Vol.4, No.7, Page 2292- 2305, 2026 https://doi.org/10.59628/jast.v4i7.3475 ISSN 2958-9568 JAST

SUN PUBLISHER

مجلة جامعة صنعاء للعلوم التطبيقية والتكنولوجيا

Sana'a University Journal of Applied Sciences and Technology

https://journals.su.edu.ye/index.php/jast/

SEAF-FL-MLP: Audited Federated Edge-AI Anomaly Detection with Client Heterogeneity and Shortcut-Sensitivity Analysis

Hussein Azzan [1]* **and Abdulsalam Alkholidi** [1,2]

[1]**Department of Communication and Networks Engineering, Faculty of Engineering, Sana'a University, Sana'a, Yemen,**
[2]**Canadian Institute of Technology Tirana - Albania**

*Corresponding author: s.h.a.t.87019@gmail.com

Abstract

Smart agriculture and industrial Internet-of-Things (IoT) systems require privacy-preserving and communication-efficient anomaly detection under heterogeneous non-IID client conditions. This paper presents SEAF-FL-MLP, a leakage-safe federated edge-AI framework for binary anomaly detection across five IoT clients: AGRI, SWAT, WADI, WUSTL-IIoT, and TON-IoT. The framework uses split-first preprocessing, train-only imputation/encoding/scaling/feature selection, client-specific processing, a compact multilayer perceptron, FedAvg aggregation, validation-only threshold selection, and test-only final evaluation. In the audited V9 results, the balanced validation-selected threshold achieved Accuracy = 96.32%, Precision = 96.67%, Recall = 93.74%, F1-score = 95.18%, ROC-AUC = 98.86%, and PR-AUC = 98.03% on the global held-out test set, with a 0.0372 MB model and 4.4682 MB total communication across 12 federated rounds. Client-wise analysis showed substantial heterogeneity: WUSTL-IIoT, TON-IoT, and SWAT performed strongly, whereas AGRI and WADI remained stress cases. Shortcut-reduced WUSTL analysis confirmed strong client-level robustness after removing high-risk shortcut-sensitive features but revealed partial global-pipeline sensitivity. FedProx and mutual-information feature selection did not materially improve performance in this configuration. SEAF-FL-MLP is positioned as a compact, reproducible federated edge-AI baseline that emphasizes transparent client-level fairness, data-source auditing, and shortcut-robustness analysis.

Article Info

Keywords:
Federated learning, edge AI, smart agriculture, IoT security, anomaly detection, non-IID data, WUSTL-IIoT, TON-IoT, fairness, shortcut sensitivity.

Article History:
Received: 4-July-2026,
Revised: 5-July-2026,
Accepted: 6-July-2026,
Published: 28 July 2026.

1. INTRODUCTION

Smart farming is evolving from isolated monitoring systems toward interconnected, data-driven cyber-physical environments in which farms, irrigation infrastructures, water-treatment facilities, industrial control systems, and edge gateways continuously generate large-scale sensor, process, and network telemetry. This transformation enables more precise irrigation control, soil monitoring, process supervision, and resource management, but it also expands the cyber-attack surface. Recent smart-agriculture surveys emphasize that IoT-enabled farms depend on heterogeneous sensors, wireless communication, and automated decision loops, while communication-layer security studies show that wireless links and field devices remain exposed to spoofing, tampering, eavesdropping, and denial-of-service conditions [1, 2].

Traditional cloud-based anomaly detection usually requires transmitting raw telemetry to a centralized server for model training or inference. Although this paradigm simplifies model development, it introduces privacy and security concerns because operational data may reveal production schedules, crop conditions, infrastructure status, or device behavior. It also incurs communication overhead and may not scale effectively in resource-constrained edge environments. Federated learning (FL) addresses this limitation by allowing clients to train locally and exchange model updates rather than raw data; this privacy-preserving motivation and the broader open problems of operational FL are well documented in recent FL and FL-IoT surveys [3, 4].

Deploying FL in realistic IoT environments remains difficult because clients differ in dataset size, feature distributions, attack prevalence, label quality, noise level, and device capacity. These heterogeneous and non-independent and identically distributed (non-IID) conditions may cause client drift, unstable convergence, and misleading global averages. FedProx was proposed specifically for heterogeneous federated optimization, while IoT-focused FL and FL-security surveys emphasize constraints related to computation, bandwidth, privacy leakage, poisoning, and client heterogeneity [5–7]. A high global score can therefore mask weak client behavior, particularly when some clients are cleanly separable network-flow datasets while others are noisy smart-agriculture or reduced stress-case datasets. Consequently, a credible FL anomaly-detection study should report not only global accuracy but also client-wise metrics, fairness diagnostics, validation curves, shortcut-sensitivity tests, communication cost, and reproducibility evidence. Shortcut-sensitivity analysis is especially important for IoT-security designers because models can otherwise learn dataset identifiers or flow artifacts that create inflated accuracy without reliable deployment robustness.

1.1. Research Gap

Recent FL-based intrusion-detection studies increasingly address privacy-preserving anomaly detection in agricultural IoT and heterogeneous IoT networks. FELIDS demonstrates the relevance of FL for agricultural-IoT intrusion detection, while heterogeneous FL-IDS work based on TON-IoT highlights the difficulty of learning under realistic client-level non-IID conditions. The TON-IoT telemetry dataset further provides an important public IIoT/IoT security benchmark for evaluating data-driven intrusion-detection systems [8–10]. However, three gaps remain important for smart-agriculture and heterogeneous edge-IoT anomaly detection. First, many studies emphasize aggregate detection accuracy while providing limited client-wise analysis of weak clients, false-positive burden, and fairness gaps. Second, edge-suitability is often discussed without jointly reporting model size, communication cost, validation curves, and latency diagnostics. Third, shortcut-sensitive tabular features, such as identifiers, timestamps, and flow-derived fields, are not always tested explicitly, although they can inflate anomaly-detection performance and encourage false confidence in security models. The present work addresses these gaps through an audited federated edge-AI evaluation that combines leakage-safe preprocessing, compact MLP modeling, client-level reporting, fairness analysis, WUSTL shortcut-reduced sensitivity, ablation evidence, and source-traceable tables and figures. The shortcut-reduced analysis is treated as a primary contribution because it helps distinguish robust anomaly-discrimination behavior from possible shortcut-driven performance.

1.2. Research Questions

This study is guided by the following research questions:

i. RQ1: Can a compact MLP-based federated model achieve strong binary anomaly-detection performance across heterogeneous smart-agriculture and IIoT clients without sharing raw data?
ii. RQ2: How does the balanced validation-selected operating threshold affect precision, recall, F1-score, false-positive rate, and specificity compared with a fixed 0.5 threshold?
iii. RQ3: To what extent do global metrics conceal client-level heterogeneity, especially for AGRI and WADI stress cases?
iv. RQ4: Does replacing the original WUSTL-IIoT client with a shortcut-reduced version materially affect client-level and global performance?
v. RQ5: Do FedProx, mutual-information feature selection, feature count, and aggregation weighting provide consistent improvements under the audited V9 configuration?

To answer these questions, this paper updates SEAF-FL-MLP as a compact federated edge-AI framework for heterogeneous IoT anomaly detection. The design uses leakage-safe split-first preprocessing, train-only feature processing, validation-only threshold selection, test-only final evaluation, and a compact MLP suitable for communication-constrained environments. The paper intentionally avoids unsupported claims: FedProx is treated as a sensitivity analysis, mutual-information feature selection is treated as a dimensionality-control mechanism, and weak clients are reported transparently.

The main contributions of this work are as follows:

- A privacy-preserving federated anomaly-detection framework based on a compact MLP is evaluated across five heterogeneous clients: AGRI, SWAT, WADI, WUSTL-IIoT, and TON-IoT.
- A leakage-safe experimental protocol is used, with split-first preprocessing, train-only imputation/encoding/scaling/feature selection, validation-only threshold selection, and test-only final evaluation.
- The audited V9 results are expanded into global, client-wise, threshold, fairness, ablation, latency, communication, and error-analysis tables, each traced to named source files.
- WUSTL-IIoT shortcut-reduced sensitivity is emphasized as a core validity contribution to assess whether near-perfect client performance depends on high-risk shortcut-sensitive features and to warn IoT-security designers against shortcut-driven accuracy.
- Client-level fairness and heterogeneity are quantified, showing that global performance must be interpreted alongside AGRI and WADI stress cases.
- Communication cost, model size, latency, and validation curves are reported to support edge-AI suitability while clearly identifying remaining validation limits.

2. RELATED WORK

The related-work discussion is deliberately claim-bounded: it positions the present work as an audited federated edge-AI evaluation rather than as a universal intrusion-detection

Table 1. Recent 2020–2026 literature positioning relative to SEAF-FL-MLP.

Theme	Representative references	Position of this paper
Smart agriculture and communication security	Smart-agriculture IoT and communication-layer security surveys [1, 2]; agricultural-IoT FL-IDS [8]	Motivates the AGRI stress client and the need for privacy-preserving, communication-efficient anomaly detection in field-oriented IoT environments.
FL for IoT and heterogeneous edge clients	General FL open problems [3], FL-IoT survey evidence [4], FedProx heterogeneity optimization [5], and resource-constrained IoT FL [6]	SEAF-FL-MLP reports global metrics together with client-level diagnostics, communication cost, latency, and heterogeneity analysis.
Security, privacy, and intrusion detection	FL security/privacy survey [7], agricultural IoT FL-IDS [8], heterogeneous IoT FL-IDS [9], TON-IoT benchmark evidence [10], and Network TON-IoT edge architecture [11]	The manuscript focuses on binary anomaly detection with explicit false-positive, false-negative, threshold, and dataset-source auditing.
Personalization, fairness, and client heterogeneity	Ditto [12], model-agnostic personalized FL [13], and adaptive local aggregation [14]	The present paper does not claim full personalization; it uses these works to motivate transparent client-wise reporting and fairness-gap analysis.
Dataset evidence sources	WUSTL-IIoT-2021 [15], official SWaT documentation [16], and official WaDi documentation [17]	The public-client sources are explicitly cited, while AGRI is treated as an audited internal stress client rather than overclaimed as a public benchmark.
Tabular learning and compact edge inference	TabNet [18] and tabular deep-learning benchmarking [19]	The model choice is framed as a compact tabular MLP baseline rather than a universal tabular-learning optimum.

solution.

2.1. Federated Learning for IoT Infrastructure

FL allows clients to train local models while keeping raw data decentralized. Recent FL surveys identify privacy, statistical heterogeneity, communication efficiency, device constraints, security, and fairness as central deployment challenges [3, 4]. FedProx introduced a proximal term to mitigate client drift under heterogeneous data [5], while IoT-focused surveys emphasize that resource-constrained clients require explicit reporting of memory, communication, and latency constraints [6]. These studies motivate the present paper's compact MLP design, validation-only thresholding, source-audited reporting, and FedAvg/FedProx sensitivity analysis.

2.2. Smart Agriculture and FL-Based Intrusion Detection

Smart-agriculture systems combine IoT sensing, wireless communication, irrigation monitoring, and cyber-physical automation. Recent surveys show that these systems can improve precision agriculture but also expose wireless and device-level attack surfaces [1, 2]. FELIDS demonstrates the relevance of FL for agricultural-IoT intrusion detection [8]. More broadly, FL-based IDS studies in heterogeneous IoT networks show that realistic non-IID client construction can significantly affect final detection performance [9]. SEAF-FL-MLP therefore reports AGRI and WADI as stress cases rather than hiding them behind the global average.

2.3. Datasets and Evidence Sources

The TON-IoT telemetry dataset is a recent IoT/IIoT benchmark for data-driven intrusion detection [10], and the Network TON-IoT architecture provides edge/fog/cloud context for evaluating AI-based security systems [11]. The WUSTL-IIoT-2021 dataset is used as a public IIoT cybersecurity source [15]. SWAT and WADI are tied to the official iTrust/SUTD secure water treatment and water distribution dataset pages [16, 17]. AGRI is treated as an audited internal smart-agriculture client in the provided V9 result package and is not overclaimed as a public benchmark unless a formal public dataset citation is later supplied.

2.4. Client Heterogeneity, Fairness, and Personalization

Personalized FL literature highlights that a single global model may not be equally suitable for all clients. Ditto explicitly addresses fairness and robustness in statistically heterogeneous federated settings [12]; model-agnostic personalized FL examines adaptation under client distribution differences [13]; and FedALA uses adaptive local aggregation to personalize client initialization [14]. This manuscript uses these works to justify client-wise reporting, fairness-gap measurement, and cautious interpretation of weak AGRI/WADI behavior. It does not claim that personalization is fully solved in the current V9 run.

2.5. Tabular Models and Shortcut Sensitivity

Most industrial IoT and network-security datasets are tabular. TabNet introduced attentive interpretable tabular learning

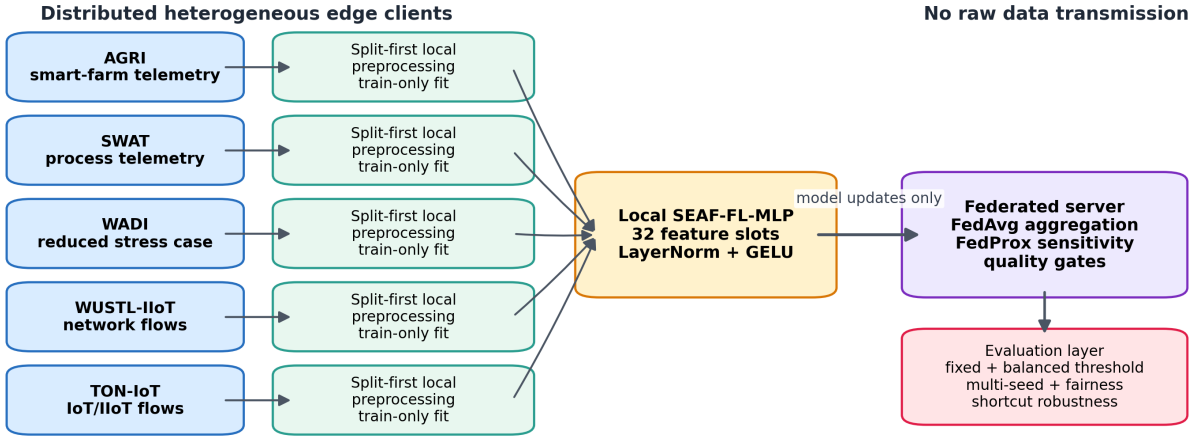

Figure 1. Overall SEAF-FL-MLP design. AGRI, SWAT, WADI, WUSTL-IIoT, and TON-IoT remain separate federated clients. The figure emphasizes the privacy-preserving data flow: local clients send model updates only, while the server performs aggregation, evaluation, quality-gate checking, fairness analysis, and WUSTL shortcut sensitivity analysis.

[18], and broader tabular deep-learning benchmarking shows that simple baselines and gradient-boosted models can remain highly competitive on tabular tasks [19]. This supports the paper's cautious framing: Local Random Forest can achieve extremely high standalone performance on the audited data, so the main contribution should not be stated as universal classifier superiority. Instead, the contribution is a compact, leakage-safe, source-audited federated edge-AI baseline with transparent client-level diagnostics and shortcut-reduced sensitivity analysis.

3. METHODOLOGY

3.1. Federated Edge-AI System

Figure 1 shows the proposed SEAF-FL-MLP workflow. Each client performs local preprocessing and model training. Only model updates are shared with the federated server; raw client data remain local. The server aggregates updates, evaluates fixed and balanced thresholds, and generates quality-gate and reproducibility reports. The use of vector PDF figures addresses the supervisor's request for clear high-resolution figures.

3.2. Data Sources and Leakage-Safe Preprocessing

The evaluation uses five heterogeneous clients: AGRI, SWAT, WADI, WUSTL-IIoT, and TON-IoT. TON-IoT, WUSTL-IIoT, SWAT, and WADI are tied to their published or institutional dataset sources [10, 11, 15–17]. AGRI is treated as a smart-agriculture client in the audited V9 result package and is not overclaimed as a public benchmark unless a formal public dataset citation is supplied. For data transparency, AGRI represents internal smart-farm telemetry prepared for stress-case evaluation from operational field-oriented sensor/gateway records; only anonymized feature-level artifacts and aggregate statistics are used in this manuscript. If institutional and privacy approvals permit, the AGRI preprocessing description, feature schema, and a release-safe anonymized version will be provided with the code package; otherwise, the private-data status and exact reproducibility limits will be documented in the public repository. All numerical tables and figures in this manuscript are traced to named V9 CSV files, and the source mapping is documented in `SOURCE_AUDIT_TABLES_FIGURES.csv`.

The pipeline uses split-first preprocessing. Each dataset is split into training, validation, and test partitions before imputation, encoding, scaling, feature selection, or threshold tuning. Imputers, encoders, scalers, and mutual-information feature selectors are fit on training data only and then applied to validation and test partitions. This protocol prevents validation or test information from influencing feature construction and supports leakage-safe evaluation.

3.3. SEAF-FL-MLP Architecture

Figure 2 presents the default edge-suitable architecture. The model consumes 32 harmonized feature slots and uses dense layers with LayerNorm and GELU activations. The architecture is intentionally compact to reduce communication and inference cost while remaining compatible with tabular IoT telemetry.

3.4. Federated Optimization and Training Objective

FedAvg is the default aggregation method. Following the FedAvg formulation [3], the global model at round $t+1$ is obtained by weighted aggregation of local client models:

$$\mathbf{w}^{t+1} = \sum_{k=1}^{K} \frac{n_k}{n} \mathbf{w}_k^{t+1}, \quad n = \sum_{k=1}^{K} n_k, \tag{1}$$

where K is the number of participating clients, n_k is the number of training samples at client k, n is the total number of

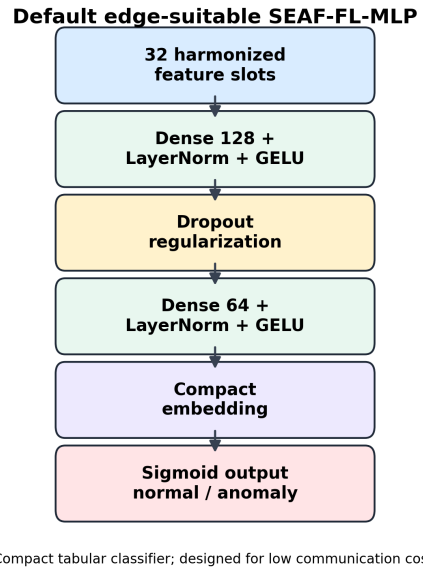

Figure 2. SEAF-FL-MLP model used as the default edge-suitable architecture. The design is a compact tabular classifier rather than a heavy sequence model, making it suitable for low-communication federated evaluation.

participating samples, and $\mathbf{w}_k^{t+1}$ is the locally updated model. FedProx is evaluated only as a sensitivity analysis because the V9 results did not show material improvement over FedAvg. Its local objective follows [5]:

$$\min_{\mathbf{w}} \; F_k(\mathbf{w}) + \frac{\mu}{2} \left\| \mathbf{w} - \mathbf{w}^t \right\|_2^2, \tag{2}$$

where $F_k(\mathbf{w})$ is the empirical client loss and μ controls the strength of the proximal term. For binary anomaly detection, the local classifier is trained using the standard binary cross-entropy objective:

$$\mathcal{L}_{\mathrm{BCE}} = -\frac{1}{m} \sum_{i=1}^{m} \left[y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \right], \tag{3}$$

where $y_i \in \{0, 1\}$ is the binary class label, $\hat{y}_i$ is the predicted anomaly probability, and m is the local mini-batch size.

3.5. Thresholding and Evaluation Metrics

The balanced operating threshold is selected on validation data only and then applied once to the held-out test set. This separation is important because threshold tuning on test data would overstate deployment performance. Operational metrics are computed from the confusion-matrix counts TP, FP, TN, and FN as follows:

$$\begin{aligned} \text{Precision} &= \frac{TP}{TP + FP}, \\ \text{Recall} &= \frac{TP}{TP + FN}, \\ F1 &= \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \end{aligned} \tag{4}$$

The evaluation reports Accuracy, Precision, Recall, F1-score, ROC-AUC, PR-AUC, false-positive rate (FPR), specificity, communication cost, latency, statistical tests, fairness gap, shortcut sensitivity, and reproducibility status. Equations Eq. (1)–Eq. (4) are included to make the optimization and evaluation definitions explicit and referenced.

4. RESULTS

This section reports only values traced to the uploaded V9 result package. To improve scientific transparency, Table and figure source mapping is provided in `SOURCE_AUDIT_TABLES_FIGURES.csv`. Each table is derived from a named CSV file in the V9 output folder, and every plotted curve or boxplot is generated from the same result files. This design avoids manually invented values and makes the manuscript reproducible from the result package. Every table and figure is introduced in the text, referred to by label, and interpreted separately, directly addressing the supervisor's request for explicit table/figure discussion.

4.1. Reproducibility and Source Audit

All expanded tables and figures in this version were regenerated from the uploaded V9 result package. The audit file `SOURCE_AUDIT_TABLES_FIGURES.csv` maps each table and figure to its source CSV. Key source files include `global_metrics.csv`, `client_metrics.csv`, `multi_seed_summary.csv`, `all_seed_metrics_raw.csv`, `validation_history.csv`, `wustl_original_vs_shortcut_reduced.csv`, `fedavg_fedprox_comparison.csv`, `ablation_results.csv`, `baseline_comparison.csv`, `error_analysis.csv`, `communication_cost.csv`, and `latency_results.csv`. The former appendix content has been moved into the Results section so that the manuscript no longer depends on an appendix for essential reproducibility information.

4.2. Global Test Performance

Table 2 summarizes the final global test performance. The balanced validation-selected threshold achieved Accuracy = 0.9632, Precision = 0.9667, Recall = 0.9374, F1 = 0.9518, ROC-AUC = 0.9886, PR-AUC = 0.9803, FPR = 0.0204, and Specificity = 0.9796. Compared with the fixed 0.5 threshold, the balanced threshold improved precision, F1-score, specificity, and false-positive control while slightly reducing recall. This is a defensible security operating point: the model gives up a small amount of recall to reduce false alarms and improve precision.

Table 3 confirms that the balanced threshold was selected using validation data only. This point is methodologically important because test-set threshold selection would lead to

Table 2. Audited global test performance copied from V9 global_metrics.csv.

Threshold	Tau	Acc.	Prec.	Recall	F1	ROC-AUC	PR-AUC	FPR	Spec.
Fixed 0.5	0.5000	0.9555	0.9367	0.9492	0.9429	0.9886	0.9803	0.0406	0.9594
Balanced validation	0.6900	0.9632	0.9667	0.9374	0.9518	0.9886	0.9803	0.0204	0.9796

Table 3. Audited balanced-threshold operating point from threshold_comparison.csv.

Scope	Tau	Acc.	Prec.	Recall	F1	ROC-AUC	PR-AUC	FPR	Spec.	Selection
GLOBAL	0.6900	0.9618	0.9632	0.9372	0.9500	0.9881	0.9799	0.0227	0.9773	validation_only

Table 4. Global multi-seed stability from V9 multi_seed_summary.csv.

Threshold	Acc. mean	Acc. std	Prec. mean	Prec. std	Recall mean	Recall std	F1 mean	F1 std	ROC mean	ROC std	PR mean	PR std
Balanced	0.9620	0.0038	0.9637	0.0047	0.9373	0.0061	0.9503	0.0051	0.9877	0.0017	0.9797	0.0034
Fixed 0.5	0.9548	0.0044	0.9305	0.0104	0.9549	0.0059	0.9425	0.0053	0.9877	0.0017	0.9797	0.0034

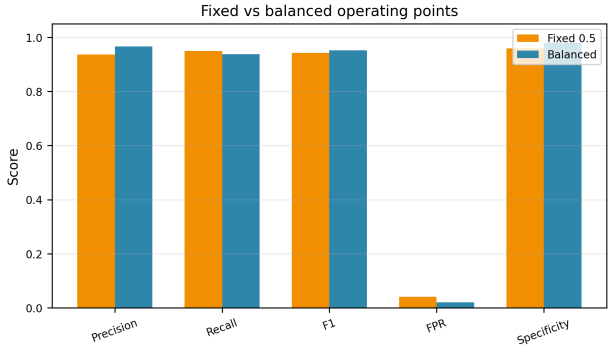

Figure 3. Threshold operating-point comparison generated from `global_metrics.csv`. Balanced validation thresholding reduces FPR and increases precision and specificity, while recall decreases slightly.

optimistic bias. The reported operating point therefore supports leakage-safe final evaluation. Figure 3 visualizes the change from the fixed to the balanced threshold. The most important change is not ROC-AUC or PR-AUC, which are threshold-independent ranking metrics and remain unchanged, but the operational trade-off among precision, recall, FPR, and specificity.

4.3. Multi-Seed Stability

The final evaluation includes multi-seed outputs rather than relying on a single favorable run. Table 4 reports global mean and standard deviation values across the available seeds. Balanced thresholding achieved mean global F1 = 0.9503 with a standard deviation of 0.0051, indicating stable overall performance. The low standard deviation of ROC-AUC and PR-AUC further suggests that the ranking behavior of the model is stable across random seeds. Figure 4 shows the same conclusion visually through boxplots. The balanced threshold is consistently concentrated at high F1 and precision values, while ROC-AUC and PR-AUC remain consistently high for both thresholds.

4.4. Federated Learning Validation Curves

A dedicated validation-curve analysis was added because it exposes the behavior of the federated model over communication rounds. Figure 5 shows that validation F1, ROC-AUC, PR-AUC, and recall generally improve across rounds, with validation F1 increasing from approximately 0.775 in round 1 to approximately 0.942 by round 12. This confirms that the federated process learns progressively rather than producing a high score only at the final test stage.

Figure 6 reports the training-loss trace. The curve is not required to be strictly monotonic because each federated round combines heterogeneous client updates; however, it provides useful diagnostic context for understanding model behavior under non-IID client conditions.

4.5. Client-Wise Results and Stress Clients

Table 5 reports held-out client-wise metrics under the balanced validation threshold. WUSTL-IIoT, TON-IoT, and SWAT drive the strong global result, while AGRI and WADI remain challenging stress-case clients. WUSTL achieved F1 = 0.9940, TON achieved F1 = 0.9538, and SWAT achieved F1 = 0.8379. In contrast, AGRI achieved F1 = 0.1997 and WADI achieved F1 = 0.1818. These values should be reported transparently rather than hidden inside the global average. Figure 7 uses seed-level client F1 values. The boxplots show that the heterogeneity pattern is not caused by one isolated run: WUSTL, TON, and SWAT remain consistently stronger than AGRI and WADI across seeds. AGRI has moderate recall but weak precision, indicating many false positives. WADI has very small support and should be interpreted as a reduced stress case rather than a benchmark-level claim.

4.6. Fairness and Client Heterogeneity

Table 6 quantifies client heterogeneity. The mean client F1 was 0.6334, the standard deviation was 0.3651, and

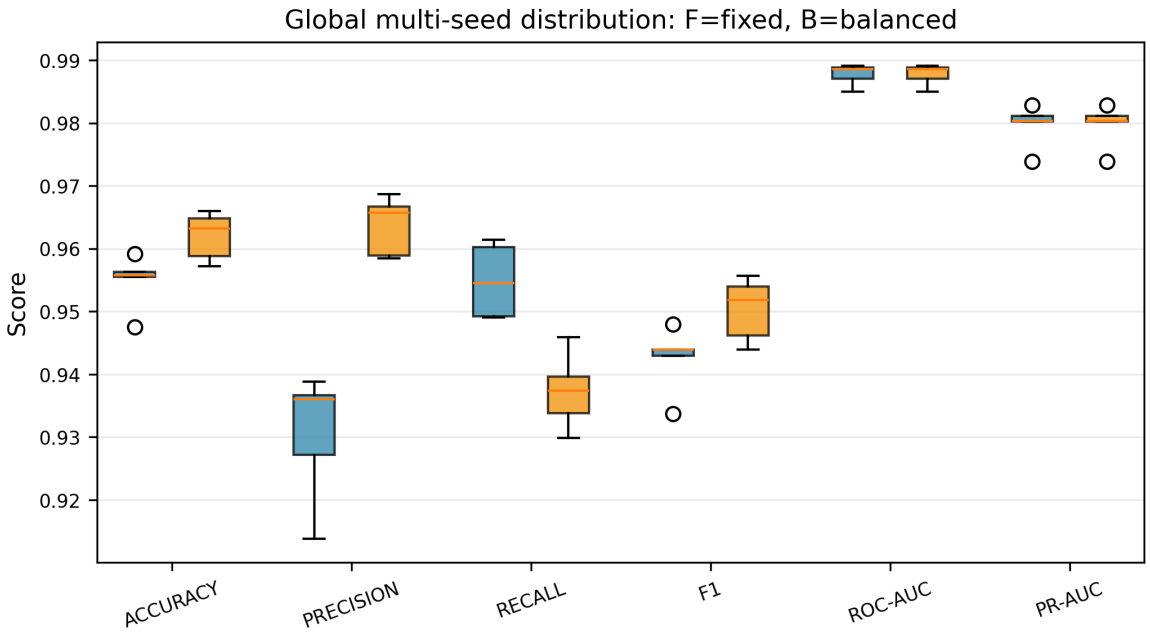

Figure 4. Global multi-seed metric distribution generated from `all_seed_metrics_raw.csv`. The figure verifies that the balanced validation-selected threshold is stable across seeds and not an isolated single-run artifact.

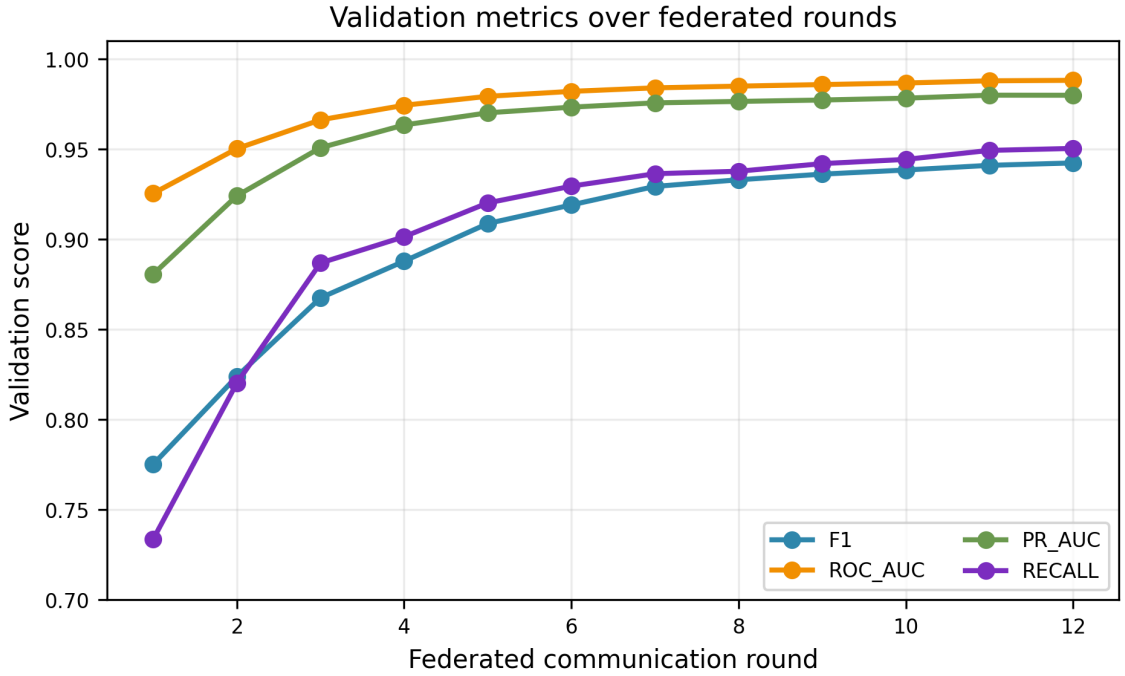

Figure 5. Federated learning validation curve over communication rounds, generated from `validation_history.csv`. F1, ROC-AUC, PR-AUC, and recall improve over rounds, supporting the stability of the federated training process.

Table 5. Audited client-wise held-out performance under the balanced validation threshold.

Client	N	Normal	Attack	Acc.	Prec.	Recall	F1	ROC-AUC	PR-AUC	FPR	Spec.
AGRI	1236	1125	111	0.5591	0.1193	0.6126	0.1997	0.5993	0.1173	0.4462	0.5538
SWAT	18000	15816	2184	0.9649	0.9544	0.7468	0.8379	0.9557	0.8850	0.0049	0.9951
WADI	75	68	7	0.5200	0.1081	0.5714	0.1818	0.5189	0.1247	0.4853	0.5147
WUSTL	22500	11250	11250	0.9940	0.9970	0.9909	0.9940	0.9999	0.9999	0.0029	0.9971
TON	22200	10950	11250	0.9546	0.9853	0.9243	0.9538	0.9879	0.9872	0.0142	0.9858

Table 6. Client fairness and heterogeneity summary from client_fairness_metrics.csv.

Mean F1	Std F1	Best client	Best F1	Worst client	Worst F1	Gap	Coef. var.	Macro F1	Weighted F1	Level
0.6334	0.3651	WUSTL	0.9940	WADI	0.1818	0.8122	0.5764	0.6334	0.1997	substantial_client_heterogeneity

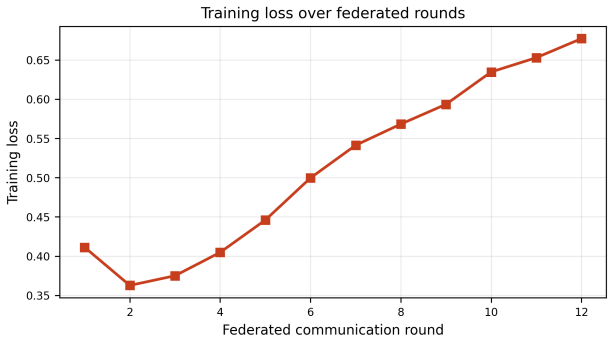

Figure 6. Training-loss trace over communication rounds generated from `validation_history.csv`. Non-monotonic behavior is expected in heterogeneous federated learning and should be interpreted together with validation metrics.

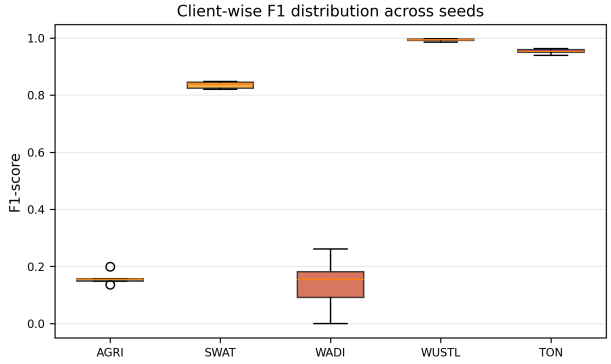

Figure 7. Client-wise F1-score distribution across seeds generated from `all_seed_metrics_raw.csv`. AGRI and WADI appear as challenging stress-case clients, while WUSTL-IIoT, TON-IoT, and SWAT are the main contributors to strong global performance.

the fairness gap between the best and worst client was 0.8122. The interpretation exported by V9 is `substantial_client_heterogeneity`. This is one of the most important scientific messages of the paper: strong global performance is real, but it is not uniformly distributed across clients.

4.7. WUSTL Shortcut-Reduced Sensitivity

WUSTL-IIoT produced near-perfect metrics, so V9 included a shortcut-reduced sensitivity test. Table 7 shows that WUSTL remained strong after removing high-risk identifier, timestamp, and flow-related shortcut-sensitive features. However, the global pipeline experienced a larger drop when the original WUSTL client was replaced with its shortcut-reduced counterpart. Thus, the precise interpretation is that WUSTL is robust at the client level, but the overall pipeline is partially sensitive to this replacement. Figure 8 visualizes the global effect of this

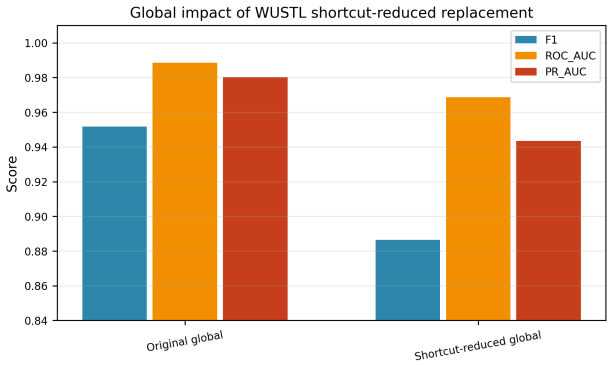

Figure 8. Global effect of replacing original WUSTL-IIoT with its shortcut-reduced version, generated from `wustl_original_vs_shortcut_reduced.csv`. WUSTL remains strong at the client level, but the global pipeline shows partial sensitivity.

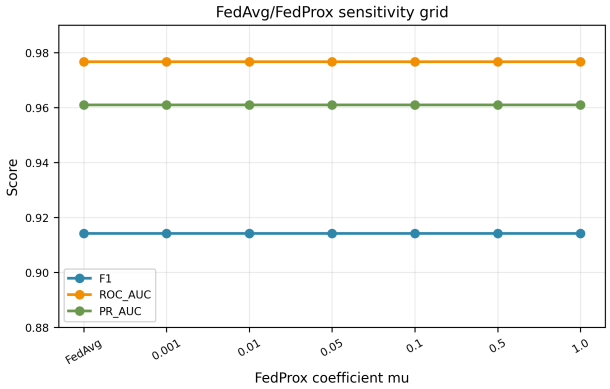

Figure 9. FedAvg/FedProx sensitivity grid generated from `fedavg_fedprox_comparison.csv`. The flat curves indicate that FedProx did not materially improve the evaluated metrics under this configuration.

replacement. The figure should be discussed as a robustness diagnostic, not as a claim of perfect external generalization.

4.8. FedAvg/FedProx Sensitivity

FedAvg is the default aggregation strategy. FedProx was evaluated as a sensitivity mechanism only. Table 8 and Figure 9 show that all tested μ values produced essentially identical global metrics under the V9 configuration. The statistical test also reported Wilcoxon $p = 1.0$ for FedAvg versus FedProx. Therefore, this paper does not claim that FedProx improves performance; it is included as a negative but useful sensitivity result.

Table 7. WUSTL shortcut-reduced sensitivity from wustl_original_vs_shortcut_reduced.csv.

Setting / Scope	Threshold	Acc.	Prec.	Recall	F1	ROC	PR	FPR	Spec.
original / WUSTL	balanced_validation	0.9940	0.9970	0.9909	0.9940	0.9999	0.9999	0.0029	0.9971
original / GLOBAL	balanced_validation	0.9632	0.9667	0.9374	0.9518	0.9886	0.9803	0.0204	0.9796
shortcut_reduced / WUSTL	balanced_validation	0.9922	0.9842	0.9972	0.9906	0.9992	0.9986	0.0113	0.9887
shortcut_reduced / GLOBAL	balanced_validation	0.9161	0.8515	0.9241	0.8864	0.9686	0.9434	0.0883	0.9117

Table 8. FedAvg/FedProx sensitivity grid from fedavg_fedprox_comparison.csv.

Mu	Acc.	Prec.	Recall	F1	ROC	PR	FPR	Train s
FedAvg	0.9329	0.9053	0.9233	0.9142	0.9767	0.9610	0.0611	49.8644
0.001	0.9329	0.9053	0.9233	0.9142	0.9767	0.9610	0.0611	47.3734
0.01	0.9329	0.9053	0.9233	0.9142	0.9767	0.9610	0.0611	51.0772
0.05	0.9329	0.9053	0.9233	0.9142	0.9767	0.9610	0.0611	51.4254
0.1	0.9329	0.9053	0.9233	0.9142	0.9767	0.9610	0.0611	52.8765
0.5	0.9329	0.9053	0.9233	0.9142	0.9767	0.9610	0.0611	48.9749
1	0.9329	0.9053	0.9233	0.9142	0.9767	0.9610	0.0611	47.4001

Table 9. Architecture comparison under balanced validation threshold from architecture_comparison.csv.

Architecture	Acc.	Prec.	Recall	F1	ROC	PR	FPR	Spec.
SEAF-FL-MLP	0.9323	0.9193	0.9046	0.9119	0.9796	0.9707	0.0502	0.9498
Lightweight_SEAF	0.9155	0.9156	0.8614	0.8877	0.9604	0.9446	0.0502	0.9498
Tabular_Ablation	0.9438	0.9604	0.8917	0.9248	0.9816	0.9742	0.0233	0.9767

Table 10. Baseline comparison from baseline_comparison.csv.

Baseline	Scope	Acc.	Prec.	Recall	F1	ROC	PR	FPR	Spec.
Local_Logistic_Regression	local_per_client	0.9567	0.9675	0.9192	0.9427	0.9708	0.9665	0.0196	0.9804
Local_Random_Forest	local_per_client	0.9986	0.9993	0.9971	0.9982	0.9999	0.9999	0.0004	0.9996
Centralized_Logistic_Regression	centralized_upper_bound	0.9504	0.9609	0.9089	0.9342	0.9553	0.9256	0.0234	0.9766

4.9. Architecture, Baseline, and Ablation Analyses

Table 9 compares the evaluated architectures. Tabular Ablation achieved the highest F1 among the architecture variants in this file, so the manuscript should not claim that SEAF-FL-MLP is universally superior to all classifier families. The defensible claim is that SEAF-FL-MLP is compact, federated, communication-efficient, and competitive under the audited configuration while maintaining strong global performance. Table 10 shows that Local Random Forest achieved extremely high standalone F1. This supports a careful interpretation: classical local models can be very strong on some tabular IoT splits, but they do not provide the same federated privacy-preserving training structure. SEAF-FL-MLP should therefore be framed as a federated edge-AI baseline rather than as a universally superior classifier. Table 11 summarizes the main ablations. The no-MI variant achieved higher F1 than the full MI-based model in this run, so MI feature selection should not be described as a performance-improving component. It is better described as a controlled dimensionality-reduction and feature-slot harmonization mechanism. Figure 10 visualizes ablation F1-scores, and Figure 11 compares local and centralized baselines. These figures make the trade-off transparent and prevent overclaiming.

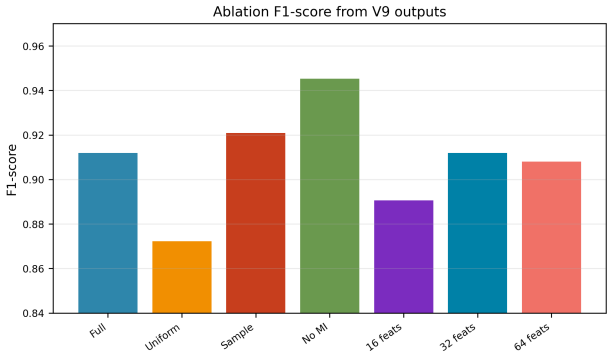

Figure 10. Ablation F1-score comparison generated from `ablation_results.csv`. The no-MI configuration outperformed the full MI configuration in this run; therefore, MI is not claimed as a performance-improving factor.

4.10. Feature Count, Weighting, and MI Ablations

Tables 12, 13, and 14 provide additional controlled comparisons from the V9 outputs. The feature-count results show that 32 feature slots remain a practical default, although 16 and 64 feature variants are close enough that feature count should not be overinterpreted as a universal optimum. The aggregation comparison indicates that sample weighting produced higher F1 than the default full configuration in the ablation file; thus,

Table 11. Ablation summary from ablation_results.csv.

Experiment	Feat.	MI	Sqrt	Acc.	Prec.	Recall	F1	ROC	PR	Comm. MB	Train s
full_seaf_fl_mlp	32	1	1	0.9323	0.9193	0.9046	0.9119	0.9796	0.9707	1.8618	34.5140
aggregation_uniform	32	1	0	0.8945	0.8221	0.9288	0.8722	0.9687	0.9543	1.8618	34.8697
aggregation_sample	32	1	0	0.9396	0.9354	0.9069	0.9209	0.9810	0.9730	1.8618	34.1086
no_MI_first_32	32	0	1	0.9578	0.9507	0.9399	0.9453	0.9890	0.9849	1.8618	35.3328
feature_count_16	16	1	1	0.9137	0.8751	0.9066	0.8906	0.9627	0.9472	1.5688	31.3321
feature_count_32	32	1	1	0.9323	0.9193	0.9046	0.9119	0.9796	0.9707	1.8618	37.7479
feature_count_64	64	1	1	0.9280	0.8986	0.9176	0.9080	0.9762	0.9666	2.4477	38.4720

Table 12. Feature-count ablation from feature_count_ablation_results.csv.

Features	Acc.	Prec.	Recall	F1	ROC	PR	FPR	Spec.	Model MB	Comm. MB
16.0000	0.9137	0.8751	0.9066	0.8906	0.9627	0.9472	0.0818	0.9182	0.0314	1.5688
32.0000	0.9323	0.9193	0.9046	0.9119	0.9796	0.9707	0.0502	0.9498	0.0372	1.8618
64.0000	0.9280	0.8986	0.9176	0.9080	0.9762	0.9666	0.0655	0.9345	0.0490	2.4477

Table 13. Aggregation weighting comparison from aggregation_weighting_comparison.csv.

Weighting	Acc.	Prec.	Recall	F1	ROC	PR	FPR	Spec.
aggregation_sqrt	0.9323	0.9193	0.9046	0.9119	0.9796	0.9707	0.0502	0.9498
aggregation_uniform	0.8945	0.8221	0.9288	0.8722	0.9687	0.9543	0.1271	0.8729
aggregation_sample	0.9396	0.9354	0.9069	0.9209	0.9810	0.9730	0.0396	0.9604

Table 14. Mutual-information feature-selection ablation from mi_feature_selection_ablation.csv.

Setting	Acc.	Prec.	Recall	F1	ROC	PR	FPR	Spec.
full_seaf_fl_mlp_with_mi	0.9323	0.9193	0.9046	0.9119	0.9796	0.9707	0.0502	0.9498
without_mutual_information_feature_selection_first_32	0.9578	0.9507	0.9399	0.9453	0.9890	0.9849	0.0308	0.9692

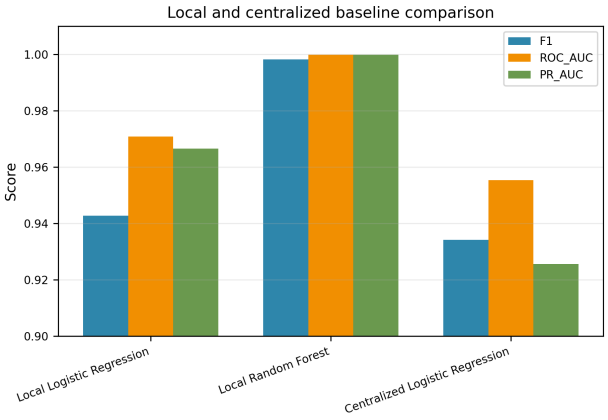

Figure 11. Baseline comparison generated from `baseline_comparison.csv`. Local Random Forest performs strongly as a standalone classifier, while SEAF-FL-MLP is positioned as a compact federated edge-AI architecture.

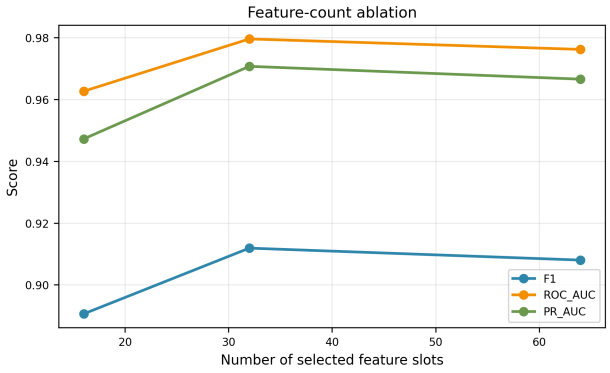

Figure 12. Feature-count ablation generated from `feature_count_ablation_results.csv`. The 32-feature setting is a reasonable compact default, but the curve does not support overclaiming a universally optimal feature count.

square-root weighting is better framed as a domination-control mechanism, not as guaranteed F1 improvement. The MI ablation confirms that MI feature selection did not consistently improve performance in this run.

4.11. Error Analysis

Table 15 and Figure 13 explain why AGRI and WADI are weak clients. AGRI has 502 false positives and only 68 true positives at the balanced threshold, which explains its low precision and F1. WADI has only 75 test samples and 7 attack samples, so its F1 is unstable and should be treated cautiously. TON has more total errors than WUSTL or WADI, but its high support and high true-positive count preserve a strong F1.

Table 15. Client-level error analysis from error_analysis.csv.

Client	FP	FN	TP	TN	Errors	FPR	Recall	Prec.	F1
TON	155.0000	852.0000	10398.0000	10795.0000	1007.0000	0.0142	0.9243	0.9853	0.9538
SWAT	78.0000	553.0000	1631.0000	15738.0000	631.0000	0.0049	0.7468	0.9544	0.8379
AGRI	502.0000	43.0000	68.0000	623.0000	545.0000	0.4462	0.6126	0.1193	0.1997
WUSTL	33.0000	102.0000	11148.0000	11217.0000	135.0000	0.0029	0.9909	0.9970	0.9940
WADI	33.0000	3.0000	4.0000	35.0000	36.0000	0.4853	0.5714	0.1081	0.1818

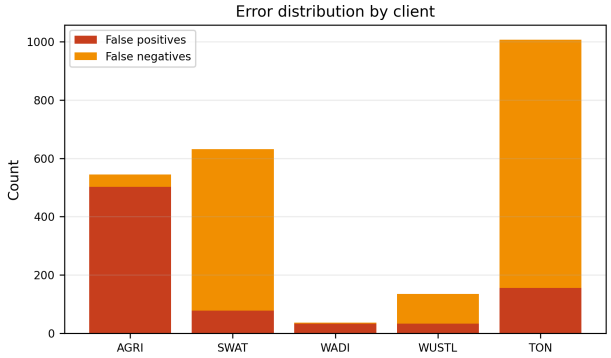

Figure 13. Client-level false-positive and false-negative distribution generated from `error_analysis.csv`. AGRI is dominated by false positives, while WADI is limited by small support.

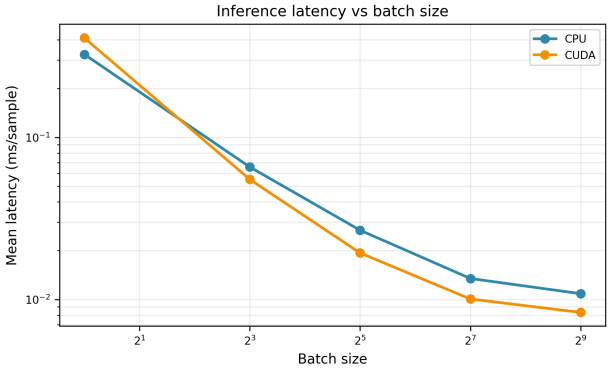

Figure 14. Mean inference latency per sample versus batch size, generated from `latency_results.csv`. Larger batches reduce per-sample latency; real edge-gateway validation remains future work.

Table 16. Communication-cost summary from communication_cost.csv.

Model MB	Clients	Rounds	Upload/round MB	Download/round MB	Total MB	Weighting
0.0372	5	12	0.1862	0.1862	4.4682	sqrt

Table 17. Latency summary aggregated from latency_results.csv.

Device	Batch	Mean ms/sample	Std ms/sample	Mean samples/s
cpu	1	0.3252	0.0121	3078.3335
cpu	8	0.0657	0.0067	15341.6985
cpu	32	0.0267	0.0066	38824.0421
cpu	128	0.0135	0.0024	75743.1023
cpu	512	0.0109	0.0041	99765.3144
cuda	1	0.4108	0.0094	2435.4570
cuda	8	0.0551	0.0028	18195.2707
cuda	32	0.0194	0.0032	52596.3135
cuda	128	0.0101	0.0013	100551.9363
cuda	512	0.0083	0.0025	128977.4640

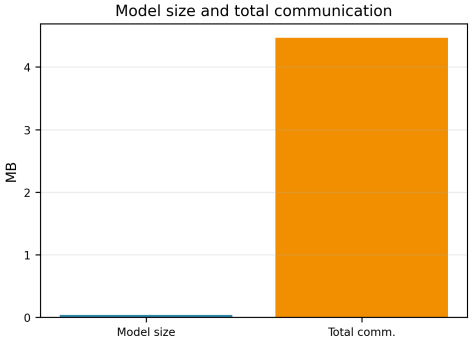

Figure 15. Model size and total communication generated from `communication_cost.csv`. The compact model and low communication footprint support edge suitability.

4.12. Latency, Communication Cost, and Edge Suitability

Table 16 reports a model size of 0.0372 MB and total estimated communication of 4.4682 MB across 12 rounds and five clients. These values support the edge-AI motivation of the framework. Table 17 and Figure 14 report measured Colab CPU/GPU inference latency. The latency improves with batch size; however, these results are Colab measurements and should not be presented as field deployment validation.

4.13. Statistical Significance and Validation Scope

Table 18 reports the available significance tests. FedAvg versus FedProx produced no evidence of improvement. Original versus shortcut-reduced analysis produced no conventional-level statistical significance in the exported test, although the global metric drop is practically important and is discussed as partial sensitivity. The Full-versus-No-MI test had insufficient paired samples, so it is interpreted descriptively rather than inferentially. Tables 19 and 20 are included to avoid overstating unexecuted V9 protocols. Personalization rows are baseline exports only, and cross-dataset generalization rows are scheduled protocols requiring full retraining. These tables are retained in the Results section as validation-scope limits and may later be moved to supplementary material if required by the target

Table 18. Statistical tests exported by V9.

Comparison	Pairs	Wilcoxon p	Paired t p	Cohen d	Interpretation
FedAvg_vs_FedProx	7	1.0000	–	0.0000	statistically_tested
Full_vs_No_MI	1	–	–	–	insufficient_paired_samples_or_scipy_missing
Original_vs_Shortcut_Reduced	8	0.1094	0.0884	0.1181	statistically_tested

Table 19. Personalization framework export. These are baseline global-model client metrics; local fine-tuning was not executed in this run.

Client	Setting	Status	Prec.	Recall	F1	ROC	PR
AGRI	global_model	baseline_exported	0.1193	0.6126	0.1997	0.5993	0.1173
SWAT	global_model	baseline_exported	0.9544	0.7468	0.8379	0.9557	0.8850
WADI	global_model	baseline_exported	0.1081	0.5714	0.1818	0.5189	0.1247
WUSTL	global_model	baseline_exported	0.9970	0.9909	0.9940	0.9999	0.9999
TON	global_model	baseline_exported	0.9853	0.9243	0.9538	0.9879	0.9872

Table 20. Cross-dataset generalization protocol export. Metrics are intentionally blank because full retraining was not executed.

Train clients	Test client	Status	Acc.	Prec.	Recall	F1	ROC	PR
SWAT+TON+AGRI	WUSTL	scheduled_requires_full_retraining	–	–	–	–	–	–
WUSTL+SWAT+TON	AGRI	scheduled_requires_full_retraining	–	–	–	–	–	–
ALL_EXCEPT_AGRI	AGRI	scheduled_requires_full_retraining	–	–	–	–	–	–
ALL_EXCEPT_SWAT	SWAT	scheduled_requires_full_retraining	–	–	–	–	–	–
ALL_EXCEPT_WADI	WADI	scheduled_requires_full_retraining	–	–	–	–	–	–
ALL_EXCEPT_WUSTL	WUSTL	scheduled_requires_full_retraining	–	–	–	–	–	–
ALL_EXCEPT_TON	TON	scheduled_requires_full_retraining	–	–	–	–	–	–

journal.

4.14. Confusion Matrix Diagnostic

Figure 16 reports the global confusion matrix under the balanced threshold. It shows 38,408 true negatives, 801 false positives, 1,553 false negatives, and 23,249 true positives. This diagnostic explains why the balanced setting achieves high precision and low FPR while preserving most attack detections.

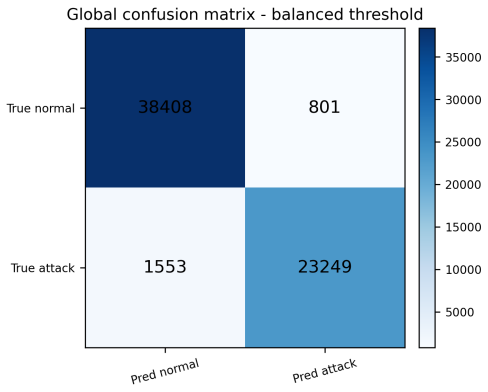

Figure 16. Global confusion matrix under the balanced threshold, reconstructed from `global_metrics.csv`. The low false-positive count relative to true negatives explains the strong FPR and specificity.

5. DISCUSSION

The audited V9 results support a strong but carefully bounded interpretation. The global metrics are high, with F1-score above 0.95 and ROC-AUC/PR-AUC near 0.99 under the balanced validation-selected threshold. This indicates that the compact SEAF-FL-MLP model can learn useful anomaly-discrimination patterns in a heterogeneous federated setting. The validation curve over communication rounds further supports this conclusion because performance improves progressively across federated rounds rather than appearing only at the final evaluation.

The most important scientific finding is client heterogeneity. WUSTL-IIoT, TON-IoT, and SWAT are strong clients, whereas AGRI and WADI remain challenging stress-case clients. AGRI has moderate recall but weak precision, indicating a high false-positive burden and limited anomaly separability. WADI has very limited test support, so its weak F1-score should not be interpreted as a definitive benchmark result. This client-level transparency strengthens the credibility of the paper because it prevents the global score from concealing weak-client behavior.

The WUSTL shortcut-reduced analysis is another important validity check. WUSTL-IIoT remained strong at the client level after shortcut-sensitive feature removal, which reduces the likelihood that its performance is entirely driven by obvious identifiers. However, the global pipeline was partially sensitive when original WUSTL was replaced by shortcut-reduced WUSTL. Therefore, WUSTL should be discussed as a strong client-level result but not as standalone proof of external

generalization.

The weak AGRI and WADI client results also identify a clear technical path for improvement rather than a failure of the overall federated protocol. Future versions should evaluate personalized FL strategies, such as local fine-tuning, Ditto-style personalized objectives, FedALA-style adaptive local aggregation, or client-specific calibration layers, so that weak clients are not forced to rely only on a single global decision boundary. For AGRI, the high false-positive burden suggests local imbalance-aware training, class-weighted or focal loss, threshold calibration per client, and augmentation or resampling of minority anomaly patterns. For WADI, the very small test support indicates the need for additional samples, uncertainty reporting, and imbalance-aware validation before making benchmark-level claims. These remedies should be tested under the same leakage-safe split-first protocol so that improvements do not come from validation or test leakage.

The ablation and baseline results require cautious wording. FedProx did not materially improve performance under the tested configuration and should remain a sensitivity analysis. Mutual-information feature selection did not improve F1-score in the reported ablation, so it should be described as a harmonization and dimensionality-control mechanism rather than as a proven performance booster. Local Random Forest achieved extremely high standalone performance, but it does not provide the same federated privacy-preserving training structure as SEAF-FL-MLP. The main claim is therefore not universal classifier superiority; it is a compact, leakage-safe, communication-efficient federated edge-AI baseline with strong global performance and transparent client-level diagnostics.

6. LIMITATIONS AND THREATS TO VALIDITY

Several limitations should be considered. First, client heterogeneity remains substantial, as shown by the large fairness gap. Second, AGRI remains challenging due to weak precision and high false-positive pressure; its internal telemetry status also requires careful documentation or a release-safe anonymized package for public replication. Third, WADI has very small test support and should be treated as a reduced stress case rather than a full benchmark. Future work should therefore combine personalized FL, client-specific threshold calibration, and local imbalance-aware objectives for AGRI and WADI. Fourth, the WUSTL shortcut-reduced experiment improves confidence but does not replace external validation on new IIoT deployments. Fifth, personalization was exported only as a baseline global-model table in V9; local fine-tuning was not executed. Sixth, cross-dataset generalization protocols were defined, but full retraining was not executed in the uploaded results. Seventh, latency results were measured in Colab CPU/GPU conditions and require validation on actual farm gateways or edge devices such as Raspberry Pi, Jetson, or industrial gateways. Finally, the V9 quality-gate file still reported missing final source-code copies inside the runtime results folder, so reproducibility packaging should be finalized before formal submission. These limitations are stated explicitly to maintain an academically cautious tone and prevent unsupported deployment claims.

7. CONCLUSION

This paper presented SEAF-FL-MLP, a leakage-safe federated edge-AI framework for anomaly detection across heterogeneous IoT and smart-agriculture-related clients. Using the audited V9 result package, the balanced validation-selected threshold achieved strong global test performance: Accuracy = 0.9632, Precision = 0.9667, Recall = 0.9374, F1 = 0.9518, ROC-AUC = 0.9886, and PR-AUC = 0.9803. The model is compact, with a size of approximately 0.0372 MB and total estimated communication of 4.4682 MB across 12 federated rounds and five clients.

The results show that SEAF-FL-MLP is a communication-efficient federated baseline with strong global performance, but they also highlight important validity constraints. Client-level analysis shows substantial heterogeneity, with WUSTL-IIoT, TON-IoT, and SWAT driving the strongest performance, while AGRI and WADI remain challenging cases. WUSTL shortcut-reduced sensitivity supports robustness at the client level but indicates partial sensitivity at the global-pipeline level, making shortcut analysis a central safeguard against overestimating IoT-security model reliability. FedProx and mutual-information feature selection should not be claimed as performance-improving mechanisms in this run. Overall, the framework demonstrates potential for privacy-preserving edge-AI intrusion detection in heterogeneous IoT environments while emphasizing the importance of transparent client-wise reporting, fairness analysis, validation curves, shortcut sensitivity, and reproducibility auditing.

Before journal submission, the remaining edge-deployment claim should be validated on actual edge hardware or industrial gateways; this recommendation is consistent with recent Edge-AI survey guidance on resource-constrained deployment evaluation [20]. Future extensions should also release or document the AGRI client as permitted by privacy constraints and evaluate personalized FL, local imbalance handling, and client-specific calibration for weak AGRI and WADI behavior.

REFERENCES

[1] O. Friha, M. A. Ferrag, L. Shu, L. Maglaras, and X. Wang, "Internet of things for the future of smart agriculture: A comprehensive survey of emerging technologies," *IEEE/CAA J. Autom. Sinica*, vol. 8, no. 4, pp. 718–752, 2021.

[2] H. M. Rouzbahani et al., "Communication layer security in smart farming: A survey on wireless technologies," *arXiv preprint arXiv:2203.06013*, 2022.

[3] P. Kairouz et al., "Advances and open problems in federated learning," *Found. Trends Mach. Learn.*, vol. 14, no. 1–2, pp. 1–210, 2021.

[4] D. C. Nguyen, M. Ding, P. N. Pathirana, A. Seneviratne, J. Li, and H. V. Poor, "Federated learning for internet of things: A comprehensive survey," *IEEE Commun. Surv. & Tutorials*, vol. 23, no. 3, pp. 1622–1658, 2021.

[5] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *Proceedings of Machine Learning and Systems (MLSys)*, 2020.

[6] A. Imteaj, U. Thakker, S. Wang, J. Li, and M. H. Amini, "A survey on federated learning for resource-constrained iot devices," *IEEE Internet Things J.*, vol. 9, no. 1, pp. 1–24, 2022.

[7] V. Mothukuri, R. M. Parizi, S. Pouriyeh, A. Dehghantanha, and G. Srivastava, "A survey on security and privacy of federated learning," *Future Gener. Comput. Syst.*, vol. 115, pp. 619–640, 2021.

[8] O. Friha, M. A. Ferrag, L. Shu, L. Maglaras, K.-K. R. Choo, and M. Nafaa, "Felids: Federated learning-based intrusion detection system for agricultural internet of things," *J. Parallel Distributed Comput.*, vol. 165, pp. 17–31, 2022.

[9] O. Belarbi, T. Spyridopoulos, E. Anthi, I. Mavromatis, P. Carnelli, and A. Khan, "Federated deep learning for intrusion detection in iot networks," *arXiv preprint arXiv:2306.02715*, 2023.

[10] A. Alsaedi, N. Moustafa, Z. Tari, A. Mahmood, and A. Anwar, "Ton_iot telemetry dataset: A new generation dataset of iot and iiot for data-driven intrusion detection systems," *IEEE Access*, vol. 8, pp. 165 130–165 150, 2020.

[11] N. Moustafa, "A new distributed architecture for evaluating ai-based security systems at the edge: Network ton_iot datasets," *Sustain. Cities Soc.*, vol. 72, p. 102 994, 2021.

[12] T. Li, S. Hu, A. Beirami, and V. Smith, "Ditto: Fair and robust federated learning through personalization," in *Proceedings of the 38th International Conference on Machine Learning (ICML)*, ser. PMLR, vol. 139, 2021, pp. 6357–6368.

[13] A. Fallah, A. Mokhtari, and A. Ozdaglar, "Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach," in *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[14] J. Zhang et al., "Fedala: Adaptive local aggregation for personalized federated learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, 2023, pp. 11 237–11 244.

[15] Washington University in St. Louis, *Wustl-iiot-2021 dataset for iiot cybersecurity research*, https://www.cse.wustl.edu/~jain/iiot2/index.html, Accessed: 2026-06-30, 2021.

[16] iTrust, Singapore University of Technology and Design, *Secure water treatment (swat) dataset characteristics*, https://www.sutd.edu.sg/itrust/itrust-labs/datasets/dataset-characteristics/swat/, Accessed: 2026-06-30, 2025.

[17] iTrust, Singapore University of Technology and Design, *Water distribution (wadi) dataset characteristics*, https://www.sutd.edu.sg/itrust/itrust-labs/datasets/dataset-characteristics/wadi/, Accessed: 2026-06-30, 2025.

[18] S. O. Arik and T. Pfister, "Tabnet: Attentive interpretable tabular learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, 2021, pp. 6679–6687.

[19] Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko, "Revisiting deep learning models for tabular data," in *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[20] R. Singh and S. S. Gill, "Edge ai: A survey," *Internet Things Cyber-Physical Syst.*, vol. 3, pp. 71–92, 2023.