AEGIS: A BYZANTINE-RESILIENT FEDERATED LEARNING PROTOCOL FOR HIGHLY NON-IID ENVIRONMENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated Learning (FL) is vulnerable to Byzantine attacks, where compromised clients submit poisoned updates to degrade the global model. Existing robust aggregators force a harsh trade-off: distance-based methods incur an $O(k^2d)$ bottleneck and penalize honest statistical drift, while efficient $O(kd)$ coordinate-wise median filters remain blind to directional poisoning. We introduce AEGIS, an $O(kd)$ aggregation protocol for high data-skew environments that replaces distance-based filtering with a two-pass dual-metric anomaly detector using adaptive Median Absolute Deviation (MAD) thresholding and cross-round reputation tracking. On CIFAR-10 under a 30% Byzantine threat, AEGIS holds accuracy within 10% of the clean baseline across Sign-Flipping, Label-Flipping, and Volume Spam while preserving linear scalability. We also document where statistical filtering breaks down: variance-envelope attacks (ALIE) collapse the protocol to chance-level accuracy, while Sybil collusion degrades it progressively, failing only once cloned identities grow numerous enough to capture the median majority.

1 INTRODUCTION

Federated Learning (FL) enables edge devices to collaboratively train a shared global model without exchanging raw, privacy-sensitive data. However, this decentralized architecture introduces a critical attack vector: Byzantine clients. Whether compromised by attackers or suffering from hardware faults, these nodes can transmit poisoned model updates to degrade accuracy, insert backdoors, or prevent convergence (Blanchard et al., 2017). Securing FL is complicated by statistical heterogeneity: in highly Non-IID settings, legitimate clients often produce weight updates that diverge from the global average purely due to local data skew, and standard robust aggregators discard these honest updates as anomalies. Distance-based methods such as Krum further scale quadratically $O(k^2d)$ with the number of clients, posing a bottleneck for massively distributed deployments.

We propose AEGIS to address these limitations. The protocol separates adversarial anomalies from natural Non-IID variance using two-pass median decontamination, adaptive credit scoring, and historical reputation tracking. Rather than relying on a single spatial metric, AEGIS jointly evaluates Euclidean distance and cosine similarity, applying MAD-based adaptive thresholds that adjust to per-round update variance. We demonstrate robust defense against conventional attacks at $O(kd)$ cost, compare empirical dynamics against baseline aggregators, and explicitly document the limits of statistical filtering against advanced optimization attacks such as ALIE and Sybil.

2 RELATED WORK

Byzantine resilience in FL has been approached from several directions, each with distinct trade-offs in computational cost, threat coverage, or privacy assumptions.

Distance-Based Aggregators: Krum and Multi-Krum (Blanchard et al., 2017) select updates using pairwise distances, providing provable resilience but requiring $O(k^2d)$ computation, and frequently failing against optimization-based attacks such as ALIE (Baruch et al., 2019). Bulyan (El Mhamdi et al., 2018) fixes some of Krum's vulnerabilities via two-phase meta-aggregation, but its $O(k^3d)$ cost is prohibitive.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Statistical and Spectral Methods: Trimmed Mean and Coordinate-wise Median (Yin et al., 2018) operate efficiently but assume coordinate-wise independence, leaving them vulnerable to Inner Product Manipulation (IPM) (Xie et al., 2019a). Divide-and-Conquer (Shejwalkar & Houmansadr, 2021) employs spectral methods to detect correlated perturbations but incurs significant overhead, and a recent meta-analysis (Fang et al., 2025) confirms that optimization-based attacks continue to bypass standard statistical rules.

Trust and Hybrid Defenses: FLTrust (Cao et al., 2021) and Zeno (Xie et al., 2019b) score updates against a server-side root dataset, violating pure FL privacy norms and struggling in extreme Non-IID settings. FLAME (Nguyen et al., 2022) combines cosine clustering and adaptive clipping but relies on $O(k^2 d)$ HDBSCAN, and CONTRA (Awan et al., 2021) tracks accumulated gradient histories via $O(k^2)$ pairwise cosine similarities to detect colluding Sybils. In contrast, AEGIS targets resilience through multi-metric filtering and scalar reputation tracking against an $O(kd)$ cleansed median, avoiding pairwise bottlenecks, global momentum buffers, and external clean datasets.

3 THREAT MODEL AND ENVIRONMENT

3.1 SYSTEM ARCHITECTURE AND DATA HETEROGENEITY

We consider a cross-device FL architecture with a central parameter server and $n = 30$ clients. To simulate realistic data skew, we use Shard-based partitioning (McMahan et al., 2017): the dataset is sorted by label and each client is allocated exactly 4 random shards, ensuring no single client holds a representative sample of the global distribution.

3.2 ADVERSARY AND PARTICIPATION MODEL

We evaluate AEGIS under a Fixed Adversary with Random Participation model. The identities of compromised nodes are fixed at the start of training. During each round, a random subset of clients $m \in [15, 25]$ is selected, so the number of participating Byzantine clients follows a Hypergeometric distribution $X \sim \text{Hypergeometric}(N = 30, f = 9, m)$, where $f = 9$ represents a 30% global Byzantine presence. In Sybil evaluations, each physical adversary operates with a multiplicity of 2, generating two distinct synthetic identities to assess resilience against identity-spoofing collusion.

3.3 ATTACK HIERARCHY

We categorize the evaluated threats into three vectors: gradient manipulation, metadata/label poisoning, and structural identity attacks.

- **Sign-Flipping:** Byzantine clients negate their local pseudo-gradients ($\Delta_k = -\Delta_{\text{local}}$) to steer the model toward divergence.
- **Inner Product Manipulation (IPM)** (Xie et al., 2019a)**:** Attackers submit a scaled inverse of the collective mean gradient, $\Delta_k = -\epsilon\mu$ with $\epsilon = 0.5$, forcing gradient ascent.
- **A Little Is Enough (ALIE)** (Baruch et al., 2019)**:** An omniscient adversary crafts poisoned vectors that lie strictly within the variance envelope of the honest distribution ($z = 1.0$), bypassing crude outlier detection.
- **Targeted Label-Flipping:** Complement label-flipping ($y_{\text{poisoned}} = (C - 1) - y$) induces misclassifications while maintaining high reported confidence.
- **Volume Spam:** Attackers inflate their reported sample size n_k to hijack the weighted average in aggregators such as FedAvg.
- **Sybil & Collusion:** Multiple adversarial identities coordinate updates (multiplicity 2) to overwhelm the median anchor in a majority-hijacking scenario.

4 THE AEGIS PROTOCOL

AEGIS executes a six-step pipeline per round (Algorithm 1). By using Introselect for median computations and avoiding pairwise distance matrices, it maintains an asymptotic complexity of $O(kd)$.

Algorithm 1 AEGIS: Two-Pass Byzantine-Resilient Aggregation with Soft Directional Penalty

Require: Updates $\{(W_k, n_k)\}_{k=1}^{K}$, global model W_{global}^{t}, round t, reputation state $\{R_k\}$
Ensure: Aggregated model W_{global}^{t+1}

// Step 1: Compute Pseudo-Gradients
for each client k **do**
  $\Delta_k \leftarrow W_k - W_{\text{global}}^{t}$
end for
// Step 2: Two-Pass Robust Center (Median Decontamination)
$\Delta_{\text{pre}} \leftarrow \text{median}_k(\Delta_k)$ ▷ Pass 1: Preliminary median
for each client k **do**
  $\cos_k^{\text{pre}} \leftarrow \dfrac{\Delta_k \cdot \Delta_{\text{pre}}}{\|\Delta_k\|_2 \|\Delta_{\text{pre}}\|_2}$
end for
$S \leftarrow \{k : \cos_k^{\text{pre}} \geq \tau_{\text{pass1}}\}$ ▷ Screen adversarial directions
$\Delta_{\text{med}} \leftarrow \text{median}_{k \in S}(\Delta_k)$ ▷ Pass 2: Debiased clean median
// Step 3: Dual Anomaly Scoring (all clients re-scored)
for each client k **do**
  $E_k \leftarrow \|\Delta_k - \Delta_{\text{med}}\|_2$ ▷ Euclidean distance
  $\text{Cos}_k \leftarrow \dfrac{\Delta_k \cdot \Delta_{\text{med}}}{\|\Delta_k\|_2 \|\Delta_{\text{med}}\|_2}$; $P_k \leftarrow 1 - \text{Cos}_k$ ▷ Cosine penalty
end for
// Step 4: Adaptive MAD Gatekeeper (Euclidean only)
$E_{\text{med}} \leftarrow \text{median}(E_k)$; $D \leftarrow \text{median}(|E_k - E_{\text{med}}|)$
$K_{\text{phase}} \leftarrow K_{\max} - (K_{\max} - K_{\min}) \cdot \min\left(\frac{t}{t_{\text{warmup}}}, 1\right)$
$CV \leftarrow D / E_{\text{med}}$; $K \leftarrow \max\big(K_{\text{floor}},\ K_{\text{phase}}(1 + \eta\, CV)\big)$
$T \leftarrow E_{\text{med}} + K \cdot D$; $A \leftarrow \{k : E_k \leq T\}$ ▷ Magnitude gate
// Step 5: Volume Bounding, Reputation & Soft Credit Scoring
for each $k \in A$ **do**
  $v_k \leftarrow \min\big(n_k,\ 2.0 \times \text{median}_{j \in A}(n_j)\big)$ ▷ Volume clip
  $R_k \leftarrow \gamma R_k + (1 - \gamma) P_k$ ▷ Reputation EMA
  $C_k \leftarrow v_k / (E_k + \alpha P_k + \lambda R_k)$ ▷ Soft suppression
end for
$\beta_k \leftarrow C_k / \sum_{j \in A} C_j$ ▷ Normalized weights
// Step 6: Aggregation
$W_{\text{global}}^{t+1} \leftarrow \sum_{k \in A} \beta_k \cdot W_k$
return W_{global}^{t+1}

Two-pass median decontamination (Steps 1–2). Optimization attacks such as IPM are designed to drag the coordinate-wise median off-center; once the reference is compromised, the aggregator begins rejecting legitimate clients whose updates reflect natural data skew. AEGIS therefore computes the robust center in two passes: a preliminary median Δ_{pre} pre-screens clients by cosine alignment, provisionally removing those with $\cos < \tau_{\text{pass1}}$ ($\tau_{\text{pass1}} = -0.3$), and a debiased median Δ_{med} is recomputed solely from the survivors S.

Dual scoring and adaptive gating (Steps 3–4). Every client is re-scored against Δ_{med} using a Euclidean distance $E_k = \|\Delta_k - \Delta_{\text{med}}\|_2$ (magnitude, detecting additive noise) and a cosine penalty $P_k = 1 - \text{Cos}_k$ (direction, detecting sign/label flips). A single hard gate rejects magnitude outliers using a self-calibrating MAD threshold $T = E_{\text{med}} + K \cdot D$, where the multiplier K decays over a warmup period and is modulated by the swarm coefficient of variation $CV = D/E_{\text{med}}$. The cosine penalty deliberately bypasses this hard gate and feeds only into the soft credit score, so honest Non-IID specialists are down-weighted rather than discarded outright.

Reputation and credit scoring (Steps 5–6). A cross-round EMA reputation $R_k^{t+1} = \gamma R_k^t + (1 - \gamma) P_k^t$ penalizes persistent stealth attackers (updated only for approved clients, so adversarial noise cannot pollute the record), and claimed sizes are clipped as $v_k = \min(n_k,\ 2.0 \times \text{median}(n_{\text{approved}}))$

to neutralize Volume Spam. The final credit score and normalized weight are

$$C_k = \frac{v_k}{E_k + \alpha P_k + \lambda R_k}, \qquad \beta_k = \frac{C_k}{\sum_{j \in A} C_j}, \tag{1}$$

and the new global model is $W_{\text{global}}^{t+1} = \sum_{k \in A} \beta_k W_k$.

Architectural note — soft directional penalty. The cosine penalty enters the credit denominator scaled by $\alpha = 30$. A sign-flip attacker incurs a large αP_k term that sharply reduces its weight, while an honest Non-IID client with slight negative alignment is only softly down-weighted rather than excluded. This continuous formulation eliminates false positives while maintaining equivalent attacker suppression.

Complexity. Introselect guarantees $O(k)$ per coordinate across d coordinates for the two-pass median; dual scoring is $O(d)$ per client; MAD thresholding and credit scoring operate on k scalars in $O(k)$. No step requires pairwise client comparisons, so the total per-round cost is $O(kd)$—identical to FedAvg and avoiding Krum's $O(k^2 d)$ bottleneck.

5 Convergence Analysis

We now sketch a non-convex convergence guarantee for AEGIS; the complete derivation is deferred to an extended technical report. Assume the global objective f is L-smooth, honest stochastic gradients are unbiased ($\mathbb{E}[g_i] = \nabla f$) with bounded noise $\mathbb{E}\|g_i - \nabla f\|^2 \le \sigma_H^2$, and honest heterogeneity is bounded by V_H. Let $\hat{g}_t = \sum_{k \in A} \beta_k \Delta_k$ be the AEGIS pseudo-gradient, $\bar{g}_H^{(t)}$ the honest mean, and define the aggregation error $e_t = \hat{g}_t - \bar{g}_H^{(t)}$.

Robustness certificate. The per-round error admits the closed-form bound

$$\|e_t\| \le B_t = \underbrace{(1 + w_B)\sqrt{\kappa_H V_H}}_{\text{honest distortion}} + \underbrace{\frac{2M|B_A|}{S}}_{\text{volume leakage}} + \underbrace{w_B \Delta_m}_{\text{residual Byz. bias}}, \tag{2}$$

where $|B_A|$ is the number of Byzantine clients surviving the filter, w_B their aggregate weight, M the volume clip, S the sample normalizer, κ_H the honest reweighting factor, and Δ_m the median displacement. Each term is governed by a distinct AEGIS mechanism: volume clipping bounds M, the adaptive gate and reputation EMA shrink w_B and $|B_A|$, and two-pass decontamination limits Δ_m.

Convergence bound. Writing $B_t^2 \le \kappa_A V_H + \zeta_A$ with $\kappa_A = 3(1 + w_B)^2 \kappa_H$ and residual bias $\zeta_A = \zeta_{\text{vol}} + \zeta_{\text{med}}$, a smoothness–telescoping argument with constant step $0 < \eta < 1/(2L)$ yields

$$\begin{aligned} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\|\nabla f(x_t)\|^2 \;\le\; & \frac{f(x_0) - f^\star}{\eta \left(\frac{1}{2} - L\eta\right) T} + \frac{L\eta\, \sigma_H^2}{\frac{1}{2} - L\eta} \\ & + \frac{\frac{1}{2} + L\eta}{\frac{1}{2} - L\eta} \Big(\kappa_A V_H + \zeta_A\Big). \end{aligned} \tag{3}$$

When filtering is perfect ($|B_A| = 0 \Rightarrow w_B = 0$, $\zeta_A = 0$, $\kappa_A = 3\kappa_H$), the bound collapses to the standard non-convex SGD rate $O(1/T) + O(\eta \sigma_H^2)$ plus an irreducible honest-heterogeneity term $3\kappa_H V_H$ that reflects the cost of reweighting under Non-IID data.

Connection to empirical limits. The residual bias ζ_A is governed by the score-suppression ratio r: for naive attacks $r \to 0$, so $w_B \to 0$ and the Byzantine terms vanish, matching AEGIS's recovery of near-clean accuracy under Sign- and Label-Flip. Under omniscient ALIE, attackers become statistically indistinguishable from honest outliers ($r \to 1$), ζ_A stays bounded away from zero, and the guarantee degrades to a non-vanishing error floor, which is the collapse we observe empirically under ALIE (Section 7.5).

6 Experimental Setup

Hyperparameter rationale: The values in Table 1 fall into three groups. The *gating* group controls the magnitude filter: the multiplier decays from a permissive $K_{\max} = 6$ (tolerant of noisy early

Table 1: AEGIS Hyperparameters

Symbol	Parameter	Value
τ_{pass1}	Pass 1 cosine threshold	-0.3
$K_{\max}$ / $K_{\min}$	Initial / final threshold multiplier	6.0 / 2.0
K_{floor}	Minimum K clamp (Non-IID)	4.0
t_{warmup}	Warmup decay period	300 rounds
η	Variance sensitivity	3.0
α	Cosine penalty weight	30.0
γ	Reputation EMA decay	0.95
λ	Reputation scaling factor	20.0

updates) to a strict $K_{\min} = 2$ at convergence over a $t_{\text{warmup}} = 300$-round schedule, floored at $K_{\text{floor}} = 4$ so natural Non-IID drift is never mistaken for an attack, with $\eta = 3$ setting how sharply the gate tightens as the swarm coefficient of variation shrinks. The *directional* group drives soft suppression: $\tau_{\text{pass1}} = -0.3$ excludes only strongly anti-aligned vectors from the debiased median, while $\alpha = 30$ makes the cosine penalty decisive against sign/label flips yet continuous enough to spare honest specialists. The *reputation* group ($\gamma = 0.95$, $\lambda = 20$) keeps a slow-moving memory of past misalignment to expose stealth attackers. These values are held fixed across every experiment below.

Model & Data: We evaluate a CNN with GroupNorm (groups = 2) and Dropout (0.25), comprising $\sim$290K parameters, trained on CIFAR-10. GroupNorm avoids the instability of BatchNorm running statistics across Non-IID clients. Local SGD uses momentum (0.8), learning rate (0.01), and batch size (32). AEGIS hyperparameters are detailed in Table 1. Evaluation uses up to 1000 communication rounds with $n = 30$ total clients and 15–28 active clients per round, with early stopping (patience 100 rounds, min-delta 0.001 on validation loss).

7 Results and Discussion

7.1 Baseline Convergence and Aggregator Comparison

In the absence of adversaries, AEGIS converges consistently despite moderate-to-severe heterogeneity (4 shards/client), achieving a mean accuracy of 75.53% with a standard deviation of only 0.18% across three seeds. We find that traditional distance-based and statistical aggregators perform far better on clean, highly Non-IID data than commonly assumed in the literature (Fig. 1, Figs. 2–3): FedAvg achieves 76.16%, AEGIS 76.08%, CWMed 72.31%, and Krum 71.30%, establishing a tight 71–76% baseline ceiling. FoolsGold (73.80%) misclassifies the natural divergence of Non-IID clients as adversarial and suppresses legitimate updates, while Bulyan (67.69%) inherits Krum's spatial limitations and loses further precision during its trimming phase.

7.2 Resilience Across Conventional Attacks

AEGIS was evaluated against Sign-Flipping, Targeted Label-Flipping, and IPM (Fig. 4). The protocol neutralizes naive poisoning and preserves model integrity despite the high data skew. Even under a 30% Label-Flip attack, AEGIS retains 74.34% accuracy, matching its clean baseline: dual-metric screening filters targeted adversaries effectively, leaving honest clients to converge. The optimization-based IPM attack ($\epsilon = 0.5$) drags accuracy to 63.8%; unlike the naive flips, its severity is governed almost entirely by the perturbation magnitude ϵ (Table 2).

IPM sensitivity to ϵ. How much damage IPM does depends almost entirely on the perturbation magnitude ϵ, and the magnitude gate offers little resistance. Across the sweep in Table 2 the detection rate stays at 0% until $\epsilon = 1.0$, because a scaled-inverse gradient $-\epsilon\mu$ shifts the direction of an update rather than inflating its size. For small ϵ this poses no real threat: the attack is too weak to move the aggregate, and accuracy stays near the clean baseline (71.9% at $\epsilon = 0.01$, 70.6% at $\epsilon = 0.1$). At larger ϵ the poisoned gradient grows strong enough to pull the weighted average off course despite the cosine penalty, and accuracy drops to 43.7% ($\epsilon = 1.0$) and 17.7% ($\epsilon = 2.0$). So it

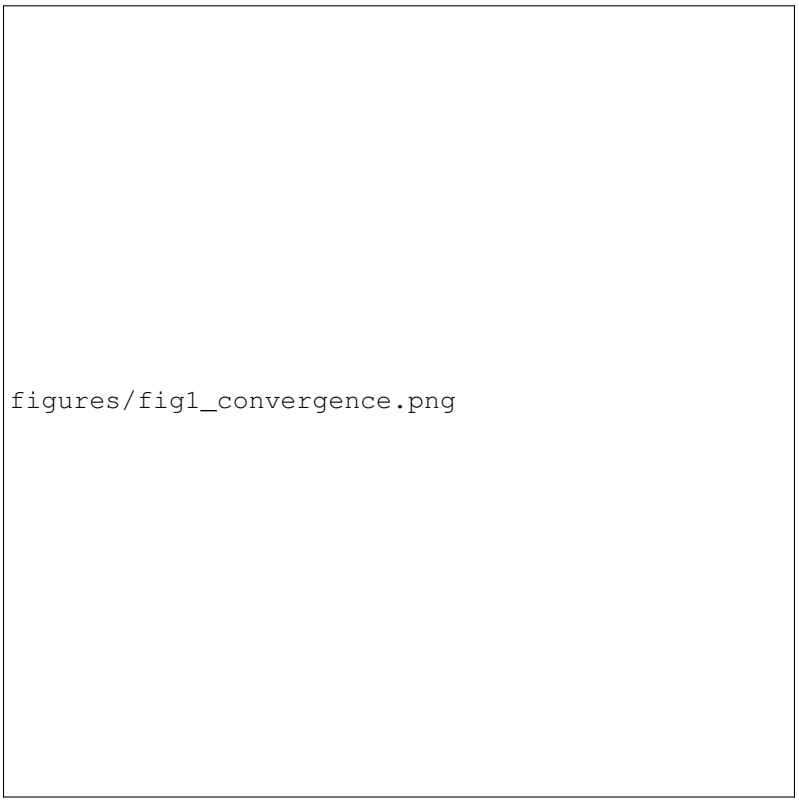


Figure 1: Test accuracy convergence of standard aggregators vs. AEGIS without attacks in a 4-shard Non-IID environment. CWMed and Krum trigger early stopping due to instability plateaus.

Table 2: AEGIS Under IPM at Varying Perturbation Magnitude ϵ ($f = 30\%$)

IPM ϵ	**Accuracy**	**Loss**	**Detection**
0.01	71.86%	0.8189	0.0%
0.1	70.59%	0.8379	0.0%
0.5	63.77%	1.0187	0.0%
1.0	43.67%	1.6033	0.0%
2.0	17.69%	2.2481	5.8%

is the cosine term, not the Euclidean gate, that holds IPM back at moderate ϵ, matching our component ablation (Section 7.4), where removing the cosine penalty rather than median decontamination is what collapses IPM resilience.

7.3 COMPARATIVE EVALUATION

To contextualize AEGIS, we subjected all baseline aggregators to identical threat models (Table 3). Under a 30% Sign-Flip attack, FedAvg collapses to 40.04% and FoolsGold fails completely (34.20%), while Krum shows the highest resilience (70.05%) as its $O(k^2 d)$ pairwise matrix isolates massive spatial outliers. The landscape shifts under Label-Flipping: doing nothing (FedAvg) yields 70.80%, indicating the 70% honest majority natively overpowers semantic poisoning in this Non-IID setting, yet AEGIS achieves the highest robustness at 74.34%. This points to a genuine trade-off: Krum's hard spatial rejection marginally outperforms AEGIS on extreme-magnitude attacks (Sign-Flip), whereas AEGIS defends better against the stealthier semantic poisoning of Label-Flip, and it does so in strictly linear $O(kd)$ time.

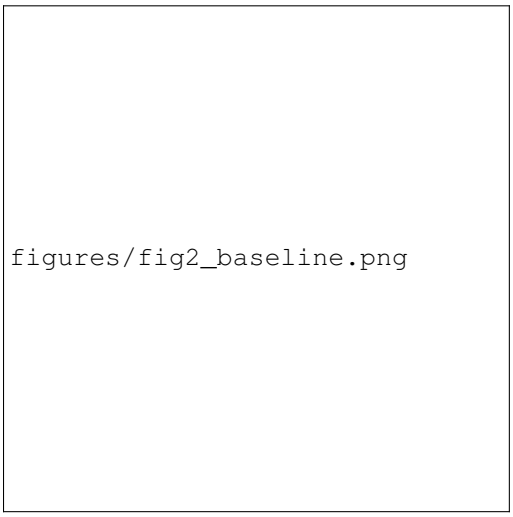


Figure 2: Final baseline accuracies without attacks. Krum and CWMed surpass 71% despite the extreme data skew.

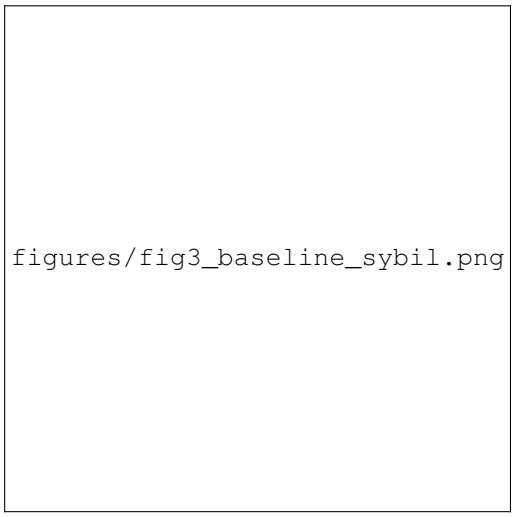


Figure 3: Final clean baseline accuracies comparing AEGIS against Sybil-centric (FoolsGold) and meta-aggregation (Bulyan) defenses.

Table 3: Final Accuracy (%) Across Baselines and Attacks ($f = 30\%$, Non-IID)

Defense	**No Atk.**	**Sign-Flip**	**Label-Flip**	**IPM**	**ALIE**	**Vol-Spam**
FedAvg	76.16	40.04	70.80	70.51	66.78	6.52
Krum	71.30	70.05	72.85	12.97	34.18	72.14
CWMed	72.31	49.64	62.07	10.87	23.21	60.47
Bulyan	67.69	62.58	68.51	10.03	20.39	50.47
FoolsGold	73.80	34.20	67.06	71.01	71.15	62.17
AEGIS	76.08	66.61	74.34	63.32	10.00	72.21

7.4 COMPONENT ABLATION

To isolate the contribution of individual mechanisms, we disable one component at a time under the attack it primarily defends (Table 4). Two components prove decisive. Under Volume Spam, removing the volume clip collapses accuracy from 73.12% to 6.56%, confirming it as the sole barrier against size-inflation hijacking. Under IPM, removing the cosine penalty nearly halves accuracy (63.39% $\rightarrow$ 36.80%), verifying that directional suppression—not magnitude filtering—carries the defense against inner-product manipulation; two-pass median decontamination contributes only marginally here (63.39% vs. 63.59%), since IPM's scaled-inverse gradients remain within magnitude bounds and are instead caught by their angular deviation.

7.5 ADVANCED THREATS: VOLUME SPAM AND VARIANCE ENVELOPES

AEGIS nullifies Volume Spam (72.21%, Table 3)—where unbounded weighting collapses FedAvg to 6.52%—via the median-based volume clipping of Step 5, which prevents adversaries from scaling their reported sizes without bound; its necessity is confirmed by ablation (Table 4). The omniscient ALIE attacker, however, exposes a fundamental vulnerability of median-based decontamination. Under a 30% ALIE attack ($z = 1.0$), AEGIS collapses to 10.00%—random guessing—whereas doing nothing (FedAvg) retains 66.78%, as the honest majority natively outweighs the poisoned vectors in a simple mean (Table 3).

Here AEGIS effectively works as an *attack amplifier*. ALIE attackers submit identically constrained vectors that lie strictly within the honest variance envelope; because these vectors are uniform, they hijack the two-pass median (Step 2). The adaptive Euclidean gatekeeper then evaluates honest Non-IID clients against this poisoned median, flags their natural drift as outliers, and discards them.

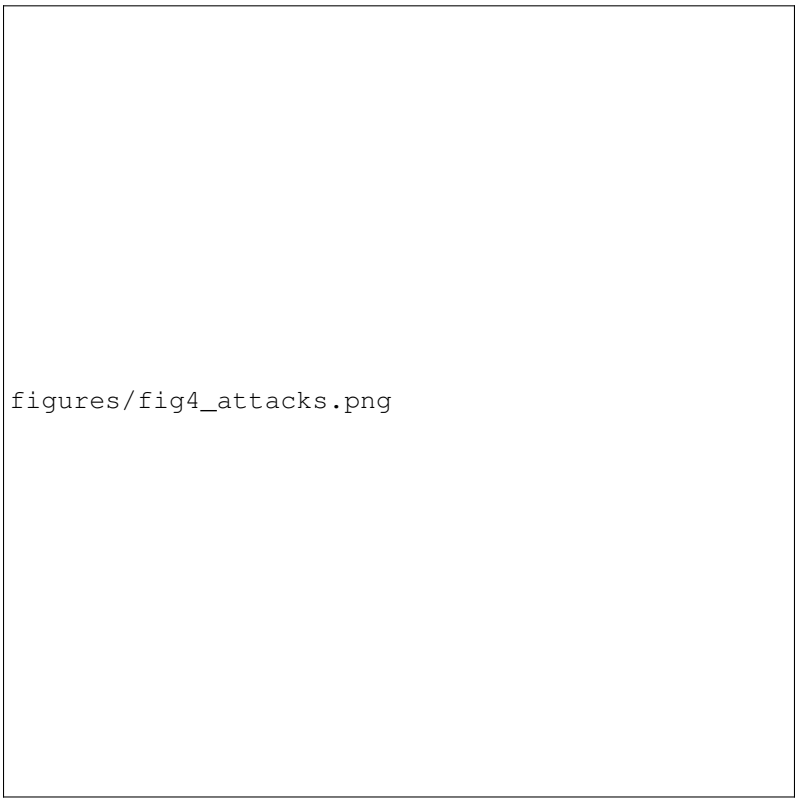


Figure 4: Test accuracy showing AEGIS's resilience across conventional and optimization-based attacks. The protocol tightly bounds the impact of both Sign-Flip and IPM.

Table 4: Component Ablation: Accuracy (%) with Individual Mechanisms Disabled ($f = 30\%$)

Attack	**Configuration**	**Accuracy**
Volume Spam	Full AEGIS	73.12
	− Volume Clipping	6.56
IPM ($\epsilon = 0.5$)	Full AEGIS	63.39
	− Median Decontamination	63.59
	− Cosine Penalty	36.80

The soft cosine penalty ($\alpha = 30$) makes matters worse: it suppresses honest clients with naturally low alignment more aggressively than the crafted ALIE gradients, which retain moderate cosine alignment, so weight is redistributed away from rare-class information and the bias compounds. This failure mode is not specific to our design; it affects any credit-scoring mechanism that treats cosine similarity as a quality signal once attacker alignment exceeds that of honest outliers. Consistently, FoolsGold achieves 71.15% against ALIE: because ALIE attackers must submit tightly coupled vectors to stay within the envelope, its historical-similarity logic correctly identifies them as collusion—a structural signal that AEGIS's per-round spatial logic misses.

7.6 STRUCTURAL ATTACKS: SYBIL MULTIPLICITY

We evaluate coordinated Sybil collusion by scaling the number of synthetic identities each physical adversary clones (k per attacker), which amplifies the effective Byzantine presence (Table 5). At $k = 1$ AEGIS retains 71.48% accuracy, as the honest majority still dominates the aggregate. Beyond $k = 1$, however, the cloned identities capture the coordinate-wise median and utility falls sharply—to 41.63% ($k = 2$), 33.52% ($k = 3$), and below 14% ($k \geq 4$). This delineates the protocol's

Table 5: AEGIS Under Sybil Collusion at Varying Multiplicity (k Clones per Attacker)

Multiplicity (k)	**Accuracy**	**Loss**
1	71.48%	0.8319
2	41.63%	1.4083
3	33.52%	1.6541
4	12.78%	2.0824
5	13.50%	2.1773

Table 6: AEGIS Performance Under Sign-Flip at Varying Fractions

Byzantine Fraction (f)	**Accuracy**	**Loss**
0% (Baseline)	75.12%	0.7358
10%	73.16%	0.7811
20%	71.31%	0.8585
30%	66.79%	1.0347
40%	30.33%	2.2319

structural breaking point: once colluding clones seize the median majority, per-round spatial filtering has no clean anchor to reference, echoing the $f = 40\%$ collapse in Table 6.

7.7 Resilience to Varying Byzantine Fractions

We subjected the protocol to a Sign-Flip attack while scaling the global Byzantine fraction $f \in \{10\%, 20\%, 30\%, 40\%\}$ (Table 6). AEGIS degrades gracefully up to 30%, but at $f = 40\%$ it suffers a collapse to 30.33%. In a highly Non-IID environment, a 40% coordinated adversarial presence combined with the natural skew of honest clients is sufficient to overpower the two-pass median (Step 2); once this anchor is compromised, the gatekeepers are evaluated against a poisoned reference, causing systemic failure.

7.8 Sensitivity to Data Heterogeneity

Varying the number of shards per client ($s \in \{2, 4, 6\}$) under benign and Label-Flip ($f = 30\%$) conditions (Table 7), we observe that extreme skew ($s = 2$) creates natural variance that overlaps with malicious distributions, lowering the baseline (68%) and widening the vulnerability gap (13%). At moderate-to-mild skew ($s \geq 4$), the two-pass system cleanly separates adversarial vectors from natural variance, holding degradation to a marginal 2–4%.

8 Limitations and Conclusion

Our evaluation exposes fundamental limits of per-round statistical filtering, consistent with known impossibility results (Baruch et al., 2019). (i) **Variance-envelope attacks:** omniscient ALIE hijacks the median anchor, tricking AEGIS into discarding honest clients and turning the protocol into an attack amplifier (10.00% accuracy). (ii) **Large-ϵ IPM:** the magnitude gate almost never flags inner-product manipulation (detection $\approx 0\%$), so the cosine penalty is its sole defense; this suffices for small-to-moderate ϵ ($\geq 70\%$ accuracy at $\epsilon \leq 0.1$) but is overwhelmed as ϵ grows, with accuracy falling to 43.7% at $\epsilon = 1.0$ and 17.7% at $\epsilon = 2.0$ (Table 2). (iii) **Reputation bootstrap failure:** under ALIE the model diverges before the EMA signal ($\gamma = 0.95$) accumulates enough to isolate compromised nodes. (iv) **Sybil thresholds:** coordinated Sybils erode the median-based anchor as their multiplicity grows; utility degrades from $k \geq 2$ onward and collapses once the clones seize the median majority at $k \geq 4$ (Table 5).

In summary, dual-metric filtering at $O(kd)$ cost is sufficient to neutralize conventional poisoning in highly Non-IID FL, while we also establish the precise conditions under which statistical filters break down. AEGIS achieves this via two-pass median decontamination and cross-round reputation tracking. Future work must bridge the gap exposed by ALIE—potentially by integrating history-

Table 7: AEGIS Accuracy Across Varying Non-IID Degrees ($f = 30\%$)

Shards per Client	No Attack	Label-Flip	Gap
2 (Extreme Skew)	68%	55%	13%
4 (Moderate Skew)	75%	71%	4%
6 (Mild Skew)	77%	74%	3%

based similarity constraints (e.g., FoolsGold logic) into the reputation system—and evaluate AEGIS at larger client scales ($k > 100$) where its $O(kd)$ complexity is most impactful.

AI USE STATEMENT

[TODO — required. State which tasks generative AI tools were used for (e.g., writing assistance, LaTeX formatting, code), which tasks they were not used for, and how the authors reviewed AI-assisted work. See the ICLR 2027 AI Policy for Authors.] We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies defenses against poisoning attacks in federated learning. All experiments use the public CIFAR-10 benchmark and simulated clients; no human subjects or private data are involved. We describe failure modes of our defense (ALIE and Sybil attacks) to help practitioners avoid overestimating its protection, rather than to enable new attacks; the attacks we evaluate are already published.

REPRODUCIBILITY STATEMENT

The full aggregation procedure is given in Algorithm 1, and all hyperparameters, which are held fixed across experiments, are listed in Table 1. The model architecture, local training settings, data partitioning, participation model and early-stopping criteria are described in Sections 3 and 6. Anonymized source code for all experiments will be provided as supplementary material.

REFERENCES

Sana Awan, Bo Luo, and Fengjun Li. CONTRA: Defending against poisoning attacks in federated learning. In *European Symposium on Research in Computer Security (ESORICS)*, pp. 455–475, 2021.

Gilad Baruch, Moran Baruch, and Yoav Goldberg. A little is enough: Circumventing defenses for distributed learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 8632–8642, 2019.

Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 119–129, 2017.

Xiaoyu Cao, Minghong Fang, Jia Liu, and Neil Zhenqiang Gong. FLTrust: Byzantine-robust federated learning via trust bootstrapping. In *Network and Distributed System Security Symposium (NDSS)*, 2021.

El Mahdi El Mhamdi, Rachid Guerraoui, and Sébastien Rouault. The hidden vulnerability of distributed learning in Byzantium. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp. 1840–1849, 2018.

Minghong Fang, Seyedsina Nabavirazavi, Zhuqing Liu, Wei Sun, Sundararaja Sitharama Iyengar, and Haibo Yang. Do we really need to design new Byzantine-robust aggregation rules? In *Network and Distributed System Security Symposium (NDSS)*, 2025.

H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1273–1282, 2017.

Thien Duc Nguyen et al. FLAME: Taming backdoors in federated learning. In *31st USENIX Security Symposium*, pp. 1415–1432, 2022.

Virat Shejwalkar and Amir Houmansadr. Manipulating the Byzantine: Optimizing model poisoning attacks and defenses for federated learning. In *Network and Distributed System Security Symposium (NDSS)*, 2021.

Cong Xie, Oluwasanmi Koyejo, and Indranil Gupta. Fall of empires: Breaking Byzantine-tolerant SGD by inner product manipulation. In *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 261–270, 2019a.

Cong Xie, Oluwasanmi Koyejo, and Indranil Gupta. Zeno: Distributed stochastic gradient descent with suspicion-based fault-tolerance. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pp. 6893–6902, 2019b.

Dong Yin, Yudong Chen, Kannan Ramchandran, and Peter Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp. 5650–5659, 2018.