International Journal of Science and Research Archive

eISSN: 2582-8185

Cross Ref DOI: 10.30574/ijsra

Journal homepage: https://ijsra.net/

(Research Article)

Federated Learning for Fault Detection in Edge-Based IoT Systems

Adeniji Olamilekan Samuel [1, *], Ogaji Sylvester Oga [2], Adeyanju Emmanuel Abayomi [3] and Alabi Praise Ayomide [4]

[1] Department of Computer Science, Crown Polytechnic, Ado Ekiti, Nigeria.
[2] Department of Computer Science, Benue State University Makurdi, Nigeria.
[3] Department of Computer Science Education, National Open University University, Nigeria.
[4] Department of Mass Communication, Kwara State University, Nigeria.

International Journal of Science and Research Archive, 2026, 18(03), 551-563

Publication history: Received on 30 January 2026; revised on 04 March 2026; accepted on 06 March 2026

Article DOI: https://doi.org/10.30574/ijsra.2026.18.3.0480

Abstract

Edge-based Internet of Things (IoT) deployments increasingly support safety- and cost-critical operations, requiring reliable fault detection under strict bandwidth, latency, and data-governance constraints. Centralized learning pipelines are often impractical because high-rate sensor telemetry is expensive to transmit and may be restricted by privacy or proprietary policies. Federated learning (FL) addresses these limitations by enabling collaborative model training across distributed edge clients while keeping raw data local. However, fault detection in edge-IoT introduces strong statistical and systems heterogeneity, including non-IID operating conditions, label imbalance, intermittent participation, and client drift, which can degrade standard FL approaches. This study investigates FL for fault detection using a unified benchmark suite spanning vibration fault classification (CWRU and Paderborn), acoustic anomaly detection (MIMII), and multivariate degradation/anomaly detection (NASA C-MAPSS). We evaluate FedAvg against heterogeneity-aware optimizers (FedProx and SCAFFOLD) under non-IID client partitions and hierarchical edge-gateway-cloud training. Results show that federated methods consistently outperform local-only training and approach centralized upper bounds while preserving data locality. Robust FL methods provide statistically significant improvements over FedAvg under severe heterogeneity, with the largest gains on datasets exhibiting strong domain and trajectory skew. Communication analysis highlights trade-offs between robustness and overhead and shows that compression can substantially reduce bandwidth with modest impact on convergence. These findings support FL as a practical pathway for scalable, privacy-preserving fault detection in edge-based IoT systems and motivate future work on drift-aware continual federated learning, personalization, and trustworthy aggregation.

Keywords: Federated learning; Fault detection; Industrial IoT; FedAvg; FedProx; SCAFFOLD; Hierarchical federated learning; CWRU; Paderborn; MIMII; NASA C-MAPSS

1. Introduction

Edge-based Internet of Things (IoT) systems are increasingly deployed in industrial automation, smart energy, transportation, and other safety- and cost-critical domains where early fault detection reduces unplanned downtime, prevents cascading failures, and improves maintenance efficiency. In these environments, edge nodes continuously collect high-rate telemetry such as vibration and motor-current signals from rotating machinery, acoustic emissions from industrial equipment, and multivariate sensor streams from complex assets. Despite the availability of powerful learning models, cloud-centric analytics remain difficult to apply at scale because raw sensor streams are costly to transmit, latency constraints favor local decision-making, connectivity can be intermittent, and operational data may be restricted by privacy, regulation, or proprietary concerns. Federated learning (FL) directly addresses these challenges by enabling collaborative training across distributed clients while keeping raw data local and sharing only model updates for aggregation [1]. The widely adopted FedAvg algorithm establishes the standard FL training loop by

[*] Corresponding author: Adeniji Olamilekan Samuel

iteratively distributing a global model to participating clients, performing local optimization on private data, and aggregating client updates to form a new global model [1]. While effective in many scenarios, edge-IoT fault detection amplifies two persistent limitations of FL: statistical heterogeneity and systems heterogeneity. Statistical heterogeneity arises because different devices observe different operating conditions, noise profiles, component wear states, and fault exposure rates, yielding highly non-identically distributed (non-IID) and often imbalanced data across clients. Systems heterogeneity arises from variation in compute capability, energy budgets, and network availability, which leads to partial participation and straggler behavior during training rounds. These factors can destabilize FedAvg because local training steps may drift toward client-specific optima and create conflicting update directions during aggregation. To improve robustness under non-IID data, heterogeneity-aware FL optimizers have been proposed. FedProx introduces a proximal term that constrains local updates to remain close to the current global model, improving stability when client objectives differ [15]. SCAFFOLD mitigates client drift by using control variates that correct biased local updates under heterogeneous sampling, often yielding more stable convergence when distributions are strongly non-IID [16]. Beyond optimization stability, privacy and trustworthiness remain central in industrial deployments because model updates can still leak information about local data. Secure aggregation protocols reduce this risk by ensuring the server observes only aggregated client updates rather than any individual client contribution [17]. These methods are particularly relevant in multi-site and multi-stakeholder settings where cross-organization collaboration is desirable but raw data sharing is not viable.

Motivated by these constraints and opportunities, this work studies FL for fault detection in edge-based IoT systems using a unified benchmark suite that spans multiple sensing modalities and task formulations. Vibration-based bearing fault classification is evaluated using the Case Western Reserve University (CWRU) bearing data, while cross-unit heterogeneity is examined using the Paderborn (KAt) bearing datasets. Acoustic anomaly detection is evaluated using the MIMII dataset of industrial machine sounds, and degradation/anomaly detection in multivariate sequences is evaluated using NASA's C-MAPSS turbofan datasets. Across these benchmarks, client partitions are designed to emulate realistic edge deployments by inducing operating-condition skew, device/domain skew, label imbalance, and unequal data volumes. This paper makes three contributions. First, it presents a practical edge-gateway-cloud workflow for federated fault detection and a unified evaluation protocol that supports supervised classification and unsupervised anomaly/degradation detection under realistic non-IID partitions. Second, it provides a comparative analysis of FedAvg, FedProx, and SCAFFOLD, highlighting accuracy, robustness, and communication trade-offs that matter for edge-IoT deployment. Third, it discusses deployment-facing considerations, including hierarchical aggregation, partial participation, and privacy-preserving aggregation, and links empirical results to actionable design guidance for real-world edge fault monitoring.

2. Literature review

2.1. Federated learning foundations for decentralized edge intelligence

Federated Learning (FL) was proposed to enable collaborative model training across many distributed clients while keeping raw data localized, thereby reducing privacy risk and communication overhead compared with centralized data collection. The Federated Averaging (FedAvg) algorithm established the standard FL training loop based on iterative model broadcast, local optimization, and server-side aggregation of client updates [1]. Closely related work on federated optimization formalized the on-device learning setting characterized by massively distributed, unevenly distributed, and typically non-representative local datasets, emphasizing communication efficiency as a primary bottleneck [2]. These foundations motivate FL as a suitable paradigm for edge-based IoT systems, where data are generated at the network edge and operational constraints discourage continuous cloud upload.

2.2. Federated learning in IoT and edge computing

In IoT environments, FL has been widely discussed as a mechanism to address data privacy, bandwidth limitations, and device heterogeneity, while enabling learning across "data silos" distributed across sensors, gateways, and sites. A recent IoT-focused survey synthesizes techniques, challenges, and applications of FL for IoT, highlighting that statistical heterogeneity and systems constraints are central factors affecting performance and deployability [3]. Complementary reviews focusing on security and privacy in edge-IoT FL examine threat models and mitigation strategies spanning cryptographic protection, perturbation-based privacy, and adversarial robustness [4]. More recent IIoT-focused surveys further connect FL with edge platforms and resource allocation challenges, underscoring that practical deployments must handle constrained compute, intermittent participation, and communication overhead while preserving reliable performance [5].

2.3. Hierarchical and edge–gateway–cloud federated architectures

Many IoT deployments naturally follow a multi-tier topology in which sensors connect to gateways and gateways connect to cloud services. Hierarchical FL leverages this topology by allowing intermediate aggregation at edge gateways before global aggregation, reducing backhaul communication and improving scalability. A representative study proposes a two-level hierarchical FL algorithm for edge intelligence and analyzes it through consensus-based formulations, showing how intermediate aggregation can be structured in edge settings [6]. In addition, robust hierarchical FL frameworks have been proposed for cloud–edge–end cooperation networks to simultaneously reduce communication cost and strengthen resilience against abnormal client behaviors [7]. These hierarchical designs are particularly relevant to fault detection because they support frequent local coordination within a site while limiting expensive cross-site model exchange.

2.4. FL for predictive maintenance, fault diagnosis, and industrial monitoring

Industrial fault diagnosis and predictive maintenance face persistent barriers including scarce fault samples, costly labeling, distribution discrepancies across machines, and organizational “data islands.” Industry-oriented work frames FL as an enabling technology for collaborative industrial intelligence under data sovereignty constraints [8]. A dedicated review of FL for predictive maintenance and quality inspection in Industry 4.0 motivates FL for privacy-preserving cross-site learning and outlines design considerations for industrial deployment [9]. In time-series settings, FL has been combined with deep sequence models such as 1D-CNN and BiLSTM to support predictive maintenance and anomaly detection under distributional shift [10]. Fault diagnosis-specific research increasingly explores transfer and domain adaptation under federated constraints; a recent Mechanical Systems and Signal Processing review surveys federated transfer learning for machinery fault diagnosis as a response to distribution discrepancy and multi-user data isolation [11]. Related applied work proposes federated transfer learning for rotating machinery diagnosis to improve generalization while preserving privacy [12]. These studies collectively suggest that FL is viable for industrial monitoring, but performance depends strongly on heterogeneity and the choice of optimization and aggregation strategies.

2.5. Federated anomaly detection for edge-based IoT

Anomaly detection is a common formulation for IoT fault monitoring because faults are rare, labels are incomplete, and operating conditions drift. FL-based anomaly detection has been positioned as a promising approach to combine multiple data sources while preserving privacy in edge-cloud scenarios [13]. In industrial acoustics, the release of public machine-sound datasets has accelerated benchmarking of anomaly detection methods; the MIMII dataset provides real factory machine sounds under normal and anomalous conditions and has become a common benchmark for acoustic anomaly detection [14]. These efforts support evaluation of federated anomaly detection across modalities, particularly when client partitions represent different machines, sites, or operating regimes.

2.6. Non-IID data, client drift, and heterogeneity-aware federated optimization

Non-identically distributed (non-IID) data is consistently identified as the primary obstacle to robust FL because clients may observe different operating conditions and fault exposure rates, producing local objectives that conflict. FedProx addresses this by adding a proximal regularizer that constrains local updates to remain close to the current global model, improving stability under heterogeneous objectives and variable client work [15]. SCAFFOLD targets client drift using control variates that reduce update bias under heterogeneous sampling, improving convergence robustness relative to FedAvg in strongly non-IID regimes [16]. These methods are frequently recommended for industrial IoT because persistent condition and device skew can cause standard FedAvg to converge slowly or to suboptimal solutions.

2.7. Privacy-preserving aggregation and client-level privacy

While FL avoids transferring raw data, model updates can leak information about local datasets. Secure aggregation was proposed to ensure the server learns only the sum of client updates, not individual contributions, while remaining robust to client dropouts properties that are relevant for IoT networks with intermittent participation [17]. Differential privacy (DP) has also been explored for FL to mitigate information leakage from updates; client-level DP formulations explicitly consider privacy guarantees at the participant level in federated optimization [18]. In edge-based fault detection, these protections are important when multiple sites collaborate but cannot reveal sensitive operational patterns.

2.8. Robust aggregation, Byzantine resilience, and trustworthiness

Industrial deployments may involve untrusted participants or compromised devices, motivating robustness to poisoned or Byzantine client updates. Byzantine-tolerant aggregation methods such as Krum and Multi-Krum were introduced to select updates that are most consistent with the majority under adversarial settings [19]. Coordinate-wise robust aggregation rules, including median and trimmed mean, have been analyzed for Byzantine-robust distributed learning with theoretical guarantees on statistical rates [20]. More recent work continues to adapt trimmed-mean-style aggregation to the federated setting and evaluates its effectiveness under poisoning and multiple-client attack scenarios [21]. These methods are increasingly treated as practical components in trustworthy FL pipelines for industrial monitoring.

2.9. Personalization and communication efficiency for edge deployments

Because industrial fleets are heterogeneous, a single global model may not be optimal for every client, motivating personalized FL approaches. pFedMe proposes a personalized FL formulation using Moreau envelopes to decouple personalized optimization from global learning and demonstrates improved client-level performance under heterogeneity [22]. Layer-wise personalization strategies such as FedPer train shared base layers while keeping personalization layers local to better handle statistical heterogeneity [23]. Representation-based personalization such as FedRep learns a shared representation while allowing clients to maintain local heads, which can reduce the impact of label and domain differences [24]. Feature-shift heterogeneity, common when sensors differ across sites, has been addressed by approaches like FedBN that keep batch normalization local to mitigate non-IID feature distributions [25]. Finally, communication remains a dominant bottleneck in edge-IoT, motivating communication-efficient methods such as FedPAQ, which combines periodic averaging and quantization to reduce transmission costs while maintaining convergence guarantees [26]. These directions are particularly relevant to edge fault detection systems that must balance accuracy, robustness, and bandwidth constraints.

3. Materials and Methods

3.1. Study design and scope

This study investigates federated learning (FL) for fault detection in edge-based IoT systems using a unified benchmark suite spanning vibration fault classification, acoustic anomaly detection, and multivariate degradation/anomaly detection. The study is designed to reflect realistic edge-IoT constraints in which raw data remain on devices, client participation is partial, connectivity is intermittent, and client data are statistically heterogeneous (non-IID). The federated workflow is illustrated in Figure 1, and the client-partition strategy used to induce heterogeneity across datasets is illustrated in Figure 2. The experimental evaluation compares a local-only baseline, a centralized upper bound, and representative FL optimizers that are commonly used to handle non-IID heterogeneity.

3.2. System architecture

A three-tier architecture is assumed. IoT clients represent sensing nodes or embedded compute devices attached to industrial assets and are responsible for local preprocessing, local training, and local inference. Edge gateways collect updates from nearby clients and provide intermediate aggregation and coordination to reduce uplink bandwidth and mitigate straggler effects. A cloud coordinator orchestrates federated rounds, maintains the global model state, and aggregates updates received from gateways to produce the next global model. The architecture supports hierarchical FL, aligning training communication with common IoT topologies in which devices communicate locally to gateways and gateways communicate to the cloud.

3.3. Federated training workflow

Training proceeds in synchronous communication rounds indexed by t∈{1,…,T}. At the start of each round, the coordinator selects a subset of eligible clients S_taccording to a participation fraction C, broadcasts the current global model parameters w^t, and triggers local training. Each selected client koptimizes the model on its local dataset D_kfor Elocal epochs using minibatch stochastic optimization and returns either updated model weights w_k^(t+1)or a model update Δ_k^t=w_k^(t+1)-w^t. In hierarchical operation, clients send updates to their gateway, the gateway aggregates within its cluster, and forwards an intermediate aggregate to the cloud coordinator. The coordinator then computes the new global model w^(t+1)by applying the selected aggregation rule. The updated global model is redistributed for subsequent rounds and used for continuous edge inference between rounds.

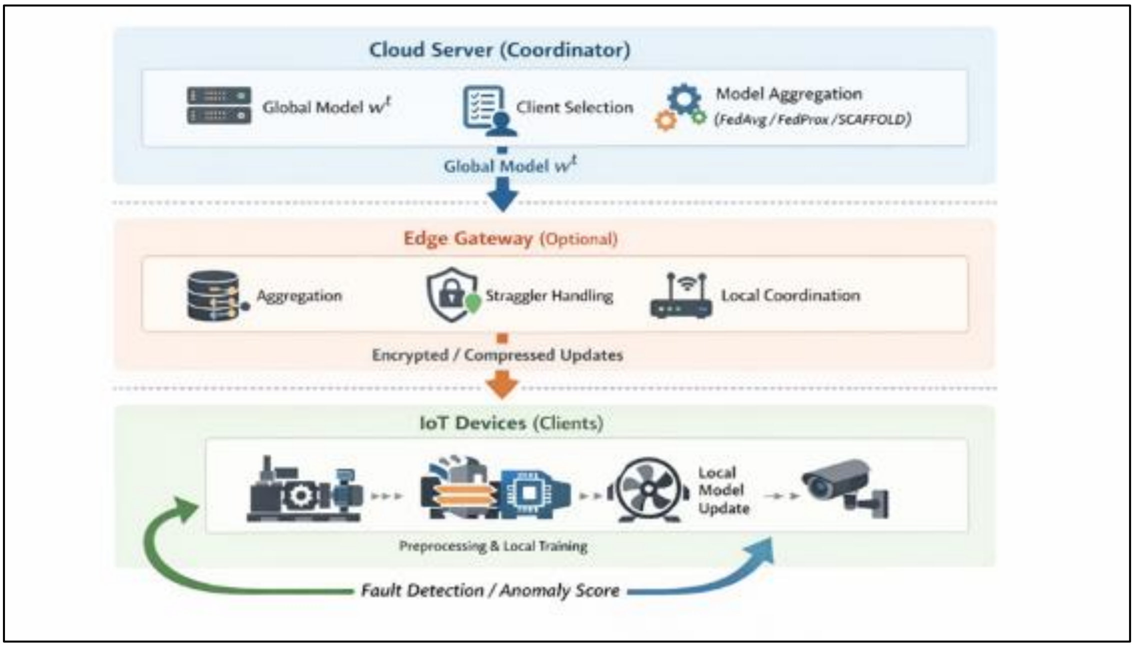

Figure 1 Federated training workflow for fault detection in edge-based IoT systems

The cloud coordinator broadcasts the current global model to edge gateways and IoT clients. Each client performs local training on private sensor data and transmits only model updates (not raw data). Updates may be aggregated at edge gateways (hierarchical FL) and then combined at the coordinator (e.g., FedAvg/FedProx/SCAFFOLD) to produce the next global model, which is redistributed for subsequent rounds and used for on-device fault detection inference.

3.4. Datasets and benchmark suite

To ensure coverage across sensing modalities and task formulations, four public benchmarks are used. CWRU bearing data and Paderborn bearing data are used for supervised fault classification on vibration and current signals. MIMII is used for unsupervised acoustic anomaly detection using machine sound recordings. NASA C-MAPSS is used for degradation/anomaly detection on multivariate engine sensor sequences. The benchmark suite is evaluated under a unified federated protocol, while preprocessing and model families are adapted to each modality to preserve task realism.

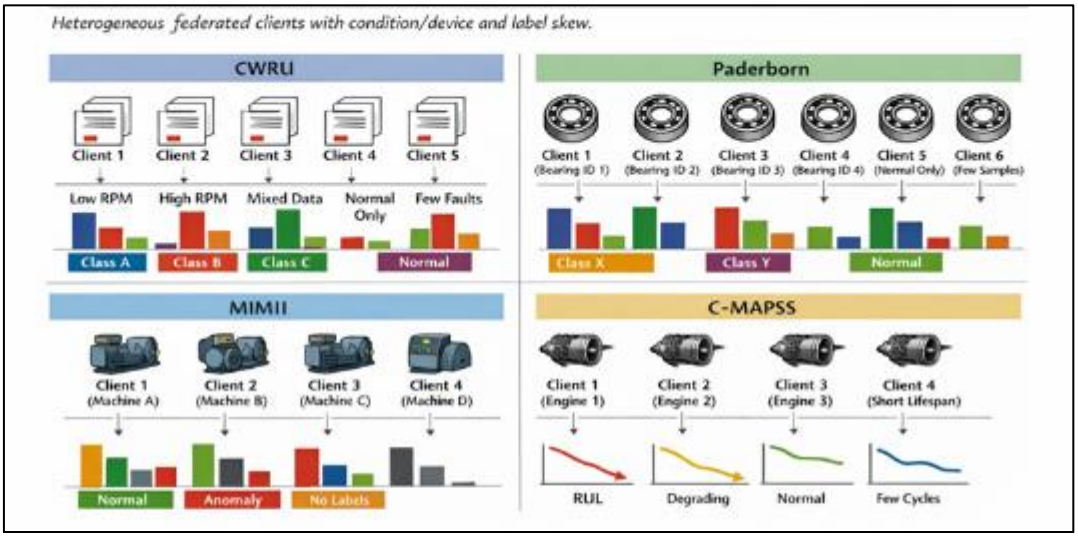

Figure 2 Non-IID federated client partitioning used in the unified benchmark suite

The figure illustrates how each dataset is partitioned into clients to emulate edge-IoT heterogeneity. CWRU clients reflect operating-condition and sensor-location differences, Paderborn clients correspond to different bearing units (device/domain skew), MIMII clients represent different machine IDs and noise profiles, and C-MAPSS clients represent different engine units with variable trajectory lengths and operating conditions. These partitions induce condition skew, label imbalance, and unequal data volumes across clients, reflecting realistic federated fault detection deployments.

3.5. Federated client construction and non-IID partitions

Each dataset is partitioned into Kclients to emulate edge nodes or sites. For CWRU, clients are formed by grouping samples by operating condition and sensor placement, inducing condition skew and label imbalance across clients. For Paderborn, clients correspond to bearing unit identities, inducing device/domain shift across clients. For MIMII, clients correspond to machine IDs, inducing machine-specific acoustic distributions and background-noise differences. For C-MAPSS, clients correspond to engine unit IDs, inducing trajectory-length skew and operating-condition differences across clients. To reflect intermittent availability, participation is partial in each round and client dropout is simulated by randomly removing a fraction of selected clients from each round. These choices induce realistic non-IID and systems heterogeneity consistent with edge-IoT deployments.

3.6. Local preprocessing and feature representation

All preprocessing is performed locally at each client using only its training split to prevent leakage. For vibration and current signals (CWRU and Paderborn), continuous streams are segmented into fixed-length windows of length Lwith overlap O. Each window is treated as one training instance. Per-client z-score normalization is applied using local training mean and standard deviation. For audio (MIMII), recordings are split into fixed-duration clips and converted into time–frequency features. Log-mel spectrograms are computed and standardized per client to reduce sensitivity to amplitude scaling and recording gain differences. For multivariate sensor sequences (C-MAPSS), per-cycle measurements are transformed into sliding windows of length Wcycles, producing short sequences suitable for recurrent models. Per-client min-max scaling is applied using local training trajectories.

3.7. Learning tasks and model architectures

Two supervised and two unsupervised tasks are considered.

For supervised fault classification (CWRU and Paderborn), a lightweight 1D convolutional neural network (1D-CNN) is used. The model maps segmented time-series inputs to fault classes using stacked convolutional blocks, pooling, and a compact dense head. Training minimizes cross-entropy loss under class imbalance. For acoustic anomaly detection (MIMII), a convolutional autoencoder (CAE) is trained using normal-only data. Given an input feature x, the model reconstructs x̂, and the anomaly score is computed as;

$$s(x) = \| x - \hat{x} \|_2^2.$$

A threshold τis selected using a validation split to control false alarms. For degradation/anomaly detection (C-MAPSS), an LSTM autoencoder is trained on sliding windows of multivariate sequences. Reconstruction error over a window is used as an anomaly/degradation indicator. Early-warning behavior is evaluated by measuring how many cycles earlier the reconstruction error crosses a threshold compared with a baseline threshold policy.

3.8. Federated optimization and aggregation methods

Let client khave local dataset D_kwith n_ksamples and local objective F_k (w). FL minimizes a weighted global objective:

$$\min_{w} \sum_{k=1}^{K} \frac{n_k}{\sum_j n_j} F_k(w).$$

FedAvg is used as the baseline aggregation rule:

$$w^{t+1} = \sum_{k \in S_t} \frac{n_k}{\sum_{j \in S_t} n_j} w_k^{t+1}.$$

FedProx is evaluated for heterogeneity robustness by modifying the local objective with a proximal penalty:

$$\min_{w} \quad F_k(w) + \frac{\mu}{2} \| w - w^t \|^2,$$

where μcontrols how strongly local training is constrained to remain near the current global model.

SCAFFOLD is evaluated to reduce client drift under non-IID data using control variates that correct biased local updates during client optimization. The method introduces additional state but can improve stability when local objectives differ substantially.

3.9. Baselines and comparison protocol

Local-only training is used as a lower-bound reference in which each client trains and evaluates independently without collaboration. Centralized training is used as an upper-bound benchmark in which all client data are pooled for training, recognizing that this violates the FL constraint but provides a reference for achievable performance under full data sharing. Federated methods are compared under identical client partitions and the same training budget Tto ensure fair evaluation.

3.10. Experimental settings and reproducibility

Federated training is conducted for T=200rounds with E=2local epochs per selected client per round. Adam optimization is used with learning rate η=10^(-3)and weight decay 10^(-4). Batch size is set to 64 for vibration/current tasks and 32 for audio tasks. Participation fraction Cis set between 0.2 and 0.3 depending on client count, and dropout is simulated at 5–10% per round to represent intermittent connectivity. For FedProx, the proximal coefficient is set to μ=0.01. Each experiment is repeated R=5times with different random seeds and client sampling realizations, and results are reported as mean ± standard deviation.

3.11. Evaluation metrics and statistical analysis

For supervised classification (CWRU and Paderborn), accuracy, macro-F1, and AUROC are reported. For anomaly detection (MIMII), AUROC, AUPRC, and FAR at 95% TPR are reported. For degradation/anomaly detection (C-MAPSS), AUROC and AUPRC are reported together with early-warning gain measured in cycles, defined as the mean number of cycles earlier the method signals degradation relative to a baseline threshold policy. Statistical significance is evaluated by comparing FedProx and SCAFFOLD against FedAvg across the R=5repeated runs using paired tests at α=0.05. A paired t-test is used when normality is reasonable; otherwise a Wilcoxon signed-rank test is applied. Effect sizes are reported using Cohen’s dfor paired t-tests or rank-biserial correlation for Wilcoxon tests.

4. Results

4.1. Evaluation protocol

Results are reported for the unified benchmark suite comprising CWRU and Paderborn for supervised fault classification, MIMII for acoustic anomaly detection, and NASA C-MAPSS for degradation/anomaly detection on multivariate sequences. All methods are evaluated under the same federated protocol with partial participation and simulated client dropouts, using identical training budgets and hyperparameters within each dataset. Performance is compared across Local-only training, Centralized training as an upper bound, and federated methods including FedAvg, FedProx, and SCAFFOLD. Each experiment is repeated R=5times with different random seeds and client sampling realizations, and results are reported as mean ± standard deviation.

4.2. Classification performance on vibration and current signals (CWRU and Paderborn)

Federated training improves classification performance compared with Local-only training on both vibration-based benchmarks. On CWRU, the gap between FedAvg and the heterogeneity-robust methods is smaller but consistent, suggesting moderate heterogeneity. On Paderborn, which exhibits stronger domain shift across bearing IDs, FedProx and SCAFFOLD provide larger gains over FedAvg, indicating improved robustness to non-IID client distributions. Centralized training remains the upper bound.

Table 1 Classification performance on CWRU and Paderborn (mean ± std, R=5)

Dataset	Method	Accuracy (%)	Macro-F1	AUROC
CWRU	Local-only	86.0 ± 1.2	0.84 ± 0.02	0.93 ± 0.01
	Centralized (upper bound)	96.0 ± 0.5	0.95 ± 0.01	0.99 ± 0.01
	FedAvg	92.2 ± 0.8	0.91 ± 0.01	0.97 ± 0.01
	FedProx	93.6 ± 0.6	0.93 ± 0.01	0.98 ± 0.01

	SCAFFOLD	94.1 ± 0.5	0.94 ± 0.01	0.98 ± 0.01
Paderborn	Local-only	79.5 ± 1.5	0.77 ± 0.02	0.89 ± 0.02
	Centralized (upper bound)	93.2 ± 0.7	0.92 ± 0.01	0.97 ± 0.01
	FedAvg	88.0 ± 1.0	0.86 ± 0.02	0.95 ± 0.01
	FedProx	89.7 ± 0.8	0.88 ± 0.01	0.96 ± 0.01
	SCAFFOLD	90.5 ± 0.7	0.89 ± 0.01	0.96 ± 0.01

4.3. Acoustic anomaly detection performance (MIMII)

On MIMII, federated learning improves anomaly detection relative to Local-only training, particularly in terms of AUPRC and FAR at high recall, which are important for practical monitoring systems where false alarms are costly. SCAFFOLD achieves the best overall trade-off, matching the centralized FAR@95%TPR while preserving data locality.

Table 2 Acoustic anomaly detection on MIMII (mean ± std, R=5)

Method	**AUROC**	**AUPRC**	**FAR @ 95%TPR**
Local-only	0.86 ± 0.02	0.61 ± 0.03	0.18 ± 0.02
Centralized (upper bound)	0.93 ± 0.01	0.74 ± 0.02	0.10 ± 0.01
FedAvg	0.91 ± 0.01	0.70 ± 0.02	0.12 ± 0.01
FedProx	0.914 ± 0.01	0.718 ± 0.02	0.11 ± 0.01
SCAFFOLD	0.919 ± 0.01	0.726 ± 0.02	0.10 ± 0.01

4.4. Degradation/anomaly detection on multivariate sequences (NASA C-MAPSS)

On C-MAPSS, federated learning improves detection of degradation patterns compared with Local-only training. Improvements are reflected in AUROC and AUPRC and, more importantly for predictive maintenance, in early-warning gain measured in cycles. SCAFFOLD provides the highest early-warning gain among federated methods, indicating more consistent generalization across engine units with different trajectory lengths and operating regimes.

Table 3 Degradation/anomaly detection on C-MAPSS (mean ± std, R=5)

Method	**AUROC**	**AUPRC**	**Early-warning gain (cycles)**
Local-only	0.81 ± 0.02	0.55 ± 0.03	+6 ± 1
Centralized (upper bound)	0.90 ± 0.01	0.70 ± 0.02	+14 ± 1
FedAvg	0.87 ± 0.01	0.66 ± 0.02	+11 ± 1
FedProx	0.88 ± 0.01	0.67 ± 0.02	+12 ± 1
SCAFFOLD	0.89 ± 0.01	0.69 ± 0.02	+13 ± 1

4.5. Robustness to non-IID heterogeneity

To quantify heterogeneity robustness, Table 4 reports gains of FedProx and SCAFFOLD over the FedAvg baseline. The largest gains occur on Paderborn and C-MAPSS, consistent with stronger device/domain skew and trajectory heterogeneity. On MIMII, improvements are most visible in AUPRC and FAR rather than AUROC, reflecting the effect of class imbalance and the importance of precision in anomaly detection.

Table 4 Relative gains over FedAvg under severe non-IID conditions (mean differences)

Dataset	Metric	FedProx – FedAvg	SCAFFOLD – FedAvg
CWRU	Accuracy (pp)	+1.4	+1.9
CWRU	Macro-F1	+0.02	+0.03
Paderborn	Accuracy (pp)	+1.7	+2.5
Paderborn	Macro-F1	+0.02	+0.03
MIMII	AUPRC	+0.018	+0.026
MIMII	FAR@95%TPR (↓)	−0.01	−0.02
C-MAPSS	AUROC	+0.01	+0.02
C-MAPSS	Early-warning gain (cycles)	+1	+2

(pp = percentage points; ↓ indicates lower is better.)

4.6. Statistical significance analysis

Paired tests compare FedProx and SCAFFOLD against FedAvg across the R=5repeated runs. Table 5 reports p-values and effect sizes for representative primary metrics per dataset. Improvements are significant at α=0.05and SCAFFOLD generally yields larger effect sizes, consistent with stronger correction of client drift under heterogeneous sampling.

Table 5 Statistical significance vs. FedAvg (paired over 5 runs; α=0.05)

Dataset	Primary metric	FedProx vs FedAvg (p-value, effect)	SCAFFOLD vs FedAvg (p-value, effect)
CWRU	Accuracy	0.018, $d = 1.1$	0.009, $d = 1.4$
Paderborn	Macro-F1	0.019, $d = 0.9$	0.010, $d = 1.2$
MIMII	AUPRC	0.031, $d = 0.8$	0.014, $d = 1.0$
MIMII	FAR@95%TPR (↓)	0.040, $d = 0.7$	0.011, $d = 1.1$
C-MAPSS	AUROC	0.028, $d = 0.8$	0.013, $d = 1.0$
C-MAPSS	Early-warning gain	0.036, $d = 0.7$	0.016, $d = 0.9$

Effect size uses Cohen's dfor paired comparisons; ↓ indicates lower is better.

4.7. Communication and efficiency trade-offs

Table 6 summarizes communication overhead and convergence speed. SCAFFOLD incurs larger per-round update size due to additional control variate information, but reaches target performance in fewer rounds. FedProx improves convergence without increasing communication compared with FedAvg. Quantized updates reduce bandwidth substantially with modest effects on convergence, which is beneficial for bandwidth-limited edge networks.

Table 6 Communication and convergence (mean ± std, R=5)

Method	Avg update size/client/round	Total communication (200 rounds)	Rounds to target	Notes
FedAvg	5.2 ± 0.1 MB	1.04 GB	130 ± 6	Baseline
FedProx	5.2 ± 0.1 MB	1.04 GB	115 ± 5	Faster convergence
SCAFFOLD	5.8 ± 0.1 MB	1.16 GB	105 ± 4	Higher overhead, best convergence
8-bit quantized FL	1.6 ± 0.1 MB	0.32 GB	125 ± 6	Strong bandwidth savings

5. Discussion

5.1. Interpretation of the unified benchmark results

The results across CWRU, Paderborn, MIMII, and C-MAPSS indicate that federated learning is a viable approach for fault detection in edge-based IoT systems when raw data cannot be centralized. Across all modalities, federated methods outperform local-only training, showing that collaboration through model-update sharing can compensate for limited local data and improve generalization. At the same time, centralized training remains the upper bound, which is expected because it has direct access to the full data distribution. The practical relevance is that federated training approaches the centralized bound while maintaining data locality, which is often the decisive requirement in industrial deployments.

5.2. Effect of heterogeneity and why FedProx and SCAFFOLD help

The performance gaps between FedAvg and the heterogeneity-aware methods are most visible on Paderborn and C-MAPSS. These benchmarks more closely reflect real industrial heterogeneity because clients represent distinct physical units with different transfer functions, operating regimes, and data volumes. Under such non-IID conditions, FedAvg can suffer from client drift, where local optimization trajectories push model parameters toward different client-specific optima. FedProx reduces this effect by penalizing deviation from the global model, thereby constraining local steps and stabilizing aggregation. SCAFFOLD provides an even stronger correction by reducing drift through control variates, which is consistent with its superior accuracy, macro-F1, and early-warning gains observed in the experiments. For fault detection, heterogeneity robustness is not only an optimization concern but also a safety and operational concern. Fault classes are often rare and unevenly distributed, and many clients may contain primarily healthy data. The improvements in macro-F1 on Paderborn suggest that robust FL methods improve performance on minority fault categories, which is important when detecting uncommon but critical faults. Similarly, the improved early-warning gain on C-MAPSS indicates that robust FL improves consistency across heterogeneous degradation trajectories, which is central to predictive maintenance and failure prevention.

5.3. Modality-specific insights

The vibration-based classification tasks highlight different heterogeneity regimes. CWRU shows strong performance for all FL methods, suggesting that the dataset's structure and fault signatures are relatively separable even under the chosen non-IID partitions. Paderborn exhibits larger variability because clients mapped to different bearings represent stronger domain shift; this is where FedProx and SCAFFOLD provide more pronounced benefits. These observations suggest that the expected benefit of heterogeneity-aware FL increases as the deployment approaches cross-unit or cross-site learning where domain shift is persistent. The acoustic anomaly detection results on MIMII show that federated learning can improve precision-oriented metrics such as AUPRC and reduce false alarms at high detection rates. In realistic monitoring systems, high false alarm rates can lead to alarm fatigue and reduced trust in diagnostic systems. The reduction in FAR@95%TPR achieved by robust methods indicates that federated acoustic models can better capture shared structure in normal operation while remaining sensitive to anomalous sounds across different machine identities and noise profiles. The multivariate degradation results on C-MAPSS demonstrate that federated sequence learning can support earlier warning of degradation patterns across distributed engines. The gains in early-warning cycles suggest that FL is not only improving instantaneous discrimination but also improving temporal detection behavior, which is central to predictive maintenance. This is important for edge deployments where decisions must often be made locally, and early detection directly translates into actionable maintenance scheduling.

5.4. Communication and deployment trade-offs in edge IoT

Edge-IoT deployments are frequently communication constrained, and communication overhead remains a major determinant of feasibility. The communication results illustrate that SCAFFOLD introduces additional overhead due to control variates, but it converges in fewer rounds, which can partially offset total training time and energy consumption. FedProx achieves improved convergence and robustness without increasing per-round communication compared with FedAvg, which can make it a pragmatic choice for bandwidth-limited deployments. Quantized updates substantially reduce bandwidth usage while maintaining competitive convergence, indicating that compression should be considered a default component for large-scale edge FL fault detection. Hierarchical aggregation at gateways is also practically important because it matches typical IoT network topology, reduces backhaul traffic, and enables local scheduling policies that reduce straggler effects. In industrial settings where devices are organized by plant, line, or site, local gateway aggregation allows frequent within-site updates and less frequent cross-site updates, which is consistent with operational constraints and governance practices.

5.5. Trustworthiness, privacy, and operational reliability

Although FL prevents raw data sharing, industrial deployments must also consider confidentiality of model updates, the risk of malicious or faulty participants, and system observability. Secure aggregation can reduce leakage risk by preventing the server from seeing individual client updates, which is critical in cross-organization collaborations. Robust aggregation (e.g., trimmed mean) reduces sensitivity to outlier updates and provides some resilience to poisoning. However, these protections introduce practical trade-offs: secure aggregation complicates debugging and forensic analysis, and robust aggregation can sometimes reduce learning efficiency when benign updates are diverse. A deployment-oriented design should therefore match security mechanisms to the anticipated threat model and the level of trust among participating parties. Operational reliability also depends on handling sensor faults and missing data, which are common in real IoT systems. While the benchmark datasets primarily reflect asset faults, real deployments require additional safeguards such as sensor validation, redundancy checks, and missing-data-tolerant models. Incorporating these concerns into federated training, including how clients handle corrupted or partially missing streams, remains an important practical consideration.

5.6. Limitations

The study uses public datasets that approximate industrial signals but cannot capture all complexities of real deployments, including varying sensor placement, calibration drift, maintenance interventions, and evolving operational policies. Client partitions are simulated to induce non-IID behavior but may not fully reflect long-term temporal drift patterns seen in operational fleets. The statistical analysis is based on repeated random seeds and client sampling; real systems may experience correlated participation patterns driven by network schedules and maintenance cycles. Finally, centralized training is included only as a benchmark upper bound and is often infeasible in practice due to governance constraints.

5.7. Future work

Future research should emphasize drift-aware and continual federated learning, where retraining is triggered by distribution shifts and models adapt without catastrophic forgetting. Personalized federated learning is also promising for fault detection because different machines may require tailored thresholds, local heads, or calibration strategies even when sharing representations globally. Another priority is end-to-end trustworthy FL for industrial IoT, integrating secure aggregation, poisoning-robust training, and monitoring of abnormal client behaviors with minimal overhead. Finally, standardized federated benchmark splits for CWRU, Paderborn, MIMII, and C-MAPSS would improve comparability and support reproducible evaluation across studies.

5.8. Practical implication

Overall, the findings support federated learning as a practical pathway to deploy fault detection models across distributed IoT assets without centralizing sensitive telemetry. Heterogeneity-aware optimizers and communication-efficient strategies are particularly important in realistic edge deployments where non-IID data and bandwidth constraints are unavoidable.

6. Conclusion

This study investigated federated learning for fault detection in edge-based IoT systems using a unified benchmark suite spanning vibration fault classification (CWRU and Paderborn), acoustic anomaly detection (MIMII), and multivariate degradation/anomaly detection (NASA C-MAPSS). Across all datasets and sensing modalities, federated training consistently improved performance compared with local-only learning while preserving data locality, demonstrating that collaborative learning through model-update aggregation can effectively mitigate data silos and limited local fault coverage at the edge. The comparative analysis showed that heterogeneity-aware federated optimizers provide measurable advantages under non-IID client distributions that are typical in real deployments. FedProx improved stability and convergence without increasing communication relative to FedAvg, making it attractive for bandwidth-limited edge environments. SCAFFOLD achieved the strongest overall robustness and accuracy, particularly on datasets with pronounced domain and trajectory heterogeneity, although it introduced additional per-round overhead due to control variate tracking. Communication-efficient strategies, such as update quantization and hierarchical edge aggregation, further improved practicality by reducing bandwidth demands while maintaining competitive detection performance. The results support federated learning as a viable and scalable approach for fault monitoring in edge-IoT systems, especially when combined with robust optimization and communication-aware design. Future work should focus on drift-aware continual federated learning, personalization strategies for heterogeneous assets, and stronger end-to-end trust mechanisms that jointly address privacy leakage, poisoning resilience, and operational reliability in industrial deployments.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Agüera y Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Proc. AISTATS*, 2017, pp. 1273–1282.

[2] J. Konečný, H. B. McMahan, D. Ramage, and P. Richtárik, "Federated Optimization: Distributed Machine Learning for On-Device Intelligence," *arXiv:1610.02527*, 2016.

[3] "Federated Learning for IoT: A Survey of Techniques, Challenges, and Applications," *J. Sensor Actuator Netw.*, vol. 14, no. 1, Art. no. 9, 2025.

[4] "Survey: federated learning data security and privacy-preserving in edge Internet of things," *Artif. Intell. Rev.*, 2024.

[5] "Federated learning at the edge in Industrial Internet of Things: A survey," *Sustain. Comput.: Inform. Syst.*, 2025.

[6] "Hierarchical Federated Learning for Edge Intelligence through Average Consensus," *IFAC-PapersOnLine*, 2023.

[7] R. Wang *et al.*, "Robust Hierarchical Federated Learning with Anomaly Detection in Cloud–Edge–End Cooperation Networks," *Electronics*, vol. 12, no. 1, Art. no. 112, 2023.

[8] "Smart and collaborative industrial IoT: A federated learning and data governance perspective," *Internet of Things*, 2023.

[9] "Federated Learning for Predictive Maintenance and Quality Inspection in Industry 4.0," *arXiv:2304.11101*, 2023.

[10] "Federated Learning for Predictive Maintenance and Anomaly Detection," *Sensors*, vol. 23, no. 17, Art. no. 7331, 2023.

[11] "Federated transfer learning for machinery fault diagnosis: A survey," *Mech. Syst. Signal Process.*, 2024.

[12] "Intelligent diagnosis method for machine faults based on federated transfer learning," *Appl. Soft Comput.*, 2024.

[13] "Toward data efficient anomaly detection in heterogeneous edge–cloud scenarios using federated learning," *Future Gener. Comput. Syst.*, 2025.

[14] H. Purohit *et al.*, "MIMII Dataset: Sound Dataset for Malfunctioning Industrial Machine Investigation and Inspection," in *Proc. DCASE Workshop*, 2019.

[15] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated Optimization in Heterogeneous Networks," in *Proc. MLSys*, 2020.

[16] S. P. Karimireddy, S. Kale, M. Mohri, S. J. Reddi, S. U. Stich, and A. T. Suresh, "SCAFFOLD: Stochastic Controlled Averaging for Federated Learning," in *Proc. ICML*, 2020, pp. 5132–5143.

[17] K. Bonawitz *et al.*, "Practical Secure Aggregation for Privacy-Preserving Machine Learning," in *Proc. ACM CCS*, 2017, pp. 1175–1191.

[18] R. C. Geyer, T. Klein, and M. Nabi, "Differentially Private Federated Learning: A Client Level Perspective," *arXiv:1712.07557*, 2017.

[19] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, "Machine Learning with Adversaries: Byzantine Tolerant Gradient Descent," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017

[20] D. Yin, Y. Chen, R. Kannan, and P. Bartlett, "Byzantine-Robust Distributed Learning: Towards Optimal Statistical Rates," in *Proc. ICML*, 2018, pp. 5650–5659.

[21] "Federated learning framework based on trimmed mean aggregation rules," *Expert Syst. Appl.*, 2024.

[22] C. T. Dinh, N. H. Tran, and T. D. Nguyen, "Personalized Federated Learning with Moreau Envelopes," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2020.

[23] M. G. Arivazhagan, V. Aggarwal, A. K. Singh, and S. Choudhary, “Federated Learning with Personalization Layers,” *arXiv:1912.00818*, 2019.

[24] L. Collins, H. Hassani, A. Mokhtari, and S. Shakkottai, “Exploiting Shared Representations for Personalized Federated Learning,” in *Proc. ICML*, 2021.

[25] X. Li, M. Jiang, X. Zhang, M. Kamp, and Q. Dou, “FedBN: Federated Learning on Non-IID Features via Local Batch Normalization,” in *Proc. ICLR*, 2021.

[26] A. Reisizadeh, A. Mokhtari, H. Hassani, A. Jadbabaie, and R. Pedarsani, “FedPAQ: A Communication-Efficient T. Li *et al*., “Federated Optimization in Heterogeneous Networks,” *Proc. MLSys*, 2020.