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.

A Severity-Calibrated Adversarial Benchmark for Federated Learning: From Label Corruption to Structured Model-Update Injection

Md Shahanur Islam Shagor

smshagor.ru@gmail.com

Voronezh State University of Forestry and Technologies https://orcid.org/0009-0003-7730-3202

Research Article

Keywords: federated learning, poisoning attacks, adversarial machine learning, Byzantine robustness, security benchmarking, model integrity

Posted Date: September 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10976314/v1

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

Additional Declarations: The authors declare no competing interests.

Abstract

Federated learning reduces the need to centralize raw training data, but it does not by itself establish the integrity of client-side computation. Security studies consequently report data poisoning, model poisoning, Byzantine manipulation, and backdoor behavior under widely different threat models, attacker fractions, and perturbation magnitudes, making cross-attack comparison difficult. This study presents a severity-calibrated adversarial benchmark that places five manipulation families in one controlled federated-learning protocol: probabilistic label flipping, additive Gaussian update noise, sign-flip poisoning, norm-bounded adaptive poisoning, and a benchmark-specific structured low-rank matrix injection stressor. The main experiment comprises 640 runs over two classification datasets, five random seeds, four server aggregation rules, three attack severities, and a 20% malicious-client population; an additional 240 runs vary the malicious population from 6.7% to 40%. Across 17,600 federated rounds, the results show that attack magnitude alone is a poor indicator of operational damage. At high severity, arithmetic-mean aggregation lost 87.39 percentage points of final accuracy under sign flipping and 11.47 points under adaptive poisoning, while Gaussian noise produced only a 0.36-point mean drop despite a larger update-distortion ratio than sign flipping. Trimmed mean, coordinate-wise median, and Multi-Krum sharply reduced degradation in the tested 20% regime, although their behavior is not interpreted as universal immunity. The benchmark therefore argues for security evaluation that reports attack geometry, population share, clean-normalized utility loss, and aggregation assumptions together rather than treating a single perturbation norm as a sufficient measure of adversarial severity.

1. Introduction

Federated learning (FL) changes where training data are processed, but it does not remove the security consequences of accepting computations from remote participants. In the canonical setting, a coordinating server distributes a model, clients train locally, and the server aggregates their returned updates [1, 2]. This architecture can reduce raw-data movement, yet the server generally cannot reconstruct how a particular client produced an update. A compromised participant can therefore preserve protocol correctness while corrupting the semantics of the contribution. The resulting integrity problem is different from confidentiality: an update may arrive over an authenticated and encrypted channel and still be malicious.

The literature has established several ways in which this integrity boundary can be abused. Data-poisoning attacks modify local examples or labels before training [3, 4]. Model-poisoning attacks bypass local data manipulation and directly alter the model update sent to the server [5–8]. Byzantine attacks broaden the adversary further by allowing arbitrary client behavior, motivating aggregation rules such as Krum, coordinate-wise median, and trimmed mean [9–11]. More recent work has explored trust bootstrapping, attack-model-agnostic defenses, and backdoor-specific mechanisms [12–16]. These results make clear that poisoning is not a single attack class. Different attacks create different directions, magnitudes, coordination patterns, and target effects in parameter space.

A persistent experimental problem follows from that diversity. Attack A may be reported with 10% compromised clients, Attack B with 30%, and Attack C with a perturbation scaled to a different quantity altogether. Some studies report final accuracy; others emphasize attack success rate, convergence failure, or distance between updates. An attack described as 'strong' in one paper may therefore be incomparable with a 'strong' attack in another. This matters for cybersecurity evaluation because defenders are often selected using aggregate claims such as 'robust against poisoning' even though the tested attacks may occupy only a narrow portion of the threat surface.

This paper addresses that comparability problem with a controlled benchmark rather than a new aggregation algorithm. Five attack families are evaluated within the same client population, model family, training schedule, clean baseline, severity vocabulary, and set of server aggregators. The benchmark separates data-level label corruption from four update-level manipulations: random Gaussian noise, coherent sign reversal, norm-bounded opposite-direction poisoning, and a structured low-rank matrix injection designed as an additional stressor. Low, medium, and high severity are calibrated within each family and are intentionally not presented as physically equivalent units. The point is to compare security outcomes under a common protocol while preserving the different geometry of each attack.

The study makes four connected contributions. First, it defines a reproducible threat matrix that makes attack type, severity, attacker fraction, and aggregation rule explicit. Second, it evaluates clean-normalized accuracy loss and loss inflation rather than comparing attacked models against unrelated baselines. Third, it adds a structured matrix-injection stressor to test whether a large, coordinated low-rank perturbation is necessarily more damaging than smaller but directionally aligned attacks. Fourth, it quantifies the effect of malicious population size and uses paired analysis to distinguish large robustness gains from small fluctuations. The main finding is security-relevant: perturbation magnitude is not a reliable surrogate for attack effectiveness. Directional coherence and the aggregation rule can dominate raw update norm.

The work is intended as an adversarial benchmark for information-security analysis, not as evidence that one aggregation rule solves federated-learning security. The tested defenses operate only at aggregation. They do not prevent compromised endpoints, guarantee privacy, detect all backdoors, or remain valid when their Byzantine assumptions are violated. Those boundaries are central to the interpretation of the results.

2. Related work

2.1 Federated learning and poisoning

McMahan et al. established practical model averaging for decentralized learning [1], and later surveys catalogued the optimization, privacy, systems, and security challenges that arise when clients retain local data [2, 3]. From a security perspective, the same decentralization that limits direct data exposure weakens the server's ability to verify local training behavior. Tolpegin et al. showed that label-based data

poisoning can produce targeted degradation even when only a subset of FL participants is malicious [4]. Bagdasaryan et al. demonstrated that directly manipulating model updates can be substantially more powerful than poisoning local training data, including for backdoor insertion [6].

Fang et al. systematically studied local model poisoning against Byzantine-robust federated learning and showed that defenses designed around statistical outliers can still be manipulated [5]. Baruch et al. further demonstrated that comparatively small crafted changes can circumvent distributed-learning defenses [7]. Shejwalkar and Houmansadr formulated optimization-based poisoning against federated defenses and reinforced the need to evaluate attacks adaptively rather than assuming a fixed, naive adversary [8]. These studies motivate a benchmark that keeps the learning protocol fixed while changing the geometry and severity of the manipulation.

2.2 Robust aggregation as a benchmark dimension

Krum selects an update whose neighborhood is most consistent with other submitted updates under a bounded-Byzantine assumption [9]. Coordinate-wise median and trimmed mean offer different robustness mechanisms by suppressing coordinate-level extremes [10]. Bulyan and related work exposed weaknesses that can remain in high-dimensional Byzantine settings even when formal convergence conditions hold [11]. FLTrust uses a server-held root dataset to establish a trust direction and normalize client influence [12], while FoolsGold targets correlated sybil behavior [13].

The present study does not introduce a new defense. Mean, trimmed mean, median, and Multi-Krum are used as controlled benchmark factors. This is important because a security benchmark that tests only arithmetic averaging cannot distinguish between intrinsic attack potency and an aggregation-specific failure mode. Conversely, observing low degradation under a robust aggregator cannot justify a general claim of security because the guarantee may depend on the assumed number of Byzantine clients, independence of honest updates, and attacker knowledge.

2.3 Recent information-security studies and the benchmarking gap

Recent information-security literature continues to treat poisoning and backdoors in FL as active problems. Xu and Shu proposed attack-model-agnostic mechanisms and a variance-minimization defense in the Journal of Information Security and Applications (JISA) [14, 15]. FedESP focused on effective, stealthy, and persistent backdoor behavior [16], while recent privacy-preserving work has combined gradient verification with lightweight perturbation to resist poisoning [17]. These approaches address important defensive questions, but their objectives and attack models differ.

A benchmark serves a complementary role. It should not declare a winner from one attack configuration; it should make the conditions under which a defense appears effective inspectable. The recent systematic review literature likewise identifies benchmark standardization as an open direction in adversarial FL research [18]. Our design responds by pairing every attacked run with a clean run under

the same dataset, seed, and aggregation rule; reporting both attack severity and malicious population share; and distinguishing update magnitude from update direction.

3. Threat model and benchmark design

The benchmark assumes a centralized FL coordinator that is not compromised. Client identities and communication are outside the attack surface considered here: the adversary already controls a bounded subset of legitimate clients and can change those clients' local labels or submitted updates. Honest clients follow the training protocol. Malicious clients know the current global model, their own local data, and the attack configuration, but they do not receive honest clients' private training samples. The main benchmark uses 3 malicious clients among 15 participants (20%) with full participation in every round. A separate sensitivity study varies the malicious population to 1, 3, 5, and 6 clients (6.7%, 20%, 33.3%, and 40%).

The benchmark studies untargeted global utility degradation as the primary outcome. A class-0-to-class-1 confusion diagnostic is recorded, but the experiments are not designed as a persistent backdoor benchmark. This separation is deliberate: backdoor stealth and trigger persistence require additional metrics that are beyond the scope of the present cross-attack comparison.

3.1 Federated update model

Let w(t) denote the global parameter vector at round t and let u_i(t) denote the local update returned by client i. With arithmetic-mean aggregation, the server update is the average of the n submitted vectors. Robust aggregators replace that average with a selection or coordinate-wise robust statistic. The benchmark keeps the local optimization rule fixed so that changes in the global trajectory can be attributed to the injected attack and aggregation rule rather than to different client optimizers.

$$w(t+1) = w(t) + A(u_1(t), u_2(t), \ldots, u_n(t))$$

Here A is one of mean, coordinate-wise trimmed mean, coordinate-wise median, or Multi-Krum. Attack transformations are applied only to malicious clients. For data poisoning, the transformation acts before local-gradient computation; for model poisoning, it acts on the clean local update after computation.

3.2 Label-flipping data poisoning

Label flipping changes the training signal rather than the transmitted update directly. For a C-class task, selected malicious labels are mapped cyclically as $y' = (y + 1) \bmod C$. The low, medium, and high severity levels flip 25%, 50%, and 100% of the malicious client's labels, respectively. This all-class mapping avoids defining one particular target class as the only security objective and produces a broad data-integrity stressor. Because the resulting model update depends on the client's data distribution and current model state, its update distortion is measured empirically rather than fixed in advance.

3.3 Gaussian update injection

For unstructured noise injection, a random Gaussian direction z is normalized and added to the clean local update u. If $\|u\|_2$ is the clean-update norm, the injected component has norm $\rho\|u\|_2$ with rho in {0.5, 3, 8} for low, medium, and high severity. This controls perturbation magnitude while keeping the direction stochastic across malicious clients and rounds.

$u' = u + \rho\|u\|_2 (z / \|z\|_2)$

3.4 Sign-flip model poisoning

Sign flipping reverses the direction of an update and scales it by γ. The transformed update is $u' = -\gamma u$, with γ in {1, 3, 5}. Unlike Gaussian noise, the perturbation is maximally aligned against the client's honest descent direction. This provides a simple but strong test of whether the server is sensitive to coherent adversarial direction rather than merely large norms.

$u' = -\gamma u$

3.5 Norm-bounded adaptive poisoning

The adaptive stressor also moves opposite to the honest update but uses smaller output norms than the strongest sign-flip condition. The malicious vector is $u' = -ru$ with r in {0.5, 1.5, 3}. In operational terms this family represents an attacker that prioritizes a constrained perturbation envelope rather than maximal update magnitude. The label 'adaptive' refers to its dependence on the current honest local update and its norm; it is not an omniscient attack with access to honest peers' data.

3.6 Structured low-rank matrix injection

To include the requested matrix-injection threat without conflating it with the existing poisoning primitives, the benchmark adds a specific structured stressor. At each round, malicious clients share a deterministic round-specific rank-one matrix $M = ab^\top$. The matrix is flattened into the same vector space as the model parameters, normalized, and scaled relative to each malicious client's clean-update norm. The injected component uses norm ratios s in {2, 8, 20}. Sharing the round-specific direction makes the perturbation structured and coordinated, while the rank-one construction prevents it from degenerating into independent Gaussian noise.

$u' = u + s\|u\|_2 \, \mathrm{vec}(M) / \|\mathrm{vec}(M)\|_2$

This matrix injection is a benchmark-specific stressor, not a claim of a new state-of-the-art attack. Its purpose is to test a common intuition: that a much larger structured perturbation should necessarily be more damaging than a smaller one. The results show why that intuition can be misleading.

3.7 Severity semantics and metrics

The labels low, medium, and high are ordinal within an attack family. They are not meant to equate, for example, a 100% label-flip probability with a matrix perturbation of twenty update norms. A cross-family severity scale based on one scalar would erase the geometry that this benchmark is designed to expose.

Accordingly, the primary reported effect is the paired clean-normalized final-accuracy drop, measured in percentage points.

Drop_pp = 100 × (Accuracy_clean − Accuracy_attacked)

Secondary outcomes are final cross-entropy loss, mean accuracy across the 20-round trajectory, loss inflation relative to the paired clean run, gain of a robust aggregator over the attacked mean baseline, and the mean malicious-update distortion $\|u' - u\|_2 / \|u\|_2$. For each high-severity attack/aggregator condition, a 95% bootstrap confidence interval is computed over the ten paired dataset-seed observations using 5,000 resamples. Paired Wilcoxon signed-rank tests compare each robust aggregator with arithmetic mean under the same dataset, seed, attack, and severity. These tests are exploratory and no multiple-testing correction is applied.

Table 1
Attack families and severity calibration.

Attack family	Manipulated object	Low	Medium	High
Label flip	Local labels	p = 0.25	p = 0.50	p = 1.00
Gaussian	Submitted update	0.5× norm	3× norm	8× norm
Sign flip	Submitted update	γ = 1	γ = 3	γ = 5
Adaptive	Submitted update	r = 0.5	r = 1.5	r = 3
Low-rank matrix	Submitted update	+ 2× norm	+ 8× norm	+ 20× norm

4. Experimental setup

4.1 Datasets and model

The evaluation uses two compact, well-characterized classification datasets available through scikit-learn [19]. The handwritten Digits task contains 1,797 samples represented by 64 numerical features and ten classes. The Breast Cancer Wisconsin (Diagnostic) task contains 569 samples with 30 real-valued features and two diagnostic classes [20, 21]. These datasets are not proposed as a complete proxy for production FL. They were selected because they allow a large factorial attack study to be repeated deterministically while preserving multiclass and binary classification settings.

For each dataset, 75% of the samples are used for training and 25% for testing with stratification. Standardization parameters are fit on the training partition only and then applied to the held-out test set. The model is a linear softmax classifier with a bias term. The deliberately low-capacity model is a methodological choice: the benchmark focuses on poisoning geometry and aggregation behavior rather than claiming state-of-the-art predictive performance. A more complex neural network would add representational and optimization confounders that are not necessary to demonstrate the security phenomenon studied here.

4.2 Client partitioning and local training

Each training set is partitioned into 15 disjoint clients using class-wise shuffled round-robin assignment, producing approximately balanced IID-style client partitions. All 15 clients participate in each round. The global model starts from zero parameters. Each client performs one full-batch gradient step per round with learning rate 0.12 and L2 regularization coefficient 10^{-4}. Training runs for 20 federated rounds. Five independent seeds (11, 29, 47, 71, and 101) determine client partition shuffling and malicious-client selection in the main benchmark.

4.3 Aggregation rules

Arithmetic mean is the non-robust reference. Coordinate-wise trimmed mean removes f largest and f smallest values in each parameter coordinate before averaging, where f equals the number of malicious clients. Coordinate-wise median takes the median independently in each coordinate. Multi-Krum computes pairwise squared Euclidean distances, scores each update by its nearest n-f-2 neighbors, and averages up to five lowest-scoring candidates. The benchmark enforces the standard Krum safety condition $n \geq 2f + 3$. Consequently, Multi-Krum is included in the 20% main benchmark, while the separate population-sensitivity experiment is restricted to mean and median so that it does not silently cross a Krum fault-boundary at the higher attacker fractions.

4.4 Run matrix and reproducibility

The main matrix contains 600 adversarial runs: two datasets × five seeds × five attacks × three severities × four aggregators. Forty clean runs provide paired baselines for the same two datasets, five seeds, and four aggregators. The main study therefore contains 640 runs, or 12,800 federated rounds. The population-sensitivity study adds 240 high-severity runs using the first three seeds, five attacks, attacker counts of 1, 3, 5, and 6, and mean/median aggregation. The complete study comprises 880 runs and 17,600 federated rounds.

5. Results

5.1 Clean baselines

Before comparing attacks, it is necessary to establish that the aggregation rules do not begin from materially different clean regimes. On Breast Cancer, mean and trimmed mean both reached 0.9650 final accuracy; median reached 0.9622 and Multi-Krum 0.9608. On Digits, the corresponding means were 0.9031, 0.9031, 0.9053, and 0.8987. Across the five seeds, clean variability was small for mean and trimmed mean and modest for the selection-based rules. All attack effects below are therefore normalized to the clean run using the same dataset, seed, and aggregation rule rather than to a single global baseline.

Table 2
Clean final accuracy across five seeds (mean ± SD).

Dataset	Mean	Trimmed mean	Median	Multi-Krum
Breast Cancer	0.9650 ± 0.0000	0.9650 ± 0.0000	0.9622 ± 0.0106	0.9608 ± 0.0063
Digits	0.9031 ± 0.0012	0.9031 ± 0.0012	0.9053 ± 0.0020	0.8987 ± 0.0081

5.2 Severity response under arithmetic mean

Arithmetic mean exposes the clearest difference between attack families because every submitted update contributes directly to the global step. Figure 2 shows the clean-normalized accuracy drop as severity increases. Sign flipping exhibits the steepest response: the mean drop rises from 1.61 percentage points at low severity to 11.47 points at medium severity and 87.39 points at high severity. The adaptive opposite-direction attack increases from 1.01 to 2.21 and then 11.47 points. Label flipping grows from 0.13 to 0.80 and 2.02 points, while matrix injection grows from 0.14 to 0.26 and 2.53 points.

Gaussian noise behaves differently. The low- and medium-severity means fluctuate by -0.02 and − 0.05 percentage points relative to their paired clean baselines, and even high severity produces only a 0.36-point mean loss. These small negative drops at lower severity are ordinary finite-sample fluctuations, not evidence that the attack improves security or learning. The important observation is the contrast with sign flipping: high Gaussian injection is large in norm but weakly aligned across clients, whereas sign flipping is directionally coherent against local descent.

5.3 Robust aggregation at high severity

At the main 20% malicious fraction, robust aggregation substantially changes the outcome for the strongest attacks. Under arithmetic mean, high-severity sign flipping causes an average 87.39-point accuracy loss (95% bootstrap CI 85.88 to 88.68), while adaptive poisoning causes an 11.47-point loss (5.63 to 17.38). The corresponding mean losses for label flipping and matrix injection are 2.02 and 2.53 points. Gaussian injection remains small at 0.36 points with an interval crossing zero.

Trimmed mean reduces the high-severity average drops to 0.58 points for label flipping, -0.03 for Gaussian injection, 0.77 for sign flipping, 0.81 for adaptive poisoning, and 0.20 for matrix injection. Median gives 0.60, 0.29, 0.74, 0.81, and 0.31 points, respectively. Multi-Krum produces a mean clean-normalized drop of approximately 0.07 points for each high-severity attack in this controlled setting. The near-constant Multi-Krum result should be read cautiously: f is correctly specified, all 15 clients participate, the data partitions are approximately IID, and the attacks are not optimized specifically to evade Krum. It is evidence of low degradation in this experiment, not of general immunity.

Table 3
High-severity final-accuracy drop in percentage points; mean [95% bootstrap CI], n = 10 paired dataset-seed observations.

Attack	Mean	Trimmed mean	Median	Multi-Krum
Label flip	2.02 [1.46, 2.55]	0.58 [0.17, 1.00]	0.60 [-0.02, 1.21]	0.07 [-0.34, 0.52]
Gaussian	0.36 [-0.11, 0.86]	-0.03 [-0.25, 0.20]	0.29 [-0.20, 0.71]	0.07 [-0.34, 0.52]
Sign flip	87.39 [85.88, 88.71]	0.77 [0.29, 1.27]	0.74 [0.33, 1.16]	0.07 [-0.34, 0.52]
Adaptive	11.47 [5.63, 17.38]	0.81 [0.33, 1.30]	0.81 [0.36, 1.23]	0.07 [-0.34, 0.52]
Matrix injection	2.53 [0.73, 4.31]	0.20 [-0.05, 0.44]	0.31 [-0.20, 0.78]	0.07 [-0.34, 0.52]

The paired Wilcoxon analysis supports the large robustness differences for the strongest coherent attacks. Relative to mean aggregation, trimmed mean improves final accuracy by 86.62 percentage points under sign flipping ($p = 0.00195$) and 10.66 points under adaptive poisoning ($p = 0.00391$). Median improves by 86.61 and 10.64 points for the same conditions ($p = 0.00195$ and $p = 0.00391$). Multi-Krum improves by 86.89 and 10.98 points ($p = 0.00195$ and $p = 0.00391$). Smaller effects are less consistent: median's 0.04-point Gaussian gain is not significant ($p = 0.9297$), and its 2.19-point matrix-injection gain does not reach the conventional 0.05 threshold ($p = 0.1055$). Because multiple exploratory tests are reported without multiplicity correction, these p-values are supporting descriptors rather than confirmatory claims.

5.4 Dataset-level outcomes

The catastrophic sign-flip effect under mean aggregation appears on both tasks rather than being driven by one dataset. At high severity, Breast Cancer final accuracy falls to 0.0951 under mean aggregation but remains between 0.9580 and 0.9608 under the three robust rules. On Digits, mean aggregation falls to 0.0253, whereas median, trimmed mean, and Multi-Krum remain between 0.8933 and 0.8973. Adaptive poisoning is less destructive but still materially affects the Digits task under mean aggregation, reducing final accuracy to 0.6960; the robust rules remain near 0.89–0.90.

Label flipping and structured matrix injection produce more moderate global damage at 20% malicious participation. This does not make them benign. Label corruption can be targeted toward particular classes, and the benchmark's global accuracy metric can hide localized harm. Likewise, a structured update may become more effective if aligned with a model's sensitive subspace. The present results only establish their effect under the specified cyclic mapping and shared rank-one direction.

5.5 Malicious-population sensitivity

The sensitivity study asks whether conclusions at 20% malicious participation remain qualitatively stable as the attacker population changes. Results are averaged across all five high-severity attack families, both datasets, and three seeds. Under arithmetic mean, the average clean-normalized accuracy drop rises from 0.99 percentage points with one malicious client (6.7%) to 21.12 points with three clients

(20%), 39.56 points with five clients (33.3%), and 41.80 points with six clients (40%). Median remains much less affected in this experiment: 0.08, 0.46, 0.44, and 1.34 points at the same attacker fractions.

The population curve should not be interpreted as a median breakdown theorem. It averages attacks with very different geometries, uses IID-style partitions, and stops at 40% malicious clients. The experiment instead demonstrates a practical benchmarking principle: attacker share is a first-class security variable. Reporting an attack without the fraction of participating malicious clients omits information that can dominate the observed outcome.

5.6 Why update magnitude is an incomplete severity measure

Figure 4 provides the central cross-attack comparison. At high severity, the mean malicious-update distortion ratios are approximately 4.0 for adaptive poisoning, 8.0 for Gaussian injection, 6.70 for label flipping, 20.0 for matrix injection, and 6.0 for sign flipping. If update norm alone determined attack severity, matrix injection should dominate sign flipping. It does not. Matrix injection causes a 2.53-point mean-aggregation accuracy drop, while sign flipping causes 87.39 points. Gaussian injection is also larger in distortion than sign flipping but causes only 0.36 points of average degradation.

The difference is geometric. Independent random directions partially cancel under averaging. The low-rank matrix stressor is coordinated, but the chosen rank-one direction is not optimized to oppose the true descent direction. Sign flipping, by contrast, reverses each compromised client's local descent vector and therefore places malicious mass in a direction systematically antagonistic to the honest optimization trajectory. The benchmark thus separates two concepts often collapsed into one scalar: how far a malicious update moves from its clean counterpart, and how effectively that movement changes the global optimization direction.

6. Security interpretation

The results support three security conclusions. First, federated-learning integrity cannot be assessed by naming an attack family alone. The same aggregation rule that appears unaffected by high-norm Gaussian noise can fail catastrophically under a smaller but directionally coherent sign reversal. A useful benchmark must therefore report the transformation itself, not just its norm or a qualitative 'strong attack' label.

Second, robust aggregation meaningfully changes the observed attack surface, but only under assumptions that must be disclosed. Trimmed mean and median suppress coordinate-level extremes; Multi-Krum relies on a bounded f and distance-based neighborhood consistency. Their strong performance in the main experiment is informative precisely because f is known, participation is complete, and the partition structure is controlled. Real deployments may face non-IID data, partial participation, adaptive collusion, stale updates, secure aggregation, or inaccurate estimates of the

malicious population. Those factors can make benign updates look anomalous or malicious updates look statistically ordinary.

Third, benchmark design should distinguish defensive robustness from attack stealth. A server can preserve overall test accuracy while still harboring a targeted backdoor, class-specific bias, or privacy leakage. This paper measures global integrity degradation and a limited class-confusion diagnostic. It does not claim to evaluate trigger persistence, inference attacks, or covert model replacement. Security claims should remain attached to the measured property.

For practitioners, the immediate implication is to avoid using one random-noise stress test as evidence that an FL deployment is robust to malicious clients. Random perturbation is useful for checking numerical stability, but it may substantially underestimate adversarial risk when malicious clients can align updates against optimization. At minimum, deployment testing should include a data-corruption case, a coherent directional update attack, an unstructured noise case, and an attacker-population sweep. Robust aggregation should be evaluated against its stated Byzantine assumptions rather than enabled as an undocumented default.

7. Limitations and threats to validity

The benchmark intentionally favors control and comparability over task scale. It uses two tabular classification datasets and a linear softmax model. The findings therefore establish a security mechanism in a controlled setting rather than predict the exact degradation of large vision, language, or foundation models. Higher-dimensional non-convex models may create additional attack surfaces and can alter the relative behavior of distance-based aggregators.

Client partitions are approximately IID and balanced, all clients participate in every round, and the main malicious fraction is fixed at 20%. Production FL frequently combines non-IID data with partial availability and asynchronous execution. Under such conditions, honest updates can be more dispersed, potentially reducing the separability on which robust aggregation relies. The separate population study varies attacker share but not heterogeneity.

The matrix-injection condition is a benchmark-specific low-rank structured stressor. It is not presented as an optimized or stealthy attack, and its weak effect relative to sign flipping should not be generalized to every structured perturbation. An attacker that learns a sensitive subspace, coordinates against the server rule, or optimizes for a particular target may produce a substantially stronger matrix-structured attack.

Similarly, the adaptive attack uses the malicious client's own current update and a norm envelope. It is not omniscient. No malicious client is given honest clients' private data or exact honest updates. Backdoor persistence, stealth, and targeted attack success are not primary endpoints. Privacy mechanisms, secure aggregation, client authentication, and trusted execution are also outside the experiment.

Finally, the paired Wilcoxon tests are exploratory and uncorrected for multiple comparisons. Bootstrap intervals summarize the tested dataset-seed pairs but do not turn the two datasets into a statistically representative sample of all FL tasks. Reproduction on additional datasets, deeper models, heterogeneous partitions, and adaptive server-aware attackers is necessary before drawing broader quantitative conclusions.

8. Conclusion

This study introduced a severity-calibrated adversarial benchmark that evaluates label corruption, Gaussian noise injection, sign-flip poisoning, norm-bounded adaptive poisoning, and structured low-rank matrix injection within one federated-learning protocol. Across 880 runs and 17,600 rounds, coherent direction emerged as a more useful explanation of attack damage than perturbation magnitude alone. Under arithmetic mean at 20% malicious participation, high-severity sign flipping caused an 87.39-percentage-point average final-accuracy loss even though its update distortion was smaller than the Gaussian and matrix-injection stressors.

Robust aggregation sharply reduced the observed degradation in the controlled main regime, but the result is conditional on known attacker bounds, approximately IID data, full participation, and attacks that were not optimized specifically to evade each defense. The benchmark therefore does not identify a universally secure aggregator. Its contribution is a security evaluation method: compare attacks under matched baselines, make severity semantics explicit, report malicious population share, and characterize update geometry rather than reducing adversarial strength to one norm. These practices make federated-learning robustness claims more interpretable and easier to reproduce.

Declarations

Competing interests:

The author declares no competing interests.

Ethics approval:

Not applicable. The study used existing benchmark datasets and did not involve recruitment of human participants, clinical intervention, or animal experimentation.

Declarations

of generative AI and AI-assisted technologies in the manuscript preparation process

Funding:

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

Author contributions:

M.S.I.S. conceived and designed the study, developed the adversarial benchmark methodology and attack-severity framework, implemented and executed the experiments, analyzed and interpreted the results, prepared the figures and tables, and wrote the manuscript. M.S.I.S. reviewed and revised the manuscript and approved the final version.

Data availability:

The data generated and analyzed during this study consist of per-run benchmark outputs, per-round training trajectories, paired clean baselines, attack-severity summaries, and malicious-population sensitivity results. No proprietary dataset was used. The generated benchmark data and supporting materials are available from the corresponding author upon reasonable request.

References

1. McMahan B, Moore E, Ramage D, Hampson S, Aguera y Arcas B (2017) Communication-Efficient Learning of Deep Networks from Decentralized Data. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics. PMLR 54; :1273–1282
2. 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
3. Lyu L, Yu H, Yang Q (2020) Threats to Federated Learning: A Survey. arXiv:2003.02133
4. Tolpegin V, Truex S, Gursoy ME, Liu L (2020) Data Poisoning Attacks Against Federated Learning Systems. arXiv:2007.08432
5. Fang M, Cao X, Jia J, Gong NZ (2020) Local Model Poisoning Attacks to Byzantine-Robust Federated Learning. 29th USENIX Security Symposium; :1605–1622
6. Bagdasaryan E, Veit A, Hua Y, Estrin D, Shmatikov V (2020) How To Backdoor Federated Learning. Proceedings of AISTATS. PMLR 108; :2938–2948
7. Baruch M, Baruch G, Goldberg Y (2019) A Little Is Enough: Circumventing Defenses For Distributed Learning. Adv Neural Inf Process Syst. ;32
8. Shejwalkar V, Houmansadr A (2021) Manipulating the Byzantine: Optimizing Model Poisoning Attacks and Defenses for Federated Learning. Network and Distributed System Security Symposium (NDSS)
9. Blanchard P, El Mhamdi EM, Guerraoui R, Stainer J (2017) Machine Learning with Adversaries: Byzantine Tolerant Gradient Descent. Adv Neural Inf Process Syst. ;30

10. Yin D, Chen Y, Ramchandran K, Bartlett P (2018) Byzantine-Robust Distributed Learning: Towards Optimal Statistical Rates. Proceedings of the 35th International Conference on Machine Learning. PMLR 80; :5650–5659
11. El Mhamdi EM, Guerraoui R, Rouault S (2018) The Hidden Vulnerability of Distributed Learning in Byzantium. Proceedings of the 35th International Conference on Machine Learning. PMLR 80
12. Cao X, Fang M, Liu J, Gong NZ, FLTrust (2021) Byzantine-robust Federated Learning via Trust Bootstrapping. Network and Distributed System Security Symposium (NDSS); https://doi.org/10.14722/ndss.2021.24434
13. Fung C, Yoon CJM, Beschastnikh I (2020) The Limitations of Federated Learning in Sybil Settings. 23rd International Symposium on Research in Attacks, Intrusions and Defenses (RAID); :301–316
14. Xu H, Shu T (2024) Attack-model-agnostic defense against model poisonings in distributed learning. J Inform Secur Appl 82:103739. https://doi.org/10.1016/j.jisa.2024.103739
15. Xu H, Shu T (2024) Defending against model poisoning attack in federated learning: A variance-minimization approach. J Inform Secur Appl 82:103744. https://doi.org/10.1016/j.jisa.2024.103744
16. Wang S, Li X, Yu M, Yuan S, Guan Z (2025) FedESP: Effective, Stealthy, and Persistent backdoor attack on federated learning. J Inform Secur Appl 94:104223. https://doi.org/10.1016/j.jisa.2025.104223
17. Peng C, Xian H (2026) Lightweight orthogonal perturbation for privacy-preserving federated learning against poisoning attacks. J Inform Secur Appl 97:104345. https://doi.org/10.1016/j.jisa.2025.104345
18. Farid Babar F, Khan S, Parkinson S (2026) Defending federated learning against adversarial attacks: A systematic literature review. Cyber Secur Appl 4:100127. https://doi.org/10.1016/j.csa.2026.100127
19. Pedregosa F, Varoquaux G, Gramfort A et al (2011) Scikit-learn: Machine Learning in Python. J Mach Learn Res 12:2825–2830
20. Alpaydin E, Kaynak C (1998) Optical Recognition of Handwritten Digits [Dataset]. https://doi.org/10.24432/C50P49. UCI Machine Learning Repository
21. Wolberg WH, Street WN, Heisey DM, Mangasarian OL (1995) Computer-derived nuclear features distinguish malignant from benign breast cytology. Hum Pathol 26(7):792–796. https://doi.org/10.1016/0046-8177(95)90229-5
22. Nawshin F, Unal D, Suganthan PN (2026) Novel defense strategies for concurrent data and model poisoning attacks in federated learning. Knowl Based Syst 116057. https://doi.org/10.1016/j.knosys.2026.116057

Figures

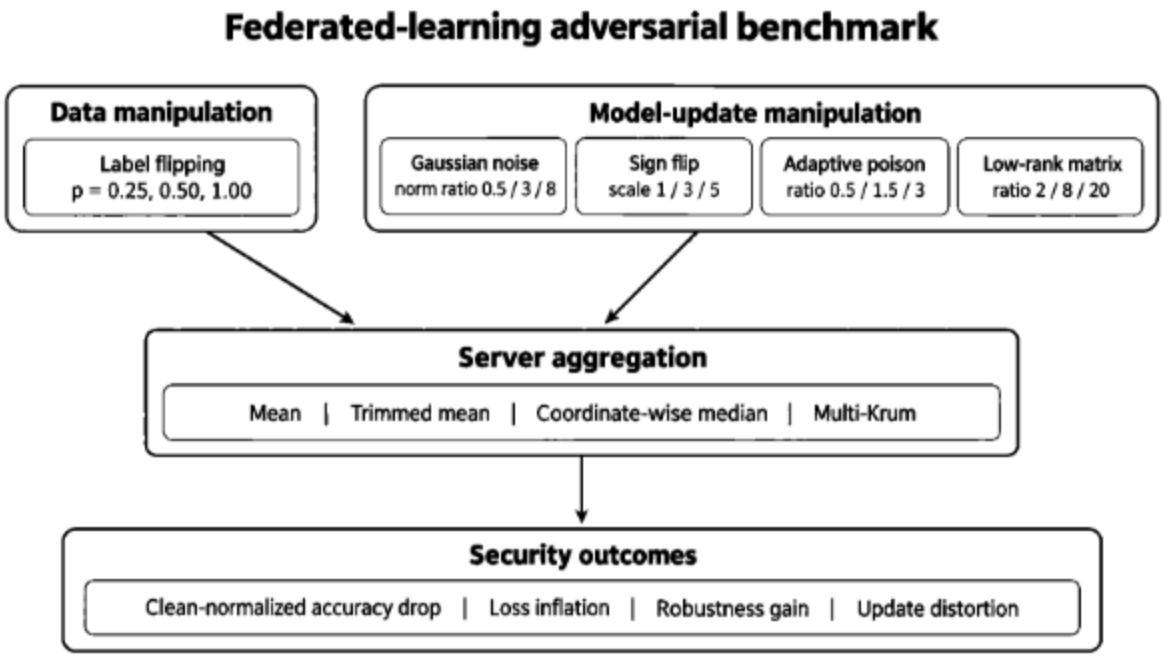


Figure 1

Benchmark threat taxonomy. Severity is calibrated within each attack family rather than treated as a common physical unit.

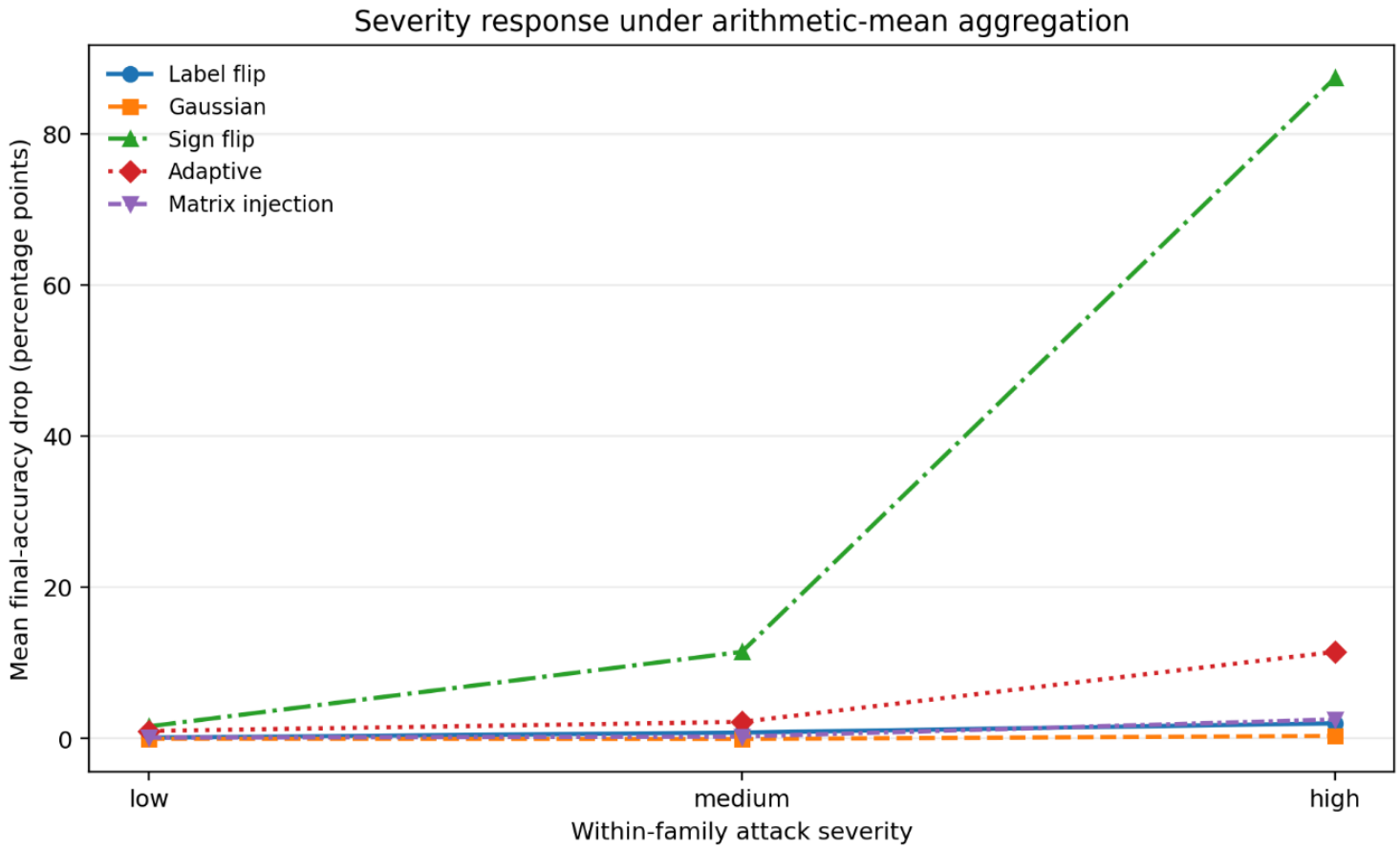


Figure 2

Mean-aggregation accuracy degradation across the three within-family severity levels. Values are averaged over both datasets and five seeds.

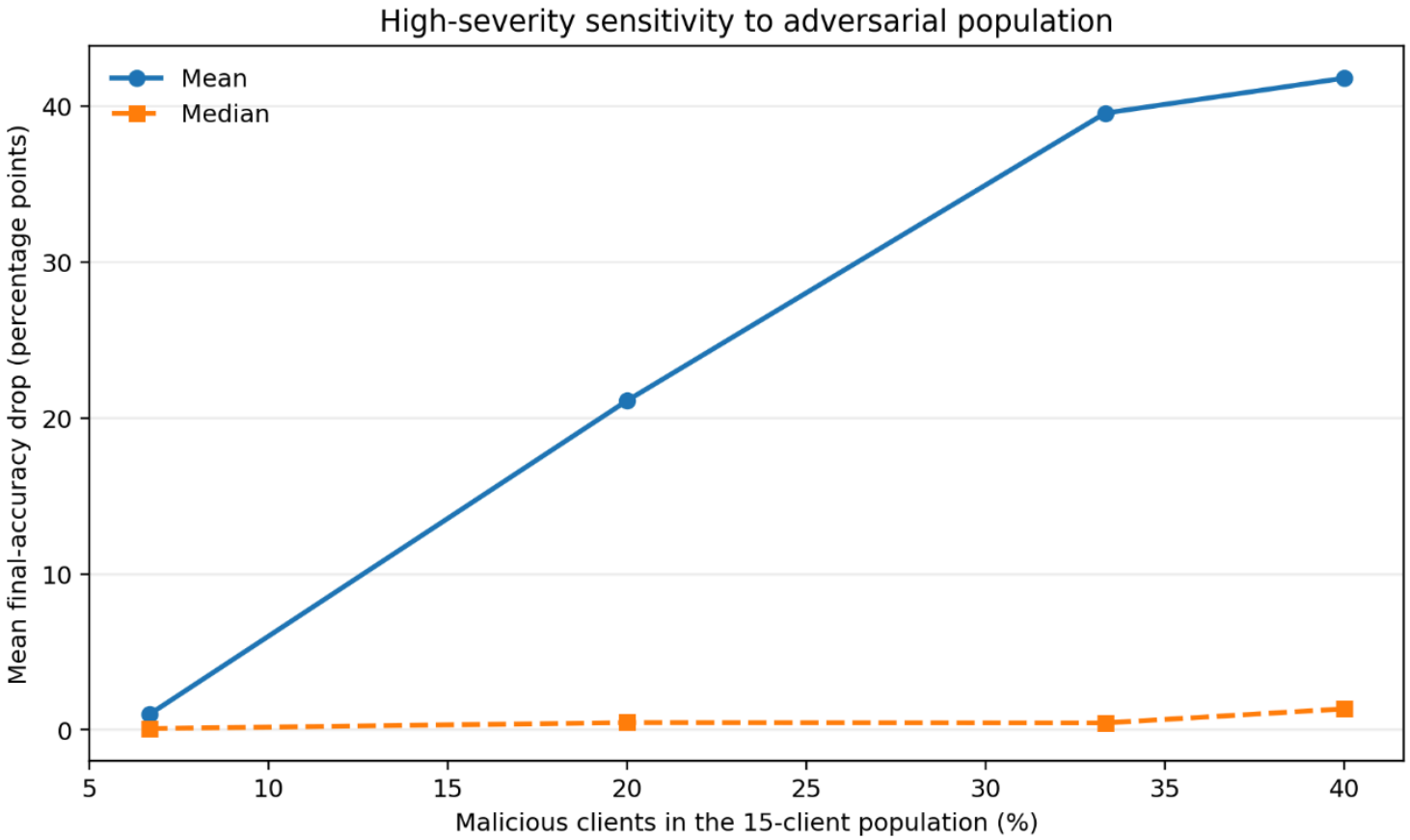


Figure 3

Mean accuracy degradation at high severity as the malicious population increases. Curves average five attacks, two datasets, and three seeds.

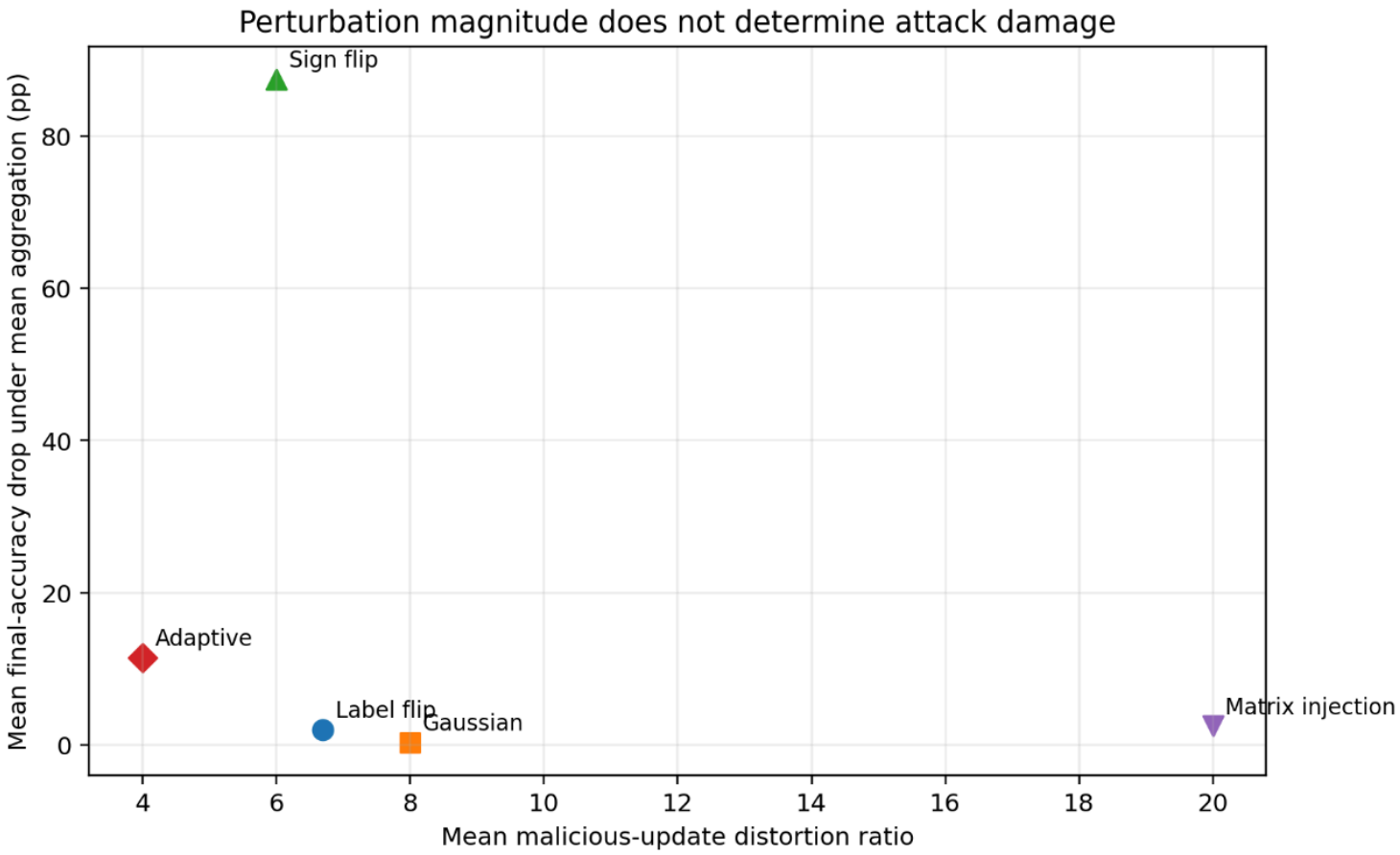


Figure 4

High-severity update distortion versus final-accuracy damage under mean aggregation. A larger perturbation norm does not imply a more effective attack.