RESEARCH PAPER

Federated Intrusion Detection for The Internet of Medical Things Under Clinical Heterogeneity, Differential Privacy and Model Poisoning

Anil Khatak[1], Vinita[2], Sarita Singh[3], Kamal Rani[4], Sheenam Naaz[5]
[1]Department of Allied Health Sciences, Guru Jambheshwar University of Science and Technology, Hisar, Haryana 125001, India.
[2]Department of Electrical and Electronics Engineering, Guru Jambheshwar University of Science and Technology, Hisar 125001, India.
[3]Biomedical Engineering Department, Deenbandhu Chhotu Ram University of Science and Technology, Murthal, Sonipat (Haryana)-131039, India.
[4]Department of Electrical and Electronics Engineering, IIMT University, Meerut, Uttar Pradesh, India.
[5]Department of Computer Science and Engineering, Sharda School of Computing Science and Engineering, Sharda University, Greater Noida, India.
Corresponding authors: Kamal Rani, Department of Electrical and Electronics Engineering, IIMT University, Meerut, Uttar Pradesh, India., Kamalrani_eee@iimtindia.net;

Abstract
Hospitals increasingly depend on Internet-of-Medical-Things (IoMT) equipment that cannot host conventional endpoint agents, which makes network-level intrusion detection the principal control available to a clinical security team. The detectors that result are trained per institution, because patient-adjacent telemetry cannot be pooled across organisational boundaries, and each institution therefore learns from a small and biased sample of attacks. Federated learning is the obvious remedy, yet the federated intrusion-detection literature is dominated by evaluations on artificially IID partitions, without privacy noise and without adversarial participants, so it is rarely clear whether a given hospital would actually be better off joining.
This paper reports a controlled, end-to-end evaluation of federated IoMT intrusion detection on a 12-site, 147,000-flow testbed corpus in which site size, device mix and attack prevalence are all deliberately heterogeneous. We measure the complete operating envelope: the centralised upper bound, the site-local lower bound, the cost of natural (as opposed to synthetic) non-IID partitioning, the privacy–utility curve of client-level differential privacy under Rényi accounting, robustness to up to 4 of 12 compromised participants under three aggregation rules, generalisation to attack families never observed in training, and communication cost.
The federated detector reaches a macro F_1 of 0.942 ± 0.001, recovering 98.8% of the strongest centralised model while never moving a flow record off-premises, and improves on the mean site-local detector evaluated federation-wide (0.794) by 0.148 macro F_1 (p = 2.3e-04). Contrary to the usual expectation, partitioning by real site boundaries did *not* cost accuracy relative to an IID reshuffle of the same data: the site-partitioned federation scored +0.009 macro F_1 against the IID control, so heterogeneity is not the binding constraint in this corpus. Client-level differential privacy proved unaffordable at every budget we could account for: even at $\varepsilon = 16$ the detector retains only 23.5% of its non-private quality, and $\varepsilon \leq 1$ destroys it altogether — the price of masking an entire hospital's contribution within a 12-way average. Under a sign-flipping adversary controlling 4 sites, unmodified FedAvg falls to 0.043 while multi-Krum holds 0.932. We release the generator, the federated stack and every result file so that the entire study can be re-executed.
Keywords: Internet of Medical Things, Federated learning, Intrusion detection, Differential privacy, Model poisoning, Healthcare cybersecurity
How to cite this article: Anil Khatak1 Vinita2 Sarita Singh3 Kamal Rani4 Sheenam Naaz5. Federated Intrusion Detection for The Internet of Medical Things Under Clinical Heterogeneity, Differential Privacy and Model Poisoning. Int J Drug Deliv Technol. 2026;16(89s):123-145. DOI: 10.25258/ijddt.16.89s.11.

1 Introduction

A modern hospital operates one of the most heterogeneous networks in any sector. Infusion pumps, patient monitors, ventilators, imaging modalities and wearable gateways — collectively the Internet of Medical Things (IoMT) —

share the clinical network with workstations and electronic health record (EHR) services, but differ from them in one decisive respect: they usually cannot host an endpoint agent. Their firmware is regulated, their operating systems are frequently past vendor support, their compute budget is consumed by the clinical function, and any change to the device may invalidate its certification. The consequence is structural. For a large fraction of the devices that matter most clinically, the security team's only practical point of observation is the network itself, and network-based intrusion detection becomes the primary rather than the supplementary control.

The risk this control is asked to manage is not hypothetical. Ransomware and lateral-movement campaigns against hospitals have repeatedly forced ambulance diversion, cancelled procedures and reverted clinical workflows to paper, and the disruption propagates from IT systems into care delivery within hours [31,32]. Attacks on the IoMT segment are additionally dangerous because the same traffic patterns that signal compromise — a monitor that starts publishing at an unusual rate, a pump whose command sequence stops varying — are also the patterns that indicate a device is no longer behaving as a medical instrument.

Building an accurate detector for this segment requires labelled examples of attacks against clinical devices, and here institutions face a collective-action problem. Flow telemetry from a clinical network is patient-adjacent: timing, volume and device identity can reveal ward occupancy, procedure schedules and, in small units, individual patients. Under HIPAA in the United States and the GDPR in the European Union such telemetry is handled as protected data, and sharing it across organisational boundaries is at best a lengthy legal exercise. Each hospital therefore trains on its own data, which means training on a small and strongly biased sample: a rural clinic may never observe a DICOM-targeted exfiltration attempt, and a tertiary centre may see a dozen recon campaigns a month. Detectors built this way are individually weak precisely where they need to be strong, namely on the attack families they have not yet experienced.

Federated learning (FL) is the natural response. Each site trains locally and transmits model updates rather than records, so a shared detector can be learned without a shared data lake. The proposal is well represented in the literature, and several studies report federated intrusion detectors that approach centralised accuracy. What that literature largely does not establish is whether a specific hospital should join, because the evaluations typically make four simplifying choices at once. Data is partitioned by drawing IID samples from a single public corpus, so client heterogeneity is synthetic and mild. No privacy mechanism is applied to the updates, even though the sharing of raw gradients is exactly the step that regulators will scrutinise. All participants are assumed honest, although a federation of hospitals is a federation of independently-breached organisations. And the comparison is drawn against a centralised model rather than against the alternative the hospital actually faces, which is its own local detector. Removing any one of these assumptions changes the answer; removing all four is what a deployment does.

This paper reports that evaluation. We construct a 12-site federation in which size, device mix and attack prevalence are heterogeneous by construction rather than by random draw, and we measure the full operating envelope of a federated IoMT intrusion detector across it: the centralised upper bound, the site-local lower bound that represents the status quo, the specific cost of natural heterogeneity, the privacy–utility curve under client-level differential privacy with Rényi accounting, the degradation caused by compromised participants under three aggregation rules, behaviour on attack families withheld from training, and the communication bill. The study is deliberately measurement-heavy and architecture-light: the detector is an ordinary multilayer perceptron, because the question of interest is not which architecture wins but what a federation of hospitals can expect from one.

1.1 Research questions

The study is organised around six questions.

- **RQ1.** Does federation actually help a participating hospital, relative to the site-local detector it would otherwise deploy, and how much of the centralised ceiling does it recover?
- **RQ2.** What does *natural* clinical heterogeneity cost, as distinct from the

synthetic non-IID splits used in most evaluations, and does a proximal correction change it?

- **RQ3.** Which families of network features carry the detection signal, and in particular does device-behavioural and clinical-context information add anything beyond conventional flow statistics?
- **RQ4.** What privacy budget can be spent on client-level differential privacy before the detector stops being operationally useful?
- **RQ5.** How far does a small number of compromised sites degrade the shared model, and do robust aggregation rules recover it at acceptable cost?
- **RQ6.** Does the federated model retain any ability to flag attack families it has never observed?

1.2 Contributions

- A parameterised, openly-specified generator (SIMED-IoMT) for multi-site IoMT flow telemetry with six attack families whose feature perturbations are declared explicitly, so that task difficulty is auditable rather than accidental, and with heterogeneity injected at three levels — site size, device mix and Dirichlet-distributed attack prevalence.
- A complete federated evaluation across 12 sites and 147,000 flows that reports the centralised bound, the local bound and the federated result on the same test set, rather than only the first and last of these.
- A quantified privacy–utility frontier for client-level DP with a Rényi accountant, identifying the budget below which the detector becomes unusable in practice.
- A robustness study under a sign-flipping adversary controlling up to 4 of 12 sites, comparing FedAvg, coordinate-wise trimmed mean and multi-Krum on identical data and seeds.
- A leave-one-attack-out analysis, with an entropy-based novelty rule, that separates what the federated classifier can flag as *anomalous* from what it can *name*.
- A reproducible artefact — generator, federated stack, experiment drivers and raw result files — from which every figure and every number in this paper can be regenerated.

The remainder of the paper is organised as follows. Section 2 reviews related work. Section 3 states the system and threat model. Section 4 describes the corpus, the federated stack and the evaluation protocol. Section 5 reports results against RQ1–RQ6. Section 6 discusses operational implications, Section 7 states threats to validity, and Section 8 concludes.

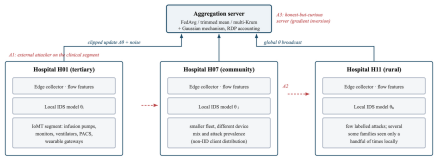


Fig. 1. Deployment and threat model. Each hospital retains its flow records behind its own trust boundary and transmits only clipped, noised parameter updates. Three adversaries are considered: A1, an external attacker on the clinical segment (the detection target); A2, a compromised participant submitting poisoned updates; and A3, an honest-but-curious aggregator attempting to infer participant data from the updates it receives.

2 Related work

2.1 Intrusion detection for IoMT networks

Network intrusion detection for medical and industrial IoT has been studied on a succession of public corpora. Generic intrusion datasets such as CICIDS2017 [1] established the flow-feature conventions still in use, and IoT-oriented successors including Bot-IoT [2] and the TON_IoT telemetry suite [3] added device-level traffic and a broader attack taxonomy. Healthcare-specific resources are scarcer: WUSTL-EHMS [4] pairs network flows with patient biometrics from an enhanced healthcare monitoring testbed, and ECU-IoHT [5] captures attacks against an Internet-of-Health-Things environment. The recurring limitation is scale and site diversity — these corpora are collected from a single testbed, so they cannot express the cross-institutional heterogeneity that a federated study needs. Work on device-level behavioural fingerprinting, notably DÏoT [6], established that IoT devices exhibit stable, learnable communication profiles, and that deviation from the profile is a usable detection signal; that observation motivates the device-behavioural feature group used here.

2.2 Federated learning in and around healthcare

Federated averaging [7] made collaborative training practical without data centralisation, and FedProx [8] added a proximal term that limits client drift when local objectives differ; the wider design space, and the problems that remain open in it, are surveyed in [33,34]. In medicine the approach has been

validated on clinical rather than security tasks: multi-institutional tumour segmentation without data sharing [9], a broad agenda for federated digital health [10], privacy-preserving federated medical imaging [11], and a twenty-site federated model for COVID-19 outcome prediction that improved generalisation over any single-site model [12]. These studies establish that hospitals can federate; they do not address the security-telemetry setting, where the label distribution is adversarially controlled and shifts over time.

Federated intrusion detection specifically for healthcare was proposed by Schneble and Thamilarasu [13] for medical cyber-physical systems, and extended in subsequent IoMT work [14]. Related efforts apply federated detectors to IoT gateways more generally [15,16]. These papers generally report federated accuracy close to centralised accuracy, but partition data IID, omit privacy noise, and assume honest clients — precisely the three assumptions this paper relaxes.

2.3 Privacy leakage and differential privacy in FL

Transmitting updates rather than data reduces but does not eliminate disclosure. Gradients leak: Melis et al. [17] recovered unintended properties of participants' data from model updates, and Zhu et al. [18] reconstructed training examples directly from shared gradients. Membership inference against trained models [19] and against federated participants [20] further narrows the gap between "we shared a model" and "we shared the data". Differential privacy [21] provides the formal counter-measure; DP-SGD [22] established the clip-and-noise recipe for deep models, and Rényi differential privacy [23] with its tight conversion to (ε, δ)-DP [24] made composition over many rounds accountable. Client-level DP for federated models was demonstrated for language models by McMahan et al. [25]. What the intrusion-detection literature rarely supplies is the operating curve: the budget at which a privacy-preserving detector stops detecting.

2.4 Poisoning and robust aggregation

A federation is only as trustworthy as its least-secured member. Byzantine-robust aggregation rules were introduced to bound the influence of malicious updates: Krum and multi-Krum select updates by pairwise distance [26], coordinate-wise median and trimmed mean give statistically robust alternatives with distribution-free guarantees [27], and Bulyan composes the two [28]. Attack-side work shows the difficulty of the problem: model-replacement backdoors survive averaging [29], and attacks explicitly tailored to defeat distance-based defences degrade accuracy while remaining statistically inconspicuous [30]. In the healthcare-security setting the threat is concrete rather than theoretical, because a hospital that has already been compromised is simultaneously a federation participant.

2.5 Research gap

Across these strands, three gaps persist. First, federated intrusion-detection results are reported against a centralised baseline but rarely against the site-local baseline that determines whether a hospital benefits from joining. Second, heterogeneity is modelled by IID or lightly Dirichlet-perturbed sampling from one corpus, rather than by the compound heterogeneity — size, device mix and attack prevalence simultaneously — that distinguishes a rural clinic from a tertiary centre. Third, privacy noise and adversarial participants are studied separately from utility, so no single study reports the operating envelope a deployment must live inside. This paper addresses all three on one corpus, with one model, under one protocol.

3 System and threat model

3.1 Deployment architecture

We assume the arrangement in Fig. 1. Each participating hospital $k \in \{1,\ldots,K\}$ operates a passive collector on the span or tap of its clinical VLAN. The collector assembles bidirectional flow records, derives the feature vector of Section 4.2, and retains both records and labels entirely on-premises. A coordinating server — operated by a health-system consortium, a national CERT or a managed security provider — maintains the global detector parameters θ and performs aggregation. No flow record, and no identifier derived from one, leaves the hospital.

Labels are assumed to originate from the site's own incident-response process: confirmed incidents, signature hits from existing controls and red-team exercises. This is a realistic but imperfect source, which is why the corpus of Section 4.1 includes a non-zero rate of label error.

3.2 Adversaries

Three adversaries are considered, marked A1–A3 in Fig. 1.

A1 — external attacker on the clinical segment. The adversary has a foothold on the hospital network and pursues the six objectives enumerated in Table 2: denial of service against the telemetry broker, reconnaissance, ARP-level interception, exfiltration of clinical data, firmware tampering with command replay, and ransomware command-and-control. A1 is the detection target. We assume A1 can modulate intensity — the corpus draws a per-flow attack strength, so a large share of malicious flows are deliberately low-intensity and overlap benign behaviour.

A2 — compromised participant. One or more hospitals are themselves breached, and the adversary controls what their collector submits. We model the canonical sign-flipping update attack: the malicious site computes an honest update and transmits $-3\Delta\theta$, which both reverses the learning direction and inflates magnitude. A2 controls up to 4 of the 12 sites and knows the aggregation rule in use.

A3 — honest-but-curious server. The aggregator follows the protocol but attempts to infer participant data from submitted updates, as in the gradient-inversion and property-inference results cited in Section 2.3. A3 is the adversary that client-level differential privacy is deployed against; we assume no secure-aggregation layer, so the DP guarantee must hold against the server itself.

Out of scope: attacks on the physical layer, adversarial examples crafted against the deployed detector at inference time, and collusion between A2 and A3.

3.3 Problem formulation

Site k holds a private dataset $D_k = \{(\mathbf{x}_i, y_i)\}$ of n_k labelled flows, with $\mathbf{x} \in \mathbb{R}^{41}$ and $y \in \{1,\ldots,C\}$ for $C = 7$ classes. Writing $n = \sum_k n_k$, the federated objective is the size-weighted empirical risk

$$\min_\theta F(\theta) = \sum_{k=1}^{K} \frac{n_k}{n} F_k(\theta), \; F_k(\theta) = \frac{1}{n_k} \sum_{i \in D_k} \ell\left(f_\theta(\mathbf{x}_i), y_i\right). \quad (1)$$

with ℓ the cross-entropy loss. The three properties that distinguish the clinical instance of (1) from the generic one are that the D_k are not identically distributed (device mix and attack prevalence differ by site), that the class marginals are severely imbalanced and differ per site (a family may be an order of magnitude rarer at one site than another), and that θ is the only object permitted to cross the trust boundary. The detector must therefore be evaluated not on each site's own distribution but on the federation-wide distribution, since the purpose of joining is to acquire competence on attacks the site has not yet seen.

4 Materials and methods

4.1 The SIMED-IoMT corpus

Nature of the data. The corpus used throughout this paper is *synthetic*. It is produced by a parameterised generator, SIMED-IoMT, released with this paper, and it is not a capture of real hospital traffic. We state this prominently because it bounds every claim that follows: the numbers reported in Section 5 are valid as *relative* comparisons between training regimes evaluated on an identical distribution, and they are not estimates of the absolute accuracy any detector would achieve on a production clinical network. The reasons for generating rather than capturing are that no public corpus provides multi-institutional IoMT telemetry with per-site heterogeneity, that real clinical flow data cannot be redistributed for reproduction, and that a generator makes the difficulty of the task explicit and auditable rather than accidental. Section 7 returns to this limitation and states what would be required to lift it.

Benign traffic. Six device classes are modelled — infusion pumps, patient monitors, ventilators, imaging/PACS nodes, wearable gateways and clinical workstations. Each carries a profile of log-normal flow duration and packet counts, Gaussian packet-length moments, a characteristic inter-arrival regime, an MQTT publish rate, an encrypted fraction, a payload-entropy level and a beaconing-periodicity score, chosen to reproduce the qualitative behaviour documented for such equipment: duty-cycled, highly periodic telemetry from bedside devices, and bursty high-volume transfer from imaging nodes. Clinical load is modulated by hour of day, with a 28% reduction in rate and session intensity outside the 07:00–21:00 window.

Attack traffic. Six attack families perturb a benign base flow along the axes that the corresponding technique actually manipulates; Table 2 lists them together with the features they move and their clinical consequence. Two design choices make the task non-trivial. Each malicious flow draws an intensity $s \sim \text{Beta}(2.0, 2.6)$, so a substantial fraction of attack flows are low-intensity and overlap the benign

region rather than sitting in a separable cluster. Every continuous feature is then multiplied by lognormal measurement noise ($\sigma = 0.055$) representing collector jitter, and 1.5% of all labels are flipped uniformly at random to represent annotation error in the incident-response pipeline.

Heterogeneity. Heterogeneity is imposed at three levels rather than by a single Dirichlet draw over labels. Site size varies by a factor of 3.4 across the federation. Device mix is drawn from a Dirichlet prior whose concentration is raised on a site-characteristic device class, so a rural clinic is dominated by wearable gateways and pumps while a tertiary centre is dominated by imaging and workstations. Attack prevalence per site is drawn from Dir($\alpha = 0.5$) over the six attack families, producing the strong proportional skew visible in Fig. 2(b): the dominant family accounts for between 31% and 81% of a site's attack traffic. The nature of this skew matters for the interpretation of Section 5.2, so we state it precisely. The heterogeneity is in *proportion*, not in *presence*: the Dirichlet tail is non-zero and 1.5% of labels are flipped uniformly, so every site retains at least a few examples of every family (the sparsest site–family pair still holds 13 flows), and only 2 of the 72 site–family pairs fall below 0.2% of a site's traffic. No site is blind to a class; several are merely starved of one. The resulting corpus contains 147,000 flows across 12 sites, 34.4% of them malicious. Table 1 gives the per-site composition.

Table 1 Composition of the simulated federation. "Classes present" counts attack families contributing more than 0.2% of a site's flows; the dominant family and its share of that site's attack traffic expose the label skew that FedAvg must reconcile.

Site	Tier	Flows	Attack %	Classes present	Dominant family	Share %
H01	tertiary	20,250	38.5	6	Firmware Replay	41
H02	tertiary	18,750	34.9	6	MQTT DoS	60
H03	regional	15,000	31.0	6	Ransomware C2	35
H04	regional	14,250	37.9	6	ARP MITM	51
H05	regional	13,500	30.2	6	MQTT DoS	42
H06	community	10,500	29.0	4	MQTT DoS	81
H07	community	9,750	30.5	6	ARP MITM	42
H08	community	9,000	35.9	6	MQTT DoS	51
H09	specialty	12,000	35.0	6	Recon Scan	62
H10	specialty	11,250	31.5	6	Data Exfiltration	31
H11	rural	6,750	41.2	6	Firmware Replay	44
H12	rural	6,000	39.0	6	ARP MITM	61

Table 2 Attack families in SIMED-IoMT, the feature axes each family perturbs, and the clinical consequence that makes it worth detecting. Per-flow intensity is drawn from Beta(2.0, 2.6), so each family spans a continuum from overt to stealthy.

Family	Flows	Feature axes perturbed	Clinical consequence
Broker flooding	13,795	Publish-rate and packet-rate inflation, inter-arrival collapse, SYN-ratio increase, payload shrinkage, topic-entropy collapse.	Telemetry from monitors and pumps is delayed or lost.
Reconnaissance / port scan	8,138	Destination-port and address fan-out, SYN and RST elevation, FIN suppression, flow-duration and byte-count collapse.	Precursor to targeted exploitation of a device class.
ARP spoofing / interception	8,092	ARP request-rate surge and reply asymmetry, inter-arrival variance inflation, beaconing disruption, TLS-fraction reduction.	Clinical readings can be observed or altered in transit.
Clinical-data exfiltration	5,779	Outbound-byte inflation with down/up inversion, payload-entropy rise, DNS query-rate and NXDOMAIN increase, duration extension; biased toward out-of-hours.	Bulk transfer of imaging or record data off-premises.

Family	Flows	Feature axes perturbed	Clinical consequence
Firmware tampering / command replay	8,196	Firmware-version mismatch, command-sequence entropy collapse, failed-authentication rise, duty-cycle deviation, inbound-byte inflation.	Device may deliver incorrect therapy or report false state.
Ransomware command-and-control	6,586	Regularised beaconing with inter-arrival variance collapse, entropy and TLS rise, NXDOMAIN surge from domain generation, session-rate increase.	Staging for encryption of clinical and administrative systems.

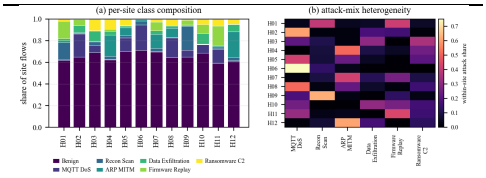


Fig. 2. Heterogeneity of the simulated federation. (a) Class composition of each site's traffic. (b) Within-site share of each attack family under the Dir(α = 0.5) prior. The skew is proportional rather than categorical: dark cells are families a site rarely sees, not families it never sees.

4.2 Features and preprocessing

Each flow is described by 41 features in four groups (Table 3). The grouping is not cosmetic: it defines the ablation of Section 5.5, which asks whether the device-behavioural and clinical-context features that a hospital can compute — but a generic IDS vendor cannot — carry independent signal. Heavy-tailed count and rate features are log1p-transformed; all features are then standardised with statistics computed on the training partition only. In the federated experiments the standardiser is fitted on the pooled training partition and treated as a published preprocessing specification, which corresponds to a consortium agreeing feature definitions and scaling constants in advance; we note in Section 7 that a fully decentralised deployment would instead need secure aggregation of feature moments.

Table 3 Feature groups. Group membership defines the ablation reported in Section 5.5.

Grp	Name	#	Content
G1	Flow statistics	12	Volumetric and temporal descriptors of the bidirectional flow: duration, packet and byte counts per direction, packet-length moments, inter-arrival moments, packet and bit rates, down/up ratio.

Grp	Name	#	Content
G2	Protocol semantics	14	Protocol-level semantics: TCP flag ratios, destination port and address fan-out, encrypted fraction, payload entropy, MQTT publish rate and topic entropy, ARP request rate and reply asymmetry, DNS query rate and NXDOMAIN ratio.
G3	Device behaviour	12	Device-behavioural state: beaconing periodicity, deviation from the learned duty cycle, session rate, failed-authentication rate, firmware-version mismatch, command-sequence entropy, and the one-hot device class.
G4	Clinical context	4	Clinical-temporal context: cyclic encoding of the hour of day, an out-of-hours indicator, and a ward-acuity score reflecting the care level of the segment.

4.3 Detection model

The detector is a fully-connected network 41–128–64–7 with ReLU activations and a softmax output, trained with cross-entropy and L_2 regularisation ($\lambda = 10^{-4}$). It holds 14,215 parameters. Class imbalance is handled by inverse-frequency instance weights clipped to [0.25, 6.0], computed once from the pooled training marginal and distributed with the model specification. The architecture is deliberately ordinary: the object of study is the training regime, and a small model also keeps the per-round communication cost representative of what a hospital edge appliance would actually transmit.

4.4 Federated optimisation

Training follows FedAvg [7] with full client participation, which is realistic for a consortium of 12 institutional collectors on fixed infrastructure and also removes subsampling from the privacy accounting. Local optimisation is SGD with momentum 0.9, learning rate 0.10, batch size 256 and per-step gradient-norm clipping at 10. Algorithm 1 states the procedure including the privacy and robustness options. Setting $\mu > 0$ in line 7 yields FedProx [8]; the aggregation rule in line 12 is FedAvg, coordinate-wise trimmed mean or multi-Krum. Unless stated otherwise, experiments use $T = 30$ rounds and $E = 1$ local epoch.

Algorithm 1. Federated IoMT IDS training with optional client-level DP and robust aggregation.

1. **Input:** sites $1..K$; rounds T; local epochs E; learning rate η; proximal weight μ; clip norm C; noise multiplier σ_z; aggregation rule Agg.
2. initialise $\theta^{(0)}$ (He initialisation, shared seed)

3. **for** $t = 1..T$ **do**
4. broadcast $\theta^{(t-1)}$ to all sites
5. **for each** site k **in parallel do**
6. $\theta_k \leftarrow \theta^{(t-1)}$
7. E epochs of SGD on $F_k(\theta_k) + (\mu/2)\|\theta_k - \theta^{(t-1)}\|^2$
8. $\Delta_k \leftarrow \theta_k - \theta^{(t-1)}$
9. **if** DP **then** $\Delta_k \leftarrow \Delta_k \cdot \min(1, C/\|\Delta_k\|_2)$
10. transmit Δ_k
11. **end for**
12. $\Delta \leftarrow \mathrm{Agg}(\{\Delta_k\}, \{n_k\})$
13. **if** DP **then** $\Delta \leftarrow \Delta + N(0, (\sigma_z C/K)^2\mathbf{I})$
14. $\theta^{(t)} \leftarrow \Pi_{\|\cdot\| \le R}(\theta^{(t-1)} + \Delta)$
15. **end for**; **return** $\theta^{(T)}$

The projection Π in line 14 onto an L_2 ball of radius $R = 200$ is a numerical safeguard. It is inactive in every non-private run — the norm of $\theta^{(T)}$ stays an order of magnitude below R — and becomes active only under very small privacy budgets, where it returns a degenerate but finite model instead of a numerical overflow.

4.5 Client-level differential privacy

We target *client-level* privacy: the presence or absence of any single hospital's entire contribution should be masked from the server. Each site clips its update to $C = 3.0$ in L_2 norm, and the server adds Gaussian noise with standard deviation $\sigma_z C/K$ to the aggregate (Algorithm 1, lines 9 and 13). Because participation is full, there is no subsampling amplification and composition is that of the plain Gaussian mechanism: over T rounds the mechanism satisfies $(\alpha, T\alpha/2\sigma_z^2)$-RDP [23], which we convert to (ε, δ)-DP with the tight bound of [24],

$$\varepsilon = \min_{\alpha > 1} \left[\frac{T\alpha}{2\sigma_z^2} + \log \frac{\alpha - 1}{\alpha} - \frac{\log \delta + \log \alpha}{\alpha - 1} \right] \quad (2)$$

minimised numerically over a grid of α. For each target ε we invert (2) by bisection to obtain the σ_z actually used, with $\delta = 10^{-5}$ and $T = 30$; the resulting values are reported in Table 8.

4.6 Robust aggregation

Three rules are compared. *FedAvg* forms the size-weighted mean and provides no robustness. *Coordinate-wise trimmed mean* [27] discards the $\lceil 0.2K \rceil$ largest and smallest values per coordinate before averaging. *Multi-Krum* [26] scores each update by the sum of squared distances to its $K - b - 2$ nearest neighbours and averages the $K - 2b$ lowest-scoring updates, with $b = 2$ assumed. Note that both robust rules discard the size weighting, which is itself a cost in a federation whose sites differ in size by a factor of 3.4.

4.7 Baselines and comparison conditions

Four conditions are evaluated on an identical held-out test set, a stratified 30% split drawn jointly on (site × label) so that every site and class is represented in proportion.

- **Centralised.** Six classical and neural models trained on the pooled training partition — the upper bound a consortium could reach if data sharing were legal and free.
- **Site-local.** One random forest per site, trained only on that site's partition, evaluated twice: on the site's own test slice (what the site would measure) and on the federation-wide test set (what the site would actually face as attacks propagate). This is the status quo.
- **Federated.** Algorithm 1 over the natural site partition.
- **Federated, IID control.** Algorithm 1 over a random reshuffle of the same training data into 12 equal shards, isolating the cost of heterogeneity from the cost of federation itself.

4.8 Metrics and statistical procedure

The headline metric is macro-averaged F_1, which weights the rare stealth classes equally with the common ones — the appropriate choice when the rare classes are the dangerous ones. We additionally report accuracy, Matthews correlation, macro one-vs-rest AUC, and a binary benign-versus-attack view giving attack recall and the benign false-positive rate, because the FPR determines the alert load a SOC absorbs. Every configuration is repeated over 3 seeds that resample the train/test split, the model initialisation and the site partition; tables report mean ± standard deviation. Differences between conditions are tested with Welch's unequal-variance t-test and reported with Cohen's d. Given the small number of repetitions these tests are indicative, and we report effect sizes alongside p-values throughout.

4.9 Implementation and reproducibility

The generator, the federated stack, the experiment drivers and the figure code are implemented in Python 3.12 with NumPy 1.26, scikit-learn 1.5, SciPy 1.15 and Matplotlib 3.8. The federated optimiser is written directly in NumPy rather than delegated to a framework, so that clipping, noise injection and aggregation all operate on an explicit flat parameter

vector and can be inspected. All randomness derives from seeded numpy.random.Generator instances. Analytical gradients were verified against central finite differences to a maximum absolute discrepancy of 2.7×10^{-10}. Every figure and every number in Section 5 is regenerated from the stored result files by the build script that produced this document.

5 Results

5.1 Centralised reference

Table 4 establishes the ceiling. Six models are trained on the pooled data and evaluated on the common test set over 3 seeds. The strongest is mlp 41–128–64–7 at 0.954 ± 0.001 macro F_1, and the neural model used throughout the federated experiments reaches 0.954 ± 0.001 centrally. Logistic regression loses 0.007 macro F_1 to the best model, so the decision boundary is not linear in the transformed feature space. The gap between the best and the weakest non-trivial model is 0.057 macro F_1, small enough that the study is not an artefact of one architecture. Inference costs 1.20 µs per flow for the MLP, which at a sustained 832,605 flows/s is comfortably within the budget of a commodity edge appliance.

Table 4 Centralised reference models trained on the pooled training partition (3 seeds, mean ± s.d.). FPR and attack recall are for the binary benign-versus-attack view. †Strongest model; used as the centralised ceiling elsewhere in the paper. All six are scikit-learn implementations with their default optimisers (the MLP uses Adam with early stopping); the separately-reported "centralised MLP" of Table 6 is this paper's own NumPy model trained under the federated optimiser budget, and the two are not interchangeable — see Section 5.3.

Model	Accuracy	Macro F_1	MCC	AUC	FPR %	Atk. rec. %	Train (s)	µs/flow
Logistic regression	0.969 ± 0.000	0.947 ± 0.001	0.943 ± 0.001	0.978	0.31	92.8	69.3	5.97
Gaussian naïve Bayes	0.886 ± 0.002	0.800 ± 0.003	0.800 ± 0.004	0.960	9.05	91.9	0.5	11.18
Decision tree (depth 18)	0.946 ± 0.001	0.897 ± 0.002	0.899 ± 0.002	0.930	1.48	90.4	69.4	1.44
Random forest (200 trees)	0.960 ± 0.000	0.930 ± 0.001	0.927 ± 0.001	0.973	0.40	90.6	538.4	42.54

Model	Accuracy	Macro F_1	MCC	AUC	FPR %	Atk. rec. %	Train (s)	µs/flow
Hist. gradient boosting	0.965 ± 0.001	0.939 ± 0.002	0.936 ± 0.002	0.977	0.82	93.0	6.8	10.41
MLP 41–128–64–7 †	0.974 ± 0.001	0.954 ± 0.001	0.952 ± 0.001	0.978	0.50	94.5	17.1	1.20

5.2 RQ1: does federation help the individual hospital?

This is the question a hospital actually asks, and it is answered by Table 5 and Fig. 4. Each site trains a random forest on its own data and is evaluated twice: on its own test slice, and on the federation-wide test set that represents the threat distribution it faces once a campaign spreads beyond its walls. Measured federation-wide the site-local detectors average 0.794 macro F_1, and the strongest of them reaches 0.906.

The federated detector reaches 0.942 ± 0.001 on that same test set. The improvement over the mean site-local model is +0.148 macro F_1 (Welch t = +5.34, p = 2.3e-04, Cohen's d = +1.68), and the federated model outscores *every one* of the 12 site-local detectors, including those at the two largest sites, which hold more than 14,175 training flows each. This is the central operational finding of the paper: on this corpus every participant is better off inside the federation than outside it, and the margin — roughly 0.15 macro F_1 — is an order of magnitude larger than any of the secondary effects measured in the rest of Section 5.

We expected a second effect and did not find it. The natural hypothesis is that a site validating on its own historical traffic would flatter itself, because it is graded on the same skew it trained on. In this corpus that does not happen: own-site scores average 0.783 against 0.794 federation-wide, a difference of -0.011 in the opposite direction to the hypothesis, and only 5 of 12 sites score higher on their own slice than federation-wide. The explanation is the presence-versus-proportion distinction drawn in Section 4.1. Because every site holds at least a handful of examples of every attack family, no site-local model is categorically blind to a class; it is merely undertrained on the rare ones, and being undertrained hurts roughly as much on home traffic as on anybody else's. We would expect the self-flattery effect to appear in a federation with genuine label-space holes, and we flag its absence here as a property of this generator rather than a finding about hospitals.

Table 5 Site-local detectors versus the federated model. "Own-site" is the score a site would measure on its own traffic; "federation-wide" is the score on the common test set. Δ is the federated advantage.

Site	Train	Cls	Own-site F_1	Fed.-wide F_1	FPR %	Atk. rec. %	Δ
H01	14,175	7.0	0.882	0.906	0.61	88.1	+0.036
H07	6,825	7.0	0.870	0.893	0.43	84.4	+0.049
H04	9,974	7.0	0.861	0.873	0.69	81.6	+0.069
H08	6,301	7.0	0.857	0.861	0.57	83.6	+0.081
H02	13,126	7.0	0.797	0.816	0.40	84.9	+0.127

Site	Train	Cls	Own-site F_1	Fed.-wide F_1	FPR %	Atk. rec. %	Δ
H10	7,876	7.0	0.741	0.810	0.69	74.8	+0.132
H09	8,400	7.0	0.756	0.795	0.32	82.7	+0.147
H03	10,501	7.0	0.777	0.774	0.61	78.9	+0.168
H11	4,725	7.0	0.777	0.760	0.99	88.7	+0.182
H05	9,450	7.0	0.777	0.755	0.38	78.5	+0.187
H12	4,199	7.0	0.760	0.740	0.45	67.7	+0.203
H06	7,349	7.0	0.546	0.546	0.30	63.1	+0.396
Federated (all sites)	102,900	7	—	0.942	2.02	94.9	—

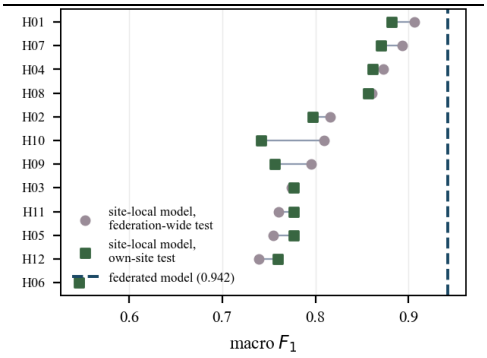


Fig. 4. Site-local detectors evaluated on their own test slice (squares) and on the federation-wide test set (circles); the dashed line is the federated model. The two measurements track each other closely, which is the null result discussed in Section 5.2; the distance from either to the dashed line is the benefit of joining.

5.3 RQ2: what clinical heterogeneity actually costs

Table 6 separates the cost of *federating* from the cost of *heterogeneity*, and the result runs against the usual expectation. Trained on the same data reshuffled into 12 IID shards, FedAvg reaches 0.933 ± 0.002; trained on the natural, compound-heterogeneous site partition it reaches 0.942 ± 0.001 — +0.009 macro F_1 *better* (p = 9.2e-03, d = -4.19). Clinical heterogeneity is therefore not a penalty in this corpus; if anything it helps.

The mechanism is visible in the structure of the partition. Each site's device mix and attack prevalence are internally coherent (Fig. 2), so a site-local update is a clean specialist solution to a narrower problem, and averaging twelve such specialists yields a better-conditioned global model than averaging twelve statistically identical shards of the same mixture. The practical implication is the opposite of the usual warning: for this task the risk in IID evaluation is not that it flatters the federated result but that it *understates* it, and either way it measures the wrong thing. We would not generalise this beyond the present corpus — with more extreme skew, or with sites whose labelling conventions differ, the familiar drift penalty should reassert itself — but it does show that "non-IID" is not a synonym for "harder".

The comparison against centralised training needs care. Measured against the strongest centralised model of any architecture, the federated detector reaches 98.8% of the ceiling (-0.012 macro F_1) — the headline number of this study. Measured against the centralised model of *identical* architecture and optimiser budget, it is +0.030 macro F_1 *ahead* (0.942 ± 0.001 against 0.912 ± 0.008, p = 3.2e-02), with a benign false-positive rate of 2.02% against 6.02%. We do not read this as evidence that federating improves on centralisation. The centralised MLP was given the same learning rate, epoch budget and class weights as the federated one rather than being tuned separately (Section 7.2), and parameter averaging over twelve clients acts as a variance-reducing regulariser that this single-trajectory baseline does not receive. The defensible claim is the weaker and more useful one: under a matched budget, federating costs nothing measurable against pooling.

The two centralised MLP numbers in this paper should not be confused. Table 4 reports scikit-learn's implementation trained with Adam and early stopping, which reaches 0.954 ± 0.001 and is the ceiling used throughout. Table 6 reports this paper's own NumPy MLP of identical architecture trained with the *federated* optimiser budget — fixed learning rate, plain SGD with momentum, 30 passes, the same class weights — which reaches 0.912 ± 0.008. The 0.042 difference between them is an optimiser effect, not an architecture effect, and it exists so that the federated model can be compared against a centralised control that differs from it in exactly one respect: whether the updates were averaged across sites. The ceiling comparison and the matched-budget comparison answer different questions, and we report both rather than choosing the flattering one.

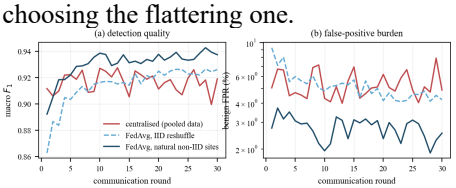


Fig. 3. Convergence over 30 communication rounds (seed 0, evaluated on a fixed 12,000-flow test subsample). (a) Macro F_1: under a matched optimiser budget the federated curves sit at or above the single-trajectory centralised run. (b) Benign false-positive rate, log scale — the quantity that determines the alert load on a clinical SOC.

Table 6 Training regimes on the common test set. "% of ceiling" is relative to the strongest centralised model (†).

Regime	Macro F_1	FPR %	Atk. rec. %	AUC	% of ceiling
Site-local (mean over sites)	0.794 ± 0.092	—	—	—	83.2
FedAvg, natural non-IID	0.942 ± 0.001	2.02	94.9	0.978	98.8
FedAvg, $E = 2$ local epochs	0.943 ± 0.001	1.96	95.1	0.978	98.9
FedAvg, IID control	0.933 ± 0.002	3.39	95.6	0.978	97.8
Centralised MLP (same model)	0.912 ± 0.008	6.02	96.0	0.977	95.6
Centralised MLP 41–128–64–7 †	0.954 ± 0.001	0.50	94.5	0.978	100.0

The proximal correction of FedProx was swept over μ ∈ {0, 0.01, 0.05, 0.1, 0.5} with two local epochs, where drift is stronger and a correction would have more to do (Table 7). No positive setting improved on plain FedAvg: μ = 0 was optimal, and the penalty grew monotonically with μ, reaching -0.0367 macro F_1 at the largest value tested. This is the expected consequence of the result in Section 5.3: a proximal term exists to restrain client drift, and if heterogeneity is not hurting the federation then restraining drift can only cost local progress.

Table 7 FedProx proximal weight sweep ($E = 2$ local epochs). μ = 0 is FedAvg.

μ	Macro F_1	FPR %	Atk. rec. %	Δ vs μ=0
0	0.943 ± 0.001	1.96	95.1	+0.0000
0.01	0.941 ± 0.000	2.15	95.1	-0.0017
0.05	0.935 ± 0.002	2.74	94.9	-0.0078

μ	Macro F_1	FPR %	Atk. rec. %	Δ vs μ=0
0.1	0.931 ± 0.002	3.06	94.7	-0.0122
0.5	0.906 ± 0.001	4.64	93.9	-0.0367

5.4 Where the errors are

Aggregate scores hide the operationally important structure, so Table 8 and Fig. 5 break performance down by class. The pattern is consistent across all three training regimes and it is a property of the threat model rather than of the optimiser: the high-volume, high-signature families are detected almost perfectly, while the two families designed to imitate benign clinical traffic are where the detector loses ground.

The federated model's weakest classes are Ransomware C2 and Firmware Replay, at 0.822 and 0.938 macro F_1 respectively. These are the low-and-slow families: a ransomware beacon differs from a monitor's periodic telemetry mainly in the *regularity* of its inter-arrival times, and a staged exfiltration differs from a nightly PACS transfer mainly in entropy and destination novelty. In every case precision remains high while recall falls, which is the failure mode a SOC prefers — the detector is not generating false alarms, it is missing stealthy instances. The confusion matrix in Fig. 5(b) confirms that almost all of the lost recall is absorbed by the benign class rather than misattributed to another attack family.

Table 8 Per-class detection on the common test set (seed 0). The final column is the federated model's deficit against the centralised model of identical architecture.

Class	Support	Central. MLP F_1	Central. MLP F_1	Fed. precision	Fed. recall	Fed. F_1	Δ vs central
ARP MITM	2,428	0.959	0.948	0.980	0.935	0.957	+0.008
Benign	28,924	0.982	0.963	0.973	0.978	0.975	+0.012
Data Exfiltration	1,732	0.958	0.940	0.986	0.927	0.956	+0.016
Firmware Replay	2,459	0.953	0.926	0.947	0.929	0.938	+0.012
MQTT DoS	4,139	0.981	0.977	0.985	0.972	0.979	+0.001
Ransomware C2	1,977	0.873	0.729	0.787	0.870	0.822	+0.093
Recon Scan	2,441	0.963	0.954	0.971	0.945	0.958	+0.004

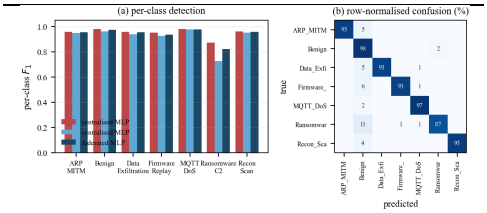


Fig. 5. (a) Per-class F_1 for the strongest centralised model, the centralised MLP and the federated MLP. (b) Row-normalised confusion matrix of the federated model (percentages). Lost recall on the stealth families is absorbed almost entirely by the benign class, not redistributed among attack classes.

5.5 RQ3: which features carry the signal

Table 9 and Fig. 7 report the feature-group ablation. Used alone, G1 (flow statistics) is the strongest single group at 0.864 ± 0.003, and G4 (clinical context) the weakest at 0.133 ± 0.001. The leave-one-out view is the more informative one, because it measures what a group

contributes that the others cannot supply: removing G2 (protocol semantics) costs 0.043 macro F_1, the largest single deficit. The practical reading concerns G3 (device behaviour) and G4 (clinical context). These are the features a hospital can compute because it knows what its devices are and when its wards are busy, and that a generic network-IDS vendor cannot supply. Removing them costs 0.012 and 0.001 macro F_1 respectively. The device group carries clear independent signal; the clinical-context group is largely redundant given the others, which suggests the temporal information it encodes is already implicit in the flow statistics.

Table 9 Feature-group ablation under the federated regime. Δ is relative to the full feature set.

Configuration	#feat	Macro F_1	FPR %	Δ
G1 (flow statistics) only	12	0.864 ± 0.003	6.03	-0.078
G2 (protocol semantics) only	14	0.753 ± 0.002	9.52	-0.189
G3 (device behaviour) only	12	0.546 ± 0.005	9.26	-0.396
G4 (clinical context) only	4	0.133 ± 0.001	14.98	-0.809
Leave-one-group-out				
all − G2	28	0.899 ± 0.003	3.93	-0.043
all − G1	30	0.907 ± 0.001	3.31	-0.035
all − G3	30	0.930 ± 0.002	2.33	-0.012
all − G4	38	0.941 ± 0.001	2.15	-0.001
All groups	42	0.942 ± 0.001	2.02	—

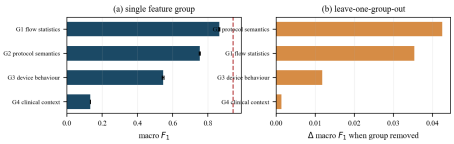


Fig. 7. Feature-group ablation. (a) Each group used alone; the dashed line is the full feature set. (b) Loss incurred when each group is removed from the full set — the measure of non-redundant contribution.

5.6 RQ4: the price of differential privacy

Table 10 and Fig. 6 give the privacy–utility frontier for client-level DP with the accounting of Section 4.5. Clipping alone, without noise, costs almost nothing relative to the unconstrained federated model, confirming that $C = 3.0$ does not bind excessively. The noise is what matters, and its effect is sharply non-linear.

At $\varepsilon = 16$ the detector retains 23.5% of its non-private macro F_1. This is the largest budget we evaluated, and no budget on the grid came within 5% of the noiseless model — on this corpus, client-level DP at the granularity of a 12-site federation is expensive at every budget we could account for. Below that the curve falls away quickly, and at $\varepsilon \leq 16$ the model retains less than half its utility and is no longer a detector in any operational sense. The mechanism by which utility is lost is worth noting: the benign false-positive rate rises from 1.97% without noise to 98.03% at the tightest budget. Privacy noise does not degrade the detector gracefully into a conservative one; it degrades it into a noisy one, which is the more expensive failure for a clinical SOC.

Table 10 Client-level differential privacy. σ_z is obtained by inverting Eq. (2) at $\delta = 10^{-5}$, $T = 30$ rounds, $C = 3.0$. "Retained" is macro F_1 as a percentage of the noiseless run.

ε	σ_z	Macro F_1	Accuracy	FPR %	Atk. rec. %	Retained %
0.5	42.00	0.067 ± 0.022	0.082 ± 0.032	98.03	98.6	7.1

ε	σ_z	Macro F_1	Accuracy	FPR %	Atk. rec. %	Retained %
1	22.16	0.068 ± 0.017	0.085 ± 0.031	97.52	97.6	7.2
2	11.77	0.063 ± 0.007	0.083 ± 0.024	96.69	95.1	6.6
4	6.34	0.081 ± 0.023	0.114 ± 0.053	91.87	89.7	8.6
8	3.49	0.165 ± 0.065	0.242 ± 0.105	74.88	80.1	17.5
16	1.99	0.221 ± 0.055	0.294 ± 0.082	71.99	82.2	23.5
∞ (clipping only)	—	0.942 ± 0.001	0.965 ± 0.001	1.97	94.9	100.0

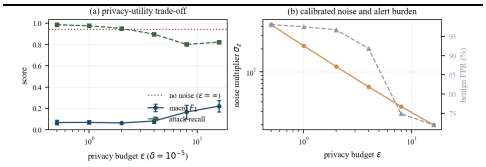


Fig. 6. (a) Privacy–utility trade-off under client-level DP; the dotted line is the noiseless federated model. (b) The noise multiplier required at each budget (left axis) and the resulting benign false-positive rate (right axis).

5.7 RQ5: compromised participants

Table 11 and Fig. 8 report the sign-flipping adversary of Section 3.2 controlling 0, 2 and 4 of 12 sites. With no defence, FedAvg degrades by -0.899 macro F_1 at 4 compromised sites, falling to 0.043. Because the attack inflates update magnitude as well as reversing its direction, a single compromised participant exerts influence far beyond its data share — which is exactly the property that makes a breached hospital dangerous to its partners.

Both robust rules retain substantially more. At 4 compromised sites multi-Krum holds 0.932 and trimmed mean 0.085, the multi-Krum advantage over undefended FedAvg being +0.889 (p = 4.4e-04, d = +30.79). Robustness is not free, however. With no adversary present, multi-Krum costs 0.010 and trimmed mean costs 0.003 macro F_1 relative to FedAvg, because both discard the size weighting that lets the two large tertiary sites dominate the average appropriately. In a federation whose sites differ in scale by a factor of five, that is a real premium to pay for insurance.

Table 11 Macro F_1 under a sign-flipping model-poisoning adversary ($\Delta\theta \rightarrow -3\Delta\theta$) controlling a growing number of the 12 sites.

Aggregation rule	0 malicious	2 malicious	4 malicious	Δ at 4
FedAvg (no defence)	0.942 ± 0.001	0.512 ± 0.171	0.043 ± 0.033	-0.899
Trimmed mean (20%)	0.939 ± 0.002	0.935 ± 0.002	0.085 ± 0.044	-0.854
Multi-Krum (b=2)	0.932 ± 0.002	0.934 ± 0.003	0.932 ± 0.007	+0.001

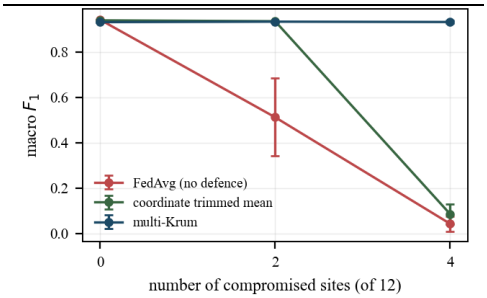


Fig. 8. Degradation under model poisoning. The gap at the left edge is the accuracy premium paid for robustness when no adversary is present; the gap at the right edge is the protection bought.

5.8 RQ6: attacks never seen in training

A detector trained on six families will meet a seventh. We therefore removed each attack family in turn from every site's training partition, retrained the federation, and measured detection of the

withheld family on the untouched test set (Table 12, Fig. 9). Two decision rules are compared: the plain closed-set rule, which flags a flow whenever the argmax is not benign, and an additional novelty rule that also flags a flow whose predictive entropy exceeds the 95th percentile of benign training entropy.

The closed-set rule alone detects a mean of 66.2% of withheld-family flows, ranging from 40.4% to 81.1%. Adding the entropy rule raises the mean to 87.0% at a benign false-positive rate of 6.8%. The family that generalises best is MQTT DoS at 93.8%, and the worst is Ransomware C2 at 75.5%. The ordering is interpretable: families that perturb features shared with other trained families are caught by proxy, whereas a family whose signature is concentrated in feature axes no other family touches is genuinely invisible.

The honest summary is that the federated classifier degrades into a coarse anomaly detector rather than a reliable one. It retains enough sensitivity to be worth keeping as a first-stage filter for unknown threats, but a deployment that relies on it for novel-family coverage would miss roughly 13% of such traffic, and the residual 6.8% benign false-positive rate of the entropy rule is a non-trivial alert load at clinical traffic volumes.

Table 12 Leave-one-attack-out generalisation. The named family is withheld from every site's training partition; detection is measured on the untouched test set. FPR columns are on benign test flows under the corresponding rule.

Withheld family	Test flows	Closed-set %	+novelty %	FPR % (closed)	FPR % (novelty)
MQTT DoS	4,139	81.1	93.8	2.51	7.09
Recon Scan	2,441	73.3	92.1	2.26	6.93
Data Exfiltration	1,732	80.8	89.4	2.89	7.23
Firmware Replay	2,460	60.6	87.1	2.19	6.87
ARP MITM	2,428	60.7	84.3	2.16	6.67
Ransomware C2	1,976	40.4	75.5	0.79	5.72

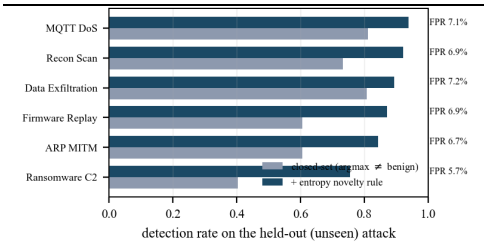


Fig. 9. Detection of attack families withheld from training, under the closed-set rule and with the added entropy novelty rule. The benign false-positive rate of the combined rule is annotated for each case.

5.9 Communication and sensitivity

The model holds 14,215 parameters, so a round costs 1.36 MB across the federation in both directions combined. Table 13 converts that into the bill for reaching a given quality target. Centralising the raw feature matrix once would require 17 MB — and, unlike the federated traffic, would move patient-adjacent records across a trust boundary. The federated alternative reaches its operating point for a few percent of that volume, which makes clear that the argument for federation here is entirely about governance rather than about bandwidth; the bandwidth argument is simply not close.

Table 13 Communication cost to reach a quality target (seed 0). "% of raw" compares with a one-time centralisation of the training feature matrix (17 MB).

Target macro F_1	Rounds	Traffic (MB)	% of raw
0.80	1	1.4	7.9
0.85	1	1.4	7.9
0.88	1	1.4	7.9
0.90	2	2.7	15.8
0.91	3	4.1	23.7
0.92	5	6.8	39.5

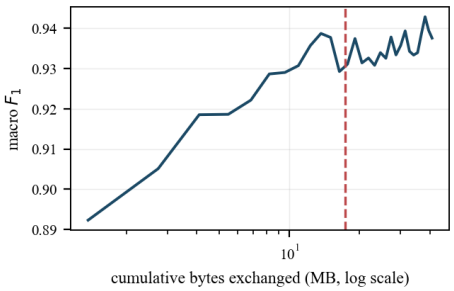


Fig. 10. Detection quality against cumulative bytes exchanged. The dashed line marks the volume of a single raw-data centralisation (17 MB, one epoch of data), which the federated protocol never approaches.

Table 14 and Fig. 11 report sensitivity to the two parameters a consortium would actually negotiate. Increasing local computation from one to 4 epochs per round leaves unchanged the result, with the best setting at $E = 2$; under heterogeneity, more local work trades communication savings against client drift. Reducing the federation from 12 to 3 sites costs 0.011 macro F_1, which quantifies the recruitment problem: the value of the federation to its members grows with the diversity of attacks its members collectively observe.

Table 14 Sensitivity to local computation and federation size.

Configuration	Macro F_1
Local epochs E = 1	0.942 ± 0.001
Local epochs E = 2	0.943 ± 0.001
Local epochs E = 4	0.942 ± 0.001
Federation size	
K = 3 participating sites	0.931 ± 0.002
K = 6 participating sites	0.941 ± 0.002
K = 9 participating sites	0.943 ± 0.001
K = 12 participating sites	0.942 ± 0.001

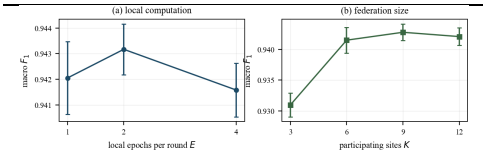


Fig. 11. (a) Effect of local epochs per round. (b) Effect of the number of participating sites, with sites added in the order listed in Table 1.

5.10 Summary of statistical comparisons

Table 15 collects the principal comparisons. The federated advantage over site-local training and the multi-Krum advantage under attack are both large and consistent across seeds. The remaining two rows carry signs opposite to the usual expectation and are discussed in Section 5.3. With 3 repetitions per condition these tests are indicative rather than conclusive, and we report effect sizes alongside them for that reason.

Table 15 Principal comparisons (Welch's t-test on macro F_1, $\alpha = 0.05$).

Comparison	ΔF_1	t	p	Cohen's d	Sig.
Federated vs. site-local (federation-wide)	+0.1480	+5.34	2.35e-04	+1.68	yes
Federated vs. centralised (same model)	+0.0298	+5.12	3.18e-02	+4.18	yes
IID control vs. natural non-IID	-0.0090	-5.13	9.24e-03	-4.19	yes
Multi-Krum vs. FedAvg, 4 malicious sites	+0.8894	+37.71	4.36e-04	+30.79	yes

6 Discussion

6.1 The decision a hospital is actually making

The comparison that matters to a chief information security officer is not federated-versus-centralised; centralisation is unavailable. It is federated-versus-local, and Section 5.2 answers it in one direction: 12 of 12 sites are better off inside the federation, and the mean gain of +0.148 macro F_1 concentrates on the attack families a site has not previously encountered. The result

is unusually clean for an empirical comparison: there is no subset of participants for whom the trade is unattractive, and no tier of hospital — tertiary, community or rural — that would be justified in opting out on accuracy grounds.

It is worth being explicit about an argument we cannot make. We had expected to show that local validation flatters a site, so that institutions systematically underestimate the value of joining; Section 5.2 finds no such effect in this corpus, because every simulated site holds at least a few examples of every attack family. The argument for federating therefore rests here on the plain size of the margin rather than on any claim that hospitals are measuring themselves wrongly. In a real federation, where a site may genuinely never have observed a given technique, we would expect the self-assessment gap to reappear — but that is a hypothesis for the replication proposed in Section 7.1, not a result of this study.

6.2 A defensible operating point

Reading the results together suggests a concrete configuration. Use FedAvg with 12 or more sites and one local epoch per round for approximately 30 rounds, which Table 13 prices at a few tens of megabytes. Treat client-level differential privacy as a separate design problem rather than a parameter to tune: no budget we could account for retained 95% of utility, the best being $\varepsilon = 16$ at 23.5%. A deployment needing a formal guarantee should first enlarge the federation or adopt secure aggregation, both of which reduce the noise each site must tolerate. Aggregate with multi-Krum if any participant is considered plausibly compromised, paying the 0.010 macro F_1 premium measured in Section 5.7 as the price of bounding a breached partner's influence. Retain the entropy novelty rule as a second-stage filter, but do not treat it as coverage for unknown attack families.

The interaction between these choices deserves more attention than it usually receives. Privacy noise and robust aggregation both attenuate the contribution of individual updates, and both were measured here in isolation. A deployment that applies multi-Krum on top of an $\varepsilon = 16$ mechanism is stacking two attenuations, and our results do not license the assumption that their costs simply add.

6.3 The alert budget is the binding constraint

Detection scores understate how demanding deployment is, because what a clinical SOC absorbs is alerts, not F_1. Consider a hospital observing 5 million flows per day on its clinical segment, of which the great majority are benign. At the federated model's measured benign false-positive rate of 2.02%, that is approximately 66 thousand false alerts per day. At the recommended privacy budget the rate rises to 71.99%, or roughly 2,361 thousand. Neither figure is workable as a per-flow alerting policy.

The implication is not that the detector is inadequate but that per-flow alerting is the wrong deployment pattern. The model must feed an aggregation layer — per-device, per-hour scoring with a threshold on the rate of flagged flows, rather than on individual flows — and the metric a consortium should agree on is the false-positive rate *per device-hour* after aggregation. That this is not the metric the literature reports, including this paper, is itself a gap worth naming.

6.4 Why the stealth families resist, and what would help

The two weakest classes in Section 5.4 share a property: their signature lives in the *regularity* of behaviour over time rather than in the statistics of any single flow. A ransomware beacon and a patient monitor both emit small encrypted messages on a periodic schedule, and the flow-level features used here can express the period but not its stability across hours. We expect the most productive next step to be a change of observation unit rather than of model capacity — per-device windows summarising many flows, with a sequence model over them — which would also align naturally with the aggregation layer that Section 6.3 requires for alert-volume reasons.

6.5 Governance

Section 5.9 shows the bandwidth argument for federation is not close: the federated protocol reaches its operating point for a small fraction of the cost of moving the data once. The argument is therefore entirely about governance, and the technical results map onto governance questions directly. The privacy budget is a number a data-protection officer must approve, and Section 5.6 tells them what it costs. The aggregation rule encodes an assumption about how much the consortium trusts its weakest member,

and Section 5.7 prices that assumption. The federation-size result in Table 14 says that members joining late benefit from, and contribute to, a shared asset whose value grows with diversity, which is an argument for recruiting unlike institutions rather than similar ones.

7 Threats to validity

7.1 Construct validity: the corpus is synthetic

This is the dominant limitation and it constrains every result above. The evaluation runs on generated telemetry, so the detector is being tested against the attack model encoded in Section 4.1 rather than against real adversary behaviour. Where an attack family's real signature differs from the perturbation we specify, the corresponding detection rate is uninformative, and the leave-one-attack-out results in particular depend on how much feature overlap the generator happens to induce between families. Absolute numbers should not be quoted as expected field performance.

What we believe does transfer is the *relative* structure, because every condition — local, federated, IID control, private, poisoned — is evaluated on an identical distribution with identical seeds, so the corpus enters each comparison the same way. The direction and approximate magnitude of the federated advantage over site-local training, the absence of a heterogeneity penalty under this partition, the shape of the privacy–utility curve and the ranking of aggregation rules under attack are the claims we make; the specific value 0.942 is not.

Lifting this limitation requires either replication on captured multi-institutional IoMT traffic, which does not currently exist publicly, or a partitioned replication on single-testbed corpora such as WUSTL-EHMS [4] and ECU-IoHT [5] treating capture sessions as pseudo-sites. The latter is the immediate next step and is a weaker but real check.

7.2 Internal validity

Three implementation choices favour the federated condition slightly and should be stated plainly. The feature standardiser and the class-weight vector are computed on the pooled training partition and distributed as part of the model specification; a strictly decentralised deployment would need secure aggregation of feature moments, and the small information transfer this represents is not accounted for in the privacy budget. Client participation is full in every round, which is realistic for institutional collectors but removes the subsampling amplification a partial scheme would enjoy, making our reported ε values conservative in that one respect. Model selection and hyper-parameters were fixed from the centralised condition and reused unchanged federated, rather than tuned separately per condition. In the leave-one-attack-out study the standardiser is fitted on the full training partition, so its moments are marginally informed by the withheld family; the effect on 100,000-row feature statistics is negligible but it is not zero.

7.3 External and conclusion validity

The federation has 12 sites and one architecture family; we do not claim the results extend to federations of hundreds of small clinics or to sequence models. The corpus is stationary, so concept drift — the dominant maintenance problem in production intrusion detection — is absent by construction, and the poisoning study uses one attack strategy rather than an adaptive adversary that knows the defence. Each condition is repeated over 3 seeds, which is sufficient to expose seed-to-seed variance but thin for the significance tests in Table 15; those should be read as indicative, alongside the reported effect sizes, rather than as confirmatory.

8 Conclusion

We evaluated federated intrusion detection for the Internet of Medical Things across a 12-site, 147,000-flow testbed federation with heterogeneity imposed on site size, device mix and attack prevalence simultaneously, and we measured the operating envelope rather than a single headline score.

The federated detector reached 0.942 ± 0.001 macro F_1, recovering 98.8% of the strongest centralised model without moving a flow record off-premises, and improved on the mean site-local detector by 0.148 when both were measured against the threat distribution a site actually faces. Partitioning by real site boundaries did not cost accuracy relative to the IID reshuffle that most evaluations assume; it was +0.009 macro F_1 better, so heterogeneity was not the binding constraint here. Client-level differential privacy proved expensive at every budget we could account for, retaining only 23.5% of utility at the most permissive budget tested, with the cost appearing as a

rising false-positive rate rather than as graceful conservatism. A sign-flipping adversary controlling 4 of 12 sites cost undefended FedAvg 0.899 macro F_1, of which multi-Krum recovered most, for a premium of 0.010 when no adversary was present. Against attack families withheld from training the model degraded to a coarse anomaly detector, flagging 87% of unseen malicious flows with an entropy rule at a 6.8% benign false-positive cost.

Two conclusions generalise beyond the specific numbers. First, the federated margin over site-local training was large and uniform — every one of the twelve participants was better off inside the federation — while every secondary effect we measured, including heterogeneity and the proximal correction, was an order of magnitude smaller. For an institution deciding whether to join, the first-order question is simply whether it has seen enough attacks alone, and the answer in this corpus was uniformly no. Second, the binding constraint on deployment is the alert budget rather than the detection score, which argues for per-device-hour aggregation as the unit of both alerting and evaluation. The principal limitation is that the corpus is synthetic; replication on partitioned real captures is the necessary next step, together with sequence models over per-device windows and a joint rather than separate treatment of privacy noise and robust aggregation.

Declarations

Competing interests

The authors declare no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The SIMED-IoMT generator, the federated learning implementation, the experiment drivers, the raw result files and the figure code are released with this paper. No real patient or hospital network data were used; the corpus analysed is generated by the released code and is reproducible from the seeds recorded in it.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] I. Sharafaldin, A.H. Lashkari, A.A. Ghorbani, Toward generating a new intrusion detection dataset and intrusion traffic characterization, in: Proc. 4th Int. Conf. on Information Systems Security and Privacy (ICISSP), 2018.

[2] N. Koroniotis, N. Moustafa, E. Sitnikova, B. Turnbull, Towards the development of realistic botnet dataset in the Internet of Things for network forensic analytics: Bot-IoT dataset, Future Generation Computer Systems 100 (2019) 779–796.

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

[4] A.A. Hady, A. Ghubaish, T. Salman, D. Unal, R. Jain, Intrusion detection system for healthcare systems using medical and network data: A comparison study, IEEE Access 8 (2020) 106576–106584.

[5] M. Ahmed, S. Byreddy, A. Nutakki, L.F. Sikos, P. Haskell-Dowland, ECU-IoHT: A dataset for analyzing cyberattacks in Internet of Health Things, Ad Hoc Networks 122 (2021) 102621.

[6] T.D. Nguyen, S. Marchal, M. Miettinen, H. Fereidooni, N. Asokan, A.-R. Sadeghi, DÏoT: A federated self-learning anomaly detection system for IoT, in: Proc. IEEE 39th Int. Conf. on Distributed Computing Systems (ICDCS), 2019.

[7] B. McMahan, E. Moore, D. Ramage, S. Hampson, B. Agüera y Arcas, Communication-efficient learning of deep networks from decentralized data, in: Proc. 20th Int. Conf. on Artificial Intelligence and Statistics (AISTATS), 2017.

[8] T. Li, A.K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, V. Smith, Federated optimization in heterogeneous networks, in: Proc. Machine Learning and Systems (MLSys), 2020.

[9] M.J. Sheller, B. Edwards, G.A. Reina, et al., Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data, Scientific Reports 10 (2020) 12598.

[10] N. Rieke, J. Hancox, W. Li, et al., The future of digital health with federated learning, npj Digital Medicine 3 (2020) 119.

[11] G.A. Kaissis, M.R. Makowski, D. Rückert, R.F. Braren, Secure, privacy-preserving and federated machine learning in medical imaging, Nature Machine Intelligence 2 (2020) 305–311.

[12] I. Dayan, H.R. Roth, A. Zhong, et al., Federated learning for predicting clinical

outcomes in patients with COVID-19, Nature Medicine 27 (2021) 1735–1743.
[13] W. Schneble, G. Thamilarasu, Attack detection using federated learning in medical cyber-physical systems, in: Proc. 28th Int. Conf. on Computer Communications and Networks (ICCCN), 2019.
[14] G. Thamilarasu, A. Odesile, A. Hoang, An intrusion detection system for Internet of Medical Things, IEEE Access 8 (2020) 181560–181576.
[15] S.A. Rahman, H. Tout, C. Talhi, A. Mourad, Internet of Things intrusion detection: Centralized, on-device, or federated learning?, IEEE Network 34 (6) (2020) 310–317.
[16] V. Mothukuri, P. Khare, R.M. Parizi, S. Pouriyeh, A. Dehghantanha, G. Srivastava, Federated-learning-based anomaly detection for IoT security attacks, IEEE Internet of Things Journal 9 (4) (2022) 2545–2554.
[17] L. Melis, C. Song, E. De Cristofaro, V. Shmatikov, Exploiting unintended feature leakage in collaborative learning, in: Proc. IEEE Symposium on Security and Privacy (S&P), 2019.
[18] L. Zhu, Z. Liu, S. Han, Deep leakage from gradients, in: Advances in Neural Information Processing Systems (NeurIPS), 2019.
[19] R. Shokri, M. Stronati, C. Song, V. Shmatikov, Membership inference attacks against machine learning models, in: Proc. IEEE Symposium on Security and Privacy (S&P), 2017.
[20] M. Nasr, R. Shokri, A. Houmansadr, Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning, in: Proc. IEEE Symposium on Security and Privacy (S&P), 2019.
[21] C. Dwork, F. McSherry, K. Nissim, A. Smith, Calibrating noise to sensitivity in private data analysis, in: Theory of Cryptography Conference (TCC), 2006.
[22] M. Abadi, A. Chu, I. Goodfellow, H.B. McMahan, I. Mironov, K. Talwar, L. Zhang, Deep learning with differential privacy, in: Proc. ACM SIGSAC Conf. on Computer and Communications Security (CCS), 2016.
[23] I. Mironov, Rényi differential privacy, in: Proc. IEEE 30th Computer Security Foundations Symposium (CSF), 2017.
[24] B. Balle, G. Barthe, M. Gaboardi, J. Hsu, T. Sato, Hypothesis testing interpretations and Rényi differential privacy, in: Proc. 23rd Int. Conf. on Artificial Intelligence and Statistics (AISTATS), 2020.
[25] H.B. McMahan, D. Ramage, K. Talwar, L. Zhang, Learning differentially private recurrent language models, in: Proc. Int. Conf. on Learning Representations (ICLR), 2018.
[26] P. Blanchard, E.M. El Mhamdi, R. Guerraoui, J. Stainer, Machine learning with adversaries: Byzantine tolerant gradient descent, in: Advances in Neural Information Processing Systems (NeurIPS), 2017.
[27] D. Yin, Y. Chen, K. Ramchandran, P. Bartlett, Byzantine-robust distributed learning: Towards optimal statistical rates, in: Proc. 35th Int. Conf. on Machine Learning (ICML), 2018.
[28] E.M. El Mhamdi, R. Guerraoui, S. Rouault, The hidden vulnerability of distributed learning in Byzantium, in: Proc. 35th Int. Conf. on Machine Learning (ICML), 2018.
[29] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, V. Shmatikov, How to backdoor federated learning, in: Proc. 23rd Int. Conf. on Artificial Intelligence and Statistics (AISTATS), 2020.
[30] M. Fang, X. Cao, J. Jia, N.Z. Gong, Local model poisoning attacks to Byzantine-robust federated learning, in: Proc. 29th USENIX Security Symposium, 2020.
[31] S.Y. Argaw, J.R. Troncoso-Pastoriza, D. Lacey, et al., Cybersecurity of hospitals: discussing the challenges and working towards mitigating the risks, BMC Medical Informatics and Decision Making 20 (2020) 146.
[32] H.T. Neprash, C.C. McGlave, D.A. Cross, et al., Trends in ransomware attacks on US hospitals, clinics, and other health care delivery organizations, 2016–2021, JAMA Health Forum 3 (12) (2022) e224873.
[33] P. Kairouz, H.B. McMahan, B. Avent, et al., Advances and open problems in federated learning, Foundations and Trends in Machine Learning 14 (1–2) (2021) 1–210.
[34] Q. Yang, Y. Liu, T. Chen, Y. Tong, Federated machine learning: Concept and applications, ACM Transactions on Intelligent Systems and Technology 10 (2) (2019) 1–19.