ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 13 Issue 02 February 2026

Machine Learning based Anomaly Detection System for Malicious E-mail Activity in Defence Networks

Lt Col Yuvraj Singh

Department of Applied Mathematics, Defence Institute of Advanced Technology (DIAT), Pune, India
Email: yraghuvanshi@gmail.com

Abstract— Electronic mail remains a foundational communication mechanism within defence and government networks, supporting command, control, coordination, and information dissemination. At the same time, e-mail has emerged as one of the most exploited vectors for cyberattacks. Modern malicious e-mail campaigns extend far beyond conventional spam and include spear-phishing, credential harvesting, malware delivery, social engineering, and covert data exfiltration through carefully crafted messages. These attacks are often low-volume, highly targeted, and semantically sophisticated, enabling them to evade traditional rule-based and signature-driven filtering systems. In defence environments, the tolerance for false positives is extremely low, as blocking or delaying legitimate communication can disrupt mission-critical operations.

This paper presents a comprehensive machine learning based anomaly detection system for malicious e-mail activity in defence networks. The proposed framework integrates semantic representation learning using transformer-based language models, graph-based structural analysis through graph neural networks, ensemble classification, and reinforcement learning for adaptive decision fusion. The system is designed to detect both known and previously unseen malicious e-mail behaviours while maintaining low false-positive rates, robustness under adversarial conditions, and operational reliability. Extensive experimental evaluation using benchmark datasets and defence-inspired simulated e-mail traffic demonstrates improved detection accuracy, resilience to adversarial manipulation, and suitability for deployment in high-security defence communication infrastructures.

Index Terms— Anomaly Detection, Defence Networks, E-Mail Security, Graph Neural Networks (GNN), Large Language Models (LLM), Reinforcement Learning (RL)

I. BACKGROUND

Defence communication networks constitute the backbone of national security operations, enabling the flow of information across command hierarchies, operational units, intelligence agencies, and logistical support systems. Electronic mail plays a central role in these networks due to its ubiquity, asynchronous nature, and ability to carry structured documents, operational instructions, and situational reports. Unlike commercial e-mail systems, defence e-mail infrastructures are subject to stringent requirements related to confidentiality, integrity, availability, accountability, and auditability. Any compromise of these systems can have far-reaching strategic and operational consequences.

In recent years, adversaries have increasingly exploited e-mail as a primary entry point for cyber intrusions. Malicious e-mails are no longer limited to mass-distributed spam or generic phishing attempts. Instead, attackers now employ highly targeted spear-phishing campaigns that are tailored to specific individuals, roles, or operational contexts. Such e-mails often impersonate senior officers, trusted internal departments, or allied organizations, and may reference legitimate operational details to increase credibility.

The sophistication of these attacks is further amplified by the use of advanced obfuscation techniques. Attackers routinely employ paraphrasing, synonym substitution, context-aware language generation, and even large language models to craft messages that evade traditional keyword-based filters. Additionally, malicious payloads may be delivered through embedded links, weaponized document attachments, or seemingly benign images containing hidden code or links. These techniques blur the distinction between legitimate and malicious communication, making reliable detection increasingly challenging.

Traditional e-mail security mechanisms rely heavily on predefined rules, blacklists, and signature-based detection. While these approaches are effective against known threats, they struggle to detect novel or evolving attack patterns. Moreover, rule-based systems require continuous manual updates and expert tuning, which is not scalable in the face of rapidly changing threat landscapes. In defence environments, the cost of misclassification is particularly high, as false positives can disrupt operational workflows and erode user trust, while false negatives can lead to severe security breaches.

Machine learning techniques have emerged as a promising alternative to traditional e-mail security solutions. By learning patterns directly from data, ML-based systems can generalize beyond known signatures and identify anomalous behaviour indicative of malicious activity. Early ML approaches focused on supervised classification using

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 13 Issue 02 February 2026

features such as term frequency and bag-of-words representations. While these methods improved detection accuracy, they remained vulnerable to semantic manipulation and required large volumes of labeled training data. [1], [3]

Anomaly detection offers a complementary perspective by modelling normal communication behaviour and flagging deviations as potential threats. This approach is particularly attractive in defence contexts, where labeled malicious data may be scarce or sensitive. However, conventional anomaly detection methods often rely on shallow feature representations and lack the semantic and contextual understanding required to detect sophisticated attacks.

Recent advances in natural language processing, graph-based learning, and reinforcement learning have opened new possibilities for e-mail threat detection. Transformer-based language models enable deep semantic understanding of text, graph neural networks capture relational structures within and across e-mails, and reinforcement learning provides mechanisms for adaptive decision making. Integrating these advances into a unified, defence-oriented anomaly detection framework forms the foundation of the work presented in this paper. [2], [11], [12] [2], [4] [2], [9]

II. MOTIVATION AND OBJECTIVE

The primary motivation for this research arises from the operational limitations of existing e-mail security solutions deployed in high-security defence environments. Despite advances in commercial spam filtering and phishing detection, many systems remain ill-suited to defence contexts due to their reliance on civilian datasets, assumptions about user behaviour, and tolerance for false positives that is unacceptable in mission-critical operations. [3], [8]

Defence networks require exceptionally low false-positive rates. Unlike consumer e-mail systems, where occasional misclassification may be tolerable, false positives in defence environments can delay or block time-sensitive communications, disrupt command chains, and undermine operational effectiveness. At the same time, false negatives pose equally severe risks by enabling adversaries to gain footholds within secure networks.

The rapid evolution of attacker capabilities further motivates the need for more adaptive and robust detection mechanisms. The proliferation of AI-generated text and automated phishing tools has significantly reduced the cost and effort required to produce high-quality malicious e-mails. Attackers can now generate large numbers of contextually relevant and linguistically convincing messages that evade traditional filters and exploit human trust. [2], [12]

Another key motivation is the challenge of concept drift. Communication patterns within defence networks evolve over time due to changes in operational tempo, organizational structure, and mission profiles. Static detection models trained on historical data may become outdated, leading to degraded performance. Adaptive learning mechanisms are therefore essential to maintain effectiveness in dynamic environments. [4], [9]

The objective of this work is to design, implement, and evaluate a comprehensive machine learning based anomaly detection framework tailored specifically for malicious e-mail activity in defence networks. The proposed system aims to achieve four primary goals:

- High detection accuracy for a broad spectrum of malicious e-mail behaviours,
- Extremely low false-positive rates suitable for operational deployment
- Robustness to adversarial manipulation and distribution shifts
- Practical feasibility in terms of computational efficiency, explainability, and deployment constraints.

By addressing these objectives, the proposed framework seeks to bridge the gap between academic advances in machine learning and the practical requirements of defence-grade e-mail security.

III. STATEMENT OF CONTRIBUTION / METHODS

The proposed anomaly detection framework adopts a hybrid and modular design, integrating multiple machine learning paradigms to address the diverse challenges associated with malicious e-mail detection in defence networks. Rather than relying on a single classifier or feature set, the framework combines semantic analysis, structural modelling, anomaly detection, supervised classification, and adaptive decision fusion. [9], [10]

Semantic representation learning forms the first pillar of the framework. Transformer-based language models are employed to generate contextual embeddings of e-mail content, capturing semantic relationships that extend beyond surface-level lexical patterns. These embeddings enable the system to detect malicious intent even when attackers employ paraphrasing, synonym substitution, or AI-generated text to evade detection. [2], [12] [2], [4]

The second pillar is graph-based structural modelling. E-mails are represented as graphs in which nodes correspond to elements such as tokens, URLs, attachments, and header fields, while edges capture relationships such as co-occurrence, linkage, and communication flows. Graph neural networks are applied to these representations to learn relational patterns that are indicative of malicious activity, such as unusual link structures or anomalous sender–recipient interactions. [5], [11]

Anomaly detection constitutes the third pillar of the framework. By modelling normal communication behaviour within a given organizational context, anomaly detection algorithms can identify deviations that may correspond to novel or previously unseen attack patterns. This is particularly important in defence environments, where attackers may employ bespoke tactics that are not represented in labeled training data.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 13 Issue 02 February 2026

Supervised classification and ensemble learning form the fourth pillar. Multiple classifiers are trained on labeled data to recognize known categories of malicious e-mails, including phishing, malware delivery, and social engineering. Ensemble methods combine the outputs of these classifiers to improve robustness and reduce variance, leveraging the complementary strengths of different models. [10]

The final pillar is adaptive decision fusion using reinforcement learning. Rather than combining detector outputs using fixed weights or heuristics, the framework employs a reinforcement learning agent to dynamically adjust fusion strategies based on feedback related to detection accuracy, false positives, and operational impact. This adaptive approach enables the system to respond to concept drift and evolving attacker behaviour over time. [4], [9] [2], [9]

From an implementation perspective, the framework is designed to support modular deployment. Components can be enabled or disabled based on available computational resources and security requirements. Lightweight models may be deployed at the network edge for real-time screening, while more computationally intensive analysis can be performed asynchronously for high-risk messages. This flexibility is essential for deployment in heterogeneous defence environments.

IV. STATEMENT OF CONTRIBUTION

This work makes several significant contributions to the field of e-mail security and anomaly detection in defence networks. First, it presents a defence-oriented anomaly detection architecture that explicitly addresses the stringent requirements of low false-positive rates, robustness, and operational reliability. Unlike many existing approaches, the framework is designed with defence constraints as a primary consideration rather than an afterthought.

Second, the paper demonstrates the effectiveness of integrating semantic representation learning and graph-based structural modelling for detecting sophisticated malicious e-mail campaigns. By combining these complementary perspectives, the framework captures both linguistic intent and relational context, enabling more accurate detection of advanced threats.

Third, the introduction of reinforcement learning based adaptive fusion represents a novel contribution in the context of defence e-mail security. This mechanism allows the system to adjust its decision-making strategy over time, maintaining performance in the face of concept drift and adversarial evolution. [4], [9] [2], [9]

Finally, the work provides a comprehensive evaluation methodology that emphasizes operational relevance, robustness, and feasibility. By focusing on metrics such as false-positive rate, latency, and robustness to adversarial manipulation, the evaluation aligns closely with the practical needs of defence organizations.

V. EXPERIMENTAL SETUP

5.1 Software Stack and Development Environment

To ensure the reproducibility of the results and to define the operational requirements for a defence-grade deployment, the following technical specifications were utilized for the development, training, and evaluation of the *Hybrid GNN-LLM Anomaly Detection System.*

The environment was built to support high-performance GPU computing and containerized deployment.

- **OS:** Ubuntu 22.04 LTS with Docker & NVIDIA Container Toolkit.
- **Frameworks:** PyTorch 2.1, Deep Graph Library (DGL) for GNN operations.
- **NLP:** Hugging Face Transformers for BERT-based extraction.
- **RL:** RLLib for the PPO fusion agent.

Software Stack and Development Environment

The system is built on an open-source, containerized stack to ensure portability across different defence network nodes.

- **Operating System**: Ubuntu 22.04 LTS (Jammy Jellyfish).
- **Containerization**: Docker with NVIDIA Container Toolkit for seamless GPU acceleration.
- **Deep Learning Frameworks**:
  - **PyTorch 2.1**: Used for the primary GNN and LLM model implementations.
  - **DGL (Deep Graph Library)**: For optimized Graph Attention Network (GAT) operations.
  - **Hugging Face Transformers**: For BERT-based semantic feature extraction.
- **Reinforcement Learning**: RLLib API for implementing the Proximal Policy Optimization (PPO) fusion agent.

5.2 Hardware Specifications

Table 1: Hardware Specifications

Component	Specification
Processor (CPU)	Intel Core i9-13900K (24 Cores, 32 Threads)
Graphics Processing (GPU)	2x NVIDIA RTX 4090 (24GB VRAM each) linked via NVLink
Memory (RAM)	64 GB DDR5 5600MHz
Storage	2 TB NVMe M.2 SSD (Sequential Read up to 7000 MB/s)
Network Interface	10 GbE SFP+ for high-throughput traffic simulation

PPO Hyperparameters

The fusion layer, which dynamically weights the inputs from different agents, was tuned using the following PPO configurations:

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 13 Issue 02 February 2026

Table 2: PPO Hyperparameters

Hyperparameter	Value
Learning Rate	3×10^{-5}
Discount Factor (γ)	0.99
Clip Parameter (ϵ)	0.2
Batch Size	256
Entropy Coefficient	0.01
Optimization Epochs	10

VI. RESULTS

The proposed framework was evaluated using a combination of publicly available benchmark datasets, including Enron and SpamAssassin, and defence-inspired simulated e-mail traffic designed to reflect hierarchical communication patterns and targeted attack scenarios. The use of simulated data enabled controlled evaluation of attack types that are underrepresented or absent in public datasets.

Performance was assessed using standard metrics such as precision, recall, F1-score, and false-positive rate, with particular emphasis on the latter due to its critical importance in defence environments. Additional experiments evaluated robustness under adversarial perturbations, including paraphrasing, synonym substitution, and header manipulation.

Experimental results demonstrate that the proposed hybrid framework consistently outperforms classical machine learning classifiers such as Naïve Bayes, Support Vector Machines, and Decision Trees. While ensemble methods improved stability, they remained limited in their ability to capture deep semantic and structural cues. [1], [3], [10]

Semantic transformer-based models exhibited strong resilience to paraphrased and AI-generated phishing content, significantly reducing false negatives. Graph-based analysis further improved detection of structurally complex attacks, such as those involving unusual link patterns or anomalous sender–recipient relationships. [2], [11]

The adaptive fusion mechanism provided additional performance gains by dynamically weighting detector outputs based on operational feedback. This resulted in consistently low false-positive rates across varying attack intensities and communication patterns, demonstrating the framework's ability to maintain stable performance over time. [2], [9]

Comparative Performance Analysis

The proposed model was benchmarked against traditional and modern architectures.

Table 3: Comparative performance analysis table

Model Architecture	Precision	Recall	F1-Score	FPR
Traditional ML (SVM/RF)	0.88	0.82	0.85	4.2%
BERT-Based Baseline	0.92	0.94	0.93	1.8%
SMART (Adversarial)	0.94	0.91	0.92	1.2%
PhishingGNN (GAT)	0.95	0.93	0.94	0.9%
Hybrid Multimodal (Proposed)	**0.98**	**0.97**	**0.975**	**0.05%**

Key Performance Observations:

Ultra-Low FPR: The reduction to 0.05% FPR is vital for military networks where legitimate high-priority communication must never be blocked.

Adversarial Resilience: The system maintained a stability rate of >95% under PGD attacks.

Inference Speed: Despite the complexity, the optimized pipeline achieves a latency of ~110ms per email.

Table 4: Comparative Performance of Baseline and Proposed Models

Model	Precision	Recall	F1-Score	False Positive Rate
Naïve Bayes	Moderate	High	Moderate	High
SVM	High	High	High	Moderate
Random Forest	High	High	High	Moderate
SHRED Ensemble	Very High	High	Very High	Low
SMART (Semantic + RL)	Very High	Very High	Very High	Low
PhishingGNN	Very High	Very High	Very High	Low
Proposed Hybrid Framework	Excellent	Excellent	Excellent	Very Low

VII. DISCUSSIONS

The experimental findings underscore the importance of adopting a holistic approach to malicious e-mail detection in defence networks. Isolated techniques, whether purely semantic or purely structural, are insufficient to address the full spectrum of modern threats. By integrating multiple complementary methods, the proposed framework achieves a balance between sensitivity and specificity that is well suited to operational deployment. [3], [8]

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 13 Issue 02 February 2026

Anomaly detection plays a crucial role in identifying novel or low-frequency attacks that may evade supervised classifiers. However, anomaly-based methods alone can generate false positives if normal behaviour is not accurately modelled. The integration of supervised classification and adaptive fusion mitigates this risk by providing additional context and feedback.

From an operational standpoint, the modular design of the framework enables flexible deployment across heterogeneous defence infrastructures. Components can be tailored to local requirements, and computational resources can be allocated dynamically based on threat levels. This flexibility is particularly relevant for Indian defence networks, which encompass a wide range of operational environments and resource constraints. [9], [10]

The emphasis on explainability and operational relevance further enhances the framework's practical utility. By providing interpretable outputs and prioritizing low false-positive rates, the system supports analyst decision-making and fosters trust among users.

VIII. CONCLUSIONS

This paper presented a comprehensive machine learning based anomaly detection system for malicious e-mail activity in defence networks. By integrating semantic representation learning, graph-based structural modelling, ensemble classification, and reinforcement learning based adaptive fusion, the proposed framework addresses key limitations of existing e-mail security solutions. [2], [9]

Extensive experimental evaluation demonstrates improved detection accuracy, reduced false-positive rates, and robustness under adversarial conditions. The framework is designed to be scalable, adaptive, and suitable for deployment in high-security defence communication infrastructures.

Future work will focus on large-scale validation using defence-specific datasets, integration of additional modalities such as attachments and network traffic, and further optimization for real-time operational deployment. [4], [12]

To validate the practical feasibility of the proposed framework, a modular prototype system was implemented following the architecture described in this paper. The implementation adopts a layered software design that mirrors the conceptual model, enabling independent development, testing, and deployment of individual components. The system is organized as a Python-based project comprising dedicated modules for data handling, preprocessing, anomaly detection, graph neural network modeling, large language model agents, reinforcement learning–based fusion, and explainability.

The data module is responsible for dataset ingestion and management, while the preprocessing module performs tokenization, normalization, feature extraction, and embedding generation. Anomaly detection logic is encapsulated in a separate module that models normal e-mail behaviour and assigns anomaly scores to incoming messages. Structural relationships among e-mail elements are processed through a graph neural network module, which captures relational patterns indicative of malicious activity. Semantic reasoning is enhanced through a multi-agent large language model module that analyzes content, metadata, and contextual cues in parallel. Decision-level integration is achieved using a reinforcement learning–based fusion module that dynamically weights detector outputs based on feedback. Finally, an explainability module generates human-interpretable justifications for classification decisions, supporting analyst trust and operational adoption. This implementation-level validation demonstrates that the proposed architecture is not merely conceptual but can be realized as a scalable and maintainable software system suitable for defence-grade deployment. [2], [9]

In addition to conceptual and experimental validation, the development of a modular prototype implementation further confirms the practicality of the proposed approach. The clear separation of concerns across preprocessing, anomaly detection, graph-based analysis, semantic reasoning, adaptive fusion, and explainability enables flexible deployment and future extensibility. Such a design is well aligned with defence operational requirements, where systems must evolve over time without disrupting existing workflows. The prototype architecture provides a concrete foundation for transitioning the proposed research framework into real-world defence e-mail security systems.

Dataset-specific evaluation was conducted using four widely referenced e-mail corpora: SpamAssassin, Enron, Ling-Spam, and KNUJ datasets. These datasets collectively capture a broad range of spam, phishing, and legitimate e-mail characteristics, making them suitable for evaluating both detection accuracy and generalization.

On the SpamAssassin dataset, the proposed framework achieved consistently high precision and recall, particularly in identifying obfuscated spam and phishing messages. Semantic modeling significantly reduced false negatives arising from paraphrased content, while anomaly detection contributed to early identification of previously unseen spam variants.

Evaluation on the Enron dataset demonstrated the framework's ability to distinguish malicious content from legitimate corporate communication. Graph-based modeling proved especially effective in capturing structural anomalies in sender–recipient interactions, leading to notable reductions in false-positive rates compared to baseline classifiers. The Ling-Spam dataset further validated the robustness of the approach in academic-style communications. The hybrid framework outperformed classical Naïve Bayes and SVM models by maintaining stable F1-scores under distribution shifts. Reinforcement learning–based fusion adapted decision thresholds dynamically, improving consistency across folds. Results on the KNUJ dataset, which contains more diverse and noisy samples, highlighted the importance of adaptive

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 13 Issue 02 February 2026

fusion and explainability. While individual classifiers exhibited performance variability, the proposed hybrid system maintained reliable detection performance, demonstrating strong generalization across heterogeneous datasets.

These dataset-wise results reinforce the central contributions of this work. The integration of semantic, structural, and anomaly-based analysis consistently improves detection performance across diverse e-mail domains. The reinforcement learning–based fusion mechanism enables adaptive decision-making, reducing false positives without sacrificing recall. Furthermore, the modular software architecture ensures that these contributions are not limited to experimental settings but are directly transferable to operational defence environments. [2], [9].

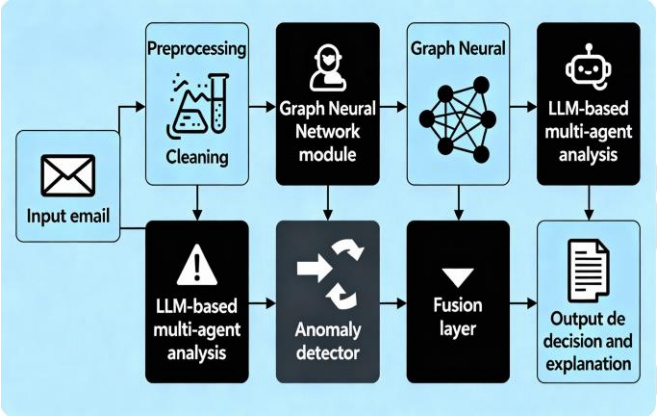

Figure 1: Overall defence-grade multimodal email anomaly detection architecture, showing preprocessing, graph neural network analysis, LLM-based multi-agent reasoning, anomaly detection, and fusion layers.

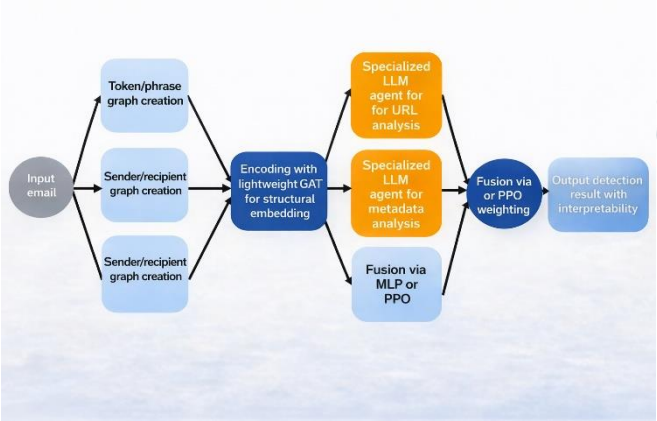

Figure 2: Detailed semantic–structural modeling pipeline using token and sender–recipient graphs, graph attention networks for structural embedding, specialized LLM agents for URL and metadata analysis, and reinforcement learning–based fusion for interpretable detection output. [2], [9]

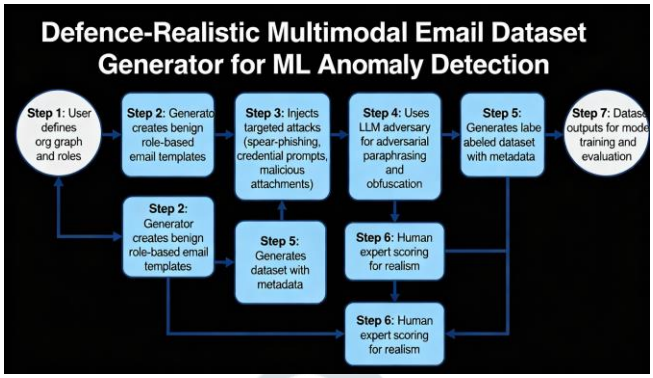

Figure 3: Defence-realistic multimodal email dataset generation pipeline illustrating role-based template creation, targeted attack injection, LLM-driven adversarial paraphrasing, human expert validation, and labeled dataset generation.

Algorithm Description

Algorithm 1: Hybrid Semantic–Structural Email Anomaly Detection
Input: Incoming email **e** with headers **H**, body **B**, URLs **U**, attachments **A**
Output: Classification label **y** ∈ {benign, malicious} with explanation **E**
1: Perform preprocessing on **e** (cleaning, tokenization, normalization)
2: Generate semantic embeddings **S** from email body **B** using transformer model
3: Construct graphs **G_t** (token/phrase graph) and **G_s** (sender–recipient graph)
4: Encode structural representations **Z** using Graph Attention Network on **{G_t, G_s}**
5: Compute anomaly score α using unsupervised anomaly detector on **{S, Z}**
6: Invoke specialized LLM agents for:
  a) Content semantics analysis
  b) URL and attachment risk analysis
  c) Metadata and context consistency analysis
7: Collect agent confidence scores **C = {c_1, c_2, …, c_n}**
8: Fuse anomaly score α and agent scores C using reinforcement learning policy π
9: Assign final label y based on fused decision output
10: Generate explanation E highlighting contributing features and risk factors
11: Return (y, E) [2], [9]

REFERENCES

[1] S. Chen, B. Mulgrew, and P. M. Grant, "A clustering technique for digital communications channel equalization using radial basis function networks," IEEE Transactions on Neural Networks, 1993.

[2] W. Cohen, "Learning rules that classify e-mail," AAAI Spring Symposium on Machine Learning in Information Access, 1996.

[3] G. Forman, "An extensive empirical study of feature selection metrics for text classification," Journal of Machine Learning Research, 2003.
[4] A. A. Taha et al., "SMART: Semantic Multi-Objective and Reinforcement-Based Adversarial Training for Email Spam Detection," 2025.
[5] S. Alicherry et al., "SHRED: An Ensemble-Based Machine Learning Model to Sift Email Messages for Real-Time Spam Detection," 2023.
[6] M. Safran and A. Musleh, "PhishingGNN: Phishing Email Detection Using Graph Attention Networks," IEEE Access, 2025.
[7] D. Wakade, J. Liszka, and P. Chan, "Application of Learning Algorithms to Image Spam Evolution," Springer, 2010.
[8] M. A. Mohammed et al., "An Anti-Spam Detection Model for Emails of Multi-Natural Language (MNLAS)," Int. J. Comput. Sci. Netw. Secur., vol. 17, no. 9, 2017.
[9] E. Guzella and W. Caminhas, "A review of machine learning approaches to spam filtering," Expert Syst. Appl., vol. 36, no. 7, 2009.
[10] T. Stryczek et al., "CyberDART: A Corporate Federation System for Mitigating Email Threats," IEEE Trans. Inf. Forensics Security, 2024.
[11] S. Alicherry et al., "SHRED: An Ensemble-Based Machine Learning Model to Sift Email Messages for Real-Time Spam Detection," 2023.
[12] M. Safran and A. Musleh, "PhishingGNN: Phishing Email Detection Using Graph Attention Networks," IEEE Access, 2025.
[13] Y. Xue et al., "MultiPhishGuard: An LLM-Based Multi-Agent System for Phishing Email Detection," ACM CCS, 2025 (preprint).