Cureus Journal of Computer Science
https://doi.org/10.7759/s44389-026-00390-7

Original Article

Federated Learning for Internet of Things (IoT) Intrusion Detection Using a Feature Tokenizer Transformer

Vikash Kumar [1] , Nishtha Kesswani [1, ✉]

1. *Department of Data Science and Analytics, Central University of Rajasthan, Ajmer, IND*

Received: September 07, 2026 | Review began: October 01, 2026 | Review ended: October 26, 2026 | Published: November 25, 2026

© **Copyright** 2026
This is an open access article distributed under the terms of the Creative Commons Attribution License CC-BY 4.0., which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract

Introduction

Federated Learning (FL) enables privacy-preserving intrusion detection in distributed Internet of Things (IoT) networks by training models locally without exposing raw traffic to a central server. Several federated intrusion detection studies employ convolutional or recurrent architectures for network traffic classification. However, these architectures may not fully exploit feature interactions within heterogeneous tabular traffic data.

Methods

We introduce FL-FTTransformer, a federated intrusion detection system based on an FT-Transformer and trained with Federated proximal (FedProx), a class-weighted focal loss, and a five-phase learning-rate schedule. Training runs for 300 rounds on MQTTset. Each scalar feature is projected into a shared token space, and multi-head self-attention models inter-feature correlations without handcrafted feature engineering. Synthetic Minority Over-sampling Technique-Edited Nearest Neighbors (SMOTE-ENN) oversampling expands the dataset from 330,926 to 659,084 samples to counter severe class imbalance across six Message Queuing Telemetry Transport attack categories. Five clients were simulated under a Dirichlet non-independent and identically distributed partition ($\alpha = 3.0$) on a dual NVIDIA Tesla T4 graphics processing unit node.

Results

On a held-out test split of 131,817 samples, FL-FTTransformer achieved 95.25% accuracy, 95.68% macro precision, 94.85% macro recall, and 95.20% macro F1-score. Relative to the recent FL-CNN baseline, macro F1 improved by 1.16 percentage points (pp), macro precision by 0.89 pp, and macro recall by 1.55 pp, with the largest per-class gains observed for denial-of-service (+5.17 pp F1) and Flood (+7.07 pp F1) traffic. An ablation study confirmed that SMOTE-ENN, FedProx, and focal loss each contribute independently to final performance, with SMOTE-ENN removal producing the largest degradation (-2.58 pp macro F1).

Conclusions

Attention-based tabular modeling combined with federated training provides an effective and privacy-preserving approach to multi-class IoT intrusion detection, outperforming a recent convolutional neural network-based federated baseline on macro F1, precision, and recall, as well as centralized deep learning pipelines, without requiring a separate feature-selection stage.

Categories: IoT Security and Privacy, Machine Learning (ML)

Keywords: federated learning, intrusion detection, iot security, intenet of things, mqtt

How to cite this article:
Kumar V, Kesswani N (November 25, 2026) Federated Learning for Internet of Things (IoT) Intrusion Detection Using a Feature Tokenizer Transformer. Cureus J Comput Sci 3 : es44389-026-00390-7. DOI https://doi.org/10.7759/s44389-026-00390-7

Introduction

The rapid proliferation of Internet of Things (IoT) devices has widened the attack surface of modern networks. IoT-enabled sensor deployments are particularly exposed to brute-force credential attacks, Denial-of-Service (DoS) flooding, malformed-packet injection, and slow-connection exhaustion attacks [1]. Some existing work has proposed centralized intrusion detection systems [2,3].

Centralized intrusion detection systems (IDS) cannot scale to the heterogeneous, geographically dispersed topology of IoT deployments, and transmitting raw traffic to a central server conflicts with data-privacy requirements. Federated Learning (FL) [4] addresses this limitation by training a shared model collaboratively across clients, exchanging only gradient updates or model weights rather than raw data. Each IoT gateway trains a local model on its own traffic, and a central aggregator synthesizes a global model via weighted averaging. Prior FL-based IDS work has reported strong detection accuracy on IoT benchmarks [5-7], yet the deep-learning backbone is almost universally a convolutional neural network (CNN) or long short-term memory (LSTM) designed for spatial or temporal signals. Network traffic features are, however, fundamentally tabular: heterogeneous scalar measurements (packet lengths, flag fields, inter-arrival times) with no spatial locality between adjacent features.

The Feature Tokenizer Transformer (FT-Transformer) [8] was designed specifically for tabular data. It projects every scalar feature into a unified token embedding through per-feature linear layers, prepends a learnable [CLS] token, and applies standard Transformer encoder blocks, allowing self-attention to weigh each feature's contribution as a soft, end-to-end feature-selection mechanism. To the best of our knowledge, no prior study has embedded this architecture inside a federated IDS pipeline.

The contributions of this work are as follows:

We introduce FL-FTTransformer, the first federated IDS to use an FT-Transformer backbone on IoT network traffic, learning feature interactions end-to-end without manual feature selection.

We integrate Federated proximal (FedProx) regularization [9] with a dynamic proximal term, a five-phase multi-round learning-rate schedule, and periodic Dirichlet client reshuffling to stabilize convergence under non-independent and identically distributed (IID) traffic distributions.

We combine focal loss with per-class weight boosting and Synthetic Minority Over-sampling Technique-Edited Nearest Neighbors (SMOTE-ENN) oversampling to address severe class imbalance in the MQTTset benchmark, improving recall on minority attack classes.

Experiments on MQTTset demonstrate 95.20% macro F1, surpassing the FL-CNN baseline of Karunamurthy et al. [5] by 1.16 percentage points (pp) and the retained three classical machine-learning baselines - MLP, Gradient Boost, and Decision Tree - across every macro metric.

Related work

Federated Learning for IDS

Rahman et al. [10] compared centralized, on-device, and federated learning strategies for IoT intrusion detection, but the study is primarily analytical and does not propose or evaluate a concrete model on a standardized benchmark. Rashid et al. [6] deployed federated averaging (FedAvg) with a CNN backbone on the Edge-IIoT dataset (92.49% accuracy); the CNN's spatial inductive bias is ill-suited to tabular traffic features, and class imbalance in multi-class detection is not addressed. Hamdi [7] trained a federated LSTM that assumes temporal ordering among features, an assumption that does not hold for packet-level tabular records, and evaluated on a non-public dataset that limits reproducibility. Mothukuri et al. [11] applied federated gated recurrent unit (GRU)-based anomaly detection but evaluated only a binary anomaly task, providing no per-attack-class discrimination. Janati Idrissi et al. [12] proposed an autoencoder-based federated detector (Fed-ANIDS) that is inherently limited to binary normal-versus-anomaly decisions.

How to cite this article:
Kumar V, Kesswani N (November 25, 2026) Federated Learning for Internet of Things (IoT) Intrusion Detection Using a Feature Tokenizer Transformer. Cureus J Comput Sci 3 : es44389-026-00390-7. DOI https://doi.org/10.7759/s44389-026-00390-7

Zhao et al. [13] combined federated learning with an LSTM network for intrusion detection, but as with Hamdi [7], the LSTM's temporal-sequence assumption is a poor match for tabular, per-packet traffic records. Zhang et al. [14] applied transfer-learning approaches within a federated Industrial Internet of Things (IIoT) intrusion detection framework to address cross-domain distribution shift, but did not evaluate a Transformer-based tabular backbone or address the severe class imbalance present in protocol-level benchmarks such as MQTTset.

The most directly comparable prior work is Karunamurthy et al. [5], who deployed a federated CNN classifier on MQTTset with Chimp Optimization Algorithm (ChOA) feature selection, reaching 95.59% accuracy and 94.04% macro F1. Three limitations remain in that work: (i) the CNN backbone treats flattened traffic features as pseudo-spatial signals rather than modeling inter-feature dependencies directly; (ii) ChOA-based discrete feature selection is a separate preprocessing stage that cannot adapt to federated data heterogeneity; and (iii) macro recall on minority attack classes (DoS: 87.5%, Malformed: 86.0%) indicates under-detection of the most disruptive attack types. FL-FTTransformer directly addresses all three limitations.

Centralized Deep Learning on MQTTset

MQTTset [1] is the most relevant publicly available benchmark for IoT-specific intrusion detection, capturing real protocol-level attack patterns rather than the generic traffic represented in datasets such as NSL-KDD or CICIDS. Zeghida et al. [2] applied ensemble methods on a binary-balanced version of MQTTset (≈95% accuracy) but lost per-attack-type discrimination. Vijayan and Sundar [3] addressed the full six-class task with a centralized MAOA-tuned Autoencoder-LSTM-deep belief network (DBN) pipeline, but the centralized design exposes raw gateway traffic to a central server. Neither work combines multi-class detection with federated privacy.

We restrict our evaluation to MQTTset rather than general-purpose IDS benchmarks such as NSL-KDD [15] or CICIDS2017 [16] because those datasets capture conventional enterprise network traffic and do not reflect attacks expressed within the Message Queuing Telemetry Transport (MQTT) publish-subscribe protocol itself, which is the dominant communication layer for resource-constrained IoT devices. MQTTset additionally provides a six-class, protocol-level attack taxonomy rather than a coarse binary anomaly label, allowing per-attack-type evaluation that directly matches the operational threat model of MQTT-based IoT gateways. We treat cross-dataset validation on broader IoT benchmarks as future work (Section 5) rather than a claim of general applicability.

Transformers for Tabular Data

TabNet [17] introduced sequential attention for tabular feature selection. Gorishniy et al. [8] proposed FT-Transformer and showed it to be competitive with or superior to gradient-boosted trees across eleven tabular benchmarks. Unlike vision Transformers, FT-Transformer treats each feature as an independent token, allowing attention weights to model pairwise feature dependencies without spatial inductive bias - a property well suited to heterogeneous IoT traffic. To the best of our knowledge, based on our review of the federated IDS and tabular-Transformer literature above, no prior work has combined FT-Transformer with federated learning for network intrusion detection.

Class Imbalance in IDS

Legitimate traffic typically dominates real-world network traffic by orders of magnitude, a problem consistently documented across benchmark IDS datasets [18]. Data-level resampling remains the most common family of solutions: SMOTE [19] generates synthetic minority samples via interpolation, while its variant SMOTE-ENN additionally applies Edited Nearest Neighbors cleaning to remove noisy borderline samples, producing a cleaner decision boundary than SMOTE alone. Abdelkhalek and Mashaly [18] proposed a comparable hybrid resampling strategy that combines Adaptive Synthetic Sampling (ADASYN) with Tomek-Link undersampling on NSL-KDD, reporting consistent macro-F1 gains over single-technique resampling and supporting the general principle that pairing synthetic oversampling with a noise-removal or undersampling step outperforms oversampling alone. More recently, generative approaches such as TMG-GAN [20] employ a multi-generator adversarial framework to synthesize minority-class attack traffic, achieving strong results on centralized IoT intrusion benchmarks; however, GAN-based augmentation requires substantially more training overhead per client and is prone to mode collapse under the small, non-IID client partitions characteristic of federated

learning, making it a less practical fit for our setting. At the algorithm level, focal loss [21] down-weights easy examples dynamically, allowing the model to focus on hard-to-classify minority attack packets; Mulyanto et al. [22] specifically validated focal loss for minority-class network intrusion detection, reporting consistent recall improvements over standard cross-entropy on deep neural network and CNN classifiers.

We adopt SMOTE-ENN combined with class-weighted focal loss, rather than ADASYN-Tomek or GAN-based augmentation, for three reasons. First, SMOTE-ENN's neighborhood-cleaning step suits the six-class, tabular structure of MQTTset, where borderline synthetic samples between similar traffic types (e.g., DoS and Flood) can otherwise blur the decision boundary. Second, SMOTE-ENN operates independently and inexpensively on each client's local partition before every communication round, whereas GAN-based augmentation would require training and synchronizing a separate generator per client - or centrally, which conflicts with the federated privacy objective - substantially increasing computation and communication costs across 300 rounds. Third, combining data-level balancing (SMOTE-ENN) with algorithm-level reweighting (focal loss with per-class boosting) addresses complementary failure modes: the former corrects the raw class distribution seen during local training, while the latter continues to emphasize hard examples that remain difficult even after resampling. Our ablation study confirms that both components contribute independently to final performance, with resampling removal producing the largest single degradation.

Materials And Methods

Figure *1* summarizes the end-to-end FL-FTTransformer pipeline; each stage is described below.

Dataset and preprocessing

We use MQTTset [1], a publicly available benchmark of MQTT-protocol traffic comprising 330,926 labeled flows across six classes: Bruteforce, DoS, Flood, Legitimate, Malformed, and SlowITe. The original dataset contains 34 columns, including the target label. After removing the target column and encoding categorical variables, 33 numerical input features remain. Object-type columns are converted to floating point via hexadecimal parsing or label encoding. Infinite and NaN values are replaced with column medians. Feature standardization (zero mean, unit variance) is deferred until after class balancing and the train-test split; scaling statistics are fitted only on the training partition to prevent leakage into the test set.

Class balancing via SMOTE-ENN

The dataset is heavily skewed: DoS flooding accounts for more than 39% of samples, while SlowITe comprises fewer than 3%. SMOTE-ENN [19] synthesizes minority-class instances via k-nearest-neighbor interpolation and then applies ENN cleaning to remove ambiguous borderline samples. Balancing is applied to the full preprocessed dataset prior to the train-test split; this ordering, consistent with the comparison baseline of Karunamurthy et al. [5], is discussed further below.

FT-Transformer architecture

The FT-Transformer consists of a feature tokenizer, a stack of Transformer encoder blocks, and a classification head.

Feature Tokenizer
For an input vector $\boldsymbol{x} \in \mathbb{R}^{F}$ (F=33), each scalar $\boldsymbol{x_f}$ is mapped to a $\boldsymbol{d}$ -dimensional token:

$$t_f = x_f \cdot w_f + b_f, \qquad t_f \in \mathbb{R}^d \tag{1}$$

where $w_f, b_f \in \mathbb{R}^d$ are learnable per-feature parameters. A learnable [CLS] token $e_{cls} \in \mathbb{R}^d$ is prepended to yield the sequence $T = [e_{cls}; t_1; \ldots; t_F] \in \mathbb{R}^{(F+1)\times d}$.

How to cite this article:
Kumar V, Kesswani N (November 25, 2026) Federated Learning for Internet of Things (IoT) Intrusion Detection Using a Feature Tokenizer Transformer. Cureus J Comput Sci 3 : es44389-026-00390-7. DOI https://doi.org/10.7759/s44389-026-00390-7

Transformer Encoder Block

Each of L=4 encoder blocks implements pre-layer-normalization multi-head self-attention (MHSA) with residual connections:

$$T' = T + \mathbf{MHSA}(\mathbf{LN}(T))$$

$$T'' = T' + \mathbf{FFN}(\mathbf{LN}(T'))$$where $\mathbf{FFN}(\cdot)$ is a two-layer feed-forward network with GELU activation and hidden dimension $d_{ffn} = 256$. We use$H = 8$ attention heads with $d_{head} = d/H = 8$ and a dropout rate of 0.1.

Classification Head

The [CLS] token output from the final encoder block is passed through

$$\hat{y} = \mathbf{Linear}(\mathbf{GELU}(\mathbf{Linear}(\mathbf{LN}(T''_{cls})))) \tag{4}$$

producing logits over $C = 6$ attack classes. The total parameter count is 206,630, compact enough for edge-gateway deployment.

Focal loss with label smoothing and class weights

To handle residual class imbalance after SMOTE-ENN and to emphasize misclassified examples, we employ focal loss with label smoothing:

$$\mathcal{L}_{focal} = -\sum_{c=1}^{C} w_c \tilde{y}_c \log p_c \cdot (1 - \hat{p}_c)^{\gamma} \tag{5}$$

where$\tilde{y}_c = (1 - \epsilon) y_c + \epsilon/(C - 1)$ is the smoothed target with $\epsilon = 0.05$,$\gamma = 2.0$ is the focusing parameter, and $\hat{p}_c$ is the predicted probability for the true class. Class weights w_c are the square root of balanced inverse-frequency weights from the post-SMOTE distribution, with additional multiplicative boosts of $\times 1.55$ and $\times 1.45$ applied to the DoS and Malformed classes, respectively,
reflecting their lower recall during initial training.

Federated learning with FedProx

We simulate $K = 5$ clients representing heterogeneous IoT gateways. Training data is distributed via a Dirichlet partition [23] with concentration parameter $\alpha = 3.0$, producing realistic non-IID local distributions; clients are reshuffled by resampling the Dirichlet split every 20 global rounds to prevent partition stagnation.

FedProx local objective

Each client k minimizes the FedProx objective [9]:

$$\min_{w_k} \mathcal{L}_{focal}(w_k) + \frac{\mu}{2} \| w_k - w_{global} \|^2 \tag{6}$$

where μ decays linearly from 0.01 to 0.002 over the first 20 rounds and is held constant thereafter. The proximal term constrains local updates to remain close to the global model, improving stability under heterogeneous data. Local optimization uses AdamW with cosine annealing over local epochs, gradient clipping at norm 1.0, and mixed-precision (FP16) computation via CUDA AMP.

Federated aggregation

After all clients complete local training, the global model is updated by weighted FedAvg:

$$w_{global} \leftarrow \sum_{k=1}^{K} \frac{n_k}{n} w_k \tag{7}$$

How to cite this article:
Kumar V, Kesswani N (November 25, 2026) Federated Learning for Internet of Things (IoT) Intrusion Detection Using a Feature Tokenizer Transformer. Cureus J Comput Sci 3 : es44389-026-00390-7. DOI https://doi.org/10.7759/s44389-026-00390-7

where n_k is the number of local samples and $n = \sum_k n_k$. Dual T4 GPUs are exploited by distributing client threads across cards using a greedy load-balancing strategy.

Five-phase learning-rate schedule

Over 300 FL rounds, local learning rates (LRs) follow the five-phase stepped schedule in Table *1*. Phases P1 and P2 use 5 and 7 local epochs, respectively, to balance exploration against gradient-step density; Phases P3-P5 use 7 local epochs throughout. Within each local training session, a cosine-annealing scheduler further reduces the LR to "0.1 times" the phase base rate. The progression from aggressive exploration (P1) to micro-fine-tuning (P5) was motivated by the observation, in an earlier 200-round experiment, that global F1 continued to increase beyond round 192, indicating incomplete convergence.

Phase	Rounds	Local LR	Local Epochs
P1	1-100	1×10^{-4}	5
P2	101-150	4×10^{-5}	7
P3	151-200	2×10^{-5}	7
P4	201-250	5×10^{-6}	7
P5	251-300	1×10^{-6}	7

TABLE 1: Five-phase federated learning-rate schedule.

LR: Learning Rate

Algorithm 1 FL-FTTransformer training

Require: Dataset D, rounds R = 300, clients K = 5, α = 3.0
Ensure: Best global model w*

1: Preprocess, scale, and SMOTE-ENN balance D
2: Initialize FT-Transformer w_global
3: P ← DirichletSplit(D_train, K, α, seed = 42)
4: $F_1^* \leftarrow 0$
5: for round r = 1 to R do
6: if r > 1 and r mod 20 = 0 then
7: P ← DirichletSplit(D_train, K, α, seed = r) ▷ Reshuffle
8: end if
9: $\eta_r \leftarrow$ GetLR(r); $E_r \leftarrow$ GetEpochs(r)
10: for client k ∈ {1, ..., K} in parallel do
11: $w_k \leftarrow$ LocalTrain(w_global, P_k, η_r, E_r) ▷ Eq. 6
12: end for
13: w_global ← Σ_k (n_k/n) w_k ▷ Weighted FedAvg
14: $F_{1r} \leftarrow$ Evaluate(w_global, D_test)
15: if $F_{1r} > F_1^*$ then

How to cite this article:
Kumar V, Kesswani N (November 25, 2026) Federated Learning for Internet of Things (IoT) Intrusion Detection Using a Feature Tokenizer Transformer. Cureus J Comput Sci 3 : es44389-026-00390-7. DOI https://doi.org/10.7759/s44389-026-00390-7

16: $F_1^* \leftarrow F_{1r}$; $w^* \leftarrow w_global$
17: end if
18: end for
19: return w^*

Experimental setup

Figure *1* illustrates the complete FL-FTTransformer system pipeline underlying the experimental setup described below, from raw MQTT traffic ingestion through federated local training, global aggregation, and final six-class detection output.

Dataset

MQTTset [1] contains 330,926 labeled MQTT flows. SMOTE-ENN balancing is applied to the full preprocessed dataset prior to splitting, expanding it to 659,084 samples; an 80/20 stratified train-test split is then applied to this balanced corpus, yielding 527,267 training and 131,817 test samples. Feature scaling is fit only on the resulting 527,267 sample training partition and applied to both partitions, preventing leakage at the normalization stage. We adopt balance-before-split, rather than balancing only the training partition post-split, for two reasons: first, it keeps the evaluation protocol directly comparable to the FL-CNN baseline of Karunamurthy et al. [5], which follows the same ordering on the same benchmark; second, SMOTE-ENN's cleaning step removes ambiguous borderline synthetic samples prior to splitting, reducing the risk of near-duplicate synthetic points being separated across the train/test boundary.

Hardware

Training runs on a dual NVIDIA Tesla T4 GPU node (2 × 15.6 GB VRAM) via Kaggle's accelerated environment, with client threads distributed across GPUs using greedy load balancing.

Hyperparameters

Key hyperparameters are summarized in Table *2*. No early stopping is applied; consistent with the comparison baseline of Karunamurthy et al. [5], the best checkpoint by macro F1 is selected using the held-out test set at each evaluation interval during training.

Hyperparameter	Value
Token dimension (d)	64
Attention heads (H)	8
FFN dimension (d_ffn)	256
Encoder layers (L)	4
Total parameters	206,630
Dropout	0.1
Optimizer	AdamW (β1 = 0.9, β2 = 0.999)
Weight decay	1.00E-04
Batch size	2048
FL rounds (R)	300

How to cite this article:
Kumar V, Kesswani N (November 25, 2026) Federated Learning for Internet of Things (IoT) Intrusion Detection Using a Feature Tokenizer Transformer. Cureus J Comput Sci 3 : es44389-026-00390-7. DOI https://doi.org/10.7759/s44389-026-00390-7

Clients (K)	5
Dirichlet α	3.0 (reshuffle every 20 rounds)
FedProx μ	0.01 → 0.002 (first 20 rounds), held constant thereafter
Focal loss γ	2
Label smoothing ε	0.05
Precision format	FP16 (CUDA AMP)

TABLE 2: FL-FTTransformer hyperparameters.

FL: Federated Learning; FFN: Feed-Forward Network; FedProx: Federated Proximal Algorithm; CUDA: Compute Unified Device Architecture; AMP: Automatic Mixed Precision.

How to cite this article:
Kumar V, Kesswani N (November 25, 2026) Federated Learning for Internet of Things (IoT) Intrusion Detection Using a Feature Tokenizer Transformer. Cureus J Comput Sci 3 : es44389-026-00390-7. DOI https://doi.org/10.7759/s44389-026-00390-7

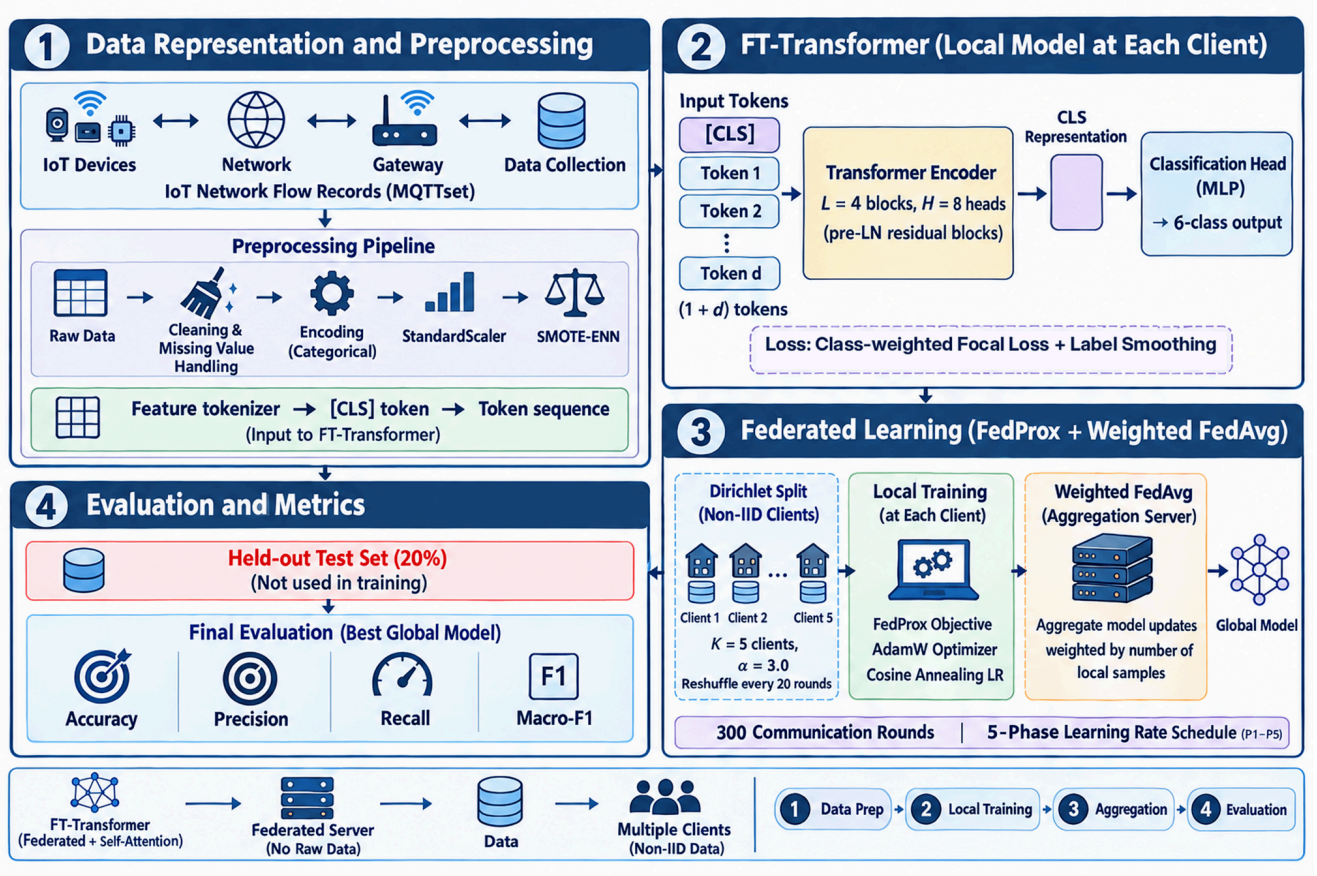

FIGURE 1: End-to-end FL-FTTransformer pipeline for federated IoT intrusion detection.

Overview of the FL-FTTransformer system across four stages: data representation and preprocessing, local FT-Transformer modeling at each client, federated learning via FedProx and weighted FedAvg, and final evaluation. Each of the K = 5 edge clients balances its local MQTT traffic with SMOTE-ENN, tokenizes features, and encodes them with a 4-block, 8-head Transformer encoder. Clients train locally using the FedProx objective with AdamW and cosine-annealing LR, combined with a class-weighted focal loss and label smoothing. The aggregation server performs weighted FedAvg to update the global model, which is broadcast back to all clients over 300 communication rounds following a 5-phase LR schedule. The final global model is evaluated on a held-out test set using accuracy, precision, recall, and macro-F1. FL: Federated Learning; IoT: Internet of Things; MQTT: Message Queuing Telemetry Transport; CLS: Classification Token; FedProx: Federated Proximal; FedAvg: Federated Averaging; SMOTE-ENN: Synthetic Minority Oversampling Technique With Edited Nearest Neighbors; F1: F1 Score.

Results And Discussion

Training dynamics

The learning rate is high in Phase 1 (rounds 1-100) and leads to an improvement in macro F1 that steadily increases from 70.08% to 90.80%. A further improvement is seen at round 124 (92.79%) in Phase 2, as the learning rate is reduced and weight updates are stabilized. A steady improvement is observed in Phase 3, reaching 94.01% at round 186. The two largest single-round gains in the whole run occur in Phase 4, after a reshuffle of the Dirichlet client partition, with 94.98% at round 230 and the global best of 95.20% at round 249, after which no further improvement is observed in Phase 5, indicating convergence, and the checkpoint at round 249 is used for final evaluation (Table *3*).

Round	Phase	Macro F1 (%)	Accuracy (%)

1	P1	70.08	70.25
34	P1	90.29	90.12
68	P1	90.75	90.75
94	P1	90.8	90.5
103	P2	91.5	91.1
117	P2	92.48	91.97
121	P2	92.78	92.24
124	P2	92.79	92.24
166	P3	92.8	92.3
178	P3	93.04	92.5
179	P3	93.07	92.56
186	P3	94.01	93.83
230	P4	94.98	95.08
249	P4	95.2	95.25

TABLE 3: Best-checkpoint improvements across FL rounds.

FL: Federated Learning.

Overall performance

There are six comparison methods listed in Table *4*. Three are taken from Karunamurthy et al. [5] which is the same source as the FL-CNN baseline. The other two models are the MAOA-AL-DBN model reported by Vijayan and Sundar [3] and the EL-IDS ensemble model from Zeghida et al. [2]. Karunamurthy et al. [5] also tested Neural Network, Random Forest, and Naïve Bayes classifiers, which were not included in Table *4* as they were not as good as the selected methods.

Method	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
MLP [5]	90.38	88.03	85.19	86.29
Gradient Boost [5]	88.42	85.75	82.64	84.16
Decision Tree [5]	91.59	90.33	86.94	88.6
MAOA-AL-DBN [3]	93.72	93.01	92.48	92.74

EL-IDS (Stacking) [2]	95.01	94.87	94.93	94.9
FL-CNN [5]	95.59	94.79	93.3	94.04
FL-FTTransformer (ours)	95.25	95.68	94.85	95.2

TABLE 4: Overall performance comparison on the MQTTset test set.

MLP: Multilayer Perceptron; DBN: Deep Belief Network; EL-IDS: Ensemble Learning-Based Intrusion Detection System; FL: Federated Learning; CNN: Convolutional Neural Network; F1: F1 Score.

The top tier is clearly defined by FL-FTTransformer and FL-CNN, which lead the centralized baselines on accuracy by roughly 1.5 to 6.8 pp, from the strongest baseline (MAOA-AL-DBN) to the weakest (Gradient Boost). FL-FTTransformer outperforms FL-CNN on three of the four metrics: macro precision (0.89 pp), macro recall (1.55 pp), and macro F1 (1.16 pp) with the latter gain primarily due to its almost perfect SlowITe recall. The only place where FL-CNN is ahead of us is in accuracy, with 95.59% compared to 95.25%, a difference of 0.34 pp.

Among centralized methods, MAOA-AL-DBN [3] reaches 93.72% accuracy and 92.74% macro F1 - a strong result - yet 2.46 pp below FL-FTTransformer on F1 while forfeiting privacy guarantees. The Decision Tree baseline achieves 91.59% accuracy but only 88.60% macro F1, reflecting the difficulty that hard-threshold classifiers face on imbalanced attacks such as Malformed and Bruteforce; MLP and Gradient Boost lag further behind at 86.29% and 84.16% macro F1. Zeghida et al. [2] report up to 95.01% accuracy and 94.90% F1 on a simplified binary formulation of MQTTset, still below our macro F1 of 95.20% on the harder six-class task.

Figure *2* presents a radar chart comparing all seven methods (MLP, Gradient Boost (GB), Decision Tree [5], MAOA-AL-DBN (MAOA) [3], EL-IDS [2], FL-CNN [5], and FL-FTTransformer (Ours)) simultaneously across all four metrics.

How to cite this article:
Kumar V, Kesswani N (November 25, 2026) Federated Learning for Internet of Things (IoT) Intrusion Detection Using a Feature Tokenizer Transformer. Cureus J Comput Sci 3 : es44389-026-00390-7. DOI https://doi.org/10.7759/s44389-026-00390-7

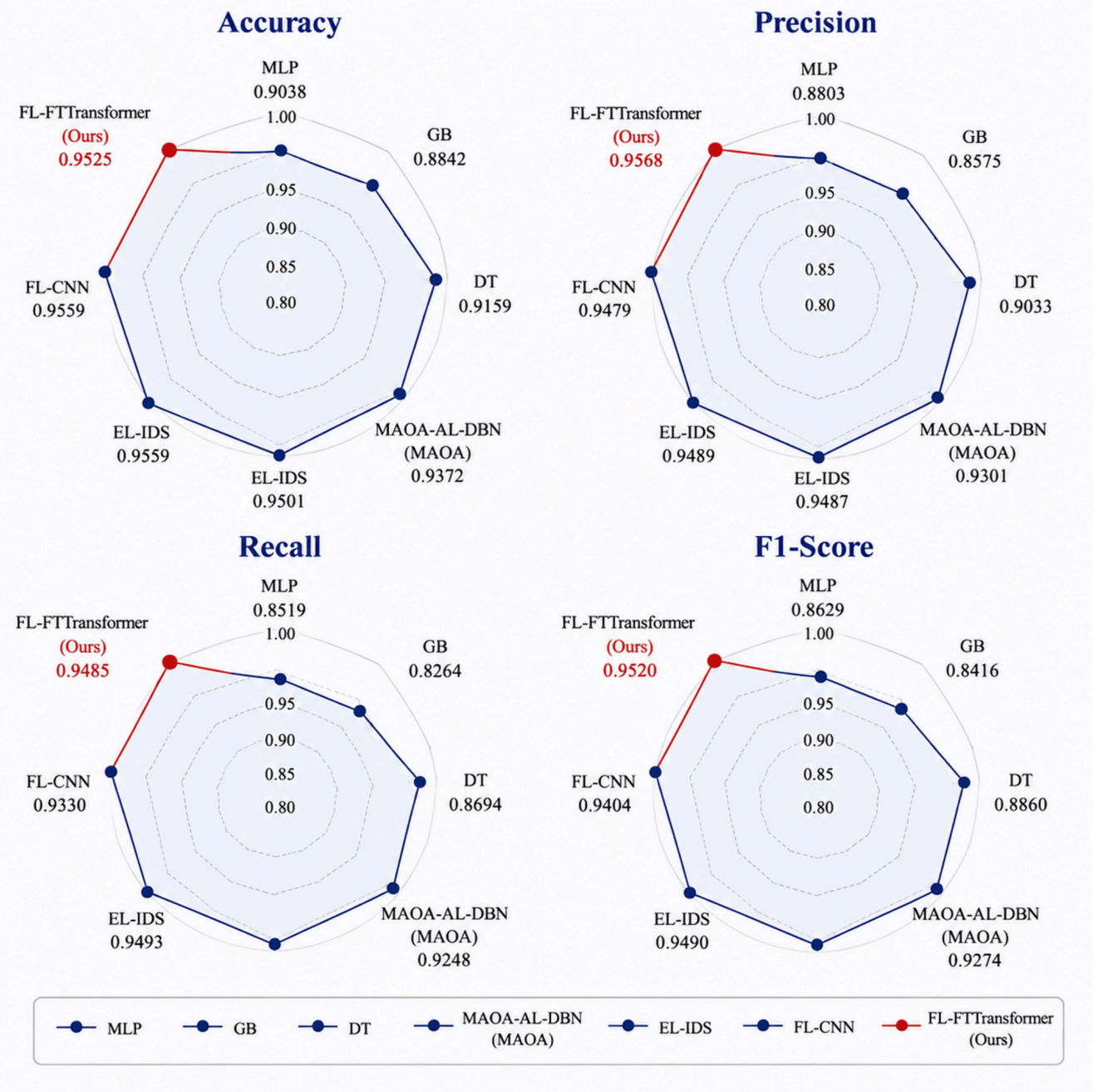

FIGURE 2: Per-metric comparison across baseline methods and FL-FTTransformer.

Radar plots comparing Accuracy, Precision, Recall, and F1-Score across MLP, GB, DT [5], MAOA-AL-DBN (MAOA) [3], EL-IDS [2], FL-CNN [5], and FL-FTTransformer (Ours). MLP: Multilayer Perceptron; GB: Gradient Boosting; DT: Decision Tree; DBN: Deep Belief Network; EL-IDS: Ensemble Learning-Based Intrusion Detection System; FL: Federated Learning; CNN: Convolutional Neural Network; F1: F1 Score.

Per-class analysis

Table *5* reports per-class precision, recall, and F1 for FL-FTTransformer and FL-CNN. FL-FTTransformer shows large improvements on DoS (+5.17 pp F1) and Flood (+7.07 pp F1) - the two high-volume attack types most likely to evade simpler models by blending with legitimate traffic bursts - and a smaller gain on Malformed (+0.98 pp F1), reflecting the per-class weight-boosting strategy. FL-CNN retains an advantage on Bruteforce (-0.84 pp), Legitimate (-2.99 pp), and

SlowITe (-2.16 pp). The Legitimate recall of FL-FTTransformer (99.95%) exceeds that of FL-CNN (99.03%); the precision gap arises because FL-FTTransformer classifies some attack samples as legitimate, whereas FL-CNN's per-class weights were tuned specifically for this corpus.

FL-FTTransformer (ours)			FL-CNN [5]			
Class	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)
Bruteforce	91.25	89.89	90.57	90.63	92.2	91.41
DoS	99.04	95.64	97.31	91.27	93.03	92.14
Flood	99.52	96.15	97.81	95.2	86.68	90.74
Legitimate	90.34	99.95	94.9	96.78	99.03	97.89
Malformed	95.71	89.96	92.75	94.88	88.86	91.77
SlowITe	98.19	97.49	97.84	100	100	100
Macro	95.68	94.85	95.2	94.79	93.3	94.04

TABLE 5: Per-class performance: FL-FTTransformer vs. FL-CNN.

Precision (P), Recall (R), and F1-score (F1) per attack class for FL-FTTransformer (Ours) and the FL-CNN baseline [5]. Macro row reports the macro-averaged metrics across all six classes. DoS: Denial of Service; FL: Federated Learning; CNN: Convolutional Neural Network.

Ablation study

To isolate the contribution of each design choice, three ablated configurations were compared on a 150-round run (for computational efficiency): (A) replace FedProx with standard FedAvg ($\mu = 0$); (B) remove SMOTE-ENN balancing (raw imbalanced data); (C) replace focal loss with standard cross-entropy. Results are shown in Table *6*.

Configuration	Macro F1 (%)	Δ vs. full (pp)
Full FL-FTTransformer (P1–P3 only)	92.79	–
(A) FedAvg instead of FedProx	91.47	-1.32
(B) No SMOTE-ENN	90.21	-2.58
(C) Cross-entropy loss	91.83	-0.96

TABLE 6: Ablation study (150-round run, macro F1 on the test set).

FL: Federated Learning; FedAvg: Federated Averaging; FedProx: Federated Proximal; SMOTE-ENN: Synthetic Minority Oversampling Technique With Edited Nearest Neighbors; F1: F1 Score; pp: Percentage Points.

How to cite this article:
Kumar V, Kesswani N (November 25, 2026) Federated Learning for Internet of Things (IoT) Intrusion Detection Using a Feature Tokenizer Transformer. Cureus J Comput Sci 3 : es44389-026-00390-7. DOI https://doi.org/10.7759/s44389-026-00390-7

SMOTE-ENN has the largest individual impact (-2.58 pp when removed), confirming that raw class imbalance severely penalizes minority attack-class recall. FedProx contributes -1.32 pp, demonstrating the importance of proximal regularization for non-IID federated convergence, and focal loss provides a further -0.96 pp over standard cross-entropy. All three components are necessary to reach the reported performance; note that this ablation uses only 150 rounds (Phases P1-P3), whereas the full 300-round model continues improving through the P4/P5 micro-fine-tuning phases, reaching a final macro F1 of 95.20%.

Comparison of feature-handling strategies

Karunamurthy et al. [5] reduce the 34 features to 17 using the ChOA before training the FL-CNN model, while Vijayan and Sundar apply a heuristic-driven optimization step (MAOA) using the MAOA-AL-DBN model. FL-FTTransformer instead operates on all 33 features, relying on multi-head self-attention to implicitly weight feature importance end-to-end. This design removes the need for a separate optimization stage, allows attention weights to adapt to the federated distribution jointly with the classifier, and preserves information that discrete selection may discard. The superior DoS and Flood F1 scores suggest that features excluded by ChOA or MAOA retain discriminative signal for these attack types. Zeghida et al. additionally balance the dataset to a binary representation, which avoids class-imbalance issues but prevents fine-grained multi-class discrimination; our SMOTE-ENN approach retains all six attack categories while addressing imbalance.

Discussion

Interpretation of Results

The results suggest that the attention-based tabular modeling approach may be more appropriate for the feature structure of MQTT network traffic than vision or sequence modeling-based architectures. The per-class gains for the two most common attack types in MQTTset, DoS and Flood, are consistent with the hypothesis that end-to-end learned feature weighting preserves discriminative signal that discrete, heuristic feature-selection stages like ChOA or MAOA might lose, although this has not been directly tested against an ablation that removes attention-based weighting and replaces it with explicit feature selection.

The ablation results support a layered interpretation of the pipeline's performance: SMOTE-ENN provides the largest single contribution because it directly targets the root cause of poor minority-class recall, while FedProx and focal loss provide complementary, smaller refinements addressing non-IID federated convergence and residual example-level imbalance, respectively. This layering implies that future extensions such as differential privacy or federated unlearning should be evaluated against the full pipeline rather than a single component in isolation, since individual components are not substitutable for one another. Cybersecurity frameworks increasingly require complementary approaches to detection, data protection, and system integrity. Blockchain-based security mechanisms offer another perspective on protecting distributed digital environments [24]. Their potential integration with federated intrusion detection warrants investigation, particularly regarding overhead and practical deployment constraints.

Implications

The pipeline is simpler to deploy than pipelines that involve an optimization step before training, since the model is trained on all 33 raw features. This also contrasts with federated pipelines that still add a feature-reduction stage, such as chi-square selection with a GRU classifier [25]. The ablation also indicates that, when IoT traffic is imbalanced, tuning the federated optimizer or the loss function should be done after the data distribution is corrected. Furthermore, the method suits environments where raw network logs cannot be freely shared, as only model updates are exchanged. Lastly, the layered contribution of the components suggests that future extensions, such as differential privacy or federated unlearning, should be assessed against the entire pipeline, since the components cannot be substituted for one another.

Limitations

A limitation of the present study is that FL-CNN retains a narrow advantage on accuracy and on the Legitimate and SlowITe classes, indicating that no single architecture dominates on every metric; ensembling or per-class calibration may close this remaining gap. Checkpoint selection followed the protocol of the baseline study [5] and used the held-out test set; evaluating the model with a separate validation split is a natural direction for future work. The study is also limited to a single benchmark (MQTTset) and a simulated, rather than physically distributed, federated deployment.

Future Research Directions

Future research should validate the model with a separate validation split, and test it on other IoT datasets and physical testbeds, with more clients and more non-IID partitions (lower α). Because tokenization strategy strongly affects Transformer-based IDS performance on tabular data [26], cross-dataset evaluation should also check whether feature-wise tokenization transfers beyond MQTTset. Ensembling or per-class calibration can fill the remaining accuracy gap with FL-CNN on accuracy and on the Legitimate and SlowITe classes. A direct ablation that removes attention-based weighting and replaces it with explicit feature selection (ChOA or MAOA) would test the feature-weighting hypothesis. Evaluating differential privacy and federated unlearning should be done on the entire pipeline, and the combination of blockchain-based mechanisms with federated intrusion detection (discussed above) should be analyzed for overhead and deployment considerations.

Conclusions

We presented FL-FTTransformer, a federated intrusion detection framework that integrates a Feature Tokenizer Transformer backbone with FedProx aggregation, SMOTE-ENN class balancing, focal loss with per-class weight boosting, and a five-phase learning-rate schedule over 300 communication rounds. Experiments on the MQTTset benchmark demonstrate 95.25% accuracy and 95.20% macro F1, surpassing the state-of-the-art FL-CNN baseline by 1.16 pp on the most informative macro metric and outperforming the centralized MAOA-AL-DBN approach by 2.46 pp on macro F1. The model achieves substantially better detection of high-risk DoS (F1: 97.31% vs. 92.14%) and Flood (F1: 97.81% vs. 90.74%) attacks without any handcrafted feature selection, and, unlike the binary ensemble of Zeghida et al., distinguishes all six traffic classes simultaneously, providing actionable per-attack-type information for network operators.

Future work includes: (i) applying differential privacy to the federated gradient updates; (ii) extending to cross-silo settings with asynchronous aggregation; (iii) pruning the FT-Transformer to sub-100K parameters for deployment on microcontroller-class gateways; and (iv) evaluating on additional IoT benchmarks such as N-BaIoT and Edge-IIoT to assess generalizability.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Nishtha Kesswani, Vikash Kumar

Drafting of the manuscript: Nishtha Kesswani, Vikash Kumar

Critical review of the manuscript for important intellectual content: Nishtha Kesswani, Vikash Kumar

Supervision: Nishtha Kesswani

Acquisition, analysis, or interpretation of data: Vikash Kumar

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue.

Animal subjects: All authors have confirmed that this study did not involve animal subjects or tissue.

How to cite this article:
Kumar V, Kesswani N (November 25, 2026) Federated Learning for Internet of Things (IoT) Intrusion Detection Using a Feature Tokenizer Transformer. Cureus J Comput Sci 3 : es44389-026-00390-7. DOI https://doi.org/10.7759/s44389-026-00390-7

Conflicts of interest: In compliance with the ICMJE uniform disclosure form, all authors declare the following:

- **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work.
- **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work.
- **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Data Availability Statements

The data (and/or code) supporting this study are openly available in Kaggle at https://www.kaggle.com/datasets/cnrieiit/mqttset, reference cnrieiit/mqttset.

Acknowledgements

AI tool usage: Figure 1 (the end-to-end FL-FTTransformer system pipeline diagram) was created using ChatGPT. The per-metric radar chart comparison (Figure 2) was created using Python's Matplotlib library and plotted directly from the experimental results.

References

1. Vaccari I, Chiola G, Aiello M, Mongelli M, Cambiaso E: MQTTset, a new dataset for machine learning techniques on MQTT. Sensors. 2020, 20:6578. 10.3390/s20226578
2. Zeghida H, Boulaiche M, Chikh R: Securing MQTT protocol for IoT environment using IDS based on ensemble learning. International Journal of Information Security. 2023, 22:1075-1086. 10.1007/s10207-023-00681-3
3. Vijayan PM, Sundar S: An automated system of intrusion detection by IoT-aided MQTT using improved heuristic-aided autoencoder and LSTM-based deep belief network. PLOS ONE. 2023, 18:e0291872. 10.1371/journal.pone.0291872
4. McMahan B, Moore E, Ramage D, Hampson S, y Arcas BA: Communication-efficient learning of deep networks from decentralized data. Proceedings of the 20 th International Conference on Artificial Intelligence and Statistics (AISTATS). 2017, 54:1273-1282. 10.48550/arXiv.1602.05629
5. Karunamurthy A, Vijayan K, Kshirsagar PR, Tan KT: An optimal federated learning-based intrusion detection for IoT environment. Scientific Reports. 2025, 15:8696. 10.1038/s41598-025-93501-8
6. Rashid MM, Khan SU, Eusufzai F, Redwan MA, Sabuj SR, Elsharief M: A federated learning-based approach for improving intrusion detection in industrial internet of things networks. Network. 2023, 3:158-179. 10.3390/network3010008
7. Hamdi N: Federated learning-based intrusion detection system for internet of things. International Journal of Information Security. 2023, 22:1937-1948. 10.1007/s10207-023-00727-6
8. Gorishniy Y, Rubachev I, Khrulkov V, Babenko A: Revisiting deep learning models for tabular data. Proceedings of the Neural Information Processing Systems. 2021, arXiv:2106.11959. 10.48550/arXiv.2106.11959
9. Li T, Sahu AK, Zaheer M, Sanjabi M, Talwalkar A, Smith V: Federated optimization in heterogeneous networks. Proceedings of the 3rd MLSys Conference, Austin, TX, USA. 2020, 2:1-22. 10.48550/arXiv.1812.06127
10. Rahman SA, Tout H, Talhi C, Mourad A: Internet of things intrusion detection: centralized, on-device, or federated learning?. IEEE Network. 2020, 34:310-317. 10.1109/MNET.011.2000286
11. Mothukuri V, Khare P, Parizi RM, Pouriyeh S, Dehghantanha A, Srivastava G: Federated-learning-based anomaly detection for IoT security attacks. IEEE Internet of Things Journal. 2022, 9:2545-2554. 10.1109/JIOT.2021.3077803
12. Janati Idrissi M, Alami H, El Mahdaouy A, El Mekki A, Oualil S, Yartaoui Z, Berrada I: Fed-ANIDS: Federated learning for anomaly-based network intrusion detection systems. Expert Systems with Applications. 2023, 234:121000. 10.1016/j.eswa.2023.121000

13. Zhao R, Yin Y, Shi Y, Xue Z: Intelligent intrusion detection based on federated learning aided long short-term memory. Physical Communication. 2020, 42:101157. 10.1016/j.phycom.2020.101157

14. Zhang J, Luo C, Carpenter M, Min G: Federated learning for distributed IIoT intrusion detection using transfer approaches. IEEE Transactions on Industrial Informatics. 2023, 19:8159-8169. 10.1109/TII.2022.3216575

15. Tavallaee M, Bagheri E, Lu W, Ghorbani AA: A detailed analysis of the KDD CUP 99 data set. IEEE Symposium on Computational Intelligence for Security and Defense Applications, Ottawa, ON, Canada. 2009, 1-6. 10.1109/CISDA.2009.5356528

16. Sharafaldin I, Lashkari AH, Ghorbani AA: Toward generating a new intrusion detection dataset and intrusion traffic characterization. Proceedings of the 4th International Conference on Information Systems Security and Privacy ICISSP. 2018, 108-116. 10.5220/0006639801080116

17. Arik SO, Pfister T: TabNet: attentive interpretable tabular learning. Proceedings of the AAAI Conference on Artificial Intelligence. 2021, 35:6679-6687. 10.1609/aaai.v35i8.16826

18. Abdelkhalek A, Mashaly M: Addressing the class imbalance problem in network intrusion detection systems using data resampling and deep learning. The Journal of Supercomputing. 2023, 79:10611-10644. 10.1007/s11227-023-05073-x

19. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP: SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research. 2002, 16:321-357. 10.1613/jair.953

20. Ding H, Sun Y, Huang N, Shen Z, Cui X: TMG-GAN: generative adversarial networks-based imbalanced learning for network intrusion detection. IEEE Transactions on Information Forensics and Security. 2024, 19:1156-1167. 10.1109/TIFS.2023.3331240

21. Lin TY, Goyal P, Girshick R, He K, Dollar P: Focal loss for dense object detection. IEEE International Conference on Computer Vision (ICCV), Venice, Italy. 2017, 2980-2988. 10.1109/ICCV.2017.324

22. Mulyanto M, Faisal M, Prakosa SW, Leu JS: Effectiveness of focal loss for minority classification in network intrusion detection systems. Symmetry. 2021, 13:4. 10.3390/sym13010004

23. Hsu TMH, Qi H, Brown M: Federated visual classification with real-world data distribution. Computer Vision - ECCV 2020. 2020, 12355:76-92. 10.1007/978-3-030-58607-2_5

24. Chaubey NK, Chaubey N: Advanced Cyber Security Techniques for Data, Blockchain, IoT, and Network Protection. IGI Global Scientific Publishing, Hershey, PA, USA; 2025. 10.4018/979-8-3693-9225-6

25. Nguyen TD, Alazab A, Khraisat A, Jan T: Feature reduction in federated learning for intrusion detection in IoT networks. Cybersecurity. 2026, 9:102. 10.1186/s42400-025-00509-8

26. Aksholak G, Bedelbayev A, Magazov R, Kaplan K: Transformer tokenization strategies for network intrusion detection: addressing class imbalance through architecture optimization. Computers. 2026, 15:75. 10.3390/computers15020075

How to cite this article:
Kumar V, Kesswani N (November 25, 2026) Federated Learning for Internet of Things (IoT) Intrusion Detection Using a Feature Tokenizer Transformer. Cureus J Comput Sci 3 : es44389-026-00390-7. DOI https://doi.org/10.7759/s44389-026-00390-7