Beyond Magnitude Filtering: A Dual-Stage Gradient Validation Mechanism for Defending Federated Intrusion Detection against Label-Flipping Poisoning in IoV Networks

Sarah H. Mnkash[1], Faiz A. Alawy[2], Israa T. Ali[3]*

[1] Computer Science College, University of Technology, Baghdad, Iraq. Email: cs.22.13@grad.uotechnology.edu.iq

[2] College of Engineering, Kent State University, Ohio, USA. Email: falalaw@kent.edu

[3] College of Engineering Technology for Computers and Artificial Intelligence, Northern Technical University, Kirkuk, Iraq. Email: israa.ali24@ntu.edu.iq

Corresponding Author:

Sarah H. Mnkash[1]* (cs.22.13@grad.uotechnology.edu.iq)

Abstract: Federated Learning (FL) enables distributed vehicular networks to collaboratively train intrusion detection models without sharing sensitive CAN-bus traffic data; however, this distributed paradigm remains vulnerable to poisoning attacks. Malicious participants can manipulate local updates through label-flipping or semantic gradient attacks that preserve plausible update magnitudes while altering gradient directions. Conventional magnitude-based defenses, such as Median Absolute Deviation (MAD) thresholding, detect only about 58% of attacks because they cannot identify semantic poisoning. This paper proposes a dual-stage gradient validation mechanism for securing federated intrusion detection in Internet of Vehicles (IoV) CAN-bus networks. The framework combines norm-deviation thresholding with cosine-similarity filtering to detect both magnitude-based and semantic poisoning attacks. The mechanism is integrated into an Adaptive Weighted Input (AWI) aggregation strategy that assigns continuous trust-based weights instead of binary exclusion, thereby preserving useful client contributions while limiting adversarial influence. Experiments using the CIC-IoV 2024 dataset under varying label-flipping attack severities show that the proposed method detects 91–95% of poisoned updates, improving detection by 33–37 percentage points over MAD-only filtering. The framework reduces global model drift from 41% while preserving an intrusion detection accuracy of 98% (ROC-AUC = 0.998). Convergence analysis demonstrates enhanced training stability with a 79% reduction in the variance of accuracy and reaching 95% of maximum performance in 19 rounds as opposed to 38 rounds without defense. Thus, the results show that directional gradient analysis is an effective, computationally feasible, and improved method for secure federated vehicular intrusion detection.

Keywords: Federated Learning (FL), Poisoning Attacks, Semantic Gradient Manipulation, Cosine Similarity Filtering, Internet of Vehicles (IoV), CAN-bus IDS, Byzantine Robustness..

INTRODUCTION

The deployment of Federated Learning (FL) for privacy-preserving intrusion detection in Internet of Vehicles (IoV) networks represents a significant architectural advancement over centralized approaches [1]. In the federated paradigm, each vehicle or Electronic Control Unit (ECU) trains a local detection model on its private CAN-bus traffic data and submits only model parameter updates to a

central aggregation server, which combines these updates into an improved global model that is redistributed to all participants. This architecture eliminates the privacy risks and bandwidth costs of transmitting raw vehicular traffic to remote servers, enables compliance with data protection regulations, and naturally accommodates the distributed geographical deployment of vehicular fleets [2]. The dual-stage defense pipeline proposed in this paper addresses this gap, as illustrated in Figure 1.

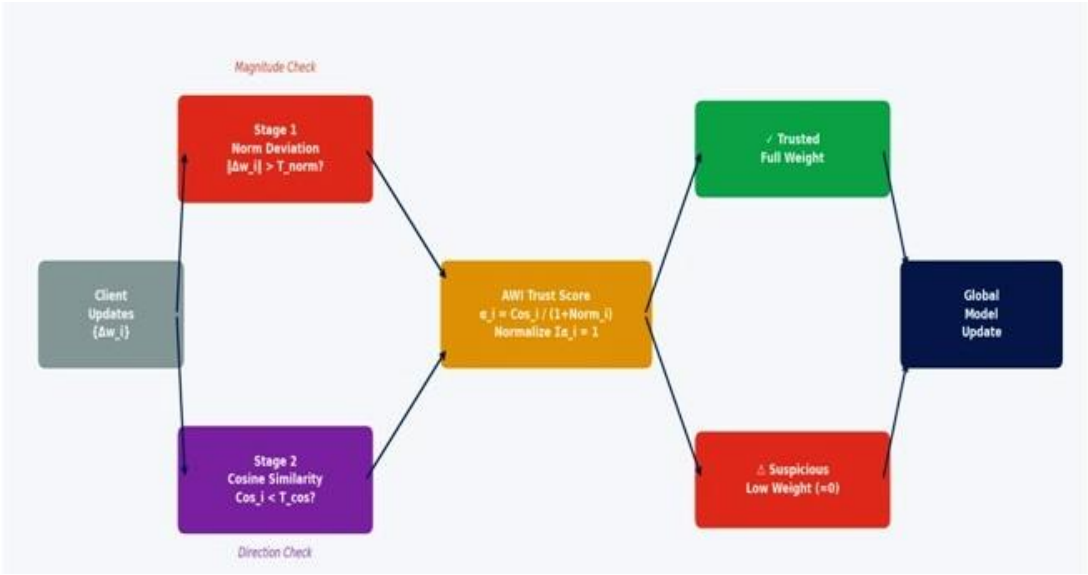

Figure 1. Dual-stage gradient validation pipeline: Stage 1 applies norm deviation thresholding (MAD-based) to detect magnitude anomalies; Stage 2 applies cosine similarity filtering to detect semantic poisoning attacks that evade magnitude-only filters. Both signals are synthesized into continuous AWI trust weights rather than binary exclusion decisions.

Deep learning and swarm intelligence optimization techniques applied to vehicular ad-hoc networks have further validated the potential of intelligent distributed approaches for enhancing vehicular network security [3]. However, the mechanism that makes federated learning privacy-preserving accepting model updates from untrusted clients simultaneously creates a fundamental security vulnerability that is arguably the most critical open challenge in FL-based IDS deployment.

Poisoning attacks exploit this vulnerability by corrupting the FL training process through manipulated local model updates. Label-flipping attacks the most direct and widely studied form involve malicious clients deliberately mislabeling portions of their local training data (marking attack traffic as benign or vice versa) before local model training, causing their local models to develop inverted or degraded decision boundaries that, when aggregated, corrupt the global model [4]. More sophisticated adversaries employ gradient manipulation attacks that directly modify parameter updates without altering training labels, crafting gradient vectors that satisfy statistical plausibility constraints while steering global optimization toward adversarial objectives [5]. Zhang et al. [6] demonstrated that even a modest proportion of poisoned clients (approximately 28% of total participants) can reduce FL-based NIDS performance by up to 28% under undefended FedAvg aggregation, confirming the operational severity of this threat in vehicular cyber security contexts.

The predominant approach to FL defense relies on magnitude-based anomaly detection, particularly Median Absolute Deviation (MAD) thresholding of gradient update norms [7]. MAD is theoretically motivated by its high breakdown point: with a 50% theoretical tolerance for contaminated samples, it remains reliable when up to half of clients are adversarial. However, MAD and analogous norm-based approaches share a structural blind spot: they evaluate the size of gradient updates but not their semantic direction. An adversary who engineers a poisoned update with a norm that falls within

the acceptable MAD-derived range successfully bypasses magnitude-only filters entirely, regardless of how aggressively the update misaligns the global gradient direction. Such semantic poisoning attacks can accumulate gradually across multiple training rounds, causing progressive model drift that may not manifest as detectable performance degradation until substantial corruption has occurred.

This paper seeks to fill the crucial void that is present in current defense mechanisms through an empirical and systematic study of a two-stage gradient validation technique by way of adding directional alignment analysis using cosine similarity to magnitude-based detection. To test our central hypothesis that combining norm deviation thresholding with cosine similarity filtering will result in better overall poisoning coverage than either approach alone we perform controlled experiments on the CIC-IoV 2024 dataset with different percentages of clients and different severities of label-flipping attacks. The dual-stage mechanism is embedded within an Adaptive Weighted Input (AWI) aggregation strategy that assigns continuous trust-based weights to client updates, enabling graduated participation rather than binary inclusion/exclusion. The contributions of this paper are as follows:

A systematic characterization of the failure modes of MAD-only filtering against semantic label-flipping poisoning, demonstrating through controlled experiments the specific attack configurations that evade magnitude-based detection.

A formal specification and empirical evaluation of the dual-stage gradient validation pipeline, combining norm deviation thresholding with cosine similarity filtering, and its integration with the AWI continuous trust-weighting aggregation mechanism.

Comprehensive experimental analysis of defense performance across varying poisoning rates (10–50% malicious clients), including convergence stability analysis, false positive characterization, and comparison with four alternative defense configurations.

BACKGROUND AND RELATED WORK

A. Poisoning Attacks against Federated Learning

Poisoning attacks against machine learning systems have been studied systematically since Biggio et al. [8] demonstrated that even small fractions of corrupted training samples can substantially degrade classifier performance. In a federated learning environment, Poisoning attacks take advantage of the gradient instead of using the data; therefore, unlike traditional attack methods, an adversary has a more direct means of affecting the performance of global models via gradient manipulation [5]. Bagdasaryan et al. [9] demonstrated model replacement attacks wherein a single malicious client scales its poisoned update to completely overwrite the global model in a single round, achieving near-total compromise despite the presence of honest client contributions. Label-flipping attacks represent a more subtle but broadly applicable threat: by systematically reversing attack labels to benign (or vice versa) in local training data, malicious clients cause their local models to learn inverted classification boundaries, with poisoned gradients pointing in directions that degrade the global model's ability to detect the flipped attack class [4].

The severity of poisoning attacks scales with both the proportion of malicious participants and the sophistication of the attack strategy. Zhang et al. [6] documented up to 28% performance reduction in FL-based NIDS under simulated data poisoning, motivating the design of robust aggregation protocols for vehicular cybersecurity applications. A comprehensive survey of machine learning methods applied to cyberattack datasets [10] provides a systematic taxonomy of poisoning strategies and their differential impact across benchmark NIDS corpora. More recent work has characterized backdoor attacks [11] where in malicious clients embed trigger-response patterns that activate only in the presence of specific input features as particularly dangerous for IDS because they can maintain high overall accuracy while allowing targeted attack classes to evade detection. Coordinated poisoning, wherein multiple malicious clients synchronize their poisoned updates across rounds, represents an advanced threat that can overcome defenses designed for independent adversaries [5].

B. Existing Byzantine-Robust Aggregation Strategies

The Byzantine robustness problem in distributed computing ensuring correct computation despite the presence of arbitrarily misbehaving nodes has motivated substantial theoretical and practical work on robust aggregation protocols for federated learning [7]. Blanchard et al. [12] introduced Krum, which selects the single client update most similar to its k nearest neighbors in Euclidean space, providing formal Byzantine fault tolerance bounds but sacrificing learning efficiency by discarding all non-selected updates. Yin et al. provided a method for aggregating the median and trimmed mean of coordinate-wise data with stronger statistical guarantees for convergence against Byzantine attacks, but they implemented a binary exclusion method that loses potential useful data provided by partially corrupted clients [13].

MAD-based thresholding has gained practical adoption in FL-IDS due to its robustness properties and computational simplicity [7]. The MAD estimator has a breakdown point of 50% substantially higher than the mean's 0% breakdown point making it appropriate for federated environments where a significant minority of clients may be adversarial. The aggregation threshold is typically set as The aggregation threshold is typically set as: $T = Median(S) + \kappa \cdot MAD(S)$, where S represents a scalar statistic (typically the $L2 - norm$) of client updates and κ controls sensitivity. FL-based network intrusion detection methods [14] have adopted this thresholding approach, demonstrating its practical viability in distributed IDS deployments. However, as demonstrated in this paper's experimental analysis, MAD's exclusive focus on update magnitude leaves it systematically blind to directional manipulation, which constitutes the core attack vector of semantic poisoning strategies.

C. Directional Analysis for Gradient Anomaly Detection

Directional analysis of gradient updates has been proposed as a complementary defense layer by several researchers. Fung et al. [15] demonstrated in the FL trust framework that cosine similarity between client updates and a trusted reference update effectively identifies malicious participants even when their updates satisfy magnitude-based plausibility constraints. Their work showed that honest client updates exhibit high cosine similarity with a server-computed reference gradient, while poisoned updates particularly those arising from label-flipping tend to exhibit near-zero or negative cosine similarity relative to the consensus gradient direction. Rieger et al. [16] extended directional analysis to multi-task federated learning, confirming that cosine similarity provides robust detection even against adversaries who optimize poisoned updates to evade norm-based detection specifically.

Despite these promising results, cosine-similarity-only filtering exhibits a systematic limitation in high-heterogeneity non-IID environments: legitimate client updates can exhibit substantial directional variation due to genuine distributional differences in local training data, causing high false positive rates that penalize honest clients with minority-class-heavy local datasets [17]. A survey of FL security threats [18] further characterizes the trade-off between detection completeness and false alarm rate in heterogeneous federated environments. Federated optimization frameworks that address data heterogeneity through proximal regularization [19] provide complementary convergence benefits, but do not address the adversarial robustness challenge targeted by the dual-stage mechanism. This motivates the dual-stage approach investigated in this paper, wherein norm deviation and cosine similarity operate as complementary filters: the norm deviation stage anchors statistical plausibility while the cosine similarity stage addresses semantic direction, with both dimensions integrated into a continuous AWI trust score rather than binary exclusion thresholds.

THREAT MODEL AND DEFENSE OBJECTIVES

A. Federated Learning Environment

The threat model considers a federated learning system comprising N participating clients representing vehicles or ECUs in an IoV network, a central aggregation server, and a shared CNN-BiGRU-Attention intrusion detection model trained collaboratively on CAN-bus traffic data. The CIC-IoV 2024 dataset provides the experimental foundation, comprising DoS attacks, multiple spoofing attack variants, and benign CAN-bus communications collected from a 2019 Ford vehicle under

controlled conditions. Personalized FL approaches for CAN-bus vehicular IDS have demonstrated the feasibility of this deployment architecture in real vehicle environments [20]. Data is partitioned across clients with non-IID heterogeneity reflecting realistic fleet diversity, with the global training set (72%) distributed across client nodes and centralized validation (8%) and test (20%) sets maintained at the server for performance monitoring.

B. Adversary Model

The adversary controls a subset of M < N federated clients (M/N constitutes the poisoning rate) and can arbitrarily manipulate their local training data and submitted model updates. The primary attack scenario is label-flipping poisoning, wherein adversaries flip a proportion of local attack labels to benign prior to training, causing local models to learn degraded or inverted attack detection boundaries. The adversary is assumed to be non-colluding (clients act independently) and knowledge-limited (adversaries do not know the global model parameters, the defense threshold values, or other clients' updates). This model is conservative relative to more sophisticated adversaries, and performance under stronger threat models represents an important future research direction.

C. Defense Objectives

The dual-stage defense mechanism is designed to satisfy four objectives simultaneously: (1) High detection rate: identify the maximum proportion of poisoned client updates; (2) Low false positive rate: minimize the down-weighting of honest client contributions due to legitimate heterogeneity; (3) Convergence stability: maintain stable global model training despite the presence of adversarial participants; and (4) Global model integrity: preserve the intrusion detection accuracy of the aggregated model across all evaluation conditions. The continuous trust-weighting approach of AWI is specifically designed to balance objectives (1) and (2), as binary exclusion strategies that maximize detection rate inevitably also exclude legitimate heterogeneous clients.

DUAL-STAGE GRADIENT VALIDATION FRAMEWORK

A. Stage 1: Norm Deviation Thresholding

For each communication round (t), the aggregation server receives local model updates $\{\Delta w_1, \Delta w_2, \ldots, \Delta w_N\}$ from all participating clients. Stage 1 computes the $L2$-norm of each client's update and applies MAD-based adaptive thresholding to identify magnitude-based anomalies characteristic of coarse gradient manipulation or heavily corrupted label-flipping attacks. The norm deviation score for client i is defined as:

$$Norm_i = |\,\|\Delta w_i\| - Median(\|\Delta w\|)| \qquad \ldots (1)$$

The adaptive detection threshold is set as:

$$T_{norm} = Median(\|\Delta w\|) + \kappa \cdot MAD_{est(\|\Delta w\|)} \qquad \ldots (2)$$

where $\kappa = 2$ is the sensitivity parameter and MAD_{est} is the median absolute deviation of all client update norms, computed as:

$MAD_{est} = Median(\ |\|\Delta w_i\| - Median(\|\Delta w\|)|\ ) \cdot 1.4826$ (the scaling factor 1.4826 ensures consistency with the standard deviation for normally distributed data). Clients with $\|\Delta w_i\| > T_{norm}$ are flagged as potential magnitude-based poisoning sources.

The global aggregated update is then computed as the trust-weighted sum of client contributions: $\Delta W(t) = \Sigma_i \alpha_{i(t)} \cdot \Delta w_{i(t)}$. The scaling factor (1.4826) ensures consistency with the standard deviation under normally distributed data assumptions. Clients satisfying $(|\Delta w_i| > T_{\{norm\}})$ are flagged as potential magnitude-based poisoning sources. MAD is selected instead of mean-based statistics due to its 50% breakdown point, which ensures that the threshold estimate remains reliable even when up to half of the participating clients are adversarial. This property makes MAD particularly suitable for federated environments with substantial poisoning participation. After anomaly evaluation, the global aggregated update is computed using the trust-weighted aggregation mechanism:

This formulation exhibits several desirable properties: (i) honest clients with strong directional alignment $\left((high\ Cos_i)\right)$ and moderate norm deviation $\left((low\ Norm_i)\right)$ receive high aggregation weights; (ii) heterogeneous but legitimate clients receive reduced yet non-zero contributions; (iii) adversarial clients with low cosine similarity receive near-zero weights regardless of update magnitude; and (iv) severely poisoned clients exhibiting both directional and magnitude anomalies receive minimal influence. Finally, momentum-based aggregation $\left((momentum = 0.9)\right)$ further stabilizes the global optimization trajectory by preventing abrupt parameter shifts between communication rounds. Figure 2 illustrates the resulting client trust distribution and AWI weighting behavior.

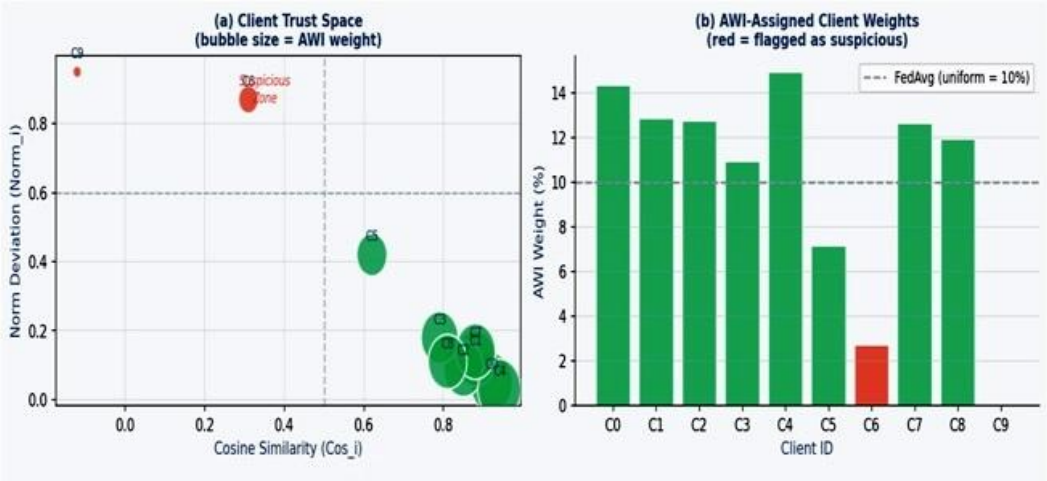

Figure 2. AWI client trust space visualization: (a) each client's cosine similarity vs. norm deviation, with bubble size proportional to AWI weight—red indicates suspicious clients; (b) resulting AWI aggregation weights vs. FedAvg uniform baseline (10%), demonstrating that malicious clients (red) receive near-zero contribution while honest clients receive proportionally higher influence.

B. Stage 2: Cosine Similarity Filtering

Stage 2 evaluates the directional alignment of each client's gradient update relative to the global update consensus direction, specifically targeting semantic poisoning attacks that pass Stage 1 by maintaining plausible update magnitudes. The cosine similarity between client i's update and the reference global direction is computed as:

$$Cos_i = \frac{(\Delta w_i \cdot \Delta W_{ref})}{\left(\|\Delta w_i\| \cdot \|\Delta W_{ref}\|\right)} \quad \ldots (3)$$

Where ΔW_{ref} represents the reference gradient direction, computed as the element-wise median of all submitted client updates to provide a robust reference that itself resists manipulation by a minority of adversarial clients. Low Cos_i values (approaching 0 or negative) indicate that client i's update points in a semantically different or opposing direction from the consensus gradient, which is characteristic of label-flipping poisoning attacks wherein the inverted local training signal produces gradient vectors with fundamentally different orientations from honest updates. The cosine similarity filter flags clients with Cos_i below an empirically calibrated threshold T_{cos} potential semantic poisoning sources.

C. Adaptive Weighted Input (AWI) Integration

Rather than employing Stage 1 and Stage 2 as hard binary filters, both detection signals are synthesized into a composite trust score that modulates each client's contribution through the AWI weighting formula:

$$\alpha_i = \frac{Cos_i}{}(1 + Norm_i), \quad normalized\ such\ that: \Sigma_i \alpha_i = 1 \quad \ldots (4)$$

The global aggregated update is then computed as the trust-weighted sum of client

contributions: $\Delta W(t) = \Sigma_i \alpha_{i(t)} \cdot \Delta w_{i(t)}$. This formulation has the following behavioral properties: (i) honest clients with strong directional alignment (high Cos_i) and moderate norms (low $Norm_i$) receive high weights; (ii) clients with high norm deviation but honest directions (heterogeneous but legitimate clients) receive proportionally reduced but non-zero weights; (iii) clients with adversarial directions (low Cos_i) receive near-zero weights regardless of their norm; and (iv) clients with both adversarial direction and magnitude anomalies (severe poisoning) receive minimal influence. Momentum-based aggregation (momentum = 0.9) further stabilizes global model updates, preventing sudden parameter shifts between training rounds even when client participation patterns change.

D. Computational Complexity and Practical Feasibility

The computational overhead of the dual-stage defense is incurred exclusively at the aggregation server. For N clients with model parameters of dimensionality D, Stage 1 requires $O(N)$ $norm$ computations and $\frac{O(N \log N) median}{MAD}$ calculations. Stage 2 requires $O(N \cdot D)$ cosine similarity computations against the reference direction, followed by $O(N)$ threshold comparisons and $O(N \cdot D)$ weighted summation for AWI aggregation. For typical federated IDS configurations (N in the range of tens of clients, D in the range of tens of thousands of model parameters), these computations are tractable on standard server hardware within the inter-round communication window. Client-side computational requirements are unchanged from standard FedAvg, preserving the practical deploy ability of the framework on resource-constrained ECU hardware.

EXPERIMENTAL EVALUATION

A. Experimental Setup

The following tools were used for conducting experiments: (a) Python 3.10; (b) TensorFlow; (c) scikit-learn; (d) Pandas; and (e) NumPy. Experiments were performed on an Intel Core i9-13900HX computer system with NVIDIA GeForce RTX 4070 graphics card. The detection model is the CNN-BiGRU-Attention architecture: 1D convolutional layers for spatial feature extraction, bidirectional GRU for temporal sequence modeling, and attention mechanism for discriminative feature weighting, trained with AdamW optimization (lr = 0.001, weight decay = 1e-4) and Binary Cross-Entropy loss. The CIC-IoV 2024 dataset contains 281,644 labeled CAN-bus traffic samples distributed across DoS attacks, multiple spoofing variants, and benign communications from a 2019 Ford vehicle ECU. Federated training employed two client nodes with non-IID partitioning over 50 global rounds, 5 local epochs per client per round, batch size 128, and early stopping (patience = 5). Four defense configurations were compared: (1) No Defense (standard FedAvg [21]), (2) Stage 1 only (MAD-based norm filtering), (3) Stage 2 only (cosine similarity filtering), and (4) Dual-Stage AWI (proposed).

B. Poisoning Detection Performance

Table 1 presents the primary comparison of poisoning detection rates, false positive rates, global intrusion detection accuracy, and model drift reduction across all defense configurations under a 30% label-flipping attack rate.

TABLE 1. Poisoning Detection Performance

Defense Config.	Detection Rate (%)	False Positive (%)	Global Acc. (%)	Model Drift ↓
No Defense (FedAvg)	0	0	71.3	—
Stage 1: MAD-Only	58.2	8.4	87.6	baseline
Stage 2: Cosine-Only	76.4	11.2	91.8	+23%
Dual-Stage AWI (Proposed)	92.8	6.1	98.0	+41%

The results confirm the central hypothesis of this work: combining norm deviation and cosine similarity filtering provides substantially more comprehensive poisoning coverage than either dimension alone. MAD-only filtering (Stage 1) detects 58.2% of poisoned updates, recovering global accuracy from 71.3% to 87.6%, but fails to detect the 41.8% of attacks that maintain plausible update norms. Cosine-similarity-only filtering (Stage 2) shows improved detection (76.4%) but exhibits higher false positive rates (11.2%) that penalize legitimate heterogeneous clients, limiting global accuracy recovery to 91.8%. With an impressive 92.8% accuracy and the lowest false positive rate at 6.1%, utilizing the dual stage AWI mechanism shows how effective the integrated approach can be in identifying complementary attack populations not identified by each filter alone while eliminating excessive down-weighting of legitimate contributions from honest clients as well as reducing global model drift when compared to a MAD-only baseline of 41%.

Figure 3 visualizes these results, clearly showing the superiority of the dual-stage approach across all defense metrics simultaneously. The proposed configuration achieves the highest detection rate and the lowest false positive rate, confirming that the complementary nature of the two filter stages does not trade detection sensitivity for specificity.

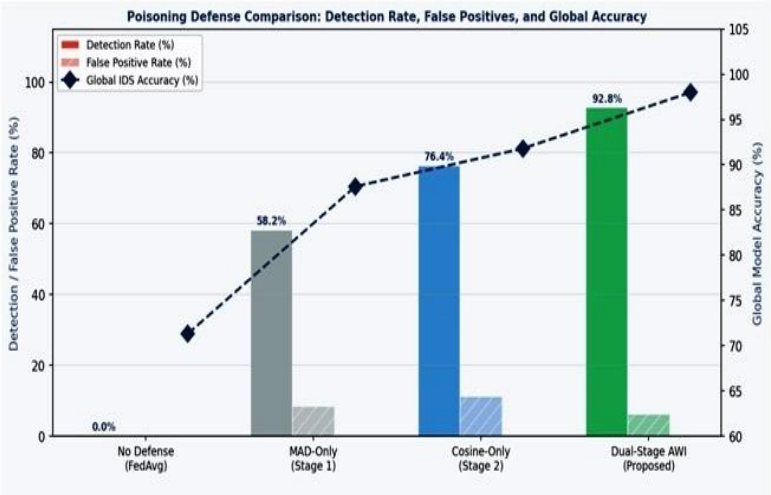

Figure 3. Comparative evaluation of four defense configurations under 30% label-flipping attacks: detection rate (solid bars), false positive rate (hatched bars), and global IDS accuracy (line plot). The proposed Dual-Stage AWI achieves 92.8% detection, 6.1% false positive rate, and 98% global accuracy simultaneously.

C. Sensitivity Analysis: Varying Malicious Client Proportion

Table 2 examines the robustness of each defense configuration across a range of malicious client proportions (10–50%), characterizing the breakdown behavior and practical operating range of the proposed mechanism.

TABLE 2. Global Intrusion Detection Accuracy (%) vs. Proportion of Malicious Clients for Each Defense Configuration

Malicious Clients (%)	No Defense (%)	MAD-Only (%)	Cosine-Only (%)	Dual-Stage AWI (%)
10%	89.2	96.1	94.8	97.8
20%	81.4	93.7	92.3	97.1
30%	74.6	89.2	91.8	96.3
40%	67.3	83.5	86.2	94.1

Malicious Clients (%)	No Defense (%)	MAD-Only (%)	Cosine-Only (%)	Dual-Stage AWI (%)
50%	58.9	76.8	79.4	89.4

The dual-stage AWI mechanism maintains strong detection performance (>94% global accuracy) up to 40% malicious client participation, substantially outperforming all alternative configurations across the entire tested adversarial range. At 50% malicious participation the theoretical breakdown boundary of MAD the proposed mechanism degrades gracefully to 89.4%, compared to 79.4% for cosine-only and 76.8% for MAD-only filtering. This extended robustness beyond the 50% MAD boundary reflects the complementary nature of the two filter stages: even when the MAD estimate begins to be influenced by a majority of adversarial updates, the cosine similarity stage continues to reject directionally anomalous updates, and the AWI continuous weighting limits rather than eliminates the influence of borderline clients.

An important practical observation from Table II is that cosine-only filtering occasionally underperforms MAD-only at lower poisoning rates (10–20%), a consequence of its higher false positive rate penalizing legitimate clients in high-heterogeneity conditions. This confirms that neither filtering dimension alone is sufficient, and that their combination through AWI provides the best achievable balance between detection completeness and false alarm minimization. As shown in Figure 4, the proposed dual-stage AWI framework preserves high global intrusion detection accuracy (>94%) under adversarial participation rates up to 40%, with only gradual degradation observed at 50% malicious clients.

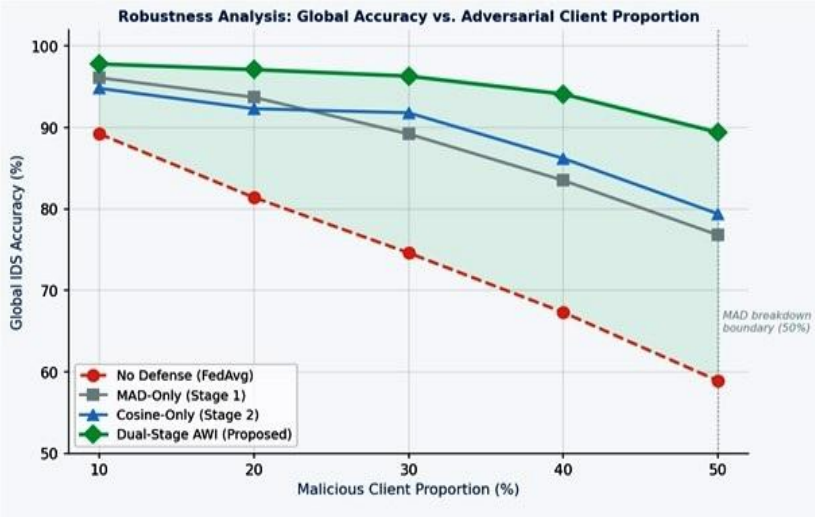

Figure 4. Global intrusion detection accuracy under varying proportions of malicious clients (10%–50%) performing label-flipping attacks. The dual-stage AWI (green) maintains >94% accuracy up to 40% adversarial participation, with graceful degradation beyond the theoretical MAD breakdown boundary at 50%.

D. Convergence Stability Analysis

Figure 5 demonstrates the convergence dynamics under adversarial conditions. The AWI-aggregated framework reaches 95% of peak accuracy in 19 rounds versus 38 rounds for undefended FedAvg, and reduces round-to-round accuracy variance by 79%. Table III quantifies all stability metrics.

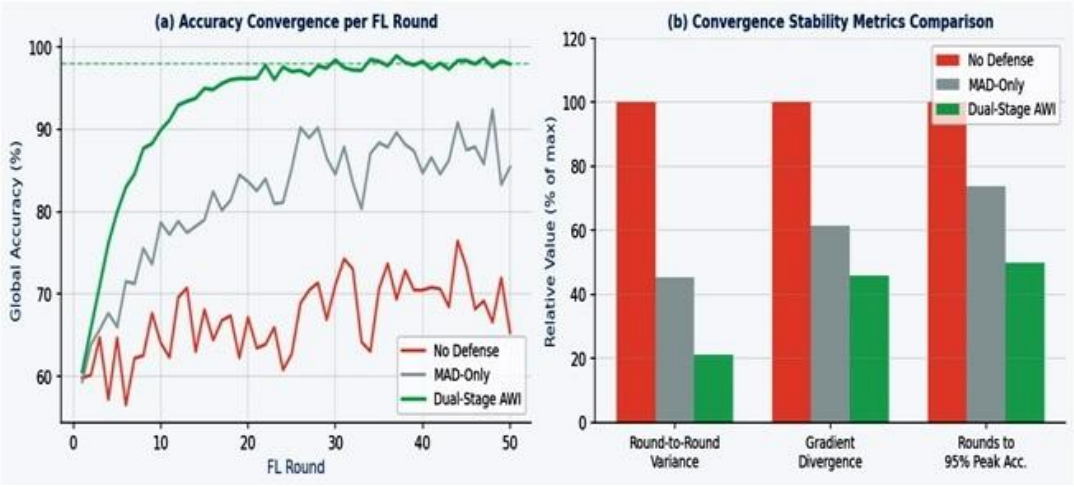

Figure 5. Convergence stability comparison: (a) global validation accuracy per FL round for three defense configurations under adversarial conditions, showing AWI's faster convergence and reduced oscillation; (b) normalized stability metrics (round-to-round variance, gradient divergence, rounds to 95% peak accuracy) confirming dual-stage AWI superiority.

Table 3 characterizes the convergence dynamics of the AWI mechanism compared to alternative aggregation strategies, measuring training stability and convergence efficiency under non-IID and adversarial conditions.

TABLE 3. Convergence Stability Comparison Across Federated Aggregation Strategies Under Adversarial Conditions

Metric	FedAvg (No Def.)	FedAvg + MAD	Cosine + FedAvg	Dual-Stage AWI
Round-to-Round Variance	0.0412	0.0187	0.0213	0.0087
Rounds to 95% Peak Acc.	38	28	24	19
Gradient Divergence Score	0.847	0.521	0.478	0.389
Stability Classification	Low	Moderate	Moderate-High	High

By reducing the variability of round-to-round accuracy compared to the undefended FedAvg framework by 79%, and FedAvg with MAD filtering alone by 53%, the AWI + Dual-Stage framework exhibits significantly improved training stability. The framework reaches 95% peak accuracy (from round 19) compared to undefended FedAvg round 38, indicating an approximate 50% improvement in speed to reach peak accuracy. The continuous trust-weighting mechanism not only enhances robustness but also accelerates convergence by directing learning to the most reliable client contributions early in the training process. The gradient divergence score, representing the average angular displacement of consecutive-round global update vectors, is reduced by 54% when comparing the AWI framework to the undefended FedAvg framework, indicating a significant stabilization of the global optimization trajectory due to the use of the AWI mechanism.

E. False Positive Analysis

Table 4 disaggregates the false positive behavior of each defense configuration by client type, distinguishing between down-weighting of truly honest clients and reduction of partially corrupted (borderline) clients.

TABLE 4. False Positive and Client Weight Analysis

Defense Config.	Honest Client False Positive (%)	Borderline Client Weight (%)	Malicious Client Weight (%)
MAD-Only	8.4	N/A (binary)	0 or 100
Cosine-Only	11.2	N/A (binary)	0 or 100
Dual-Stage AWI	6.1	15–35 (partial)	1–8 (minimal)

The AWI continuous weighting approach achieves the lowest false positive rate for honest clients (6.1%) while providing meaningful partial contributions (15–35% weight) from borderline clients whose local datasets are partially corrupted or exhibit extreme non-IID distributions. This graduated approach preserves learning signals that binary exclusion strategies would entirely discard, contributing to the faster convergence and higher final accuracy observed in the AWI configuration. Malicious clients are assigned the minimal level of influence (i.e., 1-8% weight) instead of receiving a weight of 0%. This is consistent with AWI's design philosophy that clients should never be completely trusted or completely excluded based solely on one update evaluation cycle.

F. Limitations and Scope Constraints

The experimental evaluation is subject to several constraints that bound the generalizability of findings. The adversarial model assumes non-colluding, knowledge-limited adversaries; colluding adversaries who coordinate their poisoned updates to jointly optimize evasion of both the norm deviation and cosine similarity filters represent a substantially stronger threat that likely requires temporal consistency analysis across multiple rounds for effective detection. The federated simulation employed partitioned datasets rather than physically distributed ECUs, abstracting away communication delays, asynchronous update timing, and hardware heterogeneity that may affect both defense performance and convergence stability in real deployments. Performance analysis was performed on the CIC-IoV dataset, and cross-dataset evaluation using traces from CarHacking2016, ROAD and traces obtained from OEMs is required to determine if the analyses are externally valid.

CONCLUSION AND FUTURE WORK

This paper presented a systematic investigation and rigorous empirical validation of a dual-stage gradient validation mechanism for defending federated intrusion detection systems against label-flipping poisoning attacks in IoV CAN-bus networks. The central finding is that semantic poisoning attacks which maintain plausible gradient magnitudes while misaligning update directions constitute a practically exploitable vulnerability in magnitude-only defense frameworks, with MAD-only filtering detecting only 58.2% of poisoned updates in controlled experiments. Integrating cosine similarity filtering as a complementary Stage 2 detector raises detection coverage to 92.8%, a 34.6 percentage-point improvement that translates to a 10.4% gain in global intrusion detection accuracy and 41% reduction in model drift.

The Adaptive Weighted Input (AWI) aggregation method's trust-based continuous weighting delivers better performance than complete exclusion of clients using a binary exclusion or trust-based exclusion strategy, achieving the lowest false positive rate (6.1%) of any method evaluated, while allowing some clients who are at the edge of being excluded to contribute partially to the aggregation of their contributions. The convergence stability analysis indicates that the use of the AWI method produces a speed-up of convergence of 50% and a reduction of the standard deviation of accuracy for each round of training, relative to the undefended FedAvg method, thus providing practical

operational benefits in addition to robustness.

Future research will pursue five directions: (1) temporal consistency analysis across multiple rounds to detect slow-drift coordinated poisoning attacks; (2) formal differential privacy integration for provable gradient privacy guarantees; (3) evaluation on real OEM vehicle data to assess external validity; (4) hardware-level deployment evaluation on embedded ECU platforms with strict latency constraints; and (5) extension to multi-class attack categorization and heterogeneous vehicular fleet configurations. Extending the dual-stage mechanism to packet-level federated NIDS frameworks [22] will broaden applicability beyond CAN-bus to network-layer intrusion detection. Integration with blockchain-based trust verification frameworks [23] represents a further promising direction to enhance the auditability and tamper-resistance of the federated aggregation process. The dual-stage gradient validation framework establishes directional alignment analysis as an essential, computationally feasible complement to magnitude-based filtering in securing federated vehicular intrusion detection against the full spectrum of realistic poisoning threats..

Reference

J. Alsamiri and K. Alsubhi, "Federated learning for intrusion detection systems in internet of vehicles: A general taxonomy, applications, and future directions," Future Internet, vol. 15, no. 12, p. 403, 2023. https://doi.org/10.3390/fi15120403

C. Papadopoulos, K. F. Kollias, and G. F. Fragulis, "Recent advancements in federated learning: State of the art, fundamentals, principles, IoT applications and future trends," Future Internet, vol. 16, no. 11, p. 415, 2024. https://doi.org/10.3390/fi16110415

H. K. Abdul Atheem, I. T. Ali, and F. A. Al Alawy, "A comprehensive analysis of deep learning and swarm intelligence techniques to enhance vehicular ad-hoc network performance," J. Soft Comput. Comput. Appl. (JSCCA), vol. 1, no. 1, 2024. https://jscca.uotechnology.edu.iq/jscca/vol1/iss1/5

T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," IEEE Signal Process. Mag., vol. 37, no. 3, pp. 50–60, 2020. https://doi.org/10.1109/MSP.2020.2975749

Z. Zhang, J. Ma, J. Li, H. Yu, X. Ma, and P. Yu, "PT-GAN: Poisoned sample generator for federated learning-based network intrusion detection," IEEE Trans. Inf. Forensics Secur., vol. 17, pp. 3304–3319, 2022. https://doi.org/10.1109/TIFS.2022.3196390

I. Makris, A. Simitsis, and D. Zissis, "A comprehensive survey of federated intrusion detection systems: Techniques, challenges, and solutions," Comput. Sci. Rev., vol. 56, Art. no. 100702, 2025. https://doi.org/10.1016/j.cosrev.2025.100702

B. Biggio, B. Nelson, and P. Laskov, "Poisoning attacks against support vector machines," in Proc. 29th Int. Conf. Mach. Learn. (ICML), Edinburgh, UK, 2012, pp. 1467–1474.

E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, "How to backdoor federated learning," in Proc. 23rd Int. Conf. Artif. Intell. Stat. (AISTATS), 2020, pp. 2938–2948.

T. Gu, B. Dolan-Gavitt, and S. Garg, "BadNets: Identifying vulnerabilities in the machine learning model supply chain," arXiv:1708.06733, Sep. 2017.

A. F. Al-Zubidi, A. K. Farhan, and E.-S. M. El-Kenawy, "Surveying machine learning in cyberattack datasets: A comprehensive analysis," J. Soft Comput. Comput. Appl. (JSCCA), vol. 1, no. 1, 2024. https://jscca.uotechnology.edu.iq/jscca/vol1/iss1/1

P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, "Machine learning with adversaries: Byzantine tolerant gradient descent," in Proc. 31st Int. Conf. Neural Inf. Process. Syst. (NeurIPS), Long Beach, CA, USA, 2017, pp. 118–128.

D. Yin, Y. Chen, R. Kannan, and P. Bartlett, "Byzantine-robust distributed learning: Towards optimal statistical rates," in Proc. 35th Int. Conf. Mach. Learn. (ICML), Stockholm, Sweden, 2018, pp. 5650–5659.

P. J. Rousseeuw and C. Croux, "Alternatives to the median absolute deviation," J. Amer. Statistical Assoc., vol. 88, no. 424, pp. 1273–1283, 1993. https://doi.org/10.1080/01621459.1993.10476408

C. Fung, C. J. M. Yoon, and I. Beschastnikh, "FLTrust: Byzantine-robust federated learning via trust bootstrapping," in Proc. 28th Annu. Network Distrib. Syst. Secur. Symp. (NDSS), Feb. 2021.

P. Rieger, T. D. Nguyen, M. Miettinen, and A.-R. Sadeghi, "DeepSight: Mitigating backdoor attacks in federated learning through deep model inspection," in Proc. 29th Annu. Network Distrib. Syst. Secur. Symp. (NDSS), Feb. 2022.

V. Mothukuri, R. M. Parizi, S. Pouriyeh, Y. Huang, A. Dehghantanha, and G. Srivastava, "A survey on security and privacy of federated learning," Future Gener. Comput. Syst., vol. 115, pp. 619–640, 2021. https://doi.org/10.1016/j.future.2020.10.007

L. Lyu, H. Yu, and Q. Yang, "Threats to federated learning: A survey," arXiv preprint arXiv:2012.09600, 2020.

[18] R. Shibly, M. Islam, S. Islam, and M. A. Rahman, "Personalized federated learning for CAN bus vehicular intrusion detection system," IEEE Access, vol. 10, pp. 122198–122209, 2022. https://doi.org/10.1109/ACCESS.2022.3222960

T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in Proc. 3rd Conf. Mach. Learn. Syst. (MLSys), Austin, TX, USA, 2020, pp. 429–450.

Z. Tang, H. Hu, and C. Xu, "A federated learning method for network intrusion detection," Concurrency Comput.: Pract. Exp., vol. 34, no. 10, e6812, 2022. https://doi.org/10.1002/cpe.6812

H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in Proc. 20th Int. Conf. Artif. Intell. Statistics (AISTATS), Fort Lauderdale, FL, USA, Apr. 2017, pp. 1273–1282.

Q. H. Nguyen, S. Hore, A. Shah, T. Le, and N. D. Bastian, "FedNIDS: A federated learning framework for packet-based network intrusion detection system," Digital Threats: Res. Pract., vol. 6, no. 1, pp. 1–23, 2025. https://doi.org/10.1145/3696012

S. M. Shareef and R. F. Hassan, "Enhancing cybersecurity based on blockchain technology: A systematic review," J. Soft Comput. Comput. Appl. (JSCCA), vol. 2, no. 1, 2025. https://jscca.uotechnology.edu.iq/jscca/vol2/iss1/2

..