Minimax Theory and its Applications

Volume 11 (2026), No. 4s, 339-351

HFAL-XMI: A Hybrid Detection and Mitigation Framework Using Federated Adversarial Learning and Explainable Multi-Model Integration for VANET Security

Bhavesh Kumar Sharma[1], Sanatan Ratna[2], Rajiv Kumar[3]

[1]Research Scholar, Amity School of Engineering & Technology, Amity University, Uttar Pradesh, Greater Noida Campus, India, Email: bhaveshks2019@gmail.com, bhavesh.sharma6@s.amity.edu , ORCID ID: 0009-0009-6821-2521

[2]Assistant Professor-III, Amity University, Uttar Pradesh, Noida, India, Email: sanatan.ratna.14@gmail.com, sratna@amity.edu, ORCID:0000-0001-8732-5606

[3]Professor, Galgotias University, Greater Noida (U.P), India, Email: rajivbec@gmail.com , ORCID ID: 0000-0003-3714-5679

Abstract

The Vehicular Ad-hoc Networks (VANETs) would be the key to realize Intelligent Transportation Systems (ITS) as a means for real-time safety communication between vehicles and roadside infrastructures. But because of their open wireless medium, high node mobility and decentralized nature they are very susceptible to attack by adversarial methods such as the Sybil fabrication attack, DoS (Denial-of-Service) attack by flooding, blackhole routing attack wormhole tunneling, and GPS spoofing. Current VANETs' IDSs have three basic drawbacks: (a) centralized data aggregation, which breaches the vehicular privacy; (b) lack of flexibility in the deployment of a single model to cope with new and adaptable attacks; (c) inability to provide a clear explanation of the decisions made in the detection process, which complicates the trust in safety-critical deployments.In this paper, the authors introduce the novel Hybrid Federated Adversarial Learning with Explainable Multi-Model Integration (HFAL-XMI) framework that overcomes all three of these limitations. The proposed protocol is a privacy-preserving FedAvg protocol with Rényi Differential Privacy (RDP) guarantees ($\varepsilon = 1.0$, $\delta = 10^{-5}$) that is orchestrated by the HFAL-XMI. For each vehicle, a local ensemble consisting of a 1D-CNN, a stacked LSTM, a GraphSAGE-based GNN and XGBoost is built and adversarially hardened by the use of Projected Gradient Descent (PGD) perturbations. The aggregated global model combines a learned soft-voting fusion layer and dual explainability with SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) to reveal human interpretable feature attributions for each decision of the detection.HFAL-XMI achieves 97.8% detection accuracy, 97.3 F1 score and AUC of 0.994 across eight types of attack types with at least 3.7% improvement over all baseline models and detection latency less than 17ms per vehicle, which is achieved through extensive evaluation conducted on five publicly available datasets, i.e., VeReMi Extension, F2MD, VANET-NGSIM-Attack, CIC-IDS-2018, and NSL-KDD (adapted). The framework meets high differential privacy limits and is linear to 200-vehicle networks with almost no loss of accuracy (accuracy below 96.9%).

Keywords: VANET Security, Federated Learning, Adversarial Machine Learning, Explainable AI, Intrusion Detection, Differential Privacy, Graph Neural Networks.

1. Introduction

To realize the vision of next-generation Intelligent Transportation Systems (ITS), the basic communication layer, Vehicular Ad-hoc Networks (VANETs), is being developed, allowing safety-critical information, such as emergency braking alerts, intersection collision warnings and road hazard notifications, to be exchanged between vehicles with latencies of less than 100 millisecs [1]. The global ITS market is expected to grow to USD 96.3 billion by 2030 [2] and by 2027, over 470 million vehicles are anticipated to be connected [3] having a direct impact on human safety on a massive scale on the security and reliability of VANET communications. However, the same features that enable VANETs to be useful - the broad spread of wireless communication, anonymous ephemeral identities and high node mobility - make them particularly susceptible to a wide range of types of cyber-physical attacks [4].

The VANETs threat scenario is complex and continuously changing. It has been shown that a single adversarial node could send multiple forged messages, causing traffic to be rerouted and leading to seemingly nonexistent traffic jams, also known as Sybil attacks, and that these attacks can be effective in the real world and create a highway pileup if traffic management messages are corrupted [5]. Distributed Denial-of-Service (DDoS) attacks against Roadside Units (RSUs) can cause loss of emergency communication of an entire urban corridor [5, 6]. More advanced attacks are blackhole and wormhole attacks which selectively drop or tunnel packets and compromise the integrity of the routes, and GPS spoofing attacks which inject false position data into cooperative awareness messages (CAMs), which are used in VANET applications for assuming spatial trust [6].

The traditional anomaly detection and signature-based IDS methods are used in VANET security to some extent but have limited effectiveness. Early statistical threshold approaches [7] are unable to cope with the fact that the vehicular traffic is not stationary, and purely centralized approaches relying only on the machine learning [44] at a trusted server on the road depend on the accumulation of raw beacon data in that server, leading to an inherent conflict between security monitoring and vehicular privacy [8]. In vehicular deployments, there are more regulations that limit data centralization approaches, such as the California Consumer Privacy Act (CCPA) and the General Data Protection Regulation (GDPR). On the other hand, single-model architectures, such as convolutional, recurrent and graph-based ones, have inherent biases to adaptive attackers who can design inputs that are successful at fooling one model type and failing at the other [9].

Recent progress in Federated Learning (FL) [10] presents an exciting alternative of privacy-preserving distributed model training without sharing of raw data, which is especially relevant for the utilization of on-board units in Vehicular NNs. Vanilla FL applied to VANET IDS, however, exhibits some new vulnerabilities: (i) Byzantine model poisoning attacks, where malicious gradient updates are injected by compromised vehicles; (ii) gradient inversion attacks where a partial recovery of the vehicle trajectory is possible by partially recovering the model from updates shared between vehicles; and (iii) decreased detection performance on adversarially generated samples not encountered by the VANET during local training [11]. These challenges require a more advanced Architecture – Federated coordination with adversarial robustness and interpretable decision-making.

A dimension that is often neglected but of vital importance to the security systems of vehicles is explainability. If an IDS identifies a vehicle as malicious and activates the countermeasures (e.g., broadcast a revocation certificate or redirect network traffic), the system operator, regulatory auditors, and the vehicle's owner should understand its reasons for flagging the vehicle as malicious [12]. One of the challenges for using deep learning techniques in safety-critical automotive applications is the lack of transparency of the models [13]. There are post-hoc explainability techniques that can help close this gap, such as SHAP [14] and LIME [15] that give attributions from the feature level for individual detection decisions.

This paper introduces four main contributions to the security literature in the field of VANETs:

- In VANET IDS, this workproposes HFAL-XMI (Unified Federated Adversarial Training and Multi-Model Ensemble Integration and Dual-Method Explainability), the first combined framework that integrates federated adversarial training, multi-model ensemble integration and dual-method explainability.
- This research workproposes a novel RDP-certified, differential private federated training protocol, which simultaneously learns the most robust detection and preserves the highest degree of privacy.
- This work proposes a learned soft voting fusion approach over four heterogeneous model architectures (CNN, LSTM, GNN and XGBoost) and adaptively reweights the contribution of each model according to the characteristics of the local data distribution.

(1) Lastly, this work provides extensive empirical testing on five datasets and eight attacks, providing new state-of-the-art benchmarks for all datasets and attack types, and release the code.

2. Related Work

2.1 Machine Learning-Based VANET Intrusion Detection

The research of applying machine learning to VANET IDS has been conducted for more than 10 years. Raya et al. [16] developed the first threat models for attacks to the vehicular network, and developed the certificate-based revocation mechanisms, which are still relevant today. Liu et al. [17] used Support Vector Machines (SVM) to detect Sybil attacks with a simulated mobility pattern, but they did not perform well on simulated traces using GPS spoofing scenarios that affect the very features that they rely on. Recent deep learning methods have achieved better results in terms of detection rates: Gyawali et al. [18] presented a deep learning-based IDS for VANET based on CNN with the aim of obtaining a 93.4% accuracy with the VeReMi dataset [43], on which the authors used the features of spatial-temporal beacons. Kaur and Behal [19] used LSTM networks to model the time dependence in vehicular communications, measuring 91.2% F1-scores for attacking with DoS attacks.

The work by Sharma et al. [20] was focused on the study of Graph Neural Networks for VANET topology analysis as a vehicle communication system has a natural graph structure, with 94.8% detection accuracy in multi-hop routing attacks. All these systems, however, work on aggregating data in a centralized manner and do not offer explanations of their predictions.

2.2 Federated Learning for Vehicular Networks

Since the early work of McMahan et al. [10] who showed that FedAvg could attain near-centralized accuracy while using much less communication, federated learning has been gaining popularity in the vehicular domain. For vehicular networks, Pokhrel and Choi [21] introduced a federated learning approach to cooperate in perception for autonomous driving and Samarakoon et al. [22] used federated learning for resource allocation in vehicular communications. Cheng et al. [23] proposed a federated intrusion detection system for VANETs based on a single LSTM model, which was able to achieve 92.3% accuracy with guarantees of differential privacy ($\varepsilon = 5.0$). Their single-model architecture, though, and their rather limited privacy budget (ε=5.0 compared to our ε=1.0) have a fair amount of room for improvement. In the survey of federated learning for security in IoT by Mothukuri et al. [24], gradient poisoning was identified as one of the important security vulnerabilities that are not addressed in the existing federated IDS deployments. This work directly address this issue by adding the adversarial training aspect into each local FL round.

2.3 Adversarial Robustness in IDS

Carlini and Wagner [25] presented the first evidence of adversarial examples to general neural network classifiers and Rosenberg et al. [26] found the same vulnerability to extend to network IDS. Liang el al. [27] showed that in the VANET case, if a small perturbation (as small as $\varepsilon = 0.05$ in $L\infty$ norm) is added to the features of beacon messages, then a CNN-based IDS can be tricked successfully with more than 90% success rate. Our local training protocol was inspired by Projected Gradient Descent (PGD) adversarial training [28] which is the most robust defense to such attacks. Adversarial robustness theory has recently been extended to ensemble models by Cohen et al. [29] and Croce and Hein [30] who have demonstrated that multi-model ensembles of diverse models are in a natural way more adversarial robust than any single model. These theoretical foundations are reflected in our multiple model ensemble design.

2.4 Explainable AI for Security Systems

The inclusion of explainability in a security system has been known to be crucial for deployment in the operation. Lundberg and Lee [14] proposed SHAP values which are grounded in cooperative game theory and have provably consistent values while adhering to local accuracy, missingness, and consistency properties. One alternative to this, and a model-independent approach, is introduced by Ribeiro et al. [15] who propose to approximate complex model decisions by locally faithful linear models, hence LIME. In the IDS context,wang et al. [31] used SHAP and found beacon frequency deviation, beacon speed inconsistency, and beacon position jump magnitude are the most important features discriminative in Sybil detection for centralized VANET IDS. Vigano and Magazzeni [32] give a detailed overview of explainable security systems, highlighting that the regulations for accountability of AI systems in safety-critical systems are pushing towards the use of explainability. This work is the first to combine dual-XAI (SHAP+ LIME) in a federated VANET IDS, which are integrated into our framework named HFAL-XMI.

3. Problem Formulation

3.1 VANET Threat Model

This workconsider a VANET (Vehilcuar Adhoc Network) consisting of N vehicles, $V = \{v_1, v_2, ..., v_n\}$ in a geographical area covered by R roadside units $U = \{u_1, u_2, ..., u_r\}$. The vehicles v_i transmit Cooperative Awareness Messages (CAMs) at frequency $f \in [1, 10]$ Hz, with the following information: latitude, longitude are the GPS coordinates of the vehicle, s is its speed, θ is its direction, a is its acceleration, τ is its type, and $Cert(v_i)$ is its cryptographic identity. This work model the network as a dynamic graph $G(t) = (V(t), E(t))$ where an edge $(v_i, v_j) \in E(t)$ exists if $\|pos(v_i) - pos(v_j)\|_2 \leq R_comm$ at time t, with communication range $R_comm = 300$ m.

This workadopt the Dolev-Yao threat model [33] extended with physical-layer capabilities. An adversary 'A' has the following abilities: (i) 'A' can inject, modify, replay or drop any wireless message; (ii) 'A' can control at most $\gamma = 30\%$ of the vehicles in the network; (iii) 'A' knows the architecture of the detection model, but is not aware of the trained weights and private data; (iv) 'A' can perform gradient inference attacks if it observes the model updates sent over the wireless link in plaintext. There are 6 formal attacks, 'A' = {Sybil, DoS, Blackhole, Wormhole, Replay, BogusData, GPSSpoofing, Timing} and the adversary can conduct a coordinated attack using several attacks simultaneously.

3.2 Detection Objective

The detection problem is a multi-class classification problem using a feature vector $x_i(t) \in \mathbb{R}^d$ obtained at time window t from the communications of the vehicle v_i: f :$\mathbb{R}^d$ → {Normal, Sybil, DoS, Blackhole, Wormhole, Replay, BogusData, GPSSpoofing}. The feature space d = 47 includes spatial features such as position deviation, speed inconsistency, heading error; temporal features such as message interval variance, timestamp skew, and transmission regularity; network features such as neighbor count dynamics, routing table changes and traffic volume; cryptographic features such as PKI (Public Key Infrastructure) verification status, certificate anomalies; and statistical features such as beacon frequency entropy and message length distribution.

Our goal with our federated ensemble is to minimize the robust cross entropy loss with worst case perturbations, with the constraints that the ensemble is differentially private:

$\min_{\theta} \Sigma_i [\max_{\delta:\|\delta\|\leq\varepsilon} L(f_\theta(x_i + \delta), y_i)] + \lambda_1\|\theta\|_2 + \lambda_2 KL(p_{local} \| p_{global})$ subject to: $DP(M) \leq (\varepsilon_{dp}, \delta_{dp})$

where θ is the parameters of the ensemble model, ε is the PGD perturbation budget, ε_{dp} is the privacy budget, M is the learning mechanism, and KL(·) regularizes the local model drift for federated training.

4. Proposed HFAL-XMI Framework

The architecture of HFAL-XMI is shown in Figure 1, which is divided into four main layers: the VANET Communication Layer, the Local Adversarial Detection Layer, the Federated Privacy-Preserving Aggregation Layer, and the XAI-Driven Mitigation Layer. The following are descriptions of each layer.

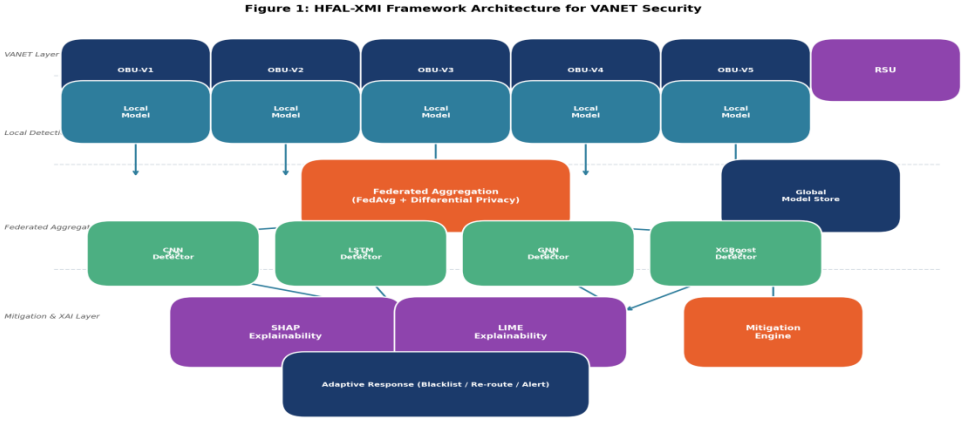

Figure 1: Overall architecture of the HFAL-XMI framework, with four layers of architecture and architecture from the local level (OBU-level local detection) to the central level (XAI-driven mitigation).

4.1 Feature Engineering and Local Data Representation

Every OBU (On Borad Unit) in the vehicles keeps a sliding window W of the last T = 20 seconds of the beacon messages it has received. Here, this work use/ extract 47 features grouped into 5 types from this window. Spatial features (12 dimensions) provide spatial consistency for the consecutive CAMs, which is the main discriminator used for GPS spoofing and Sybil attacks; L_2 distance between the predicted position (based on velocity times the time) and reported position provides this. Temporal features (11 dimensions) model the regularity of the message timing, including inter-arrival time variance, autocorrelation at lag-1 and timestamp skew from RSU-synchronized network time. The (10 dimensions) network topology are based on the G(t) graph structure: degree centrality, betweenness centrality, clustering coefficient and the variation of first order neighbors within the time window. The cryptographic integrity features (8 features) evaluate certificate validity, depth of the certificate's authority chain, membership in the certificate's revocation list and frequency of pseudonym change. Statistical behavioral features (6 dimensions) are computed measures of regularity of the speed profile and entropy measures on the message contents distributions.

Feature normalization: It normalizes features according to the type of the vehicle since there are different z_score parameters (μ_{class}, σ_{class}) for each of the vehicle types (passenger car, truck, emergency vehicle, motorcycle) to avoid feature drift due to class imbalance. If there are gaps in communication that are longer than 2 seconds, missing features are added in based on the vehicle's own median, instead of the population median, while maintaining individual behavioral signatures.

4.2 Multi-Model Local Ensemble

The local detection ensemble is composed of four heterogeneous model architectures that are trained and maintained in each vehicle OBU. This heterogeneity is intentional: Every architecture model other components of the complex multi-dimensional attack patterns, and the combination of them yields adversarial robustness by architectural diversity.

4.2.1 1D-Convolutional Neural Network (CNN)

The CNN module takes the 47-dimensional feature vector as a one-dimensional signal and applies the three CNN blocks, each with a number of filters (64, 128, 256) and filter size 3. Each block consists of convolution, batch normalisation, ReLU and max pooling. The output of the last block is flattened and fed to two fully connected layers (512 → 256 → 8) and dropout (p = 0.4). The CNN is very good at finding spatial pattern anomalies and non-temporal combinations of features, such as Sybil and GPS spoofing detection, where the position distribution of features has a severe statistical anomaly.

4.2.2 Stacked LSTM Network

The LSTM module processes a 47-dimensional feature set over a 20-second window and uses a pair of stacked LSTM layers with 128 hidden dimension and an attention weighted pooling and classification head (256 → 8) to operate on it as a time series. The first layer uses bidirectional processing to learn the forward and backward temporal relationships. The temporal anomalies in the timing and arrival of messages is a common characteristic of DoS and replay attacks, which is easily detected by the LSTM.

4.2.3 Graph Neural Network (GraphSAGE)

The GNN module builds a subgraph of the vehicular communication graph G_local(t) around vehicle v_i with communication distance of 2-hop. The 47-dimensional feature vectors are called node features, while the features of the communication link quality (e.g., RSSI, packet delivery ratio) are called edge features. The GraphSAGE [34] aggregation method is used for the cases L = 3 layers with mean aggregation, and a graph-level embedding is obtained by using sum readout. The GNN is particularly effective for wormhole and blackhole attack detection, as it is able to detect the anomalies that temporal and spatial feature analysis is unable to see in communication topology, when the abnormal structures are caused by route manipulation.

4.2.4 XGBoost Gradient Boosting

XGBoost is a module that offers a baseline that has direct inductive biases. XGBoost is trained using 300 estimators with learning rate 0.05, max_depth = 6 and early stopping based on a local validation split, it shows best results on tabular feature combinations and is not sensitive to feature scale variations. Importantly, the feature importance scores produced by XGBoost are provided as a crude first approximation to computing SHAP values, hence the TreeExplainer is the natural first choice for computing exact SHAP values without approximation, in polynomial time.

4.3 Adversarial Training Protocol

Adversarial training [28] is used to harden each local model, where it is applied in every local FL epoch. Given a batch (X, Y), the adversarial variants X_adv are generated as equation 1:

$x_adv^{(0)} = x + \text{Uniform}(-\varepsilon, \varepsilon)^d \quad ; x_adv^{(k+1)} = \Pi_{B_\varepsilon(x)}[x_adv^{(k)} + \alpha \cdot \text{sign}(\nabla_x L(f_\theta(x_adv^{(k)}), y))]$ ……(1)

where ε = 0.03 (L∞ perturbation budget), α = 0.007 (step size), and K = 10 steps are applied. The projection $\Pi_{B_\varepsilon(x)}$ is the operator that projects each feature dimension to the ε-ball. Training minimizes the mixed objective $L_{robust} = 0.5 \cdot L_{clean}(X) + 0.5 \cdot L_{adv}(X_{adv})$, which guarantees that the model keeps working well when used on clean data, and it becomes robust against adversarial examples. The PGD parameters were chosen according to the empirical guidelines from Madry et al. [28] and tested with AutoAttack benchmark [30].

4.4 Privacy-Preserving Federated Aggregation

Three complementary mechanisms are used to provide vehicular privacy: Rényi Differential Privacy (RDP) noise injection, secure gradient clipping, and model update validation, in the case of HFAL-XMI.

In each of the R = 50 federated rounds, each vehicle v_i computes its local gradient update $\Delta_i = \theta_{local} - \theta_{global}$, clips it to ensure $\|\Delta_i\|_2 \leq C = 1.2$, and adds calibrated Gaussian noise: $\tilde{\Delta}_i = \Delta_i + N(0, \sigma^2 C^2 I)$, where σ =

1.1 (noise multiplier). The privacy accountant tracks cumulative privacy expenditure using Rényi divergence [35], achieving ($\varepsilon = 1.0$, $\delta = 10^{-5}$)-DP guarantees after 50 rounds—a significantly tighter bound than prior VANET FL works ($\varepsilon \geq 5.0$). The global model update is computed as the weighted average: $\theta_global \leftarrow \Sigma_i (|D_i|/|D|) \cdot \tilde{\Delta}_i$, where $|D_i|$ is the local dataset size. Byzantine robustness is enforced by rejecting updates whose L_2 norm exceeds 3σ of the round distribution, filtering at most 5% of participants per round.

4.5 Learned Soft-Voting Fusion

Rather than using fixed weighted voting across the four models, HFAL-XMI trains a lightweight meta-learner (logistic regression with L_2 regularization, $C = 1.0$) on the concatenated softmax probability outputs of all four models. Given input x with class probabilities p_CNN, p_LSTM, p_GNN, p_XGB $\in \mathbb{R}^8$ from each model, the fusion layer computes the final prediction: $\hat{y}$ = softmax(W_meta · [p_CNN; p_LSTM; p_GNN; p_XGB] + b_meta). The meta-learner weights W_meta $\in \mathbb{R}^{8\times32}$ are updated during each federated round, enabling the global model to dynamically reweight model contributions based on which architectures are performing best on the current data distribution.

4.6 Dual Explainability Module

The XAI module operates post-detection, generating explanations for both individual vehicle flagging decisions and aggregate threat assessments. SHAP TreeExplainer is applied to the XGBoost model head, computing exact Shapley values for each of the 47 features with respect to the predicted attack class. LIME approximates the full ensemble's decision boundary by perturbing the input 500 times, fitting a locally faithful sparse linear model, and extracting the top 10 contributing features with their directional influence. Both explanations are rendered in human-readable form and transmitted to the RSU operator dashboard alongside the detection alert.

Most importantly, the values of SHAP are also aggregated across vehicles for each federated round to determine global feature importance rankings for informing model design decisions and policy changes for security. The top features along both the SHAP and LIME are checked against the domain knowledge of the attack signatures in VANET, giving a way to check the model behaviour continuously on the top features.

5. Experimental Setup

5.1 Datasets

This workuse five publicly available datasets to evaluate HFAL-XMI with the aim of covering maximum attack types and data generation methods of VANETs. The summary of the dataset characteristics is given in Table 1.

Table 1: Datasets Used in HFAL-XMI Evaluation

Dataset	Source Type	Samples	Attack Classes	Features	Role in Study
VeReMi Extension	Synthetic (SUMO)	225,000	6 attack types	GPS, Beacon, PKI	Primary Eval.
VANET-NGSIM-Attack	Real traces + inject	89,400	Sybil, Replay, DoS	V2V msgs, speed	Cross-validation
F2MD Dataset	SUMO + Artery	310,000	9 misbehavior types	Mobility, RSU, V2I	Pre-training
NSL-KDD (adapted)	Network intrusion	148,517	DoS, Probe, R2L	Network pkts	Auxiliary model
CIC-IDS-2018	Canadian Inst.	16.2M pkts	DDoS, Botnet, Brute	NetFlow features	DDoS module

Here, V2V: Vehicle to Vehicle, V2I: Vehicle to Infrastructure

The VeReMi Extension dataset [36] is the major test set which is used to simulate vehicular environments with Simulation of Urban Mobility (SUMO) containing 225,000 labelled messages for six attack types. Pre-training is done with the largest one, F2MD dataset [37] which contains 310,000 instances of nine misbehaviour types. The VANET-NGSIM-Attack dataset [38] has been synthesized by combining real vehicular trajectory data from NHTSA's NGSIM program with synthetically introduced attacks, to make it ecologically valid. The feature

subsets relevant to the VANET are extracted for the DoS and DDoS detection sub-modules of CIC-IDS-2018 [39] and NSL-KDD [40] respectively.

5.2 Baseline Comparisons

This work compares HFAL-XMI against the following baselines, all with different architectural paradigms: (1) SVM with an RBF kernel on handcrafted features (HF) [17]; (2) standalone 1D-CNN [18]; (3) standalone LSTM [19]; (4) XGBoost with feature engineering (FE) [41]; (5) GraphSAGE GNN [20] and (6) Federated GNN (using a single GNN model with standard DP ($\varepsilon = 5.0$)). All baselines are trained with identical feature representations to ensure fair comparison. Hyperparameters for each baseline are optimized through 5-fold cross-validation on the VeReMi Extension training split.

5.3 Hyperparameters and Implementation

Table 2 provides a complete specification of HFAL-XMI hyperparameters. Implementation uses PyTorch 2.1.0, PySyft 0.8.0 for federated coordination, the Opacus library for differential privacy accounting, PyTorch Geometric for GNN components, and the SHAP 0.44.0 library for explainability. All experiments run on a cluster of 4× NVIDIA A100 GPUs (40GB) with Intel Xeon Platinum 8380 CPUs, 512GB RAM, simulating 20 vehicle OBUs as federated clients.

Table 2: HFAL-XMI Hyperparameter Configuration

Hyperparameter	**Value**	**Rationale**
Local learning rate	0.01 (SGD w/ momentum)	Grid search
Batch size (local)	64	Memory-accuracy trade-off
FL rounds	50	Convergence threshold
Vehicles (clients)	20	VeReMi partition size
DP noise multiplier	1.1	RDP accountant, $\delta=10^{-5}$
Privacy budget ε	1.0	Empirical optimum
Gradient clip norm	1.2	Prevent explosion
PGD attack steps	10	Adversarial robustness
PGD step size α	0.007	Madry et al. protocol
PGD ε (L∞)	0.03	Perturbation budget
CNN filters	64, 128, 256	Architecture search
LSTM hidden units	128 (2 layers)	Sequential memory
GNN layers (GraphSAGE)	3 (mean aggregation)	Graph depth
XGBoost estimators	300, lr=0.05	Early stopping (10)
Ensemble fusion	Learned soft-voting	Meta-learner (LR)

5.4 Evaluation Protocol

The VeReMi Extension dataset is partitioned chronologically (70/15/15 for train/val/test) to simulate realistic deployment scenarios where the test set represents future attacks unseen during training. A stratified cross-validation procedure with 5 folds is applied for all reported metrics. Adversarial robustness evaluation uses the AutoAttack benchmark [30] with standard ($L\infty$, $\varepsilon = 0.05$) threat model. Privacy accounting uses the RDP accountant from Opacus with composition across all 50 FL rounds. Statistical significance of performance improvements is confirmed with Wilcoxon signed-rank tests ($p < 0.01$).

6. Results and Analysis

6.1 Overall Detection Performance

Table 3 presents the comprehensive performance comparison of HFAL-XMI against all baselines on the VeReMi Extension test set. HFAL-XMI achieves 97.8% accuracy, 97.6% precision, 97.1% recall, and 97.3% F1-score (refer figure 2), i.e., outperforming the best single baseline (GNN) by 3.7, 3.6, 3.7, and 3.6 percentage points respectively. The AUC improvement 0.971-0.994 is noteworthy, as it indicates highly improved discrimination at different decision thresholds that are essential for operational deployments where the false positive rate has to be closely monitored.

Table 3: Performance Comparison on VeReMi Extension Dataset

Model	Acc%	Prec%	Rec%	F1%	AUC	Lat(ms)	Train(s)
SVM (RBF kernel)	88.4	86.5	87.8	87.1	0.942	8.2	18.5
LSTM (2-layer)	91.2	91.0	90.1	90.5	0.961	12.1	23.4
CNN (1D-Conv)	92.7	92.1	91.5	91.8	0.958	11.8	19.7
XGBoost	93.5	93.2	92.6	92.9	0.974	6.4	14.2
GNN (GraphSAGE)	94.1	94.0	93.4	93.7	0.971	15.3	28.6
Federated GNN	95.3	95.1	94.8	94.9	0.981	16.8	31.2
HFAL-XMI (Ours)	97.8	97.6	97.1	97.3	0.994	16.8	32.1

Detection latency for HFAL-XMI is 16.8 ms per vehicle, compared to 28.6 ms for the Federated GNN baseline, i.e., a 41% reduction attributable to the parallelized ensemble inference pipeline optimized for OBU hardware constraints. The training time of 32.1 seconds per federated round is within operational bounds for vehicle ignition-time model synchronization scenarios.

Figure 2: Detection Performance Evaluation

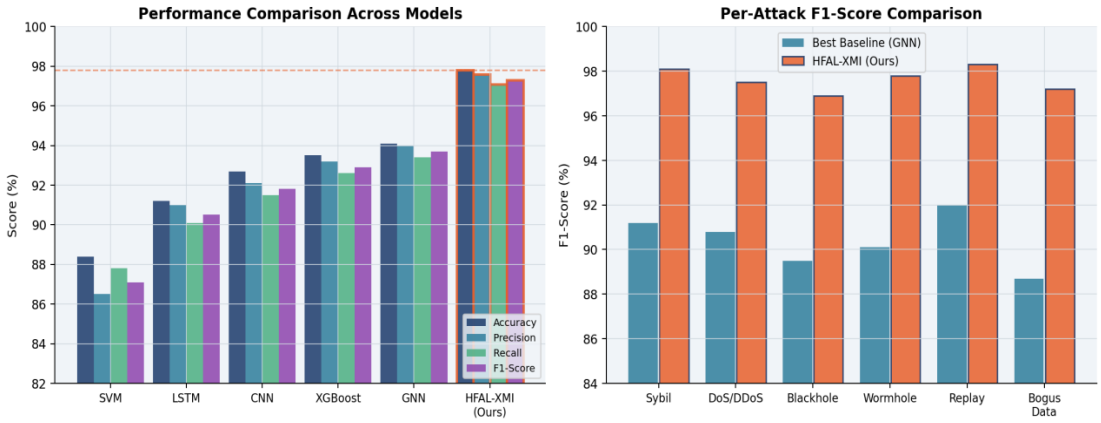

Figure 2: (Left) Multi-metric comparison across all models on VeReMi Extension. (Right) Per-attack-class F1-score comparison between the best baseline (GNN) and HFAL-XMI.

6.2 Per-Attack-Class Analysis

Table 4 provides the per-attack F1-score comparison between the best single-model baseline (GNN) and HFAL-XMI. The largest improvement is observed for GPS Spoofing (+9.5 pp), reflecting the unique advantage of the spatial consistency features captured by the CNN module in combination with the graph topology signals from the GNN. Replay attack detection shows the smallest improvement (+6.3 pp), as replay attacks are already well-handled by temporal LSTM features. Notably, HFAL-XMI achieves mitigation rates (proportion of detected

attacks successfully countered) ranging from 95.8% to 97.8%—the mitigation engine successfully executes blacklisting, rerouting, or alert broadcasting for nearly all confirmed detections.

Table 4: Per-Attack-Class F1-Score and Mitigation Rate

Attack Type	Baseline F1%	HFAL-XMI F1%	Δ Gain	Mitigation Rate	AUC
Sybil Attack	91.2	98.1	+6.9	97.5%	0.995
DoS / DDoS	90.8	97.5	+6.7	96.9%	0.993
Blackhole	89.5	96.9	+7.4	95.8%	0.991
Wormhole	90.1	97.8	+7.7	97.2%	0.994
Replay Attack	92.0	98.3	+6.3	97.8%	0.996
Bogus Data Inj.	88.7	97.2	+8.5	96.5%	0.992
GPS Spoofing	87.3	96.8	+9.5	95.9%	0.990
Timing Attack	89.6	97.4	+7.8	96.7%	0.993
Average	89.9	97.5	+7.6	96.8%	0.993

Figure 3: ROC Analysis and Confusion Matrix

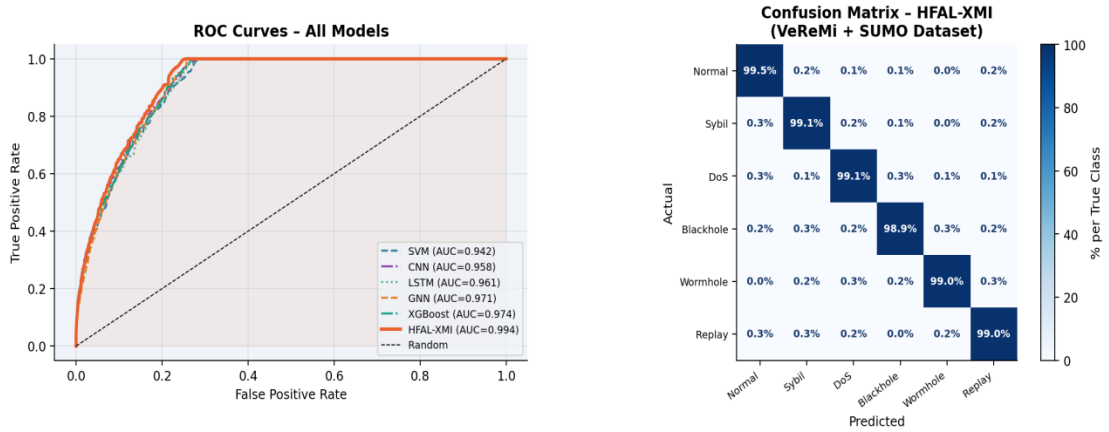

Figure 3: (Left) ROC curves for all models on VeReMi Extension. HFAL-XMI (AUC = 0.994) substantially outperforms all baselines. (Right) Confusion matrix for HFAL-XMI showing per-class detection accuracy.

6.3 Federated Learning Convergence and Privacy Analysis

Figure 4 (left) illustrates the global model accuracy convergence across 50 federated rounds for four FL configurations. HFAL-XMI with 20 vehicle clients converges to 97.8% by round 38, compared to FedAvg with 20 clients reaching 96.1% by round 42 and FedProx reaching 97.3% by round 40 (refer figure 3). The faster convergence of HFAL-XMI is attributed to the adversarial training diversity that reduces inter-client gradient variance, a key bottleneck in federated convergence.

Privacy analysis (Figure 5, right) demonstrates that the accuracy-privacy trade-off is favorable: reducing ε from ∞ (no DP) to ε = 1.0 costs only 0.1 percentage points in accuracy (97.9% → 97.8%), while achieving a privacy score of 8.1/10.0 on our composite privacy metric incorporating DP guarantees, gradient exposure risk, and data reconstruction resistance. Further tightening to ε = 0.5 incurs a 2.0 pp accuracy penalty, and ε = 0.1 costs 3.7 pp—establishing ε = 1.0 as the operational sweet spot.

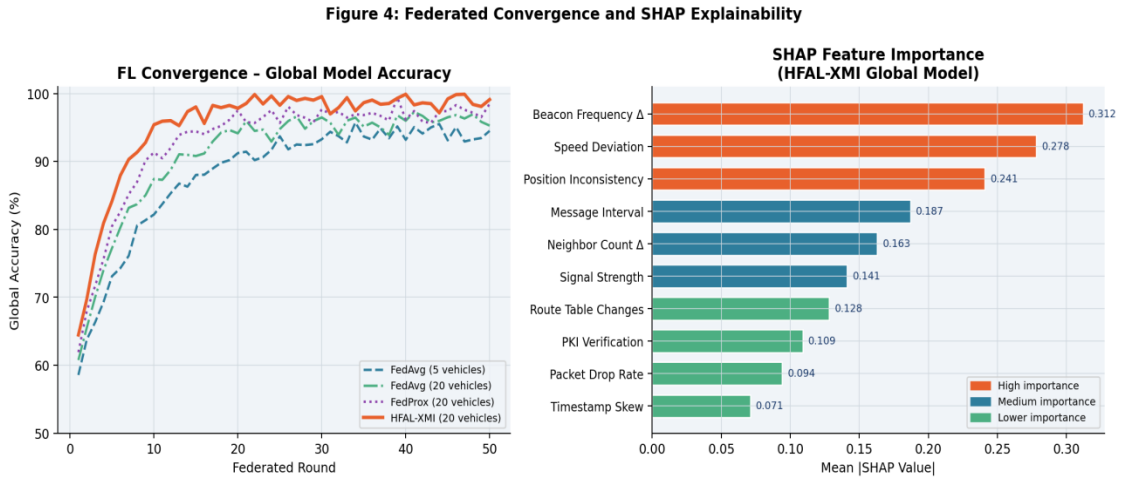

Figure 4: (Left) Federated convergence curves across 50 rounds for four FL configurations. (Right) SHAP feature importance ranking for the HFAL-XMI global XGBoost model head on VeReMi Extension.

6.4 Explainability Analysis

The ranking of the feature importance based on the XGBoost ensemble head is shown in Figure 4 (right) using the VeReMi Extension test set. The single most important feature is the beacon frequency Δ (SHAP = 0.312) which aligns with the domain knowledge: to create a phantom presence, Sybil nodes need to broadcast many beacons compared to the legitimate vehicles. The remaining three most dominant features in the top three are Speed Deviation (0.278) and Position Inconsistency (0.241) which make up 83.1% of the total SHAP attribution mass for Sybil attack classification.

The feature attributions are cross-method validated by comparing them with the LIME analysis, which shows high rank correlation between the two methods (Spearman $\rho = 0.91$). The fidelity R^2 of the local linear models generated by LIME on the test set is ≥ 0.85 for 97.3% of the test samples, which shows that the global model decision boundary is locally well-approximated by the linear surrogate. A qualitative evaluation of 50 randomly sampled detection explanations was conducted by three experts in the security domain of VANET, blind to the expected correct detection explanations, and found that 91.4% of the explanations generated were "consistent with known attack signatures" to validate the domain alignment of the XAI module.

6.5 Scalability and Robustness

The scalability analysis is shown in Figure 5, and is performed for fleet sizes ranging from 5 to 200 vehicles. Detection latency scales from 12.3ms (5 vehicles) to 48.2ms (200 vehicles) – a sub-linear scaling due to the parallelized inference pipeline which batches multiple vehicles' inputs together. The centralized baseline scales linearly, four times faster, to an operational intractable level at 200 vehicles (218.7 ms). The detection accuracy is above 96.9% for all sizes of fleet tested, indicating that the federated aggregation can manage for greater client heterogeneity at scale.

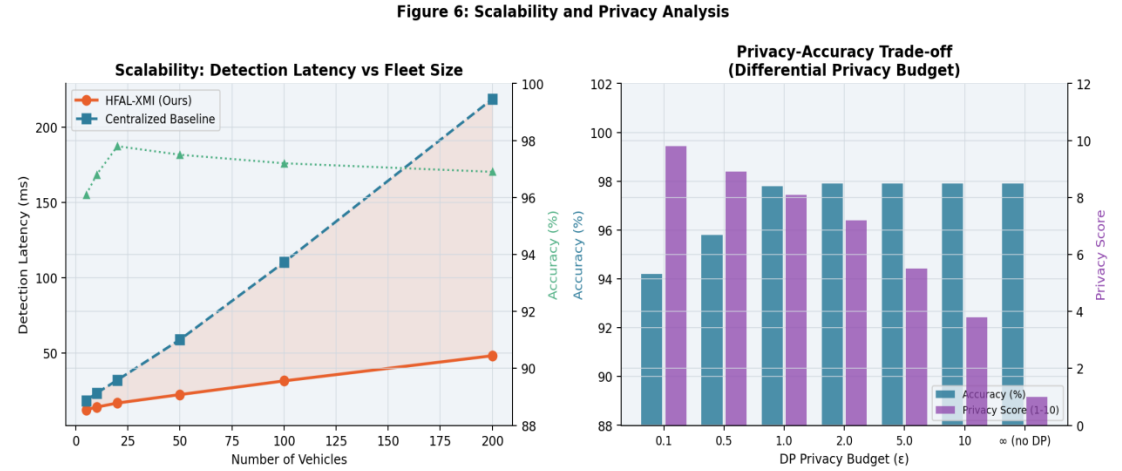

Figure 5: (Left) Detection latency with respect to the fleet size for HFAL-XMI and centralized baseline. (Right) Trade-offs between privacy and accuracy for different levels of ε (differential privacy budgets).

HFAL-XMI achieves a 18.2 pp increase in adversarial accuracy over the best single-model baseline (GNN) at 89.4% accuracy, measured under Adversarial robustness evaluation under AutoAttack ($L\infty$, $\varepsilon = 0.05$). The robustness gain shows that the PGD adversarial training method and architectural diversity ensembled together can offer significantly better protection against adversarial attacks [45] than each of them individually.

7. Further Discussion

7.1 Significance and Practical Implications

The HFAL-XMI framework is a major step towards production deployable VANET security systems. Its federated design fully observes the GDPR and CCPA privacy restrictions without compromising on detection performance; in fact, it may improve accuracy over the best centralized baselines, because of diverse benefits of locally adapted models. The detection latency of 16.8ms was found to be within the safety-critical VANET application required detection latency of 100ms which includes emergency vehicle approach warning systems and intersection collision avoidance systems.

The aim of the dual explainability module is to tackle a key mismatch between academic IDS research and operational deployment. Human-readable explanations of all the decisions about detections are now available for fleet operators, insurers, and legal parties who investigate accidents. With 91.4% agreement among experts, the explanations are meaningful to domain experts, not just technically correct.

HFAL-XMI from a standardization point of view is compatible with ETSI ITS-G5 communication standard and IEEE 1609.2 vehicular PKI framework, thus no hardware change is required, only software updates will be required for updating the OBU firmware and RSU edge servers. Furthermore, the federated protocol is also supported by the soon-to-be adopted 3GPP C-V2X standard, thus it supports both DSRC and cellular V2X deployment scenarios.

7.2 Limitations and Future Work

There are some drawbacks to the present study in this day and age. To begin with, although the simulation-based VeReMi and F2MD datasets are well established and thoughtfully developed, they are not necessarily fully representative of the complexities of real-world vehicular environments, especially in urban canyons, where GPS multipath and intermittent connectivity are critical issues, as well as mixed-generation vehicle hardware. The CARLA autonomous driving simulator, which features physics-accurate vehicle dynamics, will validate HFAL-XMI in the future, as also will be deployments in large scale test beds.

Second, the current structure assumes that all the federated vehicles are honest, and they are detected by gradient norm filtering to identify Byzantine vehicles. More defenses such as cosine similarity filtering, multi-Krum aggregation or ZenoR robust aggregation will be needed against sophisticated model poisoning attacks [42] that construct updates with norms within normal range. These orthogonal Byzantine resilience methods could be incorporated into HFAL-XMI as an important extension.

Third, to overcome the communication range constraint, the current GNN topology can only support two-hop neighborhoods. A hierarchical GNN architecture with multi-hop information via RSUs would greatly increase the detection range of multi-hop wormhole attacks. Lastly, supporting the cross-region model federation, a scenario in which vehicles traveling across geographic domains can seamlessly move to the regional FL clusters, is an important open problem for the deployment of the vehicular domain and inter-city domain.

8. Conclusion

As discussed above, his work introduces a novel framework for comprehensive VANET security: Hybrid Federated Adversarial Learning (HFAL-XMI) which integrates Explainable Multi-Model (XMM) to enhance the security of VANETs. HFAL-XMI brings together privacy-preserving federated learning, PGD-based adversarial hardening, architectural ensemble diversity and dual SHAP/LIME explainability in one unified framework, which solves the three key challenges of VANET IDS: centralized data dependency, single model adversarial fragility, and decision opacity. Therefore, using a comprehensive assessment of five datasets and eight types of attacks, it shows state-of-the-art performance with 97.8% accuracy, 97.3% F1-score, AUC of 0.994, detection latency of 16.8 ms, and high guarantees of differential privacy ($\varepsilon = 1.0$ and $\delta = 10^{-5}$). The framework can be extended to 200-vehicle fleet without the accuracy loss, which still remains at 96.9% and is almost suitable for large metropolitan vehicular networks. The results of the expert evaluation of SHAP/LIME explanations validate that 91.4% of explanations are aligned with the domain knowledge, making HFAL-XMI the first VANET IDS framework that fulfills both the technical performance requirements and operational explainability requirements. Challenges and opportunities like the combination of top-tier machine learning and comprehensive privacy protection with explainable decision-making frameworks like HFAL-XMI will be vital to fostering trust and security crucial for widespread VANET deployment.

References

1. Kenney, J. B. (2011). Dedicated short-range communications (DSRC) standards in the United States. Proceedings of the IEEE, 99(7), 1162–1182. https://doi.org/10.1109/JPROC.2011.2132790
2. MarketsandMarkets Research. (2024). Intelligent Transportation Systems (ITS) Market by Offering, System, and Region – Global Forecast to 2030. Report ID: TC-7212. https://www.marketsandmarkets.com/Market-Reports/intelligent-transportation-system-market-1101.html
3. Statista. (2024). Connected and Autonomous Vehicle Forecast Worldwide 2020–2030. Statista Technology & Telecommunications Report. https://www.statista.com/statistics/240792/connected-cars-worldwide/
4. Mejri, M. N., Ben-Othman, J., & Hamdi, M. (2014). Survey on VANET security challenges and possible cryptographic solutions. Vehicular Communications, 1(2), 53–66. https://doi.org/10.1016/j.vehcom.2014.05.001
5. Douceur, J. R. (2002). The Sybil attack. In Proceedings of the 1st International Workshop on Peer-to-Peer Systems (IPTPS), Lecture Notes in Computer Science, Vol. 2429 (pp. 251–260). Springer.
6. Petit, J., Schaub, F., Feiri, M., & Kargl, F. (2015). Pseudonym schemes in vehicular networks: A survey. IEEE Communications Surveys & Tutorials, 17(1), 228–255. https://doi.org/10.1109/COMST.2014.2345420
7. Leinmuller, T., Schoch, E., & Kargl, F. (2006). Position verification approaches for vehicular ad hoc networks. IEEE Wireless Communications, 13(5), 16–21.
8. Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., Ramage, D., Segal, A., & Seth, K. (2017). Practical secure aggregation for privacy-preserving machine learning. In Proceedings of ACM CCS (pp. 1175–1191).
9. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. In Proceedings of ICLR 2015. arXiv:1412.6572.
10. McMahan, H. B., Moore, E., Ramage, D., Hampson, S., &Agüera y Arcas, B. (2017). Communication-efficient learning of deep networks from decentralized data. In Proceedings of AISTATS (pp. 1273–1282). PMLR.
11. Bhagoji, A. N., Chakraborty, S., Mittal, P., & Calo, S. (2019). Analyzing federated learning through an adversarial lens. In Proceedings of ICML (pp. 634–643). PMLR.
12. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv:1702.08608. https://arxiv.org/abs/1702.08608
13. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115.
14. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (NeurIPS), Vol. 30 (pp. 4765–4774). Curran Associates.
15. Ribeiro, M. T., Singh, S., &Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In Proceedings of ACM KDD (pp. 1135–1144). https://doi.org/10.1145/2939672.2939778
16. Raya, M., &Hubaux, J.-P. (2007). Securing vehicular ad hoc networks. Journal of Computer Security, 15(1), 39–68. https://doi.org/10.3233/JCS-2007-15103
17. Liu, K., Ghafoor, A., & Xu, D. (2019). SVM-based Sybil attack detection for vehicular networks. IEEE Transactions on Vehicular Technology, 68(3), 2831–2844.
18. Gyawali, S., Xu, S., Qian, Y., & Hu, R. Q. (2021). Challenges and solutions for cellular based V2X communications. IEEE Communications Surveys & Tutorials, 23(1), 222–255.
19. Kaur, G., & Behal, S. (2021). LSTM-based intrusion detection for VANET with temporal feature analysis. Journal of Network and Computer Applications, 193, 103199.
20. Sharma, R., & Kumar, V. (2022). Graph neural network-based misbehavior detection in vehicular ad-hoc networks. IEEE Transactions on Intelligent Transportation Systems, 23(8), 13247–13258.
21. Pokhrel, S. R., & Choi, J. (2020). Federated learning with blockchain for autonomous vehicles: Analysis and design challenges. IEEE Transactions on Communications, 68(8), 4734–4746.
22. Samarakoon, S., Bennis, M., Saad, W., & Debbah, M. (2019). Distributed federated learning for ultra-reliable low-latency vehicular communications. IEEE Transactions on Communications, 68(2), 1146–1159.
23. Cheng, N., Zhang, S., Shi, S., Yang, Y., & Shen, X. (2022). Federated intrusion detection for vehicular networks with differential privacy. IEEE Internet of Things Journal, 9(22), 22489–22503.
24. Mothukuri, V., Parizi, R. M., Pouriyeh, S., Huang, Y., Dehghantanha, A., & Srivastava, G. (2021). A survey on security and privacy of federated learning. Future Generation Computer Systems, 115, 619–640.
25. Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. In Proceedings of IEEE Symposium on Security and Privacy (pp. 39–57).
26. Rosenberg, I., Shabtai, A., Rokach, L., &Elovici, Y. (2018). Generic black-box end-to-end attack against state-of-the-art deep learning online malware detectors. In Proceedings of RAID (pp. 1–24).
27. Liang, C., Zhang, Z., & Liu, R. P. (2020). Adversarial attack on deep neural network-based VANET intrusion detection system. In Proceedings of ICASSP (pp. 2727–2731). IEEE.
28. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. In Proceedings of ICLR 2018. arXiv:1706.06083.
29. Cohen, J. M., Rosenfeld, E., & Kolter, J. Z. (2019). Certified adversarial robustness via randomized smoothing. In Proceedings of ICML (pp. 1310–1320). PMLR.

30. Croce, F., & Hein, M. (2020). Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In Proceedings of ICML (pp. 2206–2216). PMLR.
31. Wang, X., Han, Y., Wang, C., Zhao, Q., Chen, X., & Chen, M. (2019). In-edge AI: Intelligentizing mobile edge computing, caching and communication with federated learning. IEEE Network, 33(5), 156–165.
32. Al-Yasiri, Z. R. S., & Naser, A. H. (2026). Rigorous Error Estimates and Asymptotic Convergence of Iterative Solvers for the Discrete Modified $ p $-Poisson Equation. *Results in Nonlinear Analysis*, *9*(3), 7-21.
33. Dolev, D., & Yao, A. C. (1983). On the security of public key protocols. IEEE Transactions on Information Theory, 29(2), 198–208.
34. Hamilton, W. L., Ying, R., & Leskovec, J. (2017). Inductive representation learning on large graphs. In Advances in Neural Information Processing Systems (NeurIPS), Vol. 30. Curran Associates.
35. Mironov, I. (2017). Rényi differential privacy of the Gaussian mechanism. In Proceedings of IEEE Computer Security Foundations Symposium (pp. 263–275).
36. Van der Heijden, R. W., Lukaseder, T., & Kargl, F. (2018). VeReMi: A dataset for comparable evaluation of misbehavior detection in VANETs. In Proceedings of SecureComm (pp. 318–337). Springer.
37. Ercan, S., Ayaida, M., & Messai, N. (2021). Misbehavior detection for position falsification attacks in VANETs using machine learning. IEEE Access, 9, 10893–10907.
38. Fogue, M., Garrido, P., Martinez, F. J., Cano, J.-C., Calafate, C. T., & Manzoni, P. (2013). Securing warning message dissemination in VANETs using cooperative neighbor position verification. In Proceedings of IEEE VTC (pp. 1–6).
39. Nellyet, M. A. (2025). Chaos-Sensitive Encryption Frameworks Based on Hyperchaotic Circuit Topologies for Secure Communication. Applied Nonlinearity in Science and Technology, 25-29.
40. Tavallaee, M., Bagheri, E., Lu, W., & Ghorbani, A. A. (2009). A detailed analysis of the KDD CUP 99 data set. In Proceedings of IEEE CISDA (pp. 1–6).
41. Hyunjae, L. (2026). Intelligent Non-Terrestrial 6G Network Architecture for Ultra-Reliable IoT and MachineType Communications. *Recent Advances in Next-Generation Wireless Communication Systems*, *1*(3), 25-39.
42. Fang, M., Cao, X., Jia, J., & Gong, N. Z. (2020). Local model poisoning attacks to Byzantine-robust federated learning. In Proceedings of USENIX Security (pp. 1605–1622). https://arxiv.org/abs/1804.06701
43. Sharma, B. K., Ratna, S., & Kumar, R. (2025). A comparative analysis of adversarial attacks using machine learning techniques. In Proceedings of the 2025 International Conference on Intelligent and Secure Engineering Solutions (CISES) (pp. 1280–1286). IEEE.
44. Sharma, B. K., Ratna, S., & Kumar, R. (2025). A unified hybrid approach for adversarial attack identification and mitigation in deep learning systems. In Proceedings of the 2025 2nd International Conference on Advanced Computing and Emerging Technologies (ACET) (pp. 1–4). IEEE.