© MAY 2026 | IRE Journals | Volume 9 Issue 11 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I11-1717944

Email Header Analysis for Digital Forensics Using Machine Learning

DEEPA B[1], DR. BALAMURUGAN S[2]
[1]Scholar Department of Computer Science & Information Technology Jain (Deemed-to-be University), Bangalore, Karnataka, India
[2]Professor & Research mentor Department of Computer Science & Information Technology Jain (Deemed-to-be University), Bangalore, Karnataka, India

Abstract- ***Email communication remains the dominant vector for advanced cyber-attacks, including phishing, spoofing, and Business Email Compromise (BEC). Conventional email security mechanisms rely predominantly on content-based analysis encompassing Natural Language Processing (NLP), keyword filtering, and signature-based detection which are increasingly inadequate against modern adversarial techniques such as clean-text phishing, image-based payloads, and AI-generated deceptive messages. This paper presents a novel forensic-aware, machine learning-driven framework that shifts the analytical focus from email body content to Simple Mail Transfer Protocol (SMTP) header metadata. Email headers encode verifiable forensic information relay paths (Received fields), originating IP addresses, timestamp sequences, and authentication outcomes (SPF, DKIM, DMARC) that collectively represent a tamper-resistant record of an email's transmission behavior. Unlike body content, header metadata is structurally constrained and considerably more difficult for attackers to consistently manipulate across multiple relay nodes. The proposed framework introduces a comprehensive feature engineering pipeline that extracts temporal, network, topological, and authentication-level attributes from SMTP headers. These features are processed through an ensemble of machine learning models Random Forest (RF), Isolation Forest, and XGBoost enabling both supervised classification of known attack patterns and unsupervised detection of novel anomalies. A critical contribution of this research is the integration of forensic traceability with automated detection: the system reconstructs email transmission paths and preserves evidentiary artifacts suitable for digital forensic investigation and legal proceedings. Experimental evaluations on datasets derived from SpamAssassin and PhishTank repositories demonstrate that the proposed ensemble model achieves an F1-score of approximately 0.969 and an AUC-ROC of 0.983, representing a 5–8% improvement over content-only baseline models. False positive rates are simultaneously reduced from 11.2% to 3.1%, ensuring operational reliability in enterprise environments. This research establishes a scalable, intelligent, and forensically defensible paradigm for combating sophisticated email-based cyber-threats.***

Keywords- ***Anomaly Detection, Digital Forensics, Email Spoofing, Ensemble Learning, Feature Engineering, Graph-Based Topology, SMTP Header Analysis, Digital Forensics, Phishing Detection***

I. INTRODUCTION

Email communication serves as the backbone of modern digital interaction, enabling seamless information exchange across personal, corporate, and governmental environments. With billions of emails transmitted daily, this medium has become an indispensable component of global communication infrastructure. Regrettably, the same ubiquity that makes email indispensable renders it an exceptionally attractive attack surface for cybercriminals. Phishing, email spoofing, Business Email Compromise (BEC), and malware dissemination campaigns account for a disproportionate share of organizational security breaches worldwide.

A fundamental architectural vulnerability lies in the design of the Simple Mail Transfer Protocol (SMTP), developed in an era when trusted-network assumptions prevailed and security was not a primary design consideration. SMTP lacks native authentication and message integrity mechanisms, permitting adversaries to forge sender identities, manipulate relay paths, and inject malicious content without triggering built-in safeguards. This structural weakness creates compounding challenges for both automated cybersecurity systems and digital forensic investigators tasked with attributing malicious activity.

Traditional countermeasures rule-based filters, signature databases, and NLP-driven classifiers have achieved meaningful efficacy against unsophisticated attack variants. However, the threat landscape has evolved considerably. Contemporary adversaries deploy techniques that systematically undermine content-centric defenses, including: clean-text phishing emails that impersonate legitimate organizational communication without detectable linguistic anomalies; image-based payloads that embed malicious instructions within rasterized images; domain impersonation and homograph attacks exploiting visually similar Unicode characters; and AI-generated messages whose semantic coherence renders them indistinguishable from authentic correspondence to NLP classifiers.

In contrast, the email header a structured, machine-generated record of an email's transmission journey encodes forensic artifacts that are fundamentally more difficult to spoof coherently across multiple autonomous relay systems. The header contains the originating IP address, a sequential chain of Mail Transfer Agent (MTA) relay entries (Received fields), timestamp sequences at each relay, and the results of authentication protocols including Sender Policy Framework (SPF), DomainKeys Identified Mail (DKIM), and Domain-based Message Authentication, Reporting, and Conformance (DMARC). Each of these elements contributes a verifiable data point that, in aggregate, characterizes the authenticity and integrity of an email's path from origin to destination.

Despite this forensic richness, automated machine learning analysis of SMTP header metadata remains conspicuously underutilized. Existing systems employ rule-based SPF/DKIM validation without harnessing the broader pattern-recognition capabilities of modern machine learning. Furthermore, no existing framework satisfactorily bridges the operational gap between real-time threat detection and post-incident forensic investigation functions that are treated as entirely separate domains in current practice.

This research addresses these deficiencies through the design and experimental validation of an integrated, forensic-aware machine learning framework. The system performs comprehensive SMTP header feature extraction, applies an ensemble of supervised and unsupervised learning models to classify and detect anomalous email behavior, and simultaneously preserves forensic evidence in a format suitable for downstream digital investigation. By uniting detection intelligence with forensic readiness, this work proposes a substantive advancement over the current state of the art in email security.

II. PROBLEM STATEMENT

Email-based cyber-attacks particularly phishing, spoofing, and Business Email Compromise continue to circumvent existing detection mechanisms at alarming rates. The 2024 Verizon Data Breach Investigations Report identified email as the initial access vector in over 36% of breaches involving social engineering, underscoring the urgency of more robust detection methodologies.

Existing detection systems are characterized by two fundamental limitations. First, their heavy dependence on email body content renders them susceptible to adversarial evasion techniques that produce clean, contextually plausible text. Second, they fail to leverage the rich forensic metadata embedded within SMTP headers routing behavior, authentication outcomes, and temporal transmission patterns that are far more resistant to consistent manipulation across independent relay nodes.

Furthermore, there exists a critical operational gap between detection-oriented systems and forensic investigation platforms. When a phishing campaign is identified, investigators face significant challenges in reconstructing the transmission chain, identifying originating infrastructure, and preserving admissible evidence functions that an integrated framework should support natively but that current systems do not provide.

The research problem addressed by this work is therefore three-dimensional: (1) the inadequate exploitation of SMTP header metadata in automated threat detection, (2) the absence of adaptive machine learning-based anomaly detection applied to header-

level routing behavior, and (3) the lack of unified frameworks coupling real-time detection with automated forensic evidence preservation.

III. RESEARCH OBJECTIVES

This research is guided by the following primary objectives:

- RO1: To design and implement a comprehensive SMTP header parsing and feature engineering pipeline capable of extracting temporal, network, topological, and authentication-level attributes from raw email header data at enterprise scale.
- RO2: To develop and evaluate an ensemble machine learning framework integrating Random Forest, Isolation Forest, and XGBoost for the detection of phishing and spoofing attacks using header-derived features.
- RO3: To incorporate graph-based modeling of SMTP relay topologies for the identification of abnormal routing patterns indicative of adversarial manipulation.
- RO4: To design a forensic evidence preservation module that reconstructs email transmission chains and stores evidentiary artifacts compliant with digital forensic investigation standards.
- RO5: To empirically compare the proposed framework against content-based baselines and state of the art models, demonstrating measurable improvements in detection accuracy, F1-score, and false positive reduction.

IV. RESEARCH QUESTIONS AND HYPOTHESES

A. Research Questions

- RQ1: To what extent does SMTP header metadata improve phishing and spoofing detection accuracy relative to content-only machine learning baselines?
- RQ2: Which categories of header-derived features temporal, network, topological, or authentication-based contribute most significantly to classification performance?
- RQ3: How effectively can an ensemble machine learning approach detect previously unseen (zero-day) email attack patterns through unsupervised anomaly detection on SMTP header data?
- RQ4: Can an integrated detection-forensics framework provide legally and operationally viable evidentiary reconstruction of malicious email transmission paths?

B. Hypotheses

- H1: Machine learning models trained on SMTP header features will achieve statistically significant improvements (>=5%) in F1-score over models trained exclusively on email body content.
- H2: Ensemble models combining supervised and unsupervised components will outperform individual classifiers in both detection accuracy and false positive reduction.
- H3: Graph-based topological features derived from SMTP relay chains will enable the detection of routing-based attacks that are invisible to content-centric analysis.

V. LITERATURE REVIEW

A systematic analysis of 25 peer-reviewed studies published between 2021 and 2025 reveals a research community increasingly oriented toward AI-driven email security, yet one that remains predominantly content-centric in its analytical focus. The following subsections characterize the major research streams and their respective limitations.

A. Content-Based Detection: NLP and Transformer Architectures

The preponderance of recent literature concentrates on improving classification performance through increasingly sophisticated analysis of email body content. Mohammed et al. (2025) employed TF-IDF vectorization combined with supervised classifiers including Support Vector Machines and Random Forests achieving detection accuracies of approximately 94% on benchmark phishing datasets. These approaches are inherently contingent on the presence of detectable linguistic anomalies, a condition that sophisticated adversaries have learned

to circumvent through AI-assisted message composition.

Transformer based architectures, particularly BERT and domain specific variants, have been applied to email classification with notable success in capturing contextual dependencies (Adebowale et al., 2021). However, their computational demands limit real-time deployability, and their reliance on semantic content renders them equally susceptible to clean-text evasion.

B. Generative AI and Large Language Models in Threat Detection

Ajuluchukwu et al. (2025) explored the application of Generative AI and Large Language Models for detecting malicious emails and deceptive websites. These models demonstrate impressive intent classification capabilities and can generalize across diverse phishing variants. Nevertheless, they operate purely at the semantic level, providing no mechanism for analyzing structural transmission behavior. Consequently, routing-level spoofing attacks which may carry plausible email bodies evade LLM-based classifiers entirely.

C. Authentication Protocol Integration

Several studies, including the multi-layered defense framework of LKTAU et al. (2024), integrate SPF, DKIM, and DMARC validation as primary detection mechanisms. While these protocols provide a valuable baseline, they are implemented as binary pass/fail criteria rather than as features within an adaptive learning system. Sophisticated attackers routinely exploit misconfigured servers and partial authentication failures to bypass these checks.

D. Topology and Header-Based Analysis

Lochin (2025) proposed the STAMP (SMTP Server Topological Analysis by Message Headers Parsing) framework, which analyzes relay topology encoded in email headers. This represents the closest antecedent to the approach proposed in this paper. However, STAMP operates as a rule-based analytical tool without machine learning integration, limiting its scalability and adaptability to evolving attack patterns. No existing study has combined graph-based relay topology analysis with ensemble machine learning and forensic evidence preservation in a unified framework.

E. Comparative Analysis and Research Gaps

Table I presents a structured comparison of representative studies across key dimensions. Three principal gaps emerge: (1) a methodological gap dominant reliance on content features at the expense of header metadata; (2) a contextual gap LLM-based approaches improve semantic understanding but provide no forensic level topology modeling; and (3) an integration gap no existing system unifies real-time detection with automated forensic evidence preservation.

Table I. Comparative Analysis of Representative Email Security Studies (2021–2025)

Ref.	Authors (Year)	Approach	Features Used	Accuracy	Gap
[1]	Lochin (2025)	SMTP Topology (STAMP)	Header routing paths	N/A	No ML integration
[2]	Mohammed et al. (2025)	NLP + ML Classifiers	TF-IDF, text features	~94%	Ignores header metadata

© MAY 2026 | IRE Journals | Volume 9 Issue 11 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I11-1717944

Ref.	Authors (Year)	Approach	Features Used	Accuracy	Gap
[3]	Ajuluchukwu et al. (2025)	Generative AI / LLM	Email intent, context	~96%	No routing/forensic analysis
[4]	LKTAU et al. (2024)	Multi-layer Defense	SPF/DKIM/DMARC rules	~91%	Rule-based, not adaptive
[11]	Adebowale et al. (2021)	ML-based Detection	URL + text features	~95%	No header forensics
Proposed	This Study (2025)	SMTP Header + Ensemble ML + Forensics	Header, temporal, topological, network	~97%*	Addresses all gaps

*Preliminary experimental result; detailed evaluation presented in Section IX.

VI. RESEARCH GAPS

A. Absence of SMTP Header-Centric Machine Learning Systems

Most deployed and published phishing detection systems prioritize email body content. The structured forensic richness of SMTP headers routing paths, timestamp chains, authentication outcomes remain systematically underutilized in the machine learning literature.

B. Lack of Forensic Topology Modeling

While content-based systems improve classification at the semantic level, none adequately model the topological behavior of email transmission. Hop sequence analysis, relay graph construction, and routing anomaly detection are absent from mainstream detection approaches.

C. Insufficient Adaptive Anomaly Detection on Header Data

Existing anomaly detection applications are predominantly applied to content features rather than header-level behavioral patterns. There is no established framework for adaptive, unsupervised anomaly detection on SMTP routing metadata.

D. Missing Unified Detection-Forensics Architecture

A significant operational void exists between detection tools and forensic investigation platforms. Practitioners must currently pivot between separate systems to respond to and investigate email threats, introducing latency, evidence degradation risk, and coordination overhead.

E. Limited Real-World Scalability Demonstrations

Many proposed models are validated only on small, curated benchmark datasets without demonstrating scalability to enterprise-scale email volumes or generalization across diverse SMTP server configurations.

VII. PROPOSED METHODOLOGY

The proposed framework is structured as a modular, scalable pipeline comprising four interdependent stages: (1) Data Acquisition and Preprocessing, (2) Feature Engineering, (3) Ensemble Machine

Learning Classification and Anomaly Detection, and (4) Forensic Evidence Preservation.

A. Data Acquisition and Preprocessing

The framework operates on a hybrid dataset combining publicly available email repositories with controlled experimental captures. Primary data sources include the SpamAssassin public corpus and PhishTank confirmed phishing samples. Supplementary samples are drawn from TREC 2007 Spam Track and curated enterprise email logs to ensure diversity in attack patterns and server configurations.

1. Header Parsing

Raw email files are parsed using Python's standard email library augmented by custom regular expressions to reliably extract key SMTP fields including Received (all relay instances), X-Originating-IP, Date, Message-ID, Return-Path, and Authentication-Results (SPF, DKIM, DMARC). Each field is stored in a structured schema to support downstream feature computation.

2. Hop Sequence Reconstruction

The sequential chain of Received fields is parsed to reconstruct the complete MTA relay path. Each hop entry is decomposed into sending host, receiving host, and timestamp components. Message-ID consistency across hops is verified to detect anomalies such as forged relay insertions or missing intermediate servers.

3. Timestamp Normalization

All per-hop timestamps are converted to Coordinated Universal Time (UTC) to establish a common temporal reference frame. This normalization enables computation of inter-hop delay intervals and identification of temporal paradoxes instances where an email's timestamp appears to precede its transmission a reliable indicator of header manipulation.

B. Feature Engineering

Feature engineering constitutes the analytical core of the framework. Table II enumerates the five feature categories extracted from SMTP header metadata, encompassing a 47-dimensional feature vector per email.

Table II. Feature Categories, Specific Features, and Detection Targets

Feature Category	Specific Features	Detection Target
Temporal	Date-to-first-Received delta, inter-hop delays, negative latency detection	Timestamp spoofing, time-travel anomalies
Network	IP reputation score, SPF/DKIM/DMARC pass/fail, domain mismatch flag	Spoofed sender identity, compromised relays
Topological	Hop count, path length, node centrality, relay loop indicator	Abnormal routing, relay hijacking, BEC paths
Authentication	SPF alignment, DKIM signature validity, DMARC policy compliance	Authentication bypass, header forgery
Structural	Received field count, Message-ID consistency, X-Mailer presence	Header injection, forged MTA hops

Topological features are derived by modeling the SMTP relay chain as a directed graph G = (V, E), where V represents MTA nodes and E represents transmission links. Graph metrics including path length, node degree centrality, and cycle detection are computed using the NetworkX library. Relay loops directed cycles in G constitute strong anomaly indicators associated with mail-bombing and relay hijacking campaigns.

C. Ensemble Machine Learning Models

1. Random Forest – Supervised Classification

Random Forest is deployed as the primary supervised classifier, leveraging its capacity to handle high-dimensional mixed-type feature vectors, quantify feature importance, and resist overfitting through bootstrap aggregation. RF is particularly well-suited to the structured, tabular nature of header-derived features.

2. Isolation Forest – Unsupervised Anomaly Detection

Isolation Forest complements the supervised component by identifying anomalous header configurations absent from the labeled training distribution. This model is essential for detecting zero-day attacks and novel routing manipulations. It isolates anomalies by recursively partitioning feature space; points requiring fewer partitions are flagged as outliers.

3. XGBoost – Gradient Boosting Classification

XGBoost is incorporated to capture complex non-linear interactions among header features through sequential tree boosting. Its regularization mechanisms and computational efficiency make it particularly effective at distinguishing subtle spoofing patterns from legitimate routing behavior. The final ensemble decision is produced through soft-voting aggregating posterior probabilities from all three models, with weights optimized on a held-out validation set.

D. Forensic Evidence Preservation

A dedicated forensic logging module captures and stores all evidentiary artifacts associated with each analyzed email: the complete raw header, extracted feature vectors, model confidence scores, and reconstructed relay chain graph. All artifacts are stored with SHA-256 cryptographic integrity hashes to support chain-of-custody requirements in legal and compliance contexts. Investigators can replay transmission path analysis, verify detection rationale, and export structured evidence reports compatible with standard digital forensic formats.

VIII. EXPERIMENTAL SETUP

A. Development Environment

Implementation is conducted in Python 3.10 within Jupyter Notebook environments, with computationally intensive model training offloaded to Google Colab (NVIDIA T4 GPU, 16 GB VRAM). This configuration provides iterative development flexibility and sufficient computational capacity for large-scale ensemble model training.

B. Libraries and Technologies

- Scikit-learn 1.4: Random Forest, Isolation Forest, cross-validation, and evaluation metrics
- XGBoost 2.0: Gradient boosting classifier with GPU acceleration support
- NetworkX 3.2: Graph construction and topological feature extraction from relay chains
- Pandas 2.1 / NumPy 1.26: Data preprocessing, feature matrix construction, and numerical computation
- Matplotlib / Seaborn: Visualization of feature distributions, ROC curves, and confusion matrices
- Python email / mailparser: Raw email parsing and structured header field extraction

C. Dataset Composition

The experimental dataset comprises 48,000 email samples: 22,500 legitimate (ham) emails from SpamAssassin, 18,000 confirmed phishing emails from PhishTank and TREC 2007, and 7,500 spoofing simulation emails generated through controlled laboratory conditions. The dataset is partitioned using stratified 80/20 train-test splits with five-fold cross-validation to ensure robust performance estimation and guard against class imbalance artifacts. SMOTE-based synthetic oversampling is applied to the minority class during training.

D. Evaluation Metrics

System performance is evaluated on Precision, Recall, F1-Score, False Positive Rate (FPR), and Area Under the Receiver Operating Characteristic Curve (AUC-ROC). Special emphasis is placed on FPR minimization, as false positive classifications disrupt legitimate enterprise communication and erode operator trust in automated systems.

IX. RESULTS AND DISCUSSION

Table III presents comparative performance metrics for the content-only baseline, individual header-based models, and the proposed ensemble. All values represent averages across five cross-validation folds.

© MAY 2026 | IRE Journals | Volume 9 Issue 11 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I11-1717944

Table III. Comparative Model Performance on the SMTP Header Analysis Task

Model	Precision	Recall	F1-Score	FPR	AUC-ROC
Content-Only Baseline (TF-IDF + RF)	0.891	0.883	0.887	0.112	0.921
Random Forest (Header Features)	0.934	0.928	0.931	0.068	0.952
Isolation Forest (Anomaly Detection)	0.911	0.904	0.907	0.089	0.936
XGBoost (Header Features)	0.948	0.942	0.945	0.052	0.967
Proposed Ensemble (RF + IF + XGBoost)	0.971	0.968	0.969	0.031	0.983

A. Impact of Header Metadata on Detection Performance

The most salient finding is the consistent and statistically significant performance advantage of header-based features over the content-only baseline. The proposed ensemble achieves an F1-score of 0.969 compared to 0.887 for the TF-IDF baseline—an improvement of 8.2 percentage points corroborating Hypothesis H1. This improvement is attributable to the complementary nature of header features: while content-based models succeed where linguistic anomalies are present, header-based models succeed precisely where they are absent—against clean-text and AI-generated attacks.

B. Temporal Feature Analysis

Timestamp inconsistencies emerged as the single most discriminative feature category, contributing 31.4% of the Random Forest feature importance score. Spoofed emails in the experimental corpus exhibited characteristic temporal paradoxes: 67.3% of confirmed spoofing samples demonstrated at least one negative inter-hop delay, compared to 0.8% of legitimate emails. These anomalies—arising from improper timestamp forgery across independent relay nodes—are structurally very difficult to eliminate without sacrificing the appearance of legitimate routing behavior, making them exceptionally reliable detection signals.

C. Topological Routing Anomaly Detection

Graph-based relay topology analysis identified structural irregularities including relay loops, anomalous path lengths, and disconnected relay chains in 89.1% of mail-bombing and relay hijacking samples. Relay loops were absent from all 22,500 legitimate email samples, rendering them a near-perfect indicator of specific attack categories. The incorporation of topological features increased AUC-ROC from 0.952 (RF without topology) to 0.983 (proposed ensemble), corroborating Hypothesis H3.

D. Unsupervised Anomaly Detection Performance

The Isolation Forest component, operating exclusively on unlabeled header features, successfully flagged 78.2% of zero-day simulation samples—email attacks crafted using patterns absent from the training distribution. This validates Hypothesis H3 and establishes the framework's capability for detecting previously unseen attack variants without requiring labeled training examples.

E. False Positive Reduction

The proposed ensemble achieves a False Positive Rate of 3.1%, representing a 65.2% reduction relative to the content-only baseline (11.2% FPR). In an enterprise environment processing one million emails daily, this difference equates to approximately 81,000 fewer misclassified legitimate emails per day—a difference that directly impacts operational continuity and user trust.

© MAY 2026 | IRE Journals | Volume 9 Issue 11 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I11-1717944

F. Forensic Reconstruction Accuracy

The forensic evidence preservation module successfully reconstructed complete transmission chain graphs for 96.7% of analyzed emails. The remaining 3.3% corresponded to cases where intermediary relay servers had stripped Received header fields a known limitation. All reconstructed chains were validated against ground-truth relay data for the simulation dataset, confirming structural fidelity. All exported evidence packages passed SHA-256 integrity verification.

X. SYSTEM ARCHITECTURE AND WORKFLOW

The operational pipeline proceeds through five sequential stages:

- Stage 1 – Ingestion: Raw email files are received via API endpoint, IMAP integration, or batch file upload and queued for processing.
- Stage 2 – Header Extraction and Normalization: The parsing module extracts all SMTP header fields, normalizes timestamps to UTC, and reconstructs the hop sequence.
- Stage 3 – Feature Computation: The feature engineering module generates a 47-dimensional feature vector per email encompassing all five feature categories.
- Stage 4 – Ensemble Classification: Feature vectors are simultaneously processed by RF, Isolation Forest, and XGBoost; soft-voting produces a final classification and confidence score.
- Stage 5 – Forensic Logging and Alerting: All evidentiary artifacts are stored with cryptographic integrity hashes; alerts are dispatched to configured SIEM endpoints for emails classified as malicious above a configurable confidence threshold.

XI. ADVANTAGES AND LIMITATIONS

Table IV provides a structured summary of the framework's principal strengths and acknowledged constraints.

Table IV. Summary of System Advantages and Limitations

Advantages	Limitations
Evasion-resistant: header metadata is harder to forge consistently than email body content	Header availability varies across providers; some fields may be stripped by intermediaries
Forensic-grade traceability: reconstructs full transmission chain for post-incident analysis	IP reputation databases require periodic updates; may not reflect newly malicious infrastructure
Ensemble robustness: combining RF, Isolation Forest, and XGBoost reduces individual model bias	Graph-based topology analysis adds computational overhead at high email throughput volumes
Low false positive rate: multi-dimensional features precisely balance precision and recall	Dataset diversity is limited; generalization across non-English SMTP configurations needs validation
Unsupervised anomaly detection identifies zero-day and previously unseen attack variants	Adversarial attackers may progressively craft headers to mimic legitimate relay behavior over time

XII. FUTURE SCOPE

Several promising research directions emerge from this work. First, integration of Graph Neural Networks (GNNs) particularly Graph Attention Networks (GATs) for SMTP topology analysis would enable the framework to learn structural representations of relay graphs directly, capturing higher-order routing anomalies that handcrafted graph features cannot encode.

Second, evaluation against adversarially crafted SMTP headers generated using AI models trained to mimic legitimate routing behavior represents a critical robustness test not addressed in the current work. Incorporating adversarial training into the machine learning pipeline would improve resilience against such adaptive attacks.

Third, real-time deployment integration with enterprise SIEM platforms (e.g., Splunk, Microsoft

Sentinel) and email gateway solutions would validate operational scalability and enable longitudinal performance monitoring in production environments. Fourth, expansion of the forensic evidence module to support automated chain-of-custody documentation compliant with ISO/IEC 27037 and NIST SP 800-86 digital forensic standards would enhance the legal admissibility of preserved evidence artifacts.

Fifth, cross-cultural and multilingual dataset expansion incorporating email traffic from non-English-language server configurations and international ISPs would strengthen model generalization across diverse deployment contexts globally.

CONCLUSION

This paper has presented and empirically validated an intelligent, forensic-aware machine learning framework for email threat detection centered on SMTP header metadata analysis. By redirecting analytical focus from manipulable email body content toward the structured, tamper-resistant forensic record embedded in email headers, the proposed system achieves detection performance that consistently surpasses content-only baselines most critically against the sophisticated evasion techniques that render traditional approaches increasingly inadequate.

The ensemble integration of Random Forest, Isolation Forest, and XGBoost enables the system to simultaneously classify known attack patterns and detect novel anomalies, establishing a detection capability that adapts to the evolving threat landscape. The graph-based modeling of SMTP relay topologies introduces a forensic dimension absent from all comparable systems, enabling both structural anomaly detection and reconstruction of evidentiary transmission chains.

Experimentally, the framework achieves an F1-score of 0.969, an AUC-ROC of 0.983, and a False Positive Rate of 3.1% representing improvements of 8.2, 6.2, and 65.2 percentage points, respectively, over the content-only baseline. The forensic evidence preservation module successfully reconstructs transmission chains for 96.7% of analyzed emails with verified cryptographic integrity.

This work establishes that SMTP header metadata is not merely a diagnostic artifact but a primary detection signal of sufficient richness and reliability to anchor next-generation email security architectures. As adversaries continue to weaponize AI for evasion, the forensic invariants encoded in email headers represent one of the few analytical surfaces that cannot be trivially circumvented. Future research should deepen the exploitation of this surface through graph neural networks, adversarial robustness training, and real-time enterprise-grade deployment systems.

APPENDIX

A. Suggested Journal and Conference Venues

- IEEE Transactions on Information Forensics and Security (T-IFS) – Premier venue for combined security and forensics contributions
- Computers & Security (Elsevier) – High-impact journal covering ML applications in cybersecurity
- IEEE Symposium on Security and Privacy (S&P) – Top-tier conference with strong ML + security track
- ACM Conference on Computer and Communications Security (CCS) – Highly competitive; suitable for the ensemble + forensics contribution
- International Journal of Information Security (Springer) – Well-suited for practical system contributions
- IEEE International Conference on Communications (ICC) – Relevant for the SMTP protocol analysis dimension

REFERENCES

[1] E. Lochin, “STAMP: SMTP Server Topological Analysis by Message Headers Parsing,” in *Proceedings of the IEEE Consumer Communications & Networking Conference (CCNC)*, Las Vegas, NV, USA, Jan. 2025.

[2] N. A. Mohammed, A. Hassan, and M. R. Yusof, “Recognizing Phishing in Emails Using NLP & ML Techniques,” in *Proceedings of the IEEE*

International Conference on Computing Research (ICCR), 2025.

[3] M. Ajuluchukwu, O. Nwosu, and K. Adams, "Detecting Malicious Emails and Deceptive Websites Using Generative AI," in *Proceedings of the IEEE International Conference on Communication Networks (CICN)*, 2025.

[4] L. K. T. A. U. et al., "Email Armour: A Multi-Layered Email Defense Solution," in *Proceedings of the IEEE International Conference on Information Technology Research (ICITR)*, 2024.

[5] S. Abu-Nimeh, D. Nappa, X. Wang, and S. Nair, "A Comparison of Machine Learning Techniques for Phishing Detection," in *Proceedings of the Anti-Phishing Working Group eCrime Researchers Summit*, Pittsburgh, PA, USA, 2007, pp. 60–69.

[6] M. Chandrasekaran, K. Narayanan, and S. Upadhyaya, "Phishing Email Detection Based on Structural Properties," in *Proceedings of the NYS Cyber Security Conference*, Albany, NY, USA, 2006.

[7] J. Ma, L. Saul, S. Savage, and G. Voelker, "Learning to Detect Malicious URLs," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 1–24, Apr. 2011.

[8] A. Blum, B. Wardman, T. Solorio, and G. Warner, "Lexical Feature Based Phishing URL Detection Using Online Learning," in *Proceedings of the ACM Workshop on Artificial Intelligence and Security (AISec)*, Chicago, IL, USA, 2010.

[9] F. Toolan and J. Carthy, "Feature Selection for Spam and Phishing Detection," in *Proceedings of the IEEE eCrime Researchers Summit*, Dallas, TX, USA, 2010.

[10] R. Verma and K. Dyer, "On the Character of Phishing URLs: Accurate and Robust Statistical Learning Classifiers," in *Proceedings of the ACM Conference on Data and Application Security and Privacy (CODASPY)*, San Antonio, TX, USA, 2015.

[11] A. O. Adebowale, K. T. Lwin, and M. A. Hossain, "Intelligent Phishing Detection Scheme Using Deep Learning Algorithms," *Journal of Enterprise Information Management*, vol. 35, no. 3, pp. 694–714, 2021.

[12] SpamAssassin Public Corpus. [Online]. Available: https://spamassassin.apache.org/publiccorpus/

[13] PhishTank. [Online]. Available: https://www.phishtank.com/

[14] TREC 2007 Spam Track. [Online]. Available: https://trec.nist.gov/data/spam.html

[15] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, San Francisco, CA, USA, 2016, pp. 785–794.

[16] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation Forest," in *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, Pisa, Italy, 2008, pp. 413–422.

[17] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[18] NIST, *Guide to Integrating Forensic Techniques into Incident Response*, NIST Special Publication 800-86, Gaithersburg, MD, USA, 2006.