AI-POWERED PHISHING EMAIL & MESSAGE DETECTION SYSTEM

K Shashidhar[1], Gangadhar M[2], Lokesh[3], Varshith S Patil[4], Mr. K.B. Dinesh Babu[5]
[1234] Department of Computer Science and Engineering, Don Bosco Institute of Technology, Bengaluru, India
[5] Assistant Professor, Department of Computer Science and Engineering, Don Bosco Institute of Technology, Bengaluru, India
Visvesvaraya Technological University, Belagavi, Karnataka, India

ABSTRACT: The rapid growth of digital communication has increased exposure to phishing attacks that exploit user trust through deceptive language, malicious links, fake websites, and shortened URLs. This paper presents an AI-Powered Phishing Email & Message Detection System that combines Artificial Intelligence, Machine Learning, and Natural Language Processing to identify suspicious digital messages. The proposed pipeline performs text pre-processing, feature extraction, URL and keyword analysis, and classification using machine-learning and deep-learning models including Support Vector Machine, Random Forest, Naïve Bayes, CNN, and LSTM. An ensemble decision mechanism is used to classify incoming content into Safe, Suspicious, or Phishing categories. The system also provides a user-facing interface for real-time analysis and risk scoring, while MySQL is used to maintain analysis logs for monitoring and security auditing. The approach is intended to overcome limitations of static rule-based filtering and provide an accessible, proactive layer of phishing protection.

Keywords—Artificial Intelligence, Phishing Detection, Machine Learning, Natural Language Processing, SVM, Random Forest, CNN, LSTM, Risk Scoring, Cybersecurity.

I. INTRODUCTION

Digital communication has become an essential part of personal, academic, and organizational activities. However, the same communication channels provide attackers with opportunities to deceive users and obtain sensitive information. Phishing messages are often designed to appear legitimate and may imitate trusted organizations, use urgent language, or direct users to fraudulent websites. The project report identifies traditional rule-based filtering and manual inspection as insufficient against evolving phishing strategies.

The proposed system addresses this challenge through an AI-based detection pipeline. It analyzes message content, URLs, keywords, sender-related information, and message structure before producing a risk-oriented classification. The design supports Email, SMS, and other text-based inputs and aims to provide real-time feedback without requiring extensive cybersecurity expertise.

The principal objectives are to develop an AI-based phishing detector, apply NLP to understand message content, identify suspicious URLs and keywords, classify messages into Safe, Suspicious, or Phishing categories, and provide a user-friendly interface with persistent analysis records.

II. LITERATURE SURVEY

Previous work reviewed in the project report covers data-driven phishing detection, deep-learning comparisons, real-time URL detection, hybrid machine-learning/deep-learning monitoring, and NLP-based phishing detection. These studies indicate that modern phishing defense benefits from combining multiple information sources rather than depending on a single static rule.

Data-driven approaches provide proactive and layered protection but may require substantial

computational resources. Comparative deep-learning studies evaluate architectures for email detection but do not necessarily provide a single optimized solution. Real-time URL detection is effective for malicious links but can miss social-engineering attacks that contain no obviously malicious URL. Earlier hybrid systems provide real-time monitoring but can suffer from limited accuracy, while NLP-based methods provide strong text analysis but may face computation and overfitting concerns.

TABLE I
COMPARATIVE ANALYSIS OF REVIEWED PHISHING DETECTION APPROACHES

Sl. No.	Approach	Year	Main Advantage	Main Limitation
1	AI-powered data-driven phishing detection	2025	Proactive, multi-layered defense	High computational demand
2	Deep-learning phishing email comparison	2024	Comprehensive architecture evaluation	No single optimized solution
3	Real-time phishing URL detection	2023	Fast malicious-link analysis	May miss social engineering without URLs
4	Hybrid ML and DL phishing monitoring	2022	Real-time and user-friendly	Limited intelligence/accuracy
5	NLP-based phishing email detection	2021	High prediction and scalability	Computation and overfitting risk

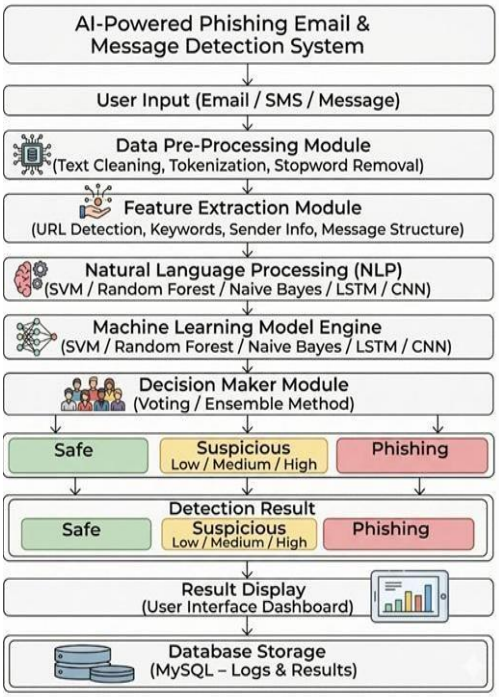
The literature review in the project report therefore motivates a combined architecture in which text analysis, URL inspection, multiple predictive models, and ensemble decision-making are integrated into one detection workflow.

III. SYSTEM DESIGN AND PROPOSED METHOD

The proposed architecture is organized as a sequential analytical pipeline. The user first submits an email, SMS, or other text-based message. The pre-processing layer cleans the text, tokenizes it, and removes stopwords. The feature-extraction layer then examines URLs, suspicious keywords, sender information, and message structure.

The processed features are supplied to the NLP and model engine. The report specifies Support Vector Machine, Random Forest, Naïve Bayes, CNN, and LSTM models. Their predictions are combined through a voting or ensemble decision mechanism. The final output is mapped to Safe, Suspicious, or Phishing, with a corresponding risk score displayed through the user interface. Analysis records are stored in MySQL for historical monitoring and auditing.

Fig. 1. System architecture of the proposed AI-powered phishing detection system.



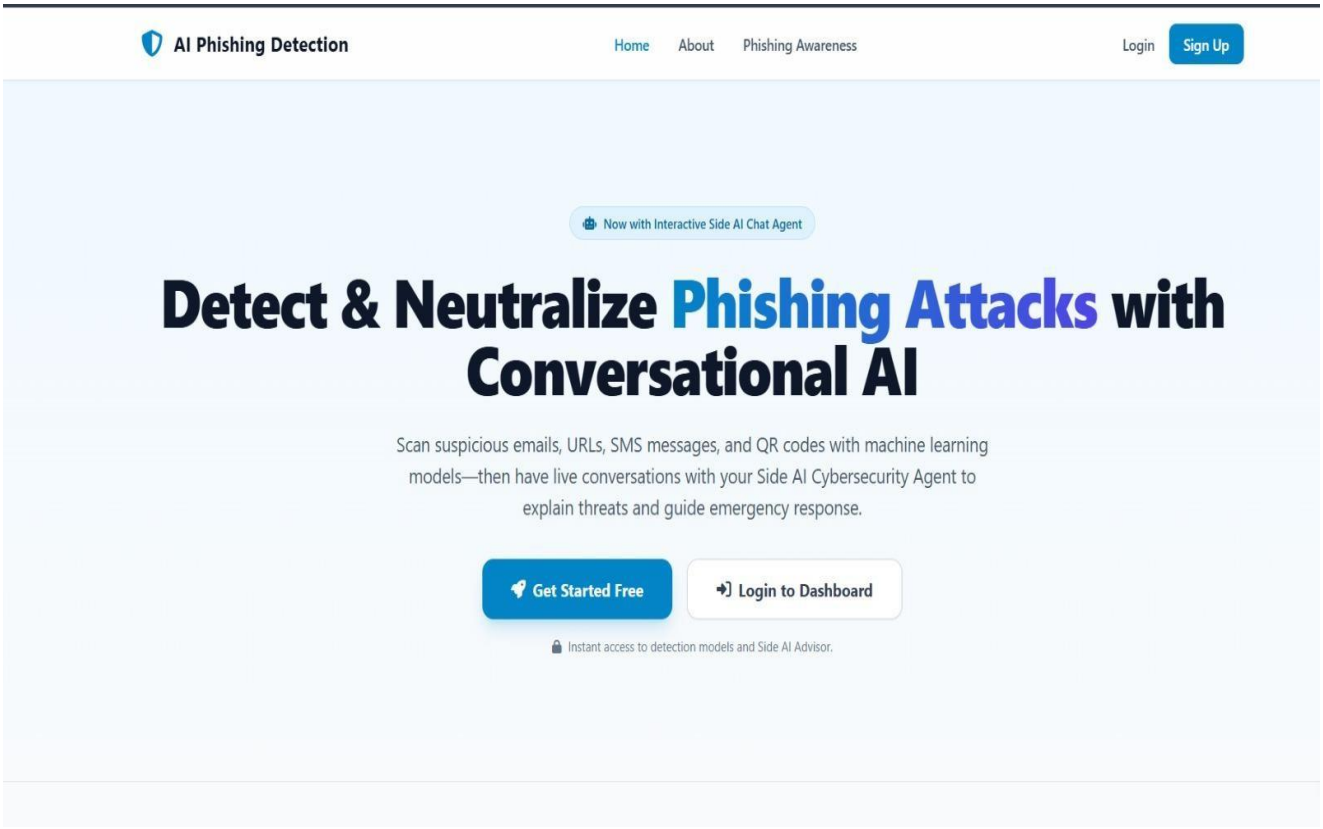
IV. IMPLEMENTATION

The implementation uses Python as the primary programming language and Flask as the web framework connecting the interface with the detection engine. Scikit-learn/TensorFlow are identified for machine-learning and deep-learning models, while NLTK/SpaCy support NLP operations such as tokenization and stopword removal. Pandas and NumPy support data manipulation and feature-processing operations. MySQL is used to store user logs, processed message data, and results.

During preprocessing, HTML tags, unnecessary special characters, and punctuation are removed where appropriate. Tokenization converts the message into analyzable units, and stopword removal reduces terms that do not contribute to threat identification. Feature extraction then detects suspicious URLs, high-risk terms such as “Urgent,” “Account Suspended,” and “Verify Now,” together with sender and structural anomalies.

For model training and classification, the project report describes SVM, Random Forest, and Naïve Bayes as machine-learning classifiers, with CNN and LSTM used for more complex pattern recognition. An ensemble voting mechanism combines model outputs. The web interface accepts manual message input or file uploads, returns a risk score and classification, and logs analyzed messages in the database.

Fig. 2. Home Page shown in the project report.



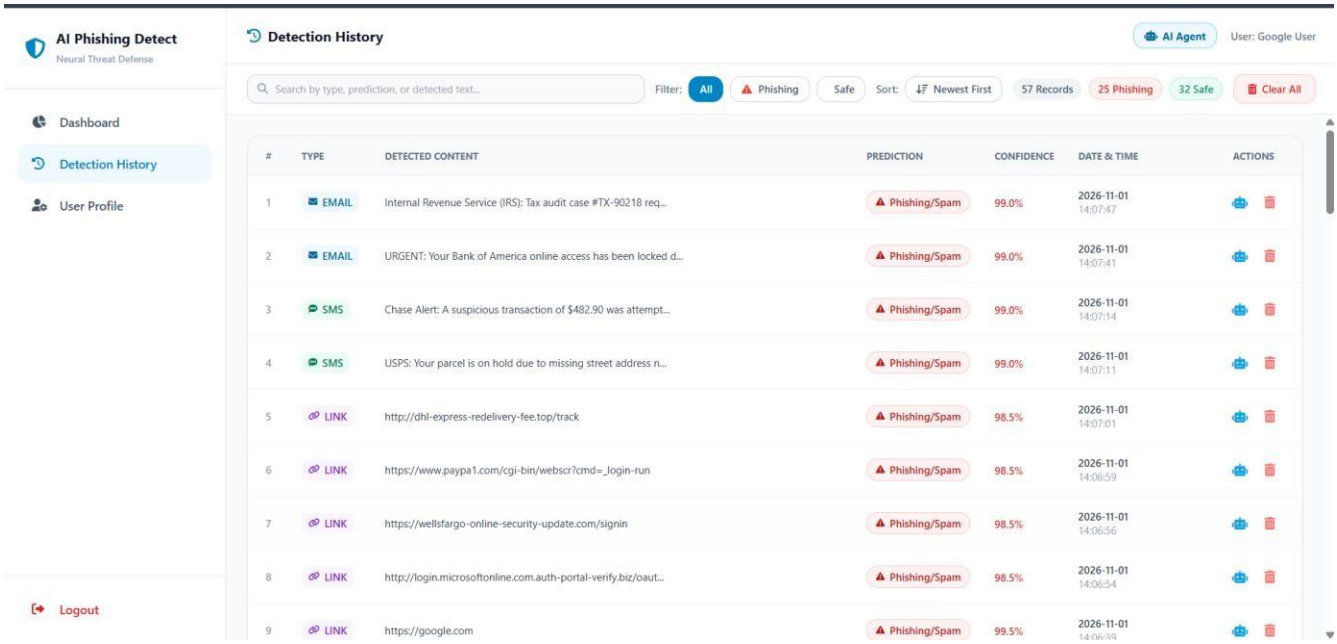
V. FINDINGS AND RESULTS

The implemented workflow demonstrates a complete path from message submission to classification: input is cleaned and tokenized, threat-related features are extracted, predictive models analyze the processed representation, and an ensemble decision produces the final class. The project report identifies the three output classes as Safe, Suspicious, and Phishing and specifies real-time risk scoring as a user-facing result.

The training-model section of the report shows a Random Forest training workflow using TF-IDF features, a train-test split, model fitting, prediction, accuracy evaluation, classification reporting, and confusion-matrix generation. The trained model and vectorizer are serialized for later use. Since the supplied report does not provide a numerical accuracy table or measured test-set results, this paper does not invent performance values.

From the literature and system design, the principal finding is that combining content-level NLP analysis with URL, keyword, metadata, and ensemble-based classification can address a broader range of phishing indicators than a single-rule or URL-only detector.

Fig. 3. Results of phishing shown in the project report.



VI. CONCLUSION AND FUTURE SCOPE

The AI-Powered Phishing Email & Message Detection System provides an integrated approach to detecting deceptive digital messages. By combining AI, machine learning, NLP, feature extraction, multiple classifiers, real-time risk scoring, and MySQL-based logging, the proposed system moves beyond purely rule-based filtering. The system is designed to help users identify fraudulent content before interaction and to improve cybersecurity awareness.

Future work identified in the project includes localized and culturally nuanced phishing detection, blockchain-based sender-authentication support, enterprise-scale deployment, browser or email-client plugin integration, and continuous learning from newly confirmed phishing patterns. These extensions can improve adaptability and provide more seamless protection in real-world communication environments.

REFERENCES

[1] S. Author et al., “AI-Powered Phishing Detection: A Data-Driven Cybersecurity Approach.”

[2] A. Researcher and B. Kumar, “Advancing Phishing Email Detection.”

[3] M. Patel, R. Singh, and K. Sharma, “PhishHaven: Real-Time AI Phishing URLs Detection System.”

[4] P. Verma and S. Gupta, “Detection of E-Mail Phishing Attacks Using Machine Learning and Deep Learning.”

[5] J. Smith and T. Roy, “Phishing Email Detection Using Natural Language Processing.”

Editorial note: *The reference entries above reproduce the bibliographic information available in the supplied Phase-1 report; complete publication metadata was not provided there*

Copyright & License:

© Authors retain the copyright of this article. This work is published under the Creative Commons Attribution 4.0 International License (CC BY 4.0), permitting unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.