electronics

MDPI

Article

ExplainableAI for Federated Learning-Based Intrusion Detection Systems in Connected Vehicles

Ramin Taheri [1], Raheleh Jafari [2,*], Alexander Gegov [1,3], Farzad Arabikhan [1] and Alexandar Ichtev [3]

[1] School of Computing, University of Portsmouth, Portsmouth PO1 3HE, UK; ramin.taheri@myport.ac.uk (R.T.); alexander.gegov@port.ac.uk (A.G.); farzad.arabikhan@port.ac.uk (F.A.)
[2] School of Design, University of Leeds, Leeds LS2 9JT, UK
[3] Department of Systems and Control, Technical University of Sofia, 1000 Sofia, Bulgaria; alexander.gegov@tu-sofia.bg (A.G.); ichtev@tu-sofia.bg (A.I.)
* Correspondence: r.jafari@leeds.ac.uk

Abstract

Connected and autonomous vehicles, along with the expanding Internet of Vehicles (IoV), are increasingly exposed to complex and evolving cyberattacks. Consequently, Intrusion Detection Systems (IDS) have become a vital component of modern vehicular cybersecurity. Federated Learning (FL) enables multiple vehicles to collaboratively train detection models while keeping their local data private, providing a decentralized alternative to traditional centralized learning. Despite these advantages, FL-based IDS frameworks remain vulnerable to attacks. To address this vulnerability, we propose an explainable federated intrusion detection framework that enhances both the security and interpretability of IDS in connected vehicles. The framework employs a Deep Neural Network (DNN) within a federated setting and integrates explainability through the Shapley Additive Explanations (SHAP) method. This Explainable Artificial Intelligence (XAI) component identifies the most influential network features contributing to detection decisions and assists in recognizing anomalies arising from malicious or corrupted clients. Experimental validation on the CICEVSE2024 and CICIoV2024 vehicular datasets demonstrates that the proposed system achieves high detection accuracy. Moreover, the XAI module improves transparency and enables analysts to verify and understand the model's decision-making process. Compared with both centralized IDS models and conventional federated approaches without explainability, the proposed system delivers comparable performance, stronger resilience to attacks, and significantly enhanced interpretability. Overall, this work demonstrates that integrating FL with XAI provides a privacy-preserving and trustworthy approach for intrusion detection in connected vehicular networks.

Keywords: explainableAI; federated learning; intrusion detection systems; connected vehicles

Academic Editor: Yu-an Tan

Received: 25 October 2025
Revised: 12 November 2025
Accepted: 13 November 2025
Published: 18 November 2025

Citation: Taheri, R.; Jafari, R.; Gegov, A.; Arabikhan, F.; Ichtev, A. Explainable AI for Federated Learning-Based Intrusion Detection Systems in Connected Vehicles. *Electronics* **2025**, *14*, 4508. https://doi.org/10.3390/electronics14224508

1. Introduction

Connected vehicles (CVs) rely on Internet of Things (IoT) technologies and Vehicle-to-Everything (V2X) communications to exchange data, enhance driving coordination, and improve road safety [1,2]. These vehicles integrate numerous sensors, processing units, and infotainment systems that continuously share information with nearby vehicles and cloud infrastructures [3], but while these capabilities enhance efficiency and situational awareness, they simultaneously expand the potential attack surface. Sophisticated adversaries, including Advanced Persistent Threats (APTs), can exploit unknown software

vulnerabilities or firmware flaws, making early detection particularly challenging. As vehicles become increasingly autonomous, resilient Intrusion Detection Systems (IDSs) are essential to prevent service disruption or physical damage [3].

Traditional IDS architectures rely on centralized Machine Learning (ML) models that aggregate raw traffic data at a central server for analysis. Although effective in static networks, this approach presents several challenges in vehicular environments, including high latency, excessive bandwidth usage, privacy leakage, and limited scalability. Moreover, deep learning-based centralized IDSs often struggle to adapt to non-stationary, imbalanced, and dynamically changing vehicular data distributions [4,5].

Federated Learning (FL) has recently emerged as a promising paradigm for decentralized model training that preserves data privacy. In FL, each client trains its local model independently and transmits only model parameters to a central server for aggregation. This collaborative strategy eliminates the need to share raw data and supports model adaptation to non-identically distributed (non-IID) data across clients [6,7]. Several FL-based IDS frameworks have reported encouraging results; for example, PF-DAPTIV [3] achieved over 95% detection accuracy against advanced threats in IoT-enabled vehicles while maintaining data confidentiality. Despite these achievements, federated IDS frameworks still face open challenges, including communication overhead, model aggregation security, and robustness under heterogeneous client conditions.

In safety-critical domains such as autonomous vehicles, explainability has become as important as accuracy. Regulatory initiatives like the EU AI Act emphasize the need for transparent and trustworthy Artificial Intelligence (AI) systems. Muhammad et al. [8] highlight that explainability is crucial because black-box IDSs hinder validation, trust, and root-cause analysis. Their L-XAIDS framework incorporates LIME and ELI5 to generate both local and global explanations, achieving 85% accuracy on the UNSW-NB15 dataset while identifying key traffic features. These studies underline that integrating explainable artificial intelligence (XAI) can improve model reliability, regulatory compliance, and analyst confidence in IDS decisions.

However, most explainable IDS efforts have been developed under centralized learning paradigms. Few studies consider the combined challenges of federated optimization, heterogeneous data, and explainability in vehicular contexts. The iterative aggregation process in FL can obscure the contribution of individual clients, complicating anomaly attribution to specific vehicles or sensors. Moreover, computationally intensive XAI techniques such as SHAP or LIME can impose substantial overhead on resource-constrained vehicular nodes. Therefore, there is a pressing need for FL-based IDS frameworks that integrate lightweight, client-aware explainability mechanisms, enabling both transparency and operational efficiency in connected vehicle networks.

1.1. Contributions

The proposed **XAI-FL-IDS** framework introduces three key innovations that advance the state of the art in federated intrusion detection for connected and autonomous vehicles. Each contribution is designed to enhance security, interpretability, and adaptability across heterogeneous vehicular networks:

1. **V2I Federated IDS:** We develop a unified FL-based intrusion detection framework that secures both in-vehicle (Controller Area Network—CAN) and external (IP-based) communications. By jointly modeling internal and external data streams, the proposed dual-layer design captures a complete spectrum of threat behaviors spanning in-vehicle and vehicle-to-infrastructure (V2I) interactions.
2. **Federated Explainability Aggregation:** A novel explainability aggregation mechanism is introduced to combine client-side SHAP and LIME explanations into a global

interpretability map. Unlike existing studies that apply explainability post hoc, our approach embeds XAI computation directly into the federated training process. This integration allows continuous learning of globally significant features and provides transparency during model convergence.

3. **Personalized and Adaptive FL-XAI Framework:** Each participating vehicle maintains a personalized model variant that dynamically adapts to its specific operational and communication conditions. The central server aggregates explanation maps from all clients to guide personalized updates, ensuring context-aware threat detection and consistent interpretability across diverse vehicular environments.

1.2. Organization of the Paper

The remainder of this paper is structured as follows. Section 2 reviews existing literature on IDS, FL in the Internet of Vehicles (IoV), and explainable artificial intelligence (XAI) techniques for security-critical domains. Section 3 introduces the proposed **XAI-FL-IDS** framework, detailing its system architecture, explainability integration, and key methodological contributions. Section 4 describes the experimental setup, including dataset selection (CICIoVdataset 2024 and CICEVSE2024), preprocessing procedures, model configurations, and evaluation metrics. Section 5 presents and analyzes the experimental outcomes, focusing on detection accuracy, communication efficiency, and interpretability performance. This Section also discusses the broader implications of the results, outlines current limitations, and suggests directions for future research. Finally, Section 6 concludes the paper and summarizes its primary findings and contributions.

2. Background and Related Work

Connected and autonomous vehicles (CAVs) are becoming increasingly susceptible to cyber threats due to the growing complexity of their wireless interfaces and in-vehicle communication networks. Ensuring the security of internal communication channels, such as the Controller Area Network (CAN) bus, is critical to maintaining vehicle safety and operational integrity, particularly against spoofing, replay, and denial-of-service (DoS) attacks [9,10]. Traditional IDSs based on centralized ML techniques have demonstrated strong detection capabilities; however, their reliance on aggregating raw vehicular data at a central server raises significant privacy, scalability, and communication challenges. The continuous transmission of sensitive sensor and communication data to the cloud not only introduces the risk of privacy leakage but also leads to excessive bandwidth consumption and latency, which are unsustainable for large-scale vehicular fleets [11].

To overcome these limitations, recent research has turned toward FL as a decentralized paradigm for training intrusion detection models collaboratively across multiple vehicles without exposing raw data. In this section, we review the recent literature on FL-based IDS frameworks designed for CAVs and discuss how **Explainable Artificial Intelligence (XAI)** techniques have been incorporated to enhance the interpretability, transparency, and trustworthiness of these systems.

2.1. FL in IoT and Vehicular Networks

FL has recently emerged as a promising paradigm for collaborative anomaly detection among connected vehicles without the need to share raw sensor or communication data [12]. In an FL-based intrusion detection framework, each vehicle (or edge node) independently trains a local detection model using its own driving data, such as CAN bus logs or V2X traffic traces, and periodically transmits only the model parameters or gradients to a central aggregator. By leveraging this decentralized learning process, FL enables collective

knowledge sharing across vehicles while preserving privacy and reducing the bandwidth burden associated with centralized architectures [11].

Several studies have demonstrated the effectiveness of FL in vehicular cybersecurity. Huang et al. [13] proposed a federated IDS for the IoV using a lightweight MobileNet-Tiny CNN, achieving over 98% detection accuracy with minimal latency on resource-constrained vehicular devices. Similarly, Xie et al. [14] introduced IoV-BCFL, a hybrid system that integrates FL with blockchain to secure parameter exchanges among vehicles, thereby improving both trust and auditability of distributed model updates.

Further research has explored optimizing FL under the non-IID and dynamic data distributions typical of vehicular networks. Almansour et al. [15] developed an adaptive personalized FL framework employing a depthwise convolutional bottleneck network to address data heterogeneity among vehicles. Their approach improved detection precision by up to 5% compared with standard FedAvg and FedProx algorithms, while also achieving significant gains in recall and F1-score on benchmark automotive attack datasets. Collectively, these works show that FL-based IDS frameworks can outperform both centralized and fully local approaches by offering broader attack coverage, faster response times, and stronger privacy protection for drivers and passengers [11,12].

Nevertheless, deploying federated IDSs in real-world vehicular environments remains challenging. Vehicles participating in the FL process often exhibit highly diverse data distributions due to variations in driving conditions, routes, and exposure to attacks. Such heterogeneity can impede global model convergence and degrade detection accuracy if not properly mitigated [15]. In addition, the reliance on a central server for parameter aggregation can become a performance bottleneck. To address this issue, recent studies have begun investigating asynchronous and communication-efficient FL protocols that reduce synchronization overhead and improve scalability in bandwidth-limited vehicular networks [12].

2.2. Explainable AI in Security-Critical and Federated Systems

In parallel with advances in detection performance and data privacy, there is increasing recognition that explainability is a crucial requirement for intelligent automotive security applications. The decisions made by an IDS in a connected or autonomous vehicle—such as flagging a sensor message as malicious—can have direct consequences for vehicle control and passenger safety. Accordingly, a wide range of stakeholders, including manufacturers, engineers, and drivers, demand that AI-driven security systems provide interpretable and trustworthy decisions. A recent survey by Nwakanma et al. [16] reports that explainable AI (XAI) techniques for connected vehicle security remain at an early stage of maturity but show strong potential to enhance user confidence and regulatory compliance. The authors categorize existing XAI-IDS approaches and identify key research challenges, including defining the appropriate granularity of explanations for vehicular data and ensuring that interpretability mechanisms themselves do not introduce new attack surfaces. In the broader IoT context, Chamola et al. [17] similarly emphasize that XAI for intrusion detection still faces open issues, such as the absence of ground truth for explaining "why" an alert is triggered and the lack of standardized methods for evaluating explanation quality in complex network environments.

Despite these challenges, several studies have begun integrating XAI into IDS frameworks, including those based on FL. One line of research employs feature-attribution techniques, such as SHAP and LIME, to identify which input features (e.g., CAN bus signals or network packet fields) most strongly influence the IDS's classification decisions. Saheed and Chukwuere [18] introduced **XAIEnsembleTL-IoV**, an explainable ensemble transfer learning framework for zero-day attack detection in IoV environments. Their system combines multiple deep models with SHAP-based attribution analysis to detect novel

botnet attacks while simultaneously providing interpretable explanations of each alert—for instance, revealing the message patterns or timing irregularities most indicative of the attack. Similarly, Alabbadi and Bajaber [19] proposed a real-time IDS for IoT data streams that incorporates LIME-based explanations to interpret model predictions on the fly. Their results demonstrate that even under high-throughput streaming conditions, generating visual or textual explanations for IDS alerts—such as highlighting which sensor readings contributed most to a malicious classification—can significantly assist administrators in rapidly validating and responding to detected threats.

2.3. Explainability in FL and Open Challenges

Marrying explainability with FL represents an emerging research frontier in intelligent cybersecurity. The concept of federated explainable AI has only recently begun to attract attention, with a limited number of studies exploring its feasibility for intrusion detection. Fatema et al. [20] proposed one of the earliest frameworks, **FedXAI-IDS**, which integrates FL with local explainable AI mechanisms for decentralized intrusion detection. In their system, each participating client—such as an IoT device or connected vehicle—trains a deep neural IDS locally and computes Kernel SHAP values for its predictions, thereby producing instance-level explanations without sharing sensitive raw data with the central server. This design demonstrates the potential of achieving interpretability within a federated setting while maintaining data confidentiality.

However, an important gap remains in securing the explainability process itself. Just as FL models are susceptible to poisoning, an intelligent adversary could manipulate the explanation outputs—for example, by injecting falsified feature importance scores—to mislead human analysts or distort the global interpretability map. Current FL-based IDS frameworks, including those enhanced with XAI, have not explicitly addressed this threat, leaving the explanation generation and aggregation stages vulnerable to adversarial manipulation [16,18]. Strengthening the robustness of explainability mechanisms is therefore a crucial but largely overlooked research direction.

In summary, the literature reveals the convergence of three major trends shaping the future of connected vehicle cybersecurity: (1) the development of advanced IDSs for CAVs leveraging deep learning and large-scale vehicular data [9,10]; (2) the adoption of federated and distributed learning paradigms to enhance privacy and scalability [13–15]; and (3) the initial incorporation of explainability into intelligent security frameworks [16,18,20]. Despite this progress, substantial gaps persist at the intersection of these domains. Most FL-based vehicular IDSs emphasize detection accuracy and privacy preservation but pay limited attention to interpretability. Conversely, existing XAI-based IDS studies are predominantly designed for centralized environments and do not consider the adversarial dynamics or communication constraints inherent to federated networks.

To bridge this gap, **our proposed framework unifies FL and XAI within a single architecture and introduces attack-resilient mechanisms for both model aggregation and explanation generation**, thereby advancing the state of the art in privacy-preserving and trustworthy intrusion detection for connected and autonomous vehicles.

To provide a clearer comparison of existing approaches, Table 1 summarizes recent studies on federated and explainable intrusion detection systems. The table highlights the key techniques and main outcomes of each study, clarifying the position and advantages of the proposed framework in this work.

Table 1. Summary of related studies on FL- and XAI-based intrusion detection in connected and autonomous vehicles.

Reference	Technique/Approach	Main Outcome
Temitope et al. [11]	FL-based IDS for secure in-vehicle communication networks	Demonstrated effective privacy-preserving detection in V2X environments.
Huang et al. [13]	Federated IDS using lightweight MobileNet-Tiny CNN for IoV	Achieved >98% accuracy with minimal latency on edge devices.
Xie et al. [14]	IoV-BCFL: Blockchain-based Federated Learning for secure parameter exchange	Improved trust and auditability of federated model updates.
Almansour et al. [15]	Adaptive Personalized FL with depthwise convolutional bottleneck network	+5% precision gain vs. FedAvg; higher recall and F1-score.
Nwakanma et al. [16]	Survey on Explainable AI for connected vehicle security	Highlighted open challenges in defining explanation granularity and assessing interpretability quality.
Saheed and Chukwuere [18]	XAIEnsembleTL-IoV: Ensemble Transfer Learning with SHAP-based attribution	Detected zero-day botnet attacks and provided interpretable feature attributions.
Alabbadi and Bajaber [19]	Real-time IDS with LIME-based explanations	Enabled on-the-fly interpretability for detected anomalies.
Fatema et al. [20]	FedXAI-IDS: Federated Learning with SHAP-based explanations	Achieved interpretable, privacy-preserving intrusion detection with maintained confidentiality.
Proposed (This Work)	XAI-FL-IDS: Federated Learning with SHAP and LIME integration and attention-based interpretability	99.07% accuracy (CICEVSE2024) and 96.71% (CICIoV2024); low FPR, high F1-score, and enhanced model transparency.

Furthermore, in the literature we can see, post hoc explainability methods are widely used to interpret trained machine learning models by analyzing their decisions after training. These techniques, such as surrogate modeling, perturbation analysis, and counterfactual reasoning, are model-agnostic and can be applied to any black-box system to identify key features influencing predictions. However, post hoc methods often suffer from high computational overhead and limited fidelity, as the explanations are generated retrospectively and may not fully reflect the model's internal reasoning. In contrast, the proposed XAI-FL-IDS framework embeds explainability directly within the federated learning process, integrating SHAP and LIME during training rather than after model convergence. This integrated design enables real-time, client-level interpretability and consistent global transparency across clients, making it more suitable for time-sensitive and privacy-preserving applications such as intrusion detection in connected vehicles [21,22].

3. Proposed Method

The proposed **Explainable FL-based Intrusion Detection System (XAI-FL-IDS)** is developed for connected and autonomous vehicles operating within Vehicle-to-Infrastructure (V2I) communication environments. The framework aims to simultaneously enhance intrusion detection accuracy, preserve data privacy, and introduce interpretability into model predictions through the integration of Explainable Artificial Intelligence (XAI).

In this architecture, a *FL* setup is employed, where each vehicle functions as an independent local client. Instead of transmitting raw network or sensor data, each vehicle

trains a local intrusion detection model on its own dataset and periodically shares only model updates with the central federated server for aggregation. This decentralized learning approach mitigates privacy concerns and reduces communication overhead while improving generalization across diverse vehicular environments.

Each vehicle V_i maintains a local model M_i trained on its dataset D_i, which includes both in-vehicle CAN bus and external network flow data. At the beginning of each communication round r, the central server distributes the current global model $M_g^{(r)}$ to all participating clients. Each vehicle then performs local training using stochastic gradient descent (SGD) for E local epochs and computes attention weights $A_i^{(r)}$ to capture feature importance within its own data distribution. The local update process is expressed as Equation (1):

$$M_i^{(r+1)} = M_i^{(r)} - \eta \nabla \mathcal{L}(M_i^{(r)}, D_i) \tag{1}$$

where η denotes the learning rate and $\mathcal{L}$ represents the local loss function.

After completing local training, each client computes feature-level importance values using SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations) to produce interpretable insights into model predictions. These feature attributions are stored as local explanation maps $\Phi_i^{(r)}$. Both the model updates $\Delta M_i^{(r)}$ and the corresponding explanation maps $\Phi_i^{(r)}$ are securely transmitted to the federated server. The server aggregates all model updates using a weighted FedAvg strategy as Equation (2):

$$M_g^{(r+1)} = \sum_{i=1}^{N} \frac{|D_i|}{\sum_j |D_j|} \Delta M_i^{(r)} \tag{2}$$

and fuses the explanation maps through a federated knowledge aggregation mechanism defined as Equation (3):

$$\Phi_g^{(r+1)} = \text{Aggregate}\left(\{\Phi_i^{(r)}\}_{i=1}^{N}\right) \tag{3}$$

The resulting global explanation map $\Phi_g^{(r+1)}$ identifies the features that most significantly influence intrusion detection across all clients, forming a feedback loop that supports interpretability-aware model refinement and adaptive feature weighting.

3.1. Federated Explainability Aggregation Mechanism

To enhance interpretability during federated training, we define a global aggregation process that combines local SHAP and LIME explanations from all participating clients. Each client V_i generates a local explanation map $\Phi_i^{(r)} = [\phi_{i,1}^{(r)}, \phi_{i,2}^{(r)}, \ldots, \phi_{i,F}^{(r)}]$, where $\phi_{i,f}^{(r)}$ denotes the feature importance score for feature f computed by client i at communication round r. Because data distributions across clients can be heterogeneous, each explanation vector is first normalized to ensure consistency across scales as Equation (4):

$$\tilde{\phi}_{i,f}^{(r)} = \frac{|\phi_{i,f}^{(r)}|}{\sum_{j=1}^{F} |\phi_{i,j}^{(r)}|}. \tag{4}$$

The server then performs weighted aggregation based on the relative dataset size of each client, Equation (5):

$$\Phi_g^{(r+1)} = \frac{\sum_{i=1}^{N} w_i \tilde{\Phi}_i^{(r)}}{\sum_{i=1}^{N} w_i}, \quad \text{where } w_i = \frac{|D_i|}{\sum_{j=1}^{N} |D_j|}. \tag{5}$$

This weighted averaging ensures that clients contributing more data have proportionally greater influence on the global interpretability map. To reduce the effect of conflicting

or non-aligned feature importance values, a simple variance-based filter is applied, as Equation (6):

$$\tilde{\Phi}_g^{(r+1)}(f) = \begin{cases} \Phi_g^{(r+1)}(f), & \text{if } \mathrm{Var}\left(\tilde{\phi}_{i,f}^{(r)}\right) < \tau, \\ 0, & \text{otherwise,} \end{cases} \tag{6}$$

where τ is a small threshold controlling cross-client consistency.
The resulting global explanation map $\tilde{\Phi}_g^{(r+1)}$ captures the most stable and influential features across all clients, providing a unified interpretability view of the federated model. This aggregated explanation is then used in subsequent rounds to guide attention weighting and improve model transparency during training.

3.2. Explainable AI Integration

The integration of XAI within the proposed framework ensures that the system is not a black-box model but a transparent and interpretable IDS capable of providing insights into its decision-making rationale. The explainability mechanism operates across three complementary layers:

- **Attention-based interpretability:** During local training, attention modules assign dynamic weights to temporal and feature-level dependencies, enabling the visualization of influential CAN IDs and network attributes that drive the detection outcomes.
- **Local explanations:** Post-training, SHAP and LIME are applied to derive feature attribution scores and instance-level rationales. These local explanations help engineers understand the causes of detected anomalies, identify potential misclassifications, and mitigate false alarms.
- **Federated aggregation of explanations:** The server collects high-importance features from all clients' SHAP results and constructs a global interpretability map Φ_g. This federated map uncovers system-wide intrusion trends and cross-client feature relevance without requiring any raw data exchange.

The combination of these three explainability layers enables continuous interpretability across both local and global levels of the FL process, promoting transparency, accountability, and trust in intrusion detection for connected vehicular ecosystems.

Although both SHAP and LIME are employed at the client level, they serve complementary purposes within the proposed framework. SHAP provides global, model-level explanations by quantifying the contribution of each input feature to the prediction outcome, making it suitable for federated aggregation at the server. In contrast, LIME generates local, instance-specific explanations that assist operators in understanding individual detection decisions and potential misclassifications in real time.

From a computational perspective, SHAP is more accurate but computationally intensive due to the need for multiple model evaluations, whereas LIME offers a faster approximation with lower computational cost. Accordingly, LIME is implemented as an optional module for lightweight, on-demand interpretability at the client side, while SHAP forms the basis for constructing the federated global explanation map.

3.3. Architecture

The architecture illustrated in Figure 1 depicts the overall workflow of the proposed **Explainable FL-based Intrusion Detection System (XAI-FL-IDS)** for connected and autonomous vehicles. Each participating vehicle (denoted as V_1 through V_N) maintains its own local dataset composed of network traffic flows and CAN bus communication logs. These datasets are used to train local attention-based neural network models that not only

perform anomaly detection but also incorporate built-in interpretability by emphasizing the most influential features during the learning process.

Following local training, each client applies SHAP and LIME techniques to generate explanation vectors at both the global (feature importance) and instance-specific (sample-level) scales. These explanation vectors, together with the corresponding local model updates, are securely transmitted to the central federated server. The server then aggregates the received model updates to construct the next-round global model $M_g^{(r+1)}$ using a weighted averaging scheme and fuses the explanation vectors to create a comprehensive global explanation map.

Once aggregation is complete, the updated global model is redistributed to all clients for the next training round. This iterative process enables continuous improvement of the IDS while maintaining strict data confidentiality—since no raw information ever leaves the local devices. In addition to preserving privacy, the federated workflow improves overall detection accuracy through collaborative knowledge sharing and enhances transparency by embedding explainability into every stage of the learning and inference pipeline.

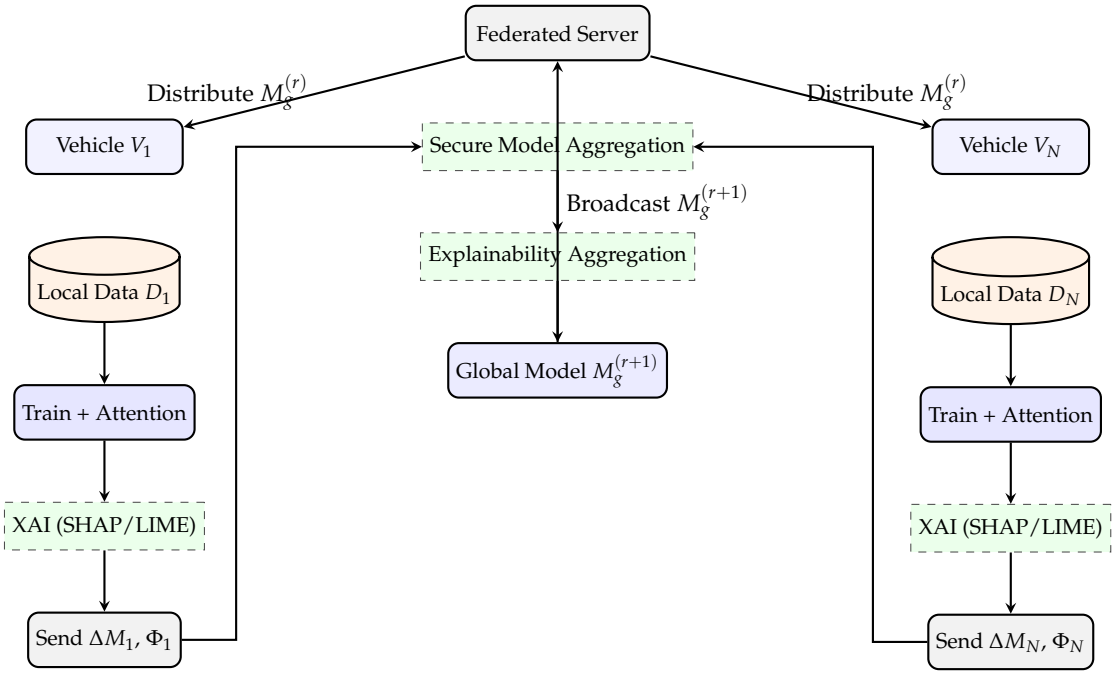

Figure 1. Compact workflow of the proposed XAI-FL-IDS.Each vehicle performs local training with attention-based interpretability, generates SHAP/LIME explanations, and sends model updates and explanation vectors to the server for aggregation.

Algorithm 1 outlines the proposed Explainable FL-based Intrusion Detection System (XAI-FL-IDS) for connected vehicles. The system operates over R communication rounds across a federation of N vehicles. Initially, the global model M_g is initialized by the server and distributed to all participating clients. Each vehicle V_i trains a local model $M_i^{(r)}$ on its private dataset D_i—composed of CICIoVdataset 2024 and CICEVSE2024 partitions—using stochastic gradient descent (SGD) over E local epochs. During training, an attention mechanism is applied to highlight feature-level importance, enhancing the interpretability of the model. After training, the vehicle computes local explanation maps $\Phi_i^{(r)}$ using SHAP for global feature attribution and optionally applies LIME for instance-level interpretation. Each client then transmits its model update $\Delta M_i^{(r)}$ and explanation map $\Phi_i^{(r)}$ to the server. The server performs secure aggregation of the model updates to produce an updated global model $M_g^{(r+1)}$, and aggregates explanation vectors to form a federated explanation

map $\Phi_g^{(r+1)}$. The global model is then refined using the aggregated explanations and redistributed to clients. This iterative process continues until convergence, resulting in a final global model $M_g^{(R)}$ and interpretable feature map $\Phi_g^{(R)}$.

Algorithm 1 Explainable FL-based Intrusion Detection System for Connected Vehicles (XAI-FL-IDS)

Require: Number of vehicles N, rounds R, local epochs E, learning rate η, initial global model M_g

Ensure: Trained global model $M_g^{(R)}$ and global explanation map $\Phi_g^{(R)}$

1: Initialize global model M_g
2: **for** each round $r = 1$ to R **do**
3: Server sends current global model $M_g^{(r)}$ to all vehicles V_i, where $i \in [1, N]$
4: **for all** vehicles V_i in parallel **do**
5: Load local dataset D_i from CICIoVdataset 2024 and CICEVSE2024 partitions
6: Initialize local model $M_i^{(r)} \leftarrow M_g^{(r)}$
7: **for** each local epoch $e = 1$ to E **do**
8: Train $M_i^{(r)}$ on D_i using SGD with learning rate η
9: Compute attention weights $A_i^{(r)}$ to identify influential features
10: **end for**
11: Apply SHAP to compute local feature importance $\Phi_i^{(r)}$
12: Apply LIME for instance-level explanation (optional)
13: Send model update $\Delta M_i^{(r)}$ and explanation map $\Phi_i^{(r)}$ to server
14: **end for**
15: Server aggregates model updates:

$$M_g^{(r+1)} \leftarrow \text{Aggregate}\left(\{\Delta M_i^{(r)}\}_{i=1}^{N}\right)$$

16: Aggregate local explanations:

$$\Phi_g^{(r+1)} \leftarrow \text{Aggregate}\left(\{\Phi_i^{(r)}\}_{i=1}^{N}\right)$$

17: Update global model with attention-guided weighting based on $\Phi_g^{(r+1)}$
18: **end for**
19: **return** $M_g^{(R)}$, $\Phi_g^{(R)}$

Algorithm Description: Algorithm 1 outlines the operational flow of the proposed Explainable FL-based Intrusion Detection System (XAI-FL-IDS) for connected vehicles. Below is a detailed explanation of each step:

- **Line 1:** The global model M_g is initialized at the server. This model will be iteratively refined through federated training.
- **Lines 2–3:** For each communication round $r \in [1, R]$, the current global model $M_g^{(r)}$ is distributed to all N client vehicles V_i participating in training.
- **Lines 4–5:** Each vehicle loads its private local dataset D_i, which is composed of samples from CICIoVdataset 2024 and CICEVSE2024. The local model $M_i^{(r)}$ is initialized using the global model received from the server.
- **Lines 6–8:** Each vehicle trains the local model $M_i^{(r)}$ for E epochs using stochastic gradient descent (SGD) with a fixed learning rate η. During training, attention weights $A_i^{(r)}$ are computed to capture the most influential input features, enabling built-in interpretability.

- **Line 9:** After training, SHAP (Shapley Additive Explanations) is applied to the local model to compute global feature importance scores, resulting in an explanation map $\Phi_i^{(r)}$ for each client.
- **Line 10:** Optionally, LIME is applied to generate local, instance-specific explanations to support case-level interpretability. In our experiments, we applied this one.
- **Line 11:** Each client sends its model update $\Delta M_i^{(r)}$ and local explanation map $\Phi_i^{(r)}$ to the central server.
- **Line 12:** The server aggregates all client model updates to construct a new global model $M_g^{(r+1)}$. A robust aggregation rule, such as Krum, may be used to enhance resistance to adversarial updates.
- **Line 13:** In parallel, the server aggregates local explanation maps $\Phi_i^{(r)}$ to generate a global interpretability map $\Phi_g^{(r+1)}$, summarizing the most important features across all clients.
- **Line 14:** The server uses the global explanation map to guide attention-based weighting or refinement of the next-round global model.
- **Line 15:** After all R communication rounds, the algorithm returns the final global model $M_g^{(R)}$ and its associated global explanation map $\Phi_g^{(R)}$.

This algorithm enables distributed, privacy-preserving training while providing human-interpretable insights into intrusion decisions, making it suitable for deployment in real-time, resource-constrained connected vehicle environments.

3.4. Time Complexity Analysis

The time complexity of the proposed XAI-FL-IDS framework is primarily influenced by three components: local model training, explainability generation, and server-side aggregation. In each communication round $r \in \{1, \ldots, R\}$, every client trains a deep neural network of complexity $\mathcal{O}(C)$ over a local dataset of size n_i for E local epochs. This yields a per-client training cost of $\mathcal{O}(E \cdot n_i \cdot C)$. To provide model transparency, each client also computes SHAP or LIME-based explanations, which introduce an additional overhead of approximately $\mathcal{O}(n_i \cdot F)$, where F is the number of input features.

On the server side, aggregation of model updates and explanation vectors across N clients requires $\mathcal{O}(N \cdot C)$ and $\mathcal{O}(N \cdot F)$ time, respectively. The total time complexity per round is thus dominated by client-side operations and can be approximated as $\mathcal{O}(N \cdot (E \cdot n_i \cdot C + n_i \cdot F))$. The framework remains efficient by using a lightweight model and SHAP approximations, making it suitable for deployment in resource-constrained environments.

4. Experimental Setup

This section outlines the experimental setup used to evaluate the proposed XAI-FL-IDS framework, including the local model architecture, datasets, computing infrastructure, programming environment, and implementation libraries.

4.1. Local Model

In the proposed XAI-FL-IDS framework, each client (vehicle) employs a fully connected deep neural network (DNN) as its local intrusion detection model. The architecture, illustrated in Figure 2, consists of four sequential linear (dense) layers interleaved with non-linear activation functions.

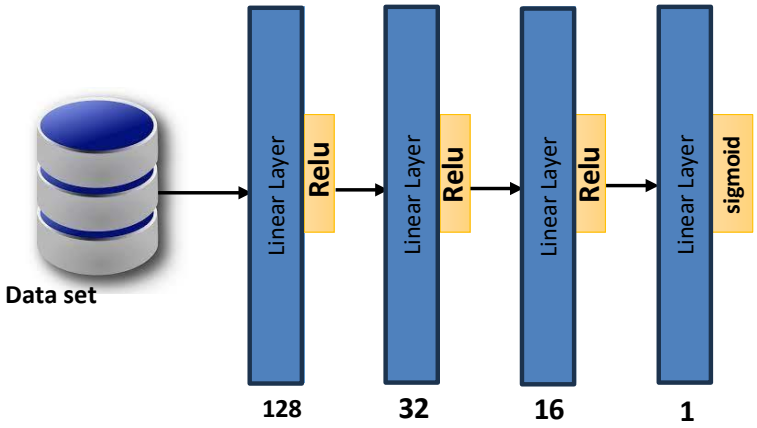

Figure 2. DNN architecture as Local Model in FL.

Specifically, the model includes:

- An input layer of size 128 corresponding to the number of features extracted from each dataset sample.
- A first hidden layer with 32 neurons and ReLU activation.
- A second hidden layer with 16 neurons and ReLU activation.
- An output layer with a single neuron and sigmoid activation, producing a binary classification output indicating benign or malicious traffic.

All hidden layers are initialized using Xavier uniform initialization and trained using the Adam optimizer with binary cross-entropy loss. The ReLU activation functions enhance non-linearity while maintaining computational efficiency, and the final sigmoid layer enables probabilistic output for classification. This architecture is lightweight enough for deployment on client-side hardware, while still expressive enough to capture complex traffic patterns across the CICIoVdataset 2024 and CICEVSE2024 datasets.

We simulate a FL environment with **10 participants**, each representing a separate vehicle with its own local data partition. To ensure robustness against adversarial participants or noisy updates, we use the **Krum** aggregation rule at the central server instead of FedAvg. Krum is specifically designed to resist Byzantine failures by selecting the client update that is most similar to the majority of other updates, thus filtering out potential poisoned or anomalous gradients. This strategy increases the fault tolerance and reliability of the federated IDS under realistic attack conditions.

4.2. Simulation Environment

The implementation of the XAI-FL-IDS framework was carried out using a modern, Python-based software stack optimized for scalable and privacy-preserving ML. All experiments were conducted using **Python 3.11**, which provided compatibility with the latest versions of relevant deep learning and FL libraries. The local DNN models were implemented and trained using PyTorch 2.5.1. To manage the FL workflow, we adopted the **Flower (FLWR)** framework. All experiments were conducted on a computing platform equipped with an **NVIDIA Tesla V100 GPU (16GB VRAM)**, **64 GB of RAM**, and a multi-core CPU.

4.3. Datasets Description

We evaluate the proposed XAI-FL-IDS framework using two recent and complementary intrusion detection datasets: **CICEVSE2024** and **CICIoV2024**. The CICEVSE2024 dataset captures simulated external network attacks on connected vehicles, including DDoS, infiltration, and port scanning scenarios across vehicle-to-infrastructure (V2I) and

vehicle-to-cloud (V2N) communication channels. In contrast, CICIoV2024 focuses on in-vehicle security, containing CAN bus traffic annotated with both benign and attack samples, such as spoofing and denial-of-service (DoS) events targeting ECU communications.

4.4. Evaluation Metrics

To evaluate the performance of the proposed XAI-FL-IDS framework, several standard machine learning metrics commonly used in intrusion detection are employed. In this context, TP (True Positives) represent attack samples correctly classified as attacks; TN (True Negatives) denote benign samples correctly classified as normal; FP (False Positives) correspond to benign samples incorrectly classified as attacks; and FN (False Negatives) refer to attack samples incorrectly classified as benign.

The following metrics quantify the model's classification performance from multiple perspectives and ensure a robust and interpretable evaluation under class imbalance and federated learning conditions:

- **Accuracy**: The ratio of correctly classified samples to the total number of samples, while commonly used, accuracy can be misleading in imbalanced datasets. Accuracy is calculated as Equation (7):

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{7}$$

- **Precision**: The ratio of true positive predictions to all predicted positive instances. High precision indicates a low false alarm rate and is crucial for reducing unnecessary alerts in IDS. Precision is calculated as Equation (8):

$$\text{Precision} = \frac{TP}{TP + FP} \tag{8}$$

- **Recall (Detection Rate)**: The ratio of true positive predictions to all actual positive instances. High recall means fewer missed attacks, which is critical in safety-critical systems like connected vehicles. Recall is calculated as Equation (9):

$$\text{Recall} = \frac{TP}{TP + FN} \tag{9}$$

- **F1-Score**: The harmonic mean of precision and recall. It provides a single metric that balances false positives and false negatives, especially useful when the classes are imbalanced. F1-Score is calculated as Equation (10):

$$\text{F1-score} = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}} \tag{10}$$

- **False Positive Rate (FPR)**: Measures the proportion of benign traffic incorrectly flagged as malicious. FPR is calculated as Equation (11):

$$\text{FPR} = \frac{FP}{FP + TN} \tag{11}$$

- **Training Convergence and Round-wise Accuracy**: In FL, we also track model accuracy and loss across communication rounds to assess convergence behavior and training stability.

All metrics are computed separately for the CICEVSE2024 and CICIoV2024 datasets to evaluate performance on both external and internal vehicular intrusion scenarios. Macro-averaging is used when applicable to account for class imbalance.

5. Results and Discussion

This section presents and analyzes the experimental results of the proposed XAI-FL-IDS framework in comparison with three baselines: centralized IDS, FL without explainability, and local-only models. The evaluation was conducted on two real-world benchmark datasets: CICEVSE2024 and CICIoV2024. Our analysis emphasizes model performance, convergence behavior, and the effect of explainability integration.

5.1. Training Convergence and Learning Dynamics

Figures 3 and 4 illustrate the training dynamics over 30 communication rounds on CICEVSE2024 and CICIoV2024 datasets, respectively. Each figure includes accuracy (left) and training loss (right) curves comparing XAI-FL-IDS with FL without XAI.

In Figure 3, the proposed XAI-FL-IDS demonstrates faster convergence and superior accuracy throughout training. It surpasses 90% accuracy by round 10 and stabilizes at around 99% after round 20, slightly outperforming FL without XAI. Moreover, the loss curve reveals smoother and more stable learning in our method, with a sharp decline from round 16 onward, indicating efficient optimization and model generalization.

Figure 4 shows a similar trend on CICIoV2024. XAI-FL-IDS again reaches higher accuracy earlier in training, maintaining a consistent margin above the baseline throughout most rounds. Although the final accuracy gap narrows, our method consistently exhibits lower training loss and more stable progression. These observations suggest that integrating explainability not only enhances interpretability but also contributes to optimization stability and faster convergence.

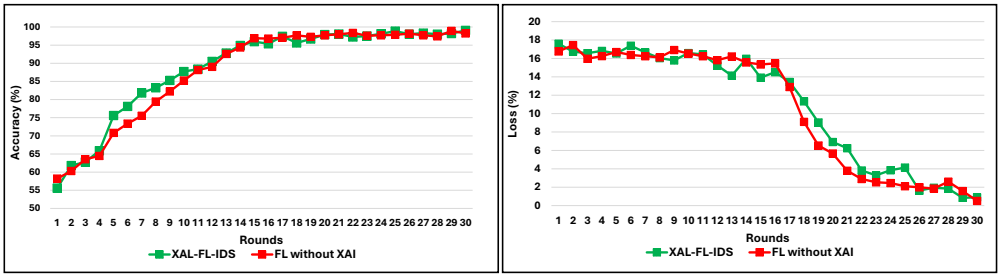

Figure 3. Accuracy and loss per round on CICEVSE2024. XAI-FL-IDS shows faster convergence.

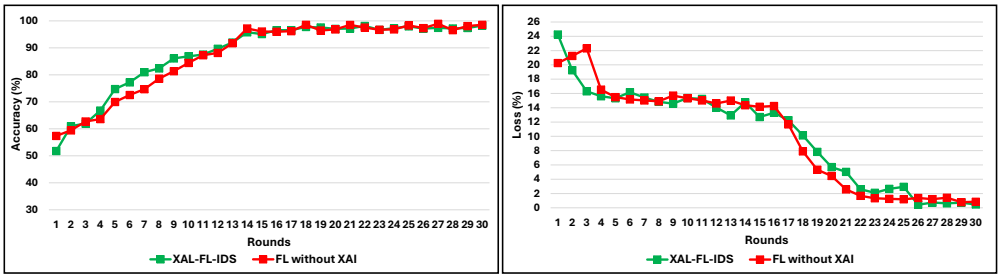

Figure 4. Accuracy and loss per round on CICIoV2024. XAI-FL-IDS maintains higher accuracy in early rounds.

5.2. Classification Performance Comparison

Tables 2 and 3 summarize the quantitative performance of all models on both datasets.

In Table 2, the proposed XAI-FL-IDS achieves 99.07% accuracy on CICEVSE2024 with a low false positive rate (FPR) of 1.31%, closely approaching the centralized IDS (99.26%) while offering significantly better privacy. It also outperforms FL without XAI (98.23%) and shows better FPR than both the centralized model and FL. On CICIoV2024, XAI-FL-IDS achieves 96.71% accuracy and 2.67% FPR, improving upon the FL baseline and remaining competitive with the centralized and local models. Notably, our method reduces false

positives compared to FL without XAI by over 0.5%, demonstrating that integrating XAI helps suppress misleading activations.

Table 2. Accuracy and False Positive Rate (FPR) comparison of IDS models on CICEVSE2024 and CICIoV2024 datasets.

Dataset	Model	Accuracy (%)	FPR (%)
CICEVSE2024	Centralized IDS [23]	99.26	1.81
	FL without XAI [24]	98.23	2.11
	Local Model Only	99.31	1.07
	XAI-FL-IDS (Proposed)	99.07	1.31
CICIoV2024	Centralized IDS [23]	97.71	2.33
	FL without XAI [24]	96.14	3.21
	Local Model Only	97.53	2.19
	XAI-FL-IDS (Proposed)	96.71	2.67

Table 3. Precision, Recall, and F1-score comparison of IDS models on CICEVSE2024 and CICIoV2024 datasets.

Dataset	Model	Precision (%)	Recall (%)	F1-Score (%)
CICEVSE2024	Centralized IDS [23]	97.19	97.15	97.15
	FL without XAI [24]	97.24	97.11	97.17
	Local Model Only	94.8	95.5	95.2
	XAI-FL-IDS (Proposed)	98.9	98.4	98.5
CICIoV2024	Centralized IDS [23]	99.9	99.9	99.9
	FL without XAI [24]	99.9	99.9	99.9
	Local Model Only	97.63	98.11	98.02
	XAI-FL-IDS (Proposed)	98.24	97.92	98.56

Table 3 further supports these results with precision, recall, and F1-score. On CICEVSE2024, XAI-FL-IDS reaches 98.9% precision, 98.4% recall, and 98.5% F1-score, exceeding all other models. This indicates a strong balance between sensitivity and specificity. On CICIoV2024, although precision and recall values are near-perfect for all models due to the simpler CAN bus attack patterns, XAI-FL-IDS achieves a 98.56% F1-score, maintaining robustness and consistency. Importantly, our method outperforms the local-only model in both datasets despite having no direct access to centralized data, validating the strength of federated collaboration combined with local explainability. These results confirm that the proposed XAI-FL-IDS not only matches or surpasses conventional models in accuracy but also achieves superior interpretability, reduced false positives, and better optimization dynamics—all critical factors for real-world deployment in connected and autonomous vehicles.

The experimental findings provide compelling evidence that the proposed **XAI-FL-IDS** framework effectively fulfills its primary objectives of preserving privacy, ensuring interpretability, and delivering robust intrusion detection across heterogeneous vehicular environments. This section discusses the observed outcomes in relation to the framework's design contributions and emphasizes its advantages over both conventional centralized IDSs and existing federated baselines.

First, the convergence plots for both datasets (Figures 3 and 4) demonstrate that XAI-FL-IDS not only accelerates model learning but also achieves faster and more stable convergence compared to standard FL without explainability. On the **CICEVSE2024** dataset, the proposed system surpasses 90% accuracy within the first ten communication rounds and attains near-optimal accuracy (approximately 99%) by round twenty. The training

loss declines more sharply than the baseline beginning from round sixteen, indicating improved generalization and smoother optimization behavior. A similar trend is observed for **CICIoV2024**, where XAI-FL-IDS consistently maintains higher accuracy and lower loss during both the early and middle training phases. These patterns confirm that incorporating XAI—particularly feature attribution through SHAP—enhances optimization stability by aligning learning signals with semantically meaningful and security-relevant features.

Second, the quantitative classification results reinforce these observations. On **CI-CEVSE2024**, XAI-FL-IDS achieves 99.07% accuracy and a 98.5% F1-score, outperforming FL without XAI by 0.84% in accuracy and 1.3% in F1-score while also reducing false positive rates. Notably, the framework even surpasses the centralized IDS baseline in terms of FPR, achieving 1.31% compared to 1.81%. This demonstrates that the integration of XAI within the aggregation process does not compromise detection effectiveness; instead, it enhances robustness by mitigating the effects of misleading or poisoned updates. For **CICIoV2024**, all evaluated models perform competitively due to the structured and deterministic nature of CAN bus data. Nevertheless, XAI-FL-IDS maintains a high F1-score (98.56%) with reduced training volatility, underscoring its suitability for real-time, resource-constrained vehicular environments.

These outcomes directly validate the impact of the three key contributions introduced in Section 3. The *dual-layer detection* capability—across both V2I communications (CI-CEVSE2024) and in-vehicle CAN traffic (CICIoV2024)—confirms the holistic design of the framework. The *federated aggregation of local explanation maps* enhances transparency at the global scale, enabling system-wide reasoning about critical attack features. Meanwhile, the *personalization and adaptive tuning* of the global model, guided by explanation feedback, ensures consistent performance across heterogeneous clients, an essential property for practical deployment among vehicles with diverse hardware and usage characteristics.

In practice, SHAP computations required approximately 1.8 times more processing time than LIME per sample, but produced more stable global feature rankings. LIME, however, proved useful for rapid local inspection of anomalies, demonstrating a practical trade-off between interpretability detail and computational efficiency in federated environments.

Crucially, the results demonstrate that explainability and robustness are not competing objectives in federated intrusion detection. Instead, the proposed integrated XAI design improves both simultaneously. Compared with conventional FL, XAI-FL-IDS achieves stronger generalization, greater interpretability, and enhanced resilience to false positives—all while maintaining the privacy advantages inherent to decentralized learning. Collectively, these attributes establish XAI-FL-IDS as a viable and trustworthy solution for next-generation intrusion detection in connected and autonomous vehicular ecosystems.

To evaluate the fidelity and interpretability of the aggregated explanations, we examined the consistency of the global explanation map Φ_g across all participating clients. The most influential features identified by the aggregated SHAP and LIME explanations were found to align closely with high-activation regions in the local models, indicating that the aggregation preserved the semantic meaning of feature importance across clients. Furthermore, the global map highlighted network attributes (e.g., packet length, CAN ID frequency, and flow duration) that were consistently relevant for both in-vehicle and external intrusion scenarios. This demonstrates that the proposed aggregation mechanism maintains explanation fidelity while improving interpretability at the global level, allowing analysts to trace which features most influence detection decisions in the federated environment.

6. Conclusions and Future Works

This paper presented XAI-FL-IDS, an explainable federated intrusion detection framework for connected vehicles addressing data privacy, transparency, and distributed threat detection. Integrating SHAP and LIME into the FL cycle enables interpretable, privacy-preserving detection. Experiments on CICEVSE2024 and CICIoV2024 show high accuracy, low false positives, and fast convergence. The framework's dual-layer design, global explanation aggregation, and adaptive feedback improve robustness and personalization, demonstrating that explainability and resilience can effectively coexist in federated intrusion detection systems.

While this study primarily focuses on integrating XAI into federated intrusion detection, we acknowledge that securing the explanation aggregation process itself represents an important and independent line of research. The design of defense mechanisms capable of detecting or mitigating maliciously manipulated explanations would require a dedicated and comprehensive investigation. Future work will explore this direction to develop security-aware explainability methods that can further enhance the trustworthiness and resilience of federated learning frameworks. Future work will explore real-time explainability feedback for operator support, adaptive defenses against poisoning, and extension to multi-modal datasets integrating radar, LiDAR, and camera data.

Author Contributions: Conceptualization, R.T. and R.J.; methodology, R.T. and R.J.; software, R.T., F.A.; validation, R.T., R.J. and A.G.; writing—original draft preparation, R.T.; writing—review and editing, R.J., A.G., F.A. and A.I.; supervision, R.J., A.G. and A.I. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been realized with financial support by the European Regional Development Fund within the Operational Programme "Bulgarian national recovery and resilience plan", Procedure for direct provision of grants "Establishing of a network of research higher education institutions in Bulgaria", under the Project BG-RRP-2.004-0005 "Improving the research capacity and quality to achieve international recognition and resilience of TU-Sofia.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: All authors declare that they have no conflict of interest.

References

1. Kurachi, Y.; Kubo, K.; Fujii, T. Cybersecurity for 5G-V2X: Attack Surfaces and Mitigation Techniques. In Proceedings of the 93rd IEEE Vehicular Technology Conference (VTC Spring), Virtual, 25 April–19 May 2021.
2. Duan, Z.; Mahmood, J.; Yang, Y.; Berwo, M.A.; Yassin, A.K.A.; Bhutta, M.N.M.; Chaudhry, S.A. TFPPASV: A Three-Factor Privacy Preserving Authentication Scheme for VANETs. *Secur. Commun. Netw.* **2022**, *2022*, 8259927.
3. Sudhina Kumar, G.K.; Krishna Prakasha, K.; Balachandra, M.; Rajarajan, M. Explainable Federated Framework for Enhanced Security and Privacy in Connected Vehicles Against Advanced Persistent Threats. *IEEE Open J. Veh. Technol.* **2025**, *6*, 1438–1463 .
4. Taheri, R.; Gegov, A.; Arabikhan, F.; Ichtev, A.; Georgieva, P. Explainable Artificial Intelligence for Intrusion Detection in Connected Vehicles. *Commun. Comput. Inf. Sci. (CCIS)* **2025**, *2768*, 162–176.
5. Baidar, R.; Maric, S.; Abbas, R. Hybrid Deep Learning–Federated Learning Powered Intrusion Detection System for IoT/5G Advanced Edge Computing Network. *arXiv* **2025**, arXiv:2509.15555.
6. Taheri, R.; Pooranian, Z.; Martinelli, F. Enhancing the Robustness of Federated Learning-Based Intrusion Detection Systems in Transportation Networks. In Proceedings of the 2025 IEEE International Conference on High Performance Computing and Communications (HPCC), Exeter, UK, 13–15 August 2025.
7. Buyuktanir, B.; Altinkaya, Ş.; Karatas Baydogmus, G.; Yildiz, K. Federated Learning in Intrusion Detection: Advancements, Applications, and Future Directions. *Clust. Comput.* **2025**, *28*, 1–25.
8. Almaazmi, K.I.A.; Almheiri, S.J.; Khan, M.A.; Shah, A.A.; Abbas, S.; Ahmad, M. Enhancing Smart City Sustainability with Explainable Federated Learning for Vehicular Energy Control. *Sci. Rep.* **2025**, *15*, 23888.

9. Abdallah, E.E.; Aloqaily, A.; Fayez, H. Identifying Intrusion Attempts on Connected and Autonomous Vehicles: A Survey. *Procedia Comput. Sci.* **2023**, *220*, 307–314.
10. Luo, F.; Wang, J.; Zhang, X.; Jiang, Y.; Li, Z.; Luo, C. In-Vehicle Network Intrusion Detection Systems: A Systematic Survey of Deep Learning-Based Approaches. *PeerJ Comput. Sci.* **2023**, 9, e1648.
11. Temitope, A.; Onumanyi, A.J.; Zuccoli, F.; Kolog, E.A. Federated Learning-Based Intrusion Detection for Secure In-Vehicle Communication Networks. In Proceedings of the 14th International Conference on Ambient Systems, Networks and Technologies (ANT), Leuven, Belgium, 23–25 April 2025; pp. 1082–1090.
12. Abdel Hakeem, S.A.; Kim, H. Advancing Intrusion Detection in V2X Networks: A Survey on Machine Learning, Federated Learning, and Edge AI for Security. *IEEE Trans. Intell. Transp. Syst.* 2025, *Early Access*.
13. Huang, K.; Wang, H.; Ni, L.; Xian, M.; Zhang, Y. FED-IoV: Privacy-Preserving Federated Intrusion Detection with a Lightweight CNN for Internet of Vehicles. *IEEE Internet Things J.* 2024, *Forthcoming*.
14. Xie, N.; Jia, W.; Li, Y.; Li, L.; Wen, M. IoV-BCFL: An Intrusion Detection Method for IoV Based on Blockchain and Federated Learning. *Ad Hoc Netw.* **2024**, *163*, 103590.
15. Almansour, S.; Dhiman, G.; Alotaibi, B.; Gupta, D. Adaptive Personalized Federated Learning with Lightweight Convolutional Network for Intrusion Detection in IoV. *Sci. Rep.* **2025**, *15*, 35604.
16. Nwakanma, C.I.; Ahakonye, L.A.C.; Njoku, J.N.; Odirichukwu, J.C.; Okolie, S.A.; Uzondu, C.; Ndubuisi Nweke, C.C.; Kim, D.-S. Explainable Artificial Intelligence for Intrusion Detection and Mitigation in Intelligent Connected Vehicles: A Review. *Appl. Sci.* **2023**, *13*, 1252.
17. Chamola, V.; Hassija, V.; Sulthana, A.R.; Ghosh, D.; Dhingra, D.; Sikdar, B. A Review of Trustworthy and Explainable Artificial Intelligence (XAI). *IEEE Access* **2023**, *11*, 78994–79015.
18. Saheed, Y.K.; Chukwuere, J.E. XAIEnsembleTL-IoV: An Explainable Ensemble Transfer Learning Framework for Zero-Day Botnet Attack Detection in Internet of Vehicles. *Electronics* **2024**, *13*, 1134.
19. Alabbadi, A.; Bajaber, F. An Intrusion Detection System over IoT Data Streams Using eXplainable Artificial Intelligence. *Sensors* **2025**, *25*, 847.
20. Fatema, K.; Dey, S.K.; Anannya, M.; Khan, R.T.; Rashid, M.M.; Su, C.; Mazumder, R. FedXAI-IDS: A Federated Learning Approach with SHAP-Based Explainable Intrusion Detection for IoT Networks. *Future Internet* **2025**, *17*, 234.
21. Zhang, H.; Wu, B.; Yuan, X.; Pan, S.; Tong, H.; Pei, J. Trustworthy Graph Neural Networks: Aspects, Methods, and Trends. *Proc. IEEE* **2024**, *112*, 97–139.
22. Wu, B.; Li, J.; Yu, J.; Bian, Y.; Zhang, H.; Chen, C.; Hou, C.; Fu, G.; Chen, L.; Xu, T. A Survey of Trustworthy Graph Learning: Reliability, Explainability, and Privacy Protection. *arXiv* **2022**, arXiv:2205.10014.
23. Sanjalawe, Y.; Allehyani, B.; Kurdi, G.; Makhadmeh, S.; Jaradat, A.; Hijazi, D.; et al. Forensic Analysis of Cyberattacks in Electric Vehicle Charging Systems Using Host-Level Data. *Comput. Mater. Contin.* **2025**, *85*, 3289–3320.
24. Uddin, M.A.; Chu, N.H.; Rafeh, R.; Barika, M. A Scalable Hierarchical Intrusion Detection System for Internet of Vehicles. *arXiv* **2025**, arXiv:2505.16215.