Journal of Computer Science and Technology Studies
ISSN: 2709-104X
DOI: 10.32996/jcsts
Journal Homepage: www.al-kindipublisher.com/index.php/jcsts

JCSTS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Federated Intrusion Detection for Internet of Medical Things Networks: Differential Privacy, Non-IID Robustness, and Cross-Device Generalization

Mahfuz Islam Khan Jabed[1]✉, Md Parvez Ahmed[2], Fardous Mir Tofa[3], Md Faisal Islam[4], Clapher Ankur Gomes[5], Md Riad Mahamud Sirazy[6]

[1]*Department of Information Technology, Washington University of Science and Technology, Virginia, USA*
[2]*Department of Information Technology, Washington University of Science and Technology, Virginia, USA*
[3]*Department of Information Technology, Washington University of Science and Technology, Virginia, USA*
[4]*Department of Business Administration, Texas A&M University, Texas, USA*
[5]*Department of Computer Science, University of the Potomac, Washington D.C, USA*
[6]*Department of Cyber Security, Washington University of Science and Technology, Virginia, USA*
Corresponding Author: Mahfuz Islam Khan Jabed, **E-mail**: mahfuzislamkhanjabed@gmail.com

| ABSTRACT

Internet of Medical Things (IoMT) deployments combine clinically relevant sensing with heterogeneous wireless protocols and resource-constrained devices, creating a difficult setting for centralized intrusion detection. Federated learning can retain network-flow records at participating sites or gateways, but statistical heterogeneity and privacy protection can reduce utility. This study evaluates centralized and federated binary intrusion-detection models on a reproducible 71,326-row sample derived from the WiFi/MQTT portion of CICIoMT2024. After excluding 14,400 profiling-only benign flows, 28,740 training and 28,186 test flows covering 18 attack types and benign traffic were modeled with 44 predictors. Seven non-identically distributed client partitions were generated with a Dirichlet concentration of 0.30 over fine-grained labels; six participated in training and one was withheld from all model fitting. Centralized logistic regression and random forest were compared with FedAvg, FedProx, and record-level differentially private FedProx over five partition seeds. The private mechanism clipped per-record gradients, added Gaussian noise calibrated to replace-one-record sensitivity, and used zero-concentrated differential privacy composition, yielding $(\varepsilon, \delta) = (4.106, 10^{-5})$ for 40 local mechanisms per client.

Centralized random forest provided the empirical ceiling (mean F1 0.975; 95% CI 0.970–0.979). FedAvg and FedProx achieved mean F1 values of 0.855 and 0.855, respectively, showing no robustness gain from the selected proximal term. Differentially private FedProx achieved mean F1 0.822 (95% CI 0.818–0.826) and held-out-partition F1 0.844 (95% CI 0.797–0.891). In the primary partition, the private model's F1 was 0.033 lower than FedProx (paired bootstrap 95% CI −0.035 to −0.031; exact McNemar $p < 10^{-200}$); across only five seeds, the paired Wilcoxon test was not significant at 0.05 ($p = 0.0625$). Critically, CICIoMT2024's released processed table contains no verified physical-device identifier. Therefore, the withheld partition measures cross-partition generalization under synthetic label shift, not validated generalization to a new physical device. The study offers an auditable baseline and identifies device-aware metadata, stronger heterogeneity methods, and operational validation as necessary next steps.

| KEYWORDS

Internet of Medical Things (IoMT); Federated Learning; Intrusion Detection; Differential Privacy; Non-IID Data; Privacy-Preserving Machine Learning; Cybersecurity; Federated Optimization

| ARTICLE INFORMATION

ACCEPTED: 15 June 2026 **PUBLISHED:** 10 August 2026 **DOI:** 10.32996/jcsts.2026.8.8.23

1. Introduction

Internet-connected medical sensors, wearable monitors, imaging systems, gateways, and hospital support devices expand the observable surface of a healthcare network. Their traffic can be heterogeneous in protocol, timing, payload size, and operational purpose. These properties complicate intrusion detection because a detector must separate attacks from legitimate device-specific behavior while avoiding excessive false alarms. Survey evidence has repeatedly identified heterogeneity, limited device resources, shifting traffic distributions, and the sensitivity of medical environments as central barriers to IoT security analytics [3]. Recent datasets such as CICIoMT2024 and IoMT-TrafficData have improved the empirical basis for evaluating IoMT intrusion-detection systems [1], [2], yet evaluation designs still frequently aggregate traffic into a single centralized table.

Centralized training creates a useful performance ceiling, but it is not always a realistic governance model. Raw flow records may remain under different institutional, gateway, or administrative controls. Federated learning addresses this coordination problem by moving model updates rather than source records [7], [10]. The approach has been investigated in medical cyber-physical systems, IoMT intrusion detection, and hierarchical edge–cloud architectures [4]–[6]. Federated learning, however, does not automatically solve privacy or robustness. Model updates may reveal information about local records, and the participating clients may have markedly different class mixtures. Medical federated-learning research likewise emphasizes that data locality is useful but is not equivalent to formal privacy [17]–[20].

Two technical tensions motivate this study. First, FedAvg can drift when local objectives differ, whereas FedProx adds a proximal penalty intended to stabilize optimization under systems and statistical heterogeneity [7], [8]. Control-variate approaches such as SCAFFOLD provide another response to client drift [9], but add state and communication complexity. Second, differential privacy offers a formal limit on the influence of a record by combining sensitivity control and randomized noise [11]–[14]. That protection can reduce detection utility, and the reduction may be uneven across groups or clients [15], particularly under non-identically distributed (non-IID) data [16].

This paper evaluates these tensions using a transparent, lightweight federated logistic-regression implementation and a reproducible source-derived sample. The title refers to "cross-device generalization" because deployment to unseen devices is an important target. Nevertheless, the released processed WiFi/MQTT table used here has no verified physical-device identifier. The experiment therefore constructs proxy clients using label-mixture partitions and withholds one proxy from training. This is a stress test of cross-partition generalization, not proof of performance on a new physical IoMT device. Making this construct limitation explicit is a central part of the study rather than a post hoc qualification.

2. Litterature Review

2.1 IoMT security data and intrusion detection

Machine-learning intrusion detection depends on representative traffic, meaningful labels, and evaluation splits that reflect deployment. Al-Garadi et al. cataloged the broad use of machine and deep learning for IoT security while emphasizing data quality and operational constraints [3]. CICIoMT2024 was designed as a multi-protocol IoMT benchmark with 40 IoMT devices, 18 attack types, and WiFi, MQTT, and Bluetooth traffic [1]. IoMT-TrafficData similarly supports benchmarking with IoMT-focused data and tools [2]. Outside healthcare, Edge-IIoTset has been used for both centralized and federated security experiments [21]. These resources improve comparability, but table schema determines which claims an experiment can support. If physical device identity is not retained, a device-level split cannot be reconstructed reliably from traffic labels alone.

2.2 Federated intrusion detection

McMahan et al. introduced the widely used FedAvg procedure, in which clients optimize locally from a common global model and the server computes a sample-weighted average [7]. IoMT-specific studies have explored federated attack detection in medical cyber-physical systems [4], transfer-based IDS designs [5], and hierarchical dew–cloud federated architectures [6]. These studies demonstrate the relevance of federated coordination to distributed healthcare environments, but results depend on client definitions, data partitions, model families, and threat assumptions.

Non-IID data is a persistent challenge. FedProx augments the local objective with a proximal term that discourages local parameters from moving far from the current global model [8]. SCAFFOLD instead estimates control variates to correct client drift [9]. The broad federated-learning literature documents additional open issues around partial participation, communication, personalization, fairness, privacy, and adversarial clients [10]. Because a method's advantage is not universal, a careful evaluation should preserve negative findings such as a failure of FedProx to improve over FedAvg, rather than assuming the intended effect.

2.3 Differential privacy and healthcare federated learning

Differential privacy (DP) limits how much an algorithm's output distribution can change when one protected unit changes [11]. Gradient clipping bounds sensitivity, while Gaussian noise provides an approximate-DP mechanism [12]. Rényi DP and concentrated-DP formulations support composition across many releases [13], [14]. In federated settings, the protected unit must be stated: record-level DP and client-level DP are different guarantees. This study protects one training record within a participating proxy client; it does not provide client-level DP.

Privacy noise can have unequal accuracy effects [15], and non-IID federated optimization can compound the trade-off [16]. In healthcare, federated learning has been proposed as a means of enabling multi-institutional collaboration without pooling source data [17], [18]. Experiments combining federated learning and differential privacy in medical imaging show the feasibility of formal privacy, but also the need to report the corresponding utility cost [19]. Security surveys further warn that federated systems remain exposed to inference, poisoning, and aggregation risks unless the system threat model and countermeasures are explicit [20].

2.4 Positioning of the present study

The present study differs from an architecture proposal in three ways. First, it uses a fixed, inspectable linear federated learner so the effects of optimization and noise can be audited. Second, it reports a matched centralized logistic baseline and a stronger centralized random-forest ceiling rather than treating any centralized score as directly comparable. Third, it separates a useful held-out-partition stress test from an unsupported physical-device claim. This emphasis on construct validity is important for both scientific reporting and any later use of the work as evidence of expertise.

3. Materials and Methods

3.1 Data source and scope

The source benchmark was CICIoMT2024 [1]. The official study describes traffic from 40 IoMT devices, 18 attacks, and WiFi, MQTT, and Bluetooth protocols. For reproducibility, this work used the publicly available processed derivative titled Core_Processed_Data.zip (Zenodo DOI: 10.5281/zenodo.19452870). The downloaded archive had size 250,599,251 bytes and MD5 checksum ff856492ab3ca66234c3b5968a74ce85. The WiFi/MQTT Parquet table contained 9,128,910 rows; the Bluetooth table had a different feature schema and was excluded from the common-feature experiment.

Only network-flow attributes and labels were used. No patient records, clinical outcomes, human-subject identifiers, or protected health information were present in the analysis sample. The work evaluates cybersecurity telemetry; it does not evaluate diagnosis, treatment, or patient outcome prediction.

3.2 Reproducible sampling and data quality

The source table included official train, test, and profiling designations. A deterministic Node.js sampler retained all 19 fine labels represented by benign traffic plus 18 attack types. For each attack label, up to 800 training and 800 test rows were sampled. Benign traffic was sampled at 14,400 rows in each of the train, test, and profiling groups. The resulting file contained 71,326 rows: 28,740 training rows, 28,186 test rows, and 14,400 profiling-only benign rows. Profiling rows were retained for provenance and descriptive analysis but excluded from every model fit and metric.

The sample contained no missing cells, non-finite numeric values, or duplicate analytical rows after excluding the generated row identifier and source-file string. Forty-four predictors were retained; drate and time_to_live were removed because they were constant in the sample. Table 1 summarizes the audit.

Table 1. Analysis-sample audit.

Check	Value	Status
Source-derived sample rows	71,326	Pass
Modeling rows (train + test)	56,926	Pass
Training rows	28,740	Pass
Test rows	28,186	Pass
Profiling-only rows excluded from modeling	14,400	Pass
Missing cells	0	Pass
Infinite numeric cells	0	Pass
Duplicate rows excluding row ID/source file	0	Pass
Fine-grained labels	19	Pass
Predictor features retained	44	Pass

Check	Value	Status
Constant features removed	2	Pass
Verified physical-device identifier	0	Limitation

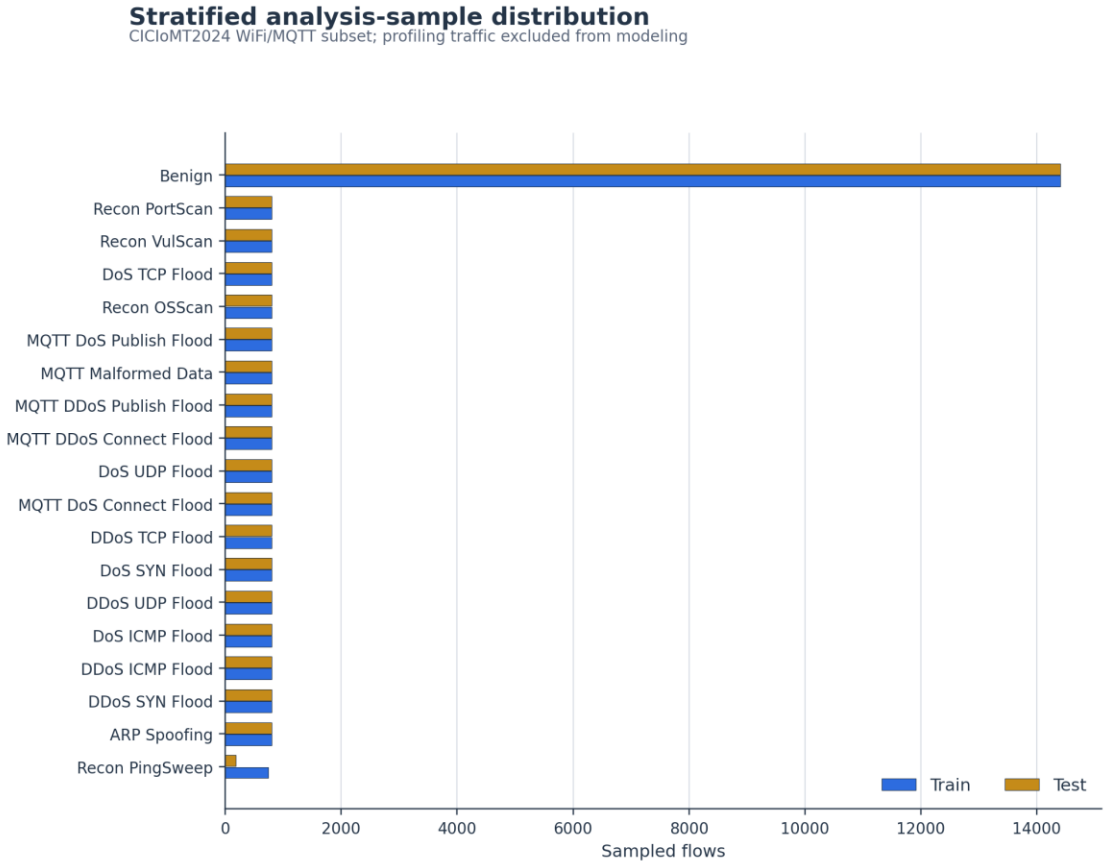


Figure 1. Distribution of the source-derived training and test sample. Profiling traffic is excluded.

3.3 Outcome and preprocessing

The model outcome was binary: benign ($y = 0$) or attack ($y = 1$). The 18 attack labels were preserved for stratified sampling and client allocation, but collapsed to the attack class for model fitting. This design tests whether a detector can identify malicious traffic, not whether it can assign an attack family.

Identifiers and descriptive columns—row_id, label, source_file, protocol, split, and binary_label—were excluded from predictors. Each numeric predictor x received a signed logarithmic transformation,

$$x' = \text{sign}(x)\log(1 + |x|),$$

which reduces extreme skew while preserving the sign of a value. A standard scaler then transformed each feature using $z = (x' - \mu)/s$. Means and standard deviations were estimated only from records belonging to the six participating training clients within each seed. The held-out client and all test rows were excluded from preprocessing estimation, preventing leakage.

3.4 Non-IID proxy-client construction

Seven client partitions were constructed independently for seeds 17, 29, 43, 71, and 101. For every fine label k, client proportions were drawn from a symmetric Dirichlet distribution,

$$\mathbf{p}_k \sim \text{Dirichlet}(\alpha\mathbf{1}), \qquad \alpha = 0.30.$$

The same label-level proportions were used to allocate training and test rows, with independent row shuffling. Candidate partitions were accepted only when every client had at least 400 rows and at least 50 benign and 50 attack rows in both train and test. Clients 0–5 participated in training; client 6 was excluded from centralized preprocessing, centralized training, and federated training. Metrics were computed overall, over seen clients 0–5, for the held-out client 6, and for every client separately.

The resulting clients are capture/protocol label-mixture proxies. They are not asserted to be actual devices, hospitals, patients, or users. Figure 2 shows the primary seed's non-IID composition.

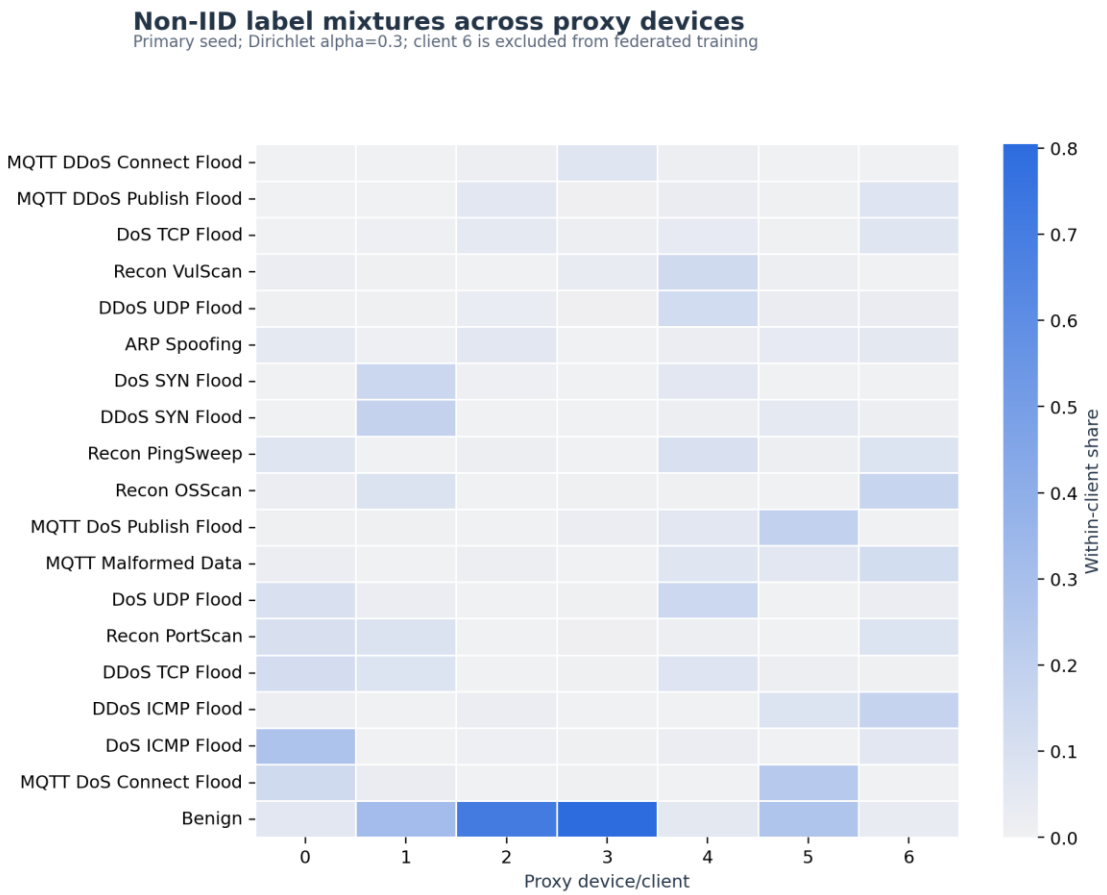


Figure 2. Fine-label mixtures across the seven proxy clients for the primary seed. Client 6 is withheld from training.

3.5 Models and optimization

Two centralized baselines and three federated models were evaluated.

Centralized logistic regression (Central LR). A logistic model with an intercept was trained on the union of clients 0–5 using the L-BFGS solver and a maximum of 2,000 iterations. This is the architecture-matched centralized comparator for the federated linear models.

Centralized random forest (Central RF). A random forest with 140 trees, square-root feature sampling, minimum leaf size 2, and balanced subsample weights was trained on the same participating records. It is an empirical nonlinear ceiling, not an architecture-matched federated comparison.

FedAvg. At communication round t, each participating client i started from the global weights w_t, performed two full-batch gradient steps, and returned its local weights $w_{t+1}^{(i)}$. The server computed

$$w_{t+1} = \sum_{i=1}^{m} \frac{n_i}{\sum_j n_j} w_{t+1}^{(i)},$$

where n_i is the client's training-row count [7].

FedProx. The local objective added the proximal penalty

$$F_i(w) + \frac{\mu}{2} \| w - w_t \|_2^2,$$

with $\mu = 0.10$ [8]. All other optimization parameters matched FedAvg.

DP-FedProx. For each local full-batch step, the per-record logistic gradient g_j was clipped to L2 norm $C = 2.0$:

$$\bar{g}_j = g_j \min\left(1, \frac{C}{\| g_j \|_2}\right).$$

The client used the noisy mean

$$\tilde{g} = \frac{1}{n_i} \sum_{j=1}^{n_i} \bar{g}_j + \mathcal{N}\left(0, \left(\sigma \frac{2C}{n_i}\right)^2 I\right),$$

where $2C/n_i$ is the L2 sensitivity of the clipped mean under replace-one-record adjacency and $\sigma = 8.0$ is the noise multiplier. The mechanism is record-level: neighboring client datasets have the same size and differ in one record. It is not client-level DP.
Each model was trained for 20 rounds, with two local steps per round and learning rate 0.25. An L2 penalty of 10^{-4} was applied to coefficients but not the intercept. All six clients participated in every round.
Table 2 lists the design.

Component	Setting
Partition seeds	17, 29, 43, 71, 101
Proxy clients	7 total; 6 train, 1 held out
Dirichlet concentration	$\alpha = 0.30$
Federated model	Logistic regression with intercept
Communication rounds	20
Local steps per round	2 full-batch steps
Learning rate	0.25
Coefficient L2 penalty	10^{-4}
FedProx coefficient	$\mu = 0.10$
DP clipping norm	$C = 2.0$
DP noise multiplier	$\sigma = 8.0$
Protected unit / adjacency	One record / replace-one
Decision threshold	0.50

3.6 Privacy accounting and threat model
One Gaussian mechanism with noise multiplier σ satisfies $\rho = 1/(2\sigma^2)$-zCDP; zCDP composes additively [14]. With $T = 20 \times 2 = 40$ local mechanisms per participating client,

$$\rho_{\text{total}} = \frac{T}{2\sigma^2} = 0.3125.$$

The standard zCDP conversion gives

$$\varepsilon = \rho_{\text{total}} + 2\sqrt{\rho_{\text{total}} \log(1/\delta)}.$$

At $\delta = 10^{-5}$, the reported guarantee is $(\varepsilon, \delta) = (4.106, 10^{-5})$ per participating client. This bound claims no privacy amplification from client sampling or record subsampling because every client and every local record participates in each full-batch step. Composition over all internal steps is conservative because the server receives only the resulting local update, which is post-processing of the noisy steps.
The threat model treats a training record's contribution to a participating client update as sensitive. It does not protect whether an entire client participated, provide secure aggregation, defend against malicious clients, or establish robustness to model poisoning. Record-level DP also does not make raw endpoint storage secure; ordinary access control, encryption, logging, and incident response remain necessary [20].

3.7 Evaluation and statistical analysis
At threshold 0.50, the study reports accuracy, balanced accuracy, precision, recall, F1, specificity, false-positive rate (FPR), Brier score, receiver-operating-characteristic area under the curve (ROC-AUC), and precision–recall area under the curve (PR-AUC). PR curves are included because they can reveal behavior hidden by ROC summaries under class imbalance [24].
For each model and scope, means, sample standard deviations, and two-sided 95% t intervals were computed across five independent partition seeds. These intervals quantify variability across the selected synthetic partitions; they do not imply population sampling from all IoMT deployments.

For the primary seed, 1,000 paired nonparametric bootstrap resamples estimated metric intervals and the F1 difference between DP-FedProx and FedProx [22]. An exact two-sided McNemar test compared their paired correctness outcomes [23]. A paired exact Wilcoxon signed-rank test compared their F1 scores across the five seeds. Because $n = 5$ seeds yields low inferential power, the seed-level test is reported alongside effect sizes and interval estimates rather than used as a binary gate.

3.8 Reproducibility controls

Sampling, partitioning, model initialization, privacy noise, and bootstrap resampling use explicit seeds. Source checksums, package versions, environment metadata, raw predictions, and intermediate tables are preserved. The central and federated models use identical transformed features and participating training records within a seed. The notebook can load all exported results without rerunning the approximately two-minute experiment, while the Python scripts reproduce the analysis from the packaged CSV.

4. Results

4.1 Overall discrimination

Table 3 and Figure 3 summarize overall test performance across the five non-IID partitions. Central RF produced the highest mean scores: balanced accuracy 0.975, F1 0.975, ROC-AUC 0.997, PR-AUC 0.997, and FPR 0.010. Because this model is nonlinear and not federated, it should be interpreted as an empirical ceiling for the selected features rather than a like-for-like federated baseline.

Central LR, the architecture-matched comparator, achieved mean F1 0.935. FedAvg achieved 0.855, a decrease of 0.080 relative to Central LR. FedProx achieved 0.855 and was 0.0007 below FedAvg at the displayed precision. Therefore, the selected proximal coefficient did not reduce the utility gap created by federated optimization and non-IID partitions.

DP-FedProx achieved mean F1 0.822, balanced accuracy 0.831, ROC-AUC 0.937, and PR-AUC 0.940. Its FPR was 0.133 compared with 0.079 for non-private FedProx. The private model therefore retained useful discrimination, but its fixed-threshold error profile would create materially more benign alerts.

Table 3. Overall test performance across five non-IID partition seeds. Values are mean (95% t interval).

Model	Balanced accuracy	F1	ROC-AUC	PR-AUC	FPR
Central LR	0.937 (0.932–0.943)	0.935 (0.930–0.940)	0.984 (0.983–0.984)	0.985 (0.984–0.986)	0.034 (0.022–0.046)
Central RF	0.975 (0.971–0.980)	0.975 (0.970–0.979)	0.997 (0.997–0.997)	0.997 (0.997–0.997)	0.010 (0.001–0.019)
FedAvg	0.865 (0.860–0.870)	0.855 (0.850–0.860)	0.961 (0.960–0.962)	0.963 (0.962–0.964)	0.078 (0.069–0.088)
FedProx	0.864 (0.859–0.870)	0.855 (0.849–0.860)	0.961 (0.960–0.962)	0.963 (0.962–0.964)	0.079 (0.069–0.089)
DP-FedProx	0.831 (0.826–0.836)	0.822 (0.818–0.826)	0.937 (0.936–0.938)	0.940 (0.939–0.941)	0.133 (0.121–0.144)

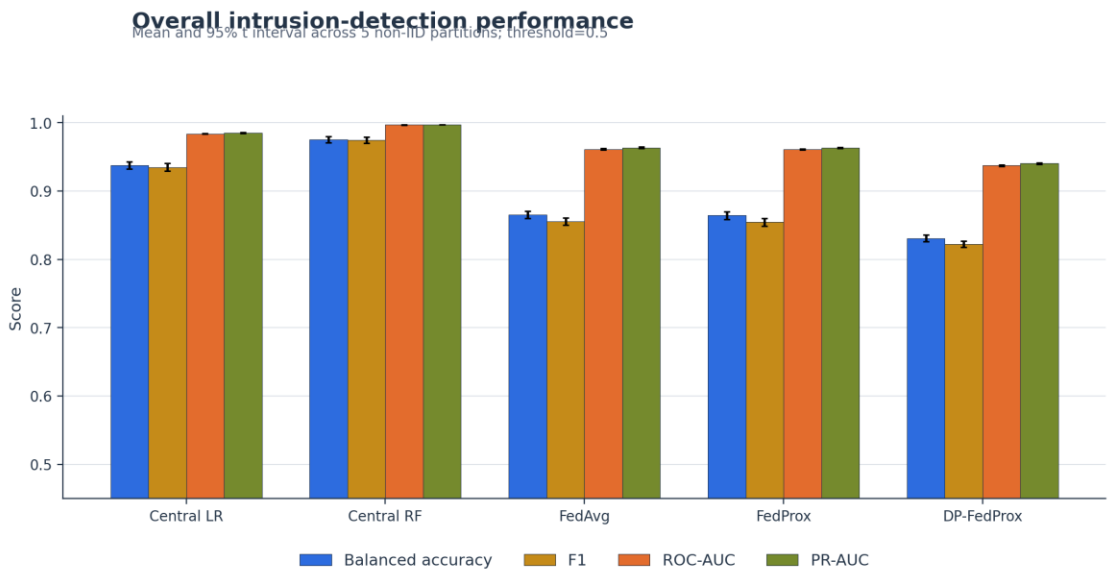
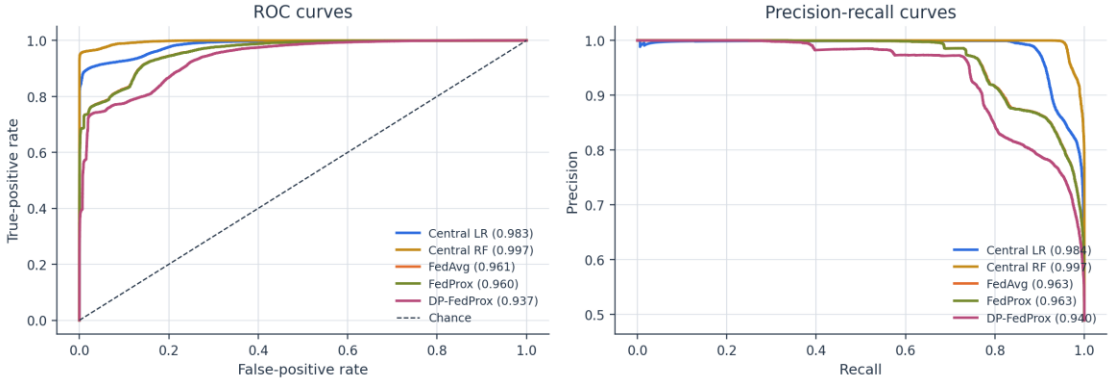


Figure 3. Mean overall performance with 95% t intervals across five non-IID partition seeds.

Primary-seed discrimination curves are shown in Figure 4. The primary ROC-AUC values were 0.983 for Central LR, 0.997 for Central RF, 0.961 for FedAvg, 0.960 for FedProx, and 0.937 for DP-FedProx. The corresponding PR-AUC values were 0.984, 0.997, 0.963, 0.963, and 0.940. These ranking metrics show that the private model's utility loss was not limited to the 0.50 threshold.



4.2 Per-client and held-out-partition generalization

Performance varied across proxy clients (Figure 5), reflecting the deliberately heterogeneous label mixtures. For the primary seed, FedProx F1 ranged from 0.745 to 0.964 across clients; DP-FedProx ranged from 0.654 to 0.962. The lowest private score occurred on proxy client 3, indicating that an overall average can conceal substantial client-specific degradation.

Across the five seeds, held-out client 6 F1 averaged 0.983 for Central RF, 0.950 for Central LR, 0.864 for FedAvg, 0.863 for FedProx, and 0.844 for DP-FedProx (Table 4). Confidence intervals were wider than the overall intervals because each seed's held-out client had a different synthetic composition and fewer observations.

Table 4. F1 on proxy client 6, excluded from all training.

Model	Mean F1	95% t interval
Central LR	0.950	0.916–0.985
Central RF	0.983	0.969–0.997
FedAvg	0.864	0.826–0.901

Model	Mean F1	95% t interval
FedProx	0.863	0.826–0.901
DP-FedProx	0.844	0.797–0.891

The held-out result answers RQ4 only within the experiment's synthetic construct. It shows transfer to a withheld label-mixture partition sampled from the same processed table. It does not demonstrate transfer to a physical device model, manufacturer, hospital, network, or future time period.

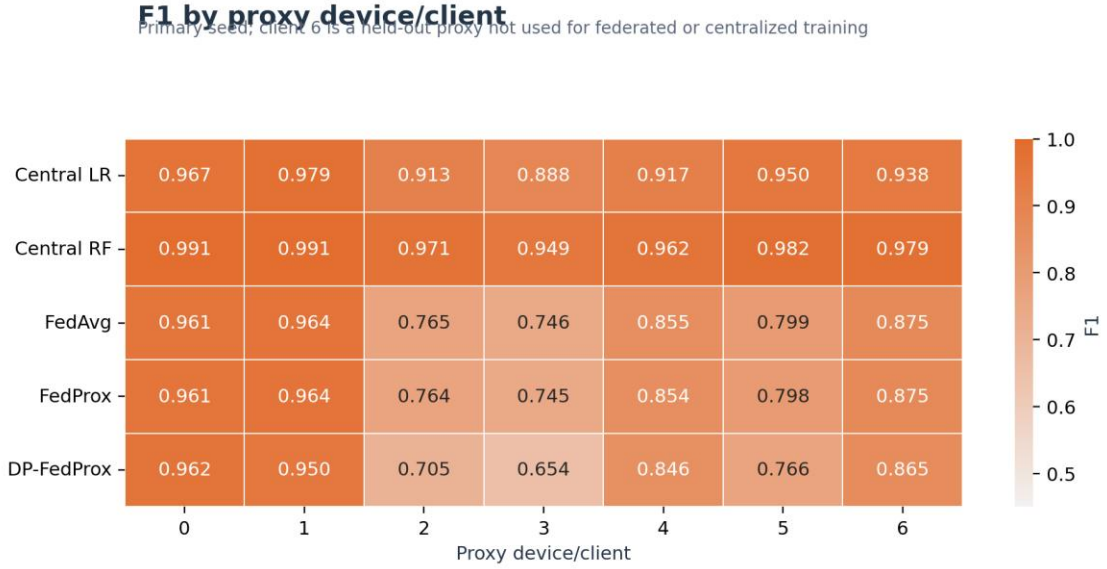


Figure 5. Primary-seed F1 by proxy client. Client 6 is excluded from centralized and federated training.

4.3 Privacy–utility analysis

The selected private model used sigma 8, rho 0.3125, and epsilon 4.106 at delta 0.00001. A primary-seed sensitivity sweep varied the noise multiplier while holding all other settings fixed (Table 5 and Figure 6). The realized F1 values were close across the four noise draws. Because each point is a single stochastic run, their small non-monotonic differences should not be interpreted as evidence that stronger privacy improves utility. The sweep primarily verifies accounting and shows that optimization and partition effects dominate small differences among these particular draws.

Table 5. Primary-seed privacy accounting and realized utility.

Noise multiplier	rho	epsilon at delta 0.00001	Overall F1	Held-out F1
4	1.2500	8.837	0.824	0.866
8	0.3125	4.106	0.823	0.866
12	0.1389	2.668	0.825	0.866
20	0.0500	1.567	0.826	0.867

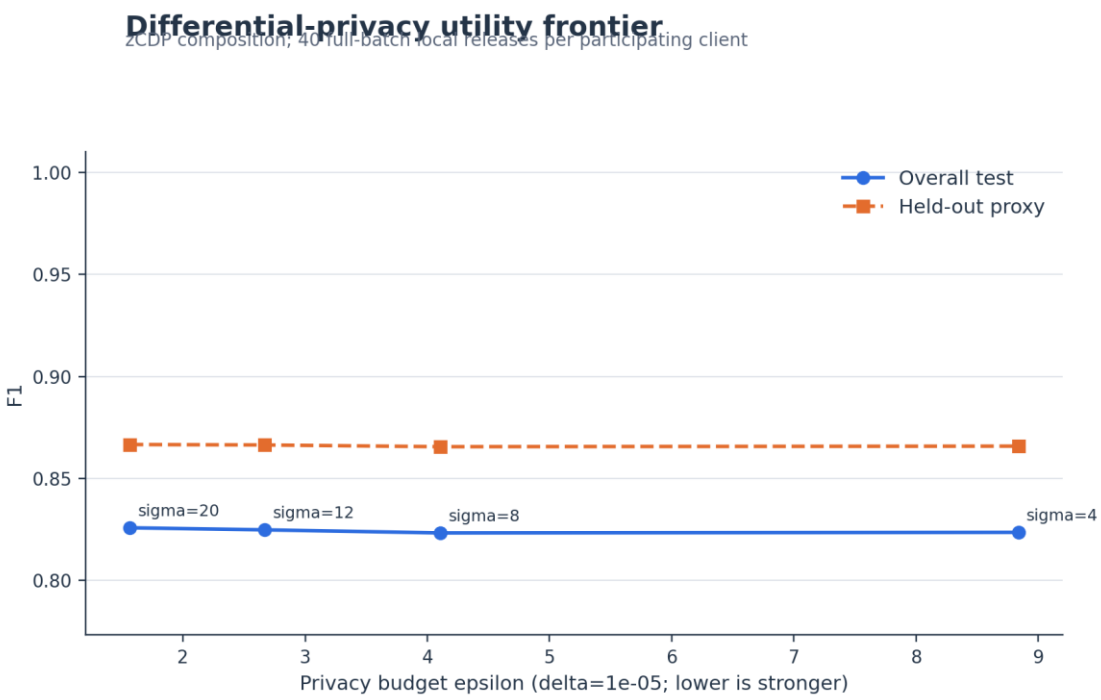


Figure 6. Privacy budget and realized F1 for the primary-seed noise sweep. Lower epsilon indicates stronger privacy.

4.4 Paired error and uncertainty analysis

In the primary seed, FedProx correctly classified 13,310 benign and 11,128 attack flows, with 1,090 false positives and 2,658 false negatives. DP-FedProx correctly classified 12,520 benign and 10,952 attack flows, with 1,880 false positives and 2,834 false negatives (Figure 7). Between the two models, 1,057 observations were correct only under FedProx and 91 were correct only under DP-FedProx. The exact McNemar test gave p = 3.09 × 10^-209, demonstrating a strong paired difference in primary-seed correctness.

The primary-seed F1 difference, DP-FedProx minus FedProx, was −0.0330. The 1,000-resample paired bootstrap interval was −0.0353 to −0.0308. Across the five independent partitions, the mean paired difference was −0.0326. The exact Wilcoxon signed-rank test gave p = 0.0625, which is the smallest attainable two-sided exact value for five nonzero differences of the same sign. This result is directionally consistent but underpowered at the conventional 0.05 threshold. Table 6 reports the analyses without converting them into a single significant/not-significant conclusion.

Table 6. Paired comparison of FedProx and DP-FedProx.

Analysis	Estimate	95% interval	p-value
Primary-seed paired bootstrap, F1 difference	−0.0330	−0.0353 to −0.0308	<0.001 by 1,000 resamples
Primary-seed exact McNemar, correctness	1,057 FedProx-only vs 91 DP-only correct	—	3.09 × 10^-209
Five-seed exact Wilcoxon, F1 difference	−0.0326	—	0.0625

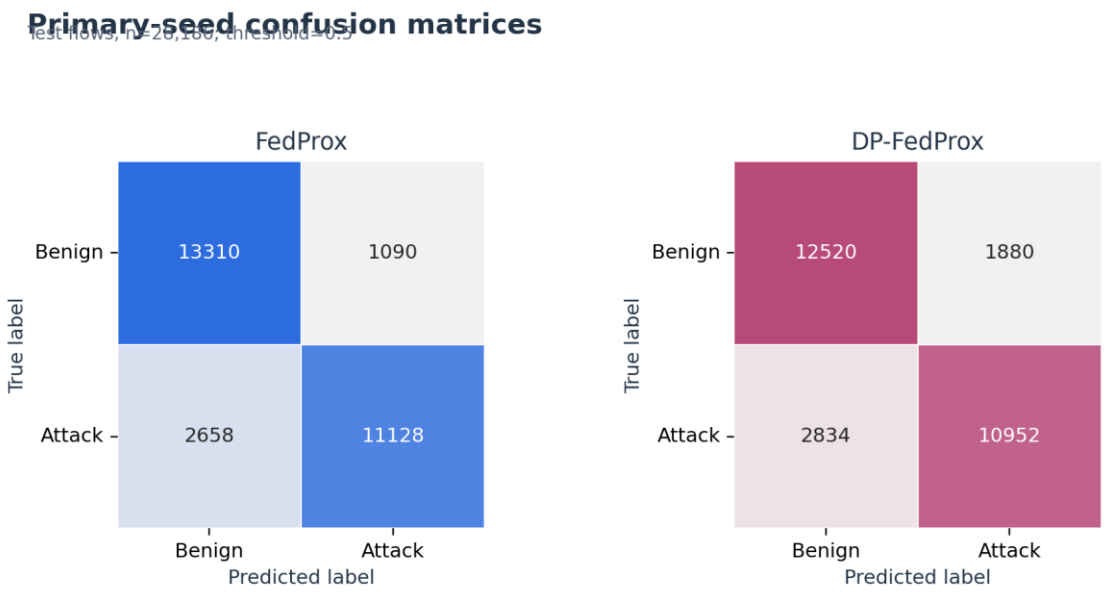


Figure 7. Primary-seed confusion matrices for non-private and private FedProx.

4.5 Optimization behavior

All three federated losses declined over 20 rounds (Figure 8). FedAvg and FedProx followed nearly identical trajectories, which is consistent with their nearly identical test metrics. DP-FedProx converged to a higher weighted training log loss because clipping and noise perturb local gradients. The smooth aggregate curve does not mean individual updates are noise-free; averaging across six clients and plotting one weighted loss per round suppresses visible step-level variation.

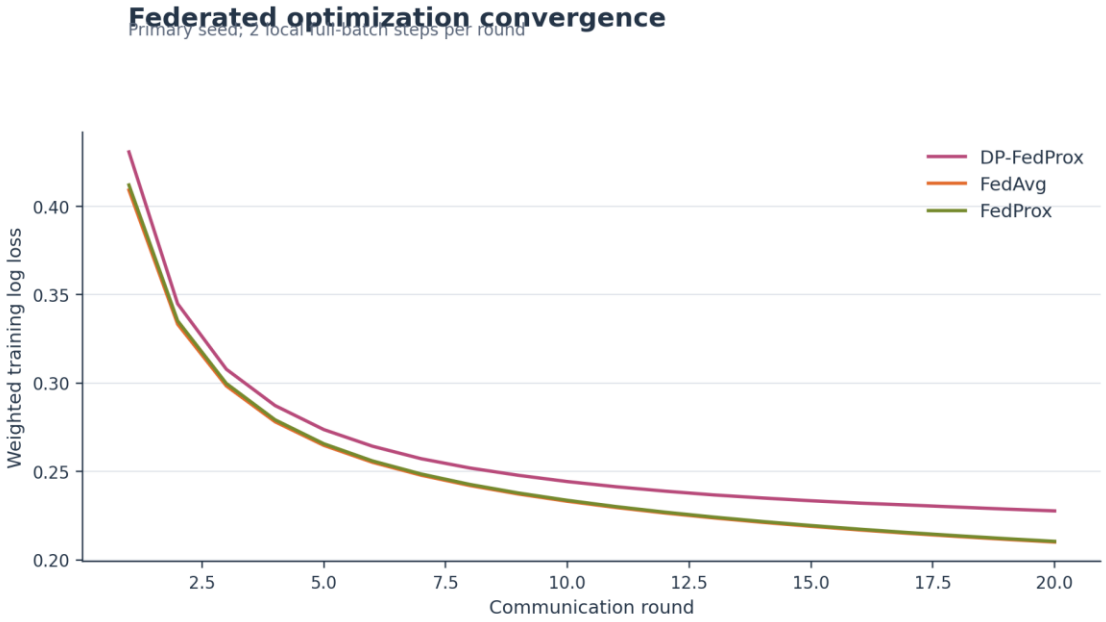


Figure 8. Primary-seed weighted training log loss over 20 communication rounds.

5. Discussion

5.1 Centralized performance

The random forest's F1 of 0.975 confirms that the sampled flow features contain strong signal for binary attack detection. It does not establish that raw records should be centralized or that the random forest can be deployed unchanged across institutions. Central RF benefits from nonlinear capacity and pooled access. The more relevant architecture-matched comparison is Central LR

versus the federated linear models, where the observed gap isolates the combined effect of distributed optimization and heterogeneous partitions more fairly.

5.2 FedProx did not help in the selected regime

FedProx is motivated by heterogeneous client objectives [8], but a proximal penalty is not guaranteed to improve every experiment. Here, only two local full-batch steps were performed before each aggregation. That limited local work may have constrained client drift enough that mu 0.10 added little. The label partitions were heterogeneous, yet all clients participated in every round and the model was convex. These factors differ from the systems heterogeneity and partial participation settings in which FedProx can be more useful.

The negative result suggests three next experiments. First, tune mu on a validation protocol nested within participating clients rather than selecting one value. Second, increase local steps to create a stronger drift regime. Third, compare control-variate or personalization methods, including SCAFFOLD [9]. None should be declared superior before a device-aware and leakage-safe evaluation.

5.3 Privacy changed the operational error profile

The private model's mean F1 reduction of approximately 0.033 is moderate in absolute terms but operationally meaningful. At threshold 0.50, its FPR rose from 0.079 for FedProx to 0.133. In a high-volume monitoring system, this difference can substantially increase alert-review load. Threshold adjustment could trade false positives against missed attacks, but should be chosen on a separate validation set with an explicit operational cost function. The present study deliberately retains one fixed threshold for comparability.

The private guarantee is also narrower than "federated privacy" in general. It protects one record within a participating proxy client under the stated mechanism and composition. It does not hide client participation, secure communications, or prevent malicious updates. A production design would normally combine record-level DP with authenticated transport, access controls, update validation, and—where the threat model requires it—secure aggregation [10], [20].

5.4 Cross-device claims require device identifiers

The strongest interpretive constraint is the absence of a verified physical-device ID. A synthetic client construction is useful for controlled stress testing: it creates repeatable label shift, supports paired comparisons, and exposes per-client performance differences. It cannot reproduce device-specific hardware, firmware, location, user behavior, temporal drift, or manufacturer protocols. Calling client 6 an "unseen device" without this caveat would overstate the evidence.

A future cross-device study should retain a physical device identifier, device category, manufacturer/model, capture session, and time window. It should then use group-aware splits: all flows from selected devices must remain outside training and preprocessing. If possible, the final test should include an external network or later capture period. Those design changes would convert the present proxy-generalization experiment into a genuine device-generalization study.

5.5 Implications for an IoMT research program

The study supports a coherent cybersecurity research direction centered on privacy-preserving, heterogeneous IoMT intrusion detection. It connects computer science, machine learning, federated optimization, privacy, and healthcare infrastructure without requiring a medical-diagnosis claim. A defensible continuation would focus on: (1) device-aware external validation; (2) adaptive or personalized federated IDS models; (3) secure aggregation and poisoning defenses; (4) privacy accounting under partial participation; and (5) cost-sensitive alert calibration. This trajectory is technically consistent with an information-technology background while addressing the operational importance of protecting connected healthcare systems.

6. Limitations and Threats to Validity

6.1 Construct validity

The term "proxy client" is essential. No verified physical-device identifier was available in the processed table, so the client partitions represent synthetic label mixtures. Cross-device generalization is a target application, not an established finding. The outcome is binary detection; multiclass attack attribution was not evaluated.

6.2 Internal validity

The train/test designations were inherited from the source-derived processed table. Although preprocessing was fitted only on participating training rows, flows from related capture processes may share latent structure across the official split. The experiment did not reconstruct session-level or temporal groups. The Dirichlet partitions also reuse the same label proportions for train and test within a seed, which creates controlled shift but may be more stable than real client evolution.

6.3 Statistical conclusion validity

Only five partition seeds were used. The t intervals summarize those five runs, and the exact Wilcoxon test has coarse attainable p-values. The primary-seed bootstrap quantifies resampling variability conditional on one test table and one partition; it does not replace independent-device replication. The privacy sweep used one stochastic run per multiplier and is descriptive.

6.4 External validity

The packaged sample is a stratified subset of the WiFi/MQTT processed table, not the full CICIoMT2024 corpus. Bluetooth was excluded because its schema differed. Findings may not transfer to raw packet data, encrypted payload features, other hospitals, unseen attack families, novel devices, or live traffic rates. No latency, memory, bandwidth, battery, or gateway throughput benchmark was conducted.

6.5 Model and privacy scope

The federated learner is logistic regression. Deep, sequential, and representation-learning models may change both utility and privacy cost. Central RF is not architecture-matched. Hyperparameters were not optimized exhaustively. The DP guarantee is record-level and relies on correct clipping, random-number generation, and noise calibration; it is not client-level privacy and provides no poisoning robustness. Secure aggregation was not implemented.

7. Ethics Statement

The analysis used public network-flow telemetry and did not involve patient records, human-subject recruitment, clinical decisions, or protected health information. The work is defensive: it evaluates detection of labeled malicious traffic and does not provide attack-generation instructions. Dataset licenses, institutional policies, and applicable privacy requirements should be reviewed before redistribution or deployment. The author should verify whether institutional review, data-use review, or cybersecurity laboratory approval is required at the intended institution, even when a dataset is public.

10. Conclusion

This study provides a reproducible evaluation of federated IoMT intrusion detection under synthetic non-IID label shift and record-level differential privacy. Centralized random forest established a strong empirical ceiling, while the matched centralized logistic model outperformed both federated linear optimizers. FedProx did not improve on FedAvg in the selected convex, full-participation regime. At epsilon 4.106 and delta 0.00001, DP-FedProx retained useful discrimination but reduced mean F1 by about 0.033 relative to FedProx and increased false positives. The private model achieved held-out proxy F1 0.844, but the dataset schema does not support a physical-device generalization claim.

The practical conclusion is not that privacy-preserving IoMT intrusion detection is ineffective. Rather, privacy, heterogeneity, and generalization must be measured together with a precise unit of analysis. The next decisive step is a device-aware, temporally separated, external validation study that combines stronger federated optimization with secure aggregation and explicit operational costs.

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest,
ORCID iDs:
Mahfuz Islam Khan Jabed: https://orcid.org/0009-0001-2141-2894
Md Parvez Ahmed: https://orcid.org/0009-0009-1981-4681
Fardous Mir Tofa: https://orcid.org/0009-0001-1874-1849
Md Faisal Islam: https://orcid.org/0000-0002-5694-1489
Clapher Ankur Gomes: https://orcid.org/0009-0004-9566-6784
Md Riad Mahamud Sirazy: https://orcid.org/0009-0009-4838-1635

References

[1] S. Dadkhah, E. C. Pinto Neto, R. Ferreira, R. C. Molokwu, S. Sadeghi, and A. A. Ghorbani, "CICIoMT2024: A benchmark dataset for multi-protocol security assessment in IoMT," Internet of Things, vol. 28, art. 101351, 2024. doi: 10.1016/j.iot.2024.101351.
[2] J. Areia, I. A. Bispo, L. Santos, and R. L. de C. Costa, "IoMT-TrafficData: Dataset and tools for benchmarking intrusion detection in Internet of Medical Things," IEEE Access, vol. 12, pp. 115370–115385, 2024. doi: 10.1109/ACCESS.2024.3437214.
[3] M. A. Al-Garadi, A. Mohamed, A. K. Al-Ali, X. Du, I. Ali, and M. Guizani, "A survey of machine and deep learning methods for Internet of Things (IoT) security," IEEE Communications Surveys & Tutorials, vol. 22, no. 3, pp. 1646–1685, 2020. doi: 10.1109/COMST.2020.2988293.

[4] W. Schneble and G. Thamilarasu, "Attack detection using federated learning in medical cyber-physical systems," in Proc. 28th International Conference on Computer Communication and Networks (ICCCN), 2019, pp. 1–8.
[5] Y. Otoum, Y. Wan, and A. Nayak, "Federated transfer learning-based IDS for the Internet of Medical Things (IoMT)," in 2021 IEEE Globecom Workshops, 2021, pp. 1–6. doi: 10.1109/GCWkshps52748.2021.9682118.
[6] P. Singh, G. Singh Gaba, A. Kaur, M. Hedabou, and A. Gurtov, "Dew-cloud-based hierarchical federated learning for intrusion detection in IoMT," IEEE Journal of Biomedical and Health Informatics, vol. 27, no. 2, pp. 722–731, 2023. doi: 10.1109/JBHI.2022.3186250.
[7] 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. 20th International Conference on Artificial Intelligence and Statistics, PMLR 54, 2017, pp. 1273–1282.
[8] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," Proceedings of Machine Learning and Systems, vol. 2, pp. 429–450, 2020.
[9] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, "SCAFFOLD: Stochastic controlled averaging for federated learning," in Proc. 37th International Conference on Machine Learning, PMLR 119, 2020, pp. 5132–5143.
[10] P. Kairouz et al., "Advances and open problems in federated learning," Foundations and Trends in Machine Learning, vol. 14, nos. 1–2, pp. 1–210, 2021. doi: 10.1561/2200000083.
[11] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in Theory of Cryptography Conference, LNCS 3876, 2006, pp. 265–284. doi: 10.1007/11681878_14.
[12] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in Proc. 2016 ACM SIGSAC Conference on Computer and Communications Security, 2016, pp. 308–318. doi: 10.1145/2976749.2978318.
[13] I. Mironov, "Rényi differential privacy," in 2017 IEEE 30th Computer Security Foundations Symposium, 2017, pp. 263–275. doi: 10.1109/CSF.2017.11.
[14] M. Bun and T. Steinke, "Concentrated differential privacy: Simplifications, extensions, and lower bounds," in Theory of Cryptography Conference, LNCS 9985, 2016, pp. 635–658. doi: 10.1007/978-3-662-53641-4_24.
[15] E. Bagdasaryan, O. Poursaeed, and V. Shmatikov, "Differential privacy has disparate impact on model accuracy," in Advances in Neural Information Processing Systems, vol. 32, 2019.
[16] L. Chen, X. Ding, Z. Bao, P. Zhou, and H. Jin, "Differentially private federated learning on non-IID data: Convergence analysis and adaptive optimization," IEEE Transactions on Knowledge and Data Engineering, vol. 36, no. 9, pp. 4567–4581, 2024. doi: 10.1109/TKDE.2024.3379001.
[17] N. Rieke et al., "The future of digital health with federated learning," npj Digital Medicine, vol. 3, art. 119, 2020. doi: 10.1038/s41746-020-00323-1.
[18] M. J. Sheller et al., "Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data," Scientific Reports, vol. 10, art. 12598, 2020. doi: 10.1038/s41598-020-69250-1.
[19] M. Adnan, S. Kalra, J. C. Cresswell, G. W. Taylor, and H. R. Tizhoosh, "Federated learning and differential privacy for medical image analysis," Scientific Reports, vol. 12, art. 1953, 2022. doi: 10.1038/s41598-022-05539-7.
[20] V. Mothukuri, R. M. Parizi, S. Pouriyeh, Y. Huang, A. Dehghantanha, and G. Srivastava, "A survey on security and privacy of federated learning," Future Generation Computer Systems, vol. 115, pp. 619–640, 2021. doi: 10.1016/j.future.2020.10.007.
[21] M. A. Ferrag, O. Friha, D. Hamouda, L. Maglaras, and H. Janicke, "Edge-IIoTset: A new comprehensive realistic cyber security dataset of IoT and IIoT applications for centralized and federated learning," IEEE Access, vol. 10, pp. 40281–40306, 2022. doi: 10.1109/ACCESS.2022.3165809.
[22] B. Efron, "Bootstrap methods: Another look at the jackknife," The Annals of Statistics, vol. 7, no. 1, pp. 1–26, 1979. doi: 10.1214/aos/1176344552.
[23] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," Psychometrika, vol. 12, pp. 153–157, 1947. doi: 10.1007/BF02295996.
[24] T. Saito and M. Rehmsmeier, "The precision–recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," PLOS ONE, vol. 10, no. 3, art. e0118432, 2015. doi: 10.1371/journal.pone.0118432.