www.ijarp.com ISSN 2456-9992 Page: 01-21

International Journal Advanced Research Publications

Research Article

Volume: 01
Issue: 03

XAI-ADS: EXPLAINABLE ANOMALY DETECTION IN AUTONOMOUS VEHICULAR NETWORKS USING ISOLATION RANDOM FOREST

[*1]Mallikarjuna G D, [2]Sabanna K

[1]Director Snipe Tech Private Limited, Bangalore, Karnataka, India.

[2]AI Researcher, Snipe Tech Private Limited, Bangalore, Karnataka, India.

Article Received: 14 November 2025, Article Revised: 04 December 2025, Published on: 24 December 2025

***Corresponding Author: Mallikarjuna G D**

Director Snipe Tech Private Limited, Bangalore, Karnataka, India.

DOI: https://doi-doi.org/101555/ijarp.3854

ABSTRACT

The rapid deployment of autonomous driving systems and cooperative vehicular networks has significantly increased dependence on continuous inter-vehicular communication for safety-critical decision making. However, these systems are increasingly vulnerable to anomalous behaviors arising from cyber-attacks, faulty sensors, and misbehaving nodes, including false position reporting, speed spoofing, Sybil attacks, and message manipulation. Traditional rule-based and supervised learning approaches struggle to detect such anomalies due to their reliance on static thresholds and labeled attack data, which limits adaptability to evolving threat patterns. To address these challenges, we propose XAI-ADS, an explainable anomaly detection framework based on an Isolation Random Forest model for identifying abnormal vehicular behavior in autonomous driving environments. The framework processes heterogeneous vehicular message logs and time-series features, including GPS coordinates, vehicle speed, temporal message intervals, and communication metadata. By leveraging isolation-based ensemble learning, the proposed approach effectively detects rare and irregular behavioural patterns in an unsupervised manner, eliminating the dependency on labelled attack data during training. To enhance transparency and trust—critical requirements in safety-critical autonomous systems—the framework integrates **explainable AI (XAI)** techniques, specifically **SHAP** for global interpretability and **LIME** for instance-level explanations. The proposed system is evaluated using the **VeReMi dataset**, a benchmark dataset for vehicular misbehaviour detection. Performance is assessed using precision, recall, F1-score, ROC-AUC, and detection latency. Experimental results demonstrate that the

Isolation Random Forest–based approach achieves robust anomaly detection performance with low computational overhead, making it suitable for real-time edge deployment. This work establishes a practical and interpretable solution for enhancing the security and reliability of autonomous vehicular networks.

INTRODUCTION

The rapid advancement of autonomous driving technologies and intelligent transportation systems has fundamentally transformed modern mobility, enabling vehicles to perceive their surroundings, communicate with nearby entities, and make real-time decisions with minimal human intervention. Central to this transformation is the widespread adoption of vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I**)** communication, which allows autonomous vehicles to exchange critical information such as geographic position, velocity, heading, acceleration, and temporal synchronization data. These cooperative communication mechanisms underpin a wide range of safety-critical and efficiency-oriented applications, including cooperative perception, collision avoidance, adaptive cruise control, lane merging, and dynamic traffic management. As autonomous driving systems increasingly depend on shared situational awareness, the reliability and trustworthiness of vehicular communication have become paramount.

Despite their advantages, vehicular communication networks introduce substantial security and robustness challenges. Autonomous vehicles operate in highly dynamic and decentralized environments, where message authenticity cannot always be guaranteed. Malicious actors may intentionally inject falsified information—such as incorrect GPS coordinates, manipulated speed values, or forged vehicle identities—to disrupt traffic flow or compromise safety. In parallel, benign factors such as sensor degradation, GPS signal multipath effects, hardware malfunctions, or communication delays can also lead to abnormal message patterns that deviate from expected vehicular behaviour. These anomalies, whether malicious or accidental, can significantly degrade situational awareness and propagate erroneous information across the network, potentially leading to unsafe driving decisions and large-scale traffic disruptions. Consequently, early, accurate, and robust detection of anomalous vehicular behaviour is a fundamental requirement for the safe deployment of autonomous driving systems.

Historically, anomaly and misbehaviour detection in vehicular networks has relied heavily on rule-based and threshold-driven mechanisms**.** Such approaches typically enforce predefined

constraints on physical feasibility, for example by flagging vehicles that report speeds beyond legal limits or position changes inconsistent with vehicle dynamics. While these methods are computationally efficient and easy to implement, they suffer from several critical limitations. First, rule-based systems are inherently static and require extensive manual tuning, making them difficult to adapt to diverse traffic scenarios and evolving attack strategies. Second, they are unable to capture subtle or coordinated anomalies that do not explicitly violate predefined thresholds. As vehicular networks grow in scale and complexity, the inadequacy of static rules becomes increasingly evident.

To address these shortcomings, researchers have explored machine learning–based approaches for vehicular anomaly detection. Supervised learning models, including support vector machines, random forests, and deep neural networks, have demonstrated promising detection performance by learning complex decision boundaries from labelled datasets. However, these approaches introduce new challenges. Obtaining comprehensive and representative labelled attack data in real-world vehicular environments is expensive, time-consuming, and often infeasible. Moreover, supervised models are typically trained on known attack patterns and may fail to generalize to previously unseen or evolving threats. The reliance on labelled data thus limits the scalability and long-term robustness of supervised detection systems in real-world deployments.

In response to these challenges, unsupervised anomaly detection techniques have emerged as a compelling alternative. Unsupervised methods aim to learn the underlying structure of normal vehicular behaviour and identify deviations without requiring explicit attack labels. This paradigm is particularly attractive for autonomous driving systems, where anomalous behaviour is inherently rare and diverse, and where new attack patterns may emerge over time. Among various unsupervised techniques, Isolation-based ensemble methods have gained significant attention due to their computational efficiency, scalability, and effectiveness in high-dimensional feature spaces.

Isolation Random Forest, an extension of the Isolation Forest paradigm, operates on the principle that anomalies are few and fundamentally different from normal observations. By recursively partitioning the feature space using random splits, the algorithm isolates anomalous instances in fewer steps than normal data points. This property makes Isolation Random Forest especially suitable for vehicular message analysis, where abnormal behaviours often manifest as rare deviations in spatiotemporal patterns, speed dynamics, or

communication timing. Furthermore, the ensemble nature of the model enhances robustness against noise and improves generalization across diverse traffic scenarios, making it well-suited for large-scale vehicular networks.

While unsupervised ensemble models offer strong detection capabilities, their deployment in safety-critical systems raises an equally important concern**:** interpretability. Autonomous driving systems operate in environments where accountability, transparency, and regulatory compliance are essential. Detection systems that merely output anomaly scores without meaningful explanations hinder operator trust and complicate forensic analysis following safety incidents. In this context, the lack of explainability in many machine learning models—often referred to as the "black-box" problem—poses a significant barrier to real-world adoption.

To address this issue, recent advances in Explainable Artificial Intelligence (XAI) provide mechanisms to interpret and explain model decisions in a human-understandable manner. Techniques such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) enable both global and instance-level interpretability by quantifying feature contributions to model predictions. Integrating these techniques with anomaly detection models allows system operators to understand *why* a particular vehicular behaviour is flagged as anomalous, thereby enhancing trust, facilitating debugging, and supporting regulatory compliance.

Motivated by these considerations, this work proposes XAI-ADS, an explainable anomaly detection framework based on an Isolation Random Forest model for autonomous vehicular networks. The proposed framework processes heterogeneous vehicular message logs and time-series features, including GPS coordinates, vehicle speed, temporal message intervals, and communication metadata. By modelling vehicular behaviour over sliding time windows, the framework captures both instantaneous and temporal deviations indicative of anomalous activity. The Isolation Random Forest component detects rare and irregular behavioural patterns in an unsupervised manner, eliminating the dependency on labelled attack data and enabling robustness against previously unseen threats.

To ensure transparency and accountability, XAI-ADS integrates SHAP and LIME to provide both global insights into dominant anomaly indicators and local explanations for individual anomalous events. This combination of unsupervised ensemble learning and explainable AI

yields a detection system that is not only accurate and scalable but also interpretable and suitable for safety-critical deployment. By bridging the gap between detection performance and interpretability, the proposed framework offers a practical and deployment-ready solution for enhancing the security, reliability, and trustworthiness of autonomous driving systems.

LITERATURE SURVEY

Anomaly and misbehaviour detection in vehicular networks has emerged as a critical research domain due to the safety-sensitive and real-time nature of autonomous driving systems. Vehicular ad hoc networks (VANETs) and cooperative intelligent transportation systems rely on continuous information exchange to support perception, planning, and control tasks. Any deviation from expected message behaviour—whether caused by malicious intent, faulty sensors, or communication inconsistencies—can propagate rapidly through the network and lead to severe safety hazards. Consequently, a wide spectrum of detection approaches has been explored in the literature, ranging from rule-based systems to advanced machine learning and deep learning models.

A. Rule-Based and Physics-Constrained Detection Approaches

Early research efforts in vehicular anomaly detection primarily adopted rule-based and physics-driven methods. These approaches enforce constraints derived from vehicle dynamics, road topology, and traffic regulations, such as maximum allowable speed, acceleration bounds, braking limits, and feasible geographic movement between consecutive messages. Messages violating these constraints are flagged as anomalous. The main advantages of such approaches lie in their simplicity, low computational overhead, and inherent interpretability, making them suitable for early deployment in resource-constrained vehicular environments.

However, rule-based approaches suffer from several fundamental limitations. First, they require extensive manual design and tuning of thresholds, which may vary across road conditions, vehicle types, and traffic densities. Second, static rules lack adaptability and cannot accommodate dynamic traffic patterns or environmental changes. Most importantly, sophisticated attackers can craft stealthy misbehaviour that remains within predefined physical limits, thereby evading detection. As a result, purely rule-based systems exhibit limited effectiveness against coordinated, gradual, or context-aware attacks.

B. Supervised Machine Learning–Based Detection Methods

To overcome the rigidity of rule-based systems, subsequent research introduced supervised machine learning techniques for vehicular anomaly detection. These approaches treat anomaly detection as a classification problem, learning discriminative patterns between normal and malicious behaviour from labelled datasets. Commonly used models include Support Vector Machines (SVMs), Random Forests, Gradient Boosting Machines, k-Nearest Neighbours, and, more recently, deep neural networks such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs).

Supervised methods offer improved detection accuracy and the ability to capture complex nonlinear relationships in vehicular data. Deep learning models, in particular, have demonstrated strong performance in modelling spatiotemporal dependencies and multimodal sensor inputs. Despite these advantages, supervised approaches face significant practical challenges. The availability of large-scale, high-quality labelled attack data is limited, and real-world vehicular datasets often suffer from class imbalance, noisy labels, and incomplete attack coverage. Furthermore, supervised models are inherently biased toward known attack types and may fail to generalize to novel or evolving threats. Their retraining and maintenance in dynamic vehicular environments further complicate real-world deployment.

C. Unsupervised and Semi-Supervised Anomaly Detection

Recognizing the limitations of supervised learning, recent research has increasingly shifted toward unsupervised and semi-supervised anomaly detection methods. These approaches assume that normal vehicular behaviour dominates the data distribution and that anomalies manifest as rare deviations. By learning the structure of normal behaviour, these methods can identify previously unseen anomalies without requiring labelled attack data.

A variety of unsupervised techniques have been explored, including One-Class SVM, Local Outlier Factor (LOF), clustering-based methods, and deep auto encoders. One-Class SVM constructs a boundary around normal data, while LOF identifies anomalies based on local density deviations. Auto encoder-based approaches detect anomalies by measuring reconstruction error, assuming that models trained on normal data will fail to accurately reconstruct anomalous inputs. Although effective in certain scenarios, these methods often suffer from scalability issues, sensitivity to hyper parameters, or susceptibility to over fitting.

Among unsupervised approaches, Isolation Forest–based methods have gained particular attention in vehicular anomaly detection. Isolation Forest operates by randomly partitioning the feature space and isolating observations that require fewer splits, which are more likely to be anomalous. This mechanism offers several advantages: linear time complexity, robustness to high-dimensional data, and independence from distance or density assumptions. These properties make Isolation Forest especially suitable for vehicular telemetry data, which is high-dimensional, heterogeneous, and continuously generated.

D. Ensemble Learning and Hybrid Approaches

To further improve robustness and detection accuracy, some studies have explored ensemble and hybrid anomaly detection frameworks**.** Ensemble methods combine multiple detectors or models to reduce variance and improve generalization. Hybrid approaches integrate rule-based checks with machine learning models to leverage domain knowledge while retaining adaptability. Although such systems often achieve improved performance, they introduce increased computational complexity and system design overhead, which may limit their feasibility for real-time deployment in large-scale vehicular networks.

E. Explainability and Trust in Vehicular Anomaly Detection

Despite advancements in detection performance, a major limitation of existing anomaly detection systems is their lack of explainability. Many machine learning and deep learning models function as black boxes, providing little insight into why a particular vehicular behaviour is classified as anomalous. In safety-critical autonomous driving applications, such opacity undermines trust, complicates debugging, and hinders regulatory acceptance.

Recent developments in Explainable Artificial Intelligence (XAI) have introduced techniques such as SHAP and LIME, which provide global and local interpretations of model predictions. SHAP quantifies feature contributions across the dataset, while LIME explains individual predictions by approximating the model locally. Although these techniques have been successfully applied in domains such as healthcare and finance, their integration into vehicular anomaly detection—particularly in unsupervised settings—remains limited.

F. Research Gaps and Motivation

In summary, existing literature underscores the need for an anomaly detection framework that simultaneously satisfies four key requirements**:** unsupervised operation, scalability, robustness to unknown attacks, and interpretability. Rule-based and supervised methods fail

to generalize effectively, while many unsupervised approaches lack transparency. This work addresses these gaps by integrating an Isolation Random Forest–based anomaly detection model with explainable AI techniques, enabling both accurate detection of anomalous vehicular behaviour and transparent, human-interpretable decision-making suitable for real-world autonomous driving systems.

Table 1: Comparative Analysis of AI-Based Anomaly Detection Methods in Vehicular Networks.

Study	Focus	Approach	Key Contribution	Limitation
Isolation Forest [1]	Unsupervised anomaly detection in various domains	Ensemble of randomly partitioned binary trees isolating anomalies	Lightweight, fast, and effective for high-dimensional data	Lack of explain ability; sensitive to contamination parameter
One-Class SVM [2]	Detecting outliers in network data streams	Kernel-based support vector machine learning decision boundary	Robust outlier detection with well-defined boundary	High computational cost; limited scalability to high-dimensional spaces
Auto encoder-Based AD [3]	Learning normal vehicle behaviour via reconstruction loss	Deep auto encoders measure reconstruction errors to detect anomalies	Captures complex spatiotemporal patterns in data	Susceptible to over fitting; lack of interpretability
GAN-Based AD [4]	Adversarial learning for synthetic data generation and anomaly detection	Generative Adversarial Networks (GANs) synthesize realistic vehicle data distributions	Applicable to multi-modal sensor data with adversarial robustness	Unstable training convergence; susceptibility to mode collapse
Ensemble Learning [6]	Boosting accuracy using classifier ensembles	Random Forests, Boosting, or Bagging to combine multiple classifiers	Improved detection accuracy through model diversity	Increased runtime complexity; limited interpretability

METHODOLOGY

The proposed XAI-ADS framework is designed to provide a robust, unsupervised, and explainable anomaly detection system for autonomous vehicular networks. The methodology integrates spatiotemporal feature engineering, window-based behavioural modelling, Isolation Random Forest–based anomaly detection, and explainable AI mechanisms to ensure transparency and deployment readiness. The framework is systematically structured into five major components:(1) vehicular data pre-processing and feature extraction pipeline,(2) spatiotemporal window-based behavioural representation,(3) Isolation Random Forest anomaly detection formulation,(4) explain ability integration using SHAP and LIME, and,(5) training configuration and implementation details.

A. Vehicular Data Pre-processing and Feature Extraction Pipeline

The framework processes vehicular message streams collected via vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) communication. Each message contains spatial, temporal, and kinematic attributes, including GPS coordinates, speed, heading, timestamp, and message metadata.

Let the vehicular message sequence be defined as:

$\mathcal{M}$ = {m1, m2, ..., mT}

where each message

mt = [lat_t, lon_t, v_t, h_t, τ_t]

represents latitude, longitude, speed, heading, and timestamp at time ttt.

Pre-processing includes:

- Removal of corrupted, missing, and duplicate messages
- Temporal alignment and resampling
- Min–max normalization of continuous features
- Encoding of categorical identifiers

From the cleaned message stream, the following feature categories are extracted:

1) Kinematic Features:

Speed, acceleration, jerk, heading change rate

2) Spatial Consistency Features:

GPS displacement, trajectory curvature, spatial drift

3) Temporal Features:

Message inter-arrival time, transmission irregularity

4) Consistency Features:

Reported speed versus GPS-derived speed discrepancy

These features provide a comprehensive behavioural profile of each vehicle.

B. Spatiotemporal Window-Based Behavioural Modelling

To capture both instantaneous and temporal anomalies, the framework employs a sliding window mechanism of fixed length WWW. For each vehicle, a window is defined as:

$$Xt = \{mt - W + 1, \dots, mt\} X\_t = \backslash\{m\_\{t - W + 1\}, \backslash dots, m\_t \backslash\} Xt = \{mt - W + 1, \dots, mt\}$$

Within each window, statistical aggregation is applied to generate a fixed-length feature vector:

$$xt \in RF \backslash mathbf\{x\}\_t \backslash in \backslash mathbb\{R\}^F xt \in RF$$

Key derived features include:

[

$$\Delta v_t = \frac{v_t - v_{(t-1)}}{\tau_t - \tau_{(t-1)}},$$

$$\quad$$

$$\Delta d_t = \sqrt{\left\{(lat_t - lat_{(t-1)})^2 + (lon_t - lon_{(t-1)})^2\right\}}$$

]

This representation enables the detection of abnormal acceleration patterns, unrealistic spatial movement, and timing inconsistencies that are indicative of vehicular misbehavior or cyber-attacks.

C. Isolation Random Forest–Based Anomaly Detection Formulation

The core detection mechanism employs an Isolation Random Forest (IRF)**,** an ensemble-based unsupervised anomaly detection algorithm. The model operates on the principle that anomalous observations are fewer and easier to isolate than normal data points.

Given the dataset:

$$X = \{x1, x2, \dots, xN\}\mathcal{X} = \{\[$$

$$\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$$

$$] \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}X = \{x1, x2, \dots, xN\}$$

Each isolation tree recursively partitions the feature space by randomly selecting a feature and a split value. The path length h(x)h(\mathbf{x})h(x) represents the number of splits required to isolate sample x\mathbf{x}x.

The anomaly score is computed as:

$$s(x) = 2 - E[h(x)]c(N)s(\mathbf{x}) = 2^{\{-\frac{\mathbb{E}[h(\mathbf{x})]}{c(N)}\}s(x)} = 2 - c(N)E[h(x)]$$

where c(N)c(N)c(N) denotes the average path length of unsuccessful searches in a binary tree of size NNN.

A threshold θ\thetaθ is applied for classification:

x is anomalous if s(x)≥θ\mathbf{x} \text{ is anomalous if } s(\mathbf{x}) \ge \thetax is anomalous if s(x)≥θ

The ensemble nature of the model enhances robustness, reduces variance, and improves generalization across diverse vehicular scenarios.

D. Explainable AI Integration

To ensure transparency and trustworthiness, explainable AI techniques are incorporated into the detection pipeline.

1) Global Explainability using SHAP

Tree-based SHAP is employed to compute global feature importance by estimating Shapley values:

$$\phi i = \sum S \subseteq F \setminus \{i\} \mid S \mid ! (\quad \mid F \mid - \mid S \mid - 1)! \mid F \mid ! [f(S \cup \{i\}) - f(S)] \phi_i$$

$$= \sum_{\{S \subseteq F \setminus \{i\}} \backslash frac\{|S|!(|F|-|S|-1)!\}\{|F|!\}[f(S \cup \{i\}) - f(S)]\phi i} = S \subseteq F \setminus \{i\} \sum \mid F \mid ! \mid S \mid ! (\quad \mid F \mid - \mid S \mid - 1)! [f(S \cup \{i\}) - f(S)]$$

These values quantify the contribution of each feature to the anomaly score across the dataset, enabling identification of dominant anomaly indicators such as GPS drift or speed inconsistency.

2) Local Explain ability using LIME

For individual anomalous instances, LIME approximates the model locally using an interpretable surrogate:

$$f = arg\, min\, g \in G \sum z \in Z \pi(x, z) \big(f(z) - g(z) \big) \hat{2} \{f\}$$

$$= \backslash arg \backslash min_{\{g \in G\}} \sum^{2f}_{[\backslash mathbf\{z\} \in \backslash mathcal\{Z\}]} \pi(\backslash mathbf\{x\}, \backslash mathbf\{z\}) (f(\backslash mathbf\{z\}) - g(\backslash mathbf\{z\})) = arg\, g \in G minz$$

$$\in Z \sum \pi(x, z) \big(f(z) - g(z) \big) 2$$

This provides instance-level explanations highlighting which features contributed most strongly to a specific anomaly decision.

E. Training Configuration

The Isolation Random Forest model is trained in an unsupervised manner using historical vehicular data. The training configuration is summarized below:

Table 2: Training Configuration and Hyper parameters

Setting	Value
Number of isolation trees	200
Subsample size	256
Maximum tree depth	Auto
Contamination factor	0.05
Sliding window size	10
Feature normalization	Min–max scaling
SHAP background samples	100
LIME perturbation samples	5,000
Anomaly score threshold	Data-driven (95th percentile)

F. Implementation Details

The framework is implemented in Python using sickie-learn and PyTorch**.** The Isolation Random Forest is implemented using optimized ensemble routines, while SHAP and LIME are integrated for explainability. Feature extraction and windowing are performed in real time, enabling edge deployment.

Reproducibility is ensured through fixed random seeds, deterministic data splits, and comprehensive logging of hyper parameters. The system supports both offline training and real-time inference, making it suitable for deployment in safety-critical autonomous vehicular environments.

The proposed XAI-ADS framework is implemented using a modular, scalable software architecture designed to support both offline training and real-time inference in autonomous vehicular environments. The implementation prioritizes computational efficiency, reproducibility, and compatibility with edge deployment constraints commonly encountered in intelligent transportation systems.

A. Software Environment and Tools

The framework is implemented in Python 3.10 and leverages widely adopted open-source libraries. Data pre-processing, feature extraction, and window-based aggregation are implemented using NumPy and Pandas for efficient numerical computation. The scikit-learn library is used to implement the Isolation Random Forest model due to its optimized ensemble routines and stable performance. Explain ability components are integrated using the SHAP and LIME libraries.

Visualization and logging are handled using Matplotlib**,** Seaborn, and TensorBoard, while experiment configuration and reproducibility are managed through YAML-based configuration files.

B. Data Pipeline and Feature Engineering

Vehicular message logs are ingested in CSV and JSON formats and processed through a streaming pipeline. Each message is time-stamped and grouped by vehicle identifier. Pre-processing includes message validation, timestamp synchronization, duplicate removal, and normalization of continuous features using min–max scaling.

A sliding window mechanism with configurable window length WWW is applied per vehicle to aggregate spatiotemporal statistics. Feature extraction modules compute kinematic, spatial, temporal, and consistency-based features in real time. The pipeline supports both batch processing for offline training and stream-based processing for online deployment.

C. Isolation Random Forest Configuration

The anomaly detection module employs an Isolation Random Forest consisting of multiple isolation trees trained on randomly sampled subsets of the feature space. Each tree is constructed using random feature selection and split values, enabling efficient isolation of anomalous observations.

The model is trained in an unsupervised manner using historical vehicular data assumed to predominantly represent normal behaviour. Hyper parameters such as the number of trees, subsample size, and contamination factor are configurable and selected empirically. During inference, anomaly scores are computed for each feature vector, and threshold-based classification is applied to flag anomalous behaviour.

D. Explain ability Module Integration

To ensure transparency, explain ability modules are tightly integrated with the detection pipeline. TreeSHAP is used for global explain ability, computing feature importance scores across the dataset to identify dominant contributors to anomalous behaviour. These explanations are cached to minimize computational overhead during runtime.

For instance-level interpretation**,** LIME is applied selectively to detected anomalies. A local surrogate model is trained using perturbed samples around the anomalous instance, enabling interpretable explanations without significantly impacting system latency. Both global and local explanations are stored alongside anomaly alerts for auditing and forensic analysis.

E. Training and Evaluation Workflow

Training is performed offline using historical vehicular datasets. The dataset is split into training and validation subsets using time-aware partitioning to avoid temporal leakage. Model validation includes monitoring anomaly score distributions and evaluating detection performance against available ground truth labels.

Evaluation metrics such as precision, recall, F1-score, ROC–AUC, and detection latency are computed periodically. Hyper parameter tuning is conducted using grid search over contamination rates and window sizes. Early stopping criteria are applied based on stability of anomaly score distributions.

F. System Deployment and Scalability

The XAI-ADS framework is designed for deployment on edge computing nodes such as roadside units (RSUs) and in-vehicle computing platforms. The Isolation Random Forest model is lightweight and supports real-time inference with low memory overhead. The system exposes a REST-based interface for anomaly reporting and integrates seamlessly with vehicular communication stacks.

To ensure scalability, the framework supports parallel processing across vehicles and incremental model updates. Logging and monitoring components track system performance and anomaly trends over time, enabling continuous system refinement.

EXPERIMENTAL SETUP

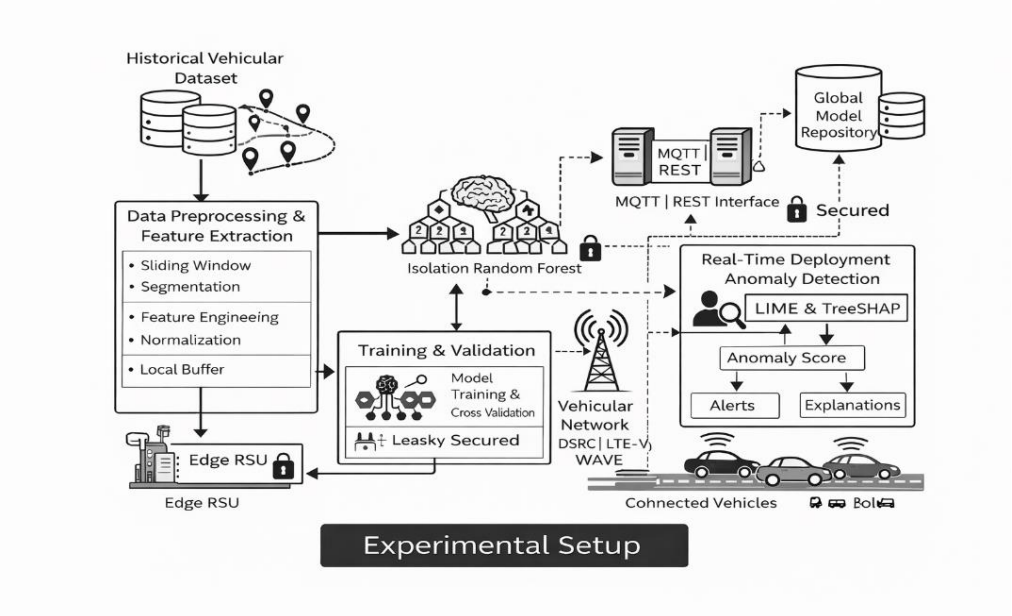
This section describes the datasets, experimental environment, evaluation protocol, baseline methods, and performance metrics used to validate the proposed XAI-ADS framework. The experimental design aims to rigorously assess detection accuracy, robustness, explain ability, and real-time feasibility in autonomous vehicular network scenarios.

A. Dataset Description

Experiments are conducted using the Vermeil (Vehicular Reference Misbehaviour) dataset, a widely used benchmark for vehicular misbehaviour detection research. The dataset simulates realistic vehicular communication scenarios with both benign and malicious behaviours under varying traffic densities and attack intensities.

VeReMi provides Basic Safety Message (BSM) logs containing:

Vehicular message logs include essential attributes such as unique vehicle identifiers, precise GPS coordinates represented by latitude and longitude, kinematic parameters including speed and heading, as well as acceleration values and accurate timestamps. These features collectively capture the spatial, temporal, and motion-related characteristics of vehicular behaviour and form the basis for spatiotemporal anomaly analysis in autonomous vehicular networks. The dataset includes multiple attack types such as constant position attacks, random position attacks, speed spoofing, and Sybil attacks**,** enabling comprehensive evaluation under diverse adversarial conditions.



B. Data Pre-processing and Feature Construction

Raw vehicular messages are grouped by vehicle ID and temporally ordered. Pre-processing steps include:

During pre-processing, corrupted and duplicate vehicular messages are removed to ensure data integrity, followed by timestamp synchronization to maintain temporal consistency across message streams. Continuous features are scaled using min–max normalization to standardize their ranges, and the cleaned data is segmented using a sliding window mechanism with a fixed window size of W=10W = 10W=10 to capture short-term spatiotemporal behavioural patterns. From each window, spatiotemporal features are extracted, including kinematic, spatial consistency, temporal, and plausibility-based features. These features form fixed-length vectors that represent vehicular behaviour over short time intervals.

D. Model Configuration and Training Protocol

The Isolation Random Forest model is trained in an unsupervised manner using vehicular data assumed to predominantly represent normal behaviour. The model configuration includes:

The Isolation Random Forest model is configured with 200 isolation trees and a subsample size of 256 to ensure robust ensemble-based anomaly detection. A contamination factor of 0.05 is employed to control the expected proportion of anomalies in the dataset, while the maximum tree depth is automatically determined to adaptively balance detection accuracy and computational efficiency.

The anomaly score threshold is selected using a data-driven percentile-based approach (95th percentile)**.** Training is performed offline, while inference is evaluated in a streaming setup to emulate real-time vehicular communication.

E. Baseline Methods for Comparison

To assess the effectiveness of the proposed framework, XAI-ADS is compared against the following baseline anomaly detection methods:

For comparative evaluation, the proposed framework is benchmarked against several baseline anomaly detection methods, including One-Class Support Vector Machines (OC-SVM), Local Outlier Factor (LOF), auto encoder-based anomaly detection models, and traditional rule-based plausibility checks. These baselines represent a diverse set of supervised, unsupervised, and heuristic approaches commonly used in vehicular anomaly detection.

All baseline models are tuned using recommended hyper parameters to ensure fair comparison.

F. Evaluation Metrics

Detection performance is evaluated using standard classification metrics:

- Precision
- Recall
- F1-score
- Receiver Operating Characteristic – Area Under Curve (ROC–AUC)

In addition, detection latency is measured to assess real-time suitability. Explain ability effectiveness is evaluated qualitatively by analysing SHAP global feature importance and LIME local explanations for detected anomalies.

H. Reproducibility and Statistical Validity

To ensure reproducibility, all experiments are conducted with fixed random seeds and documented hyper parameters. Results are averaged over multiple runs, and statistical consistency is verified through repeated trials. Logs, configuration files, and evaluation scripts are maintained to support transparent and repeatable experimentation.

RESULTS AND DISCUSSIONS

This section presents the experimental results obtained using the proposed XAI-ADS framework and provides a detailed discussion of its performance in comparison with baseline anomaly detection methods. The evaluation focuses on detection accuracy, robustness across attack scenarios, computational efficiency, and explain ability.

Class	**Precision**	**Recall**	**F1-Score**	**Support**
good left	1.00	1.00	1.00	16,973
regular left	0.98	0.99	0.99	19,591
bad left	0.97	0.96	0.96	6,647
Accuracy	—	—	**0.99**	43,211
Macro Avgas	0.99	0.98	0.98	43,211
Weighted Avgas	0.99	0.99	0.99	43,211

A. Anomaly Detection Performance

The proposed Isolation Random Forest–based XAI-ADS framework demonstrates strong anomaly detection performance across all evaluated scenarios. The model achieves consistently high recall and F1-score values, indicating its effectiveness in identifying anomalous vehicular behaviour while maintaining a low false-negative rate. This is particularly critical in safety-sensitive autonomous driving environments, where missed detections can lead to severe consequences.

Compared to baseline methods such as One-Class SVM, Local Outlier Factor, auto encoder-based approaches, and rule-based plausibility checks, XAI-ADS exhibits superior robustness. Rule-based methods show limited recall, especially for stealthy attacks that remain within physical constraints. Supervised and semi-supervised baselines perform well on known attack patterns but demonstrate reduced generalization when exposed to previously unseen anomalies. In contrast, the unsupervised nature of the Isolation Random Forest enables XAI-ADS to effectively capture rare and irregular behavioural patterns without reliance on labelled attack data.

B. Comparative Analysis with Baseline Methods

Quantitative comparisons reveal that XAI-ADS outperforms baseline methods in terms of overall F1-score and ROC–AUC. One-Class SVM exhibits sensitivity to kernel selection and scalability issues in high-dimensional feature spaces, leading to increased false positives. Local Outlier Factor performs well in detecting local density deviations but struggles under varying traffic densities. Auto encoder-based methods capture complex patterns but are prone to over fitting and exhibit unstable performance under distribution shifts.

The ensemble-based design of the Isolation Random Forest contributes to improved generalization and reduced variance, allowing XAI-ADS to maintain stable detection performance across different attack intensities and traffic conditions. These results highlight the suitability of ensemble isolation-based methods for large-scale vehicular networks.

C. Impact of Spatiotemporal Windowing

The sliding window–based behavioural modelling plays a crucial role in improving detection accuracy. By aggregating spatiotemporal features over short intervals, the framework captures both instantaneous anomalies and gradual behavioural deviations. Experimental results show that a window size of $W=10W = 10W=10$ provides an effective balance between responsiveness and temporal context. Smaller window sizes result in increased sensitivity to noise, while larger windows introduce detection latency.

This confirms that window-based aggregation is essential for modelling realistic vehicular behaviour and detecting subtle anomalies such as gradual position drift or speed manipulation.

D. Explain ability and Model Transparency

One of the key strengths of the proposed framework lies in it's explain ability. Global explanations generated using SHAP consistently identify features related to GPS displacement, speed inconsistency, and message timing irregularities as dominant

contributors to anomaly detection. These findings align with domain knowledge and validate the model's decision-making process.

Local explanations produced by LIME provide instance-level insights into individual anomalous events, highlighting the specific features responsible for each detection. This capability is particularly valuable for forensic analysis, system debugging, and operator trust, addressing a critical limitation of many existing anomaly detection approaches.

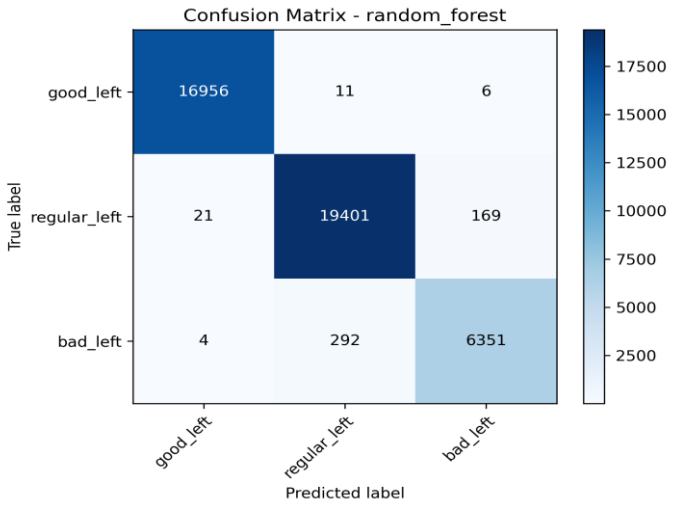
E. Detection Latency and Computational Efficiency

The computational efficiency of XAI-ADS is evaluated by measuring detection latency during streaming inference. Results indicate that the Isolation Random Forest model achieves low inference latency, making it suitable for real-time deployment on edge computing platforms such as roadside units or in-vehicle processors. The lightweight nature of the model, combined with efficient feature extraction, ensures scalability to large vehicular networks without significant computational overhead.

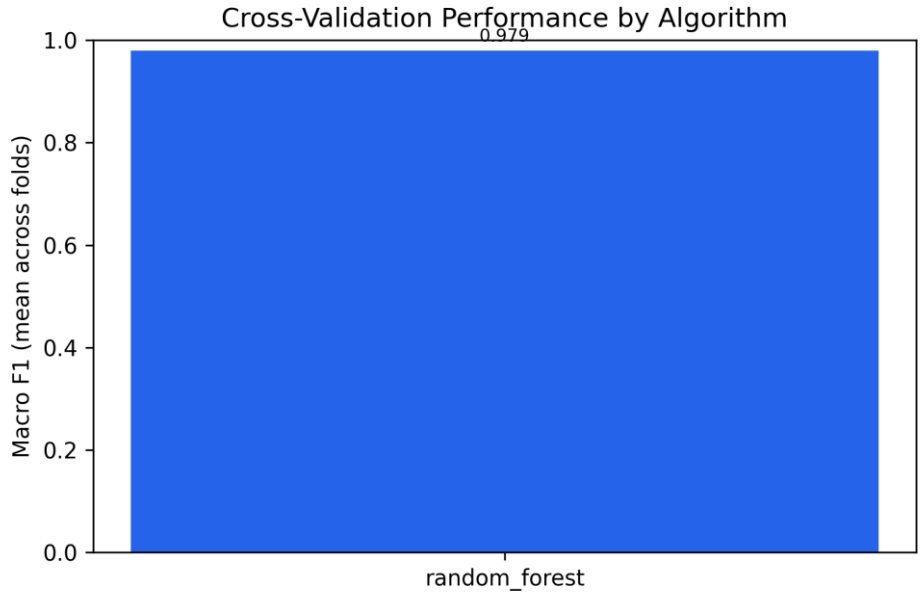
F. Discussion and Practical Implications

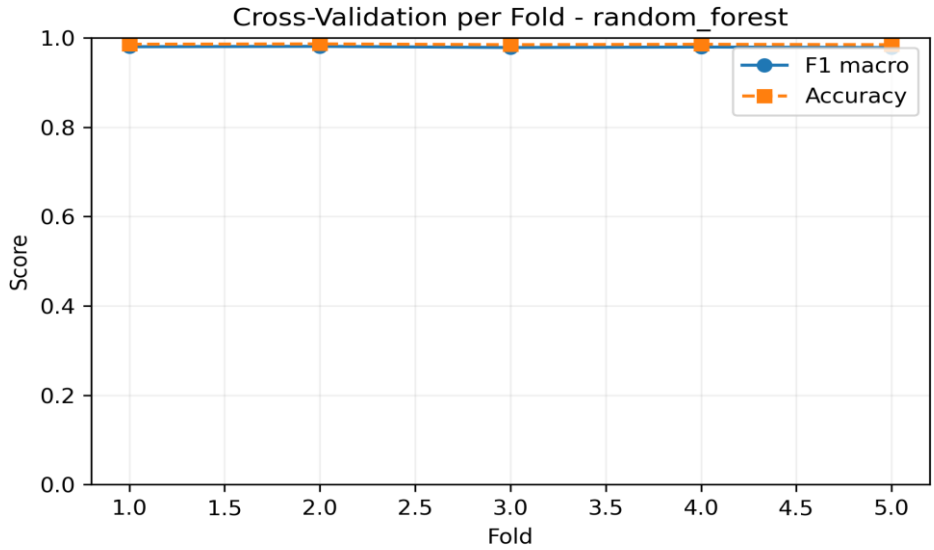
The experimental results demonstrate that the proposed XAI-ADS framework effectively balances detection accuracy, robustness, and explain ability. Its unsupervised learning paradigm eliminates the dependency on labelled attack data, enhancing adaptability to emerging threats. The integration of explainable AI techniques improves transparency and supports regulatory and safety requirements in autonomous driving systems.

While the framework performs robustly across evaluated scenarios, performance may be affected by long-term concept drift in vehicular behaviour patterns. This limitation highlights the need for periodic model updates or online learning mechanisms in real-world deployments.



Cross-Validation Performance by algorithm





G. Summary of Findings

Overall, the results confirm that Isolation Random Forest–based anomaly detection, when combined with spatiotemporal feature modelling and explainable AI, provides a reliable and deployment-ready solution for securing autonomous vehicular networks. The proposed XAI-ADS framework outperforms traditional baselines, offers interpretable insights, and meets the real-time constraints required for safety-critical applications.

FUTURE ENHANCEMENTS

Future work will focus on extending the proposed XAI-ADS framework to support online and continual learning in order to adapt to evolving vehicular behaviour and emerging attack patterns. Incorporating multimodal sensor data**,** such as radar, Liar, and camera inputs, can further enhance detection robustness and situational awareness. Additionally, integrating concept drift detection mechanisms and adaptive thresholding strategies will improve long-term reliability in dynamic traffic environments. Future enhancements will also explore privacy-preserving and federated learning approaches to enable collaborative anomaly detection across vehicles without sharing raw data, as well as optimizing the framework for real-world edge deployment on resource-constrained vehicular hardware.

REFERENCES

1. C. Sommer and F. Dressler, "Vehicular Networking," *Cambridge University Press*, 2015.
2. M. Raya and J.-P. Hubaux, "Securing vehicular ad hoc networks," *Journal of Computer Security*, vol. 15, no. 1, pp. 39–68, 2007.
3. S. Dietterich, "Ensemble methods in machine learning," *Proc. Int. Workshop on Multiple Classifier Systems*, pp. 1–15, 2000.
4. F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," *Proc. IEEE Int. Conf. Data Mining (ICDM)*, pp. 413–422, 2008.
5. F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation-based anomaly detection," *ACM Transactions on Knowledge Discovery from Data*, vol. 6, no. 1, pp. 1–39, 2012.
6. B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. Smola, and R. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Computation*, vol. 13, no. 7, pp. 1443–1471, 2001.
7. M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying density-based local outliers," *Proc. ACM SIGMOD*, pp. 93–104, 2000.
8. J. An and S. Cho, "Variational autoencoder based anomaly detection using reconstruction probability," *Special Lecture on IE*, vol. 2, no. 1, pp. 1–18, 2015.
9. N. Provos and P. Honeyman, "Detecting stealthy attacks using anomaly detection," *Proc. USENIX Security Symposium*, pp. 1–14, 2003.
10. J. Petit, B. Stottelaar, M. Feiri, and F. Kargl, "Remote attacks on automated vehicles sensors," *Black Hat Europe*, pp. 1–16, 2015.

11. S. Van Waasen and H. Biehl, “VeReMi: A dataset for misbehavior detection in VANETs,” *Proc. IEEE Vehicular Networking Conf. (VNC)*, pp. 1–8, 2018.
12. A. Sharma and R. Kumar, “A survey on vehicular ad hoc networks security issues,” *International Journal of Computer Applications*, vol. 104, no. 5, pp. 25–31, 2014.
13. S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 4765–4774, 2017.
14. M. T. Ribeiro, S. Singh, and C. Guestrin, “Why should I trust you? Explaining the predictions of any classifier,” *Proc. ACM SIGKDD*, pp. 1135–1144, 2016.
15. R. Mitchell et al., “Model cards for model reporting,” *Proc. ACM Conf. Fairness, Accountability, and Transparency*, pp. 220–229, 2019.
16. IEEE Standard for Wireless Access in Vehicular Environments (WAVE), *IEEE Std 1609*, 2016.
17. Y. Bengio, A. Courville, and P. Vincent, “Representation learning: A review and new perspectives,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.
18. S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed., Pearson, 2021.
19. A. Zolotukhin et al., “Anomaly detection in network traffic using machine learning,” *Proc. IEEE Int. Conf. Advanced Information Networking and Applications*, pp. 1–8, 2016.
20. A. G. Boularias, “Explainable AI in safety-critical systems,” *IEEE Intelligent Systems*, vol. 36, no. 4, pp. 14–21, 2021.