Phishing URL Detection Intelligent Phishing Detection System Using URL and Email Pattern Analysis

Merin c shajan[1], Umaira. K. U2, Amarnath M[3]
[1,2] *III B. Sc Digital and Cyber Forensic Science, Department of Digital and Cyber Forensic Science, Nehru Arts and Science College, Coimbatore, Tamil Nadu, India*
[3] *Assistant Professor, Department of Digital and Cyber Forensic Science, Nehru Arts and Science College, Coimbatore, Tamil Nadu, India*

***Abstract*—The rapid growth of digital communication platforms, online banking, cloud services, and e-commerce systems has significantly increased exposure to phishing attacks. Phishing remains one of the most prevalent cyber threats, exploiting deceptive URLs and fraudulent emails to steal sensitive information such as login credentials, financial data, and personal details. Traditional blacklist-based detection mechanisms are limited in their ability to identify newly generated or zero-day phishing URLs, as they rely on previously reported malicious links. To address these limitations, this paper proposes an Intelligent Phishing Detection System that integrates URL structure analysis and email pattern recognition using machine learning techniques. The proposed system employs a multi-layered architecture that collects URLs and email data from various input sources, performs preprocessing and feature extraction, and applies hybrid machine learning models for classification. URL-based features include lexical characteristics, domain age, WHOIS and DNS attributes, HTTPS certificate validation, and URL shortening detection. Email-based features analyze header anomalies, sender domain authentication (SPF, DKIM, DMARC concepts), urgency keywords, and embedded hyperlink mismatches. Multiple algorithms such as Logistic Regression, Random Forest, Support Vector Machine, Gradient Boosting, and deep learning models (LSTM/CNN) are utilized to enhance detection accuracy and zero-day threat identification. Experimental evaluation demonstrates high accuracy, strong recall, and improved detection performance compared to traditional blacklist systems. The proposed approach provides proactive, adaptive, and scalable phishing protection, significantly enhancing cybersecurity resilience across enterprise and financial environments.**

***Index Terms*—Phishing URL Detection, Machine Learning, Email Pattern Analysis, Cybersecurity.**

I. INTRODUCTION

The rapid expansion of digital technologies, online banking, cloud computing, and e-commerce platforms has significantly increased global connectivity and convenience. However, this digital transformation has also led to a dramatic rise in cyber threats, particularly phishing attacks. Phishing is a form of social engineering attack in which adversaries impersonate legitimate organizations to deceive users into revealing sensitive information such as login credentials, financial details, and personal data. Modern phishing campaigns have become highly sophisticated, utilizing deceptive URLs, domain spoofing, HTTPS certificates, URL shortening services, and well-crafted email templates that closely resemble trusted brands.

The impact of phishing attacks is especially severe in sectors such as banking, e-commerce, and enterprise environments. Financial institutions face substantial monetary losses and reputational damage due to credential theft and fraudulent transactions. E-commerce platforms are targeted to compromise customer accounts and payment information, while enterprises suffer from data breaches, ransomware infections, and unauthorized access to confidential information. The increasing use of remote work systems and digital communication tools has further expanded the attack surface, making phishing one of the most prevalent cybersecurity threats worldwide.

Traditional phishing detection mechanisms primarily rely on blacklist-based approaches, where known malicious URLs are stored in databases and blocked upon detection. Although effective against previously identified threats, blacklist systems fail to detect newly

generated or zero-day phishing URLs that have not yet been reported. Additionally, attackers frequently modify domain names and hosting servers to evade such static detection methods.These limitations highlight the urgent need for intelligent phishing detection systems that leverage URL structure analysis, email pattern recognition, and machine learning techniques. By analyzing behavioral and structural characteristics rather than relying solely on predefined signatures, intelligent systems can proactively detect emerging phishing threats and enhance overall cybersecurity resilience.

II. PROBLEM STATEMENT

Phishing attacks have evolved significantly in complexity and sophistication, making traditional detection mechanisms increasingly ineffective. One of the primary concerns is the growing sophistication of phishing URLs. Attackers now employ techniques such as domain spoofing, homograph attacks (using visually similar characters), subdomain manipulation, URL shortening services, and HTTPS certificate abuse to make malicious links appear legitimate. These deceptive tactics make it difficult for users—and even conventional security systems—to distinguish between genuine and fraudulent websites.

Another major limitation lies in signature-based and rule-based detection systems. Traditional anti-phishing solutions rely on predefined patterns, blacklists, or manually crafted rules to identify malicious URLs and emails. While these approaches are effective against previously known threats, they lack adaptability. Cybercriminals frequently generate new domains, slightly modify URL structures, or automate phishing kit deployment, rendering static signatures obsolete. As a result, rule-based systems often fail to keep pace with the dynamic nature of phishing campaigns.

A critical challenge is the detection of zero-day phishing attacks—newly created malicious URLs that have not yet been reported or added to blacklists. These attacks can remain active for hours or days before being flagged, causing significant financial and data losses during that window of exposure. Since blacklist-based systems depend on prior knowledge, they inherently struggle to detect such emerging threats in real time.

Additionally, email spoofing and domain impersonation further complicate phishing detection. Attackers manipulate email headers, forge sender identities, and create lookalike domains to impersonate trusted institutions. Display name mismatches, subtle domain alterations, and compromised legitimate accounts increase the difficulty of identifying malicious emails.

III. SYSTEM ARCHITECTURE

The proposed Intelligent Phishing Detection System is designed as a multi-layered architecture that integrates URL analysis and email pattern inspection within a unified detection framework. The architecture ensures real-time monitoring, structured feature extraction, intelligent classification, and automated alerting.

3.1 Overall System Design

Data Input Sources (URLs and Emails):The system collects data from multiple entry points, including web browsers, email servers, enterprise mail gateways, and API integrations. URLs are extracted from emails, websites, SMS messages (optional extension), and user submissions. Email data includes headers, body content, attachments, and embedded hyperlinks.

Processing Layers:Raw inputs undergo preprocessing, including text normalization, tokenization, removal of irrelevant characters, and feature extraction. URLs are decomposed into structural elements, while email contents are parsed for metadata and textual features. These structured features are then forwarded to the detection engine.

Detection Engine:The core intelligence module applies machine learning algorithms to classify inputs as legitimate or phishing. It combines URL-based features and email behavioral indicators to improve prediction accuracy. Hybrid models may include supervised classifiers and anomaly detection techniques to identify both known and zero-day threats.

Alert Generation Module:When a phishing attempt is detected, the system generates real-time alerts with risk scores. It can block access to malicious URLs, quarantine suspicious emails, and notify system administrators. Logs are stored for audit and forensic analysis.

3.2 URL Analysis Module

The URL Analysis Module focuses on identifying malicious characteristics embedded within web links.

Lexical Feature Extraction:The system analyzes structural properties such as URL length, number of subdomains, presence of special characters (@, -, //), suspicious keywords (login, verify, secure), IP-based URLs, and abnormal path depth.

Domain-Based Analysis: Domain age, registration details, top-level domain (TLD) patterns, and brand impersonation attempts are examined to detect suspicious domain usage.

WHOIS and DNS Features: WHOIS information is analyzed to verify domain registration date, registrar details, and ownership consistency. DNS-based features include TTL values, IP reputation, and hosting server location.

URL Shortening Detection: Shortened URLs are expanded and evaluated to uncover hidden malicious destinations often used to evade detection.

3.3 Email Pattern Analysis Module

Header Analysis: The system inspects email headers for anomalies such as forged sender addresses, mismatched reply-to fields, suspicious routing paths, and irregular timestamps.

Sender Domain Verification (SPF, DKIM, DMARC Concepts):

Authentication mechanisms are checked to validate whether the sender domain is authorized. Failures or inconsistencies in SPF, DKIM, or DMARC records increase the phishing risk score.

Email Body Keyword Analysis: Natural Language Processing (NLP) techniques identify urgency-based phrases (e.g., “urgent action required,” “account suspended”), financial prompts, and credential request patterns commonly used in phishing campaigns.

Attachment and Embedded Link Scanning: Attachments are scanned for malicious scripts or executables, while embedded hyperlinks are extracted and passed to the URL Analysis Module for further evaluation.

IV. FEATURE ENGINEERING

Feature engineering plays a critical role in enhancing the accuracy and reliability of the proposed Intelligent Phishing Detection System. By extracting meaningful attributes from URLs and email content, the system transforms raw input data into structured indicators that can be effectively analyzed by machine learning models. Proper feature selection reduces noise, improves classification performance, and strengthens zero-day detection capability.

4.1 URL-Based Features

URL-based features focus on identifying structural and behavioral patterns commonly associated with phishing links.

URL Length: Phishing URLs are often unusually long to obscure the actual domain name and mislead users. Excessively lengthy URLs containing multiple parameters or encoded characters are treated as high-risk indicators.

Number of Subdomains: Attackers frequently create multiple subdomains (e.g., login.verify.bank.example.com) to imitate legitimate services. A high number of subdomains can indicate suspicious intent.

Presence of Special Characters (@, -, //):Special characters such as “@” can redirect browsers to different domains, while multiple hyphens or repeated slashes may signal manipulation. These characters are extracted and quantified as lexical features.

IP Address Usage Instead of Domain Name: Legitimate organizations rarely use raw IP addresses in official URLs. The presence of numeric IP-based URLs is often associated with malicious or temporary hosting infrastructure.

HTTPS Certificate Validation: Although many phishing sites now use HTTPS, inconsistencies in certificate issuer, domain mismatch, or recently issued certificates can serve as useful detection features.

4.2 Email-Based Features

Email-based features analyze structural and textual characteristics commonly found in phishing emails.

Suspicious Subject Lines:Subjects containing phrases like “Account Suspended,” “Verify Immediately,” or “Security Alert” are flagged for further inspection.

Urgency Keywords:Natural Language Processing (NLP) techniques detect urgency-driven terms designed to pressure recipients into immediate action.

Mismatch Between Display Name and Email Address: If the visible sender name does not match the actual email domain, it raises a strong phishing indicator.

Embedded Hyperlink Mismatch Detection: The system compares the displayed hyperlink text with the

actual URL destination. A mismatch between visible text and underlying link is considered highly suspicious.

By combining these URL and email features, the system enhances predictive accuracy and strengthens resilience against evolving phishing strategies.

V. METHODOLOGY

The proposed Intelligent Phishing Detection System follows a structured methodology that integrates data collection, preprocessing, model training, and performance evaluation. The approach ensures robustness, scalability, and adaptability to evolving phishing techniques.

5.1 Dataset Collection

A diverse and well-balanced dataset is essential for accurate phishing detection. The system utilizes multiple data sources to ensure comprehensive coverage of both malicious and legitimate samples.

Phishing Datasets (Public Repositories): Phishing URLs are collected from publicly available cybersecurity repositories and threat intelligence feeds. These datasets contain labeled phishing links from real-world attack campaigns, including domain-spoofing and credential-harvesting websites.

Legitimate URL Datasets: To ensure balanced classification, legitimate URLs are gathered from trusted sources such as popular search engine listings, verified corporate websites, and open web directories. This prevents model bias toward phishing-only detection.

Email Phishing Corpora: Email datasets containing both phishing and legitimate messages are used for training the email pattern analysis module. These corpora include metadata, subject lines, body content, and embedded links, enabling comprehensive feature extraction.

5.2 Data Preprocessing

Raw data must be transformed into structured features before model training.

Cleaning and Normalization: Irrelevant characters, duplicate entries, and corrupted records are removed. URLs and email text are normalized to lowercase to maintain consistency.

Tokenization of URLs and Email Content: URLs are decomposed into meaningful components such as domain, path, query parameters, and special characters. Email text is tokenized into words or phrases for keyword and semantic analysis.

Handling Missing Values: Incomplete or missing attributes are either imputed using statistical techniques or removed to maintain dataset integrity and prevent model bias.

5.3 Model Training and Validation

The dataset is divided using a standard split ratio such as 70-15-15 (training, validation, testing) or 80-20 (training and testing). The training set is used to build the model, the validation set fine-tunes hyperparameters, and the test set evaluates final performance.

Cross-validation techniques, such as k-fold cross-validation, are applied to improve generalization and prevent overfitting. Performance metrics including accuracy, precision, recall, and F1-score are calculated to assess effectiveness.

5.4 Algorithms Used

Multiple machine learning algorithms are implemented to compare performance and enhance detection reliability:

Logistic Regression: Provides baseline binary classification with probabilistic outputs.

Random Forest: Offers robust performance through ensemble learning and handles feature interactions effectively.

Support Vector Machine (SVM): Efficient in high-dimensional feature spaces for precise classification.

Gradient Boosting: Enhances prediction accuracy by sequentially correcting model errors.

Deep Learning (LSTM / CNN): LSTM models capture sequential patterns in URLs and email text, while CNN models extract contextual text features for advanced phishing detection.

This multi-algorithm approach ensures accurate classification and resilience against both known and emerging phishing threats.

VI. IMPLEMENTATION DETAILS

The Intelligent Phishing Detection System is implemented using scalable and widely supported technologies to ensure flexibility, real-time performance, and ease of integration with existing digital infrastructures. The system follows a modular

client–server architecture, enabling seamless deployment in enterprise environments.

Programming Language (Python): Python is selected as the primary development language due to its simplicity, extensive library ecosystem, and strong support for machine learning and cybersecurity applications. Its readability and cross-platform compatibility make it suitable for rapid prototyping and large-scale deployment.

Libraries (Scikit-learn, TensorFlow, NLTK): Scikit-learn is used for implementing classical machine learning models such as Logistic Regression, Random Forest, Support Vector Machine, and Gradient Boosting. TensorFlow supports deep learning architectures including LSTM and CNN models for advanced text and URL pattern analysis. NLTK (Natural Language Toolkit) is utilized for natural language processing tasks such as tokenization, stop-word removal, keyword extraction, and semantic analysis of email content.

Web Framework (Flask/Django): Flask or Django is employed to build the web-based interface and RESTful APIs. The framework handles user input (URLs and email data), processes requests, and displays detection results in real time. It also manages authentication, logging, and administrative controls.

Database System: A relational database system such as PostgreSQL or MySQL stores extracted features, trained model outputs, phishing logs, user activity records, and alert histories. This ensures structured data management and supports forensic analysis.

API Integration for Browser/Email Clients: The system provides APIs that can be integrated into browser extensions and email clients. When a user clicks a link or receives an email, the API sends the URL or message content to the detection engine, enabling instant phishing verification and automated alert generation.

VII. RESULTS AND PERFORMANCE EVALUATION

The performance of the Intelligent Phishing Detection System was evaluated using labeled phishing and legitimate datasets for both URL and email analysis. Multiple machine learning models were tested, and their performance was measured using standard classification metrics to ensure reliability and robustness.

Accuracy: Accuracy represents the overall percentage of correctly classified instances (phishing and legitimate). The proposed system achieved high accuracy due to effective feature engineering and hybrid modeling approaches. However, since phishing detection is a security-critical task, accuracy alone is not sufficient to measure performance.

Precision, Recall, and F1-Score: Precision measures the proportion of correctly identified phishing instances among all predicted phishing cases. High precision indicates fewer false alarms.

Recall (Sensitivity) evaluates the system's ability to detect actual phishing attacks. A high recall ensures minimal missed threats.

F1-Score provides a balanced measure by combining precision and recall, making it particularly useful for imbalanced datasets where phishing samples may be fewer than legitimate ones.

The system demonstrated strong recall and F1-scores, indicating effective detection of malicious URLs and phishing emails while maintaining low false alerts.

ROC Curve Analysis: Receiver Operating Characteristic (ROC) curve analysis was used to evaluate the trade-off between True Positive Rate (TPR) and False Positive Rate (FPR). The Area Under the Curve (AUC) value was high, reflecting strong discriminative capability of the trained models.

False Positive and False Negative Rates: A low False Positive Rate reduces inconvenience to legitimate users, while a low False Negative Rate ensures phishing attempts are not overlooked. The proposed model maintained a balanced optimization of both metrics.

Comparison with Traditional Blacklist Systems: Compared to conventional blacklist-based detection methods, the intelligent system demonstrated superior performance, particularly in detecting zero-day phishing URLs. Unlike static blacklists, the proposed model identifies structural and behavioral anomalies, providing proactive and adaptive phishing protection.

VIII. CONCLUSION

Phishing attacks continue to evolve in complexity, exploiting both URL manipulation techniques and sophisticated email spoofing strategies to deceive users and compromise sensitive information. Traditional blacklist and signature-based detection systems are increasingly inadequate against dynamic

and zero-day phishing threats. This research presented an Intelligent Phishing Detection System that integrates URL analysis and email pattern recognition using machine learning and deep learning techniques.

The proposed system leverages lexical, domain-based, DNS, and HTTPS-related URL features along with comprehensive email header and content analysis to improve detection accuracy. By employing supervised learning models such as Logistic Regression, Random Forest, Support Vector Machine (SVM), and Gradient Boosting, alongside deep learning approaches like LSTM and CNN for textual pattern recognition, the system demonstrates strong classification performance. Experimental results indicate high accuracy, improved recall, and reduced false positive rates when compared with traditional blacklist-based systems.

The integration of intelligent feature engineering and adaptive model training enables proactive identification of zero-day phishing attempts. This approach enhances cybersecurity resilience for banking institutions, enterprises, e-commerce platforms, and email service providers. Future improvements may include real-time browser extensions, federated learning for distributed threat intelligence sharing, and AI-driven adaptive risk scoring systems.

Overall, the proposed intelligent phishing detection framework provides a scalable, efficient, and robust solution to mitigate phishing threats in modern digital ecosystems.

REFERENCES

[1] APWG, “Phishing Activity Trends Report,” 2023. [Online]. Available: https://apwg.org

[2] Verizon, “Data Breach Investigations Report (DBIR),” 2023.

[3] UCI Machine Learning Repository, “Phishing Websites Dataset,” University of California, Irvine, 2017.

[4] Canadian Institute for Cybersecurity, “CICIDS2017 Dataset,” University of New Brunswick, 2017.

[5] O. K. Sahingoz, E. Buber, O. Demir, and B. Diri, “Machine learning based phishing detection from URLs,” Expert Systems with Applications, vol. 117, pp. 345–357, 2019.

[6] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, “Beyond blacklists: Learning to detect malicious web sites from suspicious URLs,” in Proc. ACM SIGKDD Int. Conf., 2009.

[7] Google, “Safe Browsing Transparency Report,” 2022.

[8] Microsoft, “Digital Defense Report,” 2023.

[9] I. Fette, N. Sadeh, and A. Tomasic, “Learning to detect phishing emails,” in Proc. Int. World Wide Web Conf. (WWW), 2007.

[10] National Institute of Standards and Technology (NIST), “Digital Identity Guidelines,” 2020.