rp-马超宇0718 - 副本.docx

by User User

Submission date: 18-Jul-2025 02:03AM (UTC-0700)
Submission ID: 2714824437
File name: rp-马超宇0718_-_副本.docx (16.72K)
Word count: 563
Character count: 3753

An Explainable Time-Series Based Driving Stability Model for Fault Detection in Autonomous Vehicles

1 Introduction

With the growing adoption of autonomous vehicles (AVs), ensuring system safety in dynamic and uncertain environments has become a critical challenge. Faults arising from sensor degradation, actuator failures, or software anomalies can lead to unstable driving behavior and catastrophic consequences. While deep learning–based methods have shown promise in anomaly detection, most operate as black-box models, making them difficult to audit or trust in safety-critical applications. This has prompted growing interest in integrating Explainable Artificial Intelligence (XAI) [1] into AV fault detection frameworks to improve transparency and human interpretability.

2 Related Work

Fault detection in autonomous vehicles has evolved from rule-based and redundancy-driven methods to time-series learning approaches [2,3] that better capture complex temporal anomalies. Techniques such as DAGMM [4] and MSALSTM-CNN [5] leverage deep autoencoding, attention, and sequential modeling to improve detection accuracy. However, most remain black-box systems, underscoring the need for interpretable solutions in safety-critical scenarios.

3 Research Gaps

Current methods emphasize classification over fluctuation-aware health monitoring. They lack the ability to model normal vehicle dynamics over time and detect subtle deviations that signal early-stage faults. Moreover, there is limited integration of XAI techniques to explain model predictions, which hampers trust, debugging, and safety validation.

4 Methodology

This study proposes an explainable driving stability fault detection framework that enhances identification of gradual faults by fusing simulated and real-vehicle data through multi-scale temporal feature extraction [6] and cross-domain adaptation. It innovatively integrates attention-constrained knowledge distillation with SHAP [7] regularization to ensure interpretability reliability during model compression, and generates human-readable rules via neuro-symbolic conversion [8,9] based on high-attention features. We aim to provide local explanations of anomaly predictions in time, as well as generate global human-interpretable rules summarizing recurrent fault patterns. Evaluation will be performed on CARLA-based simulated datasets with injected fault scenarios, as well as real-world CAN bus logs from public datasets. We will assess anomaly detection accuracy (AUC, F1) and explanation fidelity (e.g., explanation consistency, domain expert validation).

5 Contributions

- Achieves improved accuracy in multi-source anomaly detection via robust time-series modeling
- Provides transparent reasoning for anomaly inference through integrated XAI techniques.
- Proposes a generalizable framework for stability-aware fault monitoring in autonomous vehicles

Enhances system trust and auditability through explanation fidelity evaluation and human-in-the-loop validation.

6 References

[1] Adadi, A., & Berrada, M.. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access. 2018.

[2] Malhotra, Pankaj, et al. "Long Short Term Memory Networks for Anomaly Detection in Time Series". *The European Symposium on Artificial Neural Networks*, 2015.

[3] Darban, Zahra Zamanzadeh, et al. "Deep Learning for Time Series Anomaly Detection: A Survey". *ACM Computing Surveys,* vol. abs/2211.05244, 2024.

[4] Zong, Bo, et al. "Deep Autoencoding Gaussian Mixture Model for Unsupervised Anomaly Detection". *International Conference on Learning Representations,* 2018.

[5] Javed, A. R., et al. "Anomaly Detection in Automated Vehicles Using Multistage Attention-Based Convolutional Neural Network". *IEEE transactions on intelligent transportation systems (Print),* 2021.

[6] Ren, Hansheng, et al. "Time-Series Anomaly Detection Service at Microsoft". *Knowledge Discovery and Data Mining*, 2019.

[7] Lundberg, Scott M., Su-In Lee "A unified approach to interpreting model predictions". *Neural Information Processing Systems,* 2017.

[8] Lakkaraju, Himabindu, et al. "Interpretable Decision Sets: A Joint Framework for Description and Prediction". *KDD,* 2016.

[9] Guidotti, Riccardo, et al. "A Survey of Methods for Explaining Black Box Models". *ACM Computing Surveys,* 2019.

rp-马超宇0718 - 副本.docx

ORIGINALITY REPORT

28%	27%	22%	22%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	export.arxiv.org Internet Source	4%
2	www.diva-portal.se Internet Source	4%
3	www.cic-chinacommunications.cn Internet Source	3%
4	ebin.pub Internet Source	3%
5	pure.southwales.ac.uk Internet Source	3%
6	www.phd-dauin.polito.it Internet Source	3%
7	Anwesha Das, Alex Aiken. "Prolego: Time-Series Analysis for Predicting Failures in Complex Systems", 2023 IEEE International Conference on Autonomic Computing and Self-Organizing Systems (ACSOS), 2023 Publication	3%
8	digitalworks.union.edu	

	Internet Source		3%
9	www.pure.ed.ac.uk Internet Source		2%
10	arxiv.org Internet Source		1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		