Intelligent Phishing URL Detection System Using Machine Learning

N. Nagamani[1], Dr.R.Prasad Rao[2], B. Siva Sai[3], B. Vedavyas Naidu[4], K. Gowthami[5], K. Pavani[6]
[1]*Assistant Professor, Department of ECE, Avanthi Institute of Engineering & Technology*
[2]*Professor, Department of ECE, Avanthi Institute of Engineering & Technology*
[3, 4, 5,6] *Student Department of ECE, Avanthi Institute of Engineering & Technology*

***Abstract*—The Intelligent Phishing URL Detection System Using Machine Learning is developed to identify malicious and fraudulent URLs through the analysis of structural and lexical characteristics of web links. Phishing attacks represent one of the most widespread cyber threats, where attackers deceive users using fake URLs that imitate legitimate websites. To address this issue, the proposed system employs machine learning algorithms to classify URLs as either legitimate or malicious based on extracted URL features.**

The application is implemented using Python and Streamlit, providing an interactive and user-friendly interface. Multiple machine learning algorithms such as Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN) are integrated into the system for classification. Automated feature extraction techniques are used to analyze parameters including URL length, special character count, IP address presence, suspicious keywords, HTTPS usage, and digit-to-letter ratio.

A synthetic dataset consisting of legitimate domains, phishing-oriented keyword domains, and suspicious URLs is generated for training and testing purposes. Each model is evaluated using performance metrics, and the final classification is obtained through a majority voting mechanism among all classifiers. The system also includes a secure login interface, real-time URL validation, and graphical visualization of model confidence levels.

By combining machine learning with cyber security principles, the proposed system offers an effective solution for phishing detection. The project demonstrates how intelligent cyber defense mechanisms can improve online security awareness and protect users from malicious websites.

***Keywords*— Phishing Detection, Machine Learning, URL Analysis, Cyber security, Ensemble Learning, Streamlit, URL Classification.**

I. INTRODUCTION

The rapid growth of internet services and online platforms has significantly increased cyber threats, particularly phishing attacks. Phishing is a social engineering attack in which cybercriminals deceive users into revealing confidential information such as usernames, passwords, banking details, and personal credentials through fake websites or malicious links[2]. These fraudulent websites are often designed to closely resemble legitimate platforms, making manual detection difficult for users.

Traditional phishing detection mechanisms mainly rely on blacklist-based approaches, where previously identified malicious URLs are stored in databases. However, these approaches are ineffective against newly generated phishing URLs, as attackers continuously create and modify malicious links to bypass detection systems[1][3]. Therefore, intelligent and adaptive phishing detection systems based on machine learning have become essential in modern cybersecurity.

The proposed system focuses on analyzing URL structures rather than depending solely on blacklist databases. It extracts important lexical and structural features from URLs, including URL length, domain structure, suspicious keywords, IP-based URLs, HTTPS verification, and character distribution patterns[6]. These extracted features are processed using multiple machine learning algorithms to predict whether a URL is legitimate or malicious.

To improve classification accuracy and reliability, the system integrates multiple machine learning models such as Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN)[9][10]. Instead of relying on a single model, the system uses an ensemble majority voting mechanism to determine the final prediction[8].

The Streamlit-based interactive dashboard enables users to input URLs and instantly obtain prediction

results along with confidence scores and graphical visualizations. The proposed system demonstrates the effective integration of artificial intelligence and cyber security techniques for real-time phishing detection and online threat prevention.

1.1 OVERVIEW

The Intelligent Phishing URL Detection System is a web-based cybersecurity application developed using Python and Streamlit for detecting phishing URLs through machine learning techniques. The system operates in three primary stages: URL validation, feature extraction, and classification.

Initially, the system verifies whether the entered URL follows a valid web format using regular expression matching. Invalid or suspiciously formatted URLs are immediately classified as malicious. During the feature extraction phase, the system analyzes several lexical and structural characteristics of the URL, including:

- URL length
- Number of dots and hyphens
- Digit-to-letter ratio
- Presence of HTTPS
- Detection of IP addresses
- Suspicious keyword identification
- Top-level domain length

These extracted features help distinguish phishing URLs from legitimate ones.

The extracted data is then processed using multiple machine learning algorithms, including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN). Each model independently predicts whether the URL is legitimate or malicious[9].

To improve prediction reliability, the system employs a majority voting ensemble mechanism, where the final decision is based on the collective output of all classifiers. The system also includes a secure login module and an interactive dashboard for real-time URL analysis. Graphical visualizations such as bar charts are used to display model confidence levels and prediction outcomes.

Additionally, the system is scalable and adaptive, allowing integration with real-world datasets, browser extensions, and APIs for practical deployment in cybersecurity environments[5].

II. LITERATURE SURVEY

Phishing detection techniques have evolved considerably over the past two decades due to the rapid expansion of internet services, e-commerce platforms, online banking systems, and cloud computing technologies. As phishing attacks continue to become more sophisticated, researchers and cybersecurity experts have proposed numerous methods to identify and prevent such attacks[3].

Earlier phishing detection approaches were primarily rule-based and reactive in nature. However, advancements in artificial intelligence, machine learning, and deep learning have transformed phishing detection into a more intelligent and proactive process[4]. Modern detection systems focus on feature extraction, pattern recognition, and predictive analysis to identify malicious URLs.

This literature survey discusses traditional phishing detection techniques, machine learning-based approaches, feature engineering methods, implementation frameworks, comparative performance studies, and research gaps identified in existing systems.

2.1 TRADITIONAL METHODS FOR PHISHING DETECTION

Traditional phishing detection methods rely on conventional security approaches, each having several limitations:

1. Blacklist-Based Detection

Blacklist-based detection systems maintain databases containing known malicious URLs and block access to those websites. Although effective against previously identified phishing sites, these systems fail to detect newly generated or zero-day phishing attacks[1].

2. Heuristic-Based Detection

Heuristic methods use predefined rules and patterns to identify suspicious URLs. Examples include detecting unusual domain names, suspicious keywords, or excessive special characters. However, heuristic methods often generate high false-positive rates by incorrectly classifying legitimate websites as malicious[6].

3. Signature-Based Detection

Signature-based approaches identify phishing websites by comparing web content with known attack signatures. While effective for known threats, these systems cannot detect newly emerging phishing techniques that do not match existing signatures[4].

4. User Awareness Approaches

Some systems rely on user awareness and manual verification of URLs. However, this approach is

unreliable because many users lack sufficient technical knowledge to identify phishing attempts accurately[7].

Overall, traditional phishing detection techniques are insufficient for handling modern and dynamically evolving phishing attacks, thereby necessitating intelligent machine learning-based approaches.

III. PROPOSED SYSTEM

The proposed system introduces an AI-driven phishing URL detection mechanism that utilizes multiple machine learning algorithms for accurate classification. The system automatically extracts URL-based features and classifies URLs as legitimate or malicious using trained models[8][10].

KEY FEATURES OF THE PROPOSED SYSTEM

1. Automated Feature Extraction

The system automatically extracts and analyzes various URL features, including domain structure, URL path characteristics, suspicious keywords, HTTPS presence, and IP-based URLs.

2. Multi-Model Classification

Multiple machine learning algorithms are employed for classification, including:

- Logistic Regression
- Decision Tree
- Random Forest
- Support Vector Machine (SVM)
- K-Nearest Neighbors (KNN)

The use of multiple classifiers enhances prediction robustness and accuracy.

3. Majority Voting Mechanism

The final prediction is generated using an ensemble majority voting mechanism, where outputs from all classifiers are combined to determine the final result. This improves overall reliability and minimizes classification errors.

4. Interactive Dashboard

A Streamlit-based interactive dashboard is developed with secure login functionality. The dashboard supports real-time URL analysis and displays prediction results along with graphical confidence visualizations.

5. Scalable and Adaptive Design

The system supports integration with real-world datasets, APIs, and browser extensions, making it suitable for large-scale practical deployment.

ADVANTAGES OF THE PROPOSED SYSTEM

- Detects unknown and newly generated phishing URLs
- Uses intelligent machine learning-based decision-making
- Provides graphical confidence visualization
- Enhances user cyber security awareness
- Supports proactive defense mechanisms
- Improves classification accuracy through ensemble learning

The proposed system effectively combines cyber security principles with machine learning intelligence to provide a robust and scalable phishing detection solution for modern digital environments.

3.1 BLOCK DIAGRAM

The Data Flow Diagram (DFD) illustrates the movement of data through the system and demonstrates the interaction between different modules.

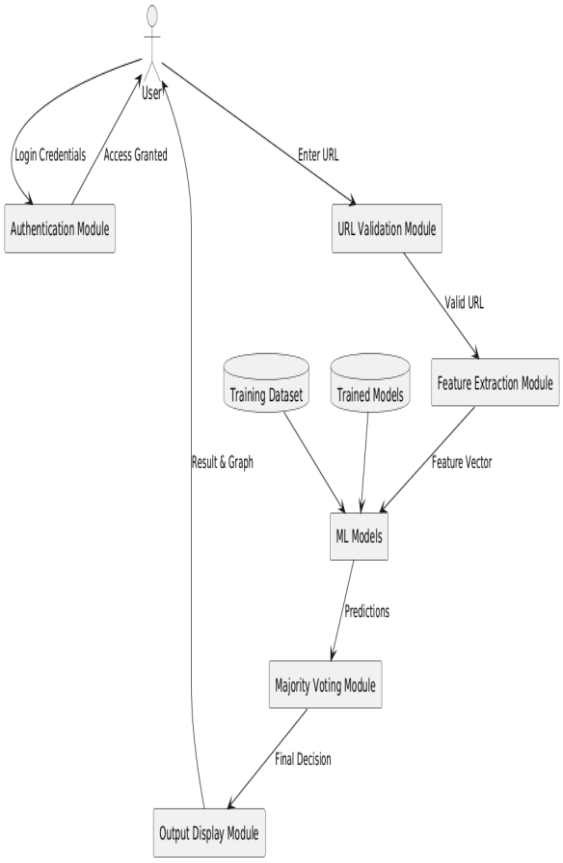

Fig 3.1. Dataflow Diagram

SYSTEM WORKFLOW

1. User enters login credentials
2. Authentication module validates the user
3. User enters the URL for analysis
4. URL validation module verifies the format

5. Feature extraction module extracts URL features
6. Machine learning models process extracted features
7. Individual predictions are generated
8. Majority voting module determines the final classification
9. Output module displays prediction results and confidence graphs

IV. RESULTS

The developed system successfully classifies URLs as phishing or legitimate based on extracted features and machine learning predictions.

Phishing Website Detection
The system accurately identifies malicious URLs containing suspicious patterns, phishing keywords, IP-based domains, or insecure structures.

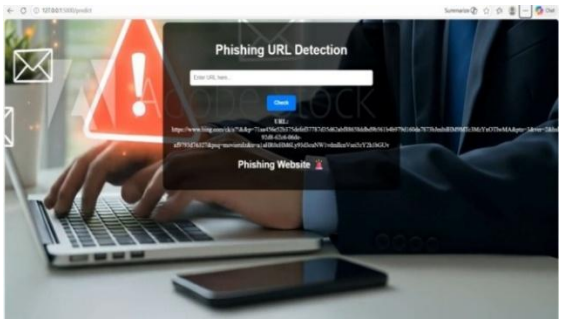

Fig 4.1. Phishing Website

Legitimate Website Detection
Legitimate websites with secure HTTPS protocols and trusted domain structures are correctly classified as safe.The experimental results demonstrate high prediction accuracy with minimal response time, confirming the efficiency of the proposed approach.

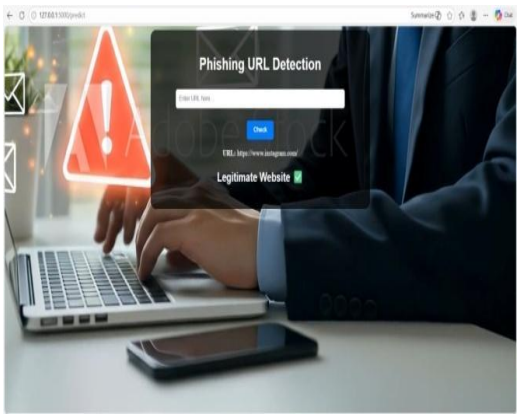

Fig 4.2.Legitimate website

V. CONCLUSION

The Intelligent Phishing URL Detection System Using Machine Learning and Cybersecurity Techniques successfully demonstrates how artificial intelligence can strengthen online security mechanisms. The system analyzes lexical and structural URL features and applies multiple machine learning algorithms to classify URLs as legitimate or malicious.

By integrating Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN), the system enhances prediction reliability through ensemble-based majority voting. This approach minimizes dependency on a single model and improves overall detection accuracy.

The project emphasizes the importance of feature extraction methods such as URL length analysis, suspicious keyword detection, HTTPS verification, and IP address identification in detecting phishing attacks. The Streamlit-based interface ensures user-friendly real-time URL analysis and confidence visualization.

Experimental testing confirms that the system performs efficiently with accurate predictions and low response times. Overall, the proposed system provides a cost-effective, scalable, and intelligent solution for phishing detection. Future enhancements may include real-world datasets, deep learning integration, and browser-based deployment for advanced cybersecurity protection.

REFERENCES

[1] E. S. Gualberto, R. T. De Sousa, T. P. De B. Vieira, J. P. C. L. Da Costa, and C. G. Duque, “From Feature Engineering and Topic Models to Enhanced Prediction Rates in Phishing Detection,” *IEEE Access*, vol. 8, pp. 76368–76385, 2020.
[2] APWG, “Phishing Activity Trends Report: Unifying the Global Response to Cybercrime,” 2024.
[3] U. M. Alhaji, S. E. Adewumi, and V. I. Yemi-Peters, “Classification of Phishing Attacks Using Machine Learning Algorithms: A Systematic Literature Review,” *Journal of*

Advances in Mathematics and Computer Science, vol. 40, no. 1, pp. 26–44, 2025.

[4] N. Q. Do, A. Selamat, O. Krejcar, E. Herrera-Viedma, and H. Fujita, "Deep Learning for Phishing Detection: Taxonomy, Current Challenges and Future Directions," *IEEE Access*, vol. 10, pp. 36429–36463, 2022.

[5] K. H. M. Gularte et al., "Safeguarding the V2X Pathways: Exploring the Cybersecurity Landscape Through Systematic Review," *IEEE Access*, vol. 12, pp. 72871–72895, 2024.

[6] M. Nanda, M. Saraswat, and P. K. Sharma, "Enhancing Cybersecurity: A Review and Comparative Analysis of CNN Approaches for URL-Based Phishing Detection," *e-Prime - Advances in Electrical Engineering, Electronics and Energy*, vol. 8, 2024.

[7] S. Zhuo, R. Biddle, Y. S. Koh, D. Lottridge, and G. Russello, "SoK: Human-Centered Phishing Susceptibility," *ACM Transactions on Privacy and Security*, vol. 26, no. 3, pp. 1–27, 2023.

[8] M. F. Alghenaim, G. Alkawsi, and C. R. Barnhart, "The State of the Art in AI-Based Phishing Detection: A Systematic Literature Review," in *Current and Future Trends on AI Applications*, Springer, 2025, pp. 431–458.

[9] A. Ozcan, C. Catal, E. Donmez, and B. Senturk, "A Hybrid DNN–LSTM Model for Detecting Phishing URLs," *Neural Computing and Applications*, vol. 35, no. 7, pp. 4957–4973, 2023.

[10] M. Bahaghighat, M. Ghasemi, and F. Ozen, "A High-Accuracy Phishing Website Detection Method Based on Machine Learning," *Journal of Information Security and Applications*, vol. 77, 2023.