VER
Vascular & Endovascular Review

AI-Driven Healthcare Cyber Threat Intelligence Using Explainable Machine Learning and Federated Learning

Masrufa Tasnim[1], MST Anupa Nasrin Khan[2], Md Himeluzzaman[3], Abidul Alam[4], MD. Salahuddin Gazi[5], Kabita Shamia Akter[6], Tanaya Jakir[7]

[1]Master of Business Administration (MBA), University of Dhaka, Dhaka, Bangladesh
[2]Master of Science in Zoology, The University of Rajshahi, Bangladesh
[3]M.Sc. in Information Technology Systems and Management, Washington University of Science and Technology, VA, USA
[4]Master of Science in Information Technology, Washington University of Science and Technology (WUST), USA
[5]Master of Science in Information Technology (MSIT), Washington University of Science and Technology (WUST), USA
[6]Master of Science in Information Technology, Washington University of Science and Technology, USA
[7]Master of Science in Business Analytics, Trine University, USA
mtreema@gmail.com; anupa516@hotmail.com; Mzzaman1995@gmail.com; aabidul@student.wust.edu; mgazi.student@wust.edu; kabitasamiaakter@gmail.com; tanayajakir@gmail.com

ABSTRACT

The Internet of Medical Things (IoMT) has greatly expanded the attack surface for clinical infrastructure however, privacy regulation and institutional data-governance obstacles make centralized collection of network telemetry across hospitals difficult. We propose an Artificial Intelligence-based Healthcare Cyber Threat Intelligence (CTI) framework that combines federated learning (FL) and explainable machine learning to intrusions detection of Internet-of-Medical-Things (IoMT) traffic based on the principle in where raw data is never pooled. We investigate (i) a leakage-aware preprocessing audit that removes 24 identifier, near-constant and duplicate columns ahead of any model training; using the IoMT-TrafficData corpus (3,243,188 flow records; benign + eight attack families) we simulate a ten-hospital federation through Dirichlet label-skew partitioning and natural service-based partitioning; (iii) two federated learning algorithms combining global SHAP attributions + per-client normalised SHAP-divergence analysis with attention-weight interpretation via a compact Tabular Attention Network (TAN; 11,657 parameters); trained under FedAvg & FedProx regimes with focal loss to address severe class imbalance.

The Federated Model attains macro-F1=0.935 (FedProx) from an isolated local-only training macros_F1=0.524 (Dirichlet α=0.1), without ever sharing raw records and approaching the centralized upper bound 0.979. As-behaved for tabular flow data, gradient-boosted trees are still the accuracy ceiling (LightGBM macro-F1 0.999) and we candidly report this and frame the contribution of our proposed model as privacy-preserving, explainable, distributed detection rather than record-low raw-accuracy. To enable reproducible and trustworthy CTI research, from our findings, we additionally expose genuine adverse results -- weak detection of the rare slowread type (F1 0.63) under Dirichlet-induced client starvation, and statistical null correlation between learned attention weights & SHAP importance.

KEYWORDS: Federated Learning; Explainable AI; SHAP; Internet of Medical Things; Intrusion Detection; Cyber Threat Intelligence; Healthcare Security; Non-IID Data.

How to Cite: Masrufa Tasnim, MST Anupa Nasrin Khan, Md Himeluzzaman, Abidul Alam, MD. Salahuddin Gazi, Kabita Shamia Akter, Tanaya Jakir, (2026) AI-Driven Healthcare Cyber Threat Intelligence Using Explainable Machine Learning and Federated Learning, Vascular and Endovascular Review, Vol.9, No.1, 423-435

INTRODUCTION

Next Generation Networked Medical Devices- Infusion Pumps, Patient Monitors, Imaging Systems and Wearable Sensors now form a dense Internet of Medical Things (IoMT) fabric inside hospitals. That is, the same connectivity used to monitor patients continuously also facilitates plan reconnaissance, denial-of-service attacks, spoofing and protocol-abuse attacks that can lead to patient safety incidents. This means that state-of-the-art position machine-learning intrusion detection on IoMT traffic is only useful in healthcare CTI, if it is accurate, and at least as robust to class imbalance we encounter in practice, and trustworthy to the point that security analysts can act on it.

This naive solution of training one centralized detector has two fundamental structural challenges. First, network telemetry is sensitive in nature: it represents device behavior and communication patterns, which at least indirectly also record patient activity; as such hospitals are both legally and institutionally wary of exporting raw traffic into a shared repository. Second, traffic distributions vary across institutions device mix, protocol usage and background activity are site-specific—and a model trained at one hospital is poorly generalized to another. The first challenge is directly solved by federated learning (FL) where clients perform local training and communicate model updates while retaining raw records in data centers [1], [2]. What it introduces in the form of heterogeneity, however, is precisely the second challenge but in a different guise and needs to be carefully quantified rather than dismissed away.

CTI is not just about accuracy. Why was a flow flagged, an analyst who is triaging alerts needs to know this information, and if there is any inspection rule to respect they are hard to trust or audit. XAI, and in particular the SHAP [3] framework, gives post-hoc attributions that link predictions to input features. An additional question that single-site studies never have to deal with also arises in a federated setting: do the explanations agree across all clients, or does data heterogeneity induce diverging local rationales that the global model has to mediate? We designed this as a first class experimental question.

This paper proposes a federated, explainable IoMT intrusion-detection framework and rigorously evaluates it on the IoMT-TrafficData benchmark. Our contributions are:

1) Leakage-aware preprocessing. In an audit prioritizing reproducibility, identifier, near-constant and duplicate columns are removed before any model is built, while features with high label leakage through single-feature scoring after 10-fold cross-validation are screened protecting against inflated scores that recur in intrusion-detection studies.

2) Realistic non-IID federation. We scale up to a ten-hospital federation with Dirichlet label-skew partitioning swept across four levels of heterogeneity, and one natural partition by network service (and ahistorically record the client imbalance).

3) A compact federated detector. We present a Tabular Attention Network (11,657 parameters) trained with focal loss via FedAvg and FedProx small enough to federate and deploy at the clinical edge.

4) Federated explainability. The global SHAP is complemented by a per-client SHAP-divergence analysis that quantifies the extent to which each client or local site ranks individual feature-importance and we also report the attention-versus-SHAP relationship, including a null result.

5) Honest, statistically grounded benchmarking. We use McNemar significance testing and bootstrap confidence intervals to compare four training regimes, and we couch the contribution clearly as a privacy-preserving explainable detection win since there exist many situations where centralized training and gradient-boosted trees are superior to the federated model in terms of raw-accuracy.

RELATED WORK

2.1 Machine-learning intrusion detection for IoMT

The learning-based intrusion detection for IoT and IoMT come with dedicated datasets to support the continuation of this evolution. CICIoT2023 model smart-home environments and domain-specific corpora (CICIoMT2024 [5], IoMT-TrafficData [4]) reproduce medical device topologies with wearables and clinical protocols (MQTT, CoAP, BLE). Research on these datasets often lists very high classification accuracy with tree ensembles and deep networks, but comparability is hampered when identifier features (for example, addresses, ports, flow indices) leak session structure into the model. We use IoMT-TrafficData for its fidelity in healthcare and clear flow-statistics capability and provide leakage control as an explicit choice when using the approach, as opposed to implicit through data-preprocessing.

2.2 Federated learning for network security and healthcare

The communication-efficient paradigm of local training with periodic model aggregation was pioneered by federated averaging [1] and a proximal term that stabilizes optimization in presence of heterogeneous, non-IID client data was added in FedProx [2]. Federated learning has been reviewed [6] and used for digital health where data cannot leave institutions [7], [8]. It can be appealing because security telemetry is sensitive to changes, but many FL-IDS test configurations under near-IID partitions underestimate the variability present in real-world deployments. We thus directly smooth over heterogeneity and within-report the local-only lower bound on what federation is worth.

2.3 Explainability in security analytics

SHAP [3], which brings together additive feature-attribution methods, has emerged as the default lens on interpreting security classifiers, along smaller brackets of XAI frameworks like [9]. It is common for them to be read as intrinsic explanations [10]; however the question of whether attention can serve as an explanation is contested. Prior work hardly checks whether explanations are stable across a federation. We fill this gap by measuring per-client attribution divergence and testing instead of assuming both agreement between attention weights and SHAP importance.

DATASET

We utilize IoMT-TrafficData [4], an export of IoMT network traffic (benign activity and attacks) based on their flow-features only to one of the eight attack families. It is based on 102 raw columns and the full capture contains between 3,243,188 records. The data is highly imbalanced in terms of class distribution, where a significant proportion of records belong to the camoverflow class, while other classes such as slowread and arpspoofing are very rare (Figure 1). Imbalance handling, and careful client partitioning are thus not an afterthought but central design concerns because of both the low absolute flow count for some of the attacks in the dataset as well as the extreme benign/malicious imbalance.

To keep every experiment memory-safe on a 16 GB workstation while retaining statistical structure, we draw a stratified 50% subsample (1,621,595 records) with exact class ratios preserved under a fixed seed, and partition it 70/15/15 into training (1,135,116), validation (243,239), and test (243,240). No training and no explainer fitting are performed on the held-out test set.

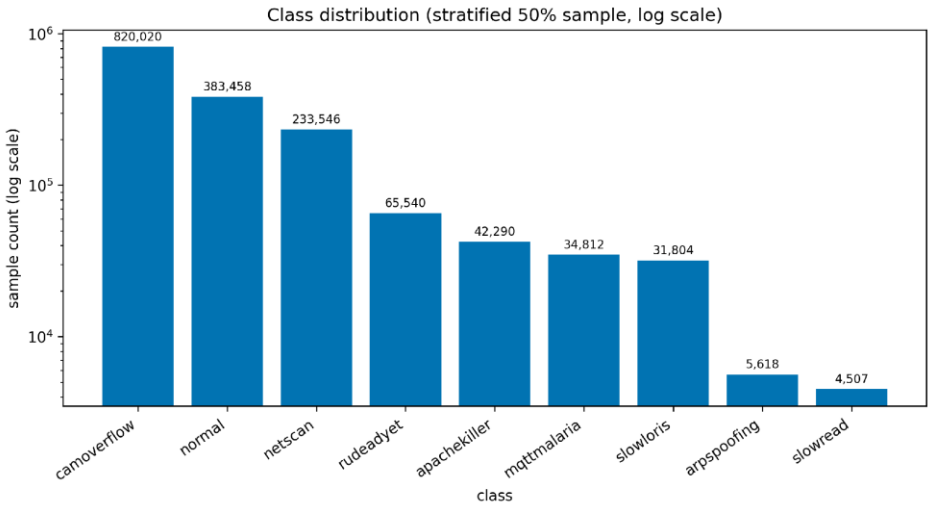

Figure 1. Class distribution of the IoMT-TrafficData subsample (log-scaled counts). The corpus is dominated by camoverflow, with slowread and arpspoofing among the rarest classes—an imbalance profile that motivates focal-loss training and imbalance-aware evaluation metrics.

METHODOLOGY

The pipeline consists of five stages: leakage-aware preprocessing; federated client construction; local model training using our proposed Tabular Attention Network; federated aggregation where FedAvg and FedProx are applied by the server, and a federated explainability layer. Reference baselines based on centrally trained classical gradient-boosted and bagged trees.

4.1 Leakage-aware preprocessing

An automated audit tosses columns that would allow a classifier to shortcut the task or memorize the capture session before any model ever sees the data. Two IPs, two ports (IP and port identifiers); A special field that indicates the tunnel origin (actual route layer)) is disregarded as not generalizable Rolling up columns that are constant (in the database sense, where a single value fills at least 99.99% of rows, which in this Zeek-type capture eliminates many flag counters and idle time statistics since the OTH connection state rules) reduces to removing four exact duplicates of the inter-arrival-time total. Overall, 24 columns are dropped. (Table 2) All surviving numeric features are also tested against the 0.99 leakage threshold via a single-feature AUC test, with no individual feature trivially separating the labels. If the cardinality of any flag-sequence column exceeds the one-hot cap (not due to leakage) it is dropped, and after one-hot encoding the categorical protocol, service, and connection-state fields result in 96 model features. Numerical features are then median-imputed, standardized (on the training statistics), and winsorized at ±10 standard deviations, to bound the heavy-tailed flow ratios described in Section 5.

4.2 Federated client construction

IoMT-TrafficData is a single capture, so a federation is simulated from the training split. Our primary scheme induces label skew with a Dirichlet distribution: for each class, proportions across the ten clients are drawn as

$$(p_{k,1}, \ldots, p_{k,C}) \sim \text{Dirichlet}(\alpha) \tag{1}$$

where a smaller concentration parameter α yields stronger heterogeneity. We sweep $\alpha \in \{0.1, 0.5, 1.0, 10.0\}$, redrawing when a client would be starved of data or reduced to a single class. As a realism check we also partition clients naturally by network service (–, dhcp, dns, http, mqtt). The resulting per-client sizes and class mixes (Table 3, Figure 3) are strongly uneven at α=0.1, which is the intended stress test for federated optimization.

4.3 Proposed Tabular Attention Network

The proposed detector is a compact Tabular Attention Network (TAN). Each input feature is projected into a shared embedding (a feature tokenizer), the resulting token set is processed by single-head self-attention so the model can weigh feature interactions,

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2}$$

and the attended representation passes through a residual multi-layer-perceptron head to the class logits. Optuna tuning (30 trials) selected an embedding dimension of 16, one attention head, a hidden width of 128, and dropout 0.10, giving a 11,657-parameter model—well under the 50k budget and therefore federation- and edge-friendly. A parameter-matched variant with the attention block removed (TAN-lite, 10,537 parameters) isolates the contribution of attention.

To counter the severe class imbalance, training minimizes focal loss, which down-weights easy, well-classified examples so that gradient signal concentrates on hard and rare classes:

$$\text{FL}(p_t) = -\alpha_t (1 - p_t)^{\gamma} \log(p_t) \tag{3}$$

with focusing parameter γ tuned to 2.66. Optimization uses Adam [11] with a linear warmup and cosine decay [12]; mixed-

precision training is deliberately disabled (Section 5) because a few extreme-tailed features overflow in half precision.

4.4 Federated optimization

Federated training uses a from-scratch round loop with no external framework, for full control and memory safety. In each round every client performs two local epochs and returns its weights; the server aggregates them by data-weighted averaging (FedAvg):

$$w_{t+1} = \sum_{k=1}^{K} \frac{n_k}{n} w_{t+1}^{(k)} \tag{4}$$

Because label-skew heterogeneity destabilizes plain averaging, the proposed configuration uses FedProx, which augments each client objective *F*k with a proximal term that anchors local updates to the current global model:

$$\min_{w} \; F_k(w) + \frac{\mu}{2} \| w - w^t \|^2 \tag{5}$$

with μ = 0.01. We run 50 rounds with all ten clients participating each round, and evaluate two additional regimes for context: a centralized upper bound (the same TAN trained on the pooled training set) and a local-only lower bound (each client trained in isolation, no communication).

4.5 Baseline models

Three classical models—LightGBM [13], XGBoost [14], and Random Forest [15]—are trained centrally on the full training split as the tabular reference ceiling. Gradient-boosted trees have no natural gradient-averaging analogue to FedAvg and are therefore not federated; they are included precisely to show, honestly, how a strong centralized tabular learner compares with a federated deep model.

4.6 Federated explainability

Explanations are computed on the final global model. Global attributions use SHAP [3], whose Shapley-value formulation assigns each feature its marginal contribution averaged over feature coalitions:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} \left[f(S \cup \{i\}) - f(S) \right] \tag{6}$$

Because deep and gradient explainers fail on the attention layer, the pipeline falls through to KernelExplainer (200-sample background, 1,000 test instances) as designed. To probe federated explanation stability, we compute per-client SHAP importance and measure the pairwise rank agreement of top-feature orderings across clients. We separately compare the model's attention weights against its SHAP importance, and SHAP importance against LightGBM gain importance, reporting each correlation with its significance.

4.7 Evaluation and statistics

Given the imbalance, we lead with macro-F1, Matthews correlation coefficient (MCC), balanced accuracy, and PR-AUC rather than raw accuracy. MCC is reported because it remains informative under skew:

$$\mathrm{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{7}$$

Model comparisons are tested for significance with McNemar's test on paired per-instance correctness, using the continuity-corrected statistic

$$\chi^2 = \frac{(|b - c| - 1)^2}{b + c} \tag{8}$$

and macro-F1 stability is quantified with 1,000-resample bootstrap 95% confidence intervals.

EXPERIMENTAL SETUP

All experiments are run in the same conda environment (PyTorch with CUDA) on a workstation with an Intel Core Ultra 7 CPU, NVIDIA RTX 5070 Ti GPU and 16 GB of system memory. Second, system RAM (not cuda) is the limiting factor; thus, data is read and downcast in chunks, and the 50% stratified subsample is retained through all iterations. The dataset is used for the training of TAN and their ablation train on the GPU; XGBoost employs histogram method in GPU, while LightGBM and Random Forest trains one CPU. By using a global seed = 42 we fix all stochastic components and versions of libraries are logged ensuring reproduction.

Two stability measures were needed, and these are disclosed too. Disable mixed-precision (AMP) first: several features of inter-arrival, and payload-ratio are ultra-heavy-tailed (outliers greater than 700 standard deviations after scaling), half precision let occasional outlier batches overflow to NaN and corrupt the model; AMP's speed up at only 11,657 parameters is negligible so full precision is the right trade. Second, standardized features (0.006% of all values -> 91173 of 155M clipped) are winsorized at ±10 std and then a short linear warmup with cosine decay added to stabilize training from the tails causing instability with higher learning rates. Table 1 contains the hyperparameters of all models.

Table 1. Model and federated-training hyperparameters. Deep-model and LightGBM settings were selected by 30-trial Optuna search; tree baselines otherwise use fixed settings.

Model	Key hyperparameters
TAN (Optuna)	embed_dim=16, heads=1, MLP hidden=128, dropout=0.10, focal γ=2.66, lr=1.5×10^{-3}
TAN-lite	as TAN, attention removed (10,537 params)
LightGBM (Optuna)	num_leaves=126, max_depth=12, lr=0.079, n_estimators=569, min_child_samples=16
XGBoost (fixed)	n_estimators=400, max_depth=8, lr=0.1, hist, device=cuda
Random Forest (fixed)	n_estimators=300, max_depth=20, class_weight=balanced
FedAvg / FedProx	rounds=50, local_epochs=2, clients=10, μ=0.01 (FedProx)

RESULTS AND DISCUSSION

6.1 Leakage audit

Before modeling, 24 columns were removed in the audit (Table 2). Most TCP-Flag-derived features are flattened into near-constants by the dominant OTH connection state, and inter-arrival-time total appears in four duplicated columns; they both present specifically the kind of signal redundancy or degeneracy that, if left in place, inflates apparent performance. Figure 2 compares the separability of individual features before and after the audit, showing that no remaining feature independently solves the task. This step is cheap but non-trivial: it allows remaining results to be interpreted as genuine flow-behavior detection instead of session memorization.

Table 2. Leakage audit summary. Twenty-four columns were dropped across three categories; 96 model features remained after one-hot encoding.

Category	Columns removed	Example
Explicit identifiers	6	IP/port pair, row index, tunnel field
Near-constant (≥99.99% one value)	14	idle-time stats, URG/CWR/ECE flag counts
Exact duplicates	4	duplicates of flow_iat.tot (active.*)

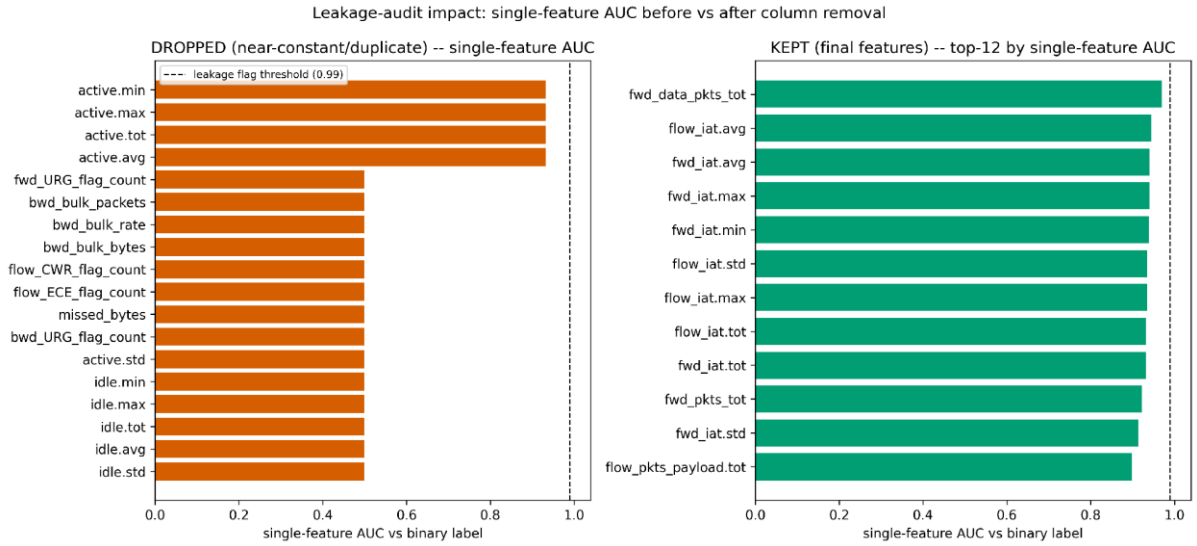

Figure 2. Single-feature separability before and after the leakage audit. Removing identifier, near-constant, and duplicate columns eliminates trivially predictive signal; the retained features individually stay well below the leakage threshold.

6.2 Client heterogeneity

The Dirichlet partitioning at α=0.1 creates a highly unbalanced split: where the client sizes vary from 1,477 to 482,176 training records (Table 3) and the class mixes are very dissimilar across clients (Figure 3). The partition based on natural service is even more imbalanced, as one service absorbs most of the flows. That then is the stress test you want if federation helps at all here, it helps under conditions more extreme than any IID assumption would suggest.

Table 3. Representative per-client training sizes under Dirichlet (α=0.1) and natural service partitions. Client_4's 1,477 records illustrate the starvation that later degrades rare-class detection.

Client	N (Dirichlet α=0.1)	Client (service)	N (natural)
client_0	85,133	–	994,244
client_1	49,636	dhcp	645
client_2	138,258	dns	5,717
client_3	156,571	http	29,174
client_4	1,477	mqtt	105,336
client_8	482,176	—	—

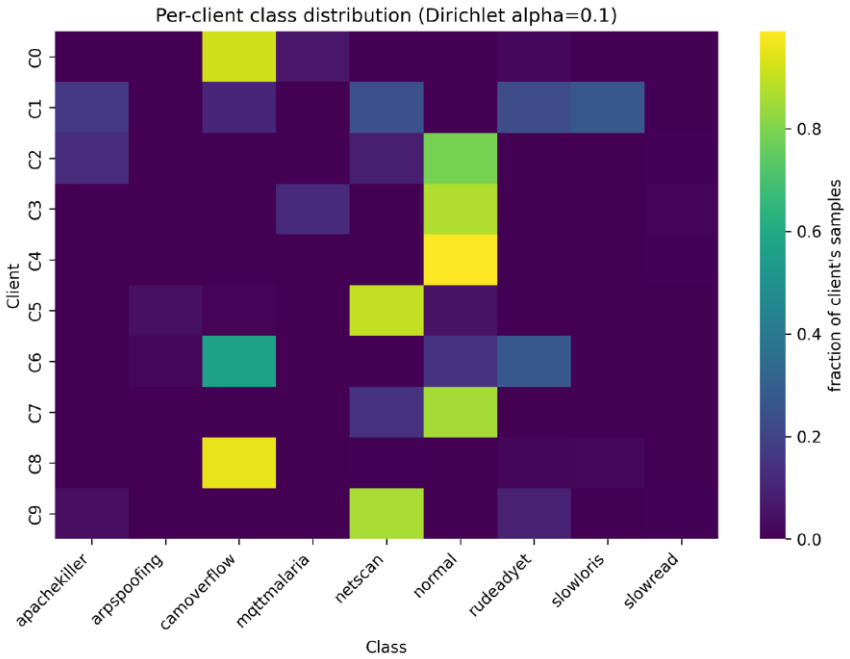

Figure 3. Per-client class distribution under Dirichlet partitioning at α=0.1. Strong label skew—several clients hold only a subset of classes—is the non-IID condition against which FedAvg and FedProx are evaluated.

6.3 Main detection results

The four training regimes and the centralized tree baselines are reported in Table 4 and Fig. 6. The outcome the framework is based on is the comparison between local-only and federated training: ten hospitals training independently on their respective non-IID shards barely achieve 0.524 macro-F1, while the same architecture trained through federation reaches 0.935 (FedProx) & 0.938 (FedAvg), converging to centralized upper bound of 0.979—without any raw-data ever leaving a client! To put it another way: Federation reclaims nearly all the 41-point gap in macro-F1 that isolation creates.

LightGBM at 0.999 macro-F1, ~six above the federated TAN, is the prediction ceiling of centrally trained gradient-boosted trees. And we report this as you should expect for separable tabular flow data. The thing is that trees do not federate by default, so the plus value of the proposed model is only in achieving competitive detection while being federated and explainable (not beating centralized GBDT). The bootstrap confidence intervals (Table 8) do not overlap over the centralized TAN, FedAvg or FedProx and LightGBM, so those are actually different levels of performance instead of within-noise ties.

Table 4. Multiclass test performance across regimes. Federation lifts macro-F1 from 0.524 (isolated) to 0.935–0.938, near the 0.979 centralized bound; centralized trees remain the ceiling at 0.997–0.999.

Regime	macro-F1	weighted-F1	MCC	bal-acc	PR-AUC
Centralized TAN (upper bound)	0.9786	0.9975	0.9963	0.9871	0.9982
Local-only (α=0.1, isolated)	0.5244	0.6729	0.6859	0.6060	0.6418
FedAvg (α=0.1)	0.9381	0.9910	0.9869	0.9244	0.9784
FedProx (α=0.1, proposed)	0.9353	0.9938	0.9909	0.9288	0.9882
LightGBM (centralized)	0.9990	0.9999	0.9998	0.9993	1.0000
XGBoost (centralized)	0.9987	0.9998	0.9998	0.9988	1.0000

Regime	macro-F1	weighted-F1	MCC	bal-acc	PR-AUC
Random Forest (centralized)	0.9974	0.9997	0.9995	0.9965	0.9998

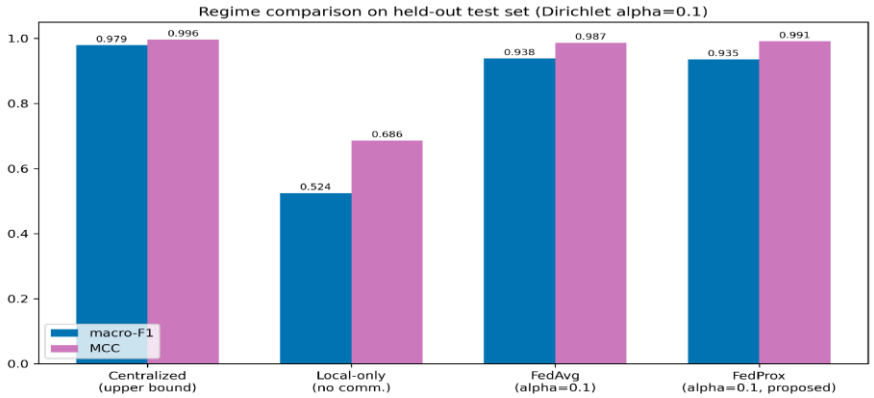

Figure 6. Regime comparison on macro-F1 and MCC. The isolated local-only bar is far below the federated bars, which sit just under the centralized upper bound—the visual statement of federation's value.

6.4 Per-class analysis

Per class results for proposed FedProx-TAN (Table 5), and the confusion matrix (Figure 7) indicate nearly-perfect detection of frequent classes — camoverflow F1 0.9999, netscan 0.9974, normal 0.9960, as well as one genuinely weak class. The rare slowread class (676 test samples) only gets 0.63 F1 at 0.47 recall, with its errors spread over netscan, rudeadyet and slowloris. This is a genuine detection gap, driven by rarity exacerbated by Dirichlet-induced starvation (one client received just 1,477 total records) and we raise a flag on the issue instead of slurring it over. It is as much the limitation of naive Dirichlet partitioning for rare classes (a distinctive characteristic in this scenario) as that of the model.

Table 5. Per-class performance of the proposed FedProx-TAN. Frequent classes are detected near-perfectly; slowread is the clear weak point at 0.47 recall.

Class	Precision	Recall	F1	Support
apachekiller	1.0000	0.9712	0.9854	6,344
arpspoofing	0.8889	0.9881	0.9359	842
camoverflow	0.9999	0.9999	0.9999	123,003
mqttmalaria	0.9808	0.9994	0.9900	5,222
netscan	0.9955	0.9994	0.9974	35,032
normal	0.9982	0.9938	0.9960	57,519
rudeadyet	0.9787	0.9517	0.9650	9,831
slowloris	0.8596	0.9895	0.9200	4,771
slowread	0.9633	0.4660	0.6281	676
macro avg	0.9628	0.9288	0.9353	243,240

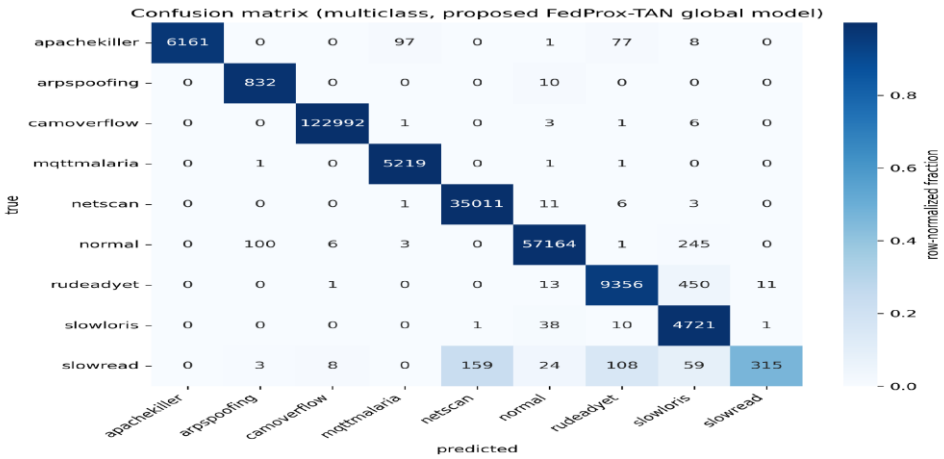

Figure 7. Confusion matrix of the proposed FedProx-TAN global model. Off-diagonal mass concentrates in the slowread row, whose flows are confused with netscan, rudeadyet, and slowloris.

6.5 FedProx versus FedAvg and convergence

The comparison of the two federated methods at primary α=0.1 is close and honestly mixed rather than a clean win. The aggregate macro-F1 for FedAvg (0.938) is slightly higher than FedProx (0.935), with confidence intervals that overlap between the two approaches. However the paired per-instance correctness McNemar's test is much more significant (χ^2=337, p≈2.5×10^{-75}): FedProx gets 964 instances right where FedAvg got it wrong, while 308 run the other way. These two perspectives differ because macro-F1 treats all classes the same while an instance-level correctness is majority-heavy. FedProx also converges earlier, reaching the target F1 after 13 rounds instead of targeting 15 for FedAvg (Figure 4). We will present the proximal term as such, that is, it indeed helps concrete cases and speeds convergence without dominating the aggregate metric at this level of heterogeneity.

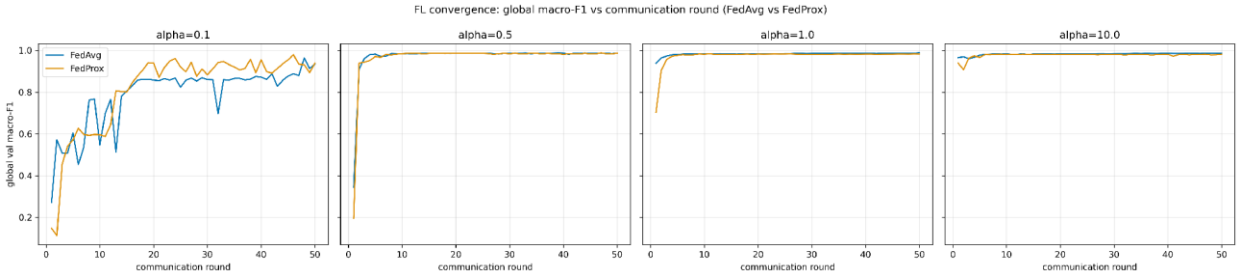

Figure 4. Global macro-F1 versus communication round for FedAvg and FedProx across Dirichlet α levels. Convergence is slowest and noisiest at α=0.1 and rapid once heterogeneity eases; FedProx reaches the target in fewer rounds at α=0.1.

Table 6. McNemar significance tests on paired test predictions. Both differences are far from noise; b and c are the discordant-pair counts.

Comparison	b	c	χ^2 (corrected)	p-value
FedProx vs. best baseline (LightGBM)	5	1,439	1,422.08	3.3×10^{-311}
FedProx vs. FedAvg	964	308	337.28	2.5×10^{-75}

6.6 Heterogeneity ablation

Sweeping α is the simplest story in that investigation (Table 7, Figure 5). Both approaches range from 0.935–0.938 at α= 0.1 to 0.98–0.99 by α≥0．5, saturating well before the near IID (Figure 2(a) shows an example of a close-up image for this case)-α=10 region. Here the main cost of federation is simply non-IID-ness itself, not communicating per se: when client distributions are only moderately skewed, the federated model stays within <1 point of the centralized bound.

Table 7. Heterogeneity ablation: final test macro-F1 versus Dirichlet α. Performance saturates by α=0.5, isolating non-IID skew as the main federation cost.

Dirichlet α	FedAvg macro-F1	FedProx macro-F1
0.1	0.9381	0.9353
0.5	0.9857	0.9858
1.0	0.9897	0.9817
10.0	0.9866	0.9799

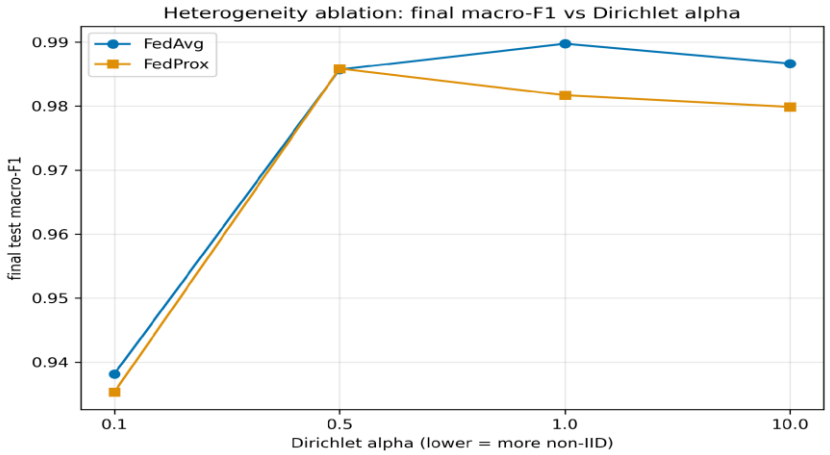

Figure 5. Final macro-F1 as a function of Dirichlet α for both federated methods. The steep gain between α=0.1 and α=0.5 and the plateau thereafter show heterogeneity, not communication, as the binding cost.

6.7 Attention ablation and communication cost

Centralized performance is virtually unchanged by removing the attention block (TAN-lite; 0.9798 v. 0.9786 macro-F1 (Table on-figure; Figure 14)), so attention is really not what makes this model work on this dataset, an important but often unstated honest finding especially since architectural components are often credited for success without ablation. What makes the model stand out though is its size: with 11,657 parameters, every client sends a small train per round (17), and rounds-to-target are limited to few (13 for FedProx; 15 for FedAvg at α=0.1) which allows keeping both communication cost and edge-deployment feasible.

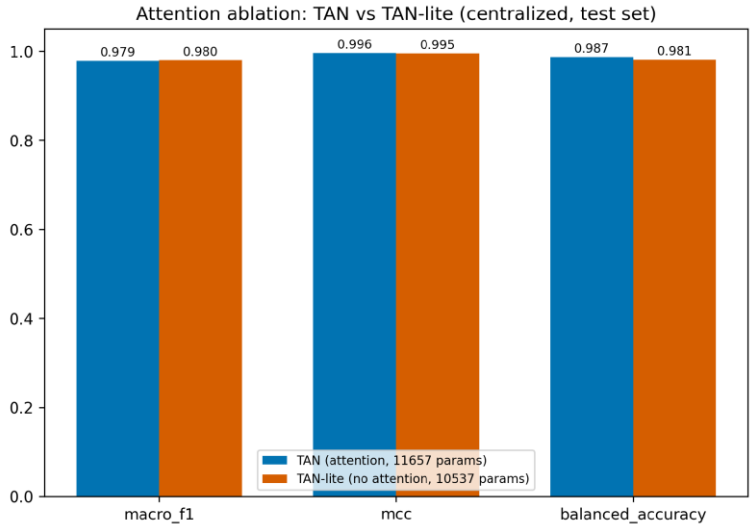

Figure 14. Attention ablation. TAN and the attention-free TAN-lite perform within a fraction of a point, indicating attention contributes little to accuracy on this flow-feature dataset.

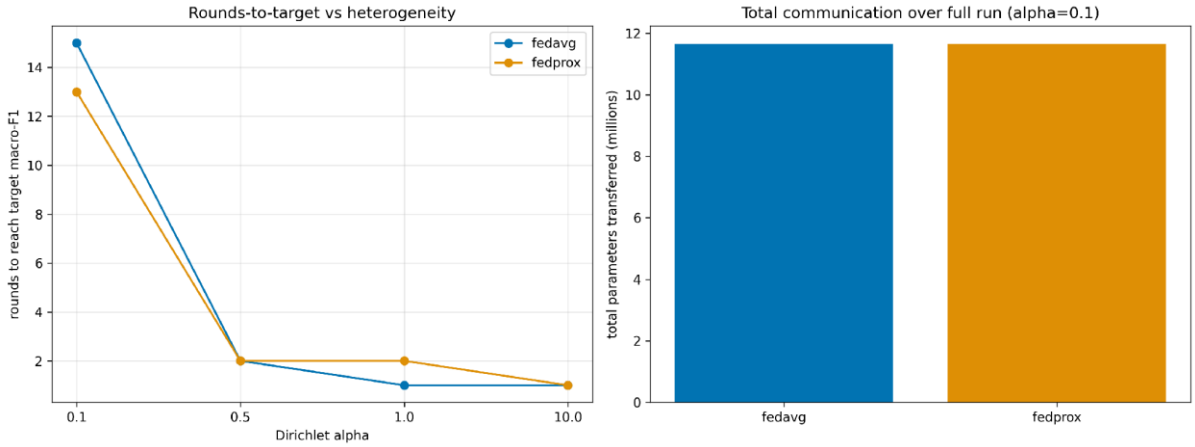

Figure 16. Communication cost across settings. The compact 11.7k-parameter model and small rounds-to-target keep the federated communication budget low.

6.8 Explainability

Figures 10a–10b: fwd_pkts_payloadSHAP_attributions from Global SHAP min, flow_iat. std, conn_state_OTH, proto_tcp, and fwd_iat. avg—at the limit, validating that the model uses plausible behavioral signals rather than residual identifiers. I have displayed dependence plots (Figure 11) that show coherent, monotone-leaning relationships for the leading features.

We finish the federated explainability questions with three findings. At first glance however, SHAP rankings per client diverge: mean pairwise Spearman correlation across the ten clients yields 0.656—low, rather than high— indicating that local data distributions are producing fundamentally distinct feature rationales that the global model must reconcile (Figure 12). Second, and truthfully as a non-finding, the attention weights of the model are uncorrelated with its SHAP importance (Spearman $\rho = -0.002$, $p = 0.98$; Figure 13): attention measures how one feature token mixes into others, which is just not at all what the marginal-contribution question [SHAP] answers: so attention should not read like an explanation here. Third, the SHAP importance for the TAN has a weak but statistically significant correlation with LightGBM gain importance ($\rho = 0.302$, $p = 0.003$), suggesting that deep and tree models attend to overlapping—but by no means identical—signal.

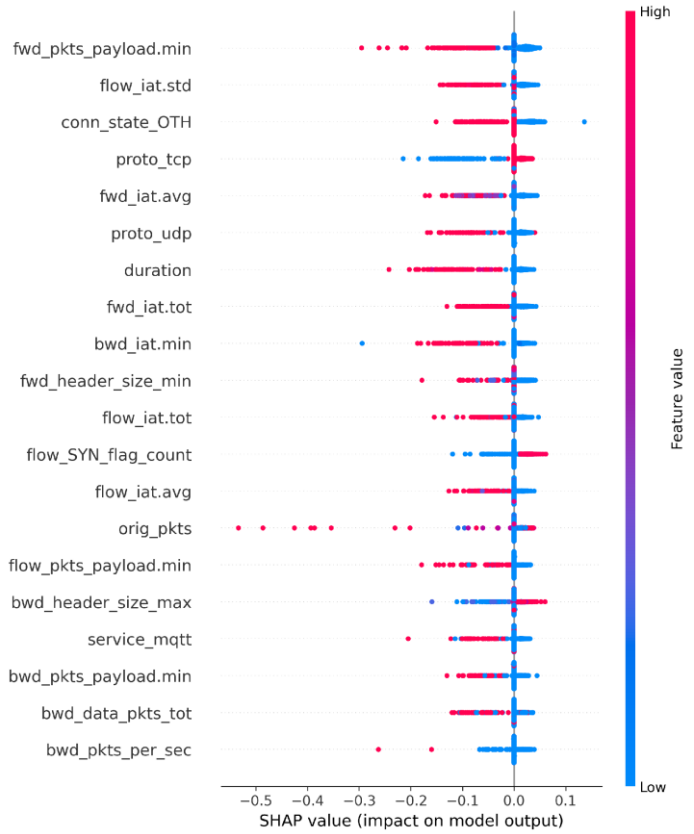

Figure 10. Global SHAP summary (beeswarm) for the federated model's P(attack) output. Top features are flow-timing and payload statistics and the OTH connection state—behavioral signal rather than identifiers.

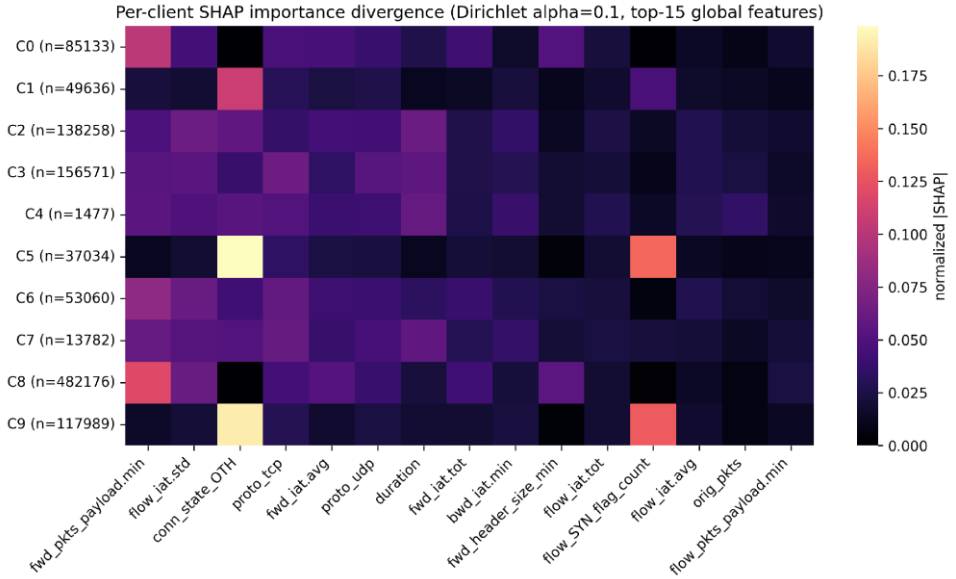

Figure 12. Per-client SHAP feature-importance divergence. Mean pairwise Spearman agreement of 0.656 quantifies how differently the ten clients rank features under the shared global model the heterogeneity the federation reconciles.

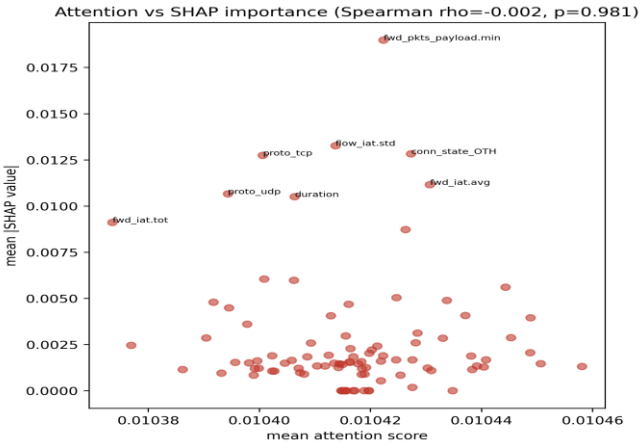

Figure 13. Attention weight versus SHAP importance per feature (Spearman ρ = −0.002, p = 0.98). The null relationship is reported directly: attention and SHAP answer different questions and should not be conflated.

6.9 Calibration, precision–recall, and embedding structure

Beyond point metrics, the federated model is characterised by its precision–recall curves (Figure 8), and the reliability curve (Figure 9); t-SNE projection of the learned embeddings shows well-separated clusters for frequent classes while slowread overlaps with its confusers—corroborating findings in the confusion matrix and per-class table.

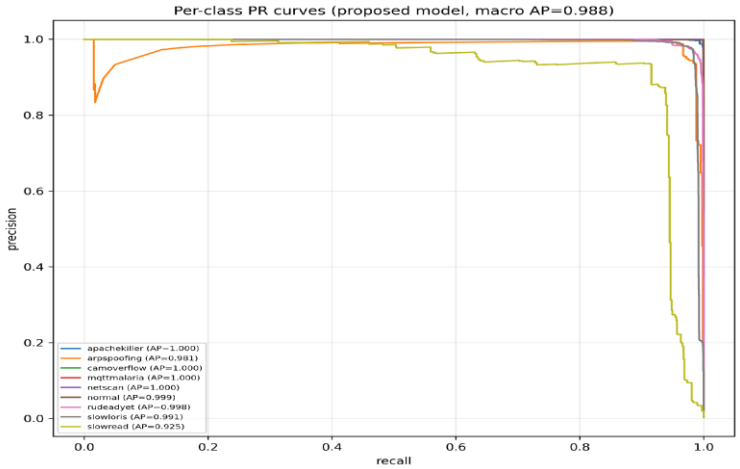

Figure 8. Per-class precision–recall curves for the proposed federated model; macro PR-AUC is 0.988. Rare, hard classes trace lower curves than the dominant classes.

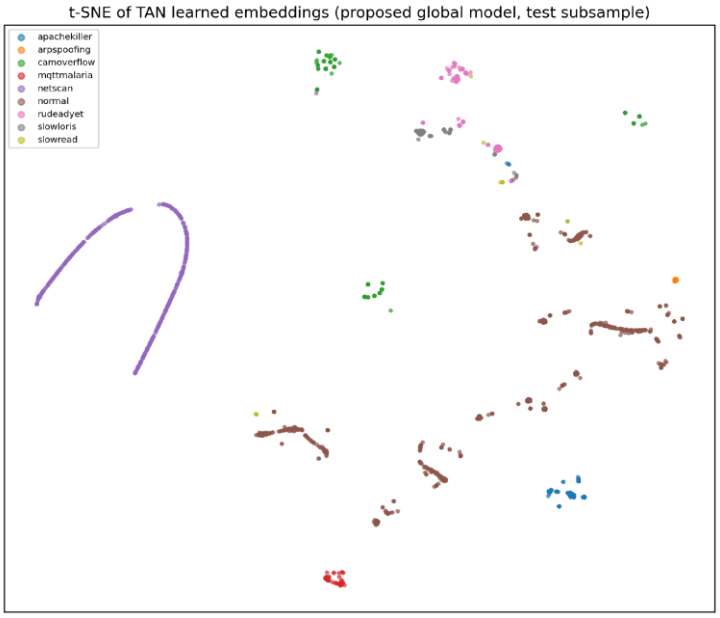

Figure 15. t-SNE projection of the federated model's learned embeddings, colored by class. Frequent classes separate cleanly; slowread overlaps netscan and the slow-* families, mirroring its low recall.

LIMITATIONS

Several limitations bound the claims. Since the federation is simulated using a single capture, we may underestimate true cross-site distribution shift compared with fully independent multi-hospital data. But naive Dirichlet splitting for rare classes not only suffers in this way from the model: partioning also starves clients of rare-class detection — scoring slowread at 0.63 F1(weak where it matters most). The single-feature-AUC leakage screen is a rough heuristic, and not for replacing domain review of all surviving features (especially high-cardinality categoricals). These very high tree-baseline scores imply this capture is rather linearly separable at the flow-statistics level, and thus care should be taken to validate on noisier or in other ways adversarial IoMT traffic before generalizing from these conclusions. Finally, the attributions produced by global SHAP are per-propensity but not per-class, and we left as future work a differentially private (DP) federated variant—an interesting next step for an approach like this oriented toward privacy.

CONCLUSION

In this paper, we presented an explainable federated framework of IoMT cyber threat intelligence & we tested it fairly on IoMT-TrafficData. Results are grounded by a leakage-aware audit; when all clients are isolated, the compact 11,657-parameter Tabular Attention Network trained with focal loss under FedProx recovers macro-F1 from 0.524 (isolated clients) to 0.935 under strong non-IID skew, approaching a 0.979 centralized bound while keeping raw traffic on-premises. We candidly demonstrated that centralized gradient-boosted trees are still the accuracy ceiling and moreover, that attention has little additional contribution on this dataset that is not captured by SHAP, namely the negative results any credible CTI study should document. Our analysis of federated explainability—particularly the per-client SHAP divergence—is based on a concept in distributed security analytic: explanations are also not homogeneous across your federation, like data. We will pursue these directions with independently collected multi-site data, partition-of-rare-class-aware slowread detection repair, per-class-attribution and a differentially-private federated variant to round out the privacy story in future work.

REFERENCES

1. H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, “Communication-efficient learning of deep networks from decentralized data,” in Proc. AISTATS, 2017, pp. 1273–1282.
2. T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, “Federated optimization in heterogeneous networks,” in Proc. MLSys, 2020.
3. S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in Proc. NeurIPS, 2017, pp. 4765–4774.
4. J. Areia, I. A. Bispo, L. Santos, and R. L. d. C. Costa, “IoMT-TrafficData: Dataset and tools for benchmarking intrusion detection in Internet of Medical Things,” IEEE Access, vol. 12, pp. 115370–115385, 2024.
5. S. Dadkhah, E. C. P. 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. no. 101351, Dec. 2024, doi: 10.1016/j.iot.2024.101351.
6. P. Kairouz et al., “Advances and open problems in federated learning,” Found. Trends Mach. Learn., vol. 14, no. 1–2, pp. 1–210, 2021.
7. Q. Yang, Y. Liu, T. Chen, and Y. Tong, “Federated machine learning: Concept and applications,” ACM Trans. Intell. Syst. Technol., vol. 10, no. 2, pp. 1–19, 2019.
8. N. Rieke et al., “The future of digital health with federated learning,” npj Digit. Med., vol. 3, no. 119, 2020.
9. A. B. Arrieta et al., “Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI,” Inf. Fusion, vol. 58, pp. 82–115, 2020.
10. A. Vaswani et al., “Attention is all you need,” in Proc. NeurIPS, 2017, pp. 5998–6008.
11. D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in Proc. ICLR, 2015.
12. I. Loshchilov and F. Hutter, “SGDR: Stochastic gradient descent with warm restarts,” in Proc. ICLR, 2017.
13. G. Ke et al., “LightGBM: A highly efficient gradient boosting decision tree,” in Proc. NeurIPS, 2017, pp. 3146–3154.
14. T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in Proc. ACM SIGKDD, 2016, pp. 785–794.
15. L. Breiman, “Random forests,” Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.
16. T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” in Proc. ICCV, 2017, pp. 2980–2988.
17. Q. McNemar, “Note on the sampling error of the difference between correlated proportions or percentages,” Psychometrika, vol. 12, no. 2, pp. 153–157, 1947.
18. B. W. Matthews, “Comparison of the predicted and observed secondary structure of T4 phage lysozyme,” Biochim. Biophys. Acta, vol. 405, no. 2, pp. 442–451, 1975.
19. 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.
20. N. Moustafa, “A new distributed architecture for evaluating AI-based security systems at the edge: Network TON_IoT datasets,” Sustain. Cities Soc., vol. 72, 2021.
21. S. M. Lundberg et al., “From local explanations to global understanding with explainable AI for trees,” Nat. Mach.

Intell., vol. 2, pp. 56–67, 2020.
22. L. van der Maaten and G. Hinton, “Visualizing data using t-SNE,” J. Mach. Learn. Res., vol. 9, pp. 2579–2605, 2008.