IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, May 2026
Impact Factor: 8.2

EmailGuard-FLIF: Privacy-Aware Federated and Hybrid Email Threat Detection Using Isolation Forest Fusion

Kunal Masurkar, Adhiraj Nimbalkar, Rushikesh Gaikwad, Aryan Raval, Smita Gumaste
Department of Computer Science
MIT ADT University Pune, India
kunalmasurkar26, adhirajrajenimbalkar24, rushikesh2264@gmail.com
aryancraval@gmail.com smita.gumaste@mituniversity.edu.in

Abstract: *Enterprise email remains the primary delivery mechanism for phishing, credential theft, and business email compromise, yet centralizing email telemetry across departments for model training violates privacy regulations. This paper introduces EmailGuard-FLIF, a hybrid threat detection system that integrates: (i) federated supervised learning combined with privacy-preserving update aggregation among 23 departmental clients, (ii) unsupervised Isolation Forest anomaly detection for zero-day sensitivity, and (iii) a cascaded three-stage decision fu- sion layer combining supervised, anomaly, and heuristic evidence streams. The supervised branch builds a multimodal feature rep- resentation including TF-IDF text features (50,000 dimensions), behavioral telemetry (7 numeric features), and identity/context signals (5 categorical features). Privacy is ensured through L2 norm clipping (C=1.0) and Gaussian perturbation (σ=0.05) on all local updates. Experiments on a large-scale enterprise email dataset (2.63M samples, 525,997 test instances) show that the federated model achieves 93.35% AUC, 93.90% threat recall, and 73.17% F1-score—outperforming the best centralized baseline by +5.12 AUC points and +14.03 F1 points. Multi-seed analysis confirms low variance (F1 std < 0.011).*

Keywords: Email threat detection, federated learning, privacy-preserving ML, Isolation Forest, anomaly detection, de- cision fusion, cybersecurity

I. INTRODUCTION

Email remains the most persistently exploited enterprise attack surface, with over 90% of successful cyber-attacks originating from email-borne vectors [1]. Contemporary at- tack campaigns combine linguistic deception with contextual camouflage—impersonating trusted senders, exploiting orga- nizational hierarchies, and timing attacks during off-hours when security oversight is diminished. Traditional rule-based filters and single-feature classifiers are insufficient against such adaptive adversaries.

Effective detection demands the integration of multiple evidence channels: content-level signals capture linguistic at- tack fingerprints, behavioral signals reveal operational anoma- lies, and identity/context signals provide organizational prior knowledge. A multimodal approach that fuses all three signal families offers substantially richer discriminative power than any single channel.

A second fundamental challenge is collaborative learning under privacy constraints. In large enterprises, departments generate sensitive email telemetry with distinct distributional characteristics. Centralizing this data violates regulatory man- dates (GDPR, HIPAA) and organizational policies. Federated learning [3], [4] enables collaborative training by exchanging only model parameter updates, preserving data locality while enabling cross-organizational knowledge transfer. However, even model updates can leak information about local training data [5], necessitating practical privacy controls such as norm clipping and noise injection.

DOI: 10.48175/IJARSCT-34746

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, May 2026 Impact Factor: 8.2

A third challenge is zero-day threat detection. Supervised classifiers are inherently bounded by historically observed attack patterns. Unsupervised anomaly detectors offer sensi- tivity to distribution shift but generate excessive false alarms. Hybrid fusion systems combining both paradigms can leverage complementary strengths with carefully calibrated fusion rules.

This work addresses these challenges through EmailGuard- FLIF, a unified, implementation-grounded architecture vali- dated on a large-scale enterprise email corpus from the CERT Insider Threat Test Dataset [14] (2.63M email records, 23 de- partments). The system is designed for practical deployment, providing serialized model artifacts, CLI-driven experimenta- tion, and transparent fusion impact diagnostics. The principal contributions are:

1) A multimodal representation jointly modeling email text (TF-IDF, 50K features), behavioral telemetry (7 numeric features), and identity/context signals (5 categorical fea- tures) through a unified ColumnTransformer pipeline.

2) A federated training procedure with FedAvg-style ag- gregation and practical privacy controls (L2 clipping + Gaussian noise) across 23 departmental clients.

3) A hybrid detection architecture combining supervised threat classification, Isolation Forest anomaly detec- tion (trained on 1.75M normal-class samples), and a cascaded three-stage decision fusion with confidence max-pooling across supervised, anomaly, and heuristic signals.

II. RELATED WORK

A. Email Threat and Insider Detection

Machine learning for phishing detection has evolved from handcrafted lexical features [1], [2] to deep contextual models based on transformer architectures [15] that capture subtle lin- guistic cues such as urgency manipulation, authority imperson- ation, and social proof exploitation. However, the predominant emphasis on content-level features has left the detection of behavioral anomalies (e.g., unusual sending times, abnormal attachment volumes) and organizational context mismatches (e.g., a low-privilege user sending high-sensitivity attachments externally) largely unaddressed. In insider threat detection, Yuan and Wu [7] provide a comprehensive survey of deep learning methodologies, identifying feature representation and labeled data scarcity as persistent challenges. Yi and Tian [8] demonstrate that hybrid unsupervised-supervised architectures outperform either paradigm in isolation—a finding that di- rectly informs our hybrid fusion design. Recent federated approaches include parameter-efficient LLM tuning with Fed- ITD [9], FedAT [10] for adversarial training robustness, multi- agent LLM collaboration [11] for log-based detection, and test- time training adaptation [12]. Homoliak et al. [13] provide a foundational taxonomy of insider threat analysis and counter- measures. Despite these advances, few approaches integrate multimodal features, practical federated privacy controls, and anomaly fusion within a single deployable pipeline.

B. Federated Learning and Anomaly Detection

Federated Averaging [3] enables decentralized training; Kairouz and McMahan [4] identify key open problems in- cluding non-IID heterogeneity and privacy guarantees. Wei et al. [5] analyze convergence under differential privacy. Isolation Forest [6] offers linear-time unsupervised anomaly detection particularly suited to high-dimensional mixed-type data, but produces high false positive rates when deployed independently—motivating its use as a complementary branch within a hybrid architecture.

C. Research Gap

Three gaps remain: (1) limited multimodal integration across text, behavior, and identity features; (2) limited practical FL with demonstrated privacy-utility tradeoffs across >20 clients; and (3) limited operational fusion of supervised and anomaly signals with diagnostic transparency. EmailGuard- FLIF addresses all three.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, May 2026
Impact Factor: 8.2

III. SYSTEM MODEL AND PROBLEM FORMULATION

A. Problem Definition

Let D = {D1, . . . , DK} denote departmental email datasets. For each sample xi:

$$x_i = (t_i, b_i, c_i),\ y_i \in \{0, 1\} \quad (1)$$

where $t_i \in \mathbb{R}^{d_t}$ is the TF-IDF text representation, $b_i \in \mathbb{R}^7$ the behavioral features, and c_i the categorical identity/context tuple.

B. Federated Learning Objective

The global objective is weighted empirical risk minimiza- tion:

where $N = \sum_k n_k$ and $L_k(\theta)$ is the local empirical loss. FedAvg approximates the solution via alternating local SGD and weighted parameter averaging.

C. Hybrid Decision Objective

The fusion layer implements a cascaded three-stage deci- sion protocol rather than a simple disjunction, reflecting the operational requirement for controlled alert escalation:

Stage 1 (Supervised gate): The supervised branch predic- tion is accepted only when both the classifier predicts threat and the fused risk confidence exceeds a fusion threshold τ_f :

$$\hat{y}_{s1} = I[\hat{y}_{sup} = 1 \wedge \hat{r} \geq \tau_f] \quad (3)$$

Stage 2 (Anomaly override): If Stage 1 does not fire, the anomaly branch can escalate the prediction under constrained conditions—requiring high anomaly confidence (≥ 0.55) and corroborating evidence from either the heuristic scorer or a relaxed supervised threshold:

$$\hat{y}_{s2} = I[\hat{y}_{s1}=0 \wedge \hat{y}_{anom}=1 \wedge p_{anom} \geq 0.55 \wedge (h=1 \vee p_{sup} \geq \tau') \quad (4)$$

where $\tau' = \max(0.35, 0.6 \cdot \tau_{sup})$ and h is the heuristic alert. Stage 3 (Heuristic backstop): A rule-based heuristic scorer independently flags samples with high phishing indicator con- fidence (≥ 0.65), providing defense-in-depth against adversarial model evasion:

$$\hat{y} = \max(\hat{y}_{s1}, \hat{y}_{s2}, h) \quad (5)$$

The risk confidence score is computed as the maximum across all three evidence streams:

$$\hat{r} = \max(p_{sup}, p_{anom}, p_{heur}) \quad (6)$$

This cascaded design ensures that each successive stage re- quires progressively stronger corroborating evidence, balanc- ing recall sensitivity with false positive control.

IV. PROPOSED METHOD

A. Data Preparation and Feature Engineering

The pipeline ingests raw email files and structured CSV data, performing header stripping, text normalization (HTML removal, URL/email/number tokenization, lowercasing), and LDAP identity enrichment. When explicit labels are absent, a heuristic scoring function assigns provisional labels based on suspicious keywords, external sender patterns, and attachment behavior.

The supervised preprocessor is a three-branch ColumnTransformer:

1) Text branch: TF-IDF with unigrams/bigrams, sublinear TF, up to 50,000 features.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, May 2026
Impact Factor: 8.2

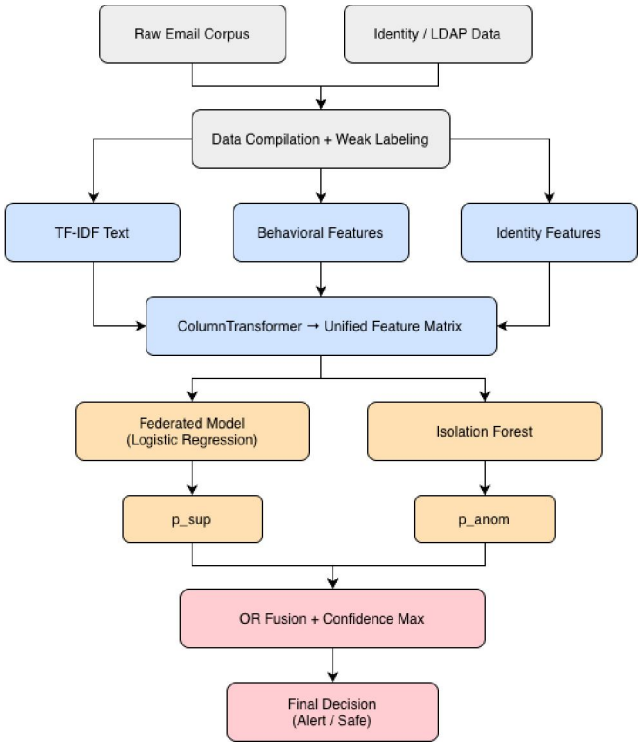

Fig. 1. End-to-end EmailGuard-FLIF architecture: data compilation, multi- modal feature engineering, federated supervised training with privacy controls, Isolation Forest anomaly branch, and cascaded three-stage decision fusion.

2) Numeric branch: StandardScaler (sparse-safe) over 7 behavioral features.
3) Categorical branch: OneHotEncoder over 5 iden- tity/context features.
The composite feature map is:
$\phi(x_i) = [\phi_{tfidf}(t_i); \phi_{num}(b_i); \phi_{cat}(c_i)]$ (7)

B. Federated Supervised Training

The federated procedure uses SGDClassifier with lo- gistic loss and L2 regularization (α=10−5). The global model is warm-started on a balanced subsample (1,024 samples per class) drawn across departments.

For each round r, the server distributes parameters w(r) to all clients. Each client k performs local SGD, computing the update delta $\Delta(r) = w(r) - w(r)$. Before transmission, privacy controls are applied:

$\tilde{\Delta}(r) = \Delta(r) \cdot \min 1, \; C \; + N(0, \sigma 2I)$ (8)

$\Delta(r) \quad 2 + \epsilon$

The server aggregates via weighted averaging:

TABLE I: MODEL AND TRAINING HYPERPARAMETERS

Parameter	Value	Description
test size	0.2	Stratified split ratio
max features	50,000	TF-IDF vocabulary size
fl rounds	5	Aggregation rounds
local epochs	1	Local SGD passes
dp clip norm (C)	1.0	L2 clipping threshold
dp noise std (σ)	0.05	Gaussian noise std
if n estimators	300	IF tree count

DOI: 10.48175/IJARSCT-34746

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, May 2026
Impact Factor: 8.2

if contamination sgd alpha	0.02	Anomaly fraction
	10−5	L2 regularization

C. Anomaly Branch and Fusion Layer

The Isolation Forest is trained on 1,748,499 normal-class samples (300 estimators, contamination 0.02), with threshold optimization via Fβ (β=0.25). At inference:

$\hat{y}_{anom} = I[s(x) < \tau_{anom}]$ (10)

Anomaly confidence is computed via sigmoid transform over the negative decision score, centered at the optimized thresh- old: $p_{anom} = \sigma(-(s(x) - \tau_{anom}) \cdot 8)$.

The fusion layer implements the cascaded three-stage protocol from Eqs. (3) – (6). Stage 1 gates the supervised prediction on fused risk confidence ($\tau_f = 0.6$). Stage 2 allows the anomaly branch to override only when supported by high anomaly confidence (≥ 0.55) and corroborating heuristic or relaxed supervised evidence. Stage 3 provides a rule-based heuristic backstop (phishing keyword confidence ≥ 0.65) as defense-in- depth. The fused risk confidence $\hat{r} = \max(p_{sup}, p_{anom}, p_{heur})$ reflects the strongest evidence across all three streams.

V. EXPERIMENTAL SETUP

A. Dataset and Implementation

The pipeline is implemented in Python 3.12 using scikit- learn (classifiers, preprocessing, metrics), pandas (data manipulation), numpy (numerical operations), and joblib (artifact serialization). The system provides a modular CLI interface supporting data compilation, federated training, model evalu- ation, single-sample inference, and exploratory data analysis. Experiments use the CERT Insider Threat Test Dataset [14] (release 4.2), which contains enterprise email, file access, device, and LDAP records. The compiled dataset comprises approximately 2.63 million email samples: 2,103,984 training samples (per federated round), 525,997 test samples (stratified 80/20 split), class distribution ∼83.1% safe / 16.9% threat, across 23 organizational departments including Accounting, Engineering, Sales, Security, Medical, Research, and others. The anomaly branch uses 1,748,499 normal-class training samples.

B. Evaluation Protocol

Threat class (label=1) is the positive class. Metrics include precision, recall, F1, AUC, accuracy, and confusion matrix. After training, the supervised threshold is optimized using Fβ (β=0.5).

The hybrid evaluator reports supervised-only, anomaly-only, and fused metrics with fusion impact diagnostics.

TABLE II: CENTRALIZED BASELINE COMPARISON (THREAT CLASS)

Model	Prec. (%)	Rec. (%)	F1 (%)	AUC (%)	Acc. (%)
Linear SVM	43.23	93.59	59.14	88.23	78.15
Logistic Reg.	43.55	91.10	58.93	88.17	78.55
Multinomial NB	45.47	54.42	49.54	70.58	81.27

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, May 2026
Impact Factor: 8.2

Branch	Prec. (%)	Rec. (%)	F1 (%)	AUC (%)	Acc. (%)
Fed. Global	59.94	93.90	73.17	93.35	88.36
Sup. (Tuned)	62.34	93.08	74.67	—	89.33
Anomaly (IF)	16.73	79.22	27.63	—	29.87

TABLE III: BRANCH-LEVEL PERFORMANCE (TEST SET: 525,997 SAMPLES)

VI. RESULTS

A. Centralized Baselines

To establish a performance reference, three centralized classifiers were trained offline on the same multimodal feature representation without federated aggregation or privacy con- trols (Table II).1 Linear SVM and Logistic Regression achieve recall above 91% but with moderate precision (~43%), while Multinomial NB achieves higher precision but only 54.42% recall. These baselines motivate the federated approach.

B. Federated Global Model

The federated model was trained across 23 clients over 5 rounds with L2 clipping (C=1.0) and Gaussian noise (σ=0.05). Table III reports branch-level metrics. The federated global model achieves 93.35% AUC, a +5.12 point improvement over the best centralized baseline, despite privacy-preserving update obfuscation. This result is particu- larly noteworthy because it suggests that cross-departmental knowledge aggregation more than compensates for the infor- mation loss introduced by clipping and noise. The 93.90% threat recall ensures that only 5,417 out of 88,872 threats are missed—a 6.1% miss rate acceptable for high-sensitivity se- curity operations. Post-training threshold optimization via Fβ (β=0.5) further improves precision from 59.94% to 62.34% with minimal recall sacrifice (93.90% to 93.08%).

The Isolation Forest anomaly branch exhibits the charac- teristic precision-recall profile of unsupervised detectors: high recall (79.22%) reflecting strong sensitivity to distributional deviations, coupled with low precision (16.73%) due to the large volume of benign-but-unusual email behavior. This is by design—the anomaly branch serves as a complementary de- tection layer that catches threats the supervised model misses, particularly zero-day attacks outside the training distribution. Multi-seed stability analysis confirms robust training: supervised F1 varies by only ±0.003 (0.39% relative) and global 1Baseline models were trained in separate offline experiments using identi- cal feature pipelines; the training scripts are not part of the current federated pipeline codebase.

Metric	Centr. SVM	Fed. Global	Δ
AUC (%)	88.23	93.35	+5.12
Recall (%)	93.59	93.90	+0.31
Precision (%)	43.23	59.94	+16.71
F1 (%)	59.14	73.17	+14.03
Accuracy (%)	78.15	88.36	+10.21

TABLE IV: CENTRALIZED VS. FEDERATED PERFORMANCE

DOI: 10.48175/IJARSCT-34746

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, May 2026
Impact Factor: 8.2

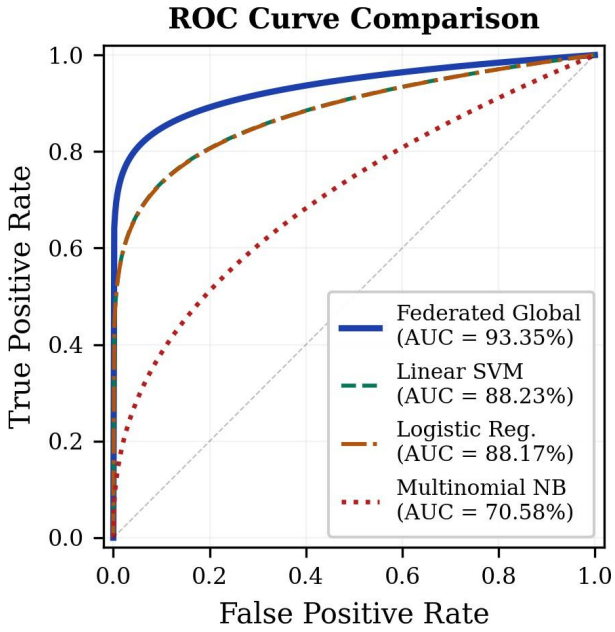

Fig. 2. ROC comparison: federated global model vs. centralized baselines.

F1 by ±0.010, validating that balanced warmup initialization and FedAvg aggregation produce consistent models regardless of random seed.

C. Federated vs. Centralized Comparison

Table IV directly compares the best centralized baseline against the federated model on identical features.

The most striking gain is in precision (+16.71 points), directly reducing false positive burden. We attribute this to:

(i) cross-departmental generalization through FedAvg aggre- gation across diverse email distributions, and (ii) implicit regularization via L2 clipping and Gaussian noise injection acting as stochastic regularization analogous to dropout.

VII. DISCUSSION

The +5.12 AUC improvement over centralized base- lines is attributable to three factors: first, cross-departmental knowledge transfer through FedAvg aggregation enables the global model to learn threat patterns spanning organizational boundaries—each department contributes gradient information from its local threat landscape, and weighted averaging com- bines these diverse perspectives into a more generalizable

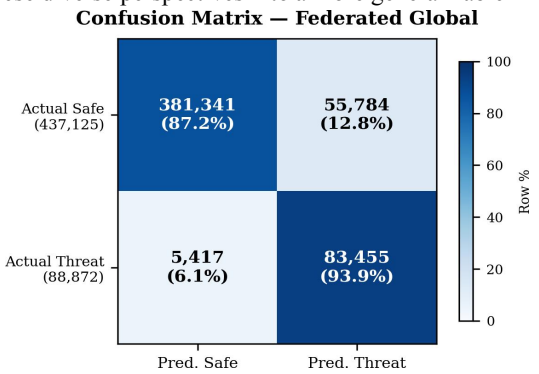

Fig. 3. Confusion matrix heatmap for the federated global classifier (525,997 test samples).

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, May 2026 Impact Factor: 8.2

decision boundary. Second, balanced warm-start initialization overcomes the cold-start problem inherent to federated training with class-imbalanced data. Third, L2 clipping and Gaus- sian noise provide beneficial implicit regularization, limiting any single department's update while introducing stochastic perturbation analogous to dropout [5]. At the chosen noise level (σ=0.05), the favorable privacy-utility tradeoff provides empirical evidence that model utility can be preserved—and even enhanced—through variance reduction, though we note this constitutes practical obfuscation rather than formal (ϵ, δ)- differential privacy.

The anomaly branch achieving 79.22% recall without any exposure to threat labels confirms that malicious emails genuinely differ from legitimate ones in multimodal feature space. While its low standalone precision (16.73%) reflects the large spectrum of unusual-but-benign email behavior, the cascaded fusion protocol ensures defense-in-depth: Stage 1 handles known threats via the supervised gate, Stage 2 allows anomaly override only with corroborating evidence, and Stage 3 pro- vides heuristic backstop coverage. This staged design avoids the false positive explosion that a naive OR rule would produce while preserving high recall through controlled escalation.

The federated global classifier produces 55,784 false pos- itives out of 437,125 safe samples (12.76% FPR). However, several factors mitigate this operational burden: (a) the fused confidence score $r\hat{}$ enables risk-stratified triage, allowing ana- lysts to prioritize high-confidence alerts; (b) the Fβ threshold optimization framework provides a principled mechanism to shift the operating point along the precision-recall curve based on operational requirements; and (c) in email security, the expected cost of a missed threat (data breach, credential compromise) typically exceeds false alarm costs by orders of magnitude, justifying the recall-oriented operating point. Key limitations include reliance on weak supervision (which may introduce label noise), evaluation on a synthetic benchmark rather than production data, and the lack of formal differential privacy accounting—all targeted for future work.

VIII. CONCLUSION

This paper presented EmailGuard-FLIF, a privacy-aware email threat detection framework that unifies three complementary detection paradigms—federated supervised learning, Isolation Forest anomaly detection, and cascaded three-stage decision fusion—into a single deployable pipeline. The archi- tecture addresses three interconnected challenges in enterprise email security: multimodal threat representation across text, behavioral, and identity signal families; privacy-preserving cross-departmental collaborative learning without raw data exchange; and zero-day threat sensitivity through unsupervised anomaly detection with controlled alert escalation.

Experimental evaluation on a large-scale enterprise email corpus (2.63 million samples, 525,997 test instances, 23 departments) demonstrates that the federated global model achieves 93.35% AUC, 93.90% threat recall, and 73.17% F1-score—outperforming the best centralized baseline by

+5.12 AUC and +14.03 F1 points despite privacy-preserving update obfuscation. The Isolation Forest anomaly branch in- dependently achieves 79.22% threat recall without exposure to any threat labels, confirming its value as a complementary zero-day detection layer.

Future work will pursue: (1) formal Re´nyi differential pri- vacy accounting with tight composition bounds across rounds and clients; (2) learned fusion weights via meta-learning and Dempster-Shafer belief combination; (3) temporal and sequen- tial modeling for slow-burn insider threats; (4) department- specific adaptive thresholding; and (5) validation on real (anonymized) production enterprise email data.

REFERENCES

[1] I. Fette, N. Sadeh, and A. Tomasic, "Learning to detect phishing emails," in Proc. 16th Int. Conf. World Wide Web, 2007.

[2] A. Bergholz et al., "New filtering approaches for phishing email," J. Comput. Security, vol. 18, no. 1, pp. 7–35, 2010.

[3] B. McMahan et al., "Communication-efficient learning of deep networks from decentralized data," in AISTATS, 2017.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, May 2026
Impact Factor: 8.2

[4] P. Kairouz and H. B. McMahan, "Advances and open problems in federated learning," Found. Trends Mach. Learn., 2021.
[5] K. Wei et al., "Federated learning with differential privacy: Algorithms and performance analysis," IEEE Trans. Inf. Forensics Security, vol. 15,
pp. 3454–3469, 2020.
[6] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation Forest," in IEEE ICDM, 2008.
[7] S. Yuan and X. Wu, "Deep learning for insider threat detection: Review, challenges and opportunities," Comput. Security, vol. 104, 2021.
[8] J. Yi and Y. Tian, "Insider threat detection model enhancement using hybrid algorithms," Electronics, vol. 13, no. 5, 2024.
[9] Z. Q. Wang, H. Wang, and A. El Saddik, "FedITD: Federated parameter- efficient tuning for insider threat detection," IEEE Access, 2024.
[10] R. G. Gayathri et al., "FedAT: Federated adversarial training for insider threat detection," in IEEE ICPADS, 2025.
[11] C. Song et al., "Audit-LLM: Multi-agent collaboration for log-based insider threat detection," arXiv:2408.08902, 2024.
[12] X. Tao et al., "Insider threat detection based on improved test-time training," High-Confidence Computing, vol. 5, no. 1, 2025.
[13] I. Homoliak et al., "Insight into insiders and IT: A survey of insider threat taxonomies," arXiv:1805.01612, 2018.
[14] CERT Insider Threat Test Dataset, Carnegie Mellon University, https:// kilthub.cmu.edu/articles/dataset/Insider Threat Test Dataset/12841247.
[15] A. Vaswani et al., "Attention is all you need," in NeurIPS, 2017.