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.

Calibrated, Feature-Attributed Accusation of Influential Response-Curve Edits in Label-Skewed Federated Intrusion Detection with Glass-Box Additive Models

Edson Pacheco

pacheco@ufpr.br

Universidade Federal do Paraná

Carlos Pedroso

Universidade Federal do Paraná

Joao Bittencourt

Universidade Federal do Paraná

Fabio Oliveira

Universidade Federal do Paraná

Research Article

Keywords: Federated learning, Intrusion detection, Poisoning attacks, Accountability, False discovery rate, Explainable boosting machine

Posted Date: November 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11676608/v1

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

Additional Declarations: No competing interests reported.

Calibrated, Feature-Attributed Accusation of Influential Response-Curve Edits in Label-Skewed Federated Intrusion Detection with Glass-Box Additive Models

Abstract

Robust aggregation limits poisoning in federated intrusion detection but names no attacker. We audit one-shot median federations of about twelve label-skewed participants uploading glass-box models. Uploads are scored by cosine distance to the consensus, with p-values from server-trained virtual participants and Benjamini–Hochberg control. Calibrated means at most 10% of clean federations have any false accusation. With the fixed sampler and skew no greater than it assumes, 5G-NIDD and ToN-IoT each had no false accusation in 40 clean federations (bound 0.088). CICIoT2023, the only dataset held out at the data level, had 1 of 40 (bound 0.132). These rates rest on over-dispersed virtual participants and in-sample scoring, and a matched sampler failed (10/30). Feature rescaling (2/90), quantity skew (2/90) and skew beyond the sampler's (1/30) caused false accusations, so members need one flow exporter. Sign-flip recall at malicious fraction 0.2 was 0.90–1.00. Accusations name the feature departing most from the consensus, an edited one for 149 of 150 accused four-random-feature edits (benign rate 0.10–0.39). Large edits ignoring the audit were accused at recall 0.92–1.00, weak ones at 0.04–0.19, and cautious colluders tuned to a surrogate audit at 0.00–0.03. Up to fraction 0.2 the median kept the colluders' damage below one point, so the audit adds accountability for non-adaptive attackers, not robustness. At 0.3–0.4 weak edits left up to 6.9 accuracy points, and other unaccused attacks caused larger class-recall losses. Because a server alone matched the federation's macro-F1 with the audit's pool, the audit suits consortia federating for coverage or data residency.

Keywords: Federated learning, Intrusion detection, Poisoning attacks, Accountability, False discovery rate, Explainable boosting machine

1 Introduction

Network intrusion detection systems for Internet of Things, fifth-generation cellular and industrial deployments are increasingly trained jointly by network operators,

device vendors and edge sites that cannot pool raw traffic. Federated learning, usually with federated averaging (FedAvg), is the mechanism most often proposed for this arrangement [23, 27, 41]. Because decentralized training removes direct control over participant contributions, a compromised or dishonest member can poison its uploaded model [19, 58]. Robust aggregation [11, 67] limits the damage while attackers remain a minority, but it does not identify the attacker, which can remain in the federation and attack again. Excluding a member or opening a formal review is a governance decision for the consortium that operates the shared detector. That decision requires evidence that a specific upload was manipulated at a known error rate.

An audit is useful for accountability only when it is both calibrated and attributable. We use *accusation* for a statistical flag that sends a participant to review rather than for a finding of misconduct. An audit is *calibrated* at level q when, in a federation without attackers, the probability that it accuses at least one participant is at most q. Because finite experiments can bound this probability but cannot prove it, we report each observed rate with an exact upper bound alongside the matching recall. An audit is *attributable* when each accusation identifies which component of the upload is anomalous.

Accountability plausibly matters most against members that do not tailor their edits to the audit, such as members compromised through commodity attack tooling, whose standard manipulations ignore the audit. Naming such a member with a stated error rate gives the consortium evidence for exclusion that a filter never supplies. Against members that know the audit, an audit can at most raise the cost of an attack.

Update-space filters such as FLAME [45] and Multi-Krum [11] give neither feature-indexed evidence nor a stated error rate. Without an attack, FLAME flags many benign clients [17] and Multi-Krum loses accuracy as its tolerance grows [66]. Explanation-based audits such as SHERPA [56] are feature-indexed, but their thresholds are not tied to an error rate. Byzantine machine identification (ByMI) with false-discovery-rate control exists for distributed learning [50], but it assumes independent and identically distributed (IID) data across machines. Under non-IID data a null estimated from the participants' own scores is wide and shifted by the attackers, and with a dozen participants it is also coarse. These limitations raise three questions. Can a server accuse participants of a one-shot federation of about a dozen, up to five of them attackers, with a bounded false-accusation rate (research question RQ1)? Can each accusation state which input feature was manipulated (RQ2)? What can an attacker that knows the audit still achieve, and which attackers does the audit name (RQ3)?

We study a federation in which each participant uploads a glass-box additive model, an Explainable Boosting Machine (EBM) [36, 46], whose per-feature response curves the server aggregates in one shot. The audit scores each upload by its cosine distance to the median consensus, a cosine score as in FLTrust [13] but with the median as reference. It accuses under Benjamini–Hochberg (BH) control and attributes every accusation to a feature. Its p-values come from virtual participants that the server trains on its own reference sample, so that no participant upload enters the null.

The audit requires two operational conditions that a consortium lacks by default, a small labeled reference pool drawn from the members' traffic and a single flow

definition across all members. Both can hold in a sector information sharing and analysis center whose members contribute a curated sample of labeled flows under a sharing agreement while full traffic stays on site. A server holding such a pool matched or exceeded the federation's macro-F1, so the case for federating cannot rest on accuracy over that pool. The members federate for coverage of their own traffic and for data residency instead. Operators, vendors and edge sites run different flow exporters, so the consortium must also standardize the flow definition. One option is a fixed set of IP Flow Information Export [15] information elements with fixed active and idle timeouts.

The main contribution is an audit that accuses participants of small one-shot label-skewed federations at an observed false-accusation rate below a stated level, and it rests on three elements:

1. A simulation-based null free of uploads for small non-IID federations, with a log-t predictive tail that avoids the resolution limit of empirical nulls with a dozen participants.
2. Feature-attributed accusations from glass-box uploads, whose cosine score decomposes exactly over response curves, with a proposition on why signed curves are needed.
3. An assessment of what the audit names and what it misses under a threat model that separates attackers that ignore the audit from attackers that tune their edits to it.

Sections 2–6 describe related work, the threat model, the audit and the experimental setup. Section 7 answers the three research questions, and Section 8 discusses operational implications and limitations.

2 Related work

Table 1 classifies existing defenses by their detection signal, audited representation, false-accusation calibration and feature-level attribution. None of the methods reviewed here combines a null that excludes participant uploads with feature-level evidence.

Optimized attacks that stay near the benign range defeat robust aggregation [6, 19, 58], whose guarantees also degrade under heterogeneous data [29]. Our score belongs to the cosine family of FLTrust [13] and FoolsGold [21], and our null makes FLTrust's assumption of labeled data at the server. The clustering, validation and history-based defenses of Table 1 suppress or down-weight suspicious updates. FLShield [26], attack-agnostic schemes [64] and behavior attestation [2] exclude them from aggregation. MESAS tests six model metrics for symmetry around their median, at a significance level that depends on the data scenario [31]. None of these rules bounds the false-accusation rate across participants, and their evidence is not indexed by input feature.

Among explanation-based audits, SHERPA clusters signed per-sample Shapley additive explanation (SHAP) attributions of each client model using hierarchical density-based spatial clustering of applications with noise (HDBSCAN) [12, 40]. It eliminates clients whose suspicion score reaches the mean [56]. MOAT aggregates

Table 1 Poisoning defenses and explanation-based federated intrusion detection systems. "Calib.": the decision rule targets a stated false-accusation level, with a proof (yes) or with an observed rate only (observed); "Attr.": an accusation names an input feature (both our classification). "Measured here": share of benign participants accused in our one-shot federations. SHAP: Shapley additive explanations; DBSCAN: density-based spatial clustering of applications with noise; HDBSCAN: its hierarchical variant; FDR: false discovery rate.

Work	Detection signal	Representation audited	Calib.	Attr.	Key limitation
Krum / Multi-Krum [11]	Update distance	Parameter vector	no	no	Keeps a fixed number of updates ($K - f$ here)
Trimmed mean, median [67]	Coordinate statistics	Parameter vector	–	no	Filters, does not identify
FLTrust [13]	Cosine to a root-data update	Parameter vector	no	no	Needs labeled root data at the server; weights, does not accuse
FoolsGold [21]	Pairwise cosine similarity	Update history	no	no	Targets sybils; down-weights
FLAME [45]	Cosine distance, clustering	Parameter vector	no	no	Excludes non-majority updates (32–42%, measured here)
DeepSight, FreqFed [20, 51]	Model statistics, frequency components	Parameter vector	no	no	Filters; no feature account
CrowdGuard [52]	Hidden-layer behavior	Model on client data	no	no	Needs client-side validation
FLDetector [70]	Cross-round consistency	Update history	no	no	Needs a history of training rounds
FedDBC [33]	Pairwise cosine, DBSCAN + trimmed mean	Parameter vector	no	no	Assumes honest majority, geometric separability
TrustFed-IDS [57]	Multi-criteria trust (cosine, validation loss, update stability)	Parameter vector, update history	no	no	Down-weights; tested with one sign-flip client of five
ByMI [50]	Split-sample gradient statistics, symmetry threshold	Gradients (two half-samples)	yes, FDR (IID)	no	IID machines; rejects zero or at least $\lceil 1/q \rceil$ participants
Importance-magnitude audit (mean-|SHAP|)	Mean-|SHAP| vector	Explainer importance	no	yes	Blind to sign-flip, nearly to permutation (Proposition 1)
SHERPA [56]	Signed SHAP vectors, HDBSCAN	Explainer attributions	no	partly	Mean-score rule ($\alpha \geq 0.5$: 37–48%, measured here)
FedXAI [28]	SHAP for transparency	Explainer importance	–	–	No poisoning detection
This work	Cosine to median consensus, simulated null	Exact additive response curves	observed	yes	Additive models; labeled server pool; common flow exporter; declared class metadata; low label-flip power; cautious edits tuned to the audit go unnamed

interpreted feature contributions against a baseline input to detect adversaries [39], whereas AgrAmplifier amplifies gradient features [22]. FedXAI securely aggregates the clients' SHAP explanations for transparency and does not audit participants [28]. Because post-hoc explanations produced by an external explainer can be manipulated [4, 5, 60], auditing the intrinsic parameters of a glass-box model removes the explainer altogether. None of these explanation-based methods ties its decision threshold to a nominal false-accusation rate.

Statistical identification of Byzantine participants, as in ByMI [50], pursues a closer goal. Each machine computes gradients $\mathbf{g}_{1k}$ and $\mathbf{g}_{2k}$ on two halves of its sample, and $W_k = (\mathbf{g}_{1k} - \hat{\mathbf{g}})^{\mathrm{T}} \mathbf{\Omega} (\mathbf{g}_{2k} - \hat{\mathbf{g}})$ is large when both deviate from a robust centre $\hat{\mathbf{g}}$ alike. Because its threshold estimates false discoveries above t by one plus the count of statistics below $-t$, any rejection requires at least $\lceil 1/q \rceil$ statistics above t. At $q = 0.1$ and $K = 12$, ByMI therefore accuses either no participant or at least 10 participants, of whom at least five are benign whenever five or fewer attack.

Beyond FLTrust's root-data update, server-side references include a detector trained on benign updates derived from public data [35] and a Weibull fit to participant scores computed on server data [55]. Forensic and provenance methods trace poisoning back to malicious clients, after deployment [24] or in each training round [69], without a false-accusation bound or feature-level evidence.

BH [9] requires valid p-values, and large-scale testing often estimates an empirical null from the bulk of the statistics [18], which in a federation includes the attackers. Conformal outlier tests obtain valid p-values from a clean reference set exchangeable with the null points [7], and our null adapts this construction (Section 4.3).

Federated intrusion detection mostly aggregates neural models by FedAvg [8, 32, 34, 44], some with poisoning-robust designs [61, 65, 71], whereas glass-box additive models need no explainer [53]. EBMs have been applied to network intrusion detection [48, 49], neural additive models expose the same curve interface [1, 14], and differentially private EBMs add formal privacy guarantees [47].

3 System and threat model

We consider cross-silo federated learning with K participants, each holding a private labeled flow-level dataset, that together train a multiclass intrusion detector with a server-side audit (Fig. 1). A subset $\mathcal{M}$ of the participants is malicious, and $\rho = |\mathcal{M}|/K$ is the malicious fraction. An honest-but-curious server performs aggregation and auditing.

The server holds a labeled reference sample of training flows from the participants' capture. It uses 2,000 rows for the model intercept and the baselines and 6,000, with at least 10 per class, for the virtual participants. Whereas FLTrust evaluates a root dataset of fewer than 100 examples [13], this larger reference pool is the primary operational requirement of our audit. In the consortium of Section 1 each member is assumed to contribute a curated sample of its own labeled flows, under an agreement that fixes its size, retention and readers. Contributing labeled attack flows discloses part of each member's attack history, which is a cost of the audit in its own right. Participants also declare their sample counts and the classes they hold, and the server uses both to shape its virtual participants. Because the audit compares response curves feature by feature, all members must share one flow exporter and feature definition, a governance requirement that the consortium imposes.

Glass-box uploads disclose more than the parameters of a black-box model, and the parties learn the following:

- the server receives every participant's response curves, declared classes and sample count, and so learns class presence and volume by protocol;

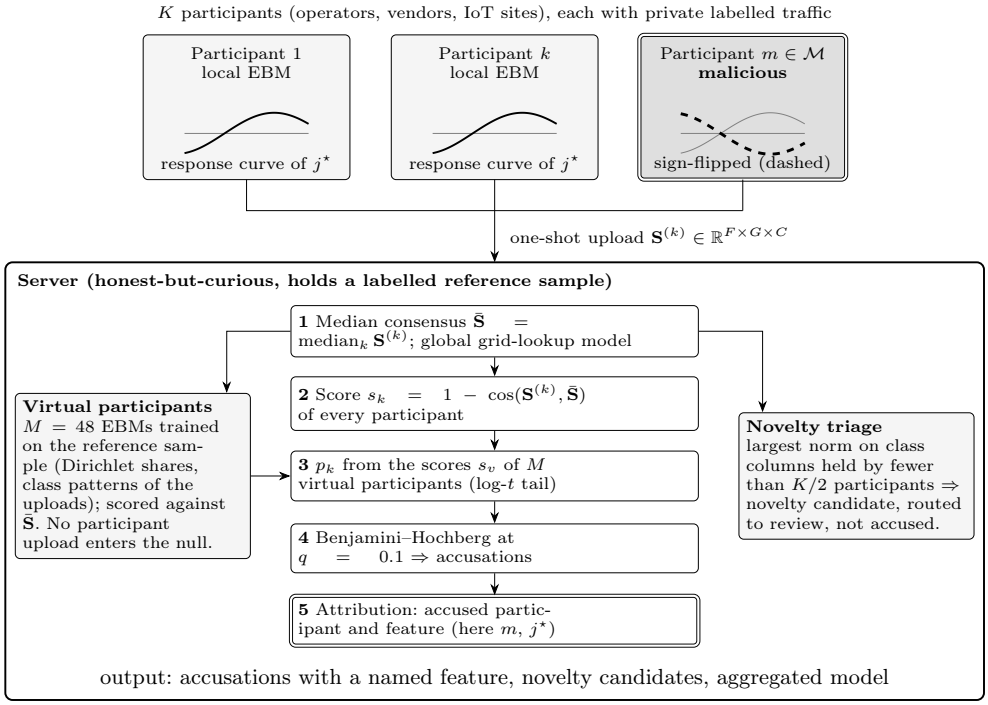

Fig. 1 System architecture and threat model for one-shot federated EBM aggregation, virtual-null calibration, feature-level accusation and novelty triage.

- the other members receive the median model, whose curves summarize the consensus, and, if the consortium circulates them, the identity and named feature of each accused participant;
- an arbiter of a dispute receives the accused upload, the consensus, the reference pool, the declared metadata and the seeds of the virtual participants (Section 8.3).

Because the audit compares individual uploads, it needs them in the clear, so it cannot run under secure aggregation [27], which reveals only the aggregate.

The adversary is a malicious participant that controls its uploaded model rather than only its local training data, as in the standard model-poisoning setting [10, 19, 58]. We evaluate two data-level attacks, label flipping and feature inversion. Our label flipping is a random-relabeling variant of targeted label flipping [62], which relabels the rows of one source class as one target class. We also evaluate two model-level attacks that edit the response curve of the participant's most influential feature $j^\star$ by default. They are curve negation, which we term sign-flip, and grid-order reversal, which we term permutation. Both model-level edits preserve the unsigned magnitude statistics of the modified curve, as shown in Section 5. A targeted evasion-window attack and a stealth blend further probe the detection limits of the audit.

We separate two tiers of attacker by what they know of the audit:

- an uninformed attacker edits its upload without regard to the audit, at a strength of its own choosing, and accountability is claimed only against this tier;
- an informed attacker knows the protocol, namely the median consensus, the grid, the score, M, the sampler and BH at $q = 0.1$, and tunes its edits to a surrogate audit.

Against the informed tier we report the damage that the median leaves unaccused and the measured cost of evading the audit, within the attack families tested. We claim no bound on what an informed attacker can achieve. Both tiers know their own data and upload, colluders share theirs, and they may falsify their declared sample count and classes. Neither knows the other participants' uploads, the server's reference sample or the seeds of the virtual participants. Beyond this threat model, we also evaluate a grid oracle that knows the detection rate of every configuration of an attack family and attacks crafted from every benign upload.

If a new attack class reaches the traffic of a single participant, its local model fits class columns that no peer holds [25]. Treating them as a poisoning anomaly would penalize the participant contributing the rarest operational data, so the server reports a novelty candidate for review alongside the accusations.

4 Method

The novelty triage was not part of the confirmation and was evaluated separately.

4.1 Participant models and one-shot aggregation

Each participant fits an EBM [36, 46], an additive model

$$\ell(\mathbb{E}[y \mid \mathbf{x}]) = \beta_0 + \sum_{j=1}^{F} f_j(x_j), \tag{1}$$

where $\mathbf{x} = (x_1, \ldots, x_F)$ is the feature vector of a flow, y its class, ℓ the link function, β_0 an intercept and f_j the response curve of feature j. We use 64 bins per feature and main effects only. Each curve is read on a shared grid of $G = 40$ points at the quantiles 0.02–0.98 of the server's reference distribution, computed on training data only, and centered to zero mean over the grid. For C classes each curve is a $G \times C$ array, with zero columns for classes absent at a participant, and a single-class participant uploads an all-zero array that the audit excludes. We write $\mathbf{S}^{(k)} \in \mathbb{R}^{F \times G \times C}$ for the upload of participant k and $\mathbf{S}_j^{(k)}$ for the slice of feature j. Because all curves share the grid, the shared detector is obtained in a single round by the coordinate-wise median [67], $\bar{\mathbf{S}} = \text{median}_k\ \mathbf{S}^{(k)}$, and an intercept fitted on the reference sample.

4.2 Score

Participant k is scored by its cosine distance to the consensus, on the flattened arrays,

$$s_k = 1 - \frac{\langle \mathbf{S}^{(k)}, \bar{\mathbf{S}} \rangle}{\|\mathbf{S}^{(k)}\|_2 \, \|\bar{\mathbf{S}}\|_2}. \tag{2}$$

This distance is the classical cosine-to-reference score [13], which FoolsGold applies pairwise [21], here with the robust median as reference. The score is scale-invariant, since multiplying an upload by $\gamma > 0$ leaves s_k unchanged, so the method relies on median aggregation to bound the influence of boosted uploads (Section 7.4). Because $\mathbf{S}^{(k)}$ is indexed by feature, grid position and class, the cosine distance decomposes exactly into a sum of per-feature terms,

$$s_k = \sum_{j=1}^{F} s_{kj}, \qquad s_{kj} = \tfrac{1}{2} \, \|\mathbf{u}_{kj} - \mathbf{a}_j\|_2^2, \tag{3}$$

where $\mathbf{u}_{kj} = \mathbf{S}_j^{(k)} / \|\mathbf{S}^{(k)}\|_2$ and $\mathbf{a}_j = \bar{\mathbf{S}}_j / \|\bar{\mathbf{S}}\|_2$ denote the normalized feature slices of participant k and of the consensus. Each term s_{kj} provides a signed, position-matched comparison between an uploaded response curve and the consensus curve.

4.3 Simulation null free of uploads

Under non-IID data, the null distribution of benign scores is wide and federation-specific, so a null estimated from the participants' own scores forms a baseline that attackers can shift. The server therefore trains $M = 48$ virtual participants on its reference sample using the participants' EBM hyperparameters. Virtual participant v adopts the declared class-presence pattern of real participant $1 + ((v-1) \bmod K)$. For each present class c, the virtual participant receives a share $\eta_c \sim \mathrm{Beta}(\alpha_\mathrm{v}, (K-1)\alpha_\mathrm{v})$ of the estimated class total. That total is the product of the participants' declared total sample count and the reference pool share of class c. This beta distribution matches the marginal share of a K-participant Dirichlet(α_v) label partition with concentration fixed at $\alpha_\mathrm{v} = 0.5$ for all datasets and partitions, except where a matched sampler sets $\alpha_\mathrm{v} = \alpha$. We fixed $\alpha_\mathrm{v} = 0.5$ a priori, when designing the null and before any null was selected. This value was not tuned on the exploration or confirmation splits. Each virtual participant receives at least one row and at most the pool's rows of c, drawn without replacement. The scores s_v of the virtual participants against the real consensus $\bar{\mathbf{S}}$ form the null. No upload enters this null, but besides the reference pool it depends on the consensus $\bar{\mathbf{S}}$ and on the declared class presence and sample counts. Attackers can influence all three.

The construction resembles conformal outlier testing [7], but the virtual participants are not exchangeable with the real ones. They are drawn at a fixed α_v and are smaller, each real upload is scored against a median that contains it, and the p-value below is parametric. No finite-sample guarantee of conformal testing or of BH therefore carries over, and calibration is an empirical property measured in Section 7.

Letting μ and σ denote the sample mean and standard deviation of $\log s_v$, we compute the predictive p-value as

$$p_k = 1 - T_{M-1}\left(\frac{\log s_k - \mu}{\sigma\sqrt{1+1/M}}\right), \tag{4}$$

where T_{M-1} is the Student-t distribution function with $M-1$ degrees of freedom and $\sqrt{1+1/M}$ accounts for estimating μ and σ from M draws.

The tail is parametric because, in a small federation, an empirical null cannot reach the significance threshold required by multiple-testing control. An empirical p-value computed from M draws is bounded below by $1/(M+1)$, which equals 0.020 for $M = 48$. Under BH control at level q across K participants, a rejection occurs only if, for some r, at least r ordered p-values do not exceed rq/K. This rule requires a threshold of 0.0083 for a single rejection at $q = 0.1$ and $K = 12$. An empirical null with $M = 48$ therefore cannot accuse a single attacker regardless of how extreme its score is. A single rejection requires $M \geq K/q - 1$ ($M \geq 119$ at $K = 12$), every draw costs a model fit, and the same limit binds any permutation or bootstrap null built from a dozen participants. A counting argument of the same kind sets ByMI's minimum rejection set of $\lceil 1/q \rceil$ participants (Section 2). The log-t tail removes this resolution floor at the cost of a parametric tail assumption.

4.4 Decision, attribution and novelty triage

The accused set is the BH rejection set at $q = 0.1$ across the K participant p-values [9], as summarized in Algorithm S1 of Online Resource 1 (Section S15). In a clean federation, every accusation is false. The BH procedure then bounds the probability of making any false accusation by q whenever the input p-values are valid and independent [9]. Here they are not exactly valid and scores against a common consensus are dependent, so the bound is checked empirically. Each real participant is scored against the median of all uploads, its own included, whereas a virtual participant is scored against a median that excludes it. We keep the in-sample score because including an upload in its own consensus lowers benign scores and adds margin against false accusation. This choice lowers recall, which Section 7.1 measures against the leave-one-out score.

Every accusation carries two attributions. The cosine attribution is $\arg\max_j s_{kj}$ in (3). Benign participants also deviate most on their own most influential feature, so the cosine attribution often names that feature with or without an attack. The standardized attribution, $\arg\max_j z_{kj}$, discounts this baseline, where z_{kj} is the robust z-score across participants of $\|\mathbf{S}_j^{(k)} - \bar{\mathbf{S}}_j\|_2$. The evidence package presents the cosine attribution, which named the edited feature more often on accused uploads (Section 7.3), and reports the standardized one beside it. It also reports the share of the largest positive z_{kj} in the sum of the positive ones, which is low when the deviation spreads over many features.

For novelty triage, the server computes for every participant the norm of its upload on the class columns held by fewer than half of the participants. The participant with the largest norm is a novelty candidate, routed to review and not accused.

5 Magnitude summaries under curve edits

A sign-flip reverses the model's response to a feature while leaving its unsigned importance unchanged. Let the effect of feature j be its centered curve $\mathbf{r} = (r_1, \ldots, r_G) \in \mathbb{R}^G$, per class column. An *importance audit* compares scalars $\phi(\mathbf{r})$ across participants, canonically $\phi_{\mathrm{SHAP}}(\mathbf{r}) = \frac{1}{G}\sum_i |r_i|$, the mean |SHAP| value [37, 38] of an additive feature on equally weighted grid cells. A functional ϕ is a *signed-multiset functional* if it depends on $\mathbf{r}$ only through the multiset $\{\{r_1, \ldots, r_G\}\}$. It is an *absolute-multiset functional* if it depends only on $\{\{|r_1|, \ldots, |r_G|\}\}$, as the mean absolute value and the Euclidean norm do. For a permutation π of the grid positions, the permutation attack is $A_\pi(\mathbf{r}) = (r_{\pi(1)}, \ldots, r_{\pi(G)})$, of which the reversal $\pi(i) = G + 1 - i$ is evaluated, and the sign-flip attack is $A_-(\mathbf{r}) = -\mathbf{r}$.

Proposition 1 (Invariance of magnitude summaries) *Let* $\mathbf{r} \in \mathbb{R}^G$.

(a) For every permutation π *and every signed-multiset functional* ϕ, $\phi(A_\pi(\mathbf{r})) = \phi(\mathbf{r})$.
(b) For every absolute-multiset functional ϕ, $\phi(A_-(\mathbf{r})) = \phi(\mathbf{r})$. *The same holds for every weighted mean* $\sum_i w_i|r_i|$ *with weights* $w_i \geq 0$, *in particular for the data-weighted mean* |SHAP| *importance.*
(c) Let $\mathbf{u}_j$ *and* $\mathbf{a}_j$ *be the normalized slices of (3),* $\mathbf{u}_j$ *that of the attacked feature. Under the sign-flip the feature term of the cosine score changes by* $2\langle \mathbf{u}_j, \mathbf{a}_j \rangle$, *under a permutation* P *by* $\langle \mathbf{u}_j - P\mathbf{u}_j, \mathbf{a}_j \rangle$, *and all other terms are unchanged.*

Proof (a) A permutation leaves the multiset of entries unchanged. (b) Negation preserves $|r_i|$ entrywise. (c) Both attacks preserve $\|\mathbf{S}_j^{(k)}\|_2$ and hence $\|\mathbf{S}^{(k)}\|_2$, so only the term of feature j moves, with $\frac{1}{2}\|-\mathbf{u}_j - \mathbf{a}_j\|_2^2 - \frac{1}{2}\|\mathbf{u}_j - \mathbf{a}_j\|_2^2 = 2\langle \mathbf{u}_j, \mathbf{a}_j \rangle$ and $\frac{1}{2}\|P\mathbf{u}_j - \mathbf{a}_j\|_2^2 - \frac{1}{2}\|\mathbf{u}_j - \mathbf{a}_j\|_2^2 = \langle \mathbf{u}_j - P\mathbf{u}_j, \mathbf{a}_j \rangle$ since $\|P\mathbf{u}_j\|_2 = \|\mathbf{u}_j\|_2$. □

When the benign curve of a feature aligns positively with the consensus so that $\langle \mathbf{u}_j, \mathbf{a}_j \rangle > 0$, a sign-flip strictly increases the cosine score. So does the reversal of a monotone curve aligned with it. In the design and exploration federations of MQTTset, 5G-NIDD and ToN-IoT the sign-flip left the importance of all 360 attacked uploads unchanged. The permutation, invariant only on grid values, changed it by 3–9% on average. The proposition covers magnitude summaries, not signed per-sample attributions such as SHERPA's. Part (c) shows that both attacks move exactly the term that the cosine attribution reports.

Table 2 Datasets: flows in the sample, features after preprocessing, classes, smallest class and role in the evaluation (Section 6.3).

Dataset	Flows	Features	Classes	Smallest class (flows)	Role
MQTTset	60,000	16	6	`flood` (121)	design set, seeds 0–9
5G-NIDD	40,000	41	9	`ICMPFlood` (44)	exploration seeds 0–9; confirmation seeds 10–19
ToN-IoT	40,000	15	10	`mitm` (213)	exploration seeds 0–9; confirmation seeds 10–19
CICIoT2023	40,000	39	8	`BruteForce` (280)	confirmation only, seeds 0–9

6 Experimental setup

6.1 Datasets

We evaluate MQTTset [63], 5G-NIDD [54, 59], ToN-IoT [42] and CICIoT2023 [43], whose sample sizes, feature counts, class counts and evaluation roles appear in Table 2. Preprocessing coerced fields to numeric values and dropped all-missing and constant or near-constant columns. It filled remaining gaps with the column median and standardized the features. In ToN-IoT, we removed addresses, ports and protocol, as well as service, connection state and the binary label, because they predict the attack type directly. The 34 CICIoT2023 labels were mapped to the standard eight families. Its 40,000 flows were drawn from a class-capped file of at most 12,000 rows per raw label, so family priors differ from the full capture. From the fixed sample of a dataset, each seed drew the train/test split and the reference sample, and drew the Dirichlet partition and the malicious set. Because 5G-NIDD concatenates flows from two base stations of a single testbed [59], its row index supports a station-level split. No other dataset provides a field that separates captures, sensors or time periods with all classes present on both sides.

6.2 Federation, attacks and baselines

The $K = 12$ participants received a Dirichlet label partition [68], in which the rows of each class are split across participants in proportions drawn from a Dirichlet distribution, with $\alpha \in \{0.1, 0.5, 1, 100\}$, from strongly non-IID to nearly IID. The malicious fraction $\rho \in \{0, 0.1, 0.2, 0.3, 0.4\}$ corresponded to 0, 1, 2, 4 or 5 attackers, with malicious sets nested across ρ within each of the 10 seeds per configuration. To evaluate feature attribution at $\rho \in \{0.1, 0.2\}$, the model-level attacks also modified either one random non-dominant feature, the least influential feature with a non-zero curve, four random features or all features. Label flipping reassigned 45% of the attacker's rows to a random alternative class, whereas feature inversion negated the input feature with the largest analysis-of-variance F-statistic in the attacker's local data. The evasion-window attack set the curve of the attacker's most frequent non-benign class to its minimum on the 8-point window of highest contribution of one feature. The stealth blend uploaded $(1-2\lambda)\mathbf{r}$ for the most influential feature, $\lambda \in \{0.1, \ldots, 1\}$. Four full-knowledge attacks, adapted to curve uploads, crafted uploads from every benign upload (Online Resource 1, Section S6). They were A Little Is Enough (ALIE) [6], Min-Max and Min-Sum [58] and the trimmed mean attack [19].

Informed colluders know the audit protocol and their own local data, but they know neither the benign uploads nor the server pool (Online Resource 1, Section S11). By fitting 16 surrogate virtual participants on their pooled rows using the server's sampler, the colluders estimate a surrogate consensus and null distribution. Each colluder adds $\lambda \|\mathbf{S}^{(k)}\|_2 \, \mathbf{d}$ to its upload along a unit direction $\mathbf{d}$ weighted by the histogram of the colluders' rows across the grid bins. In the targeted variant, $\mathbf{d}$ lowers the column of $c^\star$, the non-benign class with the most test rows, in favor of its strongest competing class. In the untargeted variant, $\mathbf{d}$ lowers every class in turn. A cautious colluder takes the largest $\lambda \leq 20$ whose surrogate p-value stays above q, and an aggressive one the largest above q/K. Fixed strengths $\lambda \in \{0.5, 1, 2, 5, 10, 20\}$, chosen without regard to the audit, complete the strategies. Damage is the mean loss of accuracy or of $c^\star$ recall of the median model against the clean federation, in points, on the 90 confirmation federations. The protection added by the audit is the maximum damage across the fixed strengths without auditing minus the maximum damage across all strategies after excluding the accused uploads.

Further conditions changed only the clean federation, the null or the server pool (Online Resource 1, Section S13). Under feature shift, each real participant multiplied three random features by log-uniform factors drawn from $[1/1.25, 1.25]$, $[1/1.5, 1.5]$ or $[0.5, 2]$ prior to local training. Under quantity skew, each participant trained on a random 25% subset of its assigned rows. Clean federations at $\alpha \in \{0.2, 0.3\}$ lay below the sampler's $\alpha_{\mathrm{v}} = 0.5$, and a matched sampler set $\alpha_{\mathrm{v}} = \alpha$ at $\alpha = 0.2$ and 100. Two colluders with unmodified uploads declared five times their sample count and every class, and two further samplers used no declarations. Server pools varied in size down to 10 rows per class, in class mix and, on 5G-NIDD, in base station.

On the same uploads, we compared the audit against cosine distance to the coordinate median with the FLTrust zero-trust threshold $\cos \leq 0$ [13], FLAME's HDBSCAN majority rule [45] and Multi-Krum with $f = 4$ [11]. We also evaluated a mean-|SHAP| importance audit, SHERPA [56] and three participant-based nulls (Online Resource 1, Section S16). These nulls were a two-pass bootstrap, a leave-one-out Student-t null and our own quantile-based variant of the empirical null idea [18]. Under the simulation null, we additionally tested the position-matched sum-L_2 distance over feature slices and the Multi-Krum distance. We adapted ByMI to federations of $K = 48$ participants that each fit two EBMs on disjoint data halves with unit-length uploads and $\mathbf{\Omega} = \mathbf{I}$ (Online Resource 1, Section S7).

For accuracy, we compared the additive model with a FedAvg multilayer perceptron tuned on the seed-0 test accuracy, which favors it. Federated logistic regression and centralized EBM and LightGBM [30] models, together with local-only and server-only models, completed the comparison (Online Resource 1, Section S16).

6.3 Design, exploration and confirmation splits

We developed the audit on a design set, selected its null on an exploration split and then evaluated it unchanged on a confirmation split. On the design set, MQTTset, none of six families of nulls combined a false-accusation rate of at most 0.10 with a sign-flip recall of at least 0.6 at $\rho = 0.2$. On the exploration split (5G-NIDD and ToN-IoT seeds 0–9) the simulation null with an empirical tail made no false accusation.

The resolution limit of Section 4.3 held its sign-flip recall at $\rho = 0.2$ to 0.00–0.10. The log-t tail with $M = 48$ had no false accusation and sign-flip recall of 0.75–1.00 there. We adopted it after seeing these results, so these seeds support no claim about it. The method was then fixed before any run on the confirmation split, which comprises seeds 10–19 of 5G-NIDD and ToN-IoT and seeds 0–9 of CICIoT2023. For 5G-NIDD and ToN-IoT the new seeds draw new splits, reference samples, partitions, malicious sets and virtual participants from the same 40,000-flow samples as the exploration seeds. No earlier step had loaded CICIoT2023, which is therefore the only dataset held out at the data level. The confirmation covers $\alpha \in \{0.5, 1, 100\}$ rather than 0.1, $\rho \in \{0, \ldots, 0.4\}$ and the sign-flip, permutation and label-flipping attacks. A further 90 clean federations on unseen seeds and 60 at $\alpha = 0.5$ on new seeds repeated the calibration test, with seeds 20–49 of 5G-NIDD and ToN-IoT and 10–39 of CICIoT2023. Some comparators and controls were added after earlier results had been read, and Online Resource 1 lists them (Section S14).

6.4 Metrics

The primary calibration metric is the any-false-flag rate, the fraction of clean federations ($\rho = 0$) with at least one accused benign participant, with exact Clopper–Pearson intervals [16]. The three α cells of a (dataset, seed) cluster share the train/test split, reference sample, grid, server pool and malicious order but not the partition or the virtual participants. We therefore count (dataset, seed) clusters with any false accusation, and we treat federations as independent only across seeds or within one α cell. Recall is the fraction of malicious participants accused, pooled over seeds, with percentile bootstrap intervals over seeds or over (dataset, seed) clusters. The hit rate of an attribution on accusations is the share of accused attackers whose top-attributed feature lies in the edited set, with exact intervals. Its benign control is the share of benign participants whose top-attributed feature lies in a set chosen by the same rule, and we report the ratio of the two. Ranking is measured by the area under the receiver operating characteristic curve (AUC) per federation, averaged over seeds. Intervals are 95%.

7 Results

7.1 Calibration

Addressing RQ1, none of the 90 clean confirmation federations produced a false accusation (Table 3). At $\alpha = 0.5$ the confirmation, unseen and new seeds give 40 independent clean federations per dataset. On 5G-NIDD and on ToN-IoT none of them had a false accusation, an exact upper bound of 0.088 below $q = 0.1$ (Table 4). CICIoT2023, the only dataset held out at the data level, had 1 of 40, with an upper bound of 0.132. Over the three sets 1 of 120 (dataset, seed) clusters had a false accusation, an upper bound of 0.046 on the average over the three samples (Online Resource 1, Table S16). In the attacked confirmation federations, the per-participant false-positive rate never exceeded 0.02 and pooled accusation precision reached 0.99.

Table 3 Results on the confirmation split ($K = 12$; cosine score, simulation null, BH at $q = 0.1$). Clean: clean federations with a false accusation, of 10 (exact 95% upper bound 0.31 for 0/10); FPR, median p: per-participant false-positive rate and median benign p-value in clean federations. SF, PM, LF: sign-flip, permutation and label-flipping recall (20 attackers per cell at $\rho = 0.2$, 50 at $\rho = 0.4$). MQTTset: design set, for reference.

		Clean federations			Recall, $\rho = 0.2$			Recall, $\rho = 0.4$	
Dataset	α	Clean	FPR	median p	SF [95% CI]	PM	LF	SF	PM
5G-NIDD	0.5	0/10	0.000	0.55	0.90 [0.70, 1.00]	0.55	0.65	0.98	0.88
	1	0/10	0.000	0.87	0.90 [0.70, 1.00]	0.70	0.50	1.00	0.90
	100	0/10	0.000	0.97	0.90 [0.70, 1.00]	0.80	0.60	1.00	0.90
ToN-IoT	0.5	0/10	0.000	0.69	1.00 [1.00, 1.00]	1.00	0.50	1.00	1.00
	1	0/10	0.000	0.85	1.00 [1.00, 1.00]	1.00	0.15	0.92	0.88
	100	0/10	0.000	0.98	1.00 [1.00, 1.00]	0.75	0.00	0.64	0.34
CICIoT2023	0.5	0/10	0.000	0.63	0.90 [0.75, 1.00]	0.55	0.00	0.84	0.36
	1	0/10	0.000	0.91	0.95 [0.85, 1.00]	0.35	0.00	0.98	0.58
	100	0/10	0.000	0.99	1.00 [1.00, 1.00]	0.50	0.00	1.00	0.82
MQTTset	0.5	0/10	0.000	0.61	0.25	0.00	–	0.20	0.00
	1	0/10	0.000	–	0.00	0.00	–	0.00	0.00
	100	0/10	0.000	–	0.00	0.00	–	0.00	0.00

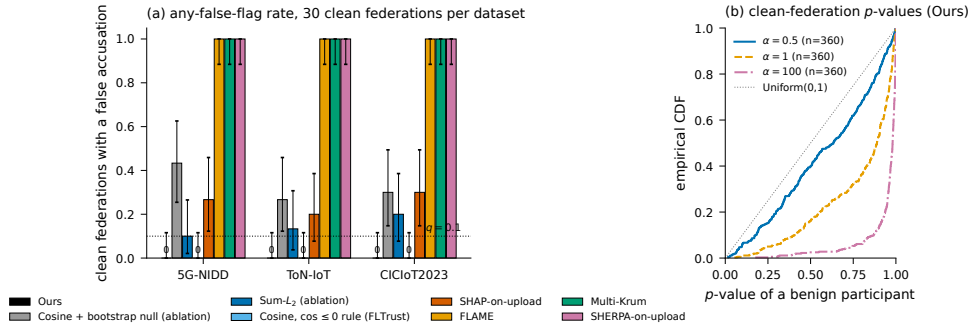

Fig. 2 Calibration on the confirmation split. (a) Share of the 30 clean federations per dataset in which each rule accuses a benign participant, with exact 95% intervals that treat them as independent; dotted: $q = 0.1$. (b) Distribution of benign p-values in clean federations under Ours, against the uniform distribution.

The benign p-values of the clean confirmation federations were larger than uniform (Fig. 2b), and their median rose with α. At $\alpha = 1$ and 100 on every dataset, and at $\alpha = 0.5$ on CICIoT2023, the median benign log-score lay below the median virtual one under both the in-sample and the leave-one-out score. At $\alpha = 0.5$ on 5G-NIDD and ToN-IoT the leave-one-out median gap was positive, 0.072 and 0.018 (Online Resource 1, Tables S13 and S14). Their 0 of 20 false accusations there come without an observed margin. Scoring each real participant against the median of the other uploads (leave-one-out) accused no benign participant in the 90 clean confirmation federations

Table 4 Clean-federation false accusations (Clean) and sign-flip recall at $\rho = 0.2$ (SF) across sampler, heterogeneity and server-pool stress conditions, with exact 95% intervals.

Condition	Units	Clean	SF
$\alpha = 0.5$, all sets, 5G-NIDD / ToN-IoT / CICIoT2023	40 federations each	0/40 [0.000, 0.088] / 0/40 [0.000, 0.088] / 1/40 [0.001, 0.132]	–
All sets, three datasets	120 clusters	1/120 [0.000, 0.046]	–
$\alpha = 0.3$ / 0.2 / 0.1, sampler $\alpha_v = 0.5$	30 / 30 / 10 federations	1/30 [0.001, 0.172] / 0/30 [0.000, 0.116] / 0/10 [0.000, 0.308]	0.75 / 0.50 / 0.05
$\alpha = 100$, matched sampler $\alpha_v = 100$	30 federations	10/30 [0.173, 0.528]	1.00
Feature shift within $[1/1.25, 1.25]$ / $[1/1.5, 1.5]$ / $[0.5, 2]$	90 federations each	2/90, 4/90, 5/90	0.94
Quantity skew (participants fit on 25% of their rows)	30 clusters	2/30 [0.008, 0.221]	0.97
Two colluders declare 5× their count and every class	45 federations	1/45 [0.001, 0.118]	0.98
Server pool of about 510 / 202 / 90 rows	15 federations each	0/15 [0.000, 0.218], 2/15 [0.017, 0.405], 1/15 [0.002, 0.319]	1.00 / 0.87 / 0.83
5G-NIDD, pool from the other base station	30 federations	0/30 [0.000, 0.116]	0.68

Note: CICIoT2023 is the only dataset held out at the data level, and Online Resource 1, Table S25, gives every condition with its reference row.

and raised pooled recall by 0.01–0.13. It raised the benign log-scores relative to the virtual ones by a median of 0.175, so the in-sample consensus supplies part of the margin.

The fixed sampler ($\alpha_v = 0.5$) provides the remainder of the margin (Table 4 and Online Resource 1, Table S17). When participant partitions were more skewed than the sampler ($\alpha < \alpha_v$), a single false accusation appeared at $\alpha = 0.3$ and sign-flip recall declined monotonically as α decreased to 0.1. A sampler matched to nearly IID federations ($\alpha_v = \alpha = 100$) falsely accused in 10 of 30 clean federations, against none with $\alpha_v = 0.5$. Matched at $\alpha = 0.2$, it kept 0 of 30 but lowered sign-flip recall to 0.27 (Online Resource 1, Table S17).

Members that departed from the virtual participants in other ways also produced false accusations (Table 4). Rescaling three features per participant by at most 25% gave 2 of 90, and wider ranges gave 4 and 5 of 90, all on ToN-IoT, with sign-flip recall of 0.94. Quantity skew gave 2 of 90 federations, or 2 of 30 clusters. Falsified declarations gave 1 of 45, against none with the true declarations, in a run that also drew new virtual participants (Online Resource 1, Section S13).

A server pool of about 500 rows with at least 10 per class gave 0 of 15 false accusations with sign-flip recall of 1.00. The 2,000-row pool gave 1 of 15, and pools of 202 and 90 rows gave 2 and 1 of 15, all on ToN-IoT, counts that fifteen federations cannot separate. With the pool from the other 5G-NIDD base station no benign participant was accused in 30 clean federations. Recall matched that of a pool from the participants' own station (Online Resource 1, Section S13).

7.2 Power and other decision rules

Sign-flip recall reached 0.90–1.00 in every confirmation cell at $\rho \in \{0.2, 0.3\}$, whereas permutation recall was lower, particularly on CICIoT2023 (Table 3 and Online Resource 1, Fig. S1). Label flipping was accused inconsistently, with recall up to 0.97 on 5G-NIDD at $\rho = 0.3$ but at most 0.12 on CICIoT2023, where its score AUC was 0.90–1.00. At $\rho = 0.4$ recall fell in some cells (Online Resource 1, Fig. S1). On the MQTTset design set, sign-flip recall at $\rho = 0.2$ did not exceed 0.25 even though the cosine score achieved an AUC of 0.98–1.00. Power also fell with K, to a sign-flip recall of 0.12 on 5G-NIDD at $K = 48$, with no false accusation in 30 clusters at $K = 24$ and 48 (Online Resource 1, Table S4).

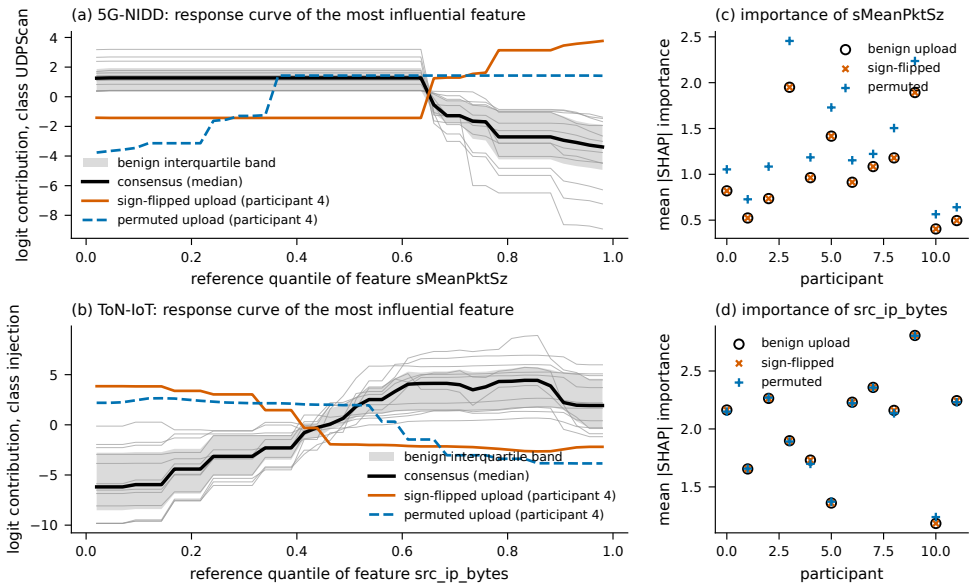

Fig. 3 Response curves of the most influential consensus feature, one class column (exploration split, $\alpha = 0.5$, seed 0): benign participants (gray, interquartile band), consensus (black) and one participant's sign-flipped and permuted uploads on 5G-NIDD and ToN-IoT; (c, d) mean-|SHAP| importance of that feature per participant.

Replacing the simulation null with a two-pass bootstrap null on the same confirmation uploads caused false accusations in 8 to 13 of the 30 clean federations per dataset (Fig. 2a). Under the simulation null, sum-L_2 had sign-flip recall of at most 0.18 at $\rho = 0.2$, and the Multi-Krum distance falsely accused in 1, 9 and 4 of 30 clean federations. At $K = 48$ our adaptation of ByMI accused a benign participant in every clean federation and the audit in none, with McNemar exact $p = 1.9 \times 10^{-9}$. The adapted rule also rejected benign participants in 2 of 9 IID controls even with a linear statistic. It is thus an adaptation that failed and says nothing about ByMI in its own setting (Online Resource 1, Section S7). By their rules FLAME, Multi-Krum and SHERPA excluded 32–42%, 33–35% and, at $\alpha \geq 0.5$, 37–48% of the benign participants in every clean federation (Fig. 2a).

7.3 Attribution

Addressing RQ2, an attribution carries evidence only where it beats the benign rate of the same top-attributed feature (Table 5). Edits of four random features were accused 150 times, and the cosine attribution named an edited feature in 149 of them, against a benign control of 0.10–0.39 per dataset. On ToN-IoT this hit count was 105 of 106 against 0.392, and edits of one random feature, accused 14 times, were named each time against a control of 0.031. For edits of the most influential feature the cosine attribution named it in all 429 accusations, but its benign control of 0.39–0.97 means

Table 5 Attribution by edited feature on the confirmation federations ($\rho \in \{0.1, 0.2\}$, sign-flip and permutation pooled). Hit: share of uploads whose top-attributed feature lies in the edited set, on accusations (exact 95% interval), on all attacked uploads and for benign participants with a rule-matched set.

				Hit on accusations		Hit, cos / std.		Ratio
Dataset (F)	Edited	Recall	Accused	Cosine	Standardized	All uploads	Benign control	cos / std.
5G-NIDD (41)	four random	0.15	26 (12)	26/26 [0.87, 1.00]	26/26 [0.87, 1.00]	0.81 / 0.93	0.106 / 0.117	9.5 / 8.6
	random	0.00	0	–	–	0.50 / 0.80	0.017 / 0.025	–
	most influential	0.70	131 (54)	131/131 [0.97, 1.00]	122/131 [0.87, 0.97]	1.00 / 0.94	0.386 / 0.017	2.6 / 55.9
ToN-IoT (15)	four random	0.58	106 (38)	105/106 [0.95, 1.00]	105/106 [0.95, 1.00]	0.90 / 0.99	0.392 / 0.378	2.5 / 2.6
	random	0.08	14 (4)	14/14 [0.77, 1.00]	14/14 [0.77, 1.00]	0.37 / 0.79	0.031 / 0.064	32.7 / 15.7
	most influential	0.95	171 (60)	171/171 [0.98, 1.00]	147/171 [0.80, 0.91]	1.00 / 0.88	0.661 / 0.161	1.5 / 5.3
CICIoT2023 (39)	four random	0.10	18 (6)	18/18 [0.81, 1.00]	17/18 [0.73, 1.00]	0.43 / 0.85	0.097 / 0.125	10.3 / 7.6
	random	0.00	0	–	–	0.12 / 0.60	0.000 / 0.028	–
	most influential	0.70	127 (57)	127/127 [0.97, 1.00]	102/127 [0.72, 0.87]	0.97 / 0.75	0.972 / 0.019	1.0 / 41.3

Note: Ratio: hit on accusations over the benign control. Accused: accused uploads of the 180 attacked uploads per row, with distinct attackers in parentheses; recall: mean over the two ρ cells. Most influential: the default attack; random: one random other feature; edits of the least influential feature were never accused (Online Resource 1, Table S19).

that it did so largely by default. The standardized attribution discounts this default, with a control of 0.017–0.161, and named the edited feature in 122 of 131, 147 of 171 and 102 of 127 of these accusations. Edits of the least influential feature were never accused, and edits of every feature always were, although no single feature is then the edited one (Online Resource 1, Table S19).

Label flipping manipulates no feature, yet its accusations name one. Among the 121 label-flipping attackers accused at their lowest malicious fraction, the attributed class column had gained rows from the label flip in 49.6% of cases. It represented a newly injected class in 4.1% of cases. The share of the largest positive standardized deviation fell below 0.12 in 131 of 252 label-flip accusations on 5G-NIDD. It did so in none on the other datasets and in 2 of 443 accusations of single-feature edits. In Fig. 3 the sign-flipped and permuted curves leave the benign band while the mean-|SHAP| importance of the attacked feature stays unchanged or almost unchanged.

7.4 Knowledgeable attackers

Addressing RQ3, whether an uninformed directional edit was accused depended directly on its magnitude λ (Tables 6 and 7, and Online Resource 1, Tables S20 and S21). Large fixed-strength edits ($\lambda = 20$) were accused at 0.92–1.00 recall across all malicious fractions, whereas weak edits ($\lambda = 1$) were accused at 0.04–0.19 recall and small edits ($\lambda = 0.5$) at 0.00–0.01 recall. When informed colluders tuned λ against their surrogate audit, cautious colluders were accused at 0.00–0.03 recall and aggressive colluders at 0.43–0.69 recall. At $\rho \leq 0.2$ the strongest edit without the audit caused at most 0.82 points of damage and the damage left under the audit was at most 0.19. At $\rho = 0.3$ and 0.4 the audit removed 4.07–13.53 points of the damage attainable without it. The colluders still removed 1.80 and 6.90 accuracy points there with the weak $\lambda = 1$ edit, which was most often the most damaging strategy under the audit. Targeted $c^\star$ recall loss after exclusion remained at or below 0.25 points at $\rho = 0.4$, partly because the server refits the global intercept on its reference pool. Holding the clean intercept fixed raised that residual targeted loss to 2.48 points.

Table 6 Informed colluders that know the protocol but not the benign uploads and tune one edit strength to a surrogate audit (90 confirmation federations per ρ). Damage in points of $c^\star$ recall (T) and of accuracy (U), with cluster-bootstrap 95% intervals.

ρ	Goal	No audit	Cautious	Aggressive	Under audit (strategy)	Audit adds
0.1	T	0.09 [−0.05, 0.27]	0.03, −0.01	0.43, 0.07	0.07 [−0.03, 0.21] (aggressive)	0.02 [−0.19, 0.22]
	U	0.13 [0.06, 0.21]	0.02, 0.07	0.46, 0.04	0.07 [0.03, 0.11] (cautious)	0.06 [0.01, 0.11]
0.2	T	0.82 [0.17, 1.69]	0.00, −0.03	0.47, 0.07	0.17 [0.02, 0.36] ($\lambda = 10$)	0.66 [0.11, 1.36]
	U	0.50 [0.28, 0.71]	0.00, 0.17	0.44, 0.14	0.19 [0.10, 0.29] ($\lambda = 1$)	0.31 [0.16, 0.45]
0.3	T	5.18 [2.04, 8.99]	0.00, 0.03	0.64, 0.04	0.09 [−0.02, 0.41] ($\lambda = 1$)	5.09 [2.02, 8.65]
	U	5.87 [3.83, 7.91]	0.00, 1.40	0.44, 1.49	1.80 [1.16, 2.45] ($\lambda = 1$)	4.07 [2.49, 5.74]
0.4	T	13.79 [7.08, 20.78]	0.02, −0.02	0.69, 0.08	0.25 [0.02, 0.69] ($\lambda = 1$)	13.53 [7.02, 20.20]
	U	18.40 [12.82, 23.95]	0.01, 4.04	0.49, 4.87	6.90 [4.43, 9.28] ($\lambda = 1$)	11.51 [7.95, 15.41]

Note: No audit: largest mean damage over the fixed edit strengths; cautious, aggressive: attacker recall and damage after excluding the accused; under audit: largest damage after exclusion over all strategies.

In the blend-and-boost family, colluders scaled the slices of their top $d \in \{1, 4, F\}$ features by $1-2\lambda$ and boosted the upload by $\gamma \in \{1, K/|\mathcal{M}|\}$ [3], for 108 configurations per federation at $\alpha = 0.5$. A grid oracle, which knows the detection rate of each configuration, picks per cell the one with the largest mean damage after exclusion (Online Resource 1, Fig. S5 and Table S9). Under median aggregation its pick had attacker recall 0.00 in 16 of the 18 cells. Its targeted losses on ToN-IoT, 4.07 and 9.13 points of $c^\star$ recall at $\rho = 0.2$ and 0.3, went unaccused. At $\rho \leq 0.2$ its accuracy damage was at most 0.58 points without the audit and 0.55 after exclusion, and in 10 of these 12 cells exclusion left the damage unchanged. Under mean aggregation boosted full sign-flips removed 58–94 accuracy points, and the boosted targeted edit was never accused, so the method relies on the median.

ALIE, Min-Max and Min-Sum, which craft uploads from the benign statistics, were never accused (Online Resource 1, Table S10). They lowered worst-class recall by up to 5.33 points at $\rho \leq 0.2$ and 24.15 points at $\rho = 0.4$. The trimmed mean attack was accused at recall 0.40–0.97 at $\rho \leq 0.3$. At $\rho = 0.4$ it was not accused on CICIoT2023 and ToN-IoT and removed 6.47 and 11.52 accuracy points there. A per-class score raised label-flip recall but falsely accused in 4 of 90 clean federations (Online Resource 1, Table S11). No evaluated audit detected the targeted evasion-window attack, which lowered target-class recall inside its window by at most 0.9 points under the median (Online Resource 1, Section S6).

7.5 Accuracy and cost

Against the better of the two FedAvg multilayer perceptrons, the additive model was 1.2, 1.5 and 3.0 accuracy points lower on MQTTset, 5G-NIDD and ToN-IoT and level on CICIoT2023. It needed 14–274 times less communication (Online Resource 1, Table S23). The federation was 7.7–19.2 accuracy points more accurate than the mean participant, but the best participant, chosen after the fact, exceeded it on every dataset (Online Resource 1, Table S24). A server training alone on 2,000–6,000 rows of an idealized IID pool, drawn from the participants' own training data, beat the federation by 1.8–4.0 accuracy points. With a pool of 500 rows, which gave no false accusation

Table 7 What median aggregation alone and median aggregation with the audit deliver per attacker class ($K = 12$). Members named: attacker recall; damage in percentage points (pp) against the clean federation.

Attacker class	Members named	Damage, median alone	Excluding the accused
Uninformed: negation or reversal of the most influential curve	SF 0.90–1.00, PM 0.35–1.00 at $\rho = 0.2$	≥ 1 pp accuracy only at $\rho = 0.4$ on CICIoT2023 (2.40 and 1.11)	recovered 0.91 and 0.28 of it
Uninformed: same edits of other features	0.00–0.08 (one feature), 0.10–0.58 (four)	not measured	–
Uninformed: label flipping	0.00–0.65 at $\rho = 0.2$	at most 1.1 accuracy	recovered 0.06 of 1.06 (ToN-IoT, $\rho = 0.4$)
Uninformed: fixed strength along the colluders' direction	0.92–1.00 at $\lambda = 20$; 0.04–0.19 at $\lambda = 1$	strongest edit ≤ 0.82 at $\rho \leq 0.2$; up to 18.40 accuracy at $\rho = 0.4$	1.80 and 6.90 accuracy at $\rho = 0.3$ and 0.4 for $\lambda = 1$
Informed colluders, one strength tuned to a surrogate audit	cautious 0.00–0.03, aggressive 0.43–0.69	as in the row above	≤ 0.19 at $\rho \leq 0.2$; 1.80 and 6.90 accuracy at $\rho = 0.3$ and 0.4
Blend-and-boost family, grid oracle	0.00 in 16 of 18 cells	≤ 0.58 accuracy and ≤ 4.07 recall of $c^\star$ at $\rho \leq 0.2$; 9.13 recall of $c^\star$ at $\rho = 0.3$	unchanged in 10 of 12 cells at $\rho \leq 0.2$
ALIE, Min-Max, Min-Sum	0.00	≤ 0.10 accuracy and ≤ 5.33 worst-class recall at $\rho \leq 0.2$; up to 24.15 worst-class recall at $\rho = 0.4$	unchanged (none accused)
Trimmed mean attack	0.40–0.97 at $\rho \leq 0.3$; 0.00–0.23 at $\rho = 0.4$	up to 14.78 worst-class recall at $\rho \leq 0.2$; 6.47 and 11.52 accuracy and 59.43 worst-class recall at $\rho = 0.4$	14.78 to 1.77 (ToN-IoT, $\rho = 0.2$); none on CICIoT2023 and ToN-IoT at $\rho = 0.4$

Note: Clean federations with an accused benign participant at $\alpha = 0.5$: 0 of 40 on 5G-NIDD and on ToN-IoT, 1 of 40 on CICIoT2023, the dataset held out at the data level.

in 15 federations, it ranged from 0.62 points below to 1.63 points above the federation in accuracy. Its macro-F1 was higher on all four datasets. With the pool from the other 5G-NIDD station the federation was 33.43 accuracy points [31.94, 34.81] ahead on the participants' traffic. Its macro-F1 lead of 0.95 points had an interval of [−0.21, 2.17]. With pools of 10 and 25 rows per class it also led in accuracy, with macro-F1 differences whose intervals include zero (Online Resource 1, Table S22). The evaluation thus does not show an accuracy benefit of federating over a server that holds an IID sample of the participants' data.

Running a complete 48-model simulation audit required 5.4–16.3 CPU-minutes on a single thread, with a mean of 7.8 CPU-minutes per one-shot federation (Online Resource 1, Table S7). All timings come from one Intel Core i7-13620H processor under Windows 11, with 6 performance and 4 efficiency cores, 16 threads and a 2.4 GHz performance-core base clock. Finally, in the novelty-triage experiments, the sole benign holder of a rare attack class was accused in 3 of 120 federations. It was identified as the top novelty candidate in all 120 runs (Online Resource 1, Table S6).

8 Discussion and limitations

8.1 Conditions supporting calibration

Within the tested conditions the first question has a positive but conditional answer, and the rates describe the sampled federations, not a guarantee. The per-dataset bounds at $\alpha = 0.5$ lie below q on 5G-NIDD and ToN-IoT, whose confirmation federations reuse the exploration samples. On CICIoT2023, the one dataset held out at the data level, the bound lies above q. The observed margin comes from two built-in mismatches, virtual participants more dispersed than the members and an in-sample score that lowers benign scores (Section 7.1). The server cannot verify from the declarations whether its sampler is more dispersed than the members. Members declare their classes and total counts, not their class proportions or feature distributions.

Under feature shift and quantity skew, benign uploads depart from the consensus in ways that virtual participants fitted on the unrescaled pool at their own sizes do not reproduce. Rescaling by at most 25% already produced false accusations on ToN-IoT, and per-feature rescaling is a weak and probably optimistic proxy for real exporter differences. Other active and idle timeouts, packet sampling or counter definitions can change which flows exist and how their features covary, which no per-feature factor reproduces. Such differences were not tested, so the common exporter remains a governance requirement. Falsified declarations and a declaration-free sampler each pushed the same borderline benign participant over the threshold, so p-values near the cut-off should be treated as uncertain.

8.2 Accountability by attacker tier

In our evaluation accountability holds for uninformed members whose edit reverses an influential response curve with large strength, an edit that the median alone would suppress without naming its sender. Weak edits of the same kind largely went unnamed, so accountability against the uninformed tier depends on the size of the edit. Naming this tier has operational value although the threat model admits stronger attackers, because members compromised through commodity tooling plausibly fall in it. For them a named member with a stated error rate is evidence for exclusion or escalation, and whether naming also deters members was not measured.

The audit offers no accountability against informed insiders, since they can choose the cautious strategy, and against them its value is limited to the cost it imposes. Colluders that tune one edit strength along a histogram-weighted direction to a surrogate audit gave up 4.1–13.5 points of damage at $\rho \geq 0.3$ to avoid accusation. They still removed up to 6.9 accuracy points. Attacks outside this family left larger unaccused losses at $\rho \geq 0.3$. The grid oracle removed 9.13 points of target-class recall at $\rho = 0.3$, and Min-Max and Min-Sum up to 24.15 points of worst-class recall at $\rho = 0.4$. The trimmed mean attack, unaccused at $\rho = 0.4$ on two datasets, removed 59.43 points of worst-class recall there. The cost measured for the tested family is therefore no bound on what an informed attacker can achieve. Up to $\rho = 0.2$ the median alone kept the damage of the tuned colluders below one point. There the audit adds accountability for attackers that do not adapt and no robustness.

On the second question, the named feature is the curve that departs most from the consensus, which for data-level attacks means no feature edit. Otherwise it carries evidence mainly when the edit avoids the most influential curve.

8.3 Evidence, disputes and disclosure

In one confirmation federation (ToN-IoT, $\alpha = 0.5$, seed 14) a participant negated its curve of source IP-layer bytes, and both attributions named that feature. Its cosine score was 0.922, against at most 0.293 for the eleven benign participants. Its p-value of 6.2×10^{-6} was below the BH cut-off of 0.0083, whereas no benign p-value fell below 0.185. The evidence package holds these values, the declared metadata and the participant's curve for the named feature against the consensus and the other uploads (Fig. 3). It also gives the share of the largest positive standardized deviation. This share marked half of the label-flip accusations on 5G-NIDD but none on ToN-IoT, so it does not identify label flipping.

A dispute needs an arbiter trusted with server-private material, such as an auditor appointed by the members. Replay needs the server to commit to its pool, seeds and code before the uploads arrive. EBM fits were deterministic given data and seed in our runs, independently of the thread count, so an arbiter can refit the virtual participants and recompute every p-value. The procedure assumes an arbiter that keeps the pool and uploads confidential and colludes with no party, and it was not evaluated with real disputes or contestability beyond replay. Differentially private EBMs [47] would limit what each curve discloses but were not evaluated.

8.4 Deployment and limitations

A consortium that adopts the audit should aggregate by the coordinate-wise median, which bounds boosted uploads that the scale-invariant score cannot detect. The level q is a policy choice, and $q = 0.1$ suits accusations that open a review rather than trigger exclusion. Each contribution reveals labeled traces of the attacks a member suffered, so the sharing agreement must state who may see the pool and for how long. A server training alone matched or exceeded the federation's macro-F1 with every pool on which the audit stayed calibrated, except a class-shifted pool on ToN-IoT. There the federation led on accuracy and macro-F1, and the audit made no false accusation in five federations with sign-flip recall 1.00. Five federations without paired intervals cannot establish a regime where federating improves accuracy. No tested regime established one, so we claim no accuracy benefit of federating.

The confirmation covers $K = 12$ participants under Dirichlet label skew with $\alpha \in \{0.5, 1, 100\}$, on new partitions of the exploration samples for 5G-NIDD and ToN-IoT. Skew stronger than the sampler's, feature shift and quantity skew produced false accusations, as did falsified declarations, and real exporters, sites and periods were not tested. The server pool came from the participants' capture except in one test with a second base station of the same testbed. Power was low on MQTTset, against label flipping and in larger federations, as well as for weak edits and edits of weak features. A server pool drawn from the participants' own training data, and therefore IID, trained a more accurate detector than the federation on every dataset. Cautious

informed colluders, ALIE, Min-Max and Min-Sum evade the audit, as do class-targeted and boosted targeted attacks, so the method relies on median aggregation. Part of the low targeted damage comes from the server's intercept refit, and the attacks were not combined with altered pools or falsified declarations.

9 Conclusion

With glass-box uploads, a server audit accused no benign participant in 40 clean federations each of 5G-NIDD and ToN-IoT at $\alpha = 0.5$, an upper bound of 0.088 per dataset. CICIoT2023, the only dataset held out at the data level, had one false accusation in 40 (bound 0.132). The cluster-level bound on the average over the three samples is 0.046. These rates rest on in-sample scoring and on virtual participants more dispersed than the members, which the server cannot verify from declarations. A matched sampler falsely accused in 10 of 30 federations. Each accusation names the feature that departs most from the consensus, which for a data-level attack means no feature edit. For edits of four random features it was an edited one in 149 of 150 accusations, against a benign base rate of 0.10–0.39. Edits of the most influential curve were named largely by default. Large curve edits made without regard to the audit were named at recall 0.92–1.00, but weak ones only at 0.04–0.19. Weak edits left 1.8 and 6.9 accuracy points of damage at malicious fractions 0.3 and 0.4. Colluders that tune one edit strength along a histogram-weighted direction to a surrogate audit gave up 4.1–13.5 points of damage there, and the cautious ones were named at recall 0.00–0.03. Unaccused grid-oracle, Min-Max and trimmed mean attacks removed up to 9.13, 24.15 and 59.43 points of class recall at these fractions. Up to 0.2 the median kept the tuned colluders' damage below one point, so the audit adds accountability for attackers that do not adapt, and no robustness. Calibration also presumes a common flow exporter and truthful declarations. A pool large enough for the audit also served a server training alone, so the audit is a means of accountability for consortia that federate for coverage and data residency. Open problems are a null matched to these forms of heterogeneity, a sampler that needs no declarations and scores with more power against informed, targeted and statistics-aware attacks.

Supplementary information. Online Resource 1 contains the supplementary tables and figures cited in the text.

Statements and Declarations

Funding. Withheld for double-blind review; see the title page.

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

Ethics approval and consent to participate. Not applicable.

Data and code availability. The four benchmark datasets are publicly available at their original repositories [42, 43, 54, 63]. The code and analysis protocol are provided

as anonymized supplementary material and, for review, at https://zenodo.org/records/24060744?token=eyJhbGciOiJIUzUxMiJ9.eyJpZCI6IjRiZmY0ODkyLWM0YjEtNDAwMi05ZTFhLTU2ODRjMzVjYTc4ZCIsImRhdGEiOnt9LCJyYW5kb20iOiIxOWUxY2YyMjU3ZDgwNjFmMGJiOTlhNGM5YWY5ZGNmMiJ9.bj_dyO1LNTX2qsRFXLI2Plj6yVKi4wo8mS2164loECcUo8Jjj8E7JXzjbrUr44E-3FxUfIBqbb9Lc39NOCphHQ, and will be made public upon acceptance.

Author contributions. Provided on the separate title page.

Use of generative AI. During the preparation of this work, the authors used generative AI tools to assist with language editing and clarity improvement. The authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] Agarwal R, Melnick L, Frosst N, et al (2021) Neural additive models: interpretable machine learning with neural nets. In: Proceedings of the 35th International Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook, NY, USA, NIPS '21

[2] Al Mallah R, López D, Badu-Marfo G, et al (2023) Untargeted poisoning attack detection in federated learning via behavior attestation. IEEE Access 11:125064–125079. https://doi.org/10.1109/ACCESS.2023.3330144

[3] Bagdasaryan E, Veit A, Hua Y, et al (2020) How To Backdoor Federated Learning. In: Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS), pp 2938–2948

[4] Baniecki H, Biecek P (2024) Adversarial attacks and defenses in explainable artificial intelligence: A survey. Information Fusion 107:102303. https://doi.org/10.1016/j.inffus.2024.102303

[5] Baniecki H, Kretowicz W, Biecek P (2023) Fooling Partial Dependence via Data Poisoning. In: Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2022), Lecture Notes in Computer Science, pp 121–136, https://doi.org/10.1007/978-3-031-26409-2_8

[6] Baruch M, Baruch G, Goldberg Y (2019) A little is enough: Circumventing defenses for distributed learning. In: Advances in Neural Information Processing Systems 32 (NeurIPS 2019)

[7] Bates S, Candès E, Lei L, et al (2023) Testing for outliers with conformal p-values. The Annals of Statistics 51(1):149–178. https://doi.org/10.1214/22-AOS2244

[8] Belarbi O, Spyridopoulos T, Anthi E, et al (2023) Federated deep learning for intrusion detection in IoT networks. In: GLOBECOM 2023 - 2023 IEEE Global

Communications Conference, pp 237–242, https://doi.org/10.1109/GLOBECOM54140.2023.10437860

[9] Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Methodological) 57(1):289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

[10] Bhagoji AN, Chakraborty S, Mittal P, et al (2019) Analyzing Federated Learning through an Adversarial Lens. In: Proceedings of the 36th International Conference on Machine Learning (ICML), pp 634–643, URL https://proceedings.mlr.press/v97/bhagoji19a.html, arXiv:1811.12470

[11] Blanchard P, El Mhamdi EM, Guerraoui R, et al (2017) Machine learning with adversaries: Byzantine tolerant gradient descent. In: Guyon I, Luxburg UV, Bengio S, et al (eds) Advances in Neural Information Processing Systems, vol 30. Curran Associates, Inc., URL https://proceedings.neurips.cc/paper_files/paper/2017/file/f4b9ec30ad9f68f89b29639786cb62ef-Paper.pdf

[12] Campello RJGB, Moulavi D, Sander J (2013) Density-Based Clustering Based on Hierarchical Density Estimates. In: Advances in Knowledge Discovery and Data Mining (PAKDD 2013), Lecture Notes in Computer Science, vol 7819. Springer, pp 160–172, https://doi.org/10.1007/978-3-642-37456-2_14

[13] Cao X, Fang M, Liu J, et al (2021) FLTrust: Byzantine-robust Federated Learning via Trust Bootstrapping. In: Network and Distributed System Security Symposium (NDSS), https://doi.org/10.14722/ndss.2021.24434

[14] Chang CH, Caruana R, Goldenberg A (2022) NODE-GAM: Neural generalized additive model for interpretable deep learning. In: International Conference on Learning Representations (ICLR), URL https://arxiv.org/abs/2106.01613, arXiv:2106.01613

[15] Claise B, Trammell B, Aitken P (2013) Specification of the IP flow information export (IPFIX) protocol for the exchange of flow information. Request for Comments 7011, Internet Engineering Task Force, https://doi.org/10.17487/RFC7011, STD 77; B. Claise and B. Trammell (eds.)

[16] Clopper CJ, Pearson ES (1934) The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika 26(4):404–413. https://doi.org/10.1093/biomet/26.4.404

[17] Dou Z, Wang J, Sun W, et al (2025) Toward malicious clients detection in federated learning. In: Proceedings of the 20th ACM Asia Conference on Computer and Communications Security. ACM, ASIA CCS '25, pp 456–472, https://doi.org/10.1145/3708821.3736194

[18] Efron B (2004) Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. Journal of the American Statistical Association 99(465):96–104. https://doi.org/10.1198/016214504000000089

[19] Fang M, Cao X, Jia J, et al (2020) Local model poisoning attacks to Byzantine-Robust federated learning. In: 29th USENIX Security Symposium (USENIX Security 20). USENIX Association, pp 1605–1622, URL https://www.usenix.org/conference/usenixsecurity20/presentation/fang

[20] Fereidooni H, Pegoraro A, Rieger P, et al (2024) FreqFed: A Frequency Analysis-Based Approach for Mitigating Poisoning Attacks in Federated Learning. In: Network and Distributed System Security Symposium (NDSS), https://doi.org/10.14722/ndss.2024.24620

[21] Fung C, Yoon CJM, Beschastnikh I (2020) The limitations of federated learning in sybil settings. In: Proc. 23rd Int. Symp. Research in Attacks, Intrusions and Defenses (RAID), pp 301–316

[22] Gong Z, Shen L, Zhang Y, et al (2024) AgrAmplifier: Defending federated learning against poisoning attacks through local update amplification. IEEE Transactions on Information Forensics and Security 19:1241–1250. https://doi.org/10.1109/TIFS.2023.3333555, URL https://doi.org/10.1109/TIFS.2023.3333555

[23] Hernandez-Ramos JL, Karopoulos G, Chatzoglou E, et al (2025) Intrusion Detection Based on Federated Learning: A Systematic Review. ACM Computing Surveys 57(12):1–65. https://doi.org/10.1145/3731596

[24] Jia Y, Fang M, Liu H, et al (2025) Tracing back the malicious clients in poisoning attacks to federated learning. In: Belgrave D, Zhang C, Lin H, et al (eds) Advances in Neural Information Processing Systems, vol 38. Curran Associates, Inc., https://doi.org/10.52202/085713-3882

[25] Jin D, Chen S, He H, et al (2023) Federated Incremental Learning based Evolvable Intrusion Detection System for Zero-Day Attacks. IEEE Network 37(1):125–132. https://doi.org/10.1109/mnet.018.2200349

[26] Kabir E, Song Z, Ur Rashid MR, et al (2024) FLShield: A Validation Based Federated Learning Framework to Defend Against Poisoning Attacks. In: 2024 IEEE Symposium on Security and Privacy (SP), pp 2572–2590, https://doi.org/10.1109/sp54263.2024.00141

[27] Kairouz P, McMahan HB, Avent B, et al (2021) Advances and open problems in federated learning. Found Trends Mach Learn 14(1–2):1–210. https://doi.org/10.1561/2200000083, URL https://doi.org/10.1561/2200000083

[28] Kalakoti R, Nõmm S, Bahsi H (2025) Federated learning of explainable AI(FedXAI) for deep learning-based intrusion detection in IoT networks. Computer Networks 270:111479. https://doi.org/10.1016/j.comnet.2025.111479, URL https://www.sciencedirect.com/science/article/pii/S1389128625004463

[29] Karimireddy SP, He L, Jaggi M (2022) Byzantine-robust learning on heterogeneous datasets via bucketing. In: International Conference on Learning Representations (ICLR), URL https://arxiv.org/abs/2006.09365, arXiv:2006.09365

[30] Ke G, Meng Q, Finley T, et al (2017) LightGBM: A highly efficient gradient boosting decision tree. In: Guyon I, Luxburg UV, Bengio S, et al (eds) Advances in Neural Information Processing Systems, vol 30. Curran Associates, Inc., URL https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

[31] Krauß T, Dmitrienko A (2023) MESAS: Poisoning Defense for Federated Learning Resilient against Adaptive Attackers. In: Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security, pp 1526–1540, https://doi.org/10.1145/3576915.3623212

[32] Lamprou V, Doukas G, Ntanos C, et al (2025) On the trustworthiness of federated learning models for 5G network intrusion detection under heterogeneous data. Computer Networks 271:111598. https://doi.org/10.1016/j.comnet.2025.111598

[33] Latif N, Ma W, Ahmad HB (2026) FedDBC: A density-based defense against collusion attacks in federated intrusion detection for IoT networks. Computer Networks 286:112497. https://doi.org/10.1016/j.comnet.2026.112497

[34] Li B, Wu Y, Song J, et al (2021) DeepFed: Federated deep learning for intrusion detection in industrial cyber–physical systems. IEEE Transactions on Industrial Informatics 17(8):5615–5624. https://doi.org/10.1109/TII.2020.3023430

[35] Li S, Cheng Y, Wang W, et al (2020) Learning to detect malicious clients for robust federated learning. https://doi.org/10.48550/arXiv.2002.00211, arXiv:2002.00211

[36] Lou Y, Caruana R, Gehrke J, et al (2013) Accurate intelligible models with pairwise interactions. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery, New York, NY, USA, KDD '13, p 623–631, https://doi.org/10.1145/2487575.2487579, URL https://doi.org/10.1145/2487575.2487579

[37] Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. In: Guyon I, Luxburg UV, Bengio S, et al (eds) Advances in Neural Information Processing Systems, vol 30. Curran Associates, Inc., URL https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

[38] Lundberg SM, Erion G, Chen H, et al (2020) From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence 2(1):56–67. https://doi.org/10.1038/s42256-019-0138-9, URL https://doi.org/10.1038/s42256-019-0138-9

[39] Manna A, Kasyap H, Tripathy S (2021) Moat: Model Agnostic Defense against Targeted Poisoning Attacks in Federated Learning. In: Information and Communications Security (ICICS 2021), Lecture Notes in Computer Science, vol 12918. Springer, pp 38–55, https://doi.org/10.1007/978-3-030-86890-1_3

[40] McInnes L, Healy J, Astels S (2017) hdbscan: Hierarchical density based clustering. Journal of Open Source Software 2(11):205. https://doi.org/10.21105/joss.00205

[41] McMahan B, Moore E, Ramage D, et al (2017) Communication-Efficient Learning of Deep Networks from Decentralized Data. In: Singh A, Zhu J (eds) Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, Proceedings of Machine Learning Research, vol 54. PMLR, pp 1273–1282, URL https://proceedings.mlr.press/v54/mcmahan17a.html

[42] Moustafa N (2021) A new distributed architecture for evaluating ai-based security systems at the edge: Network TON_IoT datasets. Sustainable Cities and Society 72:102994. https://doi.org/10.1016/j.scs.2021.102994, URL https://www.sciencedirect.com/science/article/pii/S2210670721002808

[43] Neto ECP, Dadkhah S, Ferreira R, et al (2023) CICIoT2023: A real-time dataset and benchmark for large-scale attacks in IoT environment. Sensors 23(13):5941. https://doi.org/10.3390/s23135941

[44] Nguyen TD, Marchal S, Miettinen M, et al (2019) DÏoT: A federated self-learning anomaly detection system for IoT. In: 2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS), pp 756–767, https://doi.org/10.1109/ICDCS.2019.00080

[45] Nguyen TD, Rieger P, Chen H, et al (2022) FLAME: Taming backdoors in federated learning. In: 31st USENIX Security Symposium (USENIX Security 22), Boston, MA, pp 1415–1432, URL https://www.usenix.org/conference/usenixsecurity22/presentation/nguyen, arXiv:2101.02281

[46] Nori H, Jenkins S, Koch P, et al (2019) InterpretML: A unified framework for machine learning interpretability. arXiv preprint arXiv:190909223 URL https://arxiv.org/abs/1909.09223

[47] Nori H, Caruana R, Bu Z, et al (2021) Accuracy, interpretability, and differential privacy via explainable boosting. In: Meila M, Zhang T (eds) Proceedings of the 38th International Conference on Machine Learning, Proceedings of Machine Learning Research, vol 139. PMLR, pp 8227–8237, URL https://proceedings.ml

r.press/v139/nori21a.html

[48] Pacheco EJ, Pedroso CM (2026) Operational lifecycle of interpretable multi-plane intrusion detection in 5G networks. Computer Networks 289:112672. https://doi.org/10.1016/j.comnet.2026.112672

[49] Pinto D, Vitorino J, Maia E, et al (2025) Flow Exporter Impact on Intelligent Intrusion Detection Systems. In: Proceedings of the 11th International Conference on Information Systems Security and Privacy (ICISSP), pp 289–298, https://doi.org/10.5220/0013131900003899

[50] Qian C, Wang M, Ren H, et al (2024) ByMI: Byzantine machine identification with false discovery rate control. In: Proceedings of the 41st International Conference on Machine Learning, Proceedings of Machine Learning Research, vol 235. PMLR, pp 41357–41382

[51] Rieger P, Nguyen TD, Miettinen M, et al (2022) DeepSight: Mitigating Backdoor Attacks in Federated Learning Through Deep Model Inspection. In: Network and Distributed System Security Symposium (NDSS), https://doi.org/10.14722/ndss.2022.23156

[52] Rieger P, Krauß T, Miettinen M, et al (2024) CrowdGuard: Federated Backdoor Detection in Federated Learning. In: Network and Distributed System Security Symposium (NDSS), https://doi.org/10.14722/ndss.2024.23233

[53] Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence 1(5):206–215. https://doi.org/10.1038/s42256-019-0048-x

[54] Samarakoon S, Siriwardhana Y, Porambage P, et al (2022) 5G-NIDD: A comprehensive network intrusion detection dataset generated over 5G wireless network. https://doi.org/10.21227/xtep-hv36

[55] Sameera KM, Vinod P, Rocha A, et al (2025) WeiDetect: Weibull distribution-based defense against poisoning attacks in federated learning for network intrusion detection systems. https://doi.org/10.48550/arXiv.2504.04367, arXiv:2504.04367

[56] Sandeepa C, Siniarski B, Wang S, et al (2024) SHERPA: Explainable robust algorithms for privacy-preserved federated learning in future networks to defend against data poisoning attacks. In: 2024 IEEE Symposium on Security and Privacy (SP), pp 4772–4790, https://doi.org/10.1109/SP54263.2024.00271

[57] Sankeawthong S, Cao X, Zhao S, et al (2026) TrustFed-IDS: A trust-aware federated hybrid MLP-LSTM framework for robust intrusion detection in wireless sensor networks. Computer Networks 286:112430. https://doi.org/10.1016/j.comnet.2026.112430

[58] Shejwalkar V, Houmansadr A (2021) Manipulating the Byzantine: Optimizing model poisoning attacks and defenses for federated learning. In: Proceedings of the Network and Distributed Systems Security (NDSS) Symposium 2021, Virtual, https://doi.org/10.14722/ndss.2021.24498, URL https://dx.doi.org/10.14722/ndss.2021.24498

[59] Siriwardhana Y, Samarakoon S, Porambage P, et al (2025) Descriptor: 5G wireless network intrusion detection dataset (5G-NIDD). IEEE Data Descriptions 2:358–369. https://doi.org/10.1109/IEEEDATA.2025.3592888

[60] Slack D, Hilgard S, Jia E, et al (2020) Fooling LIME and SHAP. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, pp 180–186, https://doi.org/10.1145/3375627.3375830

[61] Thein TT, Shiraishi Y, Morii M (2024) Personalized federated learning-based intrusion detection system: Poisoning attack and defense. Future Generation Computer Systems 153:182–192. https://doi.org/10.1016/j.future.2023.10.005

[62] Tolpegin V, Truex S, Gursoy ME, et al (2020) Data poisoning attacks against federated learning systems. In: Computer Security – ESORICS 2020. Springer International Publishing, Lecture Notes in Computer Science, pp 480–501, https://doi.org/10.1007/978-3-030-58951-6_24, URL https://doi.org/10.1007/978-3-030-58951-6_24

[63] Vaccari I, Chiola G, Aiello M, et al (2020) MQTTset, a new dataset for machine learning techniques on MQTT. Sensors 20(22). https://doi.org/10.3390/s20226578, URL https://www.mdpi.com/1424-8220/20/22/6578

[64] Xu H, Shu T (2024) Attack-model-agnostic defense against model poisonings in distributed learning. Journal of Information Security and Applications 82:103739. https://doi.org/10.1016/j.jisa.2024.103739, URL https://www.sciencedirect.com/science/article/pii/S2214212624000425

[65] Yang R, He H, Wang Y, et al (2023) Dependable federated learning for IoT intrusion detection against poisoning attacks. Computers & Security 132:103381. https://doi.org/10.1016/j.cose.2023.103381

[66] Yang YR, Shi CW, Li WJ (2025) On the tension between Byzantine robustness and no-attack accuracy in distributed learning. In: Proceedings of the 42nd International Conference on Machine Learning, Proceedings of Machine Learning Research, vol 267. PMLR, pp 71051–71072

[67] Yin D, Chen Y, Kannan R, et al (2018) Byzantine-robust distributed learning: Towards optimal statistical rates. In: Dy J, Krause A (eds) Proceedings of the 35th International Conference on Machine Learning, Proceedings of Machine Learning Research, vol 80. PMLR, pp 5650–5659, URL https://proceedings.mlr.press/v80/yin18a.html

[68] Yurochkin M, Agarwal M, Ghosh S, et al (2019) Bayesian Nonparametric Federated Learning of Neural Networks. In: Proceedings of the 36th International Conference on Machine Learning (ICML), pp 7252–7261

[69] Zhang X, Liu Q, Ba Z, et al (2024) FLTracer: Accurate poisoning attack provenance in federated learning. IEEE Transactions on Information Forensics and Security 19:9534–9549. https://doi.org/10.1109/TIFS.2024.3410014

[70] Zhang Z, Cao X, Jia J, et al (2022) FLDetector: Defending federated learning against model poisoning attacks via detecting malicious clients. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery, New York, NY, USA, KDD '22, p 2545–2555, https://doi.org/10.1145/3534678.3539231, URL https://doi.org/10.1145/3534678.3539231

[71] Zhang Z, Zhang Y, Guo D, et al (2022) SecFedNIDS: Robust defense for poisoning attack against federated learning-based network intrusion detection system. Future Generation Computer Systems 134:154–169. https://doi.org/10.1016/j.future.2022.04.010

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementary.pdf