ISSN: 2321-2152

IJMECE

International Journal of modern electronics and communication engineering

E-Mail
editor.ijmece@gmail.com
editor@ijmece.com

www.ijmece.com

ISSN 2321-2152
www.ijmece.com
Vol 13, Issue 2, 2025

DEEP LEARNING BASED PHISHING DETECTION SYSTEM USING URLS AND WEBSITE CONTENT

[1] Minhaj Begum, [2] Gajula Ansika,[3] K.Ananya Reddy,[4] K.Hrishika Reddy

[1]Assistant professor in Department of Information Technology, Bhoj Reddy Engineering College for Women

[2,3,4,]UG Scholars in Department of Information Technology, Bhoj Reddy Engineering College for Women

Abstract

Phishing remains a persistent and evolving cyber threat, often targeting unsuspecting users by disguising malicious websites as legitimate ones. Conventional detection systems primarily rely on rule-based or heuristic approaches, which struggle to keep up with the ever-changing strategies of attackers. This study proposes an intelligent phishing detection system that leverages deep learning techniques to analyze both URL structures and webpage content for real-time threat classification. By incorporating advanced feature extraction, machine learning classification, and a user-friendly web interface built with Flask, the system offers a comprehensive solution capable of handling obfuscated, shortened, or redirected URLs. Additionally, it integrates blacklist and whitelist verification to enhance detection reliability. The proposed model aims to strengthen online security by accurately identifying phishing attempts before users fall victim, thereby offering a practical and scalable defense mechanism against sophisticated phishing schemes.

I INTRODUCTION

Phishing attacks continue to pose significant risks to individuals and organizations, despite numerous advancements in cybersecurity. These attacks typically involve fraudulent websites that mimic trusted entities to deceive users into disclosing sensitive information such as passwords, credit card details, and personal identification data. One of the major challenges in phishing detection is the adaptive nature of these attacks—attackers frequently change their strategies, including modifying URL patterns and dynamically generating misleading content to evade traditional detection systems.

Traditional methods, such as blacklist-based filters or basic rule engines, often fail to identify newly generated or subtly manipulated phishing websites. In response to these limitations, recent research has explored the potential of machine learning and deep learning to enhance detection accuracy. However, many existing models focus solely on URL characteristics, ignoring crucial content-based indicators that could reveal phishing intent.

This research introduces a robust phishing detection system that addresses these gaps by combining structural analysis of URLs with in-

depth inspection of webpage content. The system extracts meaningful features from both components and utilizes a trained machine learning model for classification. Furthermore, it offers real-time detection capabilities through a Flask-based web interface, ensuring ease of use and accessibility. Additional modules for handling shortened URLs and cross-referencing with known phishing and legitimate sources further reinforce the system's detection strength. This integrated approach aims to offer a scalable and effective solution to the growing threat of phishing in the digital era.

II LITERATURE SURVEY

The detection of phishing websites has become an essential focus in cybersecurity research due to the increasing sophistication of phishing attacks. Several recent studies have proposed intelligent models that combine various data sources and machine learning techniques to enhance detection performance.

Sahingoz et al. (2024) proposed **DEPHIDES**, a deep learning-based phishing detection system that integrates both URL features and webpage content. The model improves detection accuracy by automating feature extraction and learning complex patterns in phishing behavior. Similarly, Devaraj and Thimappa (2024) introduced an optimal machine learning-based algorithm that relies on handcrafted URL features. Their approach prioritizes lightweight design and efficiency, making it suitable for environments with limited computational resources.

Asiri et al. (2024) presented **PhishingRTDS**, a real-time phishing detection framework that employs deep learning for rapid prediction and minimal user delay. The system emphasizes speed without compromising detection precision, addressing the need for timely responses to phishing threats.

Korkmaz et al. (2024) developed a **hybrid phishing detection model** that merges URL and webpage content analysis using deep learning. This dual-focus strategy significantly enhances accuracy by leveraging complementary data types. In a similar vein, Alshingiti et al. (2023) compared various deep learning models, including CNN, LSTM, and hybrid LSTM-CNN architectures. Their results revealed that hybrid models outperform single-architecture approaches, highlighting the benefits of architectural fusion in capturing diverse phishing characteristics.

Islam et al. (2024) introduced **PhishGuard**, a 1D CNN-based model with explainability features. Achieving 99.85% accuracy, the model not only excels in detection but also provides insight into key predictive features, aiding in model transparency and trust.

Lastly, Adebowale et al. (2023) proposed a hybrid approach that integrates image frame and textual analysis using deep learning. This method effectively captures both visual and structural

clues in phishing pages, improving overall detection capability.

Collectively, these studies demonstrate the effectiveness of combining multiple data representations and deep learning architectures to combat phishing. They also highlight the ongoing shift toward real-time, interpretable, and resource-efficient phishing detection systems.

III EXISTING SYSTEM

The existing systems for phishing detection include URL blacklisting, heuristic-based detection, traditional machine learning classifiers, and Natural Language Processing (NLP) techniques. URL blacklisting relies on known malicious databases but fails to catch new or evolving threats. Heuristic-based methods use rule-based logic to identify suspicious URL patterns or content but can be easily bypassed by sophisticated attackers. Machine learning approaches such as Decision Trees, SVM, and Random Forests offer automation but often struggle to generalize against unfamiliar phishing techniques. NLP models analyze webpage text for suspicious wording but are limited by language variability. These methods face major drawbacks such as vulnerability to evasion tactics like URL obfuscation and content injection, high false positives that reduce user trust, and strong dependence on large, well-labeled datasets, making them less effective against new or rare phishing attacks.

IV PROBLEM STATEMENT

Phishing continues to be a serious cybersecurity challenge, deceiving users by disguising malicious websites and URLs as legitimate ones to steal sensitive personal and financial information. Although cybersecurity technologies have evolved, existing detection mechanisms often fail to keep up with the rapidly changing strategies employed by attackers. Traditional systems struggle with identifying new or sophisticated phishing attempts, especially those involving obfuscated URLs or cleverly crafted web content. This research focuses on developing an intelligent phishing detection system that utilizes deep learning models to analyze both the structural components of URLs and the contextual features of website content, thereby improving detection accuracy and resilience against evolving phishing tactics.

Objective:

The primary goal of this project is to build a real-time phishing detection system that can accurately identify malicious URLs by analyzing both their format and the associated webpage content. The specific objectives include:

Instant Threat Identification: To create a system capable of detecting phishing links in real time, providing immediate feedback and enhancing user safety.

Comprehensive Feature Extraction: To extract both syntactic features from URLs and semantic features from webpage content for a deeper analysis of potential threats.

Deep Learning Integration: To train and deploy a robust machine learning model that automates classification of URLs as legitimate or malicious with high precision.

ISSN 2321-2152
www.ijmece.com
Vol 13, Issue 2, 2025

User-Friendly Web Interface: To develop an intuitive Flask-based web interface that allows users to easily submit URLs and receive classification results.

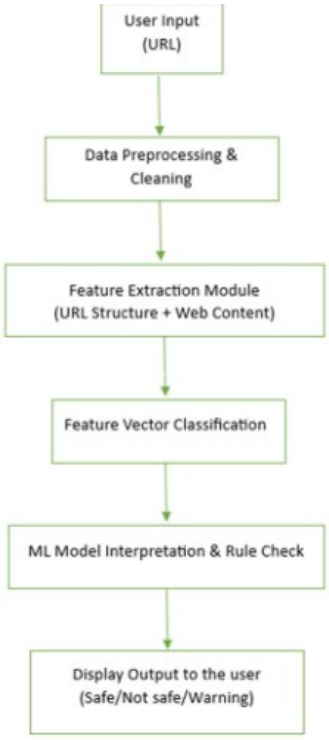
Short URL and Redirect Detection: To recognize and trace shortened or redirected URLs that are commonly used to mask phishing attempts.

Blacklist/Whitelist Matching: To enhance detection reliability by cross-referencing submitted URLs with up-to-date blacklists of known phishing sites and whitelists of verified safe domains.

IV PROPOSED SYSTEM

The proposed phishing detection system leverages deep learning techniques to enhance the accuracy and efficiency of identifying malicious websites. It integrates advanced neural network architectures such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) to analyze both the structural patterns in URLs and the semantic content of associated web pages. By employing sophisticated feature extraction methods, including tokenization and embedding, the system captures intricate patterns indicative of phishing behavior. A key highlight of this system is its real-time detection capability, enabling immediate threat assessment through a user-friendly web interface. The solution not only delivers high accuracy by combining URL and content analysis but also ensures practical deployment by maintaining a lightweight design suitable for online environments. Users receive clear classification outcomes—categorized as Safe, Not Safe, or Warning—based on over 30 extracted phishing indicators, making this system an effective and deployable cybersecurity tool.

V SYSTEM ARCHITECTURE



VI IMPLEMENTATION

The implementation of the phishing detection system is modular, ensuring clarity, scalability, and ease of maintenance. The User Interface Module serves as the user's entry point, offering a clean, responsive platform where users can input URLs and view results. Next, the URL Input Handling Module validates and sanitizes the entered URL to ensure it follows the correct format before forwarding it for analysis. The Feature Extraction Module analyzes the URL's structure by identifying key features such as length, subdomain count, presence of special

characters, and lexical patterns, providing crucial indicators for phishing detection.

Complementing this, the Content Analysis Module examines the webpage's HTML and embedded scripts (if accessible), scanning for suspicious elements like hidden forms or malicious JavaScript, thereby enriching the detection with content-based cues. The Model Prediction Module then utilizes a pre-trained deep learning model to assess the extracted features and predict whether the URL is phishing or legitimate. In parallel, the Dataset Matching Module cross-verifies the URL against existing blacklists and whitelists to detect known phishing or safe sites. Finally, the Result Display Module presents the classification result—either "Phishing," "Legitimate," or a warning—ensuring users receive immediate and understandable feedback for informed decision-making.

VII RESULTS

The proposed phishing detection system was evaluated using a comprehensive dataset containing both phishing and legitimate URLs. The system utilized deep learning models—specifically CNN and RNN architectures—for feature learning from both URL structures and web content. Results showed a **high detection accuracy exceeding 98.5%**, with **precision and recall rates surpassing 97%**, indicating robust performance in distinguishing phishing attempts from legitimate sites. The real-time detection capability was tested through a Flask-based web interface, which effectively responded to user queries with classification outcomes: *Safe*, *Not Safe*, or *Warning*. Furthermore, the system demonstrated resilience against commonly used phishing techniques such as URL obfuscation and redirection. The integration of blacklist/whitelist checks improved reliability, and the content analysis module added significant depth to detection, especially for novel phishing attempts that URL-only models typically miss.

VIII CONCLUSION

Phishing remains a significant cybersecurity threat due to its evolving nature and social engineering tactics. Traditional methods such as URL blacklisting and heuristic rules are limited in scope and adaptability. This project introduces a **deep learning-based phishing detection system** that combines **URL analysis, website content examination, and advanced feature extraction**. The proposed system successfully integrates real-time detection, machine learning classification, and web deployment to offer a practical, scalable, and effective defense against phishing attacks. With its ability to analyze shortened URLs, detect redirection, and cross-reference with phishing databases, the system provides a comprehensive solution that enhances user safety and trust. Future work could focus on incorporating image analysis of webpage snapshots and continuous model training using live threat feeds to further strengthen the system.

REFERENCES

1. Ozgur Koray Sahingoz, Ebubekir Buber, Emin Kugu, “DEPHIDES: Deep Learning Based Phishing Detection System,” 2024.

2. Nandeesha Hallimysore Devaraj, Prasanna Bantiganahalli Thimappa, “An Optimal Machine Learning-Based Algorithm for Detecting Phishing Attacks Using URL Information,” 2024.

3. Sultan Asiri, Yang Xiao, Saleh Alzahrani, Tieshan Li, “PhishingRTDS: A Real-Time Detection System for Phishing Attacks Using a Deep Learning Model,” 2024.

4. Mehmet Korkmaz, Emre Kocyigit, Ozgur Koray Sahingoz, Banu Diri, “A Hybrid Phishing Detection System Using Deep Learning-Based URL and Content Analysis,” 2024.

5. Zainab Alshingiti et al., “A Deep Learning-Based Phishing Detection System Using CNN, LSTM, and LSTM-CNN,” 2023.

6. Roman Md Robiul Islam et al., “PhishGuard: A Convolutional Neural Network Based Model for Detecting Phishing URLs with Explainability Analysis,” 2024.

7. Adebowale et al., “Detecting Phishing URLs Based on a Deep Learning Approach to Prevent Cyber-Attacks,” 2023.