



AI BASED EMAIL SECURITY SYSTEM

vhandkar P.S¹, samiksha basampure²

¹ Faculty Guide & Supervisor | ² Dattakala Group Of Institutions Faculty Of Engineering Swami - Chincholi tal. Daund dist. Pune, India

Corresponding author: basampuresamiksha@gmail.com

ARTICLE ID IJREK-2026-00140	ARTICLE TYPE Research Article	RESEARCH AREA Artificial Intelligence	PUBLICATION Volume 1, Issue Issue 1 (2026)
RECEIVED 21 October 2026	REVISED 25 October 2026	ACCEPTED 22 October 2026	PUBLISHED 28 October 2026

Abstract

Email is a central channel for personal and organizational communication, but it is also exploited to distribute spam, phishing links, fraudulent requests, and malicious attachments. This paper proposes an AI-based email security system that analyzes message text, sender-related indicators, URLs, and attachment metadata to estimate whether an email is legitimate, spam, phishing, or potentially malicious. The proposed workflow combines preprocessing, feature extraction, machine-learning classification, risk scoring, and explanations for flagged messages. A layered design is recommended because no single indicator is reliable in every case: language patterns, sender authentication signals, suspicious URL characteristics, and attachment properties should be considered together. The system should issue a warning or quarantine recommendation rather than automatically treating every flagged message as malicious. Evaluation can measure precision, recall, F1-score, false-positive rate, detection latency, and robustness to changing attack patterns. As no implemented prototype or measured dataset results were provided, this paper presents a proposed design and evaluation plan rather than claiming numerical performance.

Keywords: Email Security; Phishing Detection; Spam Classification; Machine Learning; Natural Language Processing; Cybersecurity. 2

I. INTRODUCTION

Email is widely used for education, business, banking, and personal communication. Its broad adoption makes it an attractive target for cybercriminals who send deceptive messages to steal credentials, deliver malware, or persuade recipients to transfer money. Traditional email filters use blocklists, fixed rules, sender reputation, and known signatures. These methods are useful but may struggle with new domains, altered wording, and attacks designed to avoid simple rules. A. Problem Statement An email security solution must distinguish legitimate messages from spam, phishing, impersonation attempts, and messages containing suspicious links or attachments. False positives can hide important messages, while false negatives may allow harmful content to reach users. The system should combine multiple signals, explain why an email was flagged, and allow review of uncertain messages. B. Aim and Objectives The aim is to design an AI-based system that supports email threat detection using machine learning and content analysis. Objectives are to preprocess email text; extract linguistic, sender, URL, and attachment-related features; classify messages into risk categories; generate a risk score and explanation; support safe warning or quarantine workflows; and define measurable evaluation criteria. C. Significance The system may help users recognize suspicious messages earlier and reduce reliance on manual inspection. It is a decision-support layer, not a guarantee that every malicious message will be detected. Human review, updated threat intelligence, and secure email practices remain important.

II. LITERATURE REVIEW / RELATED WORK

A. Rule-Based Filtering Conventional filters use known malicious senders, URL blocklists, spam signatures, and heuristic rules. They are fast and interpretable, but depend on current indicators and carefully maintained rules. Attackers can change domains, wording, and delivery infrastructure to bypass static checks. B. Machine Learning for Spam Detection Methods such as Naive Bayes, logistic regression, support vector machines, and tree-based classifiers can learn patterns from labelled email examples. Text representations such as TF-IDF can identify terms and phrase patterns associated with spam or phishing. Performance depends on representative

training data, reliable labels, and appropriate validation. C. NLP and URL Analysis Natural language processing can identify urgency, requests for credentials, unusual payment instructions, and persuasive patterns. URL analysis can consider IP-address links, misleading subdomains, redirects, and mismatches between displayed link text and destination. These indicators are not proof of an attack individually and should be assessed together. D. Research Gap A practical system should combine message content with sender and link indicators, explain why an email was flagged, protect message privacy, and support updates as attacks evolve. Evaluation should report false positives and false negatives rather than relying only on overall accuracy.

III. PROPOSED SYSTEM AND METHODOLOGY

A. System Architecture The proposed system contains an email input or gateway, preprocessing module, feature-extraction layer, classification engine, risk-scoring module, and user-facing alert or quarantine workflow. Email content is parsed carefully, and the classifier combines text features with available sender, authentication, URL, and attachment metadata. High-risk messages can be quarantined for review; uncertain messages should receive a warning rather than an irreversible automatic action. Incoming Email Text and Metadata Preprocessing Feature Extraction: Text • Sender • URLs • Attachments AI / Machine-Learning Classifier Risk Score and Explanation Deliver • Warn • Quarantine for Review

Incoming Email
Text and Metadata Preprocessing
Feature Extraction: Text • Sender • URLs • Attachments
AI / Machine-Learning Classifier
Risk Score and Explanation
Deliver • Warn • Quarantine for Review

Fig. 1. Proposed architecture of the AI-based email security system.

IV. IMPLEMENTATION DETAILS

TABLE I

Component	Possible implementation	Purpose
User interface	HTML, CSS, JavaScript or React	Display alerts and review status
Backend API	Python with Flask or FastAPI	Coordinate analysis
Text features	TF-IDF or text embeddings	Represent message content
Classifier	Naive Bayes, logistic regression, tree model	Classify email risk
Security signals	URL checks and SPF/DKIM/DMARC	Add sender and link evidence
Storage	Database with access controls	Store minimal metadata

V. RESULTS AND ANALYSIS

TABLE II

Test	Condition	Expected behaviour
T1	Normal email from known contact	Classifies as likely legitimate, subject to checks
T2	Credential-harvesting link	Flags likely phishing and explains indicators
T3	Bulk promotional message	Classifies as spam when supported by evidence
T4	Unknown sender with conflicting signals	Shows warning or routes to review
T5	Previously unseen suspicious URL	Combines signals and avoids certainty
T6	Classifier service unavailable	Uses safe fallback and records service error

VI. LIMITATIONS AND FUTURE SCOPE

A. Limitations The design depends on training-data quality and current security signals. Attackers may use novel wording, compromised legitimate accounts, URL shorteners, or apparently harmless attachments. Legitimate messages can resemble phishing, producing false positives. Automated detection should be combined with email authentication, endpoint protection, user education, and human review. B.

Future Scope Future work may evaluate transformer-based text classifiers, multilingual phishing detection, real-time mail-server integration, sandboxed attachment analysis, explainable alerts, and privacy-preserving learning. A dashboard could show threat trends and let analysts review reported messages. Continuous monitoring should measure model drift and support controlled updates.

VII. CONCLUSION

This paper proposes an AI-based email security system for identifying spam, phishing, and potentially malicious email content through machine-learning classification and security indicators. The design covers preprocessing, feature extraction, risk scoring, explanatory alerts, privacy safeguards, and evaluation using precision, recall, F1-score, false-positive rate, and latency. It assists users and security teams rather than guaranteeing perfect detection or replacing established controls. Its effectiveness must be established through implementation and testing on representative email data.

REFERENCES

[1] S. Abu-Nimeh, D. Nappa, X. Wang, and S. Nair, "A Comparison of Machine Learning Techniques for Phishing Detection," Proceedings of the Anti-Phishing Working Groups 2nd Annual eCrime Researchers Summit, 2007, pp. 60–69, doi: 10.1145/1299015.1299021.

[2] T. Fawcett, "An Introduction to ROC Analysis," Pattern Recognition Letters, vol. 27, no. 8, pp. 861–874, 2006, doi: 10.1016/j.patrec.2005.10.010.

[3] S. Garera, N. Provos, M. Chew, and A. D. Rubin, "A Framework for Detection and Measurement of Phishing Attacks," Proceedings of the ACM Workshop on Recurring Malcode, 2007.

[4] C. Whittaker, B. Ryner, and M. Nazif, "Large-Scale Automatic Classification of Phishing Pages," Proceedings of the Network and Distributed System Security Symposium, 2010. Note: This paper follows the compact structure of the supplied sample. Verify references and add institution name/logo if required. Include measured results only after implementation and testing.

Publication archival note: Official scholarly contribution published under CC BY 4.0 International License. Permanent record on <https://www.ijrek.org> | Article ID: IJREK-2026-00140