110398.