CERIAS

The Center for Education and Research in Information Assurance and Security

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

Sazid Nazat[1], Lingxi Li[1], and Mustafa Abdallah[1]

[1]*Indiana University-Purdue University Indianapolis, Indiana, USA*

snazat@iu.edu, ll7@iupui.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.

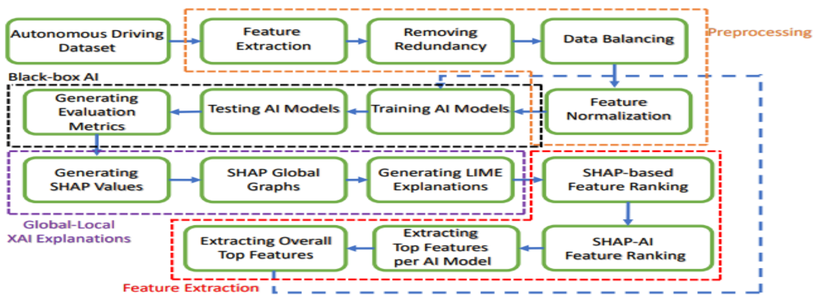
Motivation

- 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

Statistics of two datasets

	VeReMi dataset	Sensor dataset
Number of Labels	2	2
Number of Features	6	10
Dataset Size	993,834	70,000
Training Sample	695,684	49,000
Testing Sample	298,150	21,000
Normal Samples No.	664,131	35,000
Anomalous Samples No.	329,703	35,000

Description of datasets

a) VeReMi dataset

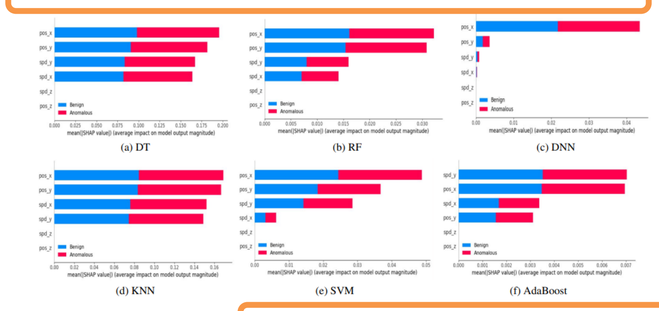
TABLE 3: Description of main features of VeReMi dataset.

Column	Description
pos_x	The x-coordinate of the vehicle position
pos_y	The y-coordinate of the vehicle position
pos_z	The z-coordinate of the vehicle position
spd_x	Speed of the vehicle in the x-direction
spd_y	Speed of the vehicle in the y-direction
spd_z	Speed of the vehicle in the z-direction

b) Sensor dataset

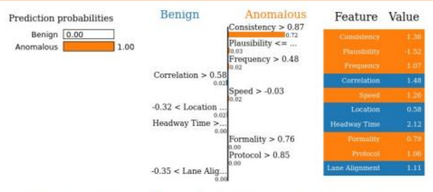
Sensor Name	Normal data range	Description
Formality	1bit – 10bit	Checks every message if it is maintaining correct formality
Location	0/1	Checks if the message reached the destined location
Frequency	1Hz – 10Hz	Checks the interval time of messages
Speed	50mph – 90mph	Checking if the AV is in the speed limit (Highway)
Correlation	0/1	Checks if several messages adheres to defined specification
Lane Alignment	1-3	Checks if the AV is in the correct lane
Headway Time	0.3s - 0.95s	Checks if the AV maintains the headway time range
Protocol	1-10000	Checks for the correct order of communication messages
Plausibility	50% - 200%	Checks if the data is plausible (relative difference between sizes of two consecutive payloads).
Consistency	0/1	Checks if all the parts of the AV are delivering consistent information about an incident

SHAP on VeReMi dataset



Key insight: "pos_x" is the top-ranked feature for the VeReMi dataset as majority of the AI models indicate.

LIME on Sensor dataset



(a) Example of a local explanation for correctly predicting an anomalous AV from the Sensor dataset.

Results

Overall Performance of AI models

a) VeReMi dataset

TABLE 5: Overall performances for AI models with top 6 features for the VeReMi dataset.

AI Model	Accuracy	Precision	Recall	F-1 score	AucRoc
DT	0.79	0.82	0.87	0.84	0.74
RF	**0.80**	0.83	0.88	**0.86**	**0.76**
DNN	0.66	0.67	**0.96**	0.79	0.50
KNN	0.78	0.82	0.85	0.84	0.75
SVM	0.55	0.54	0.66	0.60	0.54
AdaBoost	**0.80**	**0.84**	0.86	0.85	0.59

Key insight: AdaBoost and RF classifiers give the best overall performance on VeReMi dataset.

b) Sensor dataset

TABLE 6: Overall performances for AI models with top 10 features for the Sensor dataset.

AI Model	Accuracy	Precision	Recall	F-1 score	AucRoc
DT	1	1	1	1	1
RF	1	1	1	1	1
DNN	1	1	1	1	1
KNN	1	1	1	1	1
SVM	0.97	1	0.93	0.96	0.96
AdaBoost	1	1	1	1	1

Key insight: All AI models depicted high performance (except the SVM model) on Sensor dataset.

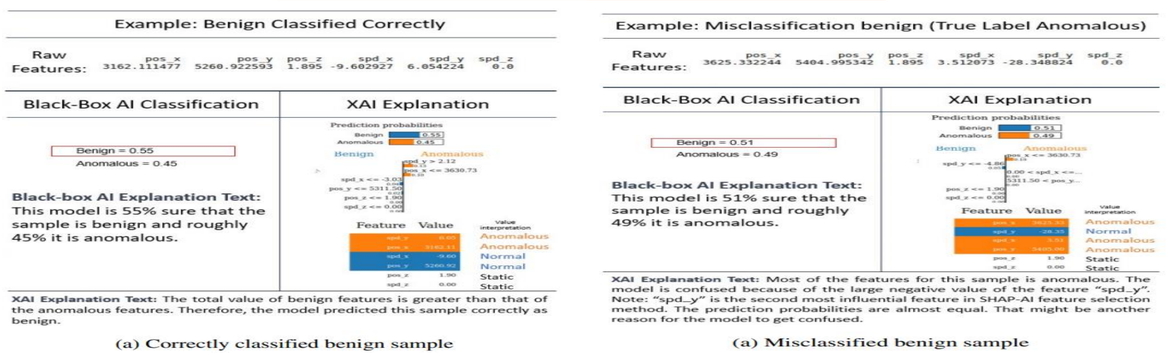
Effect of Class Balancing

	DT				RF				DNN			
Benign percentage	50%	60%	70%	80%	50%	60%	70%	80%	50%	60%	70%	80%
Acc	0.74	0.74	0.75	**0.76**	0.77	0.77	0.78	**0.78**	0.50	0.49	**0.54**	0.51
Prec	0.75	0.76	0.77	**0.79**	0.78	0.78	0.80	**0.81**	0.51	0.55	**0.58**	0.54
Rec	0.76	0.77	0.82	**0.84**	0.80	0.79	0.83	**0.85**	0.64	0.68	**0.73**	0.67
F-1	0.73	0.78	0.80	**0.81**	0.79	0.79	0.81	**0.83**	0.56	0.55	**0.65**	0.60

	KNN				SVM				AdaBoost			
Benign percentage	50%	60%	70%	80%	50%	60%	70%	80%	50%	60%	70%	80%
Acc	0.72	0.72	0.74	**0.75**	0.52	0.55	0.58	**0.61**	0.76	0.76	0.77	**0.78**
Prec	0.74	0.74	**0.76**	0.76	0.52	0.55	0.59	**0.61**	0.78	0.79	0.80	**0.82**
Rec	0.68	0.68	**0.71**	0.71	0.55	0.95	0.99	**1**	0.73	0.78	0.80	**0.83**
F-1	0.71	0.72	**0.73**	0.73	0.53	0.70	0.74	**0.76**	0.75	0.78	0.80	**0.82**

Key insight: AI models perform better when the percentage of anomalous data is **reduced** (less number of anomalous samples) except for the DNN model.

XAI Explanations vs Black Box AI Models



Evaluation on Independent Classifiers

TABLE 8: Performance of top-k features of SHAP-based and SHAP-AI feature selection methods on VeReMi dataset for the three independent classifiers (LR, LGBM, and XGBoost).

	Evaluation	LR		LGBM		XGBoost	
		SHAP-based	SHAP-AI based	SHAP-based	SHAP-AI based	SHAP-based	SHAP-AI based
k=1	Accuracy	0.50	0.50	0.67	0.67	0.67	0.67
	Precision	0.67	0.67	0.67	0.67	0.67	0.67
	Recall	0.49	0.49	0.99	0.99	0.99	0.99
	F-1 score	0.56	0.56	0.80	0.80	0.80	0.80
k=2	Accuracy	0.54	0.53	0.67	0.67	0.68	0.68
	Precision	0.67	0.68	0.67	0.67	0.68	0.68
	Recall	0.61	0.56	0.99	0.99	0.98	0.97
	F-1 score	0.64	0.61	0.80	0.80	0.80	0.80
k=3	Accuracy	0.56	0.57	0.68	0.67	0.69	0.69
	Precision	0.68	0.68	0.67	0.67	0.69	0.69
	Recall	0.65	0.67	0.99	0.99	0.97	0.97
	F-1 score	0.66	0.68	0.80	0.80	0.81	0.81
k=4	Accuracy	0.54	0.54	0.68	0.68	**0.70**	**0.70**
	Precision	0.67	0.68	0.67	0.67	**0.70**	**0.70**
	Recall	0.60	0.61	**1**	**1**	0.97	0.97
	F-1 score	0.63	0.64	0.80	0.80	**0.81**	**0.81**
k=5	Accuracy	0.54	0.54	0.67	0.67	0.69	0.69
	Precision	0.68	0.68	0.67	0.67	0.69	0.69
	Recall	0.60	0.61	1	1	0.97	0.97
	F-1 score	0.63	0.64	0.80	0.80	0.81	0.81
k=6	Accuracy	0.54	0.54	0.67	0.67	0.69	0.69
	Precision	0.67	0.67	0.67	0.67	0.69	0.69
	Recall	0.60	0.60	1	1	0.97	0.97
	F-1 score	0.64	0.64	0.80	0.80	0.81	0.81

Key insight: To rigorously evaluate the proposed feature selection approaches **without bias**, we used **three independent classifiers**, logistic regression (LR), light gradient boosting machine (LGBM), and extreme gradient boosting (XGBoost), which were not involved in feature ranking and performed well.

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.

Model		SHAP				SHAP-AI				K-best				PCA				ET				IG			
	Class	Acc	Prec	Rec	F-1	Acc	Prec	Rec	F-1	Acc	Prec	Rec	F-1	Acc	Prec	Rec	F-1	Acc	Prec	Rec	F-1	Acc	Prec	Rec	F-1
DT	Benign (0)	0.76	0.78	**0.90**	0.83	**0.79**	**0.82**	0.87	**0.84**	0.78	0.82	0.87	0.84	0.78	0.82	0.87	0.84	0.78	0.82	0.87	0.84	0.78	0.81	0.87	0.84
	Anomalous (1)		**0.70**	0.49	0.50		0.70	**0.61**	**0.66**		0.70	0.61	0.65		0.70	0.61	0.65		0.70	0.61	0.65		0.70	0.60	0.65
RF	Benign (0)	**0.80**	**0.83**	**0.88**	**0.86**	0.80	0.83	0.88	0.85	0.80	0.83	0.88	0.86	0.80	0.83	0.88	0.85	0.80	0.83	0.88	0.86	0.80	0.83	0.88	0.85
	Anomalous (1)		**0.73**	**0.64**	**0.68**		0.72	0.64	0.68		0.73	0.64	0.68		0.73	0.64	0.68		0.73	0.64	0.68		0.73	0.63	0.68
DNN	Benign (0)	0.66	0.67	0.96	0.79	0.41	0.67	0.24	0.36	**0.67**	**0.67**	**1**	**0.80**	0.54	0.67	0.64	0.65	0.65	0.67	0.96	0.79	0.64	0.67	0.92	0.77
	Anomalous (1)		0.33	0.04	0.07		**0.33**	**0.75**	**0.46**		0.33	0.27	0.29		0.33	0.36	0.34		0.33	0.04	0.08		0.30	0.07	0.11
KNN	Benign (0)	**0.78**	**0.82**	**0.85**	**0.84**	0.78	0.82	0.85	0.84	0.78	0.82	0.85	0.84	0.78	0.82	0.85	0.84	0.78	0.82	0.85	0.84	0.78	0.82	0.85	0.83
	Anomalous (1)		**0.68**	**0.63**	**0.65**		0.68	0.63	0.65		0.68	0.64	0.66		0.68	0.63	0.65		0.68	0.63	0.65		0.67	0.63	0.65
SVM	Benign (0)	0.55	0.54	0.69	0.61	0.55	0.54	0.66	0.60	**0.67**	**0.69**	**0.93**	**0.79**	0.65	0.68	0.92	0.78	0.66	0.69	0.90	0.78	0.67	0.69	0.92	0.79
	Anomalous (1)		0.57	0.41	0.48		**0.57**	**0.42**	**0.48**		0.50	0.14	0.22		0.50	0.14	0.22		0.50	0.15	0.23		0.49	0.16	0.24
AdaBoost	Benign (0)	0.79	**0.84**	0.85	0.85	**0.80**	0.83	**0.86**	**0.85**	0.79	0.83	0.86	0.85	0.79	0.84	0.86	0.85	0.78	0.84	0.85	0.84	0.79	0.84	0.85	0.84
	Anomalous (1)		0.70	0.65	0.67		**0.70**	**0.66**	**0.68**		0.70	0.65	0.68		0.70	0.66	0.68		0.70	0.66	0.62		0.69	0.66	0.68

Key insight: Compared to six baseline methods across AI models, the novel feature selection methods **excel in DT** and **AdaBoost models**. They also improve precision, recall, and F1-score for anomaly detection in most of the models.

Efficiency of XAI models

a) VeReMi dataset

XAI model	# Samples	DT	RF	DNN	KNN	SVM	ADA
SHAP	500	0.72	1.84	0.13	2.47	19.09	1.13
	1k	2.97	6.29	0.15	9.06	93.66	4.15
	10k	219.66	333.31	6.98	929.98	6711.26	477.23
	50k	8054.65	13256.81	NA	22128.2	271275	8828.6
LIME	500	0.12	2.48	4.65	4.94	25.53	1.04
	1k	0.43	3.11	5.66	5.16	24.71	0.65
	10k	27.24	43.1	95.90	89	379.38	18.88
	50k	27.25	131.88	501.67	287.75	NA	56.41

Key insight: **LIME** has better efficiency than SHAP for majority of the AI models.

b) Sensor dataset

XAI model	# Samples	DT	RF	DNN	KNN	SVM	AdaBoost
SHAP	500	10.22	20.09	**0.08**	329.04	20.43	10.43
	1k	33.56	121.20	**0.15**	866.01	61.29	50.28
	10k	3826.31	8291	**8.20**	97237.36	9513.15	3391.36
LIME	500	**0.17**	0.39	3	5.98	0.48	0.31
	1k	**0.68**	1.17	6.25	13.97	0.74	1.15
	10k	**10.27**	19.39	66.56	75.46	17.64	11.32

Key insight: **LIME** has better efficiency for Sensor dataset compared to SHAP.

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.

PURDUE UNIVERSITY

CERIAS