INTERNATIONAL JOURNAL OF RESEARCH, EXPLORATION & KNOWLEDGE (IJREK)
Official Scholarly Journal Portal: **https://www.ijrek.org** | DOI: **https://doi.org/10.5281/zenodo.23484412**

AI BASED EMAIL SECURITY SYSTEM

Dr.Bere Sir[1], Rutuja Katore[2]

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

Corresponding author: **katorerutuja4@gmail.com**

ARTICLE ID	ARTICLE TYPE	RESEARCH AREA	PUBLICATION
IJREK-2026-00126	Research Article	Artificial Intelligence	Volume 1, Issue Issue 1 (2026)
RECEIVED	**REVISED**	**ACCEPTED**	**PUBLISHED**
19 October 2026	02 November 2026	20 October 2026	03 November 2026

Abstract

Email remains a major communication channel for individuals and organizations, but it is frequently abused for phishing, spam, malware delivery, impersonation, and social engineering. Traditional filtering methods often rely on fixed rules, known signatures, or sender reputation and may struggle with new or carefully crafted attacks. This paper presents a conceptual AI-Based Email Security System that analyzes email content and available message metadata to identify suspicious messages and support safer handling. The proposed approach combines text preprocessing, machine-learning classification, phishing indicators, URL and attachment checks, and a risk-based decision layer. Emails can be categorized as legitimate, spam, phishing, or requiring further review. The design also includes user feedback, explainable warnings, privacy-conscious processing, and periodic model evaluation. The system is intended as a defensive support tool rather than a guarantee that every malicious email will be detected. Its effectiveness would depend on representative training data, secure implementation, and continuous testing.

Keywords: *Artificial Intelligence; Email Security; Phishing Detection; Spam Filtering; Machine Learning; Cybersecurity*

2. PROBLEM STATEMENT AND OBJECTIVES

The main problem is how to identify suspicious email messages accurately while keeping legitimate messages accessible. A useful system should detect common threats, reduce avoidable false alarms, explain why a message was flagged, and protect the email data used for analysis. 2.1 Objectives • To classify email messages into legitimate, spam, phishing, or review-required categories. • To analyze message text, sender-related indicators, links, and available metadata. • To identify common phishing language, suspicious URLs, and potentially unsafe attachments. • To provide understandable risk warnings and support user reporting of incorrect classifications. • To evaluate detection quality using precision, recall, F1-score, and false-positive rate. • To handle email content with appropriate privacy, access-control, and retention safeguards. 2.2 Scope The proposed system focuses on defensive analysis of email messages in an educational or organizational setting. It may be integrated with an email client or a mail-processing service. The paper describes a conceptual design; it does not claim that a deployed system, measured performance, or a security audit has already been completed.

3. RELATED WORK

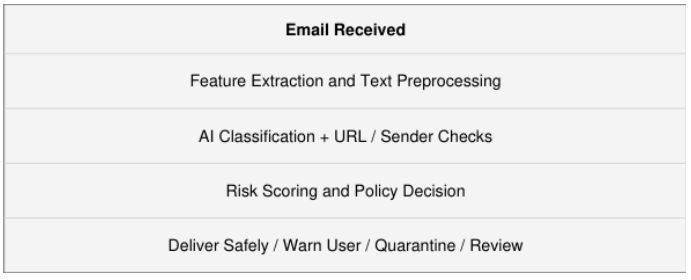
Email filtering has traditionally used rule-based checks, blocklists, sender reputation, and signature-based detection. Machine-learning approaches extend these methods by classifying messages using features learned from labelled examples. Text classification can identify patterns in subject lines and message bodies, while URL analysis and sender metadata can contribute additional signals. Research and operational guidance from cybersecurity organizations also emphasize layered defenses, user awareness, and the need to treat suspicious links and attachments carefully.

3.1 RESEARCH GAP

No single indicator is sufficient for every email threat. A message may contain convincing language but a suspicious link, or it may be written in a way that avoids simple keyword rules. A practical system therefore needs to combine multiple signals, manage uncertainty, explain alerts, and be tested on new and diverse examples rather than relying only on training accuracy.

4. PROPOSED METHODOLOGY

The proposed workflow starts when an email is received or submitted for inspection. The system extracts permitted features, normalizes the text, analyzes URLs and metadata, applies an AI classifier, and combines the outputs into a risk category. High-risk messages can be quarantined or flagged according to the organization's policy; uncertain messages can be sent for user or administrator review.



4.2 System Architecture and Activity Flow

4.1 PROCESSING STEPS

• Input: Receive email subject, body, sender fields, URLs, and permitted attachment metadata. • Preprocessing: Normalize text, remove irrelevant formatting, and extract tokens or language features. • Feature analysis: Examine suspicious phrases, sender inconsistencies, URL characteristics, and attachment indicators. • Classification: Apply a trained machine-learning or language-based classifier to estimate threat categories. • Risk decision: Combine model output with security rules and confidence thresholds. • Response: Deliver, label, quarantine, or request review based on configured policy. • Feedback: Allow authorized users to report false positives or missed threats for later evaluation.

4.2 SYSTEM ARCHITECTURE AND ACTIVITY FLOW

Email Received Feature Extraction and Text Preprocessing AI Classification + URL / Sender Checks Risk Scoring and Policy Decision Deliver Safely / Warn User / Quarantine / Review

Fig. 1. Conceptual workflow of the AI-based email security system.

4.3 FUNCTIONAL MODULES

TABLE I

Module	Purpose
Email input and parsing	Extracts relevant message fields in a controlled manner.
Text analysis	Prepares message text and identifies language patterns associated with threats.
URL and sender checks	Evaluates suspicious links and inconsistencies in available sender information.
AI classifier	Predicts a category or risk score from extracted features.
Alert and response	Displays warnings and applies delivery, quarantine, or review policies.
Feedback and reporting	Supports authorized review and monitoring of system performance.

5. TOOLS AND TECHNOLOGIES

• Python: Suitable for prototyping text processing and machine-learning workflows. • Machine-learning libraries: Scikit-learn or comparable frameworks can support classification and evaluation. • Natural Language Processing: Text normalization, tokenization, and feature extraction for email content. • URL analysis: Feature checks and reputable threat-intelligence sources, where available and permitted. • Database or logging layer: Stores limited classification outcomes and audit events under access controls. • Testing: A labelled dataset and repeatable test cases for benign and malicious email examples.

6. SECURITY, PRIVACY, AND LIMITATIONS

Email bodies and attachments may contain personal, confidential, or organizational information. Processing should therefore follow data-minimization principles, restrict access, protect data in transit and at rest, and define retention periods. Full message content should not be retained in diagnostic logs unless there is a documented need and suitable safeguards. External URL-scanning services should be assessed before message data or private links are shared with them. AI classification can produce false positives by flagging legitimate messages and false negatives by missing malicious messages. Attackers may adapt wording, use compromised trusted accounts, hide destinations behind redirects, or exploit weaknesses in the surrounding email platform. Model output should not be treated as conclusive proof. High-impact actions should use policy controls and appropriate human review, and the system should support safe recovery when a message is incorrectly quarantined.

7. EVALUATION AND DISCUSSION

Evaluation should use a separate test set that reflects realistic email variation and avoids leakage between training and testing data. Results should be reported by threat category and should include precision, recall, F1-score, and false-positive rate. Precision measures how many flagged emails are truly malicious in the labelled set; recall measures how many malicious emails are detected. Both matter because missed phishing messages and excessive false alarms create different risks. Tests should include legitimate business messages, bulk spam, credential-phishing attempts, suspicious links, attachment-related indicators, and unfamiliar wording. Evaluation should also measure processing time, user understanding of warnings, and performance drift over time. Since no experimental results are provided here, the system's effectiveness remains to be established through implementation and documented testing.

8. APPLICATIONS AND FUTURE SCOPE

Potential applications include educational institutions, small businesses, enterprise email gateways, and personal email clients. Future development may include multilingual detection, improved explainability, privacy-preserving model updates, integration with threat-intelligence feeds, and adaptive detection of new attack patterns. Further work should compare different classifiers, document dataset composition, test against adversarial examples, and assess usability with representative users. 9. Conclusion An AI-Based Email Security System can support email protection by combining message-text analysis, sender and URL indicators, machine-learning classification, and risk-based response policies. A layered design is preferable to reliance on one model or a fixed set of keywords. Privacy controls, clear alerts, human review for uncertain cases, and ongoing evaluation are essential for responsible use. The proposed system is a conceptual framework, and its real-world reliability must be demonstrated through implementation, independent testing, and continuous monitoring.

REFERENCES

[1] National Institute of Standards and Technology (NIST), Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100-1, Jan. 2023. doi: 10.6028/NIST.AI.100-1.

[2] Cybersecurity and Infrastructure Security Agency (CISA), "Recognize and Report Phishing," cybersecurity awareness guidance.

[3] Anti-Phishing Working Group (APWG), Phishing Activity Trends Reports, periodic industry reports.

[4] Scikit-learn developers, "Working With Text Data," Scikit-learn User Guide, documentation for text feature extraction and classification.

[5] National Institute of Standards and Technology (NIST), Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile, NIST AI 600-1, Jul. 2024. doi: 10.6028/NIST.AI.600-1.

Publication archival note: Official scholarly contribution published under CC BY 4.0 International License. Permanent record on **https://www.ijrek.org** | Article ID: IJREK-2026-00126