Mrs. Ch. Veera Gayathri, 2026, 14:2
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology

An Open Access Journal

Enhancing Autonomous Vehicle Security Using Explainable Artificial Intelligence for Anomaly Detection

Mrs. Ch. Veera Gayathri [1] , Bhaviri Sri Ganesha Seetha Hanuma Gowri [2], Sangani Praveen Dhana Kumar [3], Ryali Yuvaraj [4], Ganta Venkata Sridhar [5] , Gadhi Subrahmanya Krishna Teja [6]

[1] Assistant Professor, Department of CSE (Data Science) In Pragati Engineering College, Surampalem, Andhra Pradesh, India,
[2,3,4,5,6] UG Students Department of CSE (Data Science) In Pragati Engineering College, Surampalem, Andhra Pradesh, Indi

Abstract- Autonomous driving systems have emerged as a transformative technology in modern intelligent transportation, enabling vehicles to operate with minimal or no human intervention. These systems rely heavily on large volumes of sensor data, communication networks, and machine learning algorithms to make real-time driving decisions. However, the increasing integration of autonomous vehicles into vehicular networks has also introduced significant cybersecurity and safety challenges. In particular, anomalous behaviours caused by cyber-attacks, faulty sensors, or malicious vehicles in Vehicular Ad Hoc Networks (VANETs) can threaten the reliability and safety of autonomous driving environments. Detecting such anomalies using traditional monitoring approaches is difficult due to the complexity, scale, and dynamic nature of vehicular communication data. To address these challenges, this study proposes an explainable artificial intelligence (XAI)–based anomaly detection framework for autonomous driving systems. The proposed framework integrates machine learning models with explainability techniques to identify abnormal behaviours in vehicular networks while also providing transparent interpretations of model decisions. Initially, autonomous driving datasets are pre-processed through feature extraction, redundancy elimination, data balancing, and normalization to improve model performance. Several machine learning algorithms, including Decision Tree, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbour (KNN), Deep Neural Network (DNN), and AdaBoost, are implemented to classify vehicles as normal or anomalous based on their behavioural features. To enhance interpretability, the framework incorporates explainable AI techniques such as SHapley Additive explanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME). These methods provide both global and local explanations by identifying the most influential features contributing to anomaly detection decisions.

INDEX TERMS: Autonomous Driving, Anomaly Detection, Explainable Artificial Intelligence, Machine Learning, Vehicular Ad Hoc Networks (VANETs), SHAP, LIME, Feature Selection, Intelligent Transportation Systems.

I. INTRODUCTION

Autonomous driving systems represent a major advancement in modern intelligent transportation technology. These systems combine advanced sensors, communication networks, and artificial intelligence algorithms to enable vehicles to operate with minimal or no human intervention. Autonomous vehicles are capable of perceiving their surroundings, making real-time decisions, and navigating complex traffic environments using data collected from cameras, LiDAR sensors, radar systems, and vehicle-to-vehicle communication networks. As a result, autonomous driving technologies have the potential to significantly improve road safety, reduce traffic congestion, and enhance transportation efficiency in future smart cities.

Despite these advantages, autonomous driving systems also introduce several security and reliability challenges. Modern autonomous vehicles operate within Vehicular Ad Hoc Networks (VANETs), where vehicles continuously exchange information such as location, speed, and driving conditions with nearby vehicles and infrastructure. This communication framework enables cooperative driving and improves situational awareness among vehicles. However, such networks are vulnerable to various cybersecurity threats, including malicious data injection, spoofing attacks, and abnormal vehicle behavior. These anomalies can disrupt communication within the network and potentially lead to unsafe driving decisions, making reliable anomaly detection mechanisms essential for maintaining the safety and stability of autonomous transportation systems [3], [4].

Traditional anomaly detection techniques often rely on rule-based monitoring or manual inspection of network activity. While these methods may detect certain irregular behaviours, they are generally inefficient when applied to large-scale vehicular data generated by autonomous driving environments. Autonomous vehicles produce massive volumes of real-time sensor and communication data, making manual monitoring impractical. Consequently, data-driven approaches based on machine learning have gained significant attention for detecting abnormal patterns in autonomous driving systems. Machine learning algorithms can learn complex relationships within vehicular datasets and identify deviations from normal behavior, enabling early detection of malicious activities or system faults in vehicle networks [5], [7], [13].

Several machine learning models such as Decision Trees, Random Forest, Support Vector Machines (SVM), K-Nearest Neighbour (KNN), and Deep Neural Networks (DNN) have been widely applied to anomaly detection tasks in intelligent transportation systems. These models analyze multiple features related to vehicle behavior, communication patterns, and sensor measurements to classify whether a vehicle is operating normally or exhibiting anomalous activity. Ensemble learning techniques, in particular, have demonstrated strong predictive performance by combining multiple classifiers to improve detection accuracy and robustness in complex datasets [8], [11].

However, many high-performance machine learning models function as black-box systems, where the reasoning behind classification decisions is not easily interpretable. In safety-critical applications such as autonomous driving, lack of transparency can limit the reliability and trustworthiness of AI-based decision-making systems. Understanding why a model detects a vehicle as anomalous is essential for system validation, debugging, and regulatory compliance. Therefore, improving the interpretability of machine learning models has become an important research direction in the development of intelligent transportation systems [1], [2], [9].

To address this issue, Explainable Artificial Intelligence (XAI) techniques have been introduced to provide insights into the internal decision-making processes of machine learning models. XAI methods help identify the most influential features that contribute to model predictions and provide both global and local explanations of classification outcomes. Techniques such as SHapley Additive explanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME) enable researchers and system operators to better understand how machine

learning models detect anomalies within autonomous vehicle networks. These interpretability mechanisms enhance transparency, improve system reliability, and support the development of trustworthy AI-based transportation systems [10], [12].

Motivated by these challenges, this paper proposes an Explainable Artificial Intelligence framework for anomaly detection in autonomous driving systems. The proposed framework integrates machine learning models with explainability techniques to detect abnormal vehicle behavior while providing interpretable insights into model predictions. The system analyses autonomous driving datasets, extracts relevant behavioural features, and trains multiple machine learning classifiers to identify anomalous vehicles in VANET environments. In addition, explainable AI methods such as SHAP and LIME are incorporated to interpret classification decisions and identify the most influential features contributing to anomaly detection.

The remainder of this paper is organized as follows. Section II presents a review of existing research related to anomaly detection and explainable artificial intelligence in autonomous driving systems. Section III discusses the analysis of the existing system and the proposed methodology. Section IV describes the architecture and design framework of the proposed system. Section V explains the implementation modules and experimental setup. Section VI presents the experimental results and performance evaluation. Finally, Section VII concludes the paper and outlines future research directions.

II. LITERATURE SURVEY

In recent years, the rapid advancement of autonomous vehicle technologies has led to increased research interest in developing intelligent systems capable of ensuring safety and reliability in vehicular networks. Autonomous driving systems rely heavily on large volumes of sensor data and vehicle-to-vehicle communication to make real-time driving decisions. However, these systems are also vulnerable to various cyber threats and abnormal behaviours that may compromise network stability and vehicle safety. As a result, researchers have increasingly explored machine learning–based approaches for anomaly detection in autonomous driving environments and Vehicular Ad Hoc Networks (VANETs) [3], [4].

Several studies have focused on applying artificial intelligence techniques to detect abnormal behavior in autonomous vehicles. Early approaches primarily relied on statistical analysis and rule-based monitoring systems to identify irregular patterns in vehicular communication data. Although these methods provided basic anomaly detection capabilities, they were often limited in their ability to handle complex and high-dimensional datasets generated by modern autonomous vehicle systems. With the emergence of machine learning techniques, more advanced data-driven models have been introduced to analyze vehicular communication patterns and detect anomalies more efficiently.

Various machine learning algorithms have been widely applied for anomaly detection tasks in intelligent transportation systems. Classification techniques such as Decision Trees, Support Vector Machines (SVM), K-Nearest Neighbour (KNN), and Logistic Regression have been used to identify abnormal vehicle behavior by learning patterns from historical vehicular datasets. These models analyze different behavioural attributes such as vehicle position, speed, communication frequency, and sensor readings to distinguish between normal and anomalous states. Studies have shown that machine learning models are capable of identifying subtle deviations in vehicle behavior that may indicate malicious attacks or system faults within vehicular networks [5], [7].

To further improve anomaly detection performance, ensemble learning techniques have also been explored in several studies. Algorithms such as Random Forest, Gradient Boosting, and AdaBoost combine multiple classifiers to produce more accurate and robust predictions. Ensemble models are particularly effective when dealing with complex datasets because they reduce the risk of overfitting and improve generalization capability. In many

intelligent transportation applications, ensemble learning methods have demonstrated higher accuracy compared with single machine learning classifiers, making them suitable for detecting anomalies in autonomous driving systems [8], [11].

Recent research has also investigated the application of deep learning techniques for anomaly detection in vehicular networks. Deep Neural Networks (DNN) and Long Short-Term Memory (LSTM) models have been applied to analyze sequential sensor data and communication logs generated by autonomous vehicles. These models are capable of learning complex nonlinear relationships within large-scale datasets and can automatically extract meaningful features from raw input data. Although deep learning models can achieve high prediction accuracy, they often require large training datasets and substantial computational resources, which may limit their practical deployment in real-time vehicular systems.

Despite the improved performance of machine learning and deep learning models, many of these approaches operate as black-box systems, making it difficult to understand how predictions are generated. In safety-critical applications such as autonomous driving, lack of transparency in model decision-making may reduce trust and reliability in AI-based systems. To overcome this limitation, researchers have increasingly explored the use of Explainable Artificial Intelligence (XAI) techniques to interpret machine learning models used in anomaly detection tasks.

Explainable AI methods such as SHapley Additive explanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME) provide insights into how individual features contribute to model predictions. These techniques enable researchers to identify the most influential features affecting anomaly detection outcomes and provide both global and local explanations for classification decisions. By integrating explainability into anomaly detection frameworks, XAI techniques improve transparency, enhance model reliability, and support the development of trustworthy intelligent transportation systems [1], [2], [9], [12].

Although significant progress has been made in applying machine learning and explainable AI techniques for anomaly detection in autonomous driving environments, several challenges still remain. Autonomous vehicle datasets often contain high-dimensional features collected from multiple sensors and communication channels, which can increase computational complexity and affect model performance. Additionally, identifying the most relevant features for anomaly detection remains an important research problem. Therefore, there is a growing need for intelligent frameworks that combine machine learning, feature selection, and explainable AI techniques to provide accurate and interpretable anomaly detection solutions for autonomous driving systems.

Overall, existing studies demonstrate that integrating machine learning models with explainable AI methods offers a promising direction for improving anomaly detection in autonomous vehicular networks. Such approaches not only enhance detection accuracy but also provide transparent explanations for AI-based decisions, which is essential for ensuring safety, reliability, and trust in autonomous transportation technologies.

III. SYSTEM ANALYSIS

A. Existing System

Traditional security monitoring approaches for autonomous driving systems primarily rely on rule-based mechanisms and manual analysis of network activity to identify abnormal vehicle behavior. In these systems, security mechanisms monitor communication messages exchanged among vehicles in Vehicular Ad Hoc Networks (VANETs) and attempt to detect anomalies based on predefined thresholds or fixed rules. Such approaches analyze parameters such as vehicle position, speed, message frequency, and communication patterns to identify suspicious activities. Although these techniques provide a basic level of monitoring, they are often insufficient for detecting sophisticated cyber-attacks or unexpected behaviours within large-scale autonomous driving environments.

With the rapid development of intelligent transportation systems, researchers have increasingly explored the use of data-driven techniques for anomaly detection in autonomous vehicle networks. Autonomous vehicles generate massive volumes of sensor data and communication logs through cameras, LiDAR sensors, radar systems, and vehicle-to-vehicle communication channels. Analyzing these large and complex datasets using traditional monitoring approaches is extremely challenging. Consequently, machine learning techniques have been widely applied to automatically learn patterns from vehicular datasets and detect abnormal behaviours more effectively [3], [4].

Several machine learning algorithms have been used for anomaly detection in autonomous driving systems. Classification techniques such as Logistic Regression, Decision Trees, Support Vector Machines (SVM), and K-Nearest Neighbour (KNN) have been employed to identify abnormal vehicle behavior based on features such as vehicle position, speed, and sensor readings. These models are capable of learning relationships among multiple attributes and can classify vehicles as normal or anomalous by identifying deviations from learned behavioural patterns [5], [7].

In addition to individual classifiers, ensemble learning approaches have also been adopted to enhance prediction performance in anomaly detection systems. Algorithms such as Random Forest and Adaptive Boosting (AdaBoost) combine multiple weak learners to create a stronger predictive model. These ensemble techniques often provide improved classification accuracy and robustness when dealing with high-dimensional vehicular datasets. Several studies have reported that ensemble learning models are particularly effective for detecting anomalies in network-based systems because they can capture complex feature relationships and reduce the risk of overfitting [8], [11].

Despite these advancements, many existing anomaly detection systems in autonomous driving environments still face several limitations. One of the primary challenges is that many high-performing machine learning models operate as black-box systems, where the reasoning behind classification decisions is not easily interpretable. In safety-critical applications such as autonomous driving, lack of transparency can reduce trust in automated decision-making systems. Additionally, autonomous vehicle datasets often contain high-dimensional features collected from multiple sensors and communication channels, which increases computational complexity and may negatively affect model performance if appropriate feature selection techniques are not applied [9], [12].

Furthermore, autonomous driving systems operate in dynamic environments where communication patterns and sensor readings may change frequently. Variations in sensor noise, environmental conditions, and network behavior can influence the accuracy of anomaly detection algorithms. These challenges highlight the need for more robust and interpretable frameworks that combine machine learning techniques with explainable artificial intelligence methods to improve both detection accuracy and model transparency.

Limitations Of Existing System

- Traditional rule-based monitoring systems rely on predefined thresholds and may fail to detect complex or unknown anomalies in autonomous vehicle networks.
- Autonomous vehicle datasets contain large volumes of sensor and communication data, making manual analysis and traditional monitoring techniques inefficient for large-scale systems.
- High-dimensional feature spaces generated from multiple sensors can increase computational complexity and reduce model efficiency if appropriate feature selection techniques are not implemented.
- Many machine learning models used for anomaly detection operate as black-box systems, making it difficult to interpret the reasoning behind prediction results.
- Lack of interpretability in anomaly detection models reduces transparency and user trust in

AI-based decision-making systems for autonomous driving environments.

- Existing approaches often focus on improving prediction accuracy without providing clear explanations for anomaly detection decisions.

B. Proposed System

This section presents the proposed Explainable Artificial Intelligence (XAI)-based anomaly detection framework for autonomous driving systems. The proposed system integrates data preprocessing, machine learning classification, feature selection, and explainability techniques to provide accurate and interpretable anomaly detection in Vehicular Ad Hoc Networks (VANETs).

In the proposed approach, autonomous driving datasets containing vehicle behavioural information and sensor data are first collected and pre-processed. The preprocessing stage includes removing redundant records, handling missing values, balancing datasets, and normalizing feature values to improve the quality of the training data. These preprocessing steps help ensure that the machine learning models can effectively learn patterns from the dataset.

After preprocessing, relevant features such as vehicle position, speed, and communication attributes are extracted from the dataset. Feature selection techniques are applied to identify the most informative attributes that contribute to anomaly detection. Reducing unnecessary features helps improve computational efficiency while maintaining predictive accuracy.

The processed data are then used to train multiple machine learning classifiers, including Decision Tree, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbour (KNN), Deep Neural Network (DNN), and AdaBoost. These models learn patterns from the training dataset and classify vehicles as normal or anomalous based on their behavioural characteristics. Ensemble learning models such as Random Forest and AdaBoost are expected to achieve higher detection accuracy due to their ability to combine multiple decision trees and capture complex relationships within the dataset.

To improve interpretability, the proposed system integrates Explainable Artificial Intelligence (XAI) techniques such as SHapley Additive explanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME). These techniques provide insights into how machine learning models make classification decisions by identifying the most influential features contributing to anomaly detection. SHAP provides global feature importance analysis, while LIME explains individual prediction results at the local level.

The main objective of the proposed framework is to develop a secure, interpretable, and efficient anomaly detection system for autonomous driving environments. By combining machine learning algorithms with explainable AI techniques, the proposed system improves detection accuracy while also providing transparent explanations for model predictions. This approach enhances the reliability, transparency, and trustworthiness of AI-based anomaly detection systems in intelligent transportation networks.

IV.SYSTEM DESIGN

System Architecture

Below diagram depicts the whole system architecture.

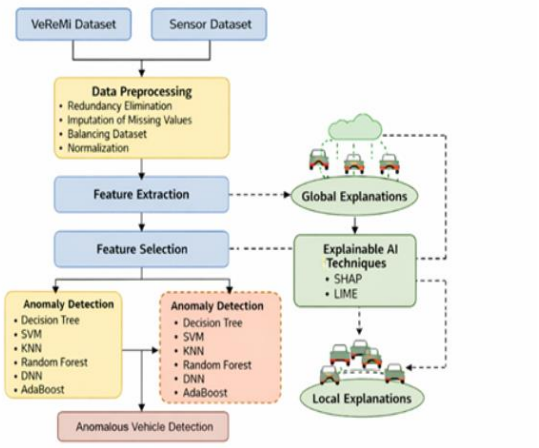

Fig 1. Methodology followed for proposed model

V. SYSTEM IMPLEMENTATION

Modules

This section describes the major implementation modules of the proposed Explainable Artificial Intelligence (XAI)-based anomaly detection framework for autonomous driving systems. The system follows a structured pipeline consisting of data acquisition, preprocessing, feature selection, machine learning model training, explainability integration, and prediction evaluation. This modular design improves the efficiency, scalability, and interpretability of the anomaly detection system in autonomous vehicle networks.

A. Data Collection Module

The Data Collection Module is responsible for gathering autonomous driving datasets containing vehicle behavioural information and communication data. In this study, datasets such as the VeReMi dataset and Sensor dataset are used to simulate realistic autonomous driving environments. These datasets contain various attributes describing vehicle behavior, including vehicle position, speed, communication parameters, and sensor measurements. Such data represent interactions among vehicles within Vehicular Ad Hoc Networks (VANETs). The collected datasets include both normal vehicle behavior and anomalous activity samples, enabling supervised machine learning models to learn patterns associated with abnormal behaviours. Autonomous driving datasets are typically high-dimensional because they contain multiple features collected from different sensors and communication sources. Additionally, these datasets may include irregularities such as redundant records, missing values, and imbalanced class distributions. Therefore, the collected data are forwarded to the preprocessing stage to ensure proper data quality before training machine learning models.

B. Data Preprocessing Module

The Data Preprocessing Module improves the quality of the collected dataset and prepares it for machine learning model training. Autonomous vehicle datasets often contain noise, redundant data, and inconsistent feature values that can affect model performance if not properly handled.

The preprocessing stage includes the following steps:

1) Redundancy Removal: Duplicate or irrelevant data entries are removed from the dataset to eliminate unnecessary information and improve training efficiency.

2) Data Balancing: Many real-world anomaly detection datasets contain significantly fewer anomalous samples compared to normal samples. To address this imbalance, dataset balancing techniques such as under sampling or oversampling are applied to ensure that the machine learning models learn effectively from both classes.

3) Feature Normalization: Feature normalization techniques such as standard scaling are applied to transform feature values into a consistent numerical range. This step prevents features with large numerical ranges from dominating the model training process and improves overall learning performance.

4) Data Cleaning: Missing values and inconsistent records are handled through appropriate data cleaning methods to ensure the dataset remains reliable and suitable for model training.

5) These preprocessing steps enhance dataset quality and ensure that machine learning models can accurately learn behavioural patterns from autonomous vehicle data.

C. Feature Selection Module

Autonomous vehicle datasets may contain a large number of features collected from various sensors and communication channels. High-dimensional datasets can increase computational complexity and may negatively affect model performance if irrelevant features are included.Therefore, a Feature Selection Module is introduced to identify the most

important features contributing to anomaly detection.Feature importance is initially evaluated using machine learning models and statistical analysis techniques. In addition, Explainable Artificial Intelligence (XAI) methods such as SHapley Additive Explanations (SHAP) are used to rank features based on their contribution to model predictions.

By selecting only the most relevant features, the system reduces dataset dimensionality while maintaining prediction accuracy. This step improves computational efficiency, accelerates model training, and enhances interpretability of the anomaly detection framework [1], [2], [8].

D. Machine Learning Training Module

The Machine Learning Training Module builds classification models to identify anomalous vehicle behavior within autonomous driving networks.
Several machine learning algorithms are implemented and evaluated in the proposed framework, including:

- Decision Tree
- Random Forest
- Support Vector Machine (SVM)
- K-Nearest Neighbour (KNN)
- Deep Neural Network (DNN)
- AdaBoost

Each algorithm is trained using the processed dataset containing both normal and anomalous vehicle samples. During training, the models learn behavioural patterns that differentiate normal vehicle operations from abnormal activities within the network.

Model evaluation is performed using training and testing datasets to ensure that the classifiers generalize well to unseen data. Among the evaluated models, ensemble learning algorithms such as Random Forest and AdaBoost often provide strong predictive performance due to their ability to combine multiple decision trees and capture complex relationships within high-dimensional datasets [5], [7].

E. Explainability Module (XAI Integration)

To improve transparency and interpretability, the proposed framework integrates Explainable Artificial Intelligence (XAI) techniques. Many machine learning models used in anomaly detection operate as black-box systems, which makes it difficult to understand how predictions are generated. In safety-critical systems such as autonomous driving, interpretability is essential for building trust in AI-based decision-making systems.

Two widely used XAI techniques are implemented in the proposed framework:

SHAP (SHapley Additive Explanations):SHAP provides global and local explanations by measuring the contribution of each feature to the prediction outcome. It calculates Shapley values based on cooperative game theory to determine the importance of individual features in model predictions.

LIME (Local Interpretable Model-Agnostic Explanations):LIME explains individual predictions by approximating the behavior of the machine learning model around a specific data instance. This method helps identify which features contributed to a particular classification result.

By integrating SHAP and LIME, the proposed system provides interpretable explanations for anomaly detection decisions and allows system operators to understand the factors influencing model predictions [1], [2], [8], [12].

F. Prediction and Evaluation Module

The Prediction and Evaluation Module generates the final anomaly detection results and evaluates model performance.

The system output includes:

- Classification result: Normal Vehicle / Anomalous Vehicle
- Prediction probability score
- Feature importance explanations from XAI techniques

To evaluate the performance of the anomaly detection models, several standard machine learning evaluation metrics are used:

- Accuracy
- Precision
- Recall
- F1-Score
- ROC–AUC Score

These evaluation metrics provide a comprehensive assessment of the model's ability to detect anomalies in autonomous driving datasets. Accurate anomaly detection is essential for ensuring the safety and reliability of autonomous vehicle networks.

By identifying abnormal vehicle behavior at an early stage, the proposed framework enhances security within vehicular networks and supports the development of reliable and trustworthy intelligent transportation systems.

VI. RESULTS AND DISCUSSION

This section presents the experimental results and performance evaluation of the proposed Explainable Artificial Intelligence (XAI)-based anomaly detection framework for autonomous driving systems. Multiple machine learning algorithms were trained and evaluated using autonomous driving datasets containing both normal and anomalous vehicle behaviours. The evaluation focuses on comparing model performance, analyzing classification accuracy, and interpreting feature contributions using explainable AI techniques.

A. Accuracy Comparison of Machine Learning Models

Several machine learning algorithms were evaluated to determine the most suitable model for anomaly detection in autonomous vehicle networks. The evaluated models include Decision Tree, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbour (KNN), Deep Neural Network (DNN), and AdaBoost. Model performance was assessed using standard evaluation metrics such as accuracy, precision, recall, and F1-score.

Table 1. Performance Comparison of Machine Learning Models

Model	**Accuracy (%)**	**Precision**	**Recall**	**F1-Score**
Decision Tree	84.3	0.82	0.81	0.81
K-Nearest Neighbour	86.1	0.84	0.83	0.83
Support Vector Machine	87.5	0.86	0.85	0.85
Random Forest	89.8	0.88	0.87	0.87
Deep Neural Network	88.6	0.87	0.86	0.86
AdaBoost	90.4	0.89	0.88	0.88

From the comparison results, AdaBoost achieved the highest classification accuracy of 90.4%, followed

closely by Random Forest and Deep Neural Networks. The superior performance of ensemble learning models can be attributed to their ability to combine multiple weak classifiers and capture complex feature relationships within autonomous driving datasets. These results demonstrate that ensemble methods are highly effective for anomaly detection in vehicular networks [5], [7].

B. ROC Curve Analysis

The Receiver Operating Characteristic (ROC) curve is used to evaluate the trade-off between the True Positive Rate (TPR) and False Positive Rate (FPR) at different classification thresholds. The Area Under the Curve (ROC–AUC) provides a comprehensive measure of a model's ability to distinguish between normal and anomalous vehicle behaviors.

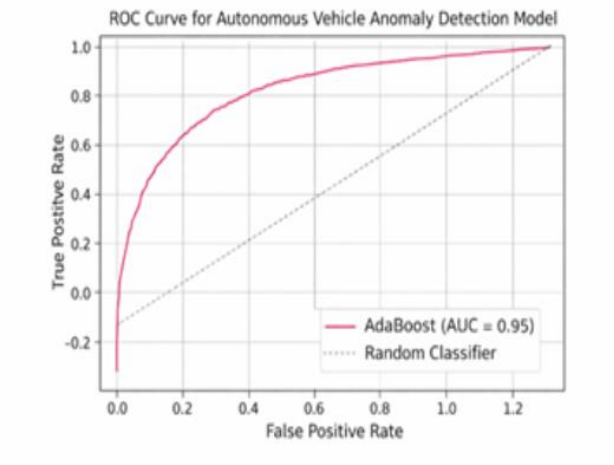

Fig 2. ROC Curve for Autonomous Vehicle Anomaly Detection Model

In this study, the AdaBoost classifier achieved a ROC–AUC score of 0.95, indicating strong discriminative capability in identifying anomalous vehicle behavior. The ROC curve is positioned close to the top-left corner of the graph, which suggests that the model performs well in distinguishing between normal and abnormal vehicle activities within the network.

The ROC analysis demonstrates that the proposed anomaly detection framework maintains reliable predictive performance even when dealing with complex and high-dimensional autonomous vehicle datasets.

C. SHAP Feature Importance Analysis

To enhance transparency and interpretability, SHapley Additive Explanations (SHAP) were used to analyze the contribution of each feature to the anomaly detection model predictions. SHAP values measure the impact of individual features on prediction outcomes based on concepts derived from cooperative game theory.

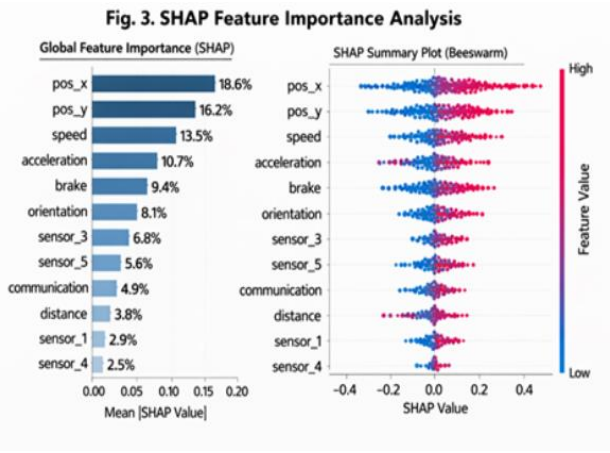

Fig. 3. SHAP Feature Importance Analysis

Fig 3. SHAP Feature Importance Analysis

The SHAP analysis revealed that several key vehicle attributes such as vehicle position coordinates (pos_x, pos_y), vehicle speed attributes, and communication consistency features had the greatest influence on anomaly detection results. Features with higher SHAP values contributed more significantly to identifying abnormal vehicle behavior in the dataset.

The global SHAP summary plot illustrates the relative importance of each feature across the entire dataset, while local SHAP explanations provide insights into how specific features influence predictions for individual vehicles.

The integration of SHAP explanations significantly improves the interpretability of the anomaly detection framework. It enables system operators and researchers to better understand the reasoning behind model predictions and identify the most influential factors contributing to anomalous vehicle behavior. This transparency is particularly important in safety-critical autonomous driving environments

where trustworthy AI decision-making is essential [1], [2], [8], [12].

VII. CONCLUSION AND FUTURE WORK

This study presented an Explainable Artificial Intelligence (XAI)-based framework for anomaly detection in autonomous driving systems operating within Vehicular Ad Hoc Networks (VANETs). Autonomous vehicles generate large volumes of sensor and communication data, which makes detecting abnormal behaviours challenging using traditional monitoring methods. To address this issue, the proposed framework integrates machine learning algorithms with explainable AI techniques to accurately detect anomalous vehicle behavior while providing interpretable insights into model predictions.

In the proposed system, autonomous driving datasets were first pre-processed using redundancy removal, dataset balancing, and feature normalization techniques to improve data quality. Feature selection methods were then applied to identify the most relevant attributes influencing anomaly detection. Several machine learning models, including Decision Tree, Support Vector Machine (SVM), K-Nearest Neighbour (KNN), Random Forest, Deep Neural Network (DNN), and AdaBoost, were implemented and evaluated for anomaly classification. Experimental results demonstrated that ensemble learning models such as AdaBoost and Random Forest achieved the highest classification accuracy, indicating their strong capability to handle high-dimensional autonomous vehicle datasets and detect abnormal vehicle behaviours effectively [5], [7].

To enhance transparency and interpretability, the proposed framework incorporated Explainable Artificial Intelligence (XAI) techniques such as SHapley Additive Explanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME). These techniques provide both global and local explanations by identifying the most influential features contributing to anomaly detection decisions. The integration of XAI improves the reliability and trustworthiness of AI-based decision-making systems, which is particularly important for safety-critical applications such as autonomous driving environments [1], [2], [8], [12].

Overall, the proposed framework demonstrates that combining machine learning with explainable AI techniques can significantly improve the effectiveness and transparency of anomaly detection systems in autonomous vehicle networks. The system not only achieves high prediction accuracy but also provides meaningful explanations for model decisions, enabling researchers and system operators to better understand abnormal vehicle behaviours.

Future research may focus on integrating real-time vehicular communication data and IoT sensor streams to enhance the system's capability for real-time anomaly detection. In addition, advanced deep learning architectures and hybrid ensemble models can be explored to further improve detection accuracy and scalability. Deployment of the proposed framework in cloud-based intelligent transportation infrastructures may also support large-scale monitoring and security management for next-generation autonomous vehicle ecosystems.

REFERENCES

1. S. M. Lundberg and S. I. Lee, “A unified approach to interpreting model predictions,” Advances in Neural Information Processing Systems, vol. 30, pp. 4765–4774, 2017.
2. M. T. Ribeiro, S. Singh, and C. Guestrin, “Why should I trust you? Explaining the predictions of any classifier,” in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, San Francisco, CA, USA, 2016, pp. 1135–1144.
3. H. Hartenstein and K. Laberteaux, “A tutorial survey on vehicular ad hoc networks,” IEEE Communications Magazine, vol. 46, no. 6, pp. 164–171, Jun. 2008.
4. J. Petit and S. E. Shladover, “Potential cyberattacks on automated vehicles,” IEEE Transactions on Intelligent Transportation Systems, vol. 16, no. 2, pp. 546–556, Apr. 2015.
5. L. Breiman, “Random forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

6. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 785–794.
7. C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.
8. D. B. Rawat, S. R. Shetty, and C. Bajracharya, "Security issues and challenges in vehicular ad hoc networks," IEEE Communications Surveys & Tutorials, vol. 19, no. 2, pp. 995–1024, 2017.
9. F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
10. A. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," arXiv preprint arXiv:1702.08608, 2017.
11. A. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.
12. B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2921–2929.