



ARTICLE

Resilient Federated Learning Framework for V2X Accident Mitigation against Sybil and Model Poisoning Threats

Vishvadharm Saminathan, Yuvaraj Kurup, Fawaz Ahamed, Samarth Maligouder, Jai Vinita L.* and Helen Vijitha P.

School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, Tamil Nadu, India

*Corresponding Author: Jai Vinita L. Email: jaivinita.l@vit.ac.in

Received: 09 June 2026; Accepted: 19 August 2026; Published: 21 October 2026

ABSTRACT: Federated learning (FL) enables vehicles to train a shared accident-detection model without transmitting raw trajectories, but coordinated Sybil identities and poisoned local updates can distort conventional Federated Averaging (FedAvg). This study evaluates a three-stage robust aggregation mechanism for synthetic V2X accident-event detection. Five SUMO simulations on an OpenStreetMap-derived road network generate 500 controlled accident-like events with explicit vehicle, time, speed, edge, and lane ground truth. Non-overlapping trajectory windows are curated into 10,000 samples and separated by simulation scenario into training (6000), validation (2000), and test (2000) sets. For each method, the standard architecture consists of an 11-16-1 multilayer perceptron, 100 FL clients, 80 communication rounds, 20% malicious activity, and validation-based selection of threshold values. In the face of combined label flip and update amplification attacks, the proposed defense strategy sustains an accuracy of 96.39% (SD 0.29), 82.24% F1 (SD 1.00), 90.84% balanced accuracy (SD 0.61), and an MCC of 0.8028 (SD 0.0111). Its attacked performance remains close to its clean performance (96.73% accuracy and 83.84% F1), whereas attacked FedAvg and FoolsGold exhibit substantially greater variability and degradation. Sensitivity experiments cover amplification factors 5–20, malicious ratios 10%–40%, and 50–500 clients. The claims are limited to controlled synthetic-event detection on one road network. A permissioned ledger is described only as an optional audit architecture and is not part of the evaluated defence.

KEYWORDS: Internet of vehicles; V2X communications; federated learning; accident detection; Sybil attack; model poisoning

1 Introduction

Every year, around 1.35 million people are killed as a result of road traffic accidents. They are also one of the leading causes of death for those aged 5 to 29 globally. Connected and automated vehicle technologies have great potential to improve safety and reduce risk by allowing for shared perception and real-time cooperative decision-making.

1.1 Background and Motivation

The Internet of Vehicles (IoV) consists of Vehicle-to-Everything (V2X) communication that allows the connection between vehicles, pedestrians, the infrastructure, and cloud services. This provides a foundation for intelligent transportation systems by providing support for semantic communications and advanced security protocols [1,2]. There are effective machine learning methods that can be used to detect anomalous

driving behaviours predictive of accidents. However, traditional centralized learning methods have very high communication costs, substantial data privacy concerns, and dependence on a central server. Federated Learning (FL) addresses the challenges of privacy and communication by making it possible to collaborate through the sharing of models while maintaining the original data in the local area. In this architecture, the RSU Aggregator continues to play an essential role in the coordination process. However, due to its openness and distributed nature, V2X networks are exposed to hostile attacks.

1.2 Problem Statement

The reliability of FL in V2X environments is compromised by two critical attack types:

- **Model Poisoning Attacks:** When using local training, the adversarial cars adopt a method referred to as “Model Poisoning Attacks” to inject harmful gradients or backdoors into the model. Apart from possibly decreasing the performance level of the global model, it is possible that this will lead to the misclassification of some objects, such as important traffic signs [3].
- **Sybil Attacks:** While in the course of aggregation, the attacker increases their influence through the creation of multiple fake identities. Moreover, an increase in the number of fake participants may overshadow the contribution of genuine participants in techniques such as Federated Average.

The exclusion of outliers from the dataset forms the basis of popular robust aggregation algorithms such as Krum and Median algorithms. These are usually not very accurate, or in worse cases will produce false positive errors that could otherwise punish any genuine participants contributing to the aggregation in case of non-independent and identically distributed (non-IID) vehicle data sources.

1.3 Contributions

Instead of proposing a completely new aggregation family, this work assesses a V2X-oriented adaptation of multi-stage robust aggregation. The contributions consist of:

- 10,000 scenario-separated windows for trajectories from five SUMO simulations and 500 manually injected events are included in this synthetic V2X accident event dataset.
- Three-stage aggregation implementation with stability-controlled global steps, history similarity suppression, and median deviation robustness.
- A leakage-proof evaluation procedure that employs the same model and threshold process for all the techniques and also distinguishes between training, validation, and testing before setting the threshold.
- The attack strength, malicious ratio, and client scalability analysis of the proposed technique are included in the five seed analyses between clean and attacked FedAvg, FoolsGold, and the proposed modification.

Hashes and aggregate metadata are conceptually recorded in the optional ledger layer; this study does not develop or benchmark a consensus network or smart contract.

1.4 Organization of the Paper

The rest of this paper is structured as follows. [Section 2](#) reviews related work on V2X accident detection, federated learning in vehicular networks, and defences against Sybil and poisoning attacks. [Section 3](#) presents the system model, including the V2X architecture and federated learning formulation. [Section 3.5](#) details the adversarial threat model. [Section 4](#) presents the evaluated three-stage aggregation and the optional audit-layer concept. [Section 5](#) describes the SUMO-OSM mobility simulation, feature engineering, experimental methodology, simulation parameters, and metrics. [Section 6](#) analyzes the results, and [Section 7](#) concludes with future research directions.

2 Related Work

2.1 V2X Security and Sybil Attacks

Vehicle-to-Everything (V2X) communication allows data exchange between vehicles, roadside infrastructure, pedestrians, and first responders, thus enhancing road safety and traffic management. The dynamic nature and open architecture of V2X networks make it challenging to ensure the reliability and security of information exchanged by these networks, especially when there is not much infrastructure available in central locations.

Vehicle-to-vehicle (V2V) safety exchange, vehicle-to-infrastructure (V2I) communication via RSUs and traffic systems, vehicle-to-pedestrian (V2P) warning messages, and vehicle-to-network (V2N) communication for more generic services are all considered as parts of the term “V2X.” The vehicles in the proposed FL scheme exchange only the models with the RSU aggregator and store their trajectory windows locally; communication means and cryptographic authentication are seen as architectural assumptions that do not belong to the scope of the quantitative analysis.

One of the major attacks that pose a danger to the security and functioning of the V2X network is a Sybil attack [3]. Such an attack involves a user who creates several false identities in order to gain access beyond what they deserve to influence the operations of the network. The manipulation of the V2X network might result in the distribution of incorrect traffic information, reporting of incorrect accidents, improper routing, and interfering with consensus mechanisms [1]. In critical scenarios where collisions need to be prevented, and emergency messages have to be sent, Sybil attacks might cause traffic jams and failures to respond to emergencies.

The first approaches towards defending against Sybil attacks in vehicular networks have been the cryptographic and identity-based mechanisms which involve Public Key Infrastructures (PKIs) and certificate authentication schemes [4]. In all these methods, a vehicle is given an identity that is unique and verifiable in order to prevent the reuse of identities by vehicles. While the techniques are secure, they usually depend on centralized entities or roadside units that handle the issuance and revocation of certificates. Recently, blockchain technology has been proposed in order to enable trust management that resists Sybil attacks in V2X networks [5] and secure, performance-oriented vehicular edge service delivery [6]. Blockchain technology allows for a decentralized trust management process and an immutable method of recording transactions, which ensure that identity fraud is minimized and accountability is increased. Distributed ledgers may store the identities or reputations of vehicles without requiring any central authority. V2X communication frameworks for off-road autonomous vehicles [7] have also been designed; yet, their reliable deployment in infrastructurally poor remote areas is difficult. Although there have been several advancements, some challenges persist. Identity-based methodologies typically require some sort of infrastructure, and thus compromise user privacy in doing so. Location-based and behavioural detection approaches, while they are less invasive than identity-based methodology, have issues with scalability and reliability in dynamic environments. Both traditional solutions (like digital certificates) and blockchain solutions have high overhead for communication and computation. Moreover, research on lightweight, fully decentralized Sybil defenses tailored for peer-to-peer V2X communication in infrastructure-less emergency scenarios remains limited.

The present work addresses the aggregation aspect of these limitations by evaluating a V2X-oriented three-stage robust aggregation method under coordinated Sybil-amplified poisoning. An optional permissioned audit ledger is described as a future traceability extension, but it is not part of the implemented defence or experimental claims.

2.2 Robust Aggregation in Federated Learning

The distributed training architecture of federated learning systems renders them susceptible to model poisoning attacks. To mitigate malicious client behavior, Byzantine-robust aggregation techniques have been extensively investigated. Initially, traditional aggregation methods such as averaging were replaced with robust statistical estimators, including trimmed mean, coordinate-wise median, and geometric median.

Aggregation algorithms that are secure, specifically, aggregation algorithms using geometric medians, have been demonstrated by Pillutla et al. to enhance robustness against corrupt updates [8]. Nevertheless, even with improvements in robustness against outliers, the effectiveness of geometric median-based algorithms diminishes with a greater departure from IID because the gradients of the benign clients tend to diverge considerably. The concept of γ -divergence-based aggregation, namely, γ -mean has been proposed by Li et al. [9]. This algorithm uses minimal estimation of γ -divergence to reduce the contribution of suspicious client updates. In order to boost robustness, the γ -mean algorithm uses a divergence-based estimator with adaptive weights. However, depending on the degree of departure, the approach might have trouble distinguishing between natural heterogeneity and malicious activity in complicated FL systems.

DeFL is an important learning period-based mechanism proposed by Yan et al. [10] for defending against any fraud client by using gradient norms in an adaptive manner. Although detection mechanisms are effective, they may also be vulnerable to validation and require proper threshold adjustment.

According to [11], there exists an extensive taxonomy on model aggregation methods in federated learning which categorizes them based on four categories of robust, hierarchical, synchronous, and asynchronous paradigms. Robust model aggregation remains one of the challenges, particularly in edge and vehicular learning systems, under adversarial and non-IID environments.

TriGuardFL [12] proposes a three-stage defense mechanism against Byzantine attacks. Different from similarity or divergence-based single-tier defenses, the mechanism filters out the Byzantine updates from benign non-IID updates in an effective way through pre-filtering using deviation, class-aware validation assessment, and Bayesian reputation monitoring. The resilience to different distributions of the vehicle data is improved by the multi-tier defense strategy.

2.3 Federated Learning and Model Poisoning in V2X

Machine learning techniques are becoming increasingly popular for V2X systems, including cooperative perception for autonomous driving [13]. One recent development in this area has come from privacy-preserving federated learning (FL), which is an open-source framework designed to provide an alternative to centralized training methods [14,15]. Through FL, each vehicle operates as an edge node by training a local model based on its own sensor data before sending only the updated model (i.e., model weights or gradients) to either a central server or group of peers.

The updated models will be aggregated to create a global model without sharing raw sensitive data (e.g., an individual vehicle's latitude and longitude), reducing the amount of privacy risk and excess bandwidth usage. Research has found that the use of vehicular federated learning (VeFL) can produce similar accuracy to that achieved through centralized training while significantly decreasing the amount of communications required and keeping drivers' identities private.

Notwithstanding the progress made in transmitting and storing data between vehicles, VeFL in an untrusted vehicular setting is still subject to security concerns, one of which is model poisoning attacks. In these situations, the attacker submits fake model updates that are intended to degrade the performance of the aggregated global model by causing an increased number of misclassifications of critical objects or inaccurate predictions of hazards [16]. Related V2X security research has also used machine learning and

vehicle consensus for false-message detection [17]. If the poisoned updates go undetected, they are aggregated with the other updates for building a single version of the global model, resulting in a negatively impacted global model with the potential to inflict serious harm with respect to safety-critical V2X applications.

In addition, model poisoning becomes more damaging when combined with a Sybil attack. In this setting, coordinated fake clients submit multiple malicious updates, thereby amplifying the adversary's influence during aggregation.

Anomalies in the update or the effect of such anomalies can be minimized or identified by employing existing Byzantine-resilient aggregation techniques such as Multi-Krum [18], trimmed mean, and coordinate-wise median. Although the trimmed mean removes outlying values in each coordinate, Multi-Krum selects updates that are consistent. Nonetheless, when an adversary has control of many Sybils working together, neither technique will work. The security measure known as FoolsGold [19] reduces strongly correlated updates from clients. As the adversaries make alterations to the updates provided to reduce correlation, this approach becomes ineffective.

The development of new resilient aggregation systems has resulted in a three-level defense system that includes influence reweighting, stability-aware update control, and gradient anomaly filtering. Other potential defenses to combat adversarial attacks would be the use of reputation systems, which provide a measure of trust for each participant based on historical performance and their effect on validation accuracy. Trust scores, as measured by reputation systems, can be utilized for the purpose of adapting weights for low-trust updates and/or excluding low-trust updates altogether. Due to the existence of an immutable ledger to record all model updates and trust evaluations, blockchain-integrated FL frameworks could represent a promising alternative for reliable and transparent monitoring of participant reliability because they provide transparency and evidence of tampering with every contribution [5]. Research has shown that having blockchain-enabled reputation systems will allow an FL to continue to maintain good model accuracy even when the number of adversarial participants reaches 40%.

Maintaining the confidentiality of individuals' information while using federated learning continues to be a major concern. Raw data remain stored by local providers and are not transmitted to the aggregator; nevertheless, the model updates sent by participants may still enable model inversion or membership inference attacks. There are techniques, such as differential privacy, where noise is added to the updates to provide a degree of privacy for the participants, to mitigate the risk of someone being able to perform these attacks, but it generally will degrade the accuracy of the end product or impact the overall processing capabilities [15]. In addition, in a V2X system, there will be issues related to location privacy that could arise as training and updating the model may inadvertently expose a participant's location or activity.

Therefore, although the use of federated learning will provide greater capabilities to maintain confidentiality while providing intelligence from a V2X network, individuals will still be at risk from model poisoning and Sybil amplification attacks. Effective mitigation necessitates robust aggregation strategies, trust mechanisms, and privacy protections. Currently, effective aggregation methods in use include detection-based exclusion methods, such as critical learning period (CLP)-aware filtering [10], and statistical deviation filtering, such as geometric median filtering and γ -divergence filtering. However, it should be noted that these methods often cannot address non-IID heterogeneity and model poisoning simultaneously, as is the case in vehicular federated learning settings. This is where the proposed system using the Tri-stage concept comes in, with its multi-stage design well-suited to various V2X scenarios.

Table 1 summarizes the settings, threats, defences, and limitations of representative related work.

Table 1: Comparison with representative related work.

Method	Setting	Threat	Defence	Non-IID Focus	Ledger	Main Limitation
Multi-Krum [18]	General FL	Byzantine updates	Distance selection	Limited	No	Can reject benign heterogeneous updates
Robust aggregation [8]	General FL	Corrupted clients	Geometric median	Partial	No	Higher cost and sensitivity to heterogeneity
γ -mean [9]	General FL	Outliers	Divergence weighting	Partial	No	Requires suitable divergence scale
FoolsGold	General FL	Sybil clients	History similarity	Partial	No	Adaptive/noisy updates can evade similarity
TriGuardFL [12]	General FL	Byzantine clients	Multi-stage screening	Yes	No	Not evaluated for this V2X event pipeline
This work	Synthetic V2X FL	Sybil + label/update poisoning	Median/MAD + history + stability	Yes	Conceptual	One map and synthetic events

3 System Model

3.1 Network Scenario

A remote Internet of Vehicles (IoV) network is considered, deployed in infrastructure-limited or off-road environments such as mining sites, disaster zones, or military operations. The network consists of N vehicles and M Roadside Units (RSUs):

- **Vehicles (Learners):** Each vehicle is equipped with an On-Board Unit (OBU) that enables local sensing, computation, and short-range wireless communication. Vehicles are further equipped with Light Detection and Ranging (LiDAR) sensors, cameras, and a limited battery. Local models are trained on private data within each vehicle.
- **RSUs (Edge Aggregators):** These fixed infrastructure components have higher computational capacity and perform the evaluated robust aggregation and global-model broadcast. A future permissioned deployment could additionally make authenticated RSUs validators of an audit ledger; that role is not implemented here.

Vehicles communicate with nearby RSUs using short-range V2X protocols, such as IEEE 802.11p or Cellular V2X (C-V2X), to periodically exchange model updates. Each vehicle operates as an independent federated learning (FL) client, while RSUs manage global model aggregation, update verification, and model broadcasting. Vehicle mobility introduces simulated traffic dynamics, resulting in non-independent and identically distributed (non-IID) and statistically heterogeneous local datasets across clients. These dynamics are modeled using mobility traces generated from OpenStreetMap (OSM) road networks integrated with the Simulation of Urban Mobility (SUMO) simulator.

3.2 Learning and Computation Model

Each vehicle $v_i \in \mathcal{V}$, where $i = 1, \dots, N$, collects private kinematic data during its travel trajectory to train a local accident-detection model. Its local dataset $\mathcal{D}_i = \{(\mathbf{x}_{ij}, y_{ij})\}_{j=1}^{n_i}$ contains 11 window-level kinematic features and binary labels $y_{ij} \in \{0, 1\}$.

Local training is performed over multiple epochs using the vehicle's onboard computational resources. Let $\mathbf{w}^{(t)}$ be the global parameters at round t and $\mathbf{w}_i^{(t)}$ the locally trained parameters. Client i submits $\Delta \mathbf{w}_i^{(t)} = \mathbf{w}_i^{(t)} - \mathbf{w}^{(t)}$ to the RSU, while raw data remain local.

Federated learning proceeds in synchronous communication rounds. In each round t :

1. Vehicles perform local training on their private data.
2. Local model updates are uploaded to associated RSUs.
3. RSUs aggregate the received updates to compute a global model $\mathbf{w}^{(t+1)}$.
4. The updated global model is broadcast to participating vehicles for the next round.

3.3 Global Optimization Objective

The following is the formulation of the federated learning objective across N vehicles:

$$\min_w F(w) = \sum_{i=1}^N \frac{|D_i|}{|D|} F_i(w) \quad (1)$$

where $|D|$ represents the total number of samples from all the clients. The empirical local loss on client dataset D_i is represented by $F_i(w)$.

In an effort to approximate this objective, the global parameter update for standard Federated Averaging (FedAvg) takes the form of a weighted sum of individual updates of the clients. In case of attacks in adversarial settings, the malicious clients will be able to tamper with the aggregation operator, and thus, the objective will not be achieved. The proposed defense method alters the aggregation process to limit this variation in the event of poisoning and Sybil assaults.

3.4 Communication Considerations

Communication is conducted over short-range V2X links between vehicles and RSUs. While packet reliability and transmission delay are influenced by channel conditions, physical-layer optimizations are not explicitly modeled in this work. Instead, communication impacts such as latency and packet loss are indirectly incorporated through simulation-based delay models that reflect realistic constraints in mobile vehicular environments.

3.5 Threat Model

The focus is on Sybil-based poisoning attacks, where a single adversary controls multiple malicious (Sybil) vehicles to manipulate the global model. Let $\mathcal{V} = \mathcal{H} \cup \mathcal{S}$ denote the complete set of vehicles, where $\mathcal{H}$ represents the set of honest vehicles and $\mathcal{S}$ denotes the set of Sybil vehicles controlled by the adversary.

Malicious clients perform binary label-flip poisoning during local training. For every malicious training example, the label is transformed as

$$\tilde{y}_{ij} = 1 - y_{ij}, \quad i \in \mathcal{S}. \quad (2)$$

After training on the flipped labels, malicious client i scales its submitted parameter update:

$$\widetilde{\Delta \mathbf{w}}_i^{(t)} = a \Delta \mathbf{w}_i^{(t)}, \quad a \geq 1, \quad i \in \mathcal{S}, \quad (3)$$

where a is the update-amplification factor. Honest clients submit unscaled updates. Sybil coordination is modelled by assigning this same attack strategy to multiple malicious client identities. The experiments do not use a trigger-based backdoor objective.

3.6 Design Rationale

Based on the attributes of vehicular data sets and the types of threats being targeted, the selection of multi-stage robust aggregation as opposed to a distance-based approach for outlier detection is suitable for use in vehicular systems. For example, in non-independently and identically distributed (non-IID) vehicular systems, while reasonable vehicles will produce a variety of gradients because of different mobility patterns and environmental conditions, Sybil (malicious) vehicles, which are controlled by an attacker or group of attackers generating the same data, produce similar or identical updates (i.e., are highly correlated).

To strengthen federated learning against coordinated poisoning attacks [12], the defence combines robust deviation weighting, history-based similarity suppression, and stability-controlled aggregation while avoiding hard exclusion based solely on one criterion.

Moreover, an optional future logging mechanism based on blockchain technology could improve confidence and accountability by creating an immutable record of aggregation weights and model updates; this mechanism is not implemented or evaluated in the present work.

4 Proposed Framework

A secure federated learning architecture is introduced for synthetic vehicular accident-event detection. The evaluated contribution is a V2X-oriented implementation of three-stage robust aggregation inspired by prior multi-stage defences [12]; an optional ledger interface is retained only as a future architectural extension.

4.1 Framework Overview

The architecture proposed by this study will work distributively across the Vehicle-to-Everything (V2X), enabling vehicles to collaboratively train an accident detection model using federated learning. Prior IoT studies of vehicle accident detection and emergency notification motivate the use of vehicle sensing data [20]. Each vehicle will locally train on its private kinematic data and send only model updates to the nearest Road-Side Unit (RSU), which acts as the edge aggregator. The overall architecture is demonstrated in Fig. 1.

RSU leverages stable control aggregation, update history similarity suppression, and median deviation robustness, unlike FedAvg. Hashing and aggregation metadata can be logged on a permissioned ledger, although that is not part of the defense being evaluated.

In a future implementation, records containing the round identifier, global-model hash, deviation statistics, and aggregation weights could be validated using authenticated RSUs. It is not claimed that there is low overhead for the blockchain; consensus protocol selection, block generation rate, smart contracts, latency, storage requirements, and RSU costs are all outside of the current scope of work.

There are four general stages in each of the rounds of Federated Learning conducted in the system:

1. Local model training on vehicles using private kinematic data.
2. Secure transmission of model updates to the nearest RSU.
3. Multi-stage robust aggregation using a three-stage defence approach at the RSU.
4. Optional logging of the global-model hash and aggregation metadata.

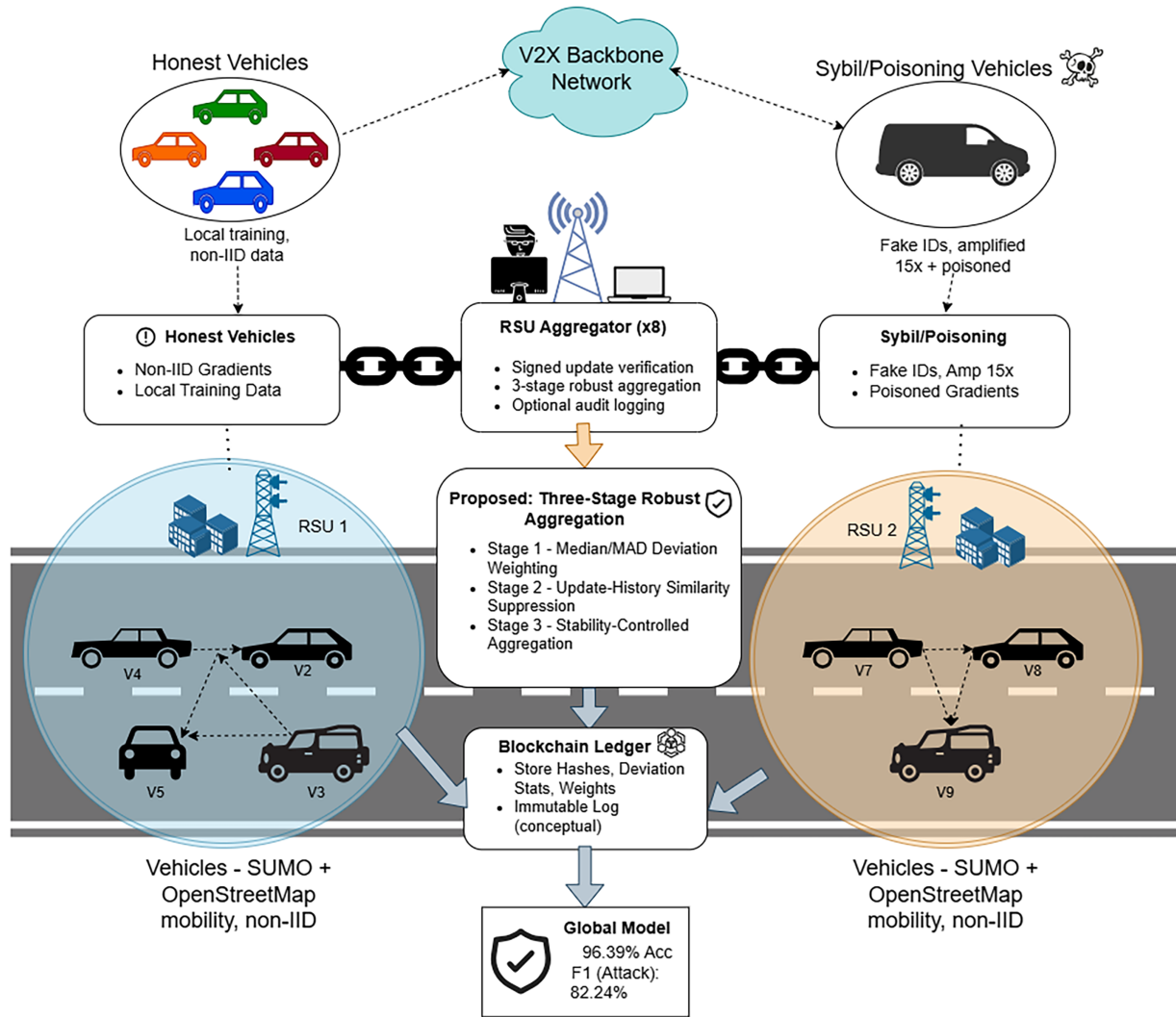


Figure 1: Evaluated V2X federated-learning flow and optional conceptual audit layer.

4.2 Local Training and Update Submission

Each vehicle $v_i \in \mathcal{V}$ initialises its local accident detection model with the current global parameters $\mathbf{w}^{(t)}$ received from the RSU. Local training is conducted for a fixed number of epochs E on the vehicle's private kinematic dataset $\mathcal{D}_i$, which includes features such as speed statistics, acceleration patterns, stop frequency, trip duration, and waiting times.

After local training, the vehicle computes the model update

$$\Delta \mathbf{w}_i = \mathbf{w}_i^{(t+E)} - \mathbf{w}^{(t)}, \quad (4)$$

$\mathbf{w}_i^{(t+E)}$ denotes the locally updated model parameters after E epochs. Only the update $\Delta \mathbf{w}_i$ is transmitted to the RSU, and the raw sensor data remain exclusively on the vehicle to ensure privacy.

The protocol design assumes authenticated transport. A deployment can hash and sign each serialized update and verify the signature at the RSU; cryptographic latency is not included in the experiments.

4.3 Three-Stage Defence-Based Robust Aggregation

The proposed architecture employs a three-stage Byzantine-resilient aggregation scheme to guard against Sybil attacks and model poisoning in a heterogeneous vehicle federated learning environment.

The method does not involve an inflexible exclusion policy based on a particular criterion but instead incorporates stability-controlled aggregation, similarity suppression based on history, and robust deviation weight.

4.3.1 Deviation-Based Gradient Filtering

Based on the local model updates $\Delta \mathbf{w}_i$ provided by the participating vehicles, the RSU computes how far the update is from the robust central tendency of the gradient distribution. The deviation measure is computed using the following formula:

$$d_i = \|\Delta \mathbf{w}_i - \text{Median}(\{\Delta \mathbf{w}_j\}_{j=1}^N)\|_2 \quad (5)$$

where the median operator ensures robustness to strong outliers.

Let $m_d = \text{median}_i(d_i)$ and $s_d = \text{median}_i |d_i - m_d| + \varepsilon$.

4.3.2 Adaptive Robustness Weighting

The defence strategy assigns a median-deviation robustness weight ρ_i to every update in order to control client influence efficiently.

$$\rho_i = \frac{1}{1 + \exp(2(d_i - m_d)/s_d)} \quad (6)$$

where ρ_i is the median-deviation robustness weight.

The exponentially smoothed update history is $\mathbf{h}_i^{(t)} = 0.8\mathbf{h}_i^{(t-1)} + 0.2\Delta \mathbf{w}_i^{(t)}$. With $q_i = \max_{j \neq i} \cos(\mathbf{h}_i^{(t)}, \mathbf{h}_j^{(t)})$, the collusion weight and normalized aggregation weight are

$$c_i = \text{clip}\left(\frac{1 - q_i}{0.10}, 0.05, 1\right), \quad \tilde{\alpha}_i = \frac{\rho_i c_i}{\sum_j \rho_j c_j + \varepsilon}. \quad (7)$$

4.3.3 Stability-Aware Aggregation

For p model parameters, the update dispersion and stability coefficient are

$$v_t = \frac{1}{p} \sum_{k=1}^p \text{Var}_i(\Delta \mathbf{w}_{i,k}^{(t)}), \quad \gamma_t = \frac{1}{1 + 5v_t}. \quad (8)$$

The global update is computed as:

$$\Delta \mathbf{w}^{(t)} = \gamma_t \sum_{i=1}^N \tilde{\alpha}_i \Delta \mathbf{w}_i^{(t)} \quad (9)$$

and the model update becomes:

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} + \Delta \mathbf{w}^{(t)} \quad (10)$$

The tri-stage architecture successfully handles a variety of client updates, prevents coordinated malevolent influence, and sustains steady convergence even in the face of an attack. The complete three-stage robust aggregation procedure is summarized in Algorithm 1.

Algorithm 1: Three-stage robust aggregation at RSU

Require: Updates $\{\Delta \mathbf{w}_i^{(t)}\}_{i=1}^N$, histories $\{\mathbf{h}_i^{(t-1)}\}_{i=1}^N$

Stage 1: Median-deviation robustness

```

1:  $\mathbf{m} \leftarrow \text{Median}(\{\Delta \mathbf{w}_i^{(t)}\}_{i=1}^N)$ 
2: for each client  $i = 1$  to  $N$  do
3:    $d_i \leftarrow \|\Delta \mathbf{w}_i^{(t)} - \mathbf{m}\|_2$ 
4: end for
5:  $m_d \leftarrow \text{Median}(\{d_i\})$ 
6:  $s_d \leftarrow \text{Median}(\{|d_i - m_d|\}) + \varepsilon$ 
7: for each client  $i = 1$  to  $N$  do
8:    $\rho_i \leftarrow [1 + \exp(2(d_i - m_d)/s_d)]^{-1}$ 
9: end for

```

Stage 2: Update-history similarity suppression

```

10: for each client  $i = 1$  to  $N$  do
11:    $\mathbf{h}_i^{(t)} \leftarrow 0.8\mathbf{h}_i^{(t-1)} + 0.2\Delta \mathbf{w}_i^{(t)}$ 
12:    $q_i \leftarrow \max_{j \neq i} \cos(\mathbf{h}_i^{(t)}, \mathbf{h}_j^{(t)})$ 
13:    $c_i \leftarrow \text{clip}((1 - q_i)/0.10, 0.05, 1)$ 
14: end for
15:  $\tilde{\alpha}_i \leftarrow \rho_i c_i / (\sum_j \rho_j c_j + \varepsilon)$ 

```

Stage 3: Stability-controlled aggregation

```

16:  $v_t \leftarrow p^{-1} \sum_{k=1}^p \text{Var}_i(\Delta \mathbf{w}_{i,k}^{(t)})$ 
17:  $\gamma_t \leftarrow (1 + 5v_t)^{-1}$ 
18:  $\Delta \mathbf{w}^{(t)} \leftarrow \gamma_t \sum_i \tilde{\alpha}_i \Delta \mathbf{w}_i^{(t)}$ 
19:  $\mathbf{w}^{(t+1)} \leftarrow \mathbf{w}^{(t)} + \Delta \mathbf{w}^{(t)}$ 

```

Ensure: Updated global model $\mathbf{w}^{(t+1)}$

Table 2 summarizes the notation used in the proposed aggregation method.

Table 2: Notation summary.

Symbol	Description
N	Number of participating vehicles
$\mathcal{D}_i$	Local dataset of vehicle i
n_i	Number of samples held by client i
$F(\mathbf{w})$	Global federated objective function
$F_i(\mathbf{w})$	Local empirical loss of client i
$\mathbf{w}^{(t)}$	Global model parameters at round t
$\Delta \mathbf{w}_i^{(t)}$	Local model update from client i
$\tilde{\Delta \mathbf{w}}_i^{(t)}$	Amplified malicious update
a	Malicious update-amplification factor
d_i	Deviation score of client i

(Continued)

Table 2 (continued)

Symbol	Description
m_d	Median of client deviation scores
s_d	Median absolute deviation scale
ρ_i	Median-deviation robustness weight
$\mathbf{h}_i^{(t)}$	Exponentially smoothed update history
q_i	Maximum history cosine similarity for client i
c_i	Similarity-based collusion weight
$\tilde{\alpha}_i$	Normalized aggregation weight
p	Number of model parameters
v_t	Mean coordinate-wise update variance
γ_t	Stability coefficient at round t
ε	Small positive numerical-stability constant

4.4 Robustness and Design Rationale

The framework takes advantage of the properties inherent to non-IID data for vehicle traffic caused by different roads, behaviors of the drivers, traffic conditions, and environment. Consequently, even though Sybil clients act in coordination and generate strongly correlated updates with a common malicious objective, legitimate vehicles could produce various gradients.

The system is able to defend against attacks that are coordinated without punishing legitimate, yet heterogeneous, honest device participation without the need for an outlier rejection mechanism based on one measure, such as Krum's distance metric or coordinate-wise median filtering. In the case of safety-critical applications that employ V2X, where failing to reject legitimate updates might be as bad as allowing tainted updates into the system, this feature is highly valuable.

4.5 Computational Complexity Analysis

Let d be the dimensionality of the model and N be the number of clients. $O(Nd)$ computations are required to determine the median deviation. The update history similarity computation takes $O(N^2d)$ computations since it measures pairwise cosine similarities between the histories of N clients. $O(Nd)$ computations are required for stability-aware aggregation and adaptive robustness weighting. Therefore, despite different operations and constants, the complexity remains the same at $O(N^2d)$, making its order asymptotically the same as that of the pairwise-distance-based approach such as Krum.

5 Experimental Setup and Evaluation Methodology

The simulation environment, dataset creation process, federated learning configuration, attack scenarios, evaluation metrics, and reproducibility methods used to evaluate the suggested secure federated learning architecture are all described in this section.

5.1 Mobility and Dataset Generation

The Simulator for Urban Mobility (SUMO) is utilized to simulate the movement patterns of vehicles utilizing the Open Street Map (OSM) data for vehicle behaviour and provide realistic vehicle movement characteristics. The SUMO lane-change and car-following models provide realistic variations in the speed,

acceleration, deceleration, stopping, and congestion for vehicles in the real world. The resulting mobility scenario is shown in [Fig. 2](#).

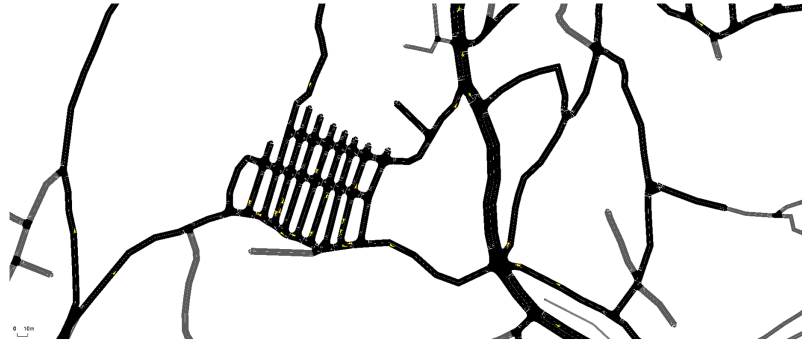


Figure 2: SUMO-based vehicular mobility scenario used for dataset generation.

Each vehicle generates detailed Floating Car Data (FCD) and trip information throughout its journey. The 11 kinematic and temporal features used by the model are listed in [Table 3](#).

Table 3: Dataset feature description.

Feature	Description
Mean Speed	Average speed within the window
Min Speed	Minimum speed within the window
Max Speed	Maximum speed within the window
Speed Standard Deviation	Standard deviation of speed
Speed Drop	Difference between maximum and minimum speed
Mean Absolute Acceleration	Mean absolute acceleration
Maximum Deceleration	Largest observed negative acceleration magnitude
Stopped Fraction	Fraction of observations at stopped speed
Longest Stop	Longest continuous stopped duration
Distance	Distance travelled within the window
Lane Changes	Number of lane changes within the window

[Table 4](#) reports the curated dataset size, class composition, simulation seeds, and scenario-separated split.

Table 4: Dataset statistics.

Total Window Samples	10,000
Controlled-accident windows	995
Normal/congestion windows	9005
Controlled events	500
Simulation seeds	5
Training/validation/test	6000/2000/2000
Window duration	10 s

The approximately 10% positive prevalence is intentionally imbalanced. It provides a more realistic rare-event evaluation than balancing every client to 50%, while not claiming to reproduce a measured population crash rate.

The scenario-separated class composition is shown in Fig. 3.

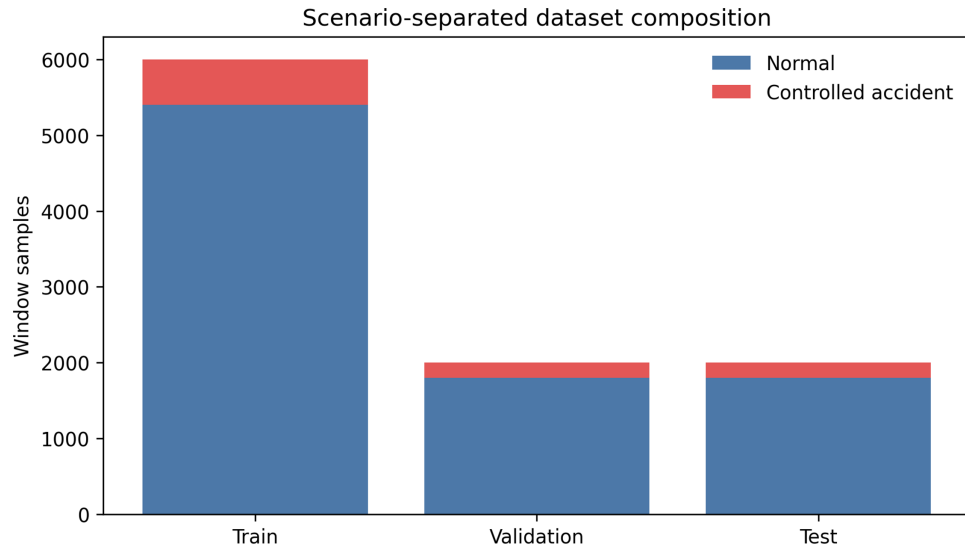


Figure 3: Scenario-separated composition of the curated synthetic-event dataset.

Fig. 4 visualizes representative kinematic separation between normal/congestion windows and controlled-event windows.

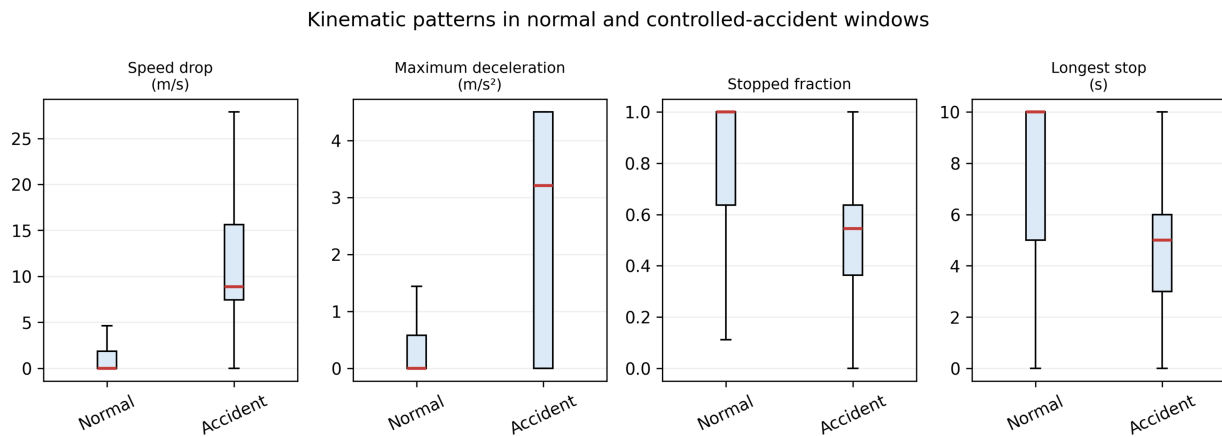


Figure 4: Feature patterns for normal and controlled-accident windows; outliers are omitted from the boxplot display only.

Through the selection of dynamic vehicles, the controlled emergency deceleration of the vehicle, and halting the vehicle for a particular period of time, TraCI creates synthetic accident-like events on purpose. A ground-truth manifest is employed to document the exact vehicle and the event window. Boundary windows are not considered, and the creation of a positive sample can neither be achieved through congestion nor by SUMO teleportation alone; it only becomes positive if 50% of the window corresponds to a triggered event.

5.2 Implementation Workflow

Client partitioning and feature extraction are carried out using traces of vehicle movement. FedAvg serves as the baseline method for federated learning. To simulate adversarial conditions, Sybil amplification and gradient manipulation are introduced during training. A tri-stage aggregation mechanism is then employed to ensure robustness against such attacks.

Deterministic seeding is employed in all experiments. The learning rate is 0.07, and each client performs eight full-batch local epochs per communication round. Binary cross-entropy uses an inverse-prevalence positive-class weight computed from the training split. For every method, the classification threshold is chosen by maximizing F1 on the validation scenario only; the test scenario is never used for model or threshold selection.

5.3 Baseline and Defense Schemes

Five federated learning scenarios are evaluated:

- **Baseline FL (FedAvg):** Standard federated averaging with no adversarial interference.
- **FL under Attack:** FedAvg under combined Sybil and model poisoning attacks.
- **FoolsGold under Attack:** History-similarity-based Sybil defence evaluated under coordinated label-flipping and update-amplification attacks.
- **Three-Stage Defence Mechanism (Clean):** The proposed three-stage robust aggregation evaluated under benign conditions to assess convergence behaviour and baseline robustness.
- **Proposed Secure FL:** Three-stage defence mechanism-based tri-stage robust aggregation applied in the presence of coordinated Sybil and poisoning attacks to evaluate suppression effectiveness and safety recovery.

This comparison highlights the impact of adversarial behaviour and assesses the effectiveness of the proposed defence mechanism.

5.4 Evaluation Metrics

These metrics are used since the detection of accidents is very crucial:

Accuracy measures the overall accuracy of the model's predictions. Whereas recall checks the extent to which the program identifies actual accidents, precision evaluates the reliability of the predicted accident scenarios. The F1 score provides an objective evaluation of the model's performance by measuring recall and precision.

The unmodified test scenario for every method is evaluated once by the decision threshold, which is selected for the validation scenario only through maximizing validation F1. Other metrics are calculated from the result files at a fixed threshold of 0.5. The choice of threshold does not depend on test labels.

5.5 Reproducibility and Fairness Considerations

All simulations use fixed random seeds, consistent client participation rates, and identical attack parameters, learning rates, and model configurations. Reproducibility and objective comparability between baselines, attacks, and defenses are ensured through these criteria.

To quantify initialisation variability, experiments use seeds 42–46 with the same scenario-separated train, validation, and test partitions. The implementation includes the controlled-event generator, feature extraction, federated training, attack simulation, result files, and figure-generation scripts. [Tables 5](#) and [6](#)

report the common simulation, system, training, and optimization settings. The complete federated learning execution command and configuration are provided in [Appendix A](#).

Table 5: Simulation and system configuration parameters.

Parameter	Value
Mobility and Network Setup	
Simulator	SUMO
Road Map Source	OpenStreetMap
Controlled Events	500 across five runs
Simulation Termination	Until no scheduled vehicles remain
Road Networks	One OSM-derived map
Scenario Seeds	42–46
Federated Learning Structure	
Model Type	Binary Classifier (MLP)
Number of FL Rounds	80
Client Data Distribution	Non-IID
Aggregation Methods Compared	FedAvg/FoolsGold/Three-stage defence mechanism
Adversarial Configuration	
Attack Type	Sybil + Model Poisoning
Sybil Amplification Factor	15
Malicious Client Ratio	20%

Table 6: Federated learning training and optimization parameters.

Parameter	Value
Learning Rate	0.07
Local Epochs per Round	8
Local Optimisation	Full local client partition
Positive Class Weight	Training-set inverse prevalence
Threshold Rule	Maximise validation F1
Hidden Layer	16 ReLU units
Input/Output Units	11/1
Experiment Seeds	42–46

5.6 Implementation of Sybil Amplification

Malicious clients in the attack setup have a significantly amplified effect in the aggregation phase by scaling their gradient updates with an amplification factor. This is done to simulate the coordinated gradient-exaggeration attacks commonly observed in Byzantine federated learning scenarios.

6 Results and Performance Analysis

6.1 Five-Seed Comparative Evaluation

Table 7 reports mean test performance across initialisation seeds 42–46. Thresholds were selected independently for each trained method using validation F1 only. The proposed defence retained performance close to its clean setting under attack: mean accuracy changed from 96.73% to 96.39%, while F1 changed from 83.84% to 82.24%. Attacked FedAvg and FoolsGold were substantially less stable, as shown by their lower mean F1 and larger seed-to-seed variation.

Table 7: Five-seed test performance (mean $\pm$ sample standard deviation).

Scenario	Accuracy (%)	F1 (%)	Bal. Acc. (%)	MCC
FedAvg (Clean)	96.86 $\pm$ 0.27	84.21 $\pm$ 1.67	91.28 $\pm$ 1.78	0.8250 $\pm$ 0.0181
FedAvg (Attack)	76.52 $\pm$ 37.24	59.17 $\pm$ 23.31	78.69 $\pm$ 16.07	0.5305 $\pm$ 0.2997
FoolsGold (Attack)	25.21 $\pm$ 32.66	24.70 $\pm$ 14.56	57.09 $\pm$ 15.16	0.1089 $\pm$ 0.2102
Proposed (Clean)	96.73 $\pm$ 0.09	83.84 $\pm$ 0.60	91.66 $\pm$ 1.40	0.8208 $\pm$ 0.0066
Proposed (Attack)	96.39 $\pm$ 0.29	82.24 $\pm$ 1.00	90.84 $\pm$ 0.61	0.8028 $\pm$ 0.0111

Fig. 5 visualizes the five-seed F1 comparison and its sample standard deviation.

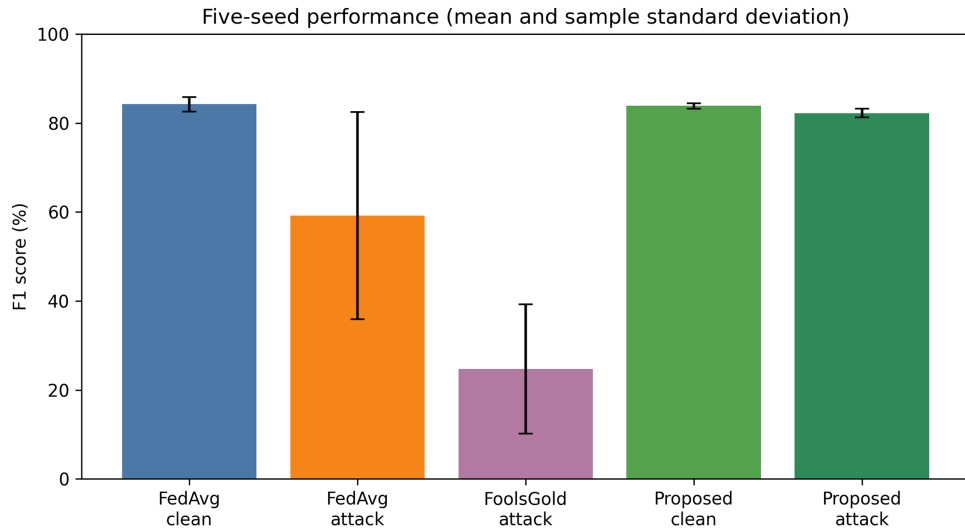


Figure 5: Five-seed F1 comparison. Error bars show sample standard deviation.

6.2 Attack Sensitivity and Scalability

At 20% malicious participation, proposed-method F1 ranged from 83.63% at amplification 5 to 82.18% at amplification 20. At amplification 15, increasing malicious participation from 10% to 40% reduced F1 from 82.35% to 80.81%. These results support resilience within the tested range but do not establish a guarantee outside it.

Fig. 6 shows the sensitivity to update amplification and malicious-client ratio.

Client-population experiments used 50, 100, 250, and 500 FL clients. Proposed attacked F1 was 82.87%, 82.67%, 82.35%, and 80.39%, respectively. The decline at 500 clients coincides with only 11–13 training samples per client, as shown in Fig. 7.

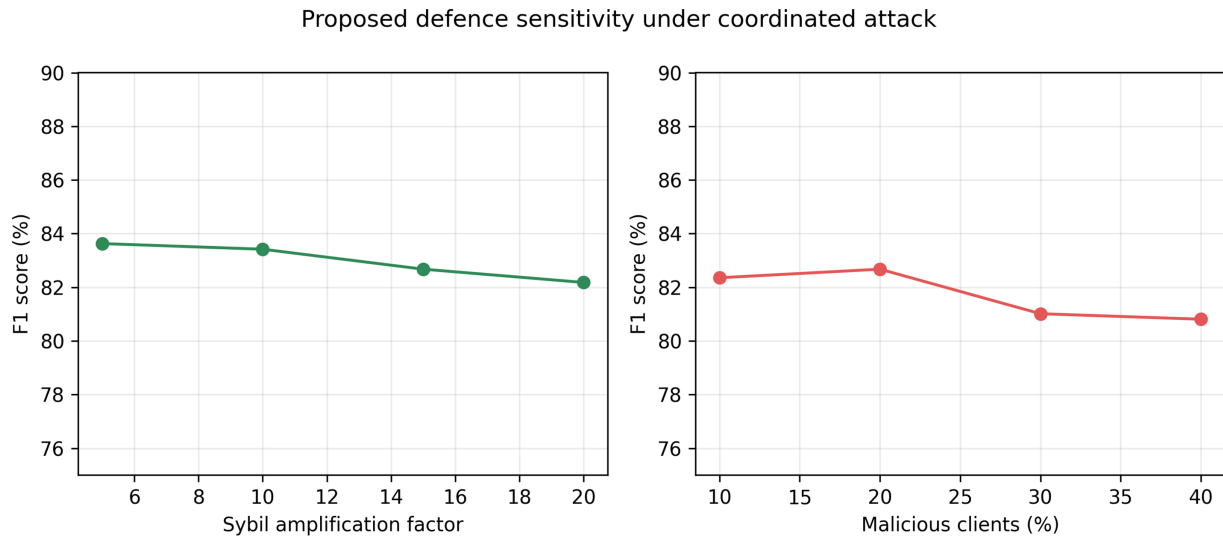


Figure 6: Sensitivity to update amplification and malicious-client ratio.

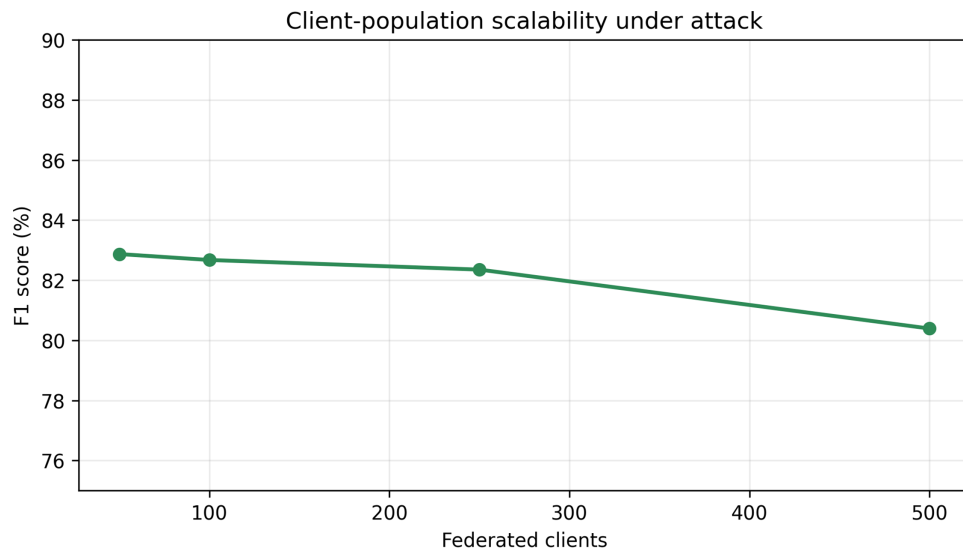


Figure 7: Scalability of the proposed defence under attack.

6.3 Ablation

The deviation-only, deviation-plus-similarity, and complete three-stage variants achieved F1 values of 84.21%, 84.42%, and 82.67%, respectively, for seed 42. Thus, similarity suppression provided a small improvement in this run, whereas stability scaling traded a small amount of final F1 for a controlled global step. This result is reported without claiming that every stage improves classification accuracy. Fig. 8 presents the ablation results.

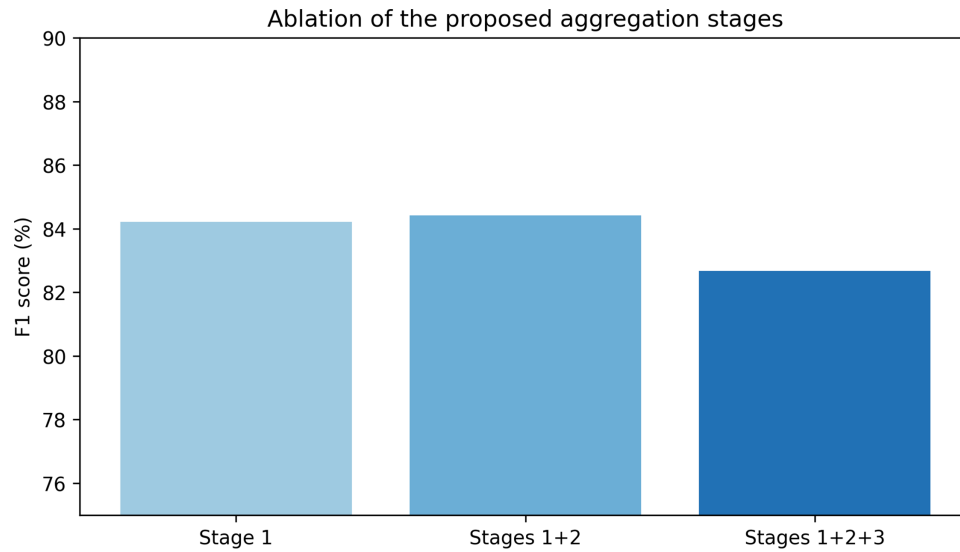


Figure 8: Ablation of the aggregation stages under attack.

6.4 Limitations

The dataset contains controlled synthetic accident-like events rather than observed real crashes and uses one OSM-derived road network. Scenario seeds provide independent mobility realisations but not geographic diversity. The backdoors, Byzantine failures, free riders, membership inference attacks, and model inversion are not tested; however, among others are label-flip poisonings with increased updates and Sybil coordination. 500 clients, not 5000, are utilized to test the scalability of the model. The latency, storage overhead, consensus overhead, and RSU overhead of the optional ledger and cryptographic protocol are not measured.

7 Conclusion and Future Work

To determine the resilience of fake V2X accident events in the presence of a coordinated Sybil and model update poisoning attack, this paper analyzes a three-stage robust aggregation method. 10,000 windows of a scenario-separated dataset and 500 simulated events were created using five SUMO runs. The proposed method outperformed the attacked versions of FedAvg and FoolsGold methods with considerably lower variance and achieved 96.39% accuracy, 82.24% F1, 90.84% balanced accuracy, and 0.8028 MCC under attack.

In terms of amplification ratios 5 to 20, malice ratios 10%–40%, and client base sizes 500, sensitivity analysis reveals degradation instead of instability. This is evident from the ablation, which shows that not all steps increase the final F1 value; there is a need for balance between precision and stability.

These results are limited to controlled occurrences from one road network simulation. Future research would have to verify the findings with other geographical maps and public/testbed V2X datasets before any claims can be made with regard to the consensus, latency, storage, energy consumption, and RSU overhead.

Acknowledgement: There are no acknowledgements in this study.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Jai Vinita L.; Methodology, Jai Vinita L., Vishvadharmam Saminathan and Yuvaraj Kurup; Software, Vishvadharmam Saminathan, Yuvaraj Kurup and Fawaz Ahamed; Validation, Vishvadharmam Saminathan, Yuvaraj Kurup, Fawaz Ahamed and Helen Vijitha P.; Formal analysis, Vishvadharmam Saminathan and Jai Vinita L.; Investigation, Vishvadharmam Saminathan, Yuvaraj Kurup and Fawaz Ahamed; Resources and supervision, Jai Vinita L.; Data curation, Vishvadharmam Saminathan and Yuvaraj Kurup; Writing—original draft preparation, Vishvadharmam Saminathan, Yuvaraj Kurup and Fawaz Ahamed; Writing—review and editing, Jai Vinita L., Vishvadharmam Saminathan, Samarth Maligouder and Helen Vijitha P.; Visualization, Vishvadharmam Saminathan and Fawaz Ahamed; Project administration, Jai Vinita L. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The controlled event, experiment code, and the synthetic data set that contains trajectory windows generated using SUMO/OSM are not available in a public repository; however, they can be accessed through the corresponding author by making a reasonable request.

Ethics Approval: Artificial data generated through SUMO/OSM traffic simulations have been used in this research; neither human nor animal participants have been utilized. For that reason, an ethics committee reference number is not relevant, and hence, ethics approval was not required.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Appendix A.1 Federated Learning Configuration

All evaluated methods are executed by one experiment driver so that they share the same dataset, client partitions, model, attack parameters, and validation-only threshold protocol.

Listing A1: Five-method revision experiment

```
python experiments/run_revision_experiments.py
--dataset datasets/curated_v2x_10000_clean.csv
--clients 100 --malicious-ratio 0.20
--rounds 80 --local-epochs 8
--learning-rate 0.07 --hidden-units 16
--attack-amplification 15 --seed 42
```

References

1. Sharma S, Kaushik B. A survey on internet of vehicles: applications, security issues & solutions. Veh Commun. 2019;20:100182.
2. Ye S, Wu Q, Fan P, Fan Q. A survey on semantic communications in internet of vehicles. Entropy. 2025;27(4):445. doi:10.3390/e27040445.
3. Afdhal A, Ahmadiar A, Adriman R. Sybil attack detection on ITS-V2X system using a realistic traffic model-based approach. In: 2022 IEEE International Conference on Communication, Networks and Satellite (COMNETSAT); 2022 Nov 3–5; Purwokerto, Indonesia. p. 1–6.
4. Hakeem SAA, El-Gawad MAA, Kim H. Comparative experiments of V2X security protocol based on hash chain cryptography. Sensors. 2020;20(19):5719. doi:10.3390/s20195719.
5. Ullah I, Deng X, Pei X, Mushtaq H, Uzair M, Qayyum S. A blockchain-based federated learning framework against poisoning attacks in the internet of vehicles. Comput Netw. 2025;272(1):111705. doi:10.1016/j.comnet.2025.111705.
6. Fardad M, Muntean GM, Tal I. A blockchain-enabled vehicular edge computing framework for secure performance-oriented V2X service delivery. IEEE Trans Veh Technol. 2024;73(9):13853–67. doi:10.1109/tvt.2024.3394150.

7. Kokuti A, Hussein A, Marín-Plaza P, de la Escalera A, García F. V2X communications architecture for off-road autonomous vehicles. In: 2017 IEEE Intelligent Vehicles Symposium (IV); 2017 Jun 11–14; Los Angeles, CA, USA. p. 696–701.
8. Pillutla K, Kakade SM, Harchaoui Z. Robust aggregation for federated learning. *IEEE Trans Signal Process.* 2022;70:1142–54. doi:10.1109/tsp.2022.3153135.
9. Li C, Huang P, Ma Y, Hung H, Huang S. Robust aggregation for federated learning by minimum Γ -divergence estimation. *Entropy.* 2022;24(5):686. doi:10.3390/e24050686.
10. Yan G, Wang H, Yuan X, Li J. DeFL: defending against model poisoning attacks in federated learning via critical learning periods awareness. In: Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence; 2023 Feb 7–14; Washington, DC, USA. p. 10711–9.
11. Qi P, Chiaro D, Guzzo A, Ianni M, Fortino G, Piccialli F. Model aggregation techniques in federated learning: a comprehensive survey. *Future Generation Computer Systems.* 2023;150(6245):272–93. doi:10.1016/j.future.2023.09.008.
12. Ye F, Yang P, Cai L, Cao X. TriGuardFL: triple-step byzantine-robust federated learning against model poisoning attacks. 2026 [cited 2026 Aug 12]. Available from: <https://openreview.net/forum?id=c6GRR7UmzI>.
13. Huang T, Liu J, Zhou X, Nguyen DC, Azghadi MR, Xia Y, et al. V2X cooperative perception for autonomous driving: recent advances and challenges. *arXiv:2310.03525.* 2023.
14. Elbir AM, Son BG, Kim YH. Federated learning for vehicular networks: a survey. In: 2022 IEEE 96th Vehicular Technology Conference (VTC2022-Fall); 2022 Sep 26–29; London, UK; Beijing, China. p. 1–6.
15. Chellapandi VP, Yuan L, Brinton CG, Žak SH, Wang Z. Federated learning for connected and automated vehicles: a survey of existing approaches and challenges. *IEEE Trans Intell Veh.* 2024;9(1):119–34.
16. Garg S, Kumar R, Kumar S. Poisoning attacks on federated learning for autonomous driving. *arXiv:2405.01073.* 2024.
17. Zhang Y, Cheong C, Li S, Cao Y, Zhang X, Liu D. False message detection in Internet of Vehicle through machine learning and vehicle consensus. *Inf Process Manag.* 2024;61(1):103827. doi:10.1016/j.ipm.2024.103827.
18. Blanchard P, El Mhamdi EM, Guerraoui R, Stainer J. Machine learning with adversaries: byzantine tolerant gradient descent. In: NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems; 2017 Dec 4–9; Long Beach, CA, USA. p. 118–28.
19. Fung C, Yoon CJM, Beschastnikh I. Mitigating sybils in federated learning poisoning. *arXiv:1808.04866.* 2020.
20. Sahraei MA, Al Mamari SRM. A review of internet of things approaches for vehicle accident detection and emergency notification. *Sustainability.* 2025;17(14):6510. doi:10.3390/su17146510.