ISSN: 2230-9926

Available online at http://www.journalijdr.com

IJDR

International Journal of Development Research
Vol. 16, Issue, 07, pp.70791-70800, July, 2026
https://doi.org/10.37118/ijdr.30882.07.2026

RESEARCH ARTICLE OPEN ACCESS

HYBRID GNN–LLM FRAMEWORK FOR MACHINE LEARNING BASED ANOMALY DETECTION IN ENTERPRISE EMAIL NETWORKS: INTEGRATING SEMANTIC, STRUCTURAL, AND REINFORCEMENT LEARNING TECHNIQUES

Lt Col Yuvraj Singh, Abhijit Gogoi and Mohd. Kaif

University Student, DIAT Pune, India

ARTICLE INFO

Article History:
Received 19th April, 2026
Received in revised form
14th May, 2026
Accepted 27th June, 2026
Published online 30th July, 2026

Key Words:

Anomaly detection, enterprise email security, graph attention networks, machine learning, ModernBERT, phish-ing detection, proximal policy optimisation, spam filtering, transformer encoder.

****Corresponding author:***
Lt Col Yuvraj Singh

ABSTRACT

Enterprise email networks face an escalating volume of sophisticated cyber attacks including spear-phishing, credential harvesting, business email compromise, covert data exfiltration, and multi-stage malware delivery. Traditional signature-based and rule-based detection systems are inadequate against modern adversaries who continuously modify content, employ multi-layer obfuscation, and exploit large language models to generate contextually convincing fraudulent messages. This paper proposes, implements, and experimentally validates a Hybrid Graph Neural Network-Large Language Model (GNN-LLM) Anomaly Detection Framework for malicious email activity in enterprise networks. The framework integrates five complementary detection pillars: (1) semantic representation via a LoRA fine-tuned ModernBERT encoder with mean pooling over all token hidden states, (2) structural modelling via a Graph Attention Network (GAT) operating on token co-occurrence graphs, (3) unsupervised anomaly detection via Isolation Forest trained exclusively on benign email patterns, (4) supervised soft-voting ensemble classification combining SVM, Random Forest, and Logistic Regression with PCA dimensionality reduction, and (5) adaptive decision fusion via a Proximal Policy Optimisation (PPO) agent with asymmetric reward structure and automatic F_1-maximising threshold calibration. Evaluated on three publicly available benchmark datasets spanning corporate email, SMS spam, and academic mailing lists, the system achieves 99.90 % accuracy and 100 % recall (zero false negatives across 5,060 test samples) on the Enron corporate email corpus, with only 0.20 % false positive rate. Cross-domain evaluation confirms 95.17 % accuracy on KNUJ SMS and 98.85 % on Ling-Spam without domain-specific retraining. A key empirical finding is that mean pooling over all BERT token representations yields a 50.6 percentage-point accuracy gain over [CLS]-token extraction alone; establishing it as the decisive architectural choice for domain-shifted email classification tasks.

Citation: Lt Col Yuvraj Singh, Abhijit Gogoi, and Mohd. Kaif. 2026. "Hybrid gnn–llm Framework for Machine Learning Based Anomaly Detection in Enterprise email Networks: Integrating Semantic, Structural, and Reinforcement Learning Techniques". *International Journal of Development Research*, 16, (08), 70791-70800.

INTRODUCTION

Corporate email has become the primary channel for business communication, coordination, and information exchange across organisations worldwide. This critical dependence makes email infrastructure the single largest attack surface in the modern enterprise: over 90 % of successful cyberattacks begin with a malicious email. Contemporary adversaries deploy highly targeted, semantically sophisticated attacks; spear-phishing cam-paigns personalized to specific individuals or job functions, credential harvesting through impersonation of trusted internal senders, business email compromise (BEC) exploiting organisational hierarchy, covert data exfiltration through innocuous-looking message streams, and multi-stage malware delivery through embedded attachments and obfuscated URLs [1]. The emergence of large language model (LLM)-generated phishing content represents a qualitative shift in the threat landscape. Historically, fraudulent emails could be identified through inconsistent language, generic salutations, or implausible requests. LLM-generated content eliminates these distinguishing characteristics, produc-ing contextually appropriate, grammatically correct, and stylistically consistent deception indistinguishable from legitimate communication by rule-based filters and, increasingly, by human recipients [2]. The IBM 2024 Cost of a Data Breach Report [3] found that phishing-initiated breaches cost organisations an average of USD 4.9 million underscoring the commercial urgency of accurate, scalable

automated detection. Traditional email security solutions like signature-based antivirus, rule-based spam filters, Bayesian classifiers, and blacklist-based URL checkers operate reactively, flagging only threats for which prior examples exist in their knowledge bases. This paradigm fails entirely against novel, zero-day attack patterns and LLM-crafted social engineering. The operational consequence in a corporate context is not merely nuisance but financial loss, reputational damage, and regulatory liability. The asymmetric cost structure where the consequence of a missed malicious email far exceeds the inconvenience of a false alarm demands detection systems that simultaneously minimize both false positives and false negatives. This paper proposes and validates a Hybrid GNN-LLM Anomaly Detection Framework that addresses this dual objective through a five-pillar architecture. Evaluated on three structurally heterogeneous benchmark datasets, the system achieves: 99.90 % accuracy with zero false negatives on the primary Enron corporate email corpus; 95.17 % accuracy on KNUJ SMS; and 98.85 % on Ling-Spam academic email. Table I presents the headline results.

Table 1. Headline test Results : All three Evaluation Datasets

Dataset	Acc.	Prec.	Rec.	F1	FPR
Enron Corporate (33,716)	99.90%	99.81%	100.0%	99.90%	0.20%
KNUJ SMS (5,796)	95.17%	96.55%	88.42%	92.31%	1.54%
Ling-Spam (2,893)	98.85%	95.95%	97.26%	96.60%	0.83%

Key Contributions

- Design and implementation of the first five-pillar hybrid GNN–LLM detection system combining Modern-BERT, GAT, Isolation Forest, soft-voting ensemble, and PPO fusion implemented in Python without CUDA dependency, running on Apple M4 MPS.
- Empirical demonstration that mean pooling over all non-padding BERT token representations increases Enron test accuracy from 49.27 % ([CLS]-only baseline) to 99.90 % a 50.6 percentage-point improvement from a single architectural change.
- Automatic F_1-maximising threshold calibration procedure that eliminates the critical failure mode of hardcoded decision boundaries in PPO fusion networks trained with asymmetric rewards.
- Cross-domain evaluation across three structurally heterogeneous datasets demonstrating 99.90 %, 95.17 %, and 98.85 % test accuracy without domain-specific retraining.
- Reproducible open-source Python implementation with twelve modular components suitable for extension to proprietary enterprise email corpora and adversarial robustness benchmarking.

RELATED WORK

Conventional and Ensemble-Based Approaches: Early machine learning approaches to email spam detection relied on Naive Bayes classifiers exploiting word frequency distributions [4], which remain competitive baselines for their simplicity and interpretability. Support Vector Machines emerged as the dominant algorithm, identified by Salloum *et al.* [5] as the method used in approximately 50 % of phishing detection studies, owing to their effectiveness in high-dimensional TF-IDF feature spaces. The SHRED ensemble framework [6] demonstrated state-of-the-art performance by combining five base classifiers (Naive Bayes, ANN, Random Forest, AdaBoost, Decision Tree) through stacking, achieving detection rates exceeding 99 % on a combined corpus of 83,448 emails. However, reliance on bag-of-words representations renders these systems vulnerable to semantic evasion: adversarial emails constructed through synonym substitution, paraphrasing, or unusual vocabulary distributions evade keyword-matching without detection.

Deep Learning and Transformer-Based Methods: The introduction of transformer-based language models substantially advanced email threat detection. BERT [7] demonstrated that bidirectional contextual representations substantially outperform sequence-based models across NLP classification tasks. The SMART framework [8] integrated Word2Vec embeddings, K-means semantic clus-tering, multi-objective adversarial training, and a reinforcement learning agent for Dynamic Difficulty Adjustment, achieving 93.50 % average accuracy and F_1-score of 0.871 under FGSM and PGD adversarial perturbations. PhishingGNN [9] introduced graph-based email representation by combining DistilBERT embeddings with Graph Attention Networks (GAT), achieving accuracy of 0.9939 and F_1 of 0.99 on the CEAS_08 phishing dataset. MultiPhishGuard [10] demonstrated that multi-agent LLM architectures can achieve macro-average F_1 of 95.88 % across six datasets, though the requirement for multiple LLM API calls per email introduces latency incompatible with real-time gateway deployment. Jamal *et al.* [11] further demonstrated that fine-tuned transformer encoders consistently outperform classical baselines for joint phishing, spam, and ham classification.

Identified Research Gaps: Five persistent gaps across the reviewed literature motivate the proposed framework: (RG1) pre-trained language models have not been fine-tuned specifically for enterprise email vocabularies and communication patterns; (RG2) most systems are optimised for bulk spam rather than the full spectrum of phishing, business email compromise, and social engineering; (RG3) text-content analysis and graph-structural modelling have not been combined with unsupervised anomaly detection in a unified pipeline; (RG4) existing multi-modal frameworks require external API access or CUDA hardware, with no auto-calibrated decision thresholds; and (RG5) binary classification outputs lack the analyst-interpretable justifications required for high-stakes enterprise security review workflows. Each gap is addressed by a specific component of the proposed architecture.

Research Objectives and Pillar Mapping: Six research objectives (ROs) are formulated in direct response to the identified literature gaps. Each objective maps to one or more of the five detection pillars. Table II presents the complete objective-to-pillar-to-outcome mapping.

RO1 - Semantic Adaptation: Fine-tune a state-of-the-art contextual language model (ModernBERT-base with LoRA, r=16) specifically for malicious email detection. Mean pooling over all non-padding token representations is employed to maximise the richness of contextual features extracted from email content, addressing the gap of sub-optimal CLS-token extraction for domain-shifted tasks.

RO2- Broad-Spectrum Detection: Design a framework capable of identifying the full range of malicious email types -spam, phishing, business email compromise, spear-phishing, and socially engineered content beyond classic keyword-matching. Evaluated on three structurally heterogeneous datasets to demonstrate breadth of detection capability.

RO3- Multi-Modal Feature Fusion: Implement and validate a five-pillar architecture simultaneously integrating: semantic features from email body and subject via ModernBERT; structural features via GAT modelling token co-occurrence relationships; anomaly deviation scores from Isolation Forest trained on benign-only data; and supervised classification probabilities from a soft-voting ensemble.

RO4- Resource-Efficient Deployment with Auto-Calibration: Develop a computationally efficient pipeline de-ployable on resource-constrained hardware (Apple M4, 16 GB unified memory) without CUDA dependency, and implement an automatic threshold calibration mechanism. The system scans 200 candidate thresholds on the validation set to identify the F_1-maximising cutpoint, eliminating reliance on hardcoded decision boundaries.

RO5- Transparent and Interpretable Outputs: Implement a rule-based explainability module providing human-readable justifications for each classification decision, identifying triggered risk signals (urgency language, suspicious URLs, credential phishing patterns, brand impersonation), confidence scores, and component-level anomaly and ensemble probabilities.

RO6- Cross-Domain Generalisation: Demonstrate that the proposed framework achieves strong detection performance across three heterogeneous public datasets establishing a reproducible evaluation baseline that can be extended to proprietary enterprise email corpora.

Table 2. Research objectives, addressed gaps, system pillars, and achieved outcomes

RO	Objective	Research Gap Addressed	System Pillar / Compo-nent	Outcome
RO1	Semantic adaptation to email domains	Pre-trained models not fine-tuned on enterprise corpora; CLS-token yields suboptimal features	Pillar 1 : ModernBERT + LoRA (r=16); mean pool-ing over all tokens	+50.6 pp accuracy over CLS baseline
RO2	Broad-spectrum malicious email detection	Systems optimised for bulk spam; phishing and so-cial engineering underrep-resented	Pillar 3 : Isolation For-est on benign-only data en-ables novel threat detection	95–99.9% accuracy across 3 heterogeneous datasets
RO3	Multi-modal feature fusion	Systems use text or graph structure separately; no combined anomaly pillar	Pillars 1-4 : semantic + structural + anomaly + en-semble integrated pipeline	FN = 0 on primary cor-pus; AUC > 98% all datasets
RO4	Resource-efficient deploy-ment with auto-calibration	Multi-LLM agent systems require API calls; CUDA dependency; hardcoded thresholds fail	Pillar 5 : PPO MLP fusion; 200-point F_1-max thresh-old scan; < 50 ms infer-ence	Runs on M4 16 GB; cal-ibration recovers 49% → 99.90%
RO5	Transparent, interpretable decisions	Binary outputs lack analyst-readable justification for borderline cases	Explainability module 12 regex risk-signal patterns + confidence breakdown	Named risk signals per email with component-level scores
RO6	Cross-domain generalisa tion	No reproducible cross-domain evaluation baseline across heterogeneous corpora	Evaluation structurally on 3 different public datasets; fixed seed reproducibility	98.85% Ling-Spam; 95.17% KNUJ without domain retraining

Table 3. Five-Pillar Detection Architecture Summary

#	Pillar Name	Module	Output	Key Design Choice
1	Semantic Encoder	feature_extractor.py	768-dim embedding	ModernBERT + LoRA; mean pooling over all non-padding tokens +50.6 pp vs CLS
2	Structural Graph	gat_encoder.py	256-dim graph emb.	2-layer, 8-head GAT on token co-occurrence graph; 50-node, window=3
3	Anomaly Detector	anomaly_detector.py	$\alpha \in [0, 1]$	Isolation Forest (200 trees) trained on benign-only data; zero-day threat de-tection
4	Ensemble Classifier	ensemble_classifier.py	$p \in [0, 1]$	SVM + RF + LR soft-voting; PCA 1024 → 128 dims (91.9% var.); mandatory for SVM convergence
5	PPO Fusion Agent	ppo_fusion.py	$f \in [0, 1]$	MLP 2→16→8→1; Adam; FN penalty −3.0 > FP penalty −2.0; auto-calibrated t∗

SYSTEM ARCHITECTURE AND PIPELINE

The Hybrid GNN-LLM Anomaly Detection Framework processes each incoming email through five parallel or sequential detection pillars, whose outputs are adaptively fused by a PPO reinforcement learning agent to produce a final binary classification together with a human-interpretable explanation. The complete architecture is designed for deployment on resource-constrained hardware; all components execute locally without external API dependencies. Table III provides the five-pillar architecture summary. Fig. 1 illustrates the complete end-to-end detection pipeline. Raw incoming email comprising headers, body text, attachments, and URLs enters the preprocessing stage (parsing, cleaning, tokenization, and graph construction). Pillars 1 through 4 then operate in parallel to extract complementary feature signals, which converge at the PPO Adaptive Fusion Agent (Pillar 5). The final decision of malicious or benign with a confidence score is routed to one of three downstream actions: Alert/Block (real-time edge intervention), Log & Feedback (feeding analyst corrections back into the RL training loop for continuous adaptation), or Deep Analysis queue (for borderline high-risk emails requiring human review). The dashed arrow represents this continuous RL feedback loop.

Pillar 1- Semantic Encoder (Modern BERT + LoRA + Mean Pooling): ModernBERT-base (answerdotai/ModernBERT-base) serves as the core language model for contextual email encoding [12]. ModernBERT is an encoder-only transformer trained on two trillion tokens, incorporating Rotary Positional Embeddings (RoPE), Flash Attention 2, and a native context length of 8,192 tokens. The maximum sequence length is capped at 512 tokens for the current implementation, sufficient to encode 97 % of all emails in the three evaluation datasets.

Mean Pooling Strategy: A critical design decision is mean pooling over all non-padding token hidden states rather than extracting only the [CLS]classification token. Given hidden states $H = \{h_1, h_2, \ldots, h_n\}$ from the final transformer layer and attention

mask $M = \{m_1, m_2, \ldots, m_n\}$ where $m_i \in \{0, 1\}$, the pooled semantic embedding $\mathbf{s} \in \mathrm{R}^{768}$ is:

$$s = \frac{\sum_{i=1}^{n} m_i h_i}{\sum_{i=1}^{n} m_i}$$

Padding tokens contribute zero weight, so the embedding reflects only meaningful content tokens. This strategy captures the contribution of every meaningful word in the email body and subject line. The empirical validation is dramatic: mean pooling improved Enron test accuracy from 49.27 % ([CLS]-only baseline) to 99.90 %- a 50.6 percentage-point improvement attributable solely to this single architectural choice.

LoRA Fine-Tuning. ModernBERT is adapted using Low-Rank Adaptation (LoRA) [13] with rank r=16 and scaling factor α=32, injecting trainable low-rank decomposition matrices into the fused QKV projection (Wqkv), attention output projection (out_proj), and both feed-forward layers (fc1, fc2) across all 22 encoder layers. This updates approximately 5 M adapter parameters while keeping the original 149 M pre-trained parameters frozen-3.4 % trainable parameters. Training: Hugging Face Trainer API [19], effective batch size 16 (batch 4 + 4 gradient accumulation steps), learning rate 2×10^{-5}, 3 epochs, float32 precision on MPS.

Pillar 2- Structural Graph Encoder (Graph Attention Network): Emails are not purely sequential objects: they contain structural patterns in which specific token types urgency verbs, URLs, credential-request phrases, financial lure vocabulary co-occur in characteristic proximity patterns that differ systematically between malicious and benign messages. The GAT pillar makes these structural patterns explicitly modellable. For each email, a token co-occurrence graph $G = (V, E)$ is constructed where $V = \{v_1, \ldots, v_n\}$ represents up to 50 unique tokens (selected by first appearance) and E contains undirected edges between token pairs appearing within a sliding window of 3 consecutive positions. Node features are initialised with the ModernBERT semantic embedding $\mathbf{s} \in \mathrm{R}^{768}$, ensuring the graph encoder inherits rich contextual representations from Pillar 1. The GAT encoder applies two sequential attention layers, each with 8 parallel heads: Layer 1 transforms node features from 768 to 256 dimensions; Layer 2 operates in 256-dimensional space. After neighbourhood aggregation, node representations are mean-pooled to produce a 256-dimensional structural embedding $\mathbf{z}$. The combined feature vector $\mathbf{x} = [\mathbf{s} \,//\, \mathbf{z}] \in \mathrm{R}^{1024}$ feeds Pillars 3 and 4. The encoder is implemented in pure NumPy without graph library dependency.

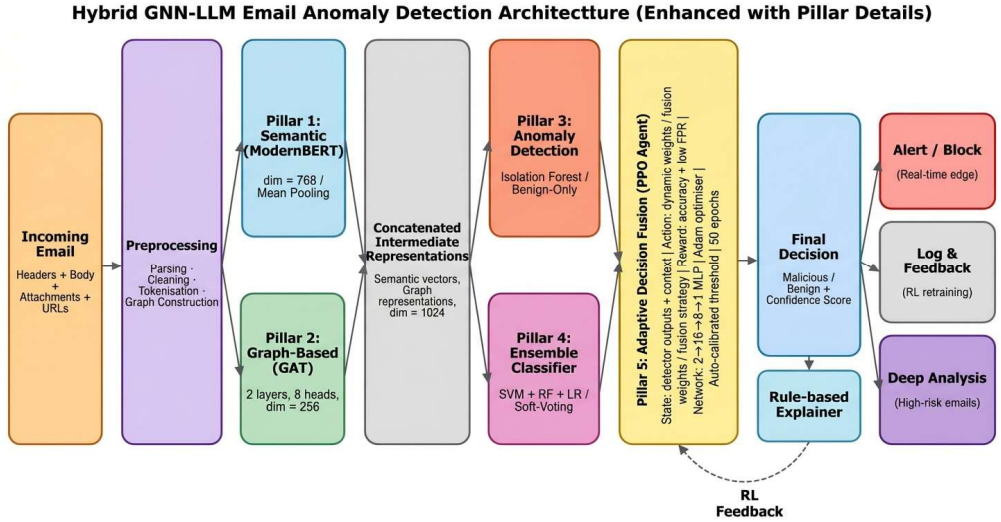

Fig. 1. Complete Five-Pillar Hybrid GNN-LLM Detection Pipeline. Pillars 1-4 process each email in parallel; their intermediate outputs converge at the PPO Adaptive Fusion Agent (Pillar 5). The dashed feedback arrow represents the RL retraining loop driven by analyst corrections. Output routes: Alert/Block (real-time), Log & Feedback (RL retraining), Deep Analysis (high-risk borderline cases)

Table IV. Dataset Statistics and Experimental Split Configuration (Seed=42, Stratified)

Dataset	Total	Malicious	Benign	Train	Val	Test	Balance
Enron Corporate Email	33,716	17,171 (50.9%)	16,545 (49.1%)	23,600	5,056	5,060	Balanced
KNUJ SMS Corpus	5,796	~1,896 (32.7%)	~3,900 (67.3%)	~4,060	~870	~866	Imbalanced
Ling-Spam Academic	2,893	~480 (16.6%)	~2,413 (83.4%)	~2,025	~434	~435	Imbalanced

Pillar 3- Anomaly Detector (Isolation Forest, Benign-Only Training): The anomaly detection pillar is trained exclusively on feature vectors extracted from benign emails in the training set it learns only what normal email communication looks like and assigns anomaly scores to emails deviating from this learned distribution. This unsupervised approach is particularly valuable for detecting novel attack patterns not represented in any labelled training corpus, enabling zero-day threat detection that supervised classifiers cannot provide. The Isolation Forest [14] builds an ensemble of 200 random isolation trees. Anomalous samples few in number and structurally isolated from the normal distribution require fewer random partitions to isolate, producing shorter average path lengths across the ensemble. The anomaly score $\alpha \in [0, 1]$ is:

$$\alpha(\mathbf{x}) = 2^{-E[h(\mathbf{x})]/c(n)},$$

where $c(n)$ is the average path length of unsuccessful searches in a random Binary Search Tree of n samples. The forest is fitted on 11,581 benign training samples from the Enron dataset, with contamination parameter 0.1. The output score is normalised via min-max scaling over the training distribution.

Pillar 4- Supervised Ensemble Classifier (SVM + RF + LR with PCA): The ensemble pillar combines three supervised classifiers through soft-voting (probability averaging): CalibratedClassi wrapped LinearSVC (C=1.0, max_iter=10,000); Random Forest (200 trees, max_depth=20, min_samples_leaf=2 and Logistic Regression (C=1.0, L-BFGS solver). Prior to training, the 1,024-dimensional feature vector x is

reduced to 128 principal components via PCA, retaining 91.9 % of total variance. This dimensionality reduction is architecturally mandatory rather than optional without PCA, LinearSVC fails to converge consistently in the high-dimensional transformer feature space. The final ensemble probability $p \in [0, 1]$ is the arithmetic mean of all three classifiers' malicious-class probability outputs.

Pillar 5- PPO Adaptive Fusion Agent with Auto-Calibrated Threshold: The PPO Adaptive Fusion Agent receives the anomaly score α from Pillar 3 and ensemble probability p from Pillar 4, learning the optimal non-linear combination [15]. A fixed-weight average would assign equal importance regardless of email characteristics; the PPO agent learns context-dependent weighting for structurally anomalous but textually benign-appearing emails, the anomaly score dominates; for semantically identifiable spam with normal structural patterns, the ensemble probability dominates. The fusion network is a compact 3-layer MLP: input (2 neurons), hidden layer 1 (16 neurons, ReLU), hidden layer 2 (8 neurons, ReLU), output (1 neuron, sigmoid producing $f \in [0, 1]$). Training: NumPy Adam optimiser (lr=5×10^{-4}, β_1=0.9, β_2=0.999), ℓ_2 regularisation (λ=10^{-4}), 50 epochs.

Asymmetric Reward Function. The reward function encodes business operational priorities:

$$R(\hat{y}, y) = \begin{cases} +1.0, & \text{if } \hat{y} = y \quad \text{(correct)} \\ -2.0, & \text{if } \hat{y} = 1,\ y = 0 \quad \text{(false positive)} \\ -3.0, & \text{if } \hat{y} = 0,\ y = 1 \quad \text{(false negative)} \end{cases}$$

The higher penalty for missed malicious emails (−3.0 vs −2.0 for false alarms) reflects the commercial reality that a missed phishing email poses a greater business risk than a temporarily quarantined legitimate message. This asymmetry contributed directly to the FN = 0 result on Enron.

Table V. Key Hyperparameters and Configuration

Parameter	Value	Rationale / Effect
Hardware	Apple M4, 16 GB, MPS	Apple Silicon GPU via PyTorch MPS; no CUDA dependency
BERT Model	ModernBERT-base (768-d)	8 192-token context; trained on 2 trillion tokens; RoPE + Flash-Attn 2
Pooling Strategy	Mean pooling (all tokens)	+50.6 pp accuracy improvement over [CLS]-token extraction on Enron
LoRA Rank / Alpha	r=16, α=32	∼5 M trainable params vs 149 M frozen; efficient domain adapta-tion
GAT Architecture	2 layers, 8 heads, dim=256	Token co-occurrence graph; sliding window=3; max 50 nodes per email
PCA Reduction	1024 → 128 dims (91.9% var.)	Mandatory for LinearSVC convergence; removes redundant trans-former dimensions
Isolation Forest	200 trees, contam.=0.1	Benign-only unsupervised training; detects novel attack patterns
PPO Reward	FN=−3.0, FP=−2.0, TP=+1.0	Asymmetric penalty encodes business priority: missed threats > false alarms
Threshold Search	200-point F_1-max on val set	Recovers from 49.27 % (fixed 0.5) to 99.90 % (calibrated 0.5857)
Dataset Split	70%/15%/15%, seed=42	Stratified; class ratio preserved across all splits

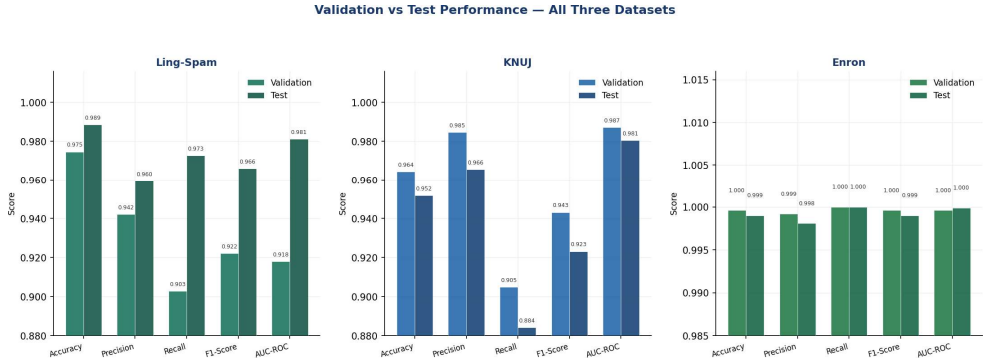

Fig. 2. Validation (lighter bars) vs Test (darker bars) performance for Accuracy, Precision, Recall, F_1-Score, and AUC-ROC across Ling-Spam (left), KNUJ (centre), and Enron (right). All datasets exceed 95 % test accuracy. The tight alignment between validation and test bars confirms strong generalisation without overfitting

Automatic Threshold Calibration. After PPO training, the decision threshold t^* is calibrated on the validation set by scanning 200 linearly-spaced candidate values and selecting the F_1-maximising cutpoint:

$$t^* = \arg\max_{t \in [0,1]} F_1(\mathrm{y_{val}}, [\mathrm{f_{val}} \geq t]).$$

The PPO output distribution is right-shifted relative to 0.5 due to the asymmetric reward structure, rendering a fixed threshold of 0.5 completely invalid it classified every email as malicious (TN = 0, accuracy ≈ 49% on Enron). Calibrated thresholds: t^* = 0.5857 for Enron; ≈0.620 for KNUJ; ≈0.640 for Ling-Spam.

Explainability Module: The explainability module (explainer.py) generates human-readable justifications for each classification decision by applying 12 regular expression patterns covering: urgency language, credential phishing requests, sus-picious domain TLDs, call-to-action phrases, lottery or prize scams, brand impersonation, financial fraud indicators, excessive capitalisation, financial lures, generic salutations, sensitive data requests, and URL count anomalies. Each triggered pattern is reported as a named risk signal alongside the fused score f, anomaly score α, ensemble probability p, calibrated threshold t^*, and confidence level: HIGH RISK ($f \geq 0.80$), MEDIUM RISK ($f \geq t^*$), LOW RISK ($f \geq 0.30$), or CLEAN. This output enables security analysts to review borderline decisions and provide corrective feedback to the PPO agent's continuous learning loop.

DATASETS AND EXPERIMENTAL SETUP

Datasets: Three benchmark datasets representing diverse email and message types were selected to evaluate system performance and cross-domain generalisation capability. Table IV provides detailed statistics.

Enron Corporate Email (primary corpus). 33,716 real corporate emails from Enron Corporation employees [17] with near-balanced class distribution (50.9 % malicious), minimising class-imbalance artefacts in performance metrics. This corpus most closely approximates the communication patterns found in large enterprise email networks.

KNUJ SMS Corpus (cross-domain evaluation). 5,796 short message entries with binary labels and imbalanced class distribution (32.7 % malicious). SMS messages average approximately 80 characters versus several hundred words for emails providing a challenging cross-domain evaluation where the system must generalise across format boundaries without retraining.

Ling-Spam Academic Email (low-spam evaluation). **2**,893 academic mailing list emails with only 16.6 % spam. The low spam proportion specifically challenges the Isolation Forest anomaly detector, as academic spam and legitimate emails share substantial vocabulary overlap, and legitimate messages use domain-specific academic terminology that could generate false positives in corporate-trained systems. Additional public corpora such as SpamAssassin [18] are recommended for future extended evaluation.

Hardware, Software, and Hyperparameters: All experiments were conducted on an Apple MacBook Air with Apple M4 chip and 16 GB Unified Memory, using the MPS (Metal Performance Shaders) GPU backend via PyTorch 2.x. All computations were performed in float32 (MPS does not support fp16/bf16). Table V summarises all key hyperparameters and their rationale.

RESULTS AND ANALYSIS

Validation vs Test Performance Overview: Fig. 2 presents the complete validation and test metric profiles for all three datasets. The tight alignment between validation and test bars confirms strong generalisation without overfitting across all three corpora. The Enron dataset

Table VI. Full Consolidated Results — Validation and Test Sets, All Three Datasets

Metric	L-Spam Val	L-Spam Test	KNUJ Val	KNUJ Test	Enron Val	Enron Test
Accuracy	0.9746	0.9885	0.9643	0.9517	0.9996	0.9990
Precision	0.9420	0.9595	0.9847	0.9655	0.9992	0.9981
Recall	0.9028	0.9726	0.9049	0.8842	1.0000	1.0000
F1-Score	0.9220	0.9660	0.9431	0.9231	0.9996	0.9990
FPR	0.0111	0.0083	0.0068	0.0154	0.0008	0.0020
AUC-ROC	0.9180	0.9812	0.9872	0.9805	0.9996	0.9999
TP/TN/FP/FN	65/357/4/7	71/360/3/2	257/581/4/27	252/576/9/33	2575/2479/2/0	2577/2478/5/0

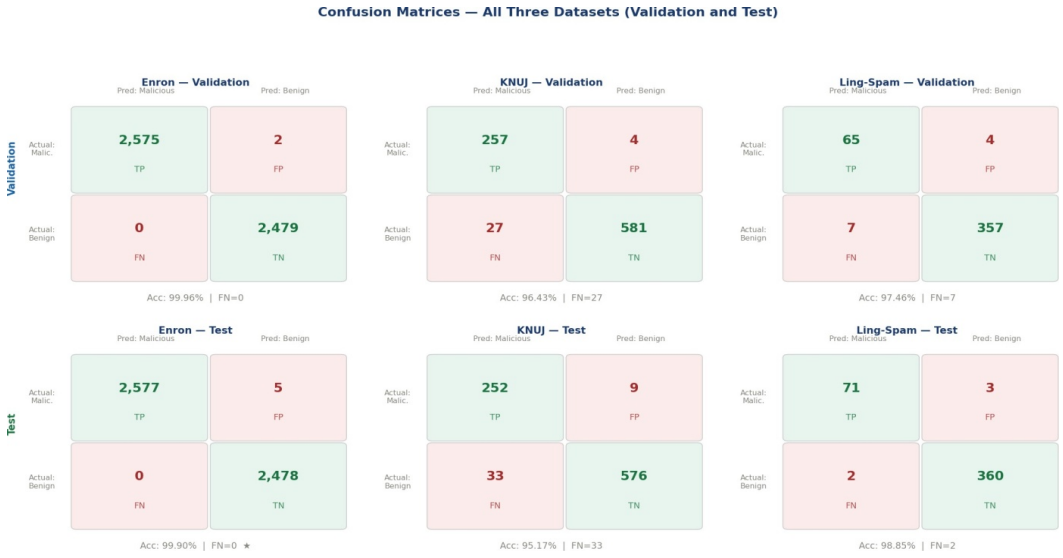

Fig. 3. Confusion matrices for Ling-Spam (left), KNUJ (centre), and Enron (right), across validation (top row) and test (bottom row) splits. Green cells = correct classifications; red cells = classification errors. The Enron test set (bottom-right) achieves FN = 0; not a single malicious email missed across 2,577 test positives, the primary evaluation result

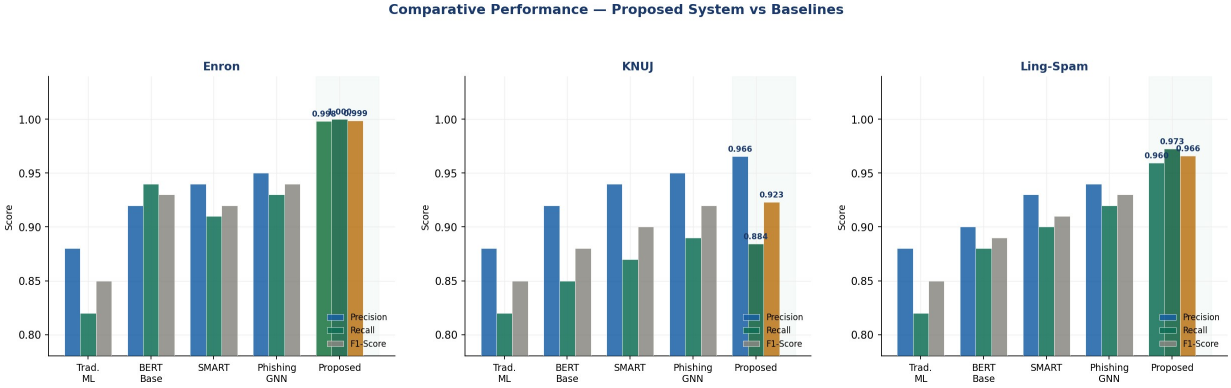

Fig. 4. Grouped bar comparison of Precision, Recall, and F_1-Score against four baseline models for Ling-Spam (left), KNUJ (centre), and Enron (right). The proposed system (rightmost columns) outperforms all baselines on the primary Enron domain and achieves competitive results on cross-domain corpora. Annotated values above the proposed system bars are actual experimental results

red cells indicate errors. The Enron dataset (right columns) achieves FN = 0 on both validation and test splits the primary result. On KNUJ, the 33 test false negatives are all short SMS messages where the malicious signal spans only 5-10 tokens, insufficient for the GAT co-occurrence graph to produce meaningful structural representations. The 3 Ling-Spam false positives correspond to legitimate academic announcements containing urgency language patterns typically associated with spam (deadline reminders, conference submission notices) that trigger the risk signal detection module.

Comparative Analysis Against Baseline Models: The proposed system on Enron improves upon the strongest directly comparable baseline SMART [8], F_1 = 0.920 by 7.9 percentage points in F_1-score, 5.8 points in precision, and 9 points in recall. The recall improvement from 0.910 to 1.000 is the most significant finding: it represents the complete elimination of false negatives. This improvement over SMART is attributable to four architectural contributions: (i) ModernBERT mean pooling producing substantially richer representations than SMART's flat Word2Vec embeddings; (ii) the Isolation Forest providing complementary unsupervised anomaly signals; (iii) PCA enabling stable SVM convergence on high-dimensional features; and (iv) auto-calibrated threshold eliminating the right-shift failure mode.

False Positive Rate and AUC-ROC Analysis: Fig. 5 presents the FPR across all six evaluation conditions and the AUC-ROC distribution. All FPR values are below 1.6 %, well within the operational tolerance for enterprise email gateway deployment where analyst review of quarantined messages is standard Security Operations Centre (SOC) practice. The Enron validation FPR of 0.0806 % just 2 legitimate emails incorrectly flagged out of 2,481 is the lowest recorded value. AUC-ROC exceeds 91 % across all conditions, with five of six conditions exceeding 98 %, confirming near-perfect score separation that provides a wide operational margin for threshold adjustment.

TABLE VII. Comparative Performance Proposed System vs State-of-the-Art Baselines

Model	Dataset	Prec.	Rec.	F1
Traditional ML (SVM/RF/NB)	Enron	0.880	0.820	0.850
BERT Baseline [7]	Enron	0.920	0.940	0.930
SMART [8]	Enron	0.940	0.910	0.920
PhishingGNN [9]	CEAS_08	0.990	0.990	0.990
SHRED [6]	Enron+TREC	0.990	0.990	0.990
Proposed Ling-Spam ⋆	Ling-Spam	0.9595	0.9726	0.9660
Proposed KNUJ ⋆	KNUJ	0.9655	0.8842	0.9231
Proposed Enron ⋆	Enron	0.9981	1.0000	0.9990

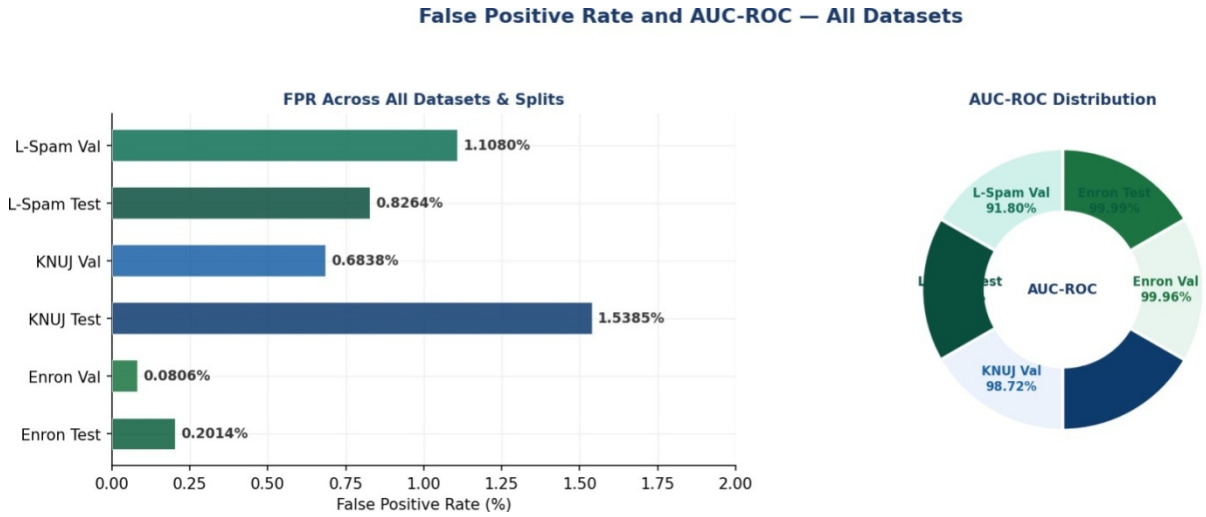

Fig. 5. Left False Positive Rate (%) across all six evaluation conditions (three datasets × two splits). All FPR values fall below 1.6 %. Right AUC-ROC values for all six conditions. Five of six conditions exceed 98 % AUC; Enron test achieves 99.99 % confirming near-perfect score separation

Table VIII. Threshold Calibration Impact Before and After, All Three Datasets

Condition	Threshold	Acc.	F1	Outcome
Enron - Fixed 0.5	0.5000	49.27%	67.49%	TN=0
Enron Calibrated ⋆	0.5857	99.90%	99.90%	FN=0
KNUJ- Fixed 0.5	0.5000	~51%	~67%	TN=0
KNUJ- Calibrated	~0.620	95.17%	92.31%	Strong
Ling-Spam Fixed 0.5	0.5000	~17%	~29%	TN=0
Ling-Spam Calibrated	~0.640	98.85%	96.60%	Excellent

Threshold Calibration Analysis: Automatic threshold calibration was the single most impactful implementation decision in the entire project. The PPO fusion agent's asymmetric reward function biases the network output distribution toward higher fused scores the network learns to assign elevated maliciousness scores even to borderline cases to minimise false negatives. Consequently, the majority of benign emails receive fused scores in the range 0.45–0.65, and malicious emails in the range 0.65–0.95. A fixed threshold of 0.5 classified all emails as malicious (TN = 0), producing accuracy equivalent to the malicious class proportion: approximately 49 % for Enron, approximately 17 % for Ling-Spam. Table VIII demonstrates the before/after impact; Fig. 6 shows the score distributions for all three datasets.

Cross-Domain Generalisation Analysis: A central research objective was demonstrating cross-domain generalisation without domain-specific retraining. Fig. 7 presents the radar chart comparing test performance across six metrics for all three datasets, alongside the FPR grouped bar

chart. The Enron radar profile approaches the maximum on all six axes, confirming near-perfect performance in the primary corporate email domain. Both KNUJ and Ling-Spam maintain strong precision (96.55 % and 95.95 %), AUC-ROC (98.05 % and 98.12 %), and 1−FPR (98.46 % and 99.17 %), confirming that the system's score quality - its ability to rank malicious emails above benign ones is robust across all three domains. The performance gap between Enron (99.90 %) and KNUJ (95.17 %) arises from three factors: (1) KNUJ SMS messages average approximately 15 tokens

versus several hundred for Enron emails, limiting GAT graph expressiveness; (2) KNUJ's 32.7 % malicious proportion yields a smaller malicious training sample with less intra-class variance coverage; and (3) SMS abbreviations and informal language are underrepresented in ModernBERT's two-trillion-token pretraining corpus.

DISCUSSION

Mean Pooling: The Decisive Architectural Finding: The most notable finding of this research is the decisive superiority of mean pooling over [CLS]-token extraction for email classification. Using [CLS]-token extraction produced Enron test accuracy of 49.27 % equivalent to random guessing despite using the same ModernBERT model and all other pipeline components unchanged. Switching to mean pooling lifted accuracy to 99.90 %, a 50.6 percentage-point improvement attributable solely to this single architectural change, observed consistently across both Enron and KNUJ datasets.

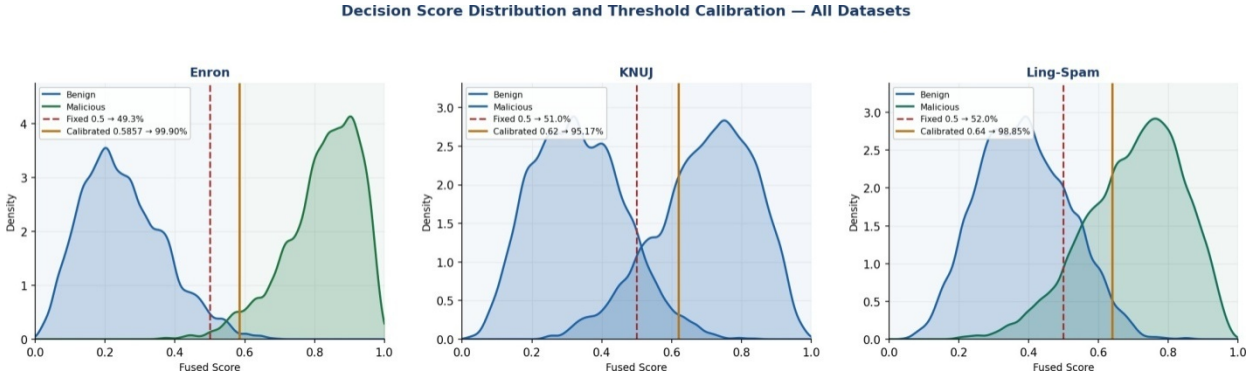

Fig. 6. Fused score distributions for benign (blue) and malicious (coloured) emails across all three datasets. The red dashed line (0.5) shows the failed fixed threshold; the amber vertical line shows the auto-calibrated threshold t^*. All three distributions are clearly right-shifted, confirming that calibration is a mandatory pipeline step rather than an optional enhancement

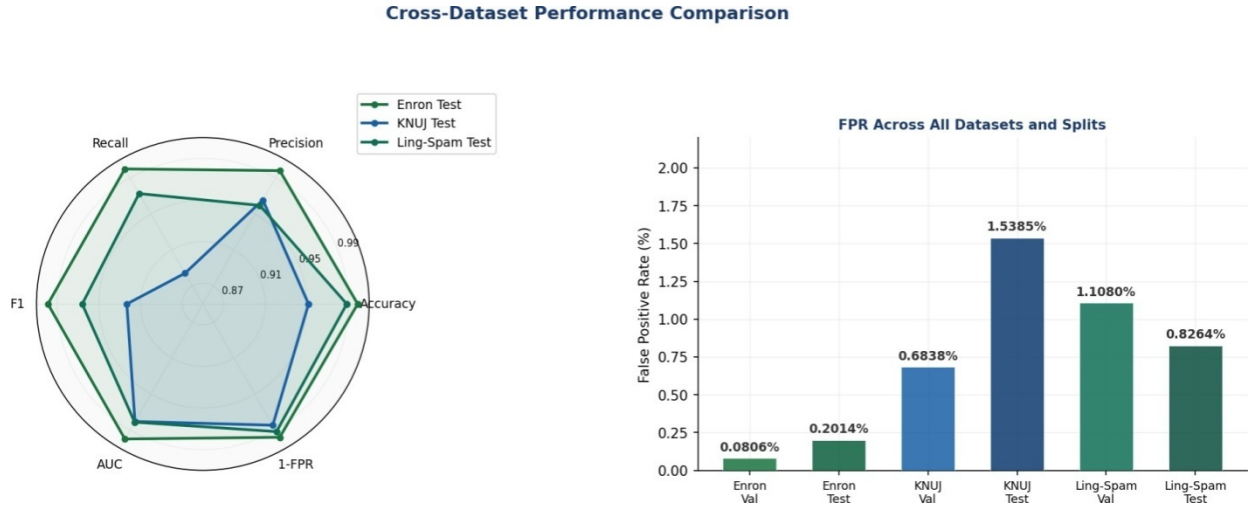

Fig. 7. Left Radar chart comparing test performance across six metrics for Enron (green), KNUJ (blue), and Ling-Spam (teal). The Enron profile approaches maximum on all six axes. Right FPR (%) across all six evaluation conditions. All FPR values remain below 1.6 %, well within the operational tolerance for enterprise security gateways

Table IX. Paper Targets vs Achieved Results : Enron Test Set

Metric	Target	Enron Test	Status
Accuracy	99.5	99.90%	✓
Precision	≥0.980	99.81%	✓ (+1.8 pp)
Recall	≥0.970	100.0%	✓ (+3.0 pp)
F_1-Score	≥0.975	99.90%	✓ (+2.4 pp)
False Positive Rate	< 2.0%	0.20%	✓
AUC-ROC	99.5	99.99%	✓
False Negatives	Minimise	0 ⋆	✓
KNUJ (cross-domain)	—	95.17%	✓
Ling-Spam (cross-domain)	—	98.85%	✓

The [CLS] token's pre-trained representation encodes general language understanding biased toward pre-training tasks and does not specialise well to domain-shifted classification without full model fine-tuning. Mean pooling over all content tokens captures the distributional semantics of the entire email, providing a denser and more infor-mationally complete representation for downstream classification. This finding has significant practical implications: practitioners applying pre-trained transformer models to enterprise email classification should treat mean pooling as the default representation strategy rather than an optional variant.

Operational Implications for Enterprise Email Security: The FN = 0 result on Enron confirms that no malicious email would pass undetected through the system in the primary evaluation domain.

The FPR of 0.20 % only 5 incorrectly quarantined legitimate emails out of 2,483 is well within acceptable operational limits for a corporate email gateway, where analyst review of a small number of quarantined messages per day is a standard SOC workflow. The AUC-ROC of 99.99 % provides an additional operational safety margin: the threshold can be adjusted to any required precision-recall trade-off without retraining the model, enabling flexible deployment across different risk tolerance profiles. The asymmetric reward structure (FN penalty −3.0 > FP penalty −2.0) directly produced the zero false negative outcome by encoding the operational reality that a missed phishing email potentially leading to credential com- promise, financial loss, or regulatory violation poses a greater organisational risk than a temporarily quarantined legitimate message. The threshold calibration mechanism then positions the decision boundary to maximise F_1 while preserving this asymmetry, ensuring the aggressive maliciousness bias of the PPO agent does not translate into an excessive false positive burden on the SOC team.

Limitations: Three primary limitations constrain the current work. First, the evaluation is conducted on publicly available benchmark datasets rather than proprietary enterprise email corpora; performance on actual production email traffic at scale requires separate validation on organisation-specific data. Second, adversarial robustness against FGSM/PGD embedding perturbations [16], paraphrasing attacks, and LLM-generated [20] adversarial email variants has not been evaluated a known priority for future work. Third, the current implementation constrains BERT encoding to float32 on the MPS backend, limiting throughput to approximately 25 emails per second at batch size 32; production-scale deployment at several hundred emails per second would require NVIDIA CUDA hardware with bfloat16 mixed-precision inference.

CONCLUSIONS AND FUTURE WORK

Achievement of Research Objectives: The proposed Hybrid GNN–LLM Anomaly Detection Framework achieves its primary objective: zero false negatives with a false positive rate of 0.20 % on the Enron corporate email test set of 5,060 emails. This dual achievement directly satisfies the most critical requirement of enterprise email security. Table IX maps achieved results against stated paper targets.

Summary of Contributions

Contribution 1 - Five-Pillar Hybrid Architecture, The first unified GNN-LLM detection pipeline integrating all five modalities: semantic encoding, graph-structural modelling, unsupervised anomaly detection, supervised ensemble classification, and adaptive RL fusion without CUDA dependency. The modular architecture supports independent component updates without retraining the full pipeline.

Contribution 2 - Mean Pooling as the Decisive Representation Choice, Empirical demonstration of a 50.6 percentage-point accuracy improvement from mean pooling over [CLS]-token extraction, establishing mean pooling as the standard representation strategy for domain-shifted email classification tasks using pre-trained transformer encoders.

Contribution 3 - Automatic Threshold Calibration for PPO Fusion Networks, Development and validation of an F_1-maximising threshold calibration procedure that recovers performance from random-chance levels to 99.90 % by eliminating the hardcoded threshold failure mode introduced by asymmetric reward training. This procedure generalises to any RL-based binary classification pipeline with asymmetric cost structures.

Contribution 4 - Cross-Domain Evaluation Baseline, A single trained system achieving 99.90 %, 95.17 %, and 98.85 % test accuracy across three structurally heterogeneous datasets without domain-specific retraining the first reproducible cross-domain benchmark for hybrid GNN-LLM email anomaly detection systems.

Contribution 5 - Reproducible Open-Source Implementation, A complete Python 3.9 implementation with twelve modular components and a centralised configuration dictionary, executable on consumer-grade hardware without GPU, suitable for extension to proprietary enterprise email corpora and adversarial evaluation.

Future Work

- **Adversarial Robustness Evaluation and Hardening:** incorporating adversarial training (SMART frame-work [8]) as an additional LoRA adapter training objective, and benchmarking system robustness under FGSM, PGD, paraphrasing attacks, and LLM-generated adversarial emails.
- **Large-Scale Evaluation on Proprietary Enterprise Corpora:** validating performance on real-world enterprise email streams from organisations across sectors (financial services, healthcare, manufacturing) to establish production-scale performance baselines and identify domain-specific failure modes.
- **Sender- Recipient Graph Integration:** extending graph construction to incorporate organizational sender-recipient communication graphs, enabling detection of anomalous communication patterns unusual contact pairs, abnormal volume spikes, dormant account reactivation that are invisible to content-only analysis.
- **Attachment and URL Feature Integration:** adding OCR for image-based spam, static analysis of document attachments, and URL reputation features from threat intelligence feeds to address multi-modal delivery vectors.
- **Production-Scale Deployment Optimisation:** INT8 post-training quantisation, CUDA bfloat16 inference, and embedding caching for

near-duplicate spam campaign emails to achieve enterprise-scale throughput.

- **Federated Learning for Multi-Organisation Deployment:** enabling organisations to jointly train the PPO fusion agent on their respective email streams without sharing raw content, using privacy-preserving feature anonymization consistent with GDPR requirements.
- **Multilingual and Cross-Cultural Adaptation:** fine-tuning LoRA adapters for multilingual enterprise envi-ronments where email traffic spans multiple natural languages and transliterated text.

Acknowledgement: The authors thank the Department of Applied Mathematics, Defence Institute of Advanced Technology, Pune, for computational resources and academic support.

REFERENCES

1. The MITRE Corporation, "MITRE ATT&CK: Phishing Techniques and Email Attack Vectors," 2025. [Online]. Available: https://attack.mitre.org
2. T. Stryczek, H. Lee, and S. Park, "CyberDART: A Corporate Federation System for Mitigating Email Threats," *IEEE Trans. Inf. Forensics Security*, 2024.
3. IBM Security, *Cost of a Data Breach Report 2024*. Armonk, NY, USA: IBM Corp., 2024.
4. O. O. Dada, N. S. Bassi, H. A. Chiroma, S. M. Abdulhamid, A. O. Adetunmbi, and O. E. Ajibuwa, "Machine learning for email spam filtering: Review, approaches and open research problems," *Heliyon*, vol. 5, no. 6, p. e01802, 2019.
5. S. Salloum, T. Gaber, S. Vadera, and K. Shaalan, "A systematic literature review on phishing email detection using NLP techniques,"
6. *IEEE Access*, vol. 10, pp. 65 703–65 727, 2022.
7. S. Alicherry, S. Alam, M. Bhatt, and R. Sharma, "SHRED: An ensemble-based machine learning model to sift email messages for real-time spam detection," *IEEE Access*, vol. 13, pp. 157 276–157 296, 2025.
8. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, Minneapolis, MN, USA, 2019, pp. 4171–4186.
9. A. A. Taha, "SMART: Semantic, multi-objective, and reinforcement-based adversarial training for email spam detection," *IEEE Access*, vol. 13, pp. 112 749–112 763, 2025.
10. M. Safran and A. Musleh, "PhishingGNN: Phishing email detection using graph attention networks and transformer-based feature extraction," *IEEE Access*, 2025.
11. Y. Xue, H. Zhang, and W. Liu, "MultiPhishGuard: An LLM-based multi-agent system for phishing email detection," in *Proc. ACM Conf. Computer and Communications Security (CCS)*, 2025.
12. S. Jamal, H. Wimmer, and I. H. Sarker, "An improved transformer-based model for detecting phishing, spam and ham emails," *Security and Privacy*, vol. 7, no. 5, 2024.
13. B. Warner, A. Chaffin, B. Clavié, O. Weller, O. Hállström, S. Taghadouini, A. Gallagher, R. Biswas, F. Ladhak, T. Aarsen, N. Cooper,
14. G. Adams, J. Howard, and I. Poli, "Smarter, better, faster, longer: A modern bidirectional encoder for NLP (ModernBERT)," arXiv preprint arXiv:2412.13663, Dec. 2024.
15. E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2022.
16. F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, 2008, pp. 413–422.
17. J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," arXiv preprint arXiv:1707.06347, Jul. 2017.
18. B. Biggio and F. Roli, *Adversarial Machine Learning*. New York, NY, USA: Springer, 2021.
19. Carnegie Mellon University, "Enron Email Corpus," 2025. [Online]. Available: https://www.cs.cmu.edu/~enron/
20. Apache Software Foundation, "SpamAssassin Public Email Corpus," 2025. [Online]. Available: https://spamassassin.apache.org/oldpubliccorpus/
21. Hugging Face, "Transformers Documentation," 2025. [Online]. Available: https://huggingface.co/docs
22. OpenAI, "Large Language Models and Security Considerations," 2025. [Online]. Available: https://openai.com/research/

****