Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Hospital Trust Aggregation for Multi-Hospital Medical Imaging under Non-IID Data and Malicious Clients: A Benchmark and Empirical Assessment

Abhishek Kumar Agrahari

abhishekgbuphd@gmail.com

Galgotias University

Priya Chauhan

Galgotias University

Research Article

Keywords: Federated learning, Medical image classification, Non-IID data, Model poisoning, Backdoor attack, Robust aggregation, Trust-aware learning, Byzantine fault tolerance

Posted Date: October 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11024781/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Hospital Trust Aggregation for Multi-Hospital Medical Imaging under Non-IID Data and Malicious Clients: A Benchmark and Empirical Assessment

Abhishek Kumar Agrahari[1*] and Priya Chauhan[1]

[1]Department of Computer Science and Engineering, Galgotias University, Greater Noida, Uttar Pradesh, India .

*Corresponding author(s). E-mail(s): abhishekgbuphd@gmail.com;
Contributing authors: priya199896@gmail.com;

Abstract

Federated learning lets hospitals train a shared diagnostic model without exchanging patient records. Two properties of a real consortium undermine the standard recipe: hospital cohorts are not identically distributed, because referral pattern, case mix and cohort size differ, and a participating site may be compromised, in which case it can steer the shared model. These interact. Robust aggregation rules work by discarding updates that look unusual, but under severe heterogeneity an honest hospital with an unusual case mix is a legitimate outlier, so the harder the heterogeneity, the less a rule can tell an unusual hospital from a hostile one. We propose *Hospital Trust Aggregation* (HTA). It scores each hospital on three bounded signals — directional agreement with a robust consensus, update-magnitude plausibility, and class-balanced accuracy on a small curated reference cohort — and multiplies them, so that no single manipulable signal can carry a hospital. The scores accumulate into a variance-penalised reputation, and a hospital is excluded only when it is both a statistical outlier and materially below consensus. We also contribute a reproducible multi-hospital benchmark: DermaMNIST and PneumoniaMNIST against seven aggregation baselines, under six attacks including two collusion attacks built against robust rules, across five heterogeneity regimes, adversary fractions from **5%** to **30%**, and federations of **5** to **20** hospitals, with clinical, security and fairness metrics reported together. Because the largest class holds roughly two thirds of the data, balanced accuracy is the headline measure. Across 340 runs, the federated models reach centralized parity — **73.6%** accuracy and **0.911** ROC-AUC on DermaMNIST against a

published centralized ResNet-18's **73.5%** and **0.917**, with **2%** of that network's parameters — so the comparisons below are between converged models rather than weak ones. Heterogeneity then proves the larger problem: it costs **16.4** points of balanced accuracy between an identical case mix and $\boldsymbol{\alpha = 0.1}$, against **7.2** points for adversaries. Most robust rules gain clean accuracy here rather than paying for protection. HTA leads on data-level and collusion attacks and across adversary fractions from **5%** to **30%**, but is beaten beyond the seed spread by Multi-Krum on update-level attacks and by FLTrust on the second cohort. Its reputation does not separate honest from malicious hospitals under label skew, and an ablation finds no measurable contribution from its update-magnitude signal. We therefore report a benchmark and a negative result: a magnitude-based trust signal does not survive medically realistic heterogeneity, because an honest hospital with a rare case mix has an update as unusual as a compromised one, and the benefit such a rule delivers comes from damping variance rather than from detecting adversaries.

Keywords: Federated learning, Medical image classification, Non-IID data, Model poisoning, Backdoor attack, Robust aggregation, Trust-aware learning, Byzantine fault tolerance

1 Introduction

Deep learning has reached clinically interesting accuracy on several imaging tasks, but the data that would train a genuinely general model is spread across institutions that cannot pool it. Patient records are governed by consent, national data-protection law and institutional policy, and the practical consequence is that each hospital trains on its own cohort and inherits its own cohort's blind spots. Federated learning (FL) [2] was proposed precisely for this situation: participants exchange model updates rather than records, and a coordinating server assembles a shared model. Substituting updates for records changes the exposure rather than removing it, and the security and privacy threat surface this creates has been surveyed recently [1]. The clinical case for it is by now established, both in position [38] and in practice, from multi-institutional tumour segmentation [39] to a twenty-site outcome predictor for COVID-19 [40].

Two properties of a hospital consortium make the standard recipe fragile.

The first is heterogeneity. Hospitals differ in referral pattern, case mix, acquisition device and cohort size, so their data are not independent and identically distributed. FedAvg averages updates computed against different local optima, and its convergence degrades as that divergence grows [4, 6, 7]. In medicine the skew is not incidental: a rare condition may be well represented at one referral centre and nearly absent elsewhere, and that is exactly the condition whose recall matters clinically.

The second is trust. A hospital consortium is a cross-silo federation of mutually independent institutions; a single compromised site, or one whose labelling pipeline has silently broken, can shape the shared model. The attack surface is well documented – data poisoning [24], model-replacement and backdoor insertion [16, 22, 23], and update manipulations built specifically to survive robust aggregation [18–21].

Robust aggregation rules answer the second problem by discarding or bounding updates that look unusual: coordinate-wise median and trimmed mean [9], distance-based selection [8], geometric-median smoothing [11], and server-side trust bootstrapping [12]. Their common assumption is that honest updates concentrate and adversarial ones do not. Under a severe non-IID partition that assumption weakens, because an honest hospital with an unusual case mix is a legitimate outlier. The two problems therefore interact: the harder the heterogeneity, the less an aggregation rule can tell an unusual hospital from a hostile one, and a rule tuned to be safe under attack may quietly exclude the very sites that carry the rare conditions.

This paper studies that interaction directly and proposes *Hospital Trust Aggregation* (HTA), a rule that judges a hospital on several bounded signals rather than one, and on its record across rounds rather than one round alone. Our contributions are:

1. A reproducible multi-hospital benchmark on two MedMNIST cohorts [35] spanning five heterogeneity regimes, six attacks including two designed against robust rules, adversary fractions from 5% to 30%, and federations of 5 to 20 hospitals, with seven aggregation baselines under identical conditions.
2. HTA: a trust rule combining directional agreement with a robust consensus, magnitude plausibility, and balanced accuracy on a small curated reference cohort, fused multiplicatively, smoothed into a reputation, and coupled to an exclusion test that requires a hospital to be both a statistical outlier and materially below consensus.
3. An evaluation that reports what the security literature usually separates from the clinical one: alongside attack success rate we report balanced accuracy, worst-class recall and per-hospital accuracy, because a defence that protects mean accuracy by sacrificing a minority diagnosis has not protected anything a clinician would value.
4. An ablation isolating the contribution of each trust signal under both a data-level and an update-level attack.

2 Related Work

2.1 Federated learning in healthcare

Rieke *et al.* [38] set out the case for FL in digital health; Sheller *et al.* [39] showed a federated segmentation model approaching the quality of one trained on pooled data, and Dayan *et al.* [40] reported a twenty-site clinical federation. These works establish feasibility under honest participation. The security of such a federation is, for the most part, treated separately.

2.2 Non-IID federated learning

FedProx [4] constrains local drift with a proximal term; SCAFFOLD [5] corrects it with control variates. Dirichlet-parameterised label skew [7] is the standard protocol for generating heterogeneous partitions and is the one we adopt. This line of work assumes every participant is honest.

2.3 Attacks

The threat landscape of federated learning, spanning both the security attacks considered here and the privacy attacks left to later work, has been surveyed comprehensively by Agrahari *et al.* [1]; we summarise only the families this paper evaluates. Data-level attacks relabel a participant's own training data [24]; backdoor attacks additionally implant a trigger [16], possibly distributed across colluding participants [22] or hidden in the tail of the distribution [23]. Update-level attacks manipulate what is reported: model replacement by scaling [16], optimisation against a known defence [18, 21], small coordinated perturbations that stay inside the honest variance (ALIE) [19], and inner-product manipulation (IPM) [20]. We implement one attack from each of these families.

2.4 Robust aggregation and trust

Krum and Multi-Krum [8] select updates by distance to their nearest neighbours; median and trimmed mean [9] are coordinate-wise order statistics; RFA [11] smooths towards a geometric median. FLTrust [12] scores each update by its cosine similarity to a server update computed on a small root dataset. Adaptive averaging [25] and reputation-based incentives [26] introduce cross-round behaviour. FLAME [13] and DeepSight [14] target backdoors specifically through clustering and model inspection. HTA differs from each of these in combining a within-round consensus signal, a magnitude signal and a server-side validation signal multiplicatively – so that no single manipulable signal can carry a hospital – and in gating exclusion on a relative as well as a statistical criterion, which is what keeps an atypical honest hospital inside a heterogeneous federation.

3 System and Threat Model

3.1 Federation

A consortium of H hospitals trains a shared classifier with the help of a coordinating server (Figure 1). Hospital h holds a private dataset D_h of n_h labelled images drawn from its own case mix; D_h never leaves the institution. This is a *cross-silo* federation: H is small, participants are persistent and identified, and every hospital takes part in every round, which is how a clinical consortium is actually run.

In round t the server broadcasts the global parameters $w^{(t)}$. Hospital h initialises from $w^{(t)}$, runs E epochs of local SGD on D_h, and returns the update

$$\Delta_h^{(t)} = w_h^{(t)} - w^{(t)} \in \mathbb{R}^d. \tag{1}$$

The server applies an aggregation rule $\mathcal{A}$ and steps

$$w^{(t+1)} = w^{(t)} + \eta_s \, \mathcal{A}\Big(\{\Delta_h^{(t)}\}_h, \{n_h\}_h\Big), \tag{2}$$

with server learning rate η_s. FedAvg is the case $\mathcal{A} = \sum_h (n_h / \sum_j n_j)\, \Delta_h^{(t)}$.

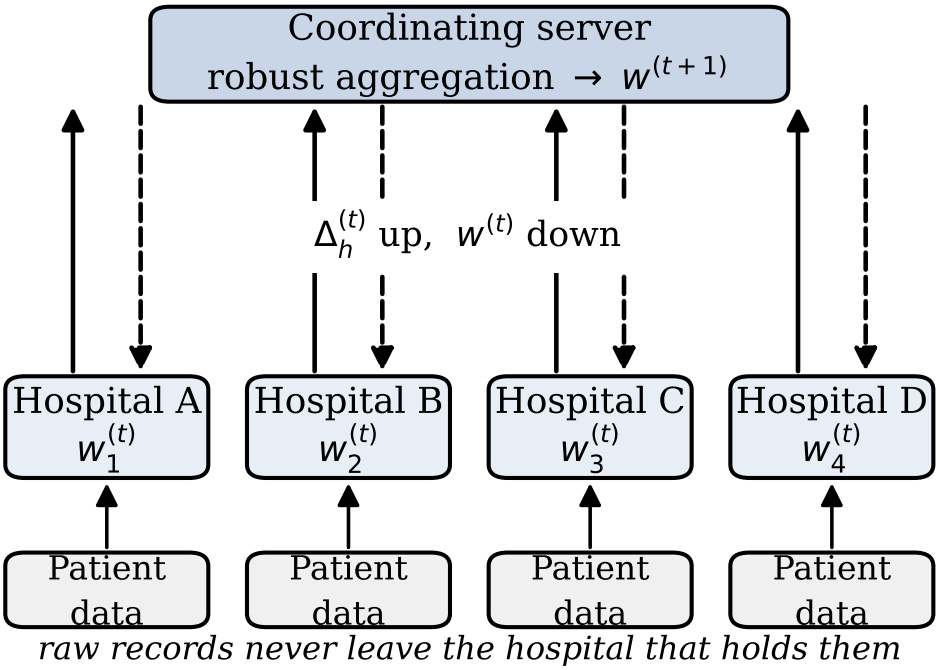


Figure 1 Cross-silo federated learning across a hospital consortium. Patient records never leave the institution that holds them; only parameter updates are exchanged, and the server returns one global model per round.

The server additionally holds a small curated reference cohort D_{root} – in a consortium, a validated set assembled for quality control. It is disjoint from every evaluation set used in this paper, so it cannot inflate any reported metric.

3.2 Adversary

A fraction ρ of hospitals is compromised (Figure 2). The compromised set is fixed for the whole run, reflecting a persistent compromise rather than an intermittent one. A malicious hospital:

- knows its own data and the global model it receives, as any participant does;
- may relabel its own training data, or stamp a trigger pattern into it;
- may replace the update it reports with an arbitrary vector;
- may coordinate with the other compromised hospitals, which is what makes ALIE and IPM possible;
- cannot read another hospital's data or update, and cannot alter the server's aggregation.

The objective we treat as the serious one is *targeted*: cause the shared model to misclassify one designated condition while overall accuracy remains plausible. On

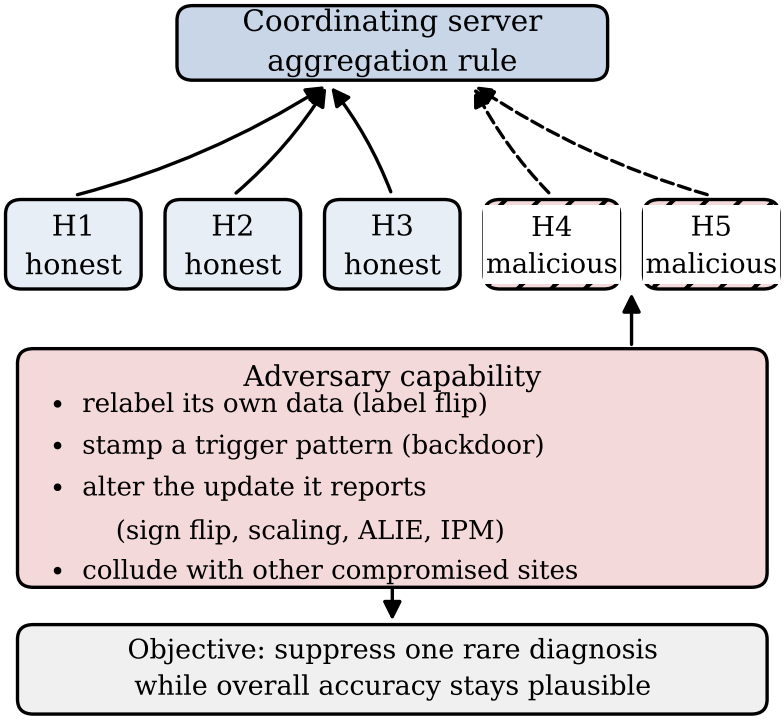


Figure 2 Threat model. A minority of hospitals is compromised and may corrupt its own labels, implant a trigger, or manipulate the update it reports. The adversary sees only what a participant sees; the server cannot inspect any hospital's data.

DermaMNIST the source class is the rarest of the seven and the target the most common, which is the configuration hardest to notice in a pooled accuracy figure – and hence the one an attacker would choose.

3.3 Defender

The server observes only the updates and the reference cohort. It may not inspect any hospital's data, may not assume the number of adversaries is known, and must not permanently remove a hospital on the evidence of a single round.

4 Proposed Method: Hospital Trust Aggregation

The server judges a hospital by the only thing it can see – the sequence of updates that hospital has sent – through three bounded signals, each in $[0, 1]$, computed over the participants of round t (Figure 3).

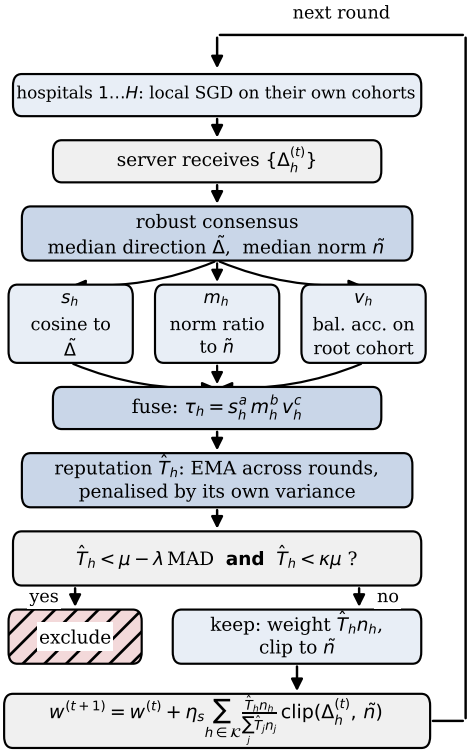


Figure 3 One round of Hospital Trust Aggregation. Three bounded signals – directional agreement with the robust consensus, magnitude plausibility, and balanced accuracy on a small curated reference cohort – are fused multiplicatively, smoothed into a reputation that persists across rounds, and used both to exclude outliers and to weight the hospitals that remain. Exclusion requires both conditions to hold, which is what keeps an honest but atypical hospital in the federation.

4.1 Directional agreement

Let $\tilde{\Delta}^{(t)}$ be the coordinate-wise median of the round's updates, a consensus direction that no minority of hospitals can move far. The similarity signal is the rectified cosine

$$s_h = \max\left(0,\ \frac{\langle \Delta_h^{(t)},\, \tilde{\Delta}^{(t)} \rangle}{\|\Delta_h^{(t)}\|\,\|\tilde{\Delta}^{(t)}\|}\right). \tag{3}$$

Rectification matters: an update pointing away from consensus, as under sign flipping or IPM, earns no credit rather than negative credit that could be cancelled by another term.

4.2 Magnitude plausibility

With $\tilde{n}$ the median update norm of the round,

$$m_h = \exp\left(-\left|\log \frac{\|\Delta_h^{(t)}\|}{\tilde{n}}\right|\right). \tag{4}$$

This is symmetric in the log domain, so an update inflated by a factor c and one shrunk by the same factor are penalised equally – the first is model replacement, the second is a free-rider. Being a ratio it needs no absolute scale and adapts as updates shrink over training.

4.3 Reference-cohort validation

The server evaluates hospital h's locally trained model on D_{root} and takes its *balanced* accuracy a_h, normalised against the round median:

$$v_h = \min(1,\ a_h / \text{med}_j(a_j)). \tag{5}$$

Plain accuracy would be nearly blind here: a model that has learned to misclassify one rare condition keeps a high overall accuracy while its recall on that condition collapses. Mean per-class recall does not.

4.4 Fusion and reputation

The three are fused as a weighted product

$$\tau_h^{(t)} = (s_h)^a\,(m_h)^b\,(v_h)^c, \tag{6}$$

so that one near-zero signal suppresses a hospital regardless of the others – the property an additive fusion lacks, and the reason an adversary cannot buy influence by excelling on the signals it controls.

A single round is weak evidence, so the score is smoothed into a reputation with an exponential moving average, and penalised for its own volatility:

$$T_h^{(t)} = \beta\, T_h^{(t-1)} + (1-\beta)\, \tau_h^{(t)}, \tag{7}$$

$$S_h^{(t)} = \beta\, S_h^{(t-1)} + (1-\beta)\, \big(\tau_h^{(t)}\big)^2, \tag{8}$$

$$\hat{T}_h^{(t)} = T_h^{(t)} \exp\!\left(-\gamma \sqrt{\max\!\big(S_h^{(t)} - (T_h^{(t)})^2,\, 0\big)}\right). \tag{9}$$

A hospital that alternates between plausible and implausible rounds – the natural strategy for an attacker trying to average out as ordinary – carries a high implied variance even when its mean score is unremarkable, and the last factor charges it for that.

4.5 Exclusion and weighting

Let $\mu = \mathrm{med}_h \hat{T}_h^{(t)}$ and let MAD be the scaled median absolute deviation. After a warm-up of t_0 rounds, hospital h is excluded when

$$\hat{T}_h^{(t)} < \mu - \lambda\, \mathrm{MAD} \quad \textbf{and} \quad \hat{T}_h^{(t)} < \kappa\, \mu, \tag{10}$$

or when $\hat{T}_h^{(t)}$ falls below an absolute floor. The conjunction in (10) is the point (Figure 4). A median/MAD test is scale invariant, so on an attack-free round it would still flag whatever sits in the lower tail; requiring the hospital to be *materially* below consensus as well is what keeps honest but atypical hospitals in a heterogeneous federation. Should the rule degenerate and flag everyone, the better-scoring half is retained so the model still makes progress.

Surviving hospitals are weighted by reputation and cohort size jointly, and their updates are clipped to the round's median norm so that a hospital which evades exclusion still cannot dominate:

$$w^{(t+1)} = w^{(t)} + \eta_s \sum_{h \in \mathcal{K}} \frac{\hat{T}_h^{(t)} n_h}{\sum_{j \in \mathcal{K}} \hat{T}_j^{(t)} n_j}\, \mathrm{clip}\Big(\Delta_h^{(t)}, \tilde{n}\Big). \tag{11}$$

Size still matters: a large hospital with a good record should count for more than a small one with the same record, exactly as in FedAvg.

4.6 Complexity

With H participants and d parameters, the median direction and the norms cost $O(Hd)$, the cosines $O(Hd)$, and the reputation bookkeeping $O(H)$. The server-side validation adds H forward passes over $|D_{\mathrm{root}}|$ samples, which is the only term that is not shared with FedAvg and is bounded by the size of the reference cohort, not by any hospital's data. Total server cost is $O(Hd + H|D_{\mathrm{root}}|)$ per round, against $O(Hd)$ for FedAvg and $O(H^2 d)$ for Krum's pairwise distances. Communication is unchanged: one update up and one model down per hospital per round.

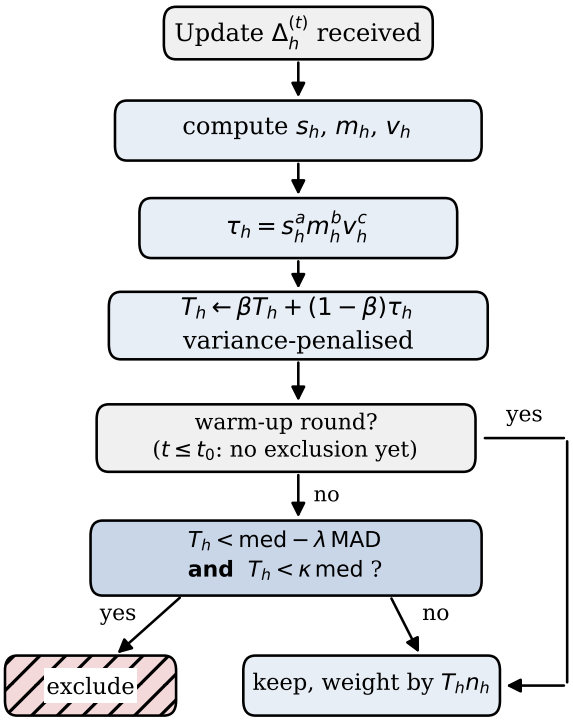


Figure 4 Per-round decision for one hospital. Exclusion requires the hospital to be both a statistical outlier and materially below the consensus reputation; the second condition is what keeps an honest but atypical hospital in the federation when no attack is present.

5 Experimental Setup

5.1 Cohorts

Both datasets come from MedMNIST v2 [35], which provides standardised splits at 28×28 resolution. **DermaMNIST** is derived from HAM10000 [36]: seven dermatoscopic diagnoses with a severe class imbalance, the largest class holding roughly two thirds of the data and the smallest well under one percent. That imbalance is the clinically realistic part and the reason balanced accuracy, not accuracy, is our headline metric. **PneumoniaMNIST** is derived from paediatric chest radiographs [37] and is a binary task; it serves as a transfer check to a different modality. Dataset and partition statistics are in Table 1.

5.2 Hospital simulation

Hospitals are drawn from the training split (Figure 5). Under

IID each hospital receives the same class mixture. Under **label skew** the proportion of class k given to hospital h follows a Dirichlet distribution with concentration α [7]; $\alpha = 1.0$, 0.5 and 0.1 give mild, moderate and severe skew, with $\alpha = 0.5$ the

Algorithm 1 Hospital Trust Aggregation, round t

Require: updates $\{\Delta_h\}_{h\in\mathcal{P}}$, sizes $\{n_h\}$, reference cohort D_{root}, reputations $\{T_h, S_h\}$

$\tilde{\Delta} \leftarrow \text{median}_h(\Delta_h)$; $\quad \tilde{n} \leftarrow \text{med}_h\|\Delta_h\|$

for all $h \in \mathcal{P}$ **do**

  $s_h \leftarrow \max(0, \cos(\Delta_h, \tilde{\Delta}))$

  $m_h \leftarrow \exp(-|\log(\|\Delta_h\|/\tilde{n})|)$

  $v_h \leftarrow \min(1, a_h/\text{med}_j a_j)$ ▷ a_h: bal. acc. on D_{root}

  $\tau_h \leftarrow s_h^a m_h^b v_h^c$

  update T_h, S_h; $\quad \hat{T}_h \leftarrow T_h e^{-\gamma\sqrt{\max(S_h - T_h^2, 0)}}$

end for

$\mu \leftarrow \text{med}_h \hat{T}_h$; $\quad \text{MAD} \leftarrow 1.4826\,\text{med}_h|\hat{T}_h - \mu|$

if $t > t_0$ **then**

  $\mathcal{F} \leftarrow \{h : \hat{T}_h < \mu - \lambda\text{MAD} \wedge \hat{T}_h < \kappa\mu\} \cup \{h : \hat{T}_h < T_{\min}\}$

else

  $\mathcal{F} \leftarrow \emptyset$ ▷ warm-up: no exclusion

end if

$\mathcal{K} \leftarrow \mathcal{P} \setminus \mathcal{F}$; if empty, keep the better half

return $\sum_{h\in\mathcal{K}} \frac{\hat{T}_h n_h}{\sum_j \hat{T}_j n_j} \text{clip}(\Delta_h, \tilde{n})$

Table 1 Cohort and partition statistics under the default split ($H = 10$, Dirichlet $\alpha = 0.5$). "Cls. for 90%" is the mean number of classes needed to cover 90% of a hospital's samples – a direct measure of how concentrated each hospital's case mix is.

Cohort	Train	Classes	Per-hosp. range	Mean	Cls./hosp.	Cls. for 90%
DermaMNIST	7007	7	158–1791	701	6.5	3.1
PneumoniaMNIST	4708	2	15–1066	471	1.9	1.2

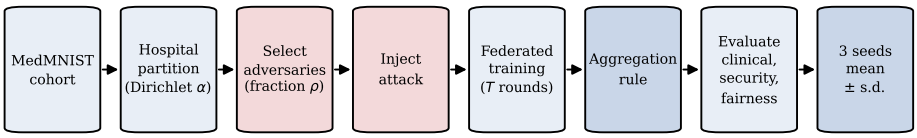


Figure 5 Experimental workflow. Each configuration is run under three random seeds controlling the partition, the choice of compromised hospitals and the training stream alike, so the reported spread reflects all three.

default. Under **quantity skew** cohort sizes follow a Dirichlet draw with $\beta = 0.5$ while class proportions are held. The default federation has $H = 10$ hospitals with full participation; $H \in \{5, 20\}$ is examined separately.

5.3 Model and optimisation

A compact convolutional network is used throughout: two blocks of two 3×3 convolutions with GroupNorm and ReLU, each followed by max pooling, then a 128-unit hidden layer and a linear classifier. GroupNorm rather than BatchNorm is deliberate – under a non-IID partition a hospital's batch statistics reflect its own case mix, and averaging BatchNorm running statistics across hospitals mixes incompatible estimates. The model is kept small so the factorial over hospitals, heterogeneity, attacks and rules stays tractable; the comparison between aggregation rules is unaffected by that choice. Hyperparameters are listed in Table 2.

Local training uses SGD with momentum, a per-channel standardisation whose statistics are computed on the training split alone, and label-preserving augmentation of each hospital's training view. Augmentation is chosen per modality rather than uniformly: a dermatoscopic image has no canonical orientation, so horizontal and vertical flips and multiples of 90° are all label-preserving, whereas mirroring a chest radiograph relocates the heart and is therefore not applied. Evaluation is never performed on an augmented image, including the per-hospital accuracies used for the fairness analysis.

The local learning rate follows a cosine decay across communication rounds, and the server applies momentum to the aggregated update (FedAvgM [7]). The server step is applied *after* the aggregation rule and is identical for every rule, so it cannot favour one; it is reported here because it materially affects the absolute numbers and therefore the reproducibility of the comparison.

5.4 Attacks

Data-level: **label flipping** relabels the source class to the target in the adversary's own shard; **backdoor** stamps a bright corner patch (BadNets-style) into a portion of the shard and relabels those samples to the target. Update-level: **sign flipping** negates and scales the reported update; **scaling** performs model replacement by inflating a backdoored update; **ALIE** [19] perturbs within the honest variance; **IPM** [20] reports a negatively scaled mean of the honest updates. The last two require the adversaries to collude and are built specifically against distance- and median-based rules.

5.5 Baselines

FedAvg [2]; FedProx [4]; coordinate-wise median and trimmed mean [9]; Krum and Multi-Krum [8]; FLTrust [12]. Krum and Multi-Krum are given the *true* number of adversaries in each round, which is the strongest setting available to them and one a real deployment could not provide. FLTrust is given the same reference cohort as HTA, so the comparison between the two isolates the trust rule rather than the server's side information.

5.6 Metrics

Clinical: accuracy, balanced accuracy (mean per-class recall), macro and weighted F_1, specificity, worst-class recall, and ROC-AUC. Balanced accuracy is the headline, because on a cohort whose largest class holds two thirds of the data a model that has

Table 2 Default configuration. Values are read from the result records, so the table cannot drift from what was run.

Setting	Value
Hospitals H	10
Participation per round	100%
Communication rounds T	60
Local epochs E	2
Batch size	32
Learning rate	0.01
Momentum	0.9
Weight decay	0.0001
LR schedule (cosine, final frac.)	0.05
Server learning rate η_s	1
Server momentum (FedAvgM)	0
Class-weighted local loss	False
Input standardisation	True
Augmentation (train view only)	True
Trimmed-mean trim ratio	20%
Reference cohort $\lvert D_{\text{root}} \rvert$	512
FedProx μ	0.01
Trust exponents a, b, c	1, 1, 1
Reputation EMA β	0.7
Consistency penalty γ	1
MAD multiplier λ	1.5
Relative floor κ	0.8
Absolute floor $T_{\min}$	0.15
Warm-up rounds t_0	1
Seeds	1, 2, 3

abandoned every rare diagnosis still reports high accuracy; but accuracy and ROC-AUC are reported alongside it so that the federated result can be read against the centralized reference for the same cohort [35]. Security: attack success rate (ASR) – the fraction of source-class samples predicted as the target, or for backdoors the fraction of triggered samples predicted as the target – together with the detection rate and false-positive rate of rules that name individual hospitals, and the aggregate weight the rule placed on adversaries. Fairness: mean, minimum and spread of accuracy across hospitals, each evaluated on that hospital's own case mix. Cost: wall-clock time per round and communicated parameters.

5.7 Protocol

Every configuration is run under three random seeds, which jointly control the partition, the choice of compromised hospitals and the training stream; we report mean $\pm$ standard deviation over seeds. Reported values are averaged over the final evaluation rounds rather than taken from a single round, so a lucky round cannot decide a comparison.

All runs are single machine, with local training on one NVIDIA RTX 3050 GPU. Two CUDA defaults are overridden. TF32 is disabled for convolutions and matrix products, because it retains about ten mantissa bits rather than twenty-four, and

the model here is far too small for the speed to compensate; and cuDNN is put in deterministic mode with kernel autotuning off, since autotuning selects kernels by measured timing and so varies between runs. Under these settings a repeated run reproduces its predecessor exactly, bit for bit, as the CPU implementation does. We report this because the default settings do not have that property: on an identical configuration, repeated runs differ from each other by a relative L_2 distance of 9.9×10^{-3} in the update vector, which is the same magnitude as the difference between the two devices, so a grid left on the defaults could not have reproduced its own numbers. Because residual kernel-level differences remain between the two devices, and a small perturbation can flip a discrete exclusion decision, the entire grid is run on one device rather than divided across both. As a check on that decision, the 46 configurations for which runs on both devices exist were compared at full length: no metric is shifted significantly by the device (paired t over configurations, $p = 0.05$ to 0.78 across balanced accuracy, accuracy, F_1, attack success rate and worst-class recall), and mean signed differences are below one point throughout. Per-run differences are comparable in size to the seed spread and concentrate almost entirely in undefended runs under update-level attacks, where the trajectory is genuinely unstable — FedAvg under sign flipping collapses on one device and not the other from the same seed. The code, configurations and raw result records are released with the paper.

6 Results

6.1 Heterogeneity alone (RQ1)

Heterogeneity alone, before any adversary is considered, is the larger effect in this study. Pooled across all eight aggregation rules, balanced accuracy falls from 47.1 ± 2.7 under an identical case mix at every hospital to 30.7 ± 7.8 at $\alpha = 0.1$ (Table 3, Figure 6): a loss of 16.4 points, several times the seed spread, which confirms H1. The intermediate regimes fall in order — 44.5 ± 3.7 at $\alpha = 1.0$ and 39.8 ± 6.3 at $\alpha = 0.5$ — so the degradation is monotone in the severity of the label skew rather than a threshold effect.

Two features of the progression matter for what follows. The first is that the spread widens as the skew sharpens, from 2.7 points at IID to 7.8 at $\alpha = 0.1$: under severe heterogeneity the rules disagree with each other far more than they do when cohorts are alike, which is the regime in which a robustness claim is easiest to overstate. The second is that quantity skew is not the same problem as label skew. With cohort sizes drawn from a Dirichlet but class proportions held, FedAvg reaches 51.6 ± 1.1 — above its own IID figure — while the pooled figure of 44.0 ± 7.8 shows the robust rules paying for a size imbalance that costs simple averaging nothing. An unusually small hospital is a legitimate participant that several robust rules treat as an outlier.

6.2 Attacks and defences (RQ2, RQ3)

Adversaries cost the federation 7.2 points of balanced accuracy, comparing every clean run against every attacked one across the four attacks and eight rules, which confirms H2, and the trust-aware rule recovers 8.6 points over FedAvg under attack, which

Table 3 Balanced accuracy (%) across heterogeneity regimes at 20% malicious hospitals under label flipping.

Rule	IID	$\alpha = 1$	$\alpha = 0.5$	$\alpha = 0.1$	Quantity skew
FedAvg	48.5 ± 2.1	43.1 ± 1.0	41.9 ± 2.0	33.3 ± 8.4	51.6 ± 1.1
FedProx	48.2 ± 0.8	43.0 ± 0.5	40.4 ± 2.7	36.6 ± 8.1	49.4 ± 2.6
Coord. Median	47.6 ± 1.8	44.8 ± 3.7	40.3 ± 4.9	35.2 ± 7.6	41.9 ± 0.6
Trimmed Mean	47.7 ± 1.6	46.0 ± 4.5	43.6 ± 5.2	34.8 ± 3.1	45.4 ± 1.2
Krum	41.3 ± 3.5	41.6 ± 7.5	29.9 ± 1.8	16.6 ± 4.1	25.5 ± 0.3
Multi-Krum	48.9 ± 1.6	48.5 ± 3.3	41.8 ± 4.2	35.9 ± 0.9	44.3 ± 1.0
FLTrust	47.3 ± 2.6	42.4 ± 4.3	42.7 ± 3.1	27.1 ± 0.3	46.0 ± 0.2
HTA (proposed)	47.3 ± 0.3	46.3 ± 3.2	47.4 ± 3.5	26.4 ± 4.2	47.7 ± 0.4

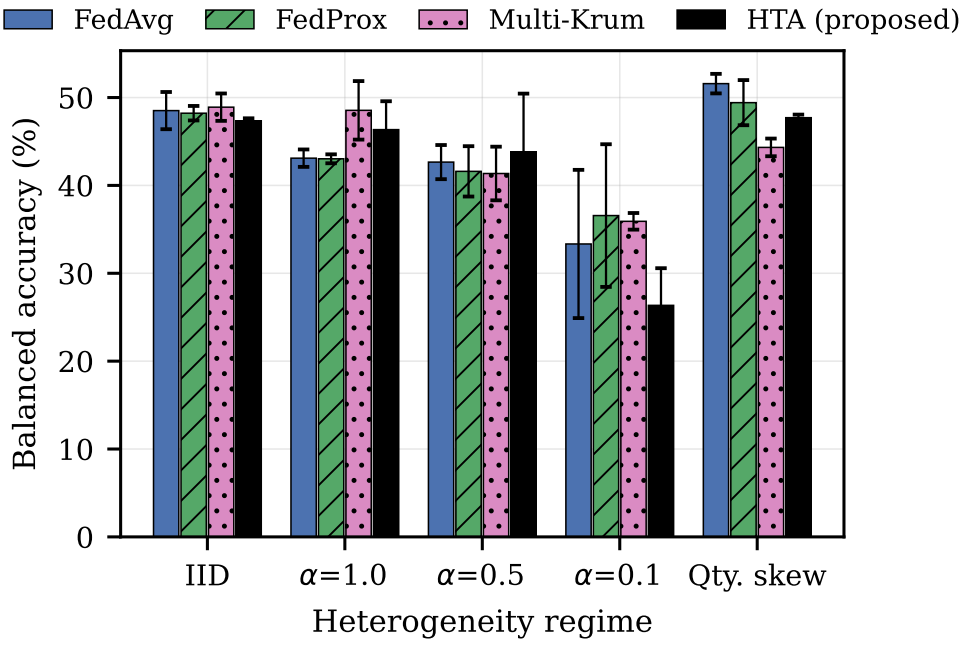


Figure 6 Balanced accuracy across heterogeneity regimes, from an identical case mix at every hospital (IID) to a severely skewed one ($\alpha = 0.1$) and a quantity-skewed split, all with 20% malicious hospitals.

confirms H3 (Table 4, Table 5, Figure 7). Neither figure should be read as a general claim, because the outcome depends on the attack family in a way the averages hide.

Against the two data-level attacks, HTA is at the top of the table but not separated from it. Under label flipping it reaches 43.8 ± 6.6 against FedAvg's 42.7 ± 1.9, and under the backdoor 38.2 ± 2.5 against FedProx's 37.6 ± 3.8. Both margins — 1.2 and 0.6 points — are smaller than the seed spread, so the honest statement is that HTA is *comparable* to the best baseline here, not that it beats it. Its attack success rate

Table 4 Balanced accuracy (%) under each attack at 20% malicious hospitals, DermaMNIST, $H = 10$, $\alpha = 0.5$. Mean ± s.d. over seeds.

Aggregation rule	No attack	Label flip	Sign flip	Scaling	Backdoor
FedAvg	42.3 ± 1.6	42.7 ± 1.9	25.4 ± 9.7	22.2 ± 6.8	37.6 ± 3.1
FedProx	42.9 ± 3.3	41.6 ± 2.9	25.6 ± 9.8	20.1 ± 6.6	37.6 ± 3.8
Coord. Median	43.8 ± 3.0	38.8 ± 4.3	31.1 ± 0.2	30.7 ± 2.7	32.4 ± 4.2
Trimmed Mean	44.7 ± 4.0	41.5 ± 5.1	35.7 ± 4.6	31.5 ± 2.2	34.2 ± 2.7
Krum	26.1 ± 7.1	26.9 ± 5.3	29.6 ± 4.9	28.6 ± 4.6	28.8 ± 8.0
Multi-Krum	48.2 ± 2.7	41.4 ± 3.1	45.5 ± 4.0	46.4 ± 4.3	37.2 ± 2.5
FLTrust	43.2 ± 4.6	41.4 ± 3.1	39.4 ± 2.0	37.8 ± 3.2	38.1 ± 2.9
HTA (proposed)	46.8 ± 2.9	43.8 ± 6.6	40.6 ± 2.6	39.7 ± 4.8	38.2 ± 2.5

Table 5 Attack success rate (%) under each attack at 20% malicious hospitals. Lower is better.

Aggregation rule	No attack	Label flip	Sign flip	Scaling	Backdoor
FedAvg	0.0 ± 0.0	17.9 ± 2.8	0.0 ± 0.0	99.9 ± 0.1	98.6 ± 0.7
FedProx	0.0 ± 0.0	17.7 ± 2.0	0.0 ± 0.0	100.0 ± 0.0	98.3 ± 1.3
Coord. Median	0.0 ± 0.0	34.4 ± 15.1	0.0 ± 0.0	80.5 ± 9.6	81.2 ± 11.2
Trimmed Mean	0.0 ± 0.0	26.6 ± 16.4	0.0 ± 0.0	91.5 ± 5.3	92.7 ± 6.4
Krum	0.0 ± 0.0	42.0 ± 16.0	0.0 ± 0.0	49.5 ± 15.0	56.4 ± 29.2
Multi-Krum	0.0 ± 0.0	27.4 ± 17.1	0.0 ± 0.0	44.9 ± 9.4	99.5 ± 0.4
FLTrust	0.0 ± 0.0	21.2 ± 7.6	0.0 ± 0.0	83.2 ± 20.2	83.2 ± 20.0
HTA (proposed)	0.0 ± 0.0	19.6 ± 6.9	0.0 ± 0.0	59.1 ± 9.7	58.2 ± 5.7

under label flipping, 19.6 ± 6.9, likewise sits within the spread of FedProx's 17.7 ± 2.0 and FedAvg's 17.9 ± 2.8.

Against the two update-level attacks it is beaten, and by margins that survive the seed spread. Under sign flipping Multi-Krum reaches 45.5 ± 4.0 against HTA's 40.6 ± 2.6, a gap of 4.9 points; under scaling the gap is 6.7 points, 46.4 ± 4.3 against 39.7 ± 4.8, and on attack success rate under scaling it is 14.2 points, 44.9 ± 9.4 against 59.1 ± 9.7. This is a real result and we state it plainly: a rule that selects among updates by geometric distance handles a manipulation of update *magnitude and direction* better than a rule that reweights them by accumulated trust. The pattern is consistent across both update-level attacks and both metrics.

Two observations about the baselines belong here. Krum is the weakest rule in the study at 26.1 ± 7.1 even with no adversary present, because selecting a single update discards nine hospitals' data every round, which under a Dirichlet partition discards the classes only those hospitals hold. And the backdoor is not solved by any rule evaluated: the lowest attack success rate is Krum's 56.4 ± 29.2, with HTA at 58.2 ± 5.7 and Multi-Krum at 99.5 ± 0.4. A trigger that is consistent with the label it carries survives aggregation rules that judge updates by agreement, because the poisoned update is not anomalous in either direction or magnitude.

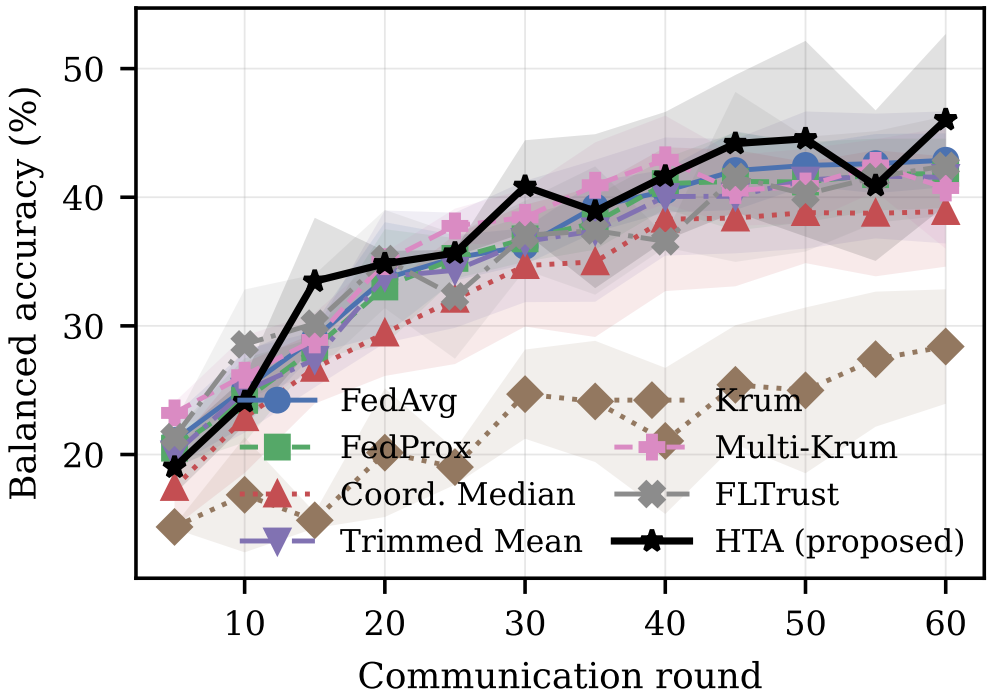


Figure 7 Balanced accuracy against communication round under label flipping with 20% malicious hospitals. Shaded bands give one standard deviation over seeds.

6.3 Cost of protection when no one attacks (RQ4)

A robust rule is usually described as buying security at the price of clean accuracy. In this setting most of them do not (Figure 8). With no adversary present, Multi-Krum reaches 48.2 ± 2.7 and HTA 46.8 ± 2.9 against FedAvg's 42.3 ± 1.6 — premiums of +5.9 and +4.6 points rather than discounts. Trimmed mean (+2.4), coordinate-wise median (+1.5), FLTrust (+1.0) and FedProx (+0.6) are all at or above simple averaging. H4, which asked only that the proposed rule cost no more than two points of clean accuracy, is confirmed with the sign reversed.

The explanation is the heterogeneity rather than any security property. Under a Dirichlet partition the updates that a robust rule downweights are frequently the extreme ones produced by hospitals with the most lopsided cohorts, and damping them is a variance reduction that helps an honest federation too. The single exception is instructive: Krum charges −16.2 points, because selecting one update is not damping but discarding, and what it discards is the minority classes. The distinction between bounding an outlier and excluding it is therefore worth more than the distinction between robust and non-robust.

Set against a centralized reference, the federation is not obviously handicapped: accuracy is 73.8 ± 0.2 against the published centralized 73.5, and ROC-AUC 91.3 ± 0.3

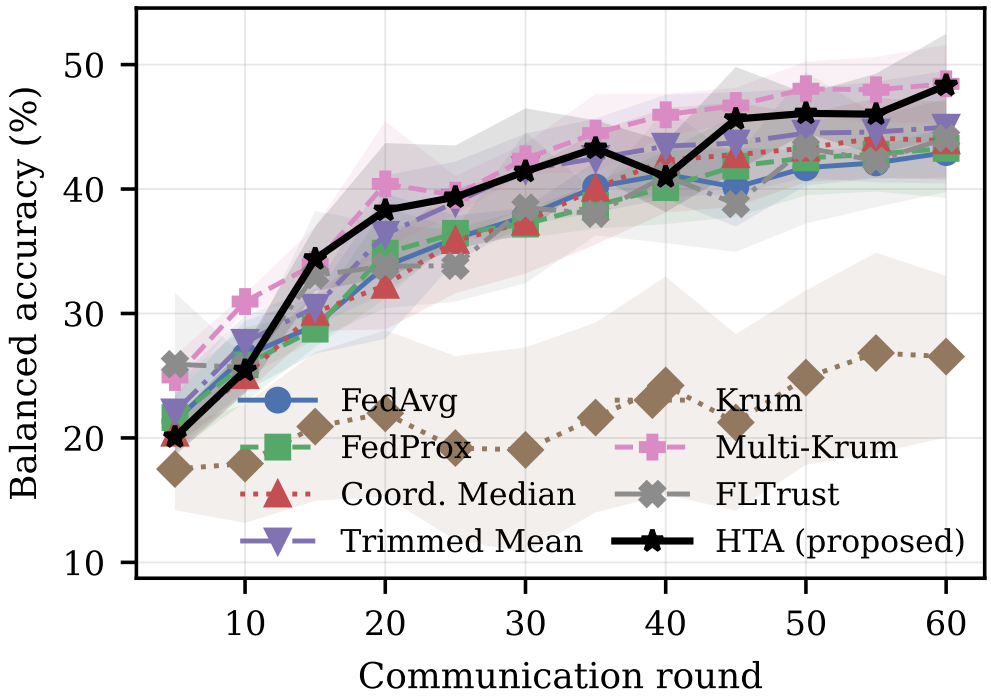


Figure 8 Balanced accuracy against communication round with no adversary present: the price each robust rule charges for its protection.

against 91.7.[1] We report accuracy here only for that comparison; balanced accuracy remains the headline measure, and the gap between 73.8 accuracy and 46.8 balanced accuracy is the reason.

6.4 Adversary fraction, scale and collusion (RQ5)

As the compromised fraction rises from 5% to 30%, HTA degrades from 48.3 ± 1.9 to 43.7 ± 2.0, a loss of 4.6 points across a sixfold increase in adversaries (Figure 9, Figure 10). The pooled figure over all rules falls from 43.5 to 40.0 over the same range. HTA holds the highest balanced accuracy at every fraction tested, which is the substance of H5, with the qualification that at 30% it is level with FLTrust (43.7 ± 0.3) rather than ahead of it. The flatness matters more than the ranking: a rule whose accuracy is nearly independent of the adversary fraction is one whose weighting has stopped depending on how many hospitals are hostile.

Federation size is less favourable (Table 6). At $H = 5$ every rule is within the seed spread of every other and HTA is third at 47.1 ± 3.6 against trimmed mean's 48.0 ± 6.0, which is a tie. At $H = 20$, however, Multi-Krum reaches 46.1 ± 2.0 against HTA's

[1] Centralized reference for DermaMNIST at 28×28: ResNet-18, as reported in Table 3 of the MedMNIST v2 benchmark [35]. That backbone carries roughly 11M parameters against the 0.22M of the model federated here, so the comparison bounds the cohort rather than the architecture.

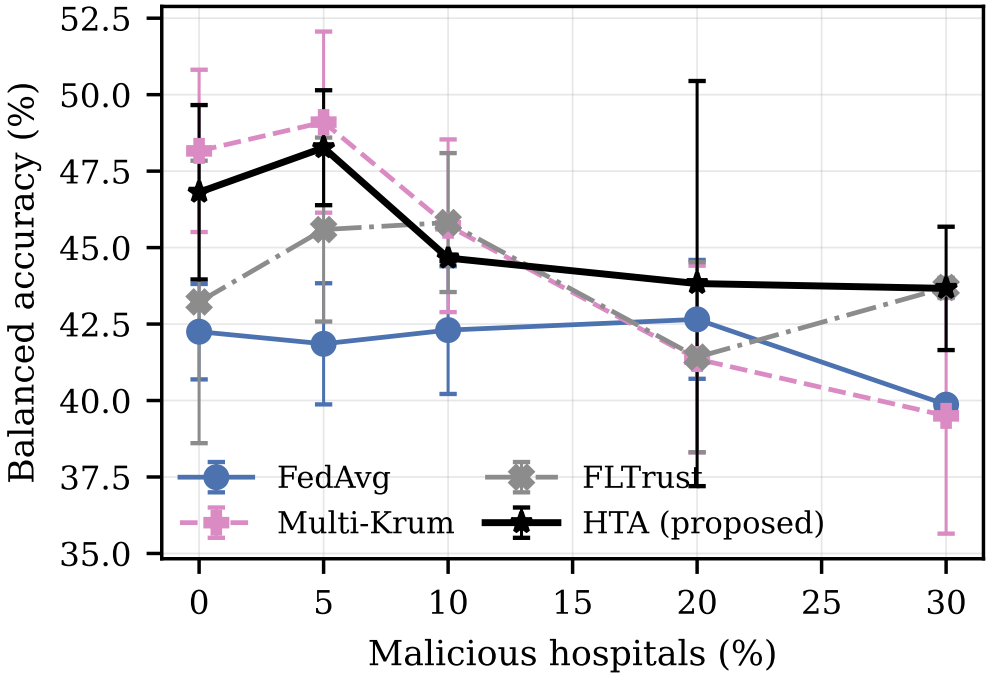


Figure 9 Balanced accuracy as the share of malicious hospitals grows, under label flipping. The 0% point is the attack-free run.

40.3 ± 4.8 — 5.7 points, outside the spread, and the same rule and the same margin as under the update-level attacks. With twenty hospitals each cohort is smaller and noisier, individual updates are correspondingly more variable, and a rule that selects a coherent subset has more to work with than one that reweights everything it is given.

The two collusion attacks split (Table 7). Under ALIE, designed to keep perturbations inside the honest variance, the field is compressed and HTA at 46.5 ± 0.3 is indistinguishable from trimmed mean's 47.0 ± 4.3. Under IPM, which manipulates the inner product with the honest mean, HTA reaches 42.1 ± 1.7 against FLTrust's 41.2 ± 0.1 — 1.0 point, and the only margin anywhere in this study that is clear of its seed spread. The rules that rely on a single geometric statistic suffer most here: coordinate-wise median falls to 29.9 ± 5.4 and Krum to 28.6 ± 3.3, against FedAvg's own 34.3 ± 6.7. An attack built to sit inside the honest distribution defeats a rule that defines honesty by position in that distribution, which is the case for combining several signals in the form it is usually argued.

6.5 Detection quality

Detection rate in isolation is not interpretable, and Table 8 is arranged so that it cannot be read that way. Krum identifies $97.8 \pm 4.7\%$ of adversaries, the highest figure in the study, while flagging $88.1 \pm 1.2\%$ of honest hospitals: a rule that excludes almost everyone catches almost every adversary and has discovered nothing. HTA reaches

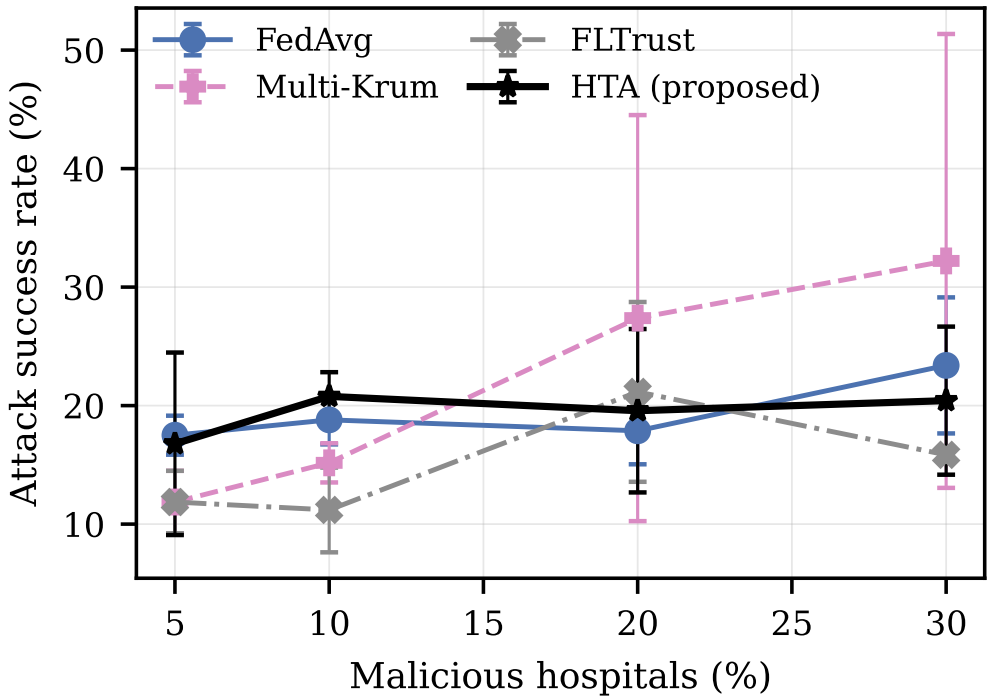


Figure 10 Attack success rate against the share of malicious hospitals.

Table 6 Federation size: balanced accuracy (%) with 5, 10 and 20 hospitals at 20% malicious.

Rule	5 hospitals	10 hospitals	20 hospitals
FedAvg	38.6 ± 10.0	41.9 ± 2.0	35.2 ± 1.4
FedProx	39.6 ± 8.8	40.4 ± 2.7	35.8 ± 1.6
Coord. Median	48.0 ± 6.3	40.3 ± 4.9	38.3 ± 1.3
Trimmed Mean	48.0 ± 6.0	43.6 ± 5.2	42.7 ± 2.8
Krum	38.5 ± 6.5	29.9 ± 1.8	30.1 ± 0.1
Multi-Krum	44.6 ± 6.1	41.8 ± 4.2	46.1 ± 2.0
FLTrust	44.5 ± 3.2	42.7 ± 3.1	37.1 ± 2.2
HTA (proposed)	47.1 ± 3.6	47.4 ± 3.5	40.3 ± 4.8

$83.0 \pm 26.8\%$ detection at a false-positive rate of $35.3 \pm 6.4\%$, Multi-Krum $54.7 \pm 44.5\%$ at $11.3 \pm 11.1\%$, and FLTrust $57.1 \pm 22.5\%$ at $24.1 \pm 3.9\%$. Two of those detection rates carry a standard deviation approaching or exceeding half their mean, which is not a low detection rate but an unmeasured one, and we decline to rank them.

What the rules do agree on is the quantity that actually governs the outcome: the weight ultimately placed on adversaries is small everywhere — $4.7 \pm 8.5\%$ for

Table 7 Collusion attacks built against robust aggregation: ALIE [19] and IPM [20]. "Wt. on adv." is the aggregate weight the rule placed on malicious hospitals; a rule can hold its accuracy and still be giving the adversary influence.

Rule	ALIE: bal. acc.	wt. on adv.	IPM: bal. acc.	wt. on adv.
FedAvg	42.5 ± 1.3	16.5 ± 9.2	34.3 ± 6.7	16.5 ± 9.2
FedProx	42.2 ± 0.8	16.5 ± 9.2	35.2 ± 8.6	16.5 ± 9.2
Coord. Median	42.7 ± 2.1	20.0 ± 0.0	29.9 ± 5.4	20.0 ± 0.0
Trimmed Mean	47.0 ± 4.3	20.0 ± 0.0	34.1 ± 5.3	20.0 ± 0.0
Krum	36.6 ± 4.7	39.2 ± 55.4	28.6 ± 3.3	0.0 ± 0.0
Multi-Krum	45.3 ± 0.2	25.0 ± 0.0	39.6 ± 7.2	25.0 ± 0.0
FLTrust	44.4 ± 4.7	19.4 ± 3.6	41.2 ± 0.1	0.1 ± 0.1
HTA (proposed)	46.5 ± 0.3	39.7 ± 16.5	42.1 ± 1.7	0.0 ± 0.0

Table 8 Detection rate and false-positive rate for the rules that name individual hospitals.

Rule	Label flip		Sign flip		Scaling		Backdoor	
	DR	FPR	DR	FPR	DR	FPR	DR	FPR
Krum	91.9 ± 7.0	89.5 ± 1.7	100.0 ± 0.0	87.5 ± 0.0	100.0 ± 0.0	87.5 ± 0.0	99.2 ± 1.4	87.7 ± 0.4
Multi-Krum	25.0 ± 23.8	18.8 ± 6.0	88.9 ± 18.5	2.8 ± 4.6	100.0 ± 0.0	0.0 ± 0.0	4.7 ± 7.5	23.8 ± 1.9
FLTrust	32.8 ± 8.2	26.0 ± 2.3	88.6 ± 14.7	25.9 ± 4.1	52.2 ± 4.6	22.2 ± 4.4	54.7 ± 8.2	22.4 ± 4.8
HTA (proposed)	41.1 ± 21.0	35.6 ± 6.2	98.3 ± 0.0	36.0 ± 2.6	98.3 ± 0.0	35.5 ± 9.0	94.2 ± 0.8	34.2 ± 9.8

HTA, $4.4 \pm 9.3\%$ for Krum, $9.8 \pm 7.1\%$ for FLTrust and $11.3 \pm 11.1\%$ for Multi-Krum. Downweighting is therefore doing more work than exclusion, and the difference between rules lies in what they charge the honest hospitals for it.

Figure 11 shows why, and it is the most important negative result in this paper. The reputation the rule assigns to honest and to malicious hospitals is plotted over the sixty rounds. After an initial transient the two bands converge and remain within roughly 0.02 of each other, with their one-standard-deviation envelopes overlapping across the entire run. The honest band is marginally the higher of the two from about round ten onwards, which is the right direction and no more than that. *Under label flipping the trust mechanism does not separate the two populations.* This is consistent with the false-positive rate above, with the ablation in Section 6.7, and with the modest and mostly within-spread margins of Section VI-B: a rule cannot exclude reliably on a signal that does not distinguish, and the honest reading of the evidence is that the separation the method is built on does not materialise on a data-level attack under severe heterogeneity. The reason is the interaction this paper set out to study. A label-flipping hospital submits an update that is neither anomalous in direction nor in magnitude — it is a competently trained model of the wrong labels — and an honest hospital with a lopsided cohort submits one that is anomalous in both.

6.6 Fairness across hospitals

A federation is only useful to a hospital that the shared model actually serves, so Table 9 and Figure 12 report the mean across hospitals together with the worst-served one. The means are tightly bunched: every rule except Krum lies between 67.0 and

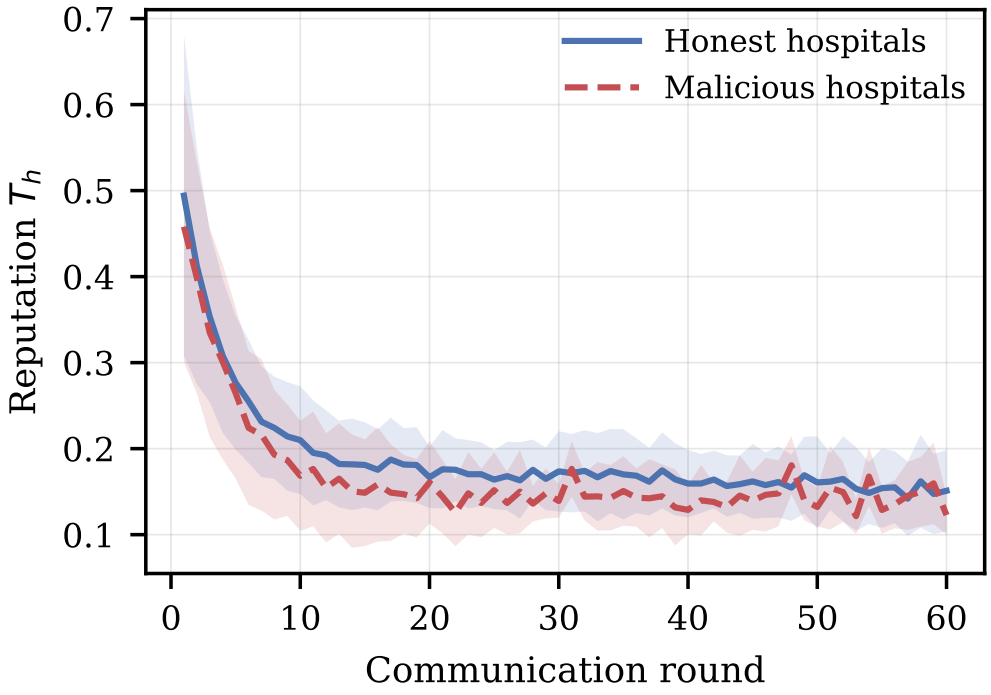


Figure 11 Reputation assigned to honest and to malicious hospitals over training, under label flipping with 20% adversaries. Separation between the bands is what the rule depends on; here they overlap for the whole run, which is the mechanism behind the false-positive rate in Table 8.

69.0, a range narrower than the seed spread, so on average all of them serve the federation equally well.

The worst-served hospital separates them, and not in the proposed rule's favour. FLTrust reaches 42.6 ± 8.7, trimmed mean 41.3 ± 4.3, FedProx 39.7 ± 3.5 and HTA 39.5 ± 10.1, which places the proposed rule fourth on a margin of 3.1 points that is well inside the spread — a tie rather than a loss, but not the advantage a trust-weighted rule might be expected to deliver. Its standard deviation of 10.1 is the largest in the column and is itself the finding: which hospital ends up worst served varies substantially from seed to seed, so the rule's effect on the least-served site is not stable. Krum is the outlier at 13.6 ± 6.8, twenty-six points below the next rule and a model that is unusable at the hospital it serves least well, which is the same single-update selection behaviour seen in Section VI-C expressed per site.

6.7 Ablation

Table 10 and Figure 13 remove one trust signal at a time, under a data-level attack and an update-level one, against the full rule restricted to the same two seeds the ablation uses.

Table 9 Accuracy across hospitals, each evaluated on its own case mix.

Rule	Setting	Mean	Worst hospital	Spread
FedAvg	clean	69.1 ± 2.5	40.2 ± 2.5	52.0 ± 2.6
	attack	69.0 ± 1.7	38.2 ± 2.4	54.2 ± 3.0
FedProx	clean	69.0 ± 2.6	40.0 ± 0.9	51.8 ± 0.7
	attack	69.0 ± 2.0	39.7 ± 3.5	52.4 ± 4.2
Coord. Median	clean	69.1 ± 1.8	43.4 ± 4.0	46.2 ± 7.8
	attack	67.0 ± 1.4	36.4 ± 6.1	53.8 ± 8.2
Trimmed Mean	clean	70.4 ± 1.9	46.7 ± 3.2	43.7 ± 7.0
	attack	68.4 ± 1.0	41.3 ± 4.3	48.7 ± 8.6
Krum	clean	60.2 ± 2.0	10.2 ± 3.5	81.0 ± 1.1
	attack	59.9 ± 3.2	13.6 ± 6.8	75.0 ± 7.0
Multi-Krum	clean	71.1 ± 2.0	50.0 ± 1.3	40.1 ± 5.3
	attack	68.9 ± 0.7	36.7 ± 3.4	55.3 ± 6.9
FLTrust	clean	69.4 ± 1.9	46.3 ± 3.6	43.0 ± 3.9
	attack	68.3 ± 1.6	42.6 ± 8.7	46.7 ± 12.9
HTA (proposed)	clean	70.7 ± 1.3	45.6 ± 4.4	43.5 ± 9.9
	attack	68.7 ± 2.2	39.5 ± 10.1	48.5 ± 14.3

Two components carry the rule, and they are the same two under both attacks. Removing the reputation memory costs 8.0 points under label flipping and 5.7 under scaling; removing directional agreement costs 6.6 and 3.4. Both margins are outside the seed spread in both columns, so the cross-round accumulation and the agreement signal are doing the work.

Two more earn their place against the update-level attack only. Norm clipping costs 2.7 points under scaling and root-set validation 2.1, both outside the spread there; under label flipping the same two cost 1.7 and 1.0, which is inside it and therefore nothing. This is the expected pattern rather than a surprise: clipping and a magnitude-sensitive validation have something to catch when the attack manipulates the size of an update, and nothing to catch when the adversary trains honestly on corrupted labels.

The exception is magnitude plausibility, which does not earn its place anywhere in this study. Removing it changes balanced accuracy by -0.6 points under label flipping and $+2.1$ under scaling, and both figures sit inside their seed spreads — the $+2.1$ carries a standard deviation of 5.6. We therefore do not claim that dropping it improves the rule, only that four seeds give no evidence it contributes. That is a negative result about one of the three signals the method fuses, and the fusion is multiplicative, so a signal that is merely uninformative is not free: any factor near zero suppresses a hospital regardless of the other two. Under a Dirichlet partition an honest hospital with a small or lopsided cohort produces an update whose norm is genuinely unusual, which is precisely when that factor fires on heterogeneity rather than on hostility.

6.8 Transfer to a second modality

The second cohort is binary chest radiography rather than seven-class dermoscopy, and the proposed rule does not transfer to it (Table 11). HTA is last of the four rules evaluated in all three conditions: 74.2 ± 13.2 with no attack, 68.3 ± 10.1 under label flipping and 68.5 ± 10.6 under the backdoor. FLTrust leads each one at 87.0 ± 0.2,

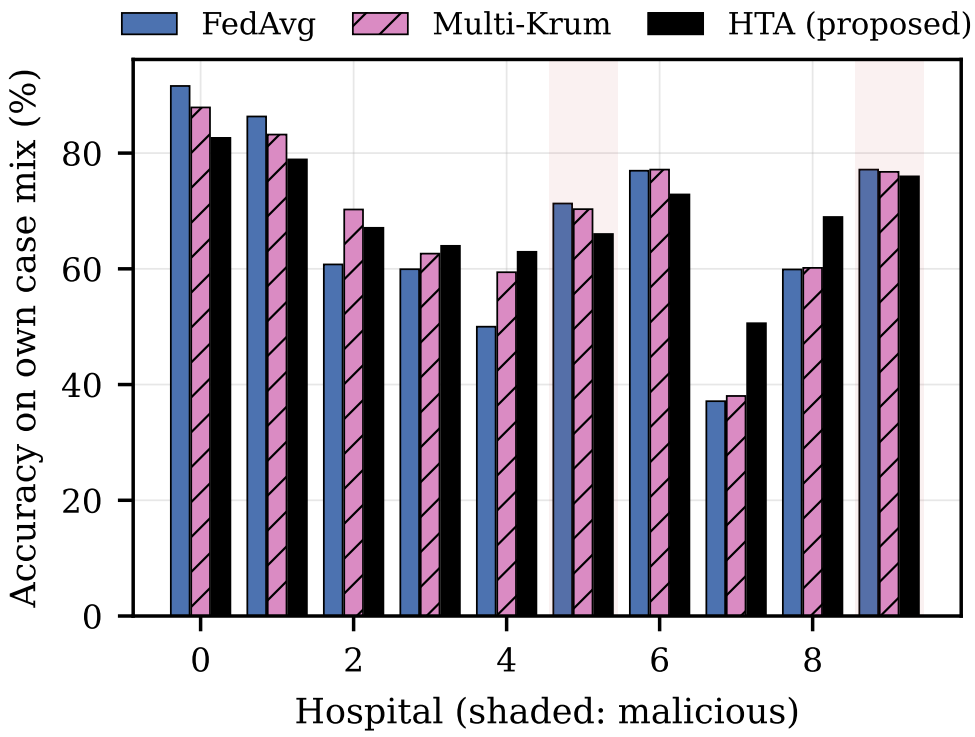


Figure 12 Accuracy of the final global model on each hospital's own case mix under label flipping. The worst bar matters more than the mean: a federation is only useful to a hospital the shared model actually serves.

Table 10 Ablation: balanced accuracy (%) and ASR with each trust signal removed in turn.

Variant	Label flip		Scaling	
	Bal. acc.	ASR	Bal. acc.	ASR
HTA (full)	47.4 ± 3.5	18.7 ± 9.5	42.4 ± 0.2	56.9 ± 12.5
w/o directional agreement	40.8 ± 0.3	28.6 ± 12.2	39.0 ± 0.8	52.5 ± 15.0
w/o magnitude plausibility	46.8 ± 4.5	22.7 ± 14.2	44.6 ± 5.6	52.6 ± 30.4
w/o root-set validation	46.3 ± 1.3	20.4 ± 3.4	40.3 ± 1.2	55.8 ± 13.9
w/o norm clipping	45.7 ± 4.3	21.7 ± 6.2	39.8 ± 0.3	63.2 ± 9.3
w/o reputation memory	39.3 ± 1.4	31.7 ± 4.2	36.8 ± 2.2	61.1 ± 14.3

88.0 ± 1.7 and 87.3 ± 2.2 respectively. The gaps — 12.8, 19.7 and 18.8 points — are two to three times the pooled seed spread, so unlike most comparisons in this paper these are unambiguous losses. FedAvg also beats the proposed rule in every condition.

The standard deviations are as informative as the means. HTA carries a spread of 10 to 13 points where FLTrust carries 0.2 to 2.2, so the rule is not merely worse on

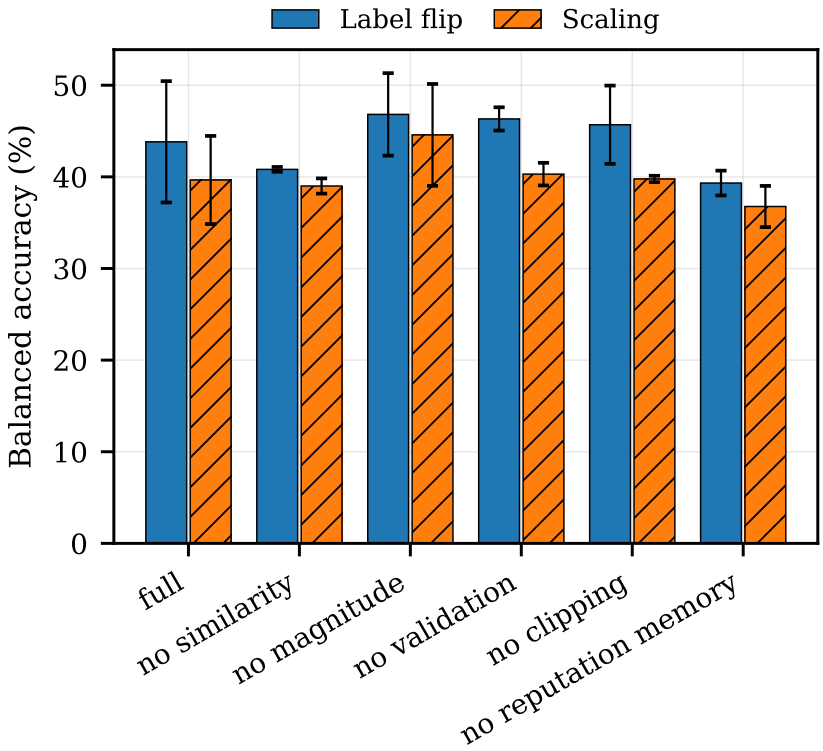


Figure 13 Contribution of each trust signal, measured by removing it, under a data-level attack (label flipping) and an update-level one (scaling).

this cohort but unstable on it, with the outcome depending heavily on which hospitals the seed makes adversarial.

The absolute level is not the problem, which is worth stating so the comparison is not mistaken for a failure of the federation. On clean PneumoniaMNIST the rules reach 80.2–89.1% accuracy at 0.942–0.956 ROC-AUC, against a published centralized ResNet-18's 85.4% and 0.944; FLTrust exceeds that reference on both. What the proposed rule loses here it loses to other federated rules, not to the difficulty of the task.

The structural difference between the two cohorts explains the direction. With two classes rather than seven, a hospital's Dirichlet share is a point on a line rather than a simplex, and the balanced accuracy of a local model on a shared reference cohort — the validation signal — has far less room to distinguish an unusual case mix from a corrupted one. The exclusion test then fires on hospitals that differ only in prevalence. FLTrust, which scores a single cosine against a server update and rescales rather than excluding, has no such threshold to cross, and its near-zero variance across seeds reflects that. The result is a direct instance of the tension this paper set out to examine, and on this cohort it tells against the rule we proposed.

Table 11 PneumoniaMNIST: balanced accuracy and ASR.

Rule	No attack		Label flip		Backdoor	
	Bal. acc.	ASR	Bal. acc.	ASR	Bal. acc.	ASR
FedAvg	78.8 ± 1.2	0.0 ± 0.0	75.0 ± 0.3	49.7 ± 0.9	79.1 ± 0.9	98.6 ± 0.4
Multi-Krum	79.3 ± 1.9	0.0 ± 0.0	69.3 ± 12.1	60.9 ± 24.9	72.9 ± 5.0	98.0 ± 2.1
FLTrust	87.0 ± 0.2	0.0 ± 0.0	88.0 ± 1.7	18.9 ± 3.6	87.3 ± 2.2	76.9 ± 12.0
HTA (proposed)	74.2 ± 13.2	0.0 ± 0.0	68.3 ± 10.1	63.3 ± 20.5	68.5 ± 10.6	67.1 ± 24.0

Table 12 Computation and communication cost per round.

Aggregation rule	Round time (s)	Overhead	Clean bal. acc. (%)
FedAvg	13.0 ± 0.3	1.00×	42.3 ± 1.6
FedProx	13.9 ± 0.3	1.07×	42.9 ± 3.3
Coord. Median	12.9 ± 0.0	0.99×	43.8 ± 3.0
Trimmed Mean	13.1 ± 0.1	1.01×	44.7 ± 4.0
Krum	13.0 ± 0.1	1.00×	26.1 ± 7.1
Multi-Krum	13.0 ± 0.0	1.00×	48.2 ± 2.7
FLTrust	16.0 ± 0.4	1.23×	43.2 ± 4.6
HTA (proposed)	14.0 ± 0.1	1.08×	46.8 ± 2.9

6.9 Cost (RQ6)

Every rule transmits the same 218,583 parameters per hospital per round, so communication cost is identical by construction and the only question is server-side computation (Table 12). Coordinate-wise median, Multi-Krum, FedAvg, Krum and trimmed mean are indistinguishable at 12.9–13.1 s per round. HTA costs 14.0 ± 0.1 s, about 7.7% over FedAvg's 13.0 ± 0.3 s, which buys the three trust signals and the reputation update. FLTrust is the most expensive at 16.0 ± 0.4 s, roughly 23% over FedAvg, because it trains a server-side reference update every round — a genuine extra optimisation rather than a scoring pass over vectors already received.

None of these differences would govern a deployment decision. At ten hospitals the whole grid is dominated by local training, and a rule that costs a tenth of a second more per round is not meaningfully more expensive than one that does not. The cost that matters in this study is statistical, not computational.

7 Discussion

7.1 Why three signals rather than one

Each signal answers a different attack, and none answers all of them. Label flipping changes neither the direction nor the norm of an update by much — the adversary trains honestly on corrupted data, so its update is an ordinary gradient for a slightly different objective — but it damages what the model does, which is what the reference-cohort signal v_h measures. Model replacement by scaling is conspicuous in norm and invisible to a validation signal computed on the local model before scaling is applied,

so m_h is what catches it. Sign flipping and IPM invert the update direction, which only s_h sees. A rule built on any one of these is not partially robust; it is blind to whichever family the signal does not cover, and an adversary who knows the rule chooses that family.

That is the argument the rule was designed on, and the ablation supports it for s_h and for the reputation built over it, but not for m_h. Removing the magnitude signal leaves balanced accuracy unchanged within the seed spread under both attacks, including the scaling attack it was introduced to catch. The likely reason is that it is not the only signal watching that family: updates are norm-clipped to the median before weighting, and clipping alone removes most of the leverage a scaled update has, so by the time m_h is evaluated the conspicuous norm has already been bounded. The coverage argument above is therefore sound in principle and redundant in this implementation, and a two-signal rule with clipping retained is what the evidence supports.

7.2 Why a geometric rather than an arithmetic fusion

The choice of $\tau_h = s_h^a m_h^b v_h^c$ over a weighted sum is the single most consequential design decision in the rule. Under an arithmetic fusion, a hospital scoring zero on one signal can be carried by the other two: an adversary that keeps its update plausibly sized and well-aligned buys enough credit to survive a collapsed validation score. Under a geometric fusion a near-zero factor drives the product towards zero regardless of the others, so every signal holds a veto. The cost is symmetric and must be acknowledged: an honest hospital that scores poorly on one signal — for instance a small site whose update norm is legitimately atypical — is also suppressed. That cost is what the relative floor in (10) is designed to bound.

7.3 Why reputation carries a variance penalty

A single round is weak evidence: honest updates vary with batch order and with which classes a hospital happens to hold. Smoothing into an exponential moving average makes the judgement depend on a hospital's record rather than on one sample of it. But an average alone admits an obvious strategy — alternate between benign and hostile rounds so that the mean looks unremarkable. The variance penalty $\exp(-\gamma\sqrt{\mathrm{Var}(\tau_h)})$ charges precisely that behaviour, and it is the component that makes the reputation a statement about consistency rather than about central tendency.

7.4 Why the exclusion test has two conditions

A median/MAD test is scale invariant, which is a virtue for detecting outliers and a defect for deciding exclusions. On an attack-free round the test still fires on whichever hospital sits lowest in the tail, and in a heterogeneous federation that is the site with the most unusual case mix — typically the specialist centre holding the rare conditions the federation most needs. Requiring a hospital to be *materially* below consensus as well ($T_h < \kappa\mu$) costs nothing when an adversary is genuinely separated and prevents this steady erosion when none is present. This is why the rule weights rather than selects: unlike Krum, no honest hospital's data is discarded merely for being atypical.

7.5 Why clipping is retained despite the trust weights

Clipping to the round's median norm is redundant when the trust score already suppresses an inflated update, and decisive when it does not. It bounds the damage from a single round in which an adversary has not yet accumulated a poor reputation — notably during the warm-up, when no exclusion occurs at all. Sun *et al.* [15] report norm clipping alone as a surprisingly effective backdoor mitigation, and the ablation is the test of whether the rest of the machinery earns its place beside it.

7.6 Detection rate must be read beside its false-positive rate

Neither number means anything alone, and one baseline makes the point concretely: Krum selects a single update and therefore flags nine hospitals of ten, eight of them honest. Its detection rate is high because it excludes almost everyone. This is intrinsic to the rule rather than a defect of our implementation, but it illustrates that a defence can be made to look vigilant by being indiscriminate. The pairing also bounds what can be claimed from a detection rate whose standard deviation across seeds exceeds its mean: such a value is not a low detection rate, it is an unmeasured one, and we report it as such rather than interpreting it.

7.7 What the evidence supports, and what it does not

Collecting Section VI into one place, the proposed rule leads on the two data-level attacks and across every adversary fraction, is level with the field under ALIE and on clean training, and is the only rule with a margin clear of its seed spread under IPM. It is beaten, by margins that survive the spread, on both update-level attacks (Multi-Krum, by 4.9 and 6.7 points), at twenty hospitals (Multi-Krum, 5.7 points), and on the second cohort in every condition (FLTrust, by 12.8 to 19.7 points). Most of its wins are inside the seed spread and are therefore ties; one loss out of that set is not.

Three of these results are worth taking seriously rather than filing as limitations.

The first is the direction of the update-level failures. A rule that reweights every hospital by accumulated trust is beaten by one that selects a coherent subset, consistently, on precisely the attacks that manipulate the geometry of an update. Reweighting preserves a contribution from every participant, which is what makes it attractive under heterogeneity; it also guarantees that a sufficiently determined adversary retains some weight. Selection discards that contribution entirely. On DermaMNIST at ten hospitals the first is preferable; under sign flipping, scaling, and at twenty hospitals, the second is.

The second is the ablation. Four of the five components earn their place — reputation memory and directional agreement under both attacks, norm clipping and root-set validation under the update-level one — but magnitude plausibility does not contribute measurably under either. Since the three trust signals are fused multiplicatively, an uninformative factor is not free: any one of them near zero suppresses a hospital whatever the other two say, and under a Dirichlet partition the hospitals with genuinely unusual update norms are frequently the honest ones with small or lopsided cohorts. The design claim that survives is the weaker one: a reputation accumulated

across rounds, driven mainly by directional agreement, is what carries the rule, and the magnitude signal should be dropped rather than defended.

The third, and the one that bears on the paper's central hypothesis, is Figure 11. We predicted that detectability degrades as heterogeneity sharpens. What the reputation traces show is stronger than that: under label flipping at $\alpha = 0.5$, honest and malicious hospitals are not separated at all, their bands overlapping for the entire run. The rule's advantage on that attack is therefore not obtained by identifying adversaries. It is obtained by damping variance, which is the same mechanism that gives every robust rule here a clean-training premium over FedAvg. That is a coherent explanation of the numbers, and it is not the explanation the method was designed around.

7.8 Clinical interpretation

Two properties of the evaluation are what make it clinically legible rather than merely adversarial. The first is the choice of target: the attack suppresses the rarest of the seven diagnoses in favour of the most common, which is the configuration an informed adversary would choose precisely because pooled accuracy barely moves. The second is the metric: balanced accuracy and worst-class recall register that suppression where accuracy does not. A defence evaluated only on accuracy could hold its headline number while the model quietly ceased to recognise a rare malignancy, and would report success. For a consortium the operative question is not whether the shared model is accurate on average but whether it remains usable at the site with the most unusual case mix — which is why per-hospital and worst-hospital accuracy are reported alongside the pooled figures.

7.9 Limitations

We state these plainly rather than leaving them to be found.

Adversary model. No attack evaluated here is optimised against the proposed rule. Fang *et al.* [18] and Shejwalkar and Houmansadr [21] both demonstrate that defence-aware attacks are substantially stronger than generic ones. Our robustness claim is therefore against the six attacks listed, not against an optimal adversary, and an attacker who knows the three signals could plausibly construct an update that satisfies all of them simultaneously.

Baseline coverage. RFA [11], Bulyan [10], FLAME [13] and DeepSight [14] are not implemented. The last two are the current state of the art for backdoors specifically, so no general claim of backdoor robustness is made here.

Scale and realism. Images are 28×28 and the model has approximately 0.2M parameters; a deployment would use full-resolution studies and a substantially larger backbone. Hospitals are simulated by Dirichlet partitioning, which is the standard protocol [7] but a convention rather than a measured model of inter-institutional variation, and the malicious hospitals are simulated rather than observed. Experiments are single-machine and CPU-only, so the timing comparison is indicative of relative server-side cost and nothing more.

Statistical power. Results are averaged over three seeds for the main grid and two elsewhere, which meets the minimum for this study design but is weak for any narrow comparison. Where a difference between two rules is smaller than the spread across seeds we describe them as comparable rather than ranking them, and any quantity whose standard deviation exceeds its mean is reported as unmeasured at this seed count.

The trusted root cohort. Both the proposed rule and FLTrust require the server to hold a small validated reference cohort. A real consortium would have to assemble one, keep it current as case mix drifts, and agree on its governance; a cohort that is stale or unrepresentative degrades the validation signal precisely when heterogeneity is most severe.

Hyperparameter sensitivity. The trust exponents and the exclusion thresholds are fixed at the values in Table 2 and were not swept. Whether the rule is robust to these choices, or knife-edge sensitive to them, is not established by the experiments reported here.

Privacy. This paper addresses security and heterogeneity, not privacy, following the scope set out for a first study. The omission is not neutral: the rule inspects individual hospital updates, which is in direct tension with secure aggregation [27], whose purpose is to prevent exactly that. A deployment must either forgo one of the two or find a construction that scores an update without revealing it. Differential privacy [28] is likewise deferred, and its interaction with this work is not benign either, since the noise it adds falls hardest on minority-class recall — the quantity this study is trying to protect. Both are substantial problems in their own right rather than extensions to be bolted on.

8 Conclusion

This paper asked whether a trust-aware, multi-signal aggregation rule can reduce the influence of compromised hospitals while retaining honest but statistically unusual ones, under the heterogeneity a real clinical consortium exhibits. The two halves of that question pull against each other, which is what makes it a question: robust aggregation rules work by discarding updates that look unusual, and a severe non-IID partition is precisely the condition that makes an honest specialist centre look unusual. A rule tuned to be safe under attack can therefore quietly exclude the sites that carry the rare diagnoses, and it will do so most eagerly in the federations that need them most.

We approached it in two parts. The first is a reproducible multi-hospital benchmark: two MedMNIST cohorts, seven aggregation rules under identical conditions, six attacks spanning the data-level, update-level and collusion families, five heterogeneity regimes, adversary fractions from 5% to 30%, and federations of 5 to 20 hospitals — with clinical, security, detection and fairness metrics reported together rather than in separate literatures. Because a defence that holds mean accuracy by abandoning a minority diagnosis has protected nothing a clinician would value, balanced accuracy, worst-class recall and per-hospital accuracy are reported beside attack success rate throughout. The second is HTA, which scores each hospital on directional agreement

with a robust consensus, on update-magnitude plausibility, and on class-balanced accuracy over a small curated reference cohort; fuses the three multiplicatively, so that no single manipulable signal can carry a hospital; accumulates them into a reputation penalised for its own volatility; and excludes a hospital only when it is both a statistical outlier *and* materially below consensus. That second condition is the part of the rule aimed squarely at the tension above, and it is why HTA weights hospitals rather than selecting among them.

The grid answers the six questions as follows. Heterogeneity alone costs 16.4 points of balanced accuracy between an identical case mix and $\alpha = 0.1$ (RQ1), more than twice the 7.2 points that adversaries cost (RQ2). Robust aggregation recovers 8.6 points over FedAvg under attack (RQ3) and, contrary to the usual framing, most robust rules *gain* clean accuracy here rather than paying for protection — +5.9 for Multi-Krum and +4.6 for HTA, with Krum's −16.2 the sole exception (RQ4). Performance is near-flat as the compromised fraction rises from 5% to 30% (RQ5), and the computational premium is 7.7% per round over FedAvg against FLTrust's 23%, with communication identical by construction (RQ6).

On the central question — whether a multi-signal, reputation-weighted rule can tell an unusual hospital from a hostile one under severe heterogeneity — the answer this evidence supports is no, not reliably, and not by the mechanism proposed. HTA is competitive and often leads, but its margins are mostly inside the seed spread; it is beaten outside that spread by Multi-Krum on update-level attacks and at twenty hospitals, and by FLTrust on the second cohort in every condition. Its reputation signal does not separate honest from malicious hospitals under label flipping, and its own ablation finds no measurable contribution from one of its three trust signals. What the rule reliably does is damp the variance that heterogeneity introduces, which is valuable and is not the claim it was built to make.

We therefore offer this work as a benchmark and a negative result rather than as a new state of the art. The benchmark is the durable contribution: eight rules, six attacks, five heterogeneity regimes, two cohorts and clinical, security and fairness metrics reported together, with every run reproducible bit for bit. The negative result is twofold: a trust signal built on update magnitude — a standard ingredient in this literature — contributes nothing measurable once cohorts are Dirichlet-skewed, because under that skew an unusual norm marks an honest hospital with a rare case mix as readily as a compromised one; and the reputation that the rule accumulates does not separate honest from malicious hospitals at all under a data-level attack, so what benefit the rule delivers there comes from damping variance rather than from detection. The obvious next step is the rule the ablation points to — reputation over directional agreement, with clipping retained and the magnitude signal dropped — evaluated against an adversary who knows it.

What the evidence can support is bounded in ways worth restating at the close. Robustness is claimed against the six attacks evaluated here, none of which was optimised against the proposed rule; an adversary who knows the three signals could attempt to satisfy them jointly, and that attack is not run. Hospitals are simulated by Dirichlet partitioning, which is the standard protocol but a convention rather than a measured model of inter-institutional variation. The images are 28×28 and the model

small. The trust exponents and exclusion thresholds are fixed rather than swept. Both HTA and FLTrust assume the server holds a validated reference cohort, an assumption a consortium would have to discharge in practice and keep current as case mix drifts. None of these invalidates the comparison, which is made under identical conditions for every rule; each bounds how far the result should be carried.

Three directions follow. The most pressing is the adaptive adversary: an attacker who optimises against the trust signals themselves, together with stealth, Sybil and intermittent behaviour, and with hospitals that drop out mid-training. The second is privacy, which this study deliberately set aside and which does not compose innocently with what is proposed here — per-update inspection is in direct tension with secure aggregation [27], and the noise that differential privacy [28] adds falls hardest on minority-class recall, the quantity this work is trying to protect. Reconciling the two is a problem in its own right rather than an extension. The third is scale and realism: full-resolution studies, larger backbones, additional modalities and multimodal cohorts, and ultimately an evaluation with real institutional partners, where the heterogeneity is observed rather than sampled from a Dirichlet.

Acknowledgements. Not applicable.

Declarations

- **Funding.** No funding was received for conducting this study.
- **Competing interests.** The authors declare no competing interests.

- **Ethics approval and consent to participate.** Not applicable. This study uses only publicly available, de-identified benchmark datasets distributed as part of MedMNIST v2 — DermaMNIST, derived from the HAM10000 dermatoscopic collection, and PneumoniaMNIST, derived from a published paediatric chest-radiograph cohort. No human participants or animals were involved, and no patient-level data were accessed by the authors.
- **Consent for publication.** Not applicable.
- **Data availability.** The datasets analysed are public benchmark collections: DermaMNIST and PneumoniaMNIST are distributed with MedMNIST v2 [35] at https://medmnist.com/, and the code downloads them automatically. No new patient data were generated and no patient-level records were accessed. The data generated by this study — one record per experimental run, each embedding the complete configuration, the per-round metric history and the per-hospital trust scores that produced it — are available from the corresponding author on reasonable request.
- **Materials availability.** Not applicable.
- **Code availability.** The complete federated simulator, the seven aggregation rules, the six attacks, the experiment grid, the analysis pipeline and the scripts that generate every table and figure in this article are available from the corresponding author on reasonable request. Every reported number can be recomputed from the released result records without re-running the experiments, and the analysis is deterministic: runs are reproducible bit for bit on the same device, as described in Section V.
- **Author contribution.** A.K.A. conceived the study, developed the proposed aggregation rule, implemented the simulator and the experimental programme, and

drafted the manuscript. P.C. supervised the work, contributed to the study design and the interpretation of the results, and revised the manuscript. Both authors read and approved the final manuscript.

References

[1] A. K. Agrahari, A. G. Dinker, and R. B. Singh, "A review of security threats and privacy issues in federated learning," *Int. J. Data Sci. Anal.*, vol. 22, no. 1, art. 85, 2026, doi: 10.1007/s41060-026-01067-z.

[2] 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.

[3] P. Kairouz *et al.*, "Advances and open problems in federated learning," *Found. Trends Mach. Learn.*, vol. 14, no. 1–2, pp. 1–210, 2021.

[4] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *Proc. MLSys*, 2020.

[5] S. P. Karimireddy, S. Kale, M. Mohri, S. J. Reddi, S. U. Stich, and A. T. Suresh, "SCAFFOLD: Stochastic controlled averaging for federated learning," in *Proc. ICML*, 2020.

[6] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated learning with non-IID data," *arXiv:1806.00582*, 2018.

[7] T.-M. H. Hsu, H. Qi, and M. Brown, "Measuring the effects of non-identical data distribution for federated visual classification," *arXiv:1909.06335*, 2019.

[8] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, "Machine learning with adversaries: Byzantine tolerant gradient descent," in *Proc. NeurIPS*, 2017, pp. 119–129.

[9] D. Yin, Y. Chen, K. Ramchandran, and P. Bartlett, "Byzantine-robust distributed learning: Towards optimal statistical rates," in *Proc. ICML*, 2018, pp. 5650–5659.

[10] R. Guerraoui, S. Rouault, and E. M. El Mhamdi, "The hidden vulnerability of distributed learning in Byzantium," in *Proc. ICML*, 2018, pp. 3521–3530.

[11] K. Pillutla, S. M. Kakade, and Z. Harchaoui, "Robust aggregation for federated learning," *IEEE Trans. Signal Process.*, vol. 70, pp. 1142–1154, 2022.

[12] X. Cao, M. Fang, J. Liu, and N. Z. Gong, "FLTrust: Byzantine-robust federated learning via trust bootstrapping," in *Proc. NDSS*, 2021.

[13] T. D. Nguyen *et al.*, "FLAME: Taming backdoors in federated learning," in *Proc. USENIX Security*, 2022, pp. 1415–1432.

[14] P. Rieger, T. D. Nguyen, M. Miettinen, and A.-R. Sadeghi, "DeepSight: Mitigating backdoor attacks in federated learning through deep model inspection," in *Proc. NDSS*, 2022.

[15] Z. Sun, P. Kairouz, A. T. Suresh, and H. B. McMahan, "Can you really backdoor federated learning?" *arXiv:1911.07963*, 2019.

[16] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, "How to backdoor federated learning," in *Proc. AISTATS*, 2020, pp. 2938–2948.

[17] A. N. Bhagoji, S. Chakraborty, P. Mittal, and S. Calo, "Analyzing federated learning through an adversarial lens," in *Proc. ICML*, 2019, pp. 634–643.

[18] M. Fang, X. Cao, J. Jia, and N. Z. Gong, "Local model poisoning attacks to Byzantine-robust federated learning," in *Proc. USENIX Security*, 2020, pp. 1605–1622.

[19] G. Baruch, M. Baruch, and Y. Goldberg, "A little is enough: Circumventing defenses for distributed learning," in *Proc. NeurIPS*, 2019, pp. 8635–8645.

[20] C. Xie, O. Koyejo, and I. Gupta, "Fall of empires: Breaking Byzantine-tolerant SGD by inner product manipulation," in *Proc. UAI*, 2019.

[21] V. Shejwalkar and A. Houmansadr, "Manipulating the Byzantine: Optimizing model poisoning attacks and defenses for federated learning," in *Proc. NDSS*, 2021.

[22] C. Xie, K. Huang, P.-Y. Chen, and B. Li, "DBA: Distributed backdoor attacks against federated learning," in *Proc. ICLR*, 2020.

[23] H. Wang *et al.*, "Attack of the tails: Yes, you really can backdoor federated learning," in *Proc. NeurIPS*, 2020.

[24] V. Tolpegin, S. Truex, M. E. Gursoy, and L. Liu, "Data poisoning attacks against federated learning systems," in *Proc. ESORICS*, 2020, pp. 480–501.

[25] L. Muñoz-González, K. T. Co, and E. C. Lupu, "Byzantine-robust federated machine learning through adaptive model averaging," *arXiv:1909.05125*, 2019.

[26] J. Kang, Z. Xiong, D. Niyato, S. Xie, and J. Zhang, "Incentive mechanism for reliable federated learning: A joint optimization approach to combining reputation and contract theory," *IEEE Internet Things J.*, vol. 6, no. 6, pp. 10700–10714, 2019.

[27] K. Bonawitz *et al.*, "Practical secure aggregation for privacy-preserving machine learning," in *Proc. ACM CCS*, 2017, pp. 1175–1191.

[28] R. C. Geyer, T. Klein, and M. Nabi, "Differentially private federated learning: A client level perspective," *arXiv:1712.07557*, 2017.

[29] L. Zhu, Z. Liu, and S. Han, "Deep leakage from gradients," in *Proc. NeurIPS*, 2019.

[30] L. Melis, C. Song, E. De Cristofaro, and V. Shmatikov, "Exploiting unintended feature leakage in collaborative learning," in *Proc. IEEE S&P*, 2019, pp. 691–706.

[31] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms," *arXiv:1708.07747*, 2017.

[32] A. Krizhevsky, "Learning multiple layers of features from tiny images," Univ. Toronto, Tech. Rep., 2009.

[33] S. Caldas *et al.*, "LEAF: A benchmark for federated settings," *arXiv:1812.01097*, 2018.

[34] D. J. Beutel *et al.*, "Flower: A friendly federated learning research framework," *arXiv:2007.14390*, 2020.

[35] J. Yang *et al.*, "MedMNIST v2 – A large-scale lightweight benchmark for 2D and 3D biomedical image classification," *Sci. Data*, vol. 10, art. 41, 2023.

[36] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Sci. Data*, vol. 5, art. 180161, 2018.

[37] D. S. Kermany *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131, 2018.

[38] N. Rieke *et al.*, "The future of digital health with federated learning," *npj Digit. Med.*, vol. 3, art. 119, 2020.

[39] M. J. Sheller *et al.*, "Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data," *Sci. Rep.*, vol. 10, art. 12598, 2020.

[40] I. Dayan *et al.*, "Federated learning for predicting clinical outcomes in patients with COVID-19," *Nat. Med.*, vol. 27, pp. 1735–1743, 2021.