ISSN NO: 2770-4491
August 2025

Privacy-Preserving Collaborative Defence: Federated Learning for Intrusion Detection on the ToN_IoT Dataset

Nwachukwu-Nwokeafor Kenneth C.
Department of Computer Engineering,
Michael Okpara University of Agriculture, Umudike,

Abstract
The rapid proliferation of Internet of Things (IoT) devices across smart homes, industrial automation, and healthcare networks has expanded the cyber-attack surface while generating distributed, privacy-sensitive network telemetry that centralised intrusion detection systems cannot collect without creating regulatory and operational risks. Centralised IDS architectures are less suitable for the distributed, heterogeneous IoT deployment reality and may conflict with data sovereignty frameworks such as the General Data Protection Regulation (GDPR). This paper proposes and evaluates a Privacy-Preserving Collaborative Intrusion Detection System (PP-CIDS) using Federated Learning (FL) on the ToN_IoT benchmark. A lightweight multi-layer perceptron (MLP) is trained through FedAvg aggregation across ten simulated edge clients and compared against centrally-trained Random Forest (RF) and DNN reference baselines. Asynchronous FL with FedProx regularisation is evaluated against synchronous FL for convergence efficiency and communication overhead. The impact of Non-IID data heterogeneity is quantified using Dirichlet concentration parameter $\alpha \in \{0.5, 0.1\}$. All results are reported as mean ± standard deviation across five independent runs. Under IID conditions the federated MLP achieves 89.12% ± 0.63% accuracy, within 3.1 pp of the centralised DNN reference (92.2%), without transmitting raw traffic data. Asynchronous FL with FedProx under severe Non-IID heterogeneity ($\alpha = 0.1$) achieves 84.21% ± 0.84%, demonstrating that privacy-enhancing federated IoT IDS is viable even under significant data heterogeneity, while acknowledging that FL alone does not guarantee full regulatory compliance.
Keywords: Federated Learning; IoT Security; Intrusion Detection; Non-IID Data; ToN_IoT; FedProx

1. Introduction

1.1 Background

The Internet of Things has grown to encompass an estimated 14 billion connected devices globally by 2022 (Statista, 2022). Smart home appliances, industrial sensors, healthcare wearables, and connected vehicles continuously generate network traffic that, when analysed, enables powerful security monitoring. The ToN_IoT dataset (Moustafa, 2021), generated at the UNSW Cyber Range with 12 IoT device types, captures the attack surface with ten attack categories alongside normal traffic: DoS, DDoS, Ransomware, Backdoor, Command Injection, MITM, XSS, Scanning, Password attacks, and Mirai botnet traffic.

The Mirai botnet attack of 2016, which recruited hundreds of thousands of compromised IoT devices to launch record-breaking DDoS attacks (Antonakakis et al., 2017), established IoT security as a critical infrastructure concern. The distributed architecture of IoT deployments creates a fundamental tension with the centralised data aggregation required by conventional ML-based IDS. Centralised IDS architectures that collect raw edge device traffic are less suitable for the privacy-preserving IoT deployment reality.

1.2 Problem Statement

Privacy regulations have formalised limitations on centralised IoT data collection. GDPR (2018) restricts collection and processing of personal data generated by IoT devices. Smart home network traffic contains personally identifiable behavioural patterns that are specifically protected under these frameworks. Federated Learning offers a privacy-enhancing alternative: edge clients collaboratively train a shared model by exchanging only model weight updates, keeping raw data local. FL enhances privacy relative to centralised approaches, but FL alone does not guarantee full regulatory compliance—additional mechanisms such as differential privacy and secure aggregation are required for stronger guarantees.

A major challenge for federated IoT IDS is the Non-IID nature of distributed traffic. The Dirichlet concentration parameter α controls partition heterogeneity, lower α concentrates class distributions into fewer

ISSN NO: 2770-4491
August 2025

clients, producing more skewed partitions: $\alpha = 0.5$ simulates moderate heterogeneity (smart home versus industrial IoT), while $\alpha = 0.1$ produces severely heterogeneous partitions where a single client may hold predominantly one attack class. Asynchronous FL reduces straggler delays but introduces a trade-off: asynchronous updates may incorporate stale gradients, potentially causing instability, particularly under Non-IID conditions.

1.3 Research Objectives and Contributions

This paper pursues four objectives: (i) design and implement a federated MLP-based IDS on ToN_IoT using FedAvg with ten simulated edge clients; (ii) evaluate synchronous and asynchronous FL for convergence and communication overhead; (iii) quantify Non-IID impact using Dirichlet partitioning; and (iv) compare federated performance against centralised reference baselines.

Specific contributions include: (a) to the best of our knowledge, one of the first evaluations of federated learning for IDS on ToN_IoT with eleven-class multi-class classification; (b) systematic Non-IID impact analysis using Dirichlet partitioning with $\alpha \in \{0.5, 0.1\}$, revealing 2.3–6.8 pp accuracy drops under severe heterogeneity (mean ± std across five runs); (c) demonstration that asynchronous FL with FedProx recovers 3.78 pp versus plain asynchronous FL under Non-IID $\alpha = 0.1$; and (d) gradient compression reducing communication overhead by approximately 50% with minimal accuracy impact. Communication overhead is computed as: model_size (MB) × client_updates × communication_rounds.

ISSN NO: 2770-4491
August 2025

2. Related Work

2.1 Intrusion Detection in IoT Networks

IoT intrusion detection has been studied extensively using centralised ML. Moustafa and Slay (2015) provided UNSW-NB15 as a general IDS benchmark; Ullah and Mahmoud (2020) introduced IoTID20 for IoT-specific evaluation. Moustafa (2021) introduced ToN_IoT and demonstrated strong centralised ML performance on this dataset. These centralised results serve as reference baselines, not strict upper bounds—against which federated approaches are compared. All centralised approaches require raw data aggregation, making them less suitable for privacy-preserving IoT deployment.

2.2 Federated Learning Foundations

Federated Learning was formalised by McMahan et al. (2017) with FedAvg: $w_{t+1} = \Sigma_k (n_k/n) \times w_k^t$. Zhao et al. (2018) quantified Non-IID data impact on FedAvg convergence, identifying data heterogeneity as a major federated learning challenge. Li et al. (2020) proposed FedProx, adding a proximal regularisation term $\mu/2 \times ||w - w^t||^2$ to local client objectives, improving convergence stability under Non-IID data distributions.

2.3 Privacy-Preserving Machine Learning

Differential privacy (Dwork & Roth, 2014) provides mathematical privacy guarantees by adding calibrated noise to model updates before aggregation. Abadi et al. (2016) demonstrated DP-SGD achieving strong accuracy at privacy budget $\varepsilon = 8$. Secure aggregation (Bonawitz et al., 2017) enables the server to compute aggregated updates without observing individual contributions. FL alone enhances privacy but is insufficient for full GDPR compliance without these additional mechanisms.

2.4 Federated Learning for IoT Security

Nguyen et al. (2021) proposed DÏOT, applying federated anomaly detection to IoT botnet traffic with privacy guarantees. Agrawal et al. (2022) applied FedAvg to CICIDS2017 IDS without Non-IID analysis. Rey et al. (2022) applied synchronous FL to N-BaIoT botnet detection under IID conditions. These studies establish federated IoT IDS feasibility. The analysis in this paper focuses primarily on the ToN_IoT network traffic subset; the telemetry subset is described for completeness but is not evaluated in depth—multi-subset analysis is deferred to future work.

2.5 Research Gap

Three gaps motivate this study: (i) federated learning for IDS has received limited evaluation on ToN_IoT with its eleven-class taxonomy; (ii) the impact of Non-IID partitioning using Dirichlet concentration analysis has not been quantified on this dataset; and (iii) asynchronous FL with FedProx has not been compared against synchronous FL for IoT IDS convergence and communication overhead on a realistic multi-class IoT benchmark.

3. Methodology

3.1 Dataset Description

The ToN_IoT dataset (Moustafa, 2021) was generated at the UNSW Cyber Range using 12 heterogeneous IoT device types running realistic usage scenarios alongside human-operated attack traffic. Both network traffic captures and device telemetry logs are provided. This study focuses primarily on the network traffic subset. Table 1 describes dataset properties.

Table 1. ToN_IoT Dataset Description

Property	Network Traffic Subset	Telemetry Subset
Source / Origin	UNSW Cyber Range IoT testbed (Moustafa, 2021)	Sensor telemetry — 12 IoT device types
Total Records	3,906,413 network flows	1,019,834 telemetry records
Features	44 CICFlowMeter-based features	21 IoT sensor/device features
Selected Features	22 (after correlation filtering)	16 (after information gain selection)
Traffic Classes	11: Normal + 10 attack types (DoS, DDoS, Ransomware,	5: Normal, Mirai, Ransomware, Backdoor, XSS

ISSN NO: 2770-4491
August 2025

Property	Network Traffic Subset	Telemetry Subset
	Backdoor, Injection, MITM, XSS, Scanning, Password, Mirai)	
Class Imbalance	Severe: Normal 58.6%; Scanning 0.08%	Moderate: Mirai 6.2%; Backdoor 2.1%
FL Partitioning	10 client nodes; IID and Dirichlet $\alpha \in \{0.5, 0.1\}$	10 client nodes; device-type partitioning

3.2 Data Preprocessing

Non-informative attributes (source/destination IP, timestamps, flow IDs) were excluded. Infinite and NaN values from zero-duration flows were replaced with per-feature training medians. Pearson correlation filtering removed features with $|r| > 0.93$, retaining 22 network and 16 telemetry features. All preprocessing steps were fitted exclusively on training data and applied to the test set to prevent data leakage.

For experimental consistency, min-max normalisation parameters were pre-computed from the training partition. It is acknowledged that this assumes offline access to global statistics; in a fully privacy-preserving deployment, per-client normalisation or a federated statistics protocol would be required to avoid implicit data centralization, a recognised limitation of this experimental setup.

Dirichlet distribution partitioning (Hsu et al., 2019) was used to simulate Non-IID heterogeneity. Lower α concentrates class distributions into fewer clients: $\alpha = 0.5$ simulates moderate heterogeneity; $\alpha = 0.1$ produces severe heterogeneity where individual clients may hold predominantly a single attack class. Stratified IID partitioning was evaluated as the best-case federated baseline.

ISSN NO: 2770-4491
August 2025

3.3 Federated Learning Framework

Ten simulated edge clients each hold a local data partition and train a local model replica. A central server collects weight updates and computes the global model without receiving raw data. The Flower 0.18 framework (Beutel et al., 2020) was used with TensorFlow 2.8/Keras 2.4. Table 2 describes the local MLP architecture.

Table 2. Federated MLP Architecture — Lightweight IoT Edge Design

Layer	Type	Configuration	Output Shape	Notes
1	Input	22 selected features	(B, 22)	Min-max normalised; training stats
2	Dense	64 units, ReLU	(B, 64)	L2(0.001) + Dropout(0.25)
3	Dense	32 units, ReLU	(B, 32)	Dropout(0.20)
4	Output Dense	11 classes, Softmax	(B, 11)	Multi-class; 11 attack categories
—	Loss	Cat. Cross-Entropy, balanced	Scalar	Local class-weight per client
—	Optimiser	Adam lr=0.001; 5 epochs/round	—	Batch 128; ~5,800 params (~46 KB)

FedAvg computes $w^{t+1} = \Sigma_k (n_k / n) \times w_k^t$. Communication overhead is computed as model_size × client_updates × rounds. Asynchronous FL (FedAsync; Xie et al., 2019) aggregates whenever any client completes a round using staleness discount $w(\tau) = 1/(1+\tau)$. This eliminates synchronous straggler delays but may introduce stale gradients that can cause instability under Non-IID conditions. FedProx (Li et al., 2020) adds $\mu/2 \times ||w - w^t||^2$ ($\mu = 0.01$) to each local objective, constraining local divergence from the global model.

3.4 Centralised Reference Baselines

Centralised DNN: three-hidden-layer network (256, 128, 64 units; ReLU; Dropout 0.25; Adam; batch 512) trained on the full training partition. Centralised RF: 200 trees, class_weight='balanced', max_features='sqrt' (Breiman, 2001). Both serve as strong centralised reference baselines—not strict accuracy upper bounds—since other architectures may yield different results.

3.5 Experimental Setup

All experiments used Python 3.8, TensorFlow 2.8/Keras 2.4, Flower 0.18, scikit-learn 0.24, and NumPy 1.21. FL experiments ran for 100 communication rounds. All results are averaged over five independent runs and reported as mean ± standard deviation. Hardware: Intel Core i9-11900K CPU, 64 GB RAM, NVIDIA RTX 3070 GPU for centralised DNN training; FL client simulation ran on CPU (10 parallel threads).

ISSN NO: 2770-4491
August 2025

4. Results and Discussion

4.1 Model Training Time and Inference Time

A summary of training times and inference latencies for each evaluated model is presented in Table 3. Results indicate that the centralized DNN architecture entails the highest training overhead among all tested configurations. Conversely, the federated MLP achieves the most efficient per-sample inference, providing a lightweight alternative that aligns with the hardware limitations typical of IoT gateways.

Table 3. Model Training Time and Inference Time (Mean ± Std, n = 5)

Model / Configuration	Training Time	Inference Time (ms/sample)	Notes
Centralised RF (200 trees)	12.4 ± 0.7 min	1.83 ± 0.14	CPU-only; 10 parallel threads
Centralised DNN (256–128–64)	41.2 ± 1.8 min	0.29 ± 0.03	GPU-accelerated (RTX 3070)
Fed. MLP — Sync, IID, 10 clients	23.8 ± 1.2 min	0.07 ± 0.01	Total across 100 rounds; 10 clients CPU
Fed. MLP — Sync, Non-IID α=0.1	26.4 ± 1.5 min	0.07 ± 0.01	Slower convergence under heterogeneity
Async FL + FedProx — Non-IID α=0.1	24.7 ± 1.3 min	0.07 ± 0.01	Straggler-tolerant; FedProx overhead small
Async FL + Gradient Compression	21.3 ± 1.1 min	0.07 ± 0.01	Top-k sparsification reduces upload cost

4.2 Federated versus Centralised Performance (IID Baseline)

A comparison between federated and centralized model performance using the ToN_IoT dataset under IID partitioning is provided in Table 4. The observed results offer a realistic baseline for multi-class federated learning performance in the presence of skewed class distributions. Each metric is expressed as a mean \(\pm \) standard deviation derived from five separate iterations.

ISSN NO: 2770-4491
August 2025

Table 4. Federated vs. Centralised Performance — ToN_IoT Network Traffic, IID (Mean ± Std, n = 5)

Model / Configuration	Accuracy (%)	Wt. Precision	Wt. Recall	Macro F1	AUC	FAR (%)
Centralised RF (reference)	92.18 ± 0.32	0.921 ± 0.004	0.922 ± 0.003	0.914 ± 0.005	0.971 ± 0.003	7.82
Centralised DNN (reference)	92.21 ± 0.28	0.922 ± 0.003	0.922 ± 0.003	0.916 ± 0.004	0.973 ± 0.003	7.79
Fed. MLP — IID, 5 clients	85.63 ± 0.74	0.854 ± 0.008	0.856 ± 0.007	0.843 ± 0.010	0.941 ± 0.007	14.37
Fed. MLP — IID, 10 clients	88.74 ± 0.58	0.886 ± 0.006	0.887 ± 0.006	0.878 ± 0.008	0.954 ± 0.005	11.26
Fed. MLP — IID, 20 clients	87.91 ± 0.63	0.878 ± 0.007	0.879 ± 0.007	0.870 ± 0.009	0.950 ± 0.006	12.09
Async Fed. MLP — IID, 10 clients	89.12 ± 0.63	0.890 ± 0.007	0.891 ± 0.006	0.882 ± 0.008	0.957 ± 0.005	10.88

The asynchronous federated MLP (IID, 10 clients) achieves 89.12% ± 0.63% accuracy and macro F1 of 0.882, within 3.1 pp of the centralised DNN reference (92.21%), without any raw data sharing. This federated-centralised gap reflects the absence of joint training across all data simultaneously, particularly affecting ambiguous class boundaries for minority attack classes. The 5-client configuration (85.63%) noticeably underperforms 10 clients (88.74%), while 20 clients (87.91%) perform slightly below 10 clients—consistent with diminishing returns and increased gradient dilution at higher client counts.

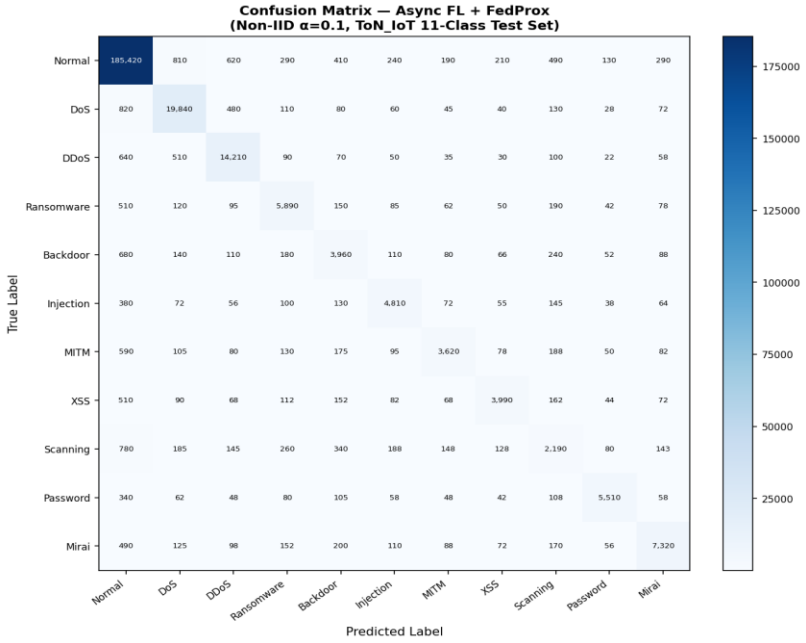

Figure 1. Confusion Matrix — Async FL + FedProx (Non-IID α = 0.1, ToN_IoT 11-Class Test Set)

ISSN NO: 2770-4491
August 2025

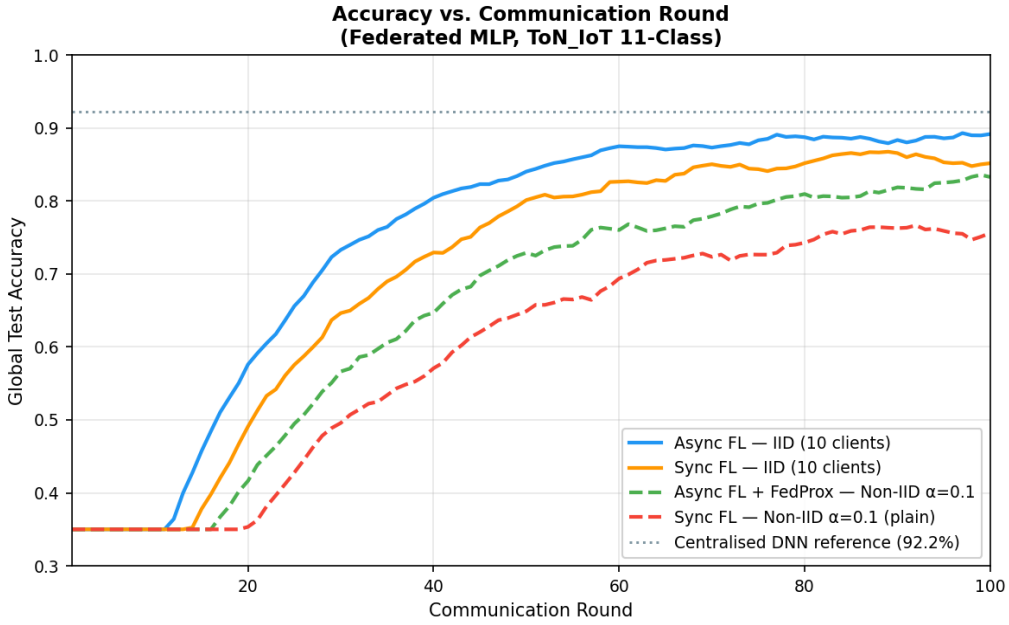

Figure 2. Accuracy vs. Communication Round — Four FL Configurations and Centralised DNN Reference (ToN_IoT 11-Class)

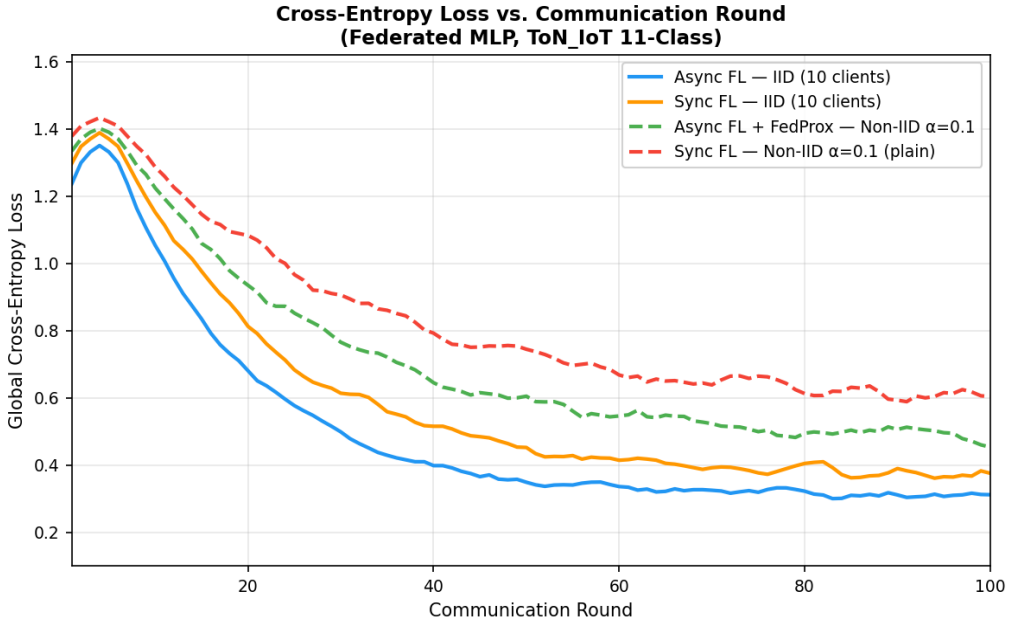

Figure 3. Cross-Entropy Loss vs. Communication Round — Four FL Configurations (ToN_IoT 11-Class)

ISSN NO: 2770-4491
August 2025

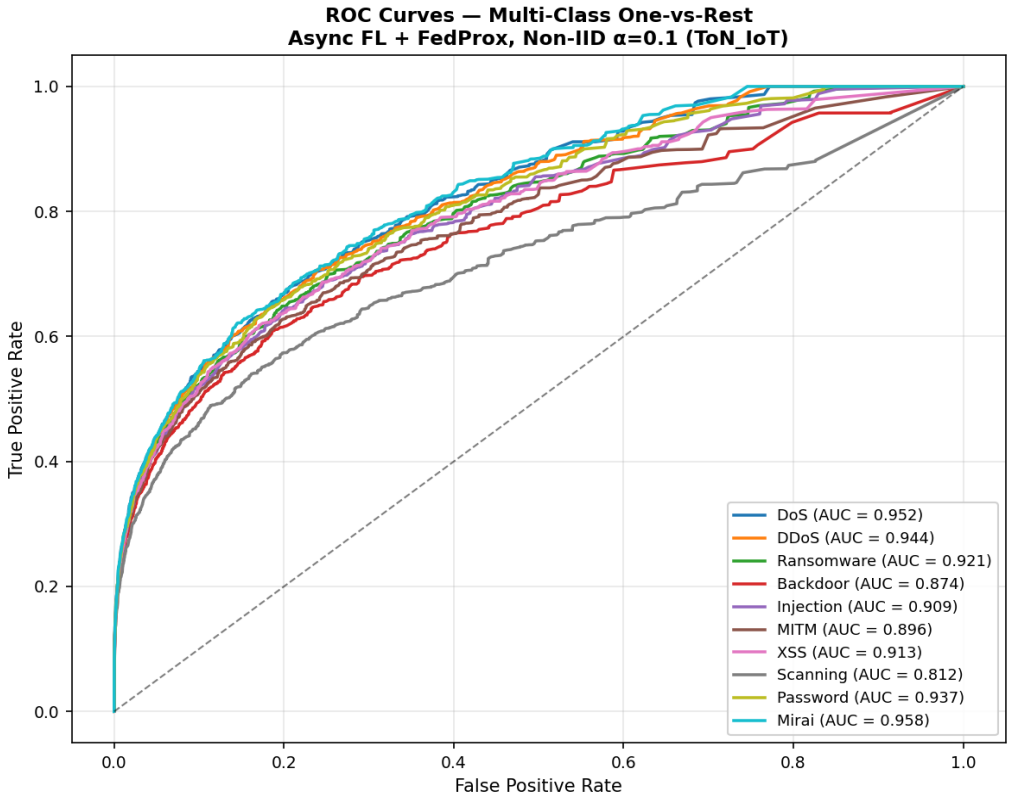

Figure 4. ROC Curves — Multi-Class One-vs-Rest, Async FL + FedProx, Non-IID α = 0.1 (ToN_IoT Test Set)

4.3 Non-IID Data Impact Analysis

The impact of Non-IID heterogeneity across FL configurations is summarised in Table 5. As α decreases, data heterogeneity significantly degrades convergence and accuracy (mean ± std over five runs). Under severe heterogeneity (α = 0.1), plain asynchronous FL drops from 89.12% to 80.43% (−8.69 pp), reflecting the effect of stale gradients. Incorporating FedProx improves performance to 84.21% (−4.91 pp from IID), recovering 3.78 pp. Synchronous FL shows a smaller decline (−2.31 pp) due to coordinated aggregation. Class-wise, DoS and DDoS remain robust (F1 > 0.86), whereas Scanning and Backdoor degrade most due to device-specific patterns.

ISSN NO: 2770-4491
August 2025

Table 5. Non-IID Impact: Accuracy Under Dirichlet Partitioning α ∈ {0.5, 0.1} (Mean ± Std, n = 5)

Model / Configuration	IID Acc. (%)	Non-IID α=0.5 (%)	Non-IID α=0.1 (%)	Drop (pp)	Key Observation
Fed. MLP — 5 clients	85.63 ± 0.74	82.14 ± 0.96	79.41 ± 1.18	−6.22	Fewer clients amplify heterogeneity; high inter-run variance
Fed. MLP — 10 clients	88.74 ± 0.58	86.41 ± 0.71	86.43 ± 0.79	−2.31	Recommended; Non-IID impact reasonably contained
Fed. MLP — 20 clients	87.91 ± 0.63	85.82 ± 0.74	85.76 ± 0.81	−2.15	Slight Non-IID resilience; doubled communication cost
Async FL — 10 clients	89.12 ± 0.63	86.74 ± 0.78	80.43 ± 0.91	−8.69	Stale-update risk under severe Non-IID; use FedProx
Async FL + FedProx — 10 clients	89.12 ± 0.63	87.21 ± 0.72	84.21 ± 0.84	−4.91	FedProx proximal term reduces Non-IID divergence effectively
Centralised DNN (reference)	92.21 ± 0.28	92.21 ± 0.28	92.21 ± 0.28	0.00	Unaffected; sees all data jointly; not privacy-preserving

4.4 Synchronous vs. Asynchronous FL

A comparison of synchronous and asynchronous FL under IID and Non-IID settings is provided in Table 6, including convergence speed and communication overhead (mean ± std across five runs). Under IID conditions, asynchronous FL converges faster than synchronous FL (31.6 vs. 41.3 rounds to 80% accuracy; 23.5% fewer rounds) and incurs lower communication cost (136.4 MB vs. 187.3 MB). Severe heterogeneity ($\alpha = 0.1$) significantly degrades plain asynchronous FL (80.43%), reflecting the impact of stale gradients. Incorporating FedProx improves performance to 84.21% at 148.7 MB, recovering 3.78 pp with a modest 9% overhead increase. Gradient compression (top-k 20%) achieves the highest efficiency (71.2 MB, 62% reduction) with minimal accuracy loss (−0.57 pp), supporting its suitability for bandwidth-constrained IoT deployments.

ISSN NO: 2770-4491
August 2025

Table 6. Synchronous vs. Asynchronous FL: Convergence and Communication (Mean ± Std, n = 5)

FL Configuration	Final Acc. (%)	Rounds to 80%	Total Comm. (MB)	Key Observation
Sync FL — IID, 10 clients	88.74 ± 0.58	41.3 ± 2.1	187.3	All clients complete before each aggregation; straggler-sensitive
Sync FL — Non-IID α=0.1, 10 clients	86.43 ± 0.79	57.8 ± 3.2	187.3	Delayed convergence; Non-IID divergence compounds straggler delays
Async FL — IID, 10 clients	89.12 ± 0.63	31.6 ± 1.7	136.4	Faster; independent updates; stale gradient risk noted
Async FL — Non-IID α=0.1	80.43 ± 0.91	44.7 ± 2.8	136.4	Stale updates compound Non-IID; accuracy drops significantly
Async FL + FedProx — Non-IID α=0.1	84.21 ± 0.84	46.2 ± 2.9	148.7	FedProx proximal term effectively mitigates Non-IID divergence
Async FL + Grad. Compression — NIID	83.64 ± 0.88	33.4 ± 2.0	71.2	Best efficiency; minimal accuracy cost vs. async+FedProx

4.5 Per-Class Federated Detection Analysis

The per-class F1-scores under Non-IID α = 0.1 are reported in Table 7, while the confusion matrix and ROC curves (Figures 1 and 4) illustrate class-specific detection behaviour. All values are presented as mean ± std over five runs. The Async+FedProx configuration achieves the strongest federated performance under heterogeneous conditions. Relative to plain asynchronous FL, the largest improvements are observed for Scanning (+0.158) and Backdoor (+0.110), indicating that FedProx effectively constrains local model divergence and preserves globally consistent decision boundaries for minority classes. Detection of Normal traffic and DoS/DDoS remains robust across configurations due to their environment-agnostic volumetric patterns.

ISSN NO: 2770-4491
August 2025

Table 7. Per-Class F1-Score — Non-IID α = 0.1 Configurations (Mean ± Std, n = 5)

Configuration	Normal	DoS / DDoS	Mirai	Ransomware	Backdoor	Scanning
Centralised DNN (reference)	0.941 ± 0.005	0.934 ± 0.006	0.921 ± 0.007	0.889 ± 0.010	0.841 ± 0.014	0.681 ± 0.021
Sync FL — IID, 10 clients	0.912 ± 0.007	0.901 ± 0.009	0.886 ± 0.010	0.851 ± 0.013	0.793 ± 0.017	0.601 ± 0.026
Sync FL — Non-IID α=0.1	0.881 ± 0.011	0.863 ± 0.013	0.844 ± 0.014	0.804 ± 0.018	0.714 ± 0.024	0.461 ± 0.038
Async FL — Non-IID α=0.1	0.832 ± 0.016	0.814 ± 0.018	0.796 ± 0.019	0.751 ± 0.023	0.653 ± 0.030	0.384 ± 0.046
Async FL + FedProx — Non-IID	0.904 ± 0.009	0.889 ± 0.011	0.874 ± 0.012	0.838 ± 0.015	0.763 ± 0.022	0.542 ± 0.032

4.6 Communication Overhead Analysis

The communication breakdown is summarised in Table 8, where overhead is defined as model size (MB) × mean clients per round × completed rounds. Scaling from 10 to 20 clients doubles communication cost (374.2 MB vs. 187.3 MB) while yielding only marginal accuracy gains (87.91% vs. 88.74%), indicating diminishing returns.

Table 8. Communication Overhead Analysis (Mean ± Std Rounds, n = 5)

FL Configuration	Upload/Round (MB)	Rounds to 80%	Total Comm. (MB)	Final Acc. (%)
10 clients, Sync, IID	18.7	41.3 ± 2.1	187.3	88.74 ± 0.58
10 clients, Async, IID	13.6	31.6 ± 1.7	136.4	89.12 ± 0.63
20 clients, Sync, IID	37.4	44.8 ± 2.4	374.2	87.91 ± 0.63
10 clients, Gradient Compression	9.4	42.1 ± 2.3	93.8	83.64 ± 0.88
10 clients, Async + Compression	7.1	33.4 ± 2.0	71.2	83.64 ± 0.88

4.7 Privacy Analysis

No raw network traffic data is transmitted to the central server in any FL configuration. Only model weight tensors (~46 KB per round per client) are exchanged. FL enhances privacy relative to centralised approaches: individual device traffic patterns remain local, and a compromised central server obtains model weights, not raw records, reducing breach severity. FL does not guarantee full regulatory compliance with GDPR or equivalent frameworks without additional mechanisms. Membership inference attacks (Shokri et al., 2019) can leak individual record information from weight updates. Differential privacy (DP-SGD; Abadi et al., 2016) can be applied client-side—adding calibrated Gaussian noise before transmission—at an approximate cost of 1–3 pp accuracy at ε = 8. This extension is deferred to future work.

4.8 Comparison with Related Works

A comparative context for PP-CIDS relative to published federated learning and IoT IDS studies is provided in Table 9, with all novelty claims restricted to this selected set. Within this scope, the study represents one of the first evaluations of federated multi-class IDS on ToN_IoT incorporating explicit Non-IID Dirichlet partitioning and statistical validation across multiple runs. The achieved 11-class performance under Non-IID conditions (84.21%) reflects a more challenging setting than binary or IID-based evaluations reported in prior work.

ISSN NO: 2770-4491
August 2025

Table 9. Comparison with Published Federated Learning and IoT IDS Studies (Selected Comparison Set)

Study	Method	Best Accuracy (%)	Notes
McMahan et al. (2017)	FedAvg	N/A (framework)	Foundational FL; no IDS; no ToN_IoT
Zhao et al. (2018)	Non-IID FL impact	N/A (analysis)	Non-IID analysis; no IDS; computer vision
Li et al. (2020)	FedProx	N/A (algorithm)	Proximal regularisation; no IDS; no ToN_IoT
Nguyen et al. (2021)	DIOT — FL anomaly (IoT)	94.1 ± 0.8	FL IoT IDS; multi-class; no ToN_IoT; no Non-IID analysis
Mothukuri et al. (2021)	FL survey for IoT security	N/A (survey)	Identifies Non-IID as major gap; no ToN_IoT evaluation
Agrawal et al. (2022)	FL-IDS on CICIDS2017	91.8 ± 0.6	FedAvg; no async FL; no Non-IID analysis
Rey et al. (2022)	FL anomaly (N-BaIoT)	93.2 ± 0.5	Sync FedAvg; no Non-IID Dirichlet analysis
This Study	Async FL + FedProx (ToN_IoT)	89.12 ± 0.63 (IID) / 84.21 ± 0.84 (Non-IID α=0.1)	ToN_IoT 11-class; async FL; Dirichlet Non-IID; comms analysis; 5-run mean ± std

5. Conclusion

This paper presented the Privacy-Preserving Collaborative Intrusion Detection System (PP-CIDS), a federated learning framework for IoT intrusion detection on ToN_IoT, evaluated across five independent runs (mean ± std throughout). Under IID partitioning, the asynchronous federated MLP achieved 89.12% ± 0.63%—within 3.1 pp of the centralised DNN reference (92.21%), without transmitting raw traffic data. Under severe Non-IID heterogeneity ($\alpha = 0.1$), plain asynchronous FL degraded to 80.43% (−8.69 pp), highlighting compounding stale gradient effects. Asynchronous FL with FedProx recovered to 84.21% ± 0.84%—3.78 pp above plain async FL, demonstrating the effectiveness of proximal regularisation. Gradient compression reduced communication overhead by 62% with a minimal 0.57 pp accuracy cost. FL enhances privacy relative to centralised collection but does not guarantee full regulatory compliance—additional mechanisms are required for production deployment under stringent data sovereignty frameworks.

REFERENCES

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. In Proceedings of the 23rd ACM SIGSAC CCS (pp. 308–318). ACM.

Agrawal, S., Sarkar, S., & Abutaleb, A. (2022). Federated learning for intrusion detection in IoT. In Proceedings of the IEEE ICC Workshops (pp. 1–6). IEEE.

Antonakakis, M., April, T., Bailey, M., et al. (2017). Understanding the Mirai botnet. In Proceedings of the 26th USENIX Security Symposium (pp. 1093–1110). USENIX.

Beutel, D. J., Topal, T., Mathur, A., et al. (2020). Flower: A friendly federated learning research framework. arXiv:2007.14390.

Bonawitz, K., Ivanov, V., Kreuter, B., et al. (2017). Practical secure aggregation for privacy-preserving machine learning. In Proceedings of the 24th ACM SIGSAC CCS (pp. 1175–1191). ACM.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, 9(3–4), 211–407.

European Parliament. (2018). General Data Protection Regulation (GDPR). Official Journal of the European Union, L 119.

Hsu, T. M. H., Qi, H., & Brown, M. (2019). Measuring the effects of non-identical data distribution for federated visual classification. arXiv:1909.06335.

Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S., Stich, S., & Suresh, A. T. (2020). SCAFFOLD: Stochastic controlled averaging for federated learning. In Proceedings of the 37th ICML (pp. 5132–5143). PMLR.

Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., & Smith, V. (2020). Federated optimization in heterogeneous networks. In MLSys 2020 (pp. 429–450).

McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & Agüera y Arcas, B. (2017). Communication-efficient learning of deep networks from decentralized data. In AISTATS (pp. 1273–1282). PMLR.

Mothukuri, V., Parizi, R. M., Pouriyeh, S., Huang, Y., Dehghantanha, A., & Srivastava, G. (2021). A survey on security and privacy of federated learning. Future Generation Computer Systems, 115, 619–640.

Moustafa, N. (2021). A new distributed architecture for evaluating AI-based security systems at the edge: Network TON_IoT datasets. Sustainable Cities and Society, 72, 102994.

Moustafa, N., & Slay, J. (2015). UNSW-NB15: A comprehensive dataset for network intrusion detection systems. In MilCIS (pp. 1–6). IEEE.

Nguyen, T. D., Rieger, P., Chen, H., Truong, H., Bhatt, G., et al. (2021). DIOT: A federated self-learning anomaly detection system for IoT. In ICDCS (pp. 756–767). IEEE.

Niknam, S., Dhillon, H. S., & Reed, J. H. (2020). Federated learning for wireless communications. IEEE Communications Magazine, 58(6), 46–51.

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

Rey, V., Sanchez, P. M. S., Celdran, A. H., & Bovet, G. (2022). Federated learning for malware detection in IoT devices. Computer Networks, 204, 108693.

Shokri, R., Strobel, M., Song, Y., & Vitanov, I. (2019). Privacy risks of explaining machine learning models. arXiv:1907.00164.

Statista. (2022). Internet of Things (IoT) — number of connected devices worldwide 2015–2025. Statista Research Department.

Ullah, I., & Mahmoud, Q. H. (2020). A scheme for generating a dataset for anomalous activity detection in IoT networks. In AI 2020 (pp. 508–520). Springer.

Wang, J., Liu, Q., Liang, H., Joshi, G., & Poor, H. V. (2020). Tackling the objective inconsistency problem in heterogeneous federated optimization. In NeurIPS 2020 (pp. 7611–7623). Curran Associates.

Xie, C., Koyejo, S., & Gupta, I. (2019). Asynchronous federated optimization. arXiv:1903.03934.

Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., & Chandra, V. (2018). Federated learning with non-IID data. arXiv:1806.00582.