XAI-ADS: An Explainable Artificial Intelligence Framework for Enhancing Anomaly Detection in Autonomous Driving Systems

Sazid Nazat1, Lingxi Li1, and Mustafa Abdallah1

1Indiana University-Purdue University Indianapolis, Indiana, USA

snazat@iu.edu, li7@pui.edu, mabdall@iu.edu

Introduction

- Autonomous vehicles (AVs) are vulnerable to many cyber-attacks, which might cause them to behave abnormally in vehicular ad hoc network (VANET). This disrupts the VANET.
- AI models are employed to classify anomalous AVs in the VANET, ensuring that it operates smoothly. XAI is then used to explain the decisions made by AI models in detecting anomalous AVs in a way that human operators (for example, safety drivers) can understand.
- To build trust between human operators and the AI models used for securing autonomous vehicles (AVs), XAI provides explanations to back up the AI predictions with evidence.

Motivation

- The complicated feature interactions in AI models make it hard to grasp how and why they make decisions. For most AI models, this black-box issue makes it hard to explain their behavior, especially when mistakes arise.
- AI models have good prediction accuracy in various autonomous driving sectors, but their decision-making process must be interpreted and explained, especially in safety-critical domains.
- XAI can build trust and ensure accountability through explainable AI model decisions and insights. Also, it can enhance security and safety by understanding AI decision-making process and providing necessary information.

Our Contribution

- Propose an end-to-end XAI framework offering global and local explanations that provide human-understandable justifications for AI models’ decisions in detecting anomalous AVs.
- Introduce two novel XAI-based feature selection approaches to identify and rank significant features contributing to anomaly classification of AVs and compare them with baseline feature selection methods, evaluated on three independent classifiers to ensure unbiased results.
- Benchmark the anomaly detection problem using six different AI models and two XAI methods on two autonomous driving datasets, analyzing the efficiency of the black-box models and the XAI techniques employed.

Framework

- Two Novel Feature Selection Methods
  - SHAP-based
  - SHAP-AI based

Evaluation

- SHAP on VeReMi dataset
- LIME on Sensor dataset

Results

- Overall Performance of AI models
  a) VeReMi dataset
  b) Sensor dataset

Effect of Class Balancing

- XAI Explanations vs Black Box AI Models

Comparison with Baseline Feature Selection Models

- TABLE 13: Results per class in every feature selection method on VeReMi dataset under k = 4 for feature selection methods.

Key insight:

- SHAP-based outperforms SHAP-AI based in feature ranking and performed equally well.
- Our XAI framework outperforms the black box AI models, with XAI models having better efficiency than SHAP.
- XAI models depicted high performance (except the SVM model) on Sensor dataset.
- AI models perform better when the percentage of anomalous data is reduced (less number of anomalous samples) except for the DNN model.

Acknowledgement

This work was supported by Lilly Endowment (AnalytiXIN); Enhanced Mentoring Program with Opportunities for Ways to Excel in Research (EMPOWER), and the 1st Year Research Immersion Program (1RIP) Grants from the Office of the Vice Chancellor for Research at Indiana University – Purdue University Indianapolis.