ANOMALY DETECTION AND ROOT-CAUSE DETERMINATION FOR AUTOMOTIVE APPLICATIONS USING DEEP LEARNING AND XAI MODELS

by
MEHMET EMİN MUMCUOĞLU

Submitted to the Faculty of Engineering and Natural Sciences
in partial fulfilment of
the requirements for the degree of Doctor of Philosophy

Sabancı University
July 2025

ANOMALY DETECTION AND ROOT-CAUSE DETERMINATION FOR AUTOMOTIVE APPLICATIONS USING DEEP LEARNING AND XAI MODELS

Approved by:

Prof. Dr. MUSTAFA ÜNEL .....
(Thesis Advisor)

Prof. Dr. KEMAL KILIÇ .....

Assoc. Prof. Dr. KEMALETTİN ERBATUR .....

Assoc. Prof. Dr. HÜSEYİN ÜVET .....

Assoc. Prof. Dr. ALİ FUAT ERGENÇ .....

Date of Approval: July 21, 2025

ABSTRACT

ANOMALY DETECTION AND ROOT-CAUSE DETERMINATION FOR AUTOMOTIVE APPLICATIONS USING DEEP LEARNING AND XAI MODELS

MEHMET EMİN MUMCUOĞLU

MECHATROCIS ENGINEERING Ph.D DISSERTATION, JULY 2025

Dissertation Supervisor: Prof. MUSTAFA ÜNEL

Keywords: Anomaly Detection, Predictive Maintenance, Explainable AI, Heavy-Duty Vehicles, Fuel Efficiency, Air Pressure System, LSTM Autoencoder, Human-in-the-Loop, Large Language Models

Anomaly detection in heavy-duty vehicles (HDVs) is crucial for predictive maintenance and efficient fleet management, yet it poses considerable challenges due to the complex interplay between mechanical systems, diverse operational conditions, and limited labeled data. Traditional diagnostic approaches often fall short, struggling with false alarms and lacking interpretability, which can undermine user trust and delay critical interventions. Addressing these challenges necessitates robust, data-driven anomaly detection frameworks that combine precision with explainability, informed by domain knowledge and human expertise. This thesis develops two tailored anomaly detection frameworks specifically designed for critical HDV applications: (1) detecting excessive fuel consumption under varying operational conditions, and (2) early detection of air pressure system (APS) failures. Excessive fuel consumption significantly impacts operational efficiency and regulatory compliance, whereas APS failures frequently result in costly breakdowns and downtime. Each application demands unique methodological considerations due to the inherent variability and complexity of the underlying data.

For fuel consumption anomaly detection, a novel quartile-based labeling method was introduced, considering weight-normalized fuel consumption and multi-level road slope segmentation. Utilizing bagged decision trees, this supervised approach clas-

sifies operational anomalies at high accuracy across diverse driving datasets from Turkey and Germany, achieving up to 92.2% accuracy and an F1 score of 0.78. An interactive fleet monitoring dashboard further provides actionable insights for fleet operators by visually identifying anomalous trips and facilitating targeted interventions. For APS failure detection, the thesis explores semi-supervised learning through Long Short-Term Memory (LSTM) Autoencoders, enhanced by a human-in-the-loop framework incorporating expert analysis. These models effectively identify subtle temporal deviations preceding mechanical failures with an overall F1 score of 0.75. Additionally, the Explainable Boosting Machine (EBM) model achieved an excellent balance of predictive accuracy (91.4%, F1 score: 0.80) and interpretability, complemented by a Large Language Model (LLM)-based agentic system that provides expert-level diagnostic reasoning and transparency.

This thesis emphasizes interpretability by integrating explainable AI techniques alongside human expertise, thus enhancing diagnostic reliability and user trust. These interpretable frameworks enable clear root-cause analysis, reduce false alarms, and improve practical decision-making across diverse operations. The developed methodologies offer versatile and adaptable solutions for sustainable fleet management, with potential future expansions toward real-time anomaly detection, multi-fault classification, and integration into automated, closed-loop predictive maintenance systems.

ÖZET

OTOMOTİV UYGULAMALARI İÇİN DERİN ÖĞRENME VE AÇIKLANABİLİR YAPAY ZEKA KULLANARAK ANOMALİ TESPİTİ VE KÖK-NEDEN ANALİZİ

MEHMET EMIN MUMCUOĞLU

MEKATRONİK MÜHENDİSLİĞİ DOKTORA TEZİ, TEMMUZ 2025

Tez Danışmanı: Prof. Dr. MUSTAFA ÜNEL

Anahtar Kelimeler: Anomali Tespiti, Kestirimci Bakım, Açıklanabilir Yapay Zekâ, Ağır Vasıtalar, Yakıt Verimliliği, Hava Basıncı Sistemi, LSTM Autoencoder, Döngüde İnsan Yaklaşımı, Büyük Dil Modelleri

Ağır vasıtalarda (HDV) anomali tespiti, kestirimci bakım ve etkin filo yönetimi için kritik öneme sahip olsa da, mekanik sistemlerin karmaşık etkileşimleri, farklı operasyonel koşullar ve sınırlı etiketlenmiş veriler nedeniyle önemli zorluklar taşımaktadır. Geleneksel tanılama yöntemleri sıklıkla yanlış alarmlarla karşılaşmakta ve açıklanabilirlikten yoksun kalmaktadır; bu da kullanıcı güvenini azaltmakta ve kritik müdahaleleri geciktirebilmektedir. Bu zorlukları aşmak için, alan bilgisi ve uzman görüşleriyle desteklenen, hassasiyet ve açıklanabilirliği bir araya getiren, veri odaklı ve güçlü anomali tespit sistemlerine ihtiyaç duyulmaktadır. Bu tez kapsamında, ağır vasıtalarda kritik uygulamalar için özel olarak tasarlanmış iki anomali tespit çerçevesi geliştirilmiştir: (1) farklı operasyonel koşullar altında aşırı yakıt tüketimini tespit etmek ve (2) hava basıncı sistemi (APS) arızalarını erken aşamada belirlemek. Aşırı yakıt tüketimi, operasyonel verimliliği ve düzenleyici uyumluluğu önemli ölçüde etkilerken; APS arızaları, yüksek maliyetli arızalara ve iş duruşlarına yol açmaktadır. Bu uygulamaların her biri, verilerin doğasındaki karmaşıklık ve değişkenlikten dolayı özgün metodolojik yaklaşımlar gerektirmektedir.

Yakıt tüketimi anomalilerinin tespiti için, ağırlık-normalize edilmiş yakıt tüketimini ve çok seviyeli yol eğimi segmentasyonunu dikkate alan, çeyrek temelli (quartile-based) yenilikçi bir etiketleme yöntemi geliştirilmiştir. Torbalanmış karar ağaçları

(bagged decision trees) kullanılarak gerçekleştirilen bu denetimli yöntem, Türkiye ve Almanya'daki çeşitli sürüş veri setlerinde %92,2'ye varan doğruluk oranı ve 0,78 F1 skoru ile operasyonel anomalileri yüksek hassasiyetle sınıflandırmaktadır. Ayrıca geliştirilen etkileşimli filo izleme paneli, operatörlere anormal seyahatleri görsel olarak belirleme ve hedefli müdahaleler yapma konusunda eyleme dönüştürülebilir içgörüler sunmaktadır. APS arıza tespiti için ise uzman analizleriyle güçlendirilmiş döngüde insan yaklaşımına (human-in-the-loop) sahip Uzun Kısa Süreli Bellek (LSTM) Autoencoder ile yarı denetimli öğrenme yaklaşımı incelenmiştir. Bu modeller, mekanik arızalardan önce ortaya çıkan ince zamansal sapmaları, toplamda 0,75 F1 skoru ile etkili şekilde belirlemektedir. Buna ek olarak, Açıklanabilir Güçlendirme Makinesi (EBM) modeli %91,4 doğruluk ve 0,80 F1 skoru ile öngörülebilirlik ve yorumlanabilirlik arasında mükemmel bir denge kurmuş; Büyük Dil Modeli (LLM) tabanlı bir aracı sistemle desteklenerek uzman düzeyinde tanısal mantık yürütme ve şeffaflık sağlanmıştır.

Bu tezde, açıklanabilir yapay zekâ teknikleri ile insan uzmanlığının bütünleştirilmesine önem verilerek tanısal güvenilirlik ve kullanıcı güveni artırılmıştır. Oluşturulan yorumlanabilir çerçeveler, net kök-neden analizini mümkün kılmakta, yanlış alarmları azaltmakta ve çeşitli operasyonlarda pratik karar almayı iyileştirmektedir. Geliştirilen yöntemler sürdürülebilir filo yönetimi için çok yönlü ve uyarlanabilir çözümler sunarken, ileride gerçek zamanlı anomali tespiti, çoklu hata sınıflandırması ve otomatik kapalı döngü kestirimci bakım sistemlerine entegrasyon gibi genişletme potansiyeline sahiptir.

ACKNOWLEDGEMENTS

First and foremost, I wish to express my deepest gratitude to my advisor, Prof. Dr. Mustafa Ünel, for his exceptional guidance, invaluable insights, and unwavering support throughout my PhD journey. His constructive feedback, visionary outlook, and continuous encouragement have significantly shaped my academic growth and research perspectives. It has truly been an honor and privilege to conduct my research under his mentorship.

I sincerely thank the members of my dissertation committee for generously dedicating their time and offering insightful suggestions and constructive critiques, which have significantly enriched the quality and depth of this dissertation.

Special thanks go to Ford Otosan for their invaluable support and provision of essential data necessary for my research. I particularly acknowledge Kerem Köprübaşı for leading the Ford Otosan team and for his continuous guidance and collaboration throughout the project, and extend my sincere appreciation to Metin Yılmaz for his valuable contributions and collaborative spirit.

I am deeply grateful to my colleague, brother, and dear friend, Shawqi Mohammed Farea, with whom I closely collaborated on the "System Health/Anomaly Prediction Based on Connected Vehicle Data" project supported by Ford Otosan.

My heartfelt appreciation also extends to my colleagues at the Control, Vision, and Robotics (CVR) group for their insightful discussions and collaborative efforts, especially Naida Fetic and Muhammed Zemzemoğlu. Special acknowledgment also goes to Harun Tolasa and Çağatay Irmak from our laboratory for their support, camaraderie and friendship. I would also like to express my sincere gratitude to my close friends Cihan Erdoğanyılmaz, Ali Yasir Naç, Fatih Emre Tosun, and Caner Dikyol, whose companionship and encouragement greatly enriched my doctoral experience.

Above all, I am profoundly indebted to my family for their patience, love, and encouragement throughout this challenging yet rewarding journey. My deepest gratitude goes to my parents, İsmail Hakkı and especially my mother Seçil Mumcuoğlu, whose boundless support has always been a source of inspiration. I am also deeply thankful to my brothers, Ali Haydar and his wife Fatıma, Hasan Hüseyin, my nephew Haydar, my beloved sister Fatıma Zehra, and my dear wife Hayrunnisa, whose love and encouragement have been indispensable.

Dedicated to my beloved family...

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATONS

LSTM Long Short-Term Memory 6

LSTM-AE LSTM-based Autoencoders 45

ML Machine Learning 7

OOB Out-of-Bag 22

PCA Principal Component Analysis 17

PDP Partial Dependence Plots 18

RNNs Recurrent Neural Networks 45

RUL Remaining Useful Life 14

SVDD Support Vector Data Description 15

SVMs Support Vector Machines 15

TN True Negative 82

TP True Positive 82

TPD Token-Per-Digit 26

TPR True Positive Rate 83

XAI Explainable AI 9

1. INTRODUCTION

HDVs have evolved into rolling sensor networks, continuously streams of telemetry data to the cloud. From powertrain signals and brake system indicators to driver behavior and environmental context, this vast data stream promises significant advances in predictive maintenance—identifying potential faults before they strand a vehicle or escalate into costly failures. However, as driving patterns diversify, road conditions fluctuate, and factors such as terrain, load, and driver routines vary significantly, distinguishing a minor outlier from a safety-critical anomaly becomes exceptionally challenging. The consequences are significant: downtime directly results in loss of transport capacity, higher repair costs, and reduced customer satisfaction.

Conventional rule-based diagnostics struggle in this context for two main reasons. First, the complexity of different vehicles and operating conditions makes manually defined thresholds unreliable. Second, when an alarm is triggered, technicians require clear explanations of the root cause; otherwise, repeated false alarms erode trust, causing unnecessary part replacements. Modern deep-learning models address the first issue by learning typical system behaviors directly from historical data. However, addressing the second challenge necessitates human expertise and model explainability to ensure the reliability and interpretability of diagnostic systems.

Figure 1.1 summarizes the critical three-way interdependency essential to robust anomaly detection: validated data streams ensure trustworthiness, human expertise provides crucial insights and accurate labeling, and intelligent models deliver precision along with explainability. Each of these elements reinforces the core anomaly detection engine, enhancing overall reliability and interpretability. Deep neural networks offer substantial predictive power by capturing complex, non-linear patterns from vehicle data. Yet, engineering expertise remains indispensable. Designing meaningful features from vehicle signals, validating that model outputs align with physical reality, and identifying the root causes behind anomalies remain persistent challenges. These difficulties arise because failures are infrequent, accurately labeled data is scarce, and operating conditions continually evolve, forcing practitioners to adopt unsupervised or semi-supervised learning methods. In such scenarios, adopt-

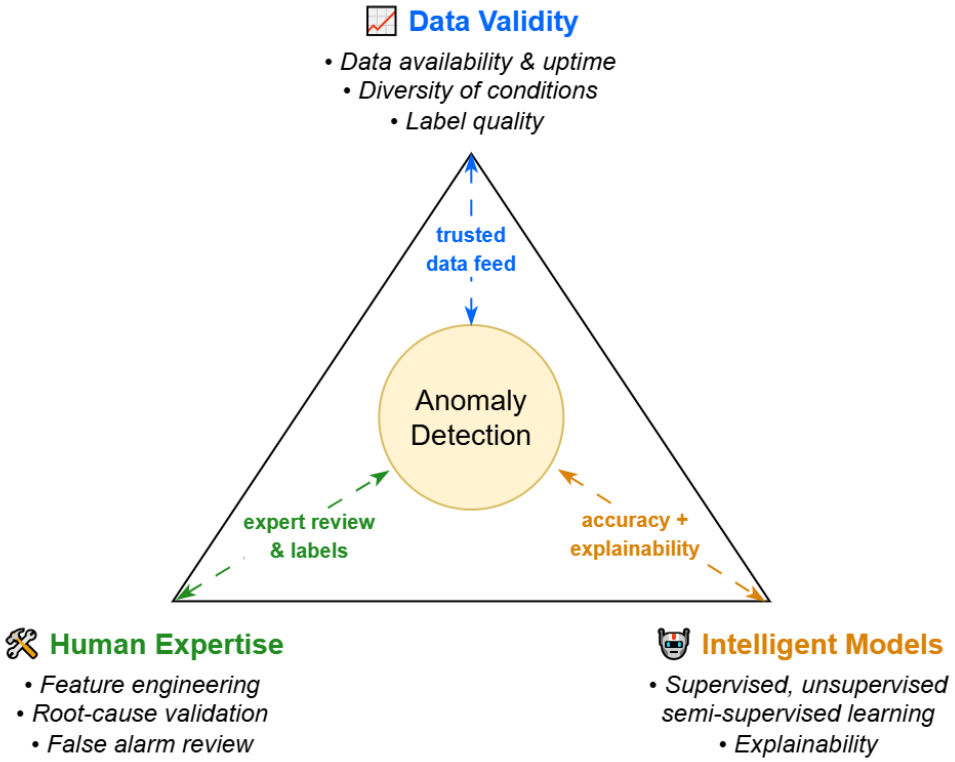

Figure 1.1 Predictive-maintenance trilemma for HDVs.

ing a human-in-the-loop approach is essential. Domain experts iteratively refine feature selection, assess model explanations, annotate edge cases, and provide feedback, enhancing model training and enabling robust and effective predictive maintenance.

Looking ahead, the rapid advancement of LLMs promises another significant leap in root-cause analysis. These models have the capacity to capture and utilize expert knowledge, embedding it within agentic systems that clearly articulate their reasoning processes. Integrated with traditional deep-learning pipelines analyzing sensor data, LLM-powered agents can interpret anomalies in plain language, suggest actionable diagnostics, and continuously refine their analyses through engineer feedback. This integration creates a collaborative, self-explaining diagnostic framework, combining the precision of data-driven methods with transparent, expert-level reasoning.

1.1 Motivation

Anomaly detection in heavy-duty vehicles poses a wide range of challenges across different systems and failure types. In this thesis, we focus on two critical and data-rich problems: excessive FC and APS failures. To tackle these tasks, we develop tailored approaches that explore effective use of ML, feature engineering, and human-in-the-loop systems—laying the groundwork for scalable and interpretable predictive maintenance solutions.

1.1.1 Detecting Excessive Fuel Consumption Levels

Anomaly detection is crucial for monitoring HDV performance, as excessive fuel consumption can arise from subtle mechanical faults or inefficient driving behaviors. Customers and fleets often complain about their vehicles' fuel consumption, yet establishing a reference for fair comparison and detecting excessive or anomalous usage is far from trivial. Given the increasing availability of on-road data, data-driven methods have become the most practical approach for early detection and timely intervention.

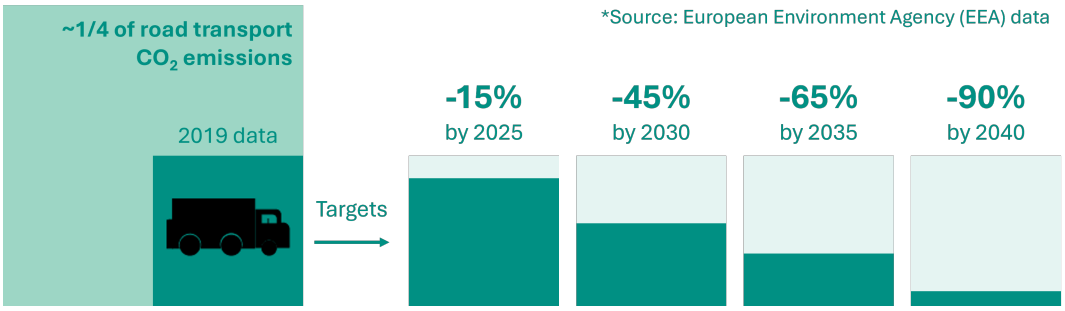

Figure 1.2 EU Fleet-Wide CO_2 Reduction Targets for HDVs.

Diesel HDVs continue to dominate long-haul freight transport due to their high energy density and durability, yet they account for approximately one-quarter of the European Union's road-transport CO_2 emissions (European Environment Agency, 2022). These emissions have increased consistently since 2014, with the exception of a temporary dip during the 2020 pandemic (European Parliament, 2018). Consequently, the EU has mandated significant fleet-wide CO_2 reductions: 15% by 2025, 45% by 2030, 65% by 2035, and 90% by 2040, relative to a 2019 baseline (European Commission, 2024) (Figure 1.2). Achieving these ambitious targets demands more

than periodic compliance testing; fleet operators require continuous monitoring to ensure real-world fuel consumption does not quietly increase over time.

Excessive fuel consumption typically results from two main sources: behavioral inefficiencies (such as aggressive acceleration, prolonged idling, and suboptimal use of cruise control) that waste fuel even when vehicles are mechanically sound, and technical inefficiencies (such as injector drift, turbocharger wear, and under-inflated tyres) that elevate fuel consumption under normal driving conditions. Identifying whether excessive consumption is driven by driver behavior or mechanical faults is essential for profitability and regulatory compliance.

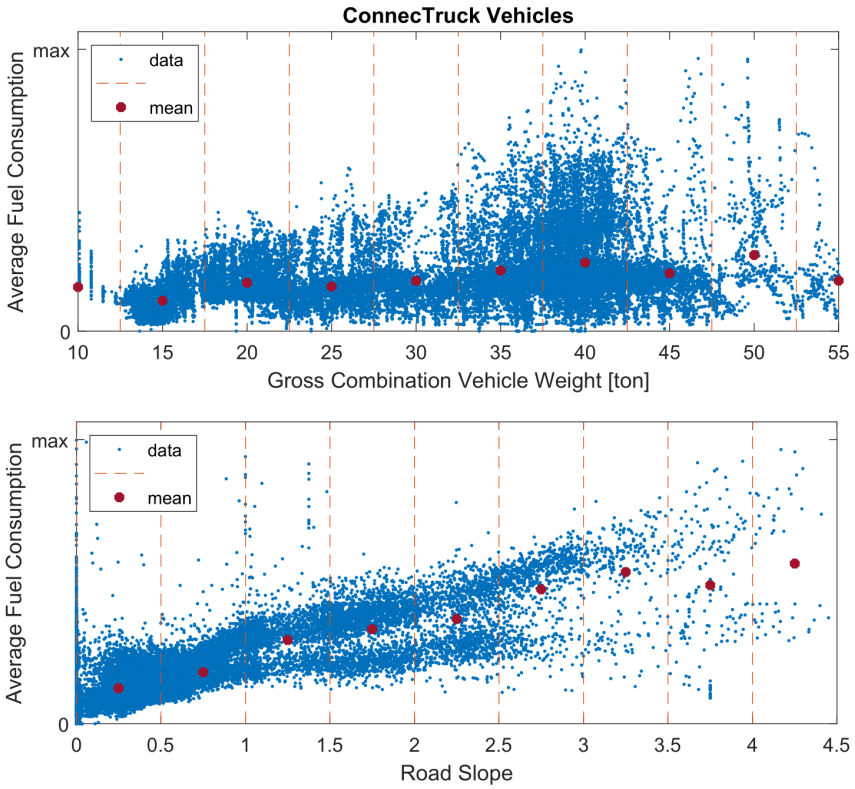

Figure 1.3 Relationship between average fuel consumption, gross combination weight, and road slope (Mumcuoglu et al., 2023).

However, the complexity and variability of operating conditions complicate anomaly detection. Factors such as gross vehicle weight, road gradient, vehicle speed, tyre pressure, ambient temperature, and seasonal variations interact in complex, nonlinear ways. Figure 1.3 illustrates how the variability in average fuel consumption increases with load and slope, making predictions increasingly uncertain. As a result, any effective detection approach must be context-aware.

Due to the scarcity of reliable ground-truth labels indicating what constitutes "excessive" fuel consumption, this challenge is framed as an unsupervised anomaly-detection problem. The goal is to establish a robust baseline model that inherently

accounts for physical drivers such as load and slope, and subsequently flags trips whose fuel consumption significantly deviates from expected levels. By leveraging physics-informed features (e.g., load-normalised fuel rate) and flexible, learning-based models capable of capturing both point-in-time and temporal consumption patterns, this approach becomes feasible. The resulting framework enables a fair, data-driven separation of behavioral inefficiencies from emerging mechanical issues, thereby guiding timely and targeted interventions.

1.1.2 Detecting Air Pressure System Failures

HDVs operate under demanding conditions, frequently leading to mechanical failures driven by inadequate maintenance planning, the inability to timely identify anomalies, and suboptimal driving habits. Among the various mechanical systems vulnerable to such failures, the APS stands out due to its critical role. Central to this system is the E-APU, which is essential for maintaining proper pressure in air brakes and suspension systems. Failures within the APS are notably significant, as they frequently cause HDVs to become stranded roadside, resulting in costly emergency interventions and diminished customer satisfaction.

Issues within the APS, whether mechanical faults or sensor malfunctions, can result in overloading and mechanical fatigue, ultimately causing premature E-APU failures. Detecting these faults at an early stage is essential to prevent vehicles from becoming immobilized during operation, thus avoiding expensive roadside assistance and minimizing downtime. Incidents that lead to such breakdowns impose considerable financial burdens on vehicle manufacturers and disrupt operations for fleet operators, negatively impacting overall productivity and profitability.

However, accurately detecting APS failures in advance is a complex challenge that typically requires substantial domain-specific expertise. Traditionally, APS issues are identified manually during routine maintenance inspections or through reactive procedures triggered by customer complaints, often due to noticeable air leaks. Figure 1.4 illustrates various examples of failed E-APU units, clearly demonstrating the severe operational stresses HDVs endure and highlighting the complexity associated with monitoring and maintaining these components.

Given these complexities and the significant risks associated with APS failures, there is an urgent need for intelligent vehicle systems equipped with advanced predictive maintenance capabilities. Despite this necessity, research specifically addressing

Figure 1.4 Service images of failed E-APU units (Mumcuoglu et al., 2024b), shown both installed and removed examples, illustrating progressive wear and operational damage.

predictive detection of APS failures—such as air compressor malfunctions using real-time operational data—is currently limited. Modern ML techniques, particularly semi-supervised approaches, offer considerable promise given the inherently uncertain nature of APS failures and the limited availability of labeled failure data.

Advanced ML architectures, such as LSTM Autoencoders, have already demonstrated notable success in anomaly detection applications across various industries, particularly in identifying failure conditions (Khalid Fahmi et al., 2024). Such sophisticated models, complemented by expert domain knowledge and careful data interpretation, have significant potential to address the specific challenges posed by APS failures. Integrating domain expertise with state-of-the-art analytical approaches can lay a robust foundation for predictive maintenance solutions, significantly enhancing the reliability, safety, and operational intelligence of HDVs.

1.2 Thesis Contributions

This thesis makes contributions to the field of anomaly detection and root-cause determination for automotive applications, specifically focusing on HDVs. Building upon the challenges highlighted in the motivation section, the research addresses two critical issues: detecting excessive FC levels and early detection of APS failures in HDVs.

Fuel Consumption Classification with Load and Slope-Aware Quartile Labeling

Fuel consumption is a vital performance indicator for HDVs, heavily influenced by various factors such as vehicle weight, road slope, and driving behavior. Accurately detecting anomalies in fuel consumption can help transportation companies and manufacturers identify system faults or driving behaviors that lead to excessive energy consumption and emissions. The detailed contributions in this area are as follows:

- *Dataset Generation*: Two comprehensive datasets were generated and utilized: Dataset A with 606 naturalistic driving records collected from 57 heavy-duty trucks (Turkish route), and Dataset B with 520 trips from 187 trucks (German route), providing diverse geographical and operational contexts for model validation.
- *Road Slope-Aware Labeling Methodology*: A novel quartile-based labeling approach was developed incorporating weight-normalized FC quartiles with multi-level slope segmentation, enabling accurate distinction between normal operational variations and genuine FC anomalies.
- *Data Segmentation and Labeling*: Each driving record was divided into 10-minute sections. Each section was labeled based on FC quartiles normalized by the combined weight of the truck with its carry load. Separate quartiles were computed based on different slope levels of the driven road.
- *Development of Classification Models*: High FC and outlier FC classification models were developed using the Bagged Decision Trees algorithm. These models classify 10-minute sections of HDV driving, using vehicle signals, in terms of fuel consumption considering vehicle weight and road slope.
- *Feature Importance Analysis*: In the design process of the ML models, a feature importance analysis was performed to select the most significant predictors,

enhancing the models' accuracy and efficiency.

- *Interactive Dashboard Development*: An interactive MATLAB dashboard was developed for fleet-level FC monitoring, providing three visualization levels: fleet overview with vehicle ranking, distribution analysis of consumption patterns, and detailed trip-level inspection with GPS mapping for actionable fleet management insights.

- *Cross-Dataset Validation*: The driving data from both Turkish and German datasets were evaluated based on the results of the FC classification system, demonstrating the system's effectiveness in detecting anomalies across different geographical and operational contexts.

This system can assist transportation companies and manufacturers in determining driving behavior anomalies or system faults that cause excessive energy consumption and emissions for HDVs.

Air Pressure System Failure Detection Using LSTM Autoencoders, Human Expert Analysis, and Explainable AI

Mechanical failures in HDVs, particularly in the APS, can lead to significant operational disruptions and costs. Early detection of APS failures, such as those involving the E-APU, is crucial for preventive maintenance and avoiding breakdown scenarios. Given the unpredictable nature of APS failures and the limited availability of labeled automotive data, semi-supervised ML techniques are well-suited to tackle this challenge. The main contributions in this area include:

- *Data Acquisition*: Acquired a dataset comprising 30 days of operational time-series data from two HDV groups: 30 vehicles that underwent E-APU replacements due to failures, and 110 vehicles with no maintenance issues.

- *Feature Engineering and Preprocessing*: Proposed several preprocessing methods to manage the large dataset and extracted engineered features that highlight specific temporal patterns indicative of APS failures. These features, created with the aid of domain knowledge, facilitated both HEA and the development of ML models.

- *Development of LSTM Autoencoder Model*: Developed an LSTM Autoencoder model to address APS failure detection as a semi-supervised anomaly detection problem. The model learns normal operational patterns of healthy vehicles and identifies deviations that may indicate failures.

- *Integration of Human Expert Analysis*: Proposed a human-in-the-loop ML

framework with enhanced HEA+ methodology, which incorporates brake usage patterns to reduce false positives and improve failure detection accuracy.

- *XAI Integration*: Developed complementary XAI modules including EBM for interpretable supervised classification and an LLM-based agentic framework using specialized AI agents for diagnostic reasoning and interpretable analysis.
- *Hybrid Detection Approaches*: Demonstrated effective combination of LSTM autoencoders with human expert analysis, achieving significant false positive reduction and improved maintenance decision-making through multi-tiered detection strategies.

By focusing on detecting APS failures, this work demonstrates the effectiveness of the proposed framework in addressing this commercial automotive issue using cloud-based operational driving data. The proposed method significantly improves failure detection rates and combines data analytics with domain expertise to enhance the performance of the ML models.

1.3 Outline

The remainder of the thesis is structured as follows:

Chapter 2 reviews anomaly detection techniques in automotive applications, including supervised, semi-supervised, and unsupervised methods, and XAI techniques relevant to automotive contexts. It further describes the primary machine-learning methods employed—Ensemble of Bagged Decision Trees, LSTM networks, Autoencoders, EBM, and LLMs for time-series data.

Chapter 3 details the HDV subsystems examined in this thesis and presents a comprehensive data acquisition architecture utilized in Ford F-MAX trucks. It introduces the datasets used: two FC datasets (from Turkish and German routes) and an APS failure dataset with run-to-failure data from 140 vehicles.

Chapter 4 introduces a novel load- and slope-aware FC classification framework. It proposes a weight-normalized quartile labeling approach using multi-level slope segmentation, implements Bagged Decision Tree classifiers for identifying high and outlier FC, describes the progressive labeling strategy across four refinement levels, and presents an interactive fleet-level FC monitoring dashboard.

Chapter 5 describes a comprehensive APS failure detection methodology, including sliding window-based feature extraction, baseline methods combining LSTM Autoencoders with HEA, and advanced XAI techniques employing EBMs and LLM-based agentic analysis with specialized diagnostic agents.

Chapter 6 provides quantitative and qualitative evaluations of the proposed FC anomaly detection methods (demonstrating effectiveness across Turkish and German datasets with load and slope considerations) and the APS failure detection framework. It evaluates individual ML models, HEA, XAI methods, and hybrid combinations.

Chapter 7 summarizes the thesis's main contributions, discusses practical implications for fleet management and predictive maintenance, addresses the limitations of the proposed methodologies, and outlines potential directions for future research in explainable anomaly detection and predictive maintenance for HDVs.

1.4 Publications

Journal Articles:

- **M.E. Mumcuoglu**, S.M. Farea, M. Unel, S. Mise, S. Unsal, E. Cevik, M. Yilmaz, K. Koprubasi, *"Detecting APS Failures Using LSTM-AE and Anomaly Transformer Enhanced with Human Expert Analysis,"* Engineering Failure Analysis, vol. 165, Pergamon, 2024.
- S.M. Farea, **M.E. Mumcuoglu**, M. Unel, *"An Explainable AI approach for detecting failures in air pressure systems,"* Engineering Failure Analysis, vol. 173, Pergamon, 2025.

Conference Papers:

- **M.E. Mumcuoglu**, S.M. Farea, M. Unel, S. Mise, S. Unsal, M. Yilmaz, K. Koprubasi, *"Fuel Consumption Classification for Heavy-Duty Vehicles: A Novel Approach to Identifying Driver Behavior and System Anomalies,"* 2023 AEIT International Conference on Electrical and Electronic Technologies for Automotive, IEEE, 2023.
- S.M. Farea, **M.E. Mumcuoglu**, M. Unel, S. Mise, S. Unsal, M. Yilmaz, K. Koprubasi, *"Towards Driving-Independent Prediction of Fuel Consumption in Heavy-Duty Trucks,"* 2023 AEIT International Conference on Electrical and Electronic Technologies for Automotive, IEEE, 2023.
- B.B. Turan, E. Genc, I.N. Akcig, N. Goztepe, **M.E. Mumcuoglu**, M. Unel, *"Detecting High Fuel Consumption in HDVs with Ensemble of Anomaly Detection Models,"* 22nd International Conference on Industrial Informatics, IEEE, 2024.
- **M.E. Mumcuoglu**, S.M. Farea, M. Unel, S. Mise, S. Unsal, E. Cevik, M. Yilmaz, K. Koprubasi, *"Air Pressure System Failures Detection Using LSTM-Autoencoder,"* 2024 IEEE International Workshop on Metrology for Automotive, IEEE, 2024.
- S.M. Farea, **M.E. Mumcuoglu**, M. Unel, S. Mise, S. Unsal, E. Cevik, M. Yilmaz, K. Koprubasi, *"Prediction of Failures in Air Pressure System: A Semi-Supervised Framework Based on Transformers,"* 22nd International Conference on Industrial Informatics, IEEE, 2024.

2. BACKGROUND AND LITERATURE REVIEW

This chapter situates the thesis within the broader context of data-driven vehicle health monitoring, emphasizing the critical role of robust anomaly detection in modern automotive systems. It begins by reviewing the existing literature categorized according to supervision level—supervised, semi-supervised, and unsupervised approaches—and subsequently introduces the three primary techniques employed in this thesis: Bagged Decision Trees, LSTM networks, and Autoencoders.

By the end of this chapter, readers will clearly understand (i) the current state of research in the field, (ii) the specific gaps that this thesis aims to address, and (iii) the rationale behind selecting methods tailored to the data characteristics and operational constraints of real-world automotive applications.

2.1 Anomaly Detection in Automotive Applications

Anomaly detection is crucial for maintaining the safety and reliability of automotive systems. With modern vehicles generating vast amounts of data from numerous sensors and control units, ML techniques have become indispensable for fault detection and diagnosis. Depending on the availability of labeled data, anomaly detection methods are classified into supervised, semi-supervised, and unsupervised learning approaches.

The key difference among these learning approaches lies in the availability of labeled data and the learning objectives. Supervised learning methods depend entirely on labeled data, where each data point is associated with a known output or class label. These methods use labeled examples to learn from data, enabling the classification or prediction of new, unseen instances. Unsupervised learning methods, on the other hand, focus on discovering inherent patterns or structures in completely

unlabeled data. They extract meaningful information by identifying anomalies or clusters without any prior knowledge of the data labels. The recently proposed semi-supervised learning approach combines both labeled and unlabeled data during the learning process. This approach leverages the abundance of unlabeled data, which is often easier and less expensive to collect, along with a smaller set of labeled data to improve learning performance.

In the context of anomaly detection for automotive applications, the choice among these learning paradigms depends on the availability and quality of labeled data. Supervised learning is effective when there is ample labeled data representing both normal and faulty conditions. However, obtaining labeled faulty data can be challenging due to the rarity of certain faults and the costs associated with data labeling. Unsupervised learning is advantageous when labeled data are scarce or unavailable, allowing models to identify anomalies based on deviations from learned patterns in the data. Semi-supervised learning strikes a balance by utilizing the limited labeled data available to guide the learning process, improving the detection of anomalies that may not be well-represented in the labeled dataset (Figure 2.1).

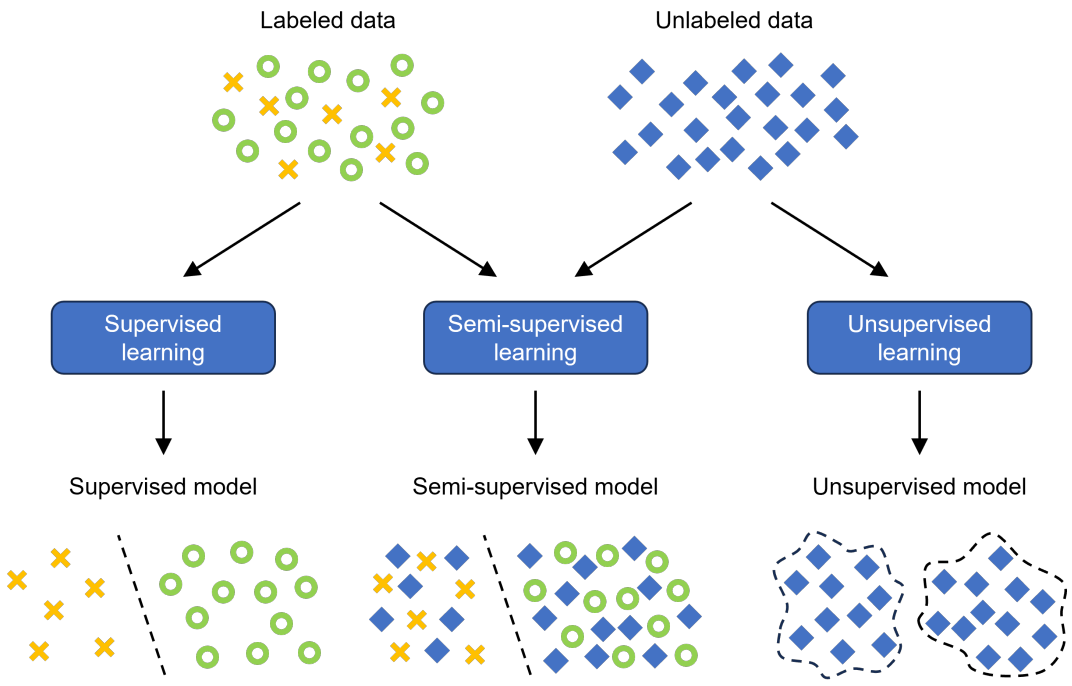

Figure 2.1 Relationship between data labeling and learning paradigms in ML methods.

2.1.1 Supervised Anomaly Detection in Automotive Applications

Supervised ML methods are applied when fault types are known and labeled data is available. These models learn from labeled examples to classify or predict faults in new, unseen data. In automotive applications, supervised learning is effective when sufficient labeled data representing both normal and faulty conditions is accessible.

Wolf et al. (2018) developed a deep learning-based pre-ignition detection framework utilizing signals from Electronic Control Units collected from a fleet of vehicles. They evaluated different deep neural network architectures, combining LSTM networks and CNN, achieving high F1 scores in classifying internal combustion engine faults. Their work demonstrated the effectiveness of deep learning models in capturing complex temporal patterns associated with engine anomalies.

Similarly, Rengasamy et al. (2020) proposed a predictive maintenance methodology using deep neural networks to detect engine faults in heavy-duty trucks. They utilized Scania's open dataset, which contains over 75,000 instances with only 1.8% representing faulty conditions, illustrating a highly imbalanced dataset common in real-world scenarios. To address this imbalance, they designed a weighted loss function to enhance fault detection accuracy. Various deep learning architectures, including ANNs, LSTMs, CNNs, and GRUs, were evaluated, showing that deep learning models can effectively handle imbalanced data in fault detection tasks.

Nowaczyk et al. (2013); Prytz et al. (2015) focused on predictive maintenance applications for air compressor systems in heavy-duty trucks using Volvo's internal service records dataset. They developed a fuzzy rule-based method, named the relaxed prediction horizon algorithm, to classify failures (Nowaczyk et al., 2013), and utilized a random forest model for RUL estimation of compressor failures (Prytz et al., 2015). Their work highlighted the applicability of supervised learning methods in forecasting maintenance needs for specific vehicle components.

Additionally, Wong et al. (2016) utilized ELMs for classifying internal combustion engine faults based on vehicle signals, while Zhong et al. (2018) implemented an ensemble of Bayesian ELMs for pre-ignition detection, demonstrating improved classification performance. These studies underscore the importance of supervised ML methods in automotive anomaly detection when labeled data is available.

A summary of supervised ML applications in automotive anomaly detection is provided in Table 2.1.

Table 2.1 Supervised ML applications, use cases, and ML methods

Reference	**Use Case**	**ML Method**
Wong et al. (2016)	IC engine faults	ELM
Zhong et al. (2018)	Pre-ignition detection	Ensemble of Bayesian ELMs
Wolf et al. (2018)	IC engine faults	CNN, LSTM
Rengasamy et al. (2020)	Air pressure system faults	ANN, LSTM, CNN, GRU
Nowaczyk et al. (2013)	Truck compressor failures	Relaxed prediction horizon algorithm
Prytz et al. (2015)	Air compressor system RUL	Random forest

These studies demonstrate that supervised learning methods, including deep neural networks, random forests, and ensemble techniques, are effective in detecting and predicting faults in various automotive systems when labeled data is available.

2.1.2 Semi-supervised Anomaly Detection in Automotive Applications

Semi-supervised methodologies are applied when only a limited amount of labeled data is available, or when the goal is to detect both known and unknown fault types. These models are trained primarily on normal (non-faulty) data, learning the patterns of normal operation, and are then used to identify deviations that may indicate anomalies.

Killeen et al. (2019) developed a semi-supervised approach to detect defective buses that differed from the rest of the fleet using statistical analysis. Jung (2020) proposed an anomaly detection algorithm for various internal combustion engine faults using multiple one-class SVMs. Theissler (2017) presented a full vehicle fault detection methodology utilizing an ensemble of one-class and two-class classifiers.

Key semi-supervised techniques used for fault detection include one-class SVMs, SVDD, and autoencoders. Theissler (2017) implemented one-class SVM and SVDD classifiers to detect simulated vehicle faults such as erroneous sensor readings. Jung (2020) utilized a one-class SVM classifier for fault detection in internal combustion engines, while Sang et al. (2020) applied a one-class SVM to detect braking system faults in electric multiple units.

Autoencoders, which are neural network architectures designed for unsupervised learning of efficient codings, have also been employed for fault detection. Zhang et al. (2023) used SVDD and autoencoder models for fault detection in lithium batteries of electric vehicles. Min et al. (2023) implemented a denoising shrinkage autoencoder to detect faulty sensors in autonomous vehicles. Similarly, Geglio et al. (2022) utilized a convolutional autoencoder to detect powertrain faults in hybrid-electric vehicles.

Table 2.2 Semi-supervised ML applications, use cases, and ML methods

Reference	**Use Case**	**ML Method**
Killeen et al. (2019)	Detection of defective buses differing from fleet	Statistical analysis (semi-supervised approach)
Jung (2020)	Anomaly detection for IC engine faults	Multiple one-class SVMs
Theissler (2017)	Full vehicle fault detection methodology	Ensemble of one-class and two-class classifiers
Jung (2020)	Fault detection in IC engines	One-class SVM classifier
Sang et al. (2020)	Detection of braking system faults in electric multiple units	One-class SVM
Zhang et al. (2023)	Fault detection in lithium batteries of electric vehicles	SVDD and autoencoder models
Min et al. (2023)	Detection of faulty sensors in autonomous vehicles	Denoising shrinkage autoencoder
Geglio et al. (2022)	Detection of powertrain faults in hybrid-electric vehicles	Convolutional autoencoder
Davari et al. (2022)	Anomaly identification in public transport bus subsystems	LSTM autoencoder
Kang et al. (2021)	Monitoring brake operating unit of metro trains	One-class LSTM autoencoder

LSTM autoencoders have been used to capture temporal dependencies in time-series data for anomaly detection. Davari et al. (2022) employed an LSTM autoencoder to identify anomalies in public transport bus subsystems, demonstrating superior precision and recall compared to multilayer autoencoders. Kang et al. (2021) adopted a one-class LSTM autoencoder to monitor the brake operating unit of metro trains. By analyzing brake cylinder pressure data, they achieved early fault detection, showcasing the method's robustness in real-world applications.

2.1.3 Unsupervised Anomaly Detection in Automotive Applications

Unsupervised ML models are utilized when only unlabeled data are available. These methods detect anomalies by identifying patterns or structures in the data that deviate from the norm, without prior knowledge of fault types. Clustering-based methods and anomaly detection algorithms can be combined with expert knowledge to interpret anomalous conditions.

Fan et al. (2015) proposed a method that compares the internal signals of a vehicle with a set of vehicles performing similar operations to detect anomalies. Jung & Sundström (2019) introduced a residual selection algorithm to detect and classify internal combustion engine faults. Tagawa et al. (2014) employed denoising autoencoders for fault detection, which are effective in learning representations that are robust to noise and can highlight anomalies. Routray et al. (2010) developed a full vehicle fault detection system using ICA, PCA, and clustering techniques.

Table 2.3 Unsupervised ML applications, use cases, and ML methods

Reference	**Use Case**	**ML Method**
Fan et al. (2015)	Detection of anomalies in vehicles by comparing internal signals with similar operations	Comparison with peer vehicles (statistical analysis)
Jung & Sundström (2019)	Detection and classification of internal combustion engine faults	Residual selection algorithm
Tagawa et al. (2014)	Fault detection and analysis	Denoising autoencoder
Routray et al. (2010)	Full vehicle fault detection system	ICA, PCA, and clustering techniques
Wang et al. (2023)	Fault prediction in lithium batteries	Transformer model
Zhao et al. (2023)	Battery fault detection in electric vehicles	Transformer-based model
Xu et al. (2021)	Time-series anomaly detection	Anomaly Transformer
Tuli et al. (2022)	Multivariate time-series anomaly detection	Transformer networks (TranAD)
Yu et al. (2023)	Path planning in autonomous vehicles	Transformer-based planning approach
Lin et al. (2023)	Autonomous collision avoidance for unmanned underwater vehicles	Transformer-based dual-channel self-attention
Wang et al. (2023)	Autonomous parking space detection	Global perception-based transformer model

Transformers have recently gained popularity in time-series anomaly detection due

to their ability to capture long-range dependencies and achieve state-of-the-art results (Tuli et al., 2022; Xu et al., 2021). Similar to autoencoders, transformers have been used in both semi-supervised and unsupervised learning schemes as reconstruction-based anomaly detection approaches. Although fault detection is considered a subfield of anomaly detection, the application of transformers to this area is still emerging.

In one of the few works in the automotive sector, Wang et al. (2023) utilized a transformer model for fault prediction in lithium batteries and compared the results with those of an autoencoder model. Their findings indicated that the transformer achieved better performance in terms of the F1 score. Similarly, Zhao et al. (2023) applied a transformer-based model to battery fault detection in electric vehicles. Beyond fault detection, transformers have demonstrated effectiveness in other automotive applications, such as path planning (Yu et al., 2023), collision avoidance (Lin et al., 2023), and autonomous parking (Wang et al., 2023).

2.2 Explainable AI in Automotive Applications

2.2.1 Traditional XAI Approaches

Explainable AI plays an increasingly vital role in automotive systems, especially in autonomous driving and EVs. XAI enhances transparency, trust, and safety by making ML models interpretable and their predictions understandable. This section outlines key XAI approaches applied in automotive contexts, including gradient-based methods, PDP, LIME, SHAP, and EBM.

Gradient-based methods provide explanations by analyzing the sensitivity of model outputs to inputs, making them particularly effective for deep learning models. Charroud et al. (2023) used SmoothGrad and VarGrad to improve the interpretability of deep learning models used for autonomous vehicle localization. Similarly, Grad-CAM has provided valuable post-hoc explanations for neural networks in semantic segmentation tasks for autonomus driving (Kolekar et al., 2022; Saravanarajan et al., 2023).

Partial Dependence Plots visualize the global relationship between model predictions and individual input features. Jafari & Byun (2024) employed PDPs to interpret predictions from models estimating the remaining useful life of lithium-ion batteries in EVs, enhancing model transparency.

LIME provides local explanations for individual predictions by approximating complex models with simpler, interpretable surrogates. Ahmad Khan et al. (2024) utilized LIME in conjunction with SHAP to explain APS failure detection models in Scania Trucks. SHAP, a game-theoretic approach, quantifies the contribution of each feature to the model's predictions. SHAP is widely adopted due to its interpretability across diverse models. Li et al. (2023) employed SHAP for interpreting lane-change detection models. Mohanty & Roy (2023) applied SHAP to study factors influencing energy consumption at EV charging stations, modeled via random forests. Konstantinou et al. (2023) used SHAP with Gradient Boosting Decision Trees to explain FC models. Additionally, SHAP has been effective in identifying risky driving behaviors, such as hard braking and speeding events (Masello et al., 2023; Zhou et al., 2024).

EBM, unlike the previously mentioned methods, inherently produce interpretable models by combining the simplicity of linear models with the accuracy of ensemble methods. Barbado & Corcho (2022) utilized EBM to predict FC anomalies, simultaneously generating clear explanations for predictions. Despite the growing use of XAI, its application to APS failure detection remains limited. Ahmad Khan et al. (2024) applied XAI methods to APS failure detection using anonymized datasets, which poses challenges in verifying explanations due to the lack of clear feature identities. Ensuring explanation validity remains critical in XAI-driven automotive applications.

2.2.2 LLM-Based Approaches

LLMs have recently gained traction in automotive research due to their exceptional zero-shot and few-shot learning capabilities, enhancing transparency, interpretability, and explainability. LLMs naturally produce textual explanations of their reasoning, complementing traditional XAI methods such as SHAP and LIME, and addressing challenges posed by limited labeled data.

Key application areas of LLMs in automotive contexts include traffic management, vehicle planning, perception, and maneuver prediction. In traffic management,

LLMs translate multimodal data streams into interpretable textual forecasts, enabling transparent decision-making in traffic flow prediction and signal control tasks (Guo et al., 2024; Wang et al., 2024). Vehicle planning and control have also benefited from LLMs; GPT-Driver and DriveGPT4 demonstrate that framing trajectory planning as a language-generation problem allows models to provide natural-language rationales alongside their control outputs (Mao et al., 2023; Xu et al., 2023). Furthermore, LLMs enhance perception and scene understanding by generating textual justifications through multimodal question-answering tasks and semantic anomaly detection (Elhafsi et al., 2023; Sima et al., 2023). Retrieval-augmented generation, such as RAG-Driver, further improves explanation accuracy by grounding explanations in external knowledge, thereby reducing misleading outputs (Yuan et al., 2024). LLM-based models, including LC-LLM, also predict maneuver intents transparently, transforming traditional classification tasks into human-readable reasoning processes (Peng et al., 2024).

While LLMs enable rapid adaptation through zero-shot or few-shot learning highlighted by methods such as BEV-CLIP and prompt-based tuning frameworks (Wang et al., 2023; Wei et al., 2024)—their deployment faces significant challenges. The reliability and accuracy of natural-language explanations remain concerns due to potential model hallucinations, latency limitations, and difficulties in certification for safety-critical automotive applications. Integrating traditional XAI techniques with LLM-derived explanations represents a promising direction for ensuring robust and auditable interpretations in automotive use cases.

LLMs for Time-Series Anomaly Detection:

Beyond automotive-specific applications, recent studies have demonstrated the potential of LLMs for zero-shot anomaly detection in general time-series data. (Xu et al., 2025) introduced a multimodal approach by transforming numeric series into visual representations, evaluating multiple multimodal LLMs. Their results demonstrated robust detection of range- and variate-level anomalies, even when data completeness dropped to 75%, and revealed comparable performance between open-source and proprietary models for univariate series. Similarly, Zhou & Yu (2024) explored text-based LLMs, reporting satisfactory anomaly detection accuracy on univariate point- and range-wise anomalies, but noting limitations in more complex multivariate scenarios. Despite these promising advances, LLM-based time-series anomaly detection has yet to be explored within automotive telemetry or safety-critical diagnostics, highlighting a significant and promising research direction.

2.3.1 Ensemble of Bagged Decision Trees

Ensembles of bagged trees are powerful machine-learning techniques known for their robustness against overfitting. In essence, they grow k decision trees using k subsets (i.e., bootstrap samples) of the input data — each one of these sampled subsets is utilized to train one decision tree and, then the output label of each data point is given according to the majority voting approach. As each decision tree in this ensemble is trained using a sample of the whole data, the unseen data – which is also known as out-of-bag data – can be used to validate that decision tree. Due to this cross-validation-like process, the decision trees become more robust against overfitting. In addition, at each node of each decision tree, a random subset of the input features is used; this, in fact, helps in providing an estimation of feature importance as well. Thus, these intrinsic abilities to prevent overfitting and provide feature importance motivate us to adopt the ensemble of bagged trees in this work as the classification algorithm.

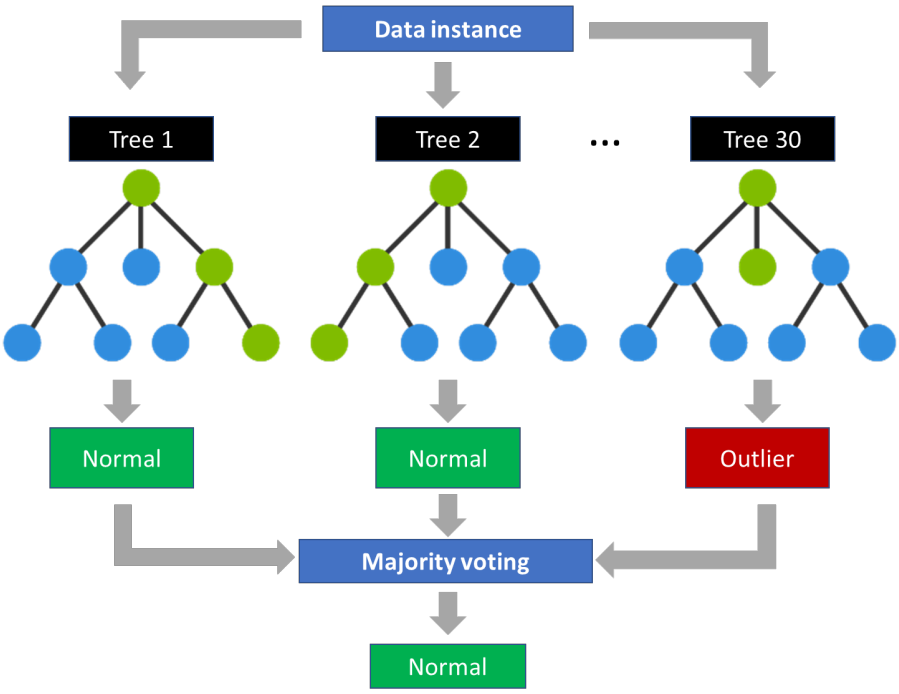

Figure 2.2 The ensemble of bagged trees.

Ensembles of bagged trees are one of the powerful machine-learning techniques for both classification and regression. In essence, they are used to overcome the overfitting problem associated with decision trees. They are based on bootstrap aggregation. That is, they firstly construct multiple bootstrap samples (i.e., subsets) from the data through sampling with replacement where the size of each subset is the same as the size of the whole data. Consequently, each bootstrap sample will contain almost two-thirds of the data, with some duplicates, while one-third of the data, known as out-of-bag data, will not be included in that sample. Then, the ensemble of bagged trees contains as many decision trees as the number of bootstrap samples such that each decision tree is trained using one bootstrap sample and validated using its corresponding out-of-bag data. By doing so, the ensemble of bagged trees is ensured not to overfit.

Out-of-Bag Feature Importance by Permutation

A crucial advantage of ensemble methods is their ability to provide reliable feature importance estimates OOB permutation testing (Breiman, 2001). This method leverages the natural cross-validation structure inherent in bagged ensembles to assess predictor influence without requiring additional validation data. The OOB permutation importance measures how much each predictor variable contributes to the model's predictive accuracy by evaluating the degradation in performance when that predictor's values are randomly shuffled.

The estimation process follows a systematic approach for each tree t in the ensemble ($t = 1, \ldots, T$) and each predictor variable x_j ($j = 1, \ldots, p$). For tree t, the out-of-bag error ε_t is first computed using the observations not included in the bootstrap sample used to train that tree. Subsequently, for each predictor variable x_j that was used for splitting in tree t, the values of x_j in the out-of-bag observations are randomly permuted, and the resulting error ε_{tj} is calculated. The difference $d_{tj} = \varepsilon_{tj} - \varepsilon_t$ quantifies the importance of predictor x_j for tree t, with larger positive values indicating greater importance (Breiman et al., 1984).

The final importance score for each predictor x_j is computed as the standardized mean difference across all trees:

$$\text{Importance}(x_j) = \frac{\bar{d}_j}{\sigma_j} \tag{2.1}$$

where $\bar{d}_j$ is the mean of differences d_{tj} across all trees, and σ_j is the standard deviation of these differences (Loh, 2002). This standardization ensures that importance

scores are comparable across predictors with different scales and variability. Predictors with higher importance scores have a greater influence on the model's predictions, making this metric valuable for feature selection and model interpretation in complex datasets.

2.3.2 Long-Short Term Memory Networks

A standard LSTM cell includes the forget gate f_t which determines the information to be discarded, a tanh gate that produces the candidate memory state $\tilde{c}t$, an update gate u_t that selects values to refresh the memory state c_t, and an output gate o_t that generates the cell's output from the input and stored memory. Inputs to these gates include the current sample x_t and the previous cell's output state a_{t-1}. The aforementioned network is illustrated in Figure 2.3 and can be implemented using following equations:

$$f_t = \sigma(W_f[a_{t-1}, x_t] + b_f) \tag{2.2}$$

$$u_t = \sigma(W_u[a_{t-1}, x_t] + b_u) \tag{2.3}$$

$$\tilde{c}_t = \tanh(W_c[a_{t-1}, x_t] + b_c) \tag{2.4}$$

$$c_t = f_t * c_{t-1} + u_t * \tilde{c}_t \tag{2.5}$$

$$o_t = \sigma(W_o[a_{t-1}, x_t] + b_o) \tag{2.6}$$

(2.7) $$a_t = o_t * \tanh(c_t)$$

where W_f, W_u, W_c, W_o are weight matrices, and b_f, b_u, b_c, b_o are the bias vectors of corresponding operations.

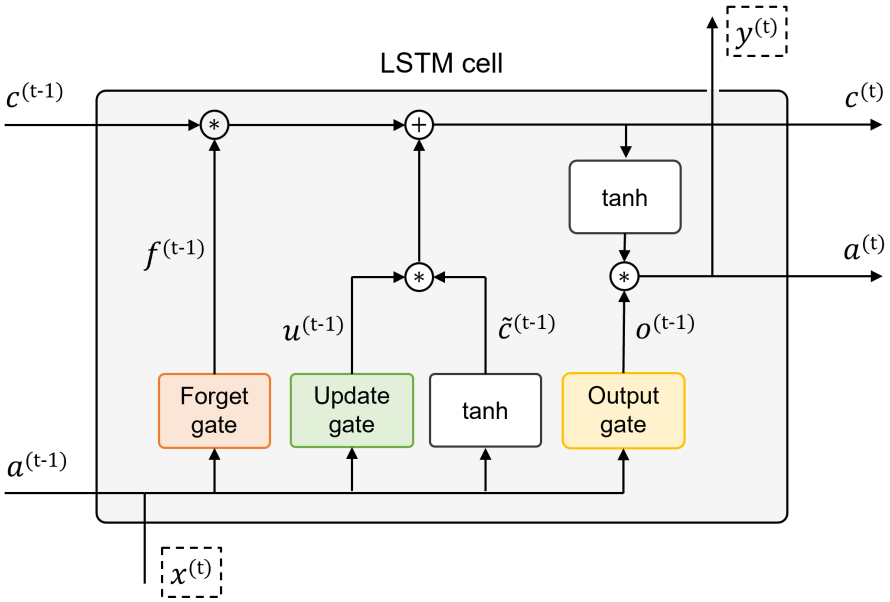

Figure 2.3 Architecture of an LSTM cell.

2.3.3 Autoencoders

Autoencoders are designed to learn representations of given data, similar to the linear PCA technique. However, as a type of artificial neural networks, autoencoders have the additional capability to capture non-linear characteristics in the data (Kramer, 1991). In an autoencoder, the encoder layer reduces the input data to a latent space representation, and the decoder layer then uses that representation to reconstruct the output (Figure 2.4). The difference between the input and reconstructed output is measured and used to update the network's weights through error backpropagation. The model is trained to minimize the reconstruction loss, which is defined as in Eq. (2.8).

(2.8) $$Loss = \frac{1}{N}\sum_{t=1}^{N} \|x_t - \hat{x}_t\|^2$$

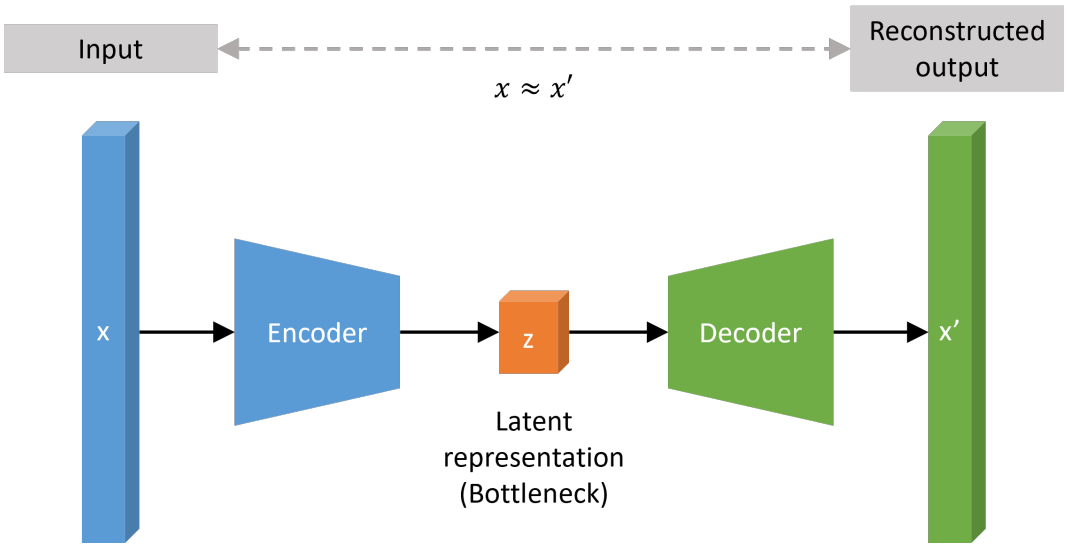

Figure 2.4 Schematic view of autoencoder architecture.

2.3.4 Explainable Boosting Machine (EBM)

EBM is a glass-box interpretable model based on GAM. Unlike traditional linear models, GAM uses flexible functions of individual features rather than strictly linear relationships. EBM further enhances GAM by incorporating pairwise feature interactions and leveraging gradient boosting and bagging techniques during training.

In EBM, each feature is modeled individually using shallow decision trees trained iteratively on residuals of previous models, thus forming feature-specific additive functions. Pairwise interactions among features are also automatically identified and learned similarly. The overall model structure is illustrated in Figure 2.5. Formally, the prediction of an EBM is computed as:

$$g(E[y]) = a_0 + \sum_i f_i(x_i) + \sum_{i<j} f_{ij}(x_i, x_j), \tag{2.9}$$

where a_0 is the intercept, $f_i(x_i)$ represents the contribution of the i^{th} feature, and $f_{ij}(x_i, x_j)$ captures the interaction effects between feature pairs. The link function $g(.)$ adapts the model for tasks such as regression or classification; for instance, $g^{-1}(.)$ is the sigmoid or softmax function for classification tasks and identity for regression.

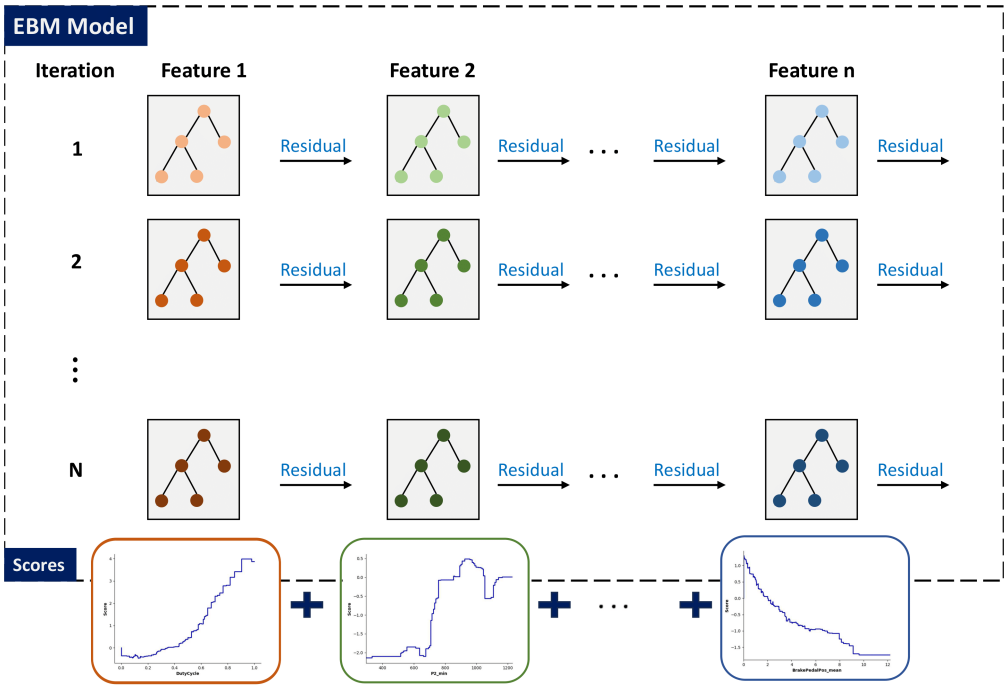

Figure 2.5 Illustration of EBM architecture (Farea et al., 2025).

EBM enables straightforward global and local interpretations by visualizing individual feature functions and interaction effects. Compared to post-hoc explainability methods like SHAP and LIME, EBM provides inherent interpretability, lower prediction latency, and competitive performance relative to complex black-box models, making it suitable for automotive fault detection tasks where transparency is crucial (Das et al., 2020; Nori et al., 2019).

2.3.5 Large Language Models for Time-Series Data

LLMs are Transformer-based architectures processing input sequences of tokens $\{u_1, \ldots, u_L\}$ via multi-head self-attention. Each token is mapped to an embedding vector $\mathbf{e}_j \in \mathbb{R}^d$ and combined with positional encodings $\mathbf{p}_j$ to maintain sequence ordering. The attention weights, defined as $\alpha_{ij} = \text{softmax}(\mathbf{q}_i^\top \mathbf{k}_j / \sqrt{d_h})$, allow each token to attend globally across the sequence, facilitating long-range contextual modeling.

Since Transformers require discrete tokens, numeric time-series data must be converted into textual representations. Common encoding strategies include: (i) direct scalar-to-string conversion (e.g., “12.34”), (ii) CSV-style tokenization, and (iii) TPD

encoding. While TPD can improve performance in certain scenarios, it may degrade accuracy in others due to increased token lengths and deviation from pretraining distributions (Zhou & Yu, 2024).

Additionally, attention mechanisms scale quadratically with sequence length L, imposing practical context-length constraints ($L_{\max}$). Empirical evidence shows that downsampling long sequences can significantly improve anomaly detection performance, underscoring a pronounced context-length sensitivity in current LLMs (Zhou & Yu, 2024). Moreover, chain-of-thought prompting, which encourages explicit step-by-step reasoning, has been shown to negatively impact numeric anomaly detection tasks, suggesting that LLMs rely more heavily on pattern matching than on logical inference in such contexts (Zhou & Yu, 2024).

3. SYSTEM DESCRIPTION AND DATA ACQUISITION

This chapter presents the acquired datasets and system descriptions used in the study. It covers datasets collected from Ford F-MAX trucks under varying drivers, routes, and operating conditions for two key problems: one related to fuel consumption and the other to air-pressure-system behavior, the latter including run-to-failure cases for vehicles with E-APU replacements. The chapter also includes technical details of the systems, providing the necessary background for the analyses in the following chapters.

The data acquisition process relies on a fleet telematics system architecture, as illustrated in Figure 3.1. This system consists of three main components: on-vehicle data collection through telematics control units that interface with the vehicle's CAN bus network, wireless transmission via cellular networks, and cloud-based data processing and analytics platforms. The telematics units continuously sample vehicle parameters and transmit this information in real-time to centralized servers for storage and analysis, enabling comprehensive monitoring of fleet operations and vehicle health status.

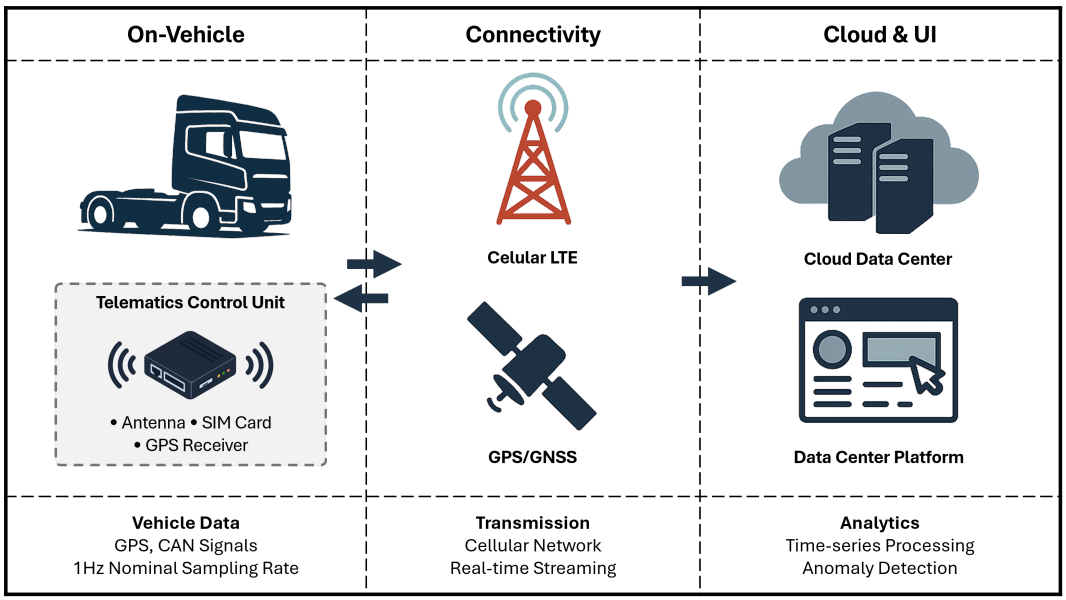

Figure 3.1 Fleet Telematics System Architecture.

Dataset A — Arifiye-İnönü route (Türkiye)

The dataset comprises 606 cloud-logged trips from 57 heavy-duty trucks that shuttle between Arifiye and İnönü. As summarised in Table 3.1, the records span 45 min – 3 h 43 min and 28.4 km – 157.8 km, were sampled at 5 Hz, and total 4 ,123 ,367 rows across 34 synchronised CAN signals. Vehicle mass varies from 8.0 tons when empty to 57.2 tons when fully loaded. Figure 3.2(a) shows the Arifiye–İnönü route from which the first dataset was collected.

Table 3.1 Dataset A summary statistics (Arifiye-İnönü).

Number of vehicles	57
Number of driving records	606
Number of data points	4,123,367
Nominal sampling rate	5 Hz
Number of signals	34
Duration (min - max)	45 min - 3 hr 43 min
Traveled distance (min - max)	28.4 km - 157.8 km
Vehicle weight (min - max)	8.0 ton - 57.2 ton

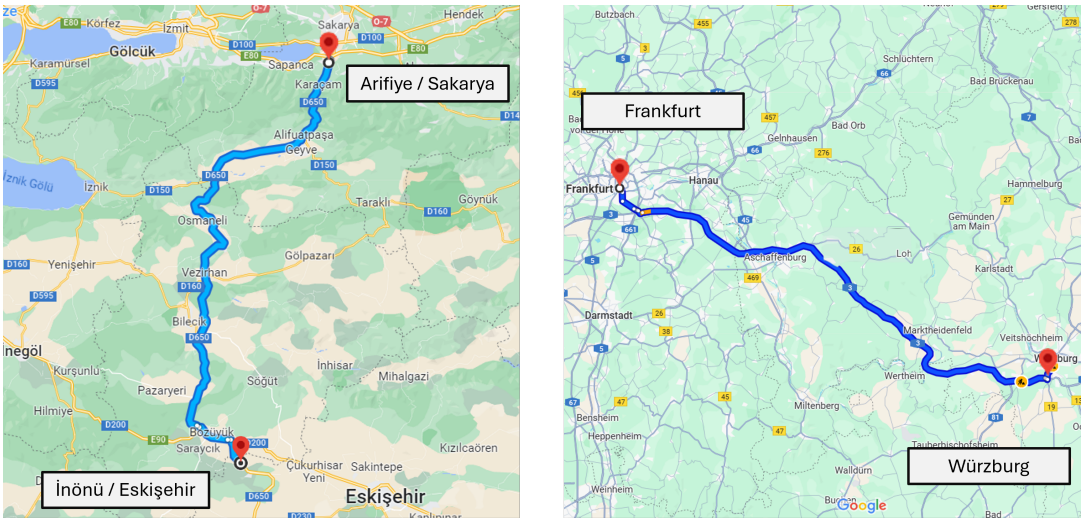

Figure 3.2 Dataset routes: (a) Arifiye-İnönü (Türkiye) and (b) Frankfurt-Würzburg (Germany).

Dataset B — Frankfurt-Würzburg route (Germany)

Figure 3.2(b) shows the Frankfurt–Würzburg route. It contains 520 trip logs from 187 heavy-duty trucks and 1.55 million rows sampled at 0.2 Hz via the fleet-telematics uplink. Trip lengths range from 50 km to 111 km, and gross-combination weight spans 9.5 to 47.8 tons (Table 3.2). Compared with Dataset A, this route is flatter, faster, and recorded at a lower sampling rate, making it an ideal test bed for assessing the robustness of the load- and slope-aware models developed in Chapter 4.

Table 3.2 Dataset B summary statistics (Frankfurt-Würzburg).

Number of vehicles	187
Number of driving records	520
Number of data points	1,546,292
Nominal sampling rate	0.2 Hz
Number of signals	34
Travel distance (min – max)	50 km – 111 km
Vehicle weight (min – max)	9.5 t – 47.8 t

For both datasets, the signals listed in Table 3.1 include time stamps, vehicle ID, cumulative distance, total fuel used (litres), vehicle speed (km/h), engine speed (rpm), engine-torque percentage, and accelerator/brake-pedal positions (%). Road slope, gross combination weight (tons), and engine-oil/coolant temperatures (^oC) describe operating context, and an AdBlue dosing rate (continuous, scales with torque demand) measures urea injection. Together, these kinematic and operating-state variables form the basis for the load- and slope-aware fuel-consumption models developed in Chapter 4.

Table 3.3 List of primary signals for the fuel-consumption dataset.

#	Signal Name	Description	Unit
1	DateTime	Date and time stamp	ISO 8601
2	VehicleID	Vehicle identifier	-
3	HRTVD	High res. tot. dist. traveled	meters
4	TachographVehicleSpeed	Vehicle speed	km/h
5	EngSpeed	Engine speed	rpm
6	ActualEngPercentTorque	Engine torque	%
7	AccelPedalPos1	Accelerator pedal position	%
8	BrakePedalPos	Brake pedal position	%
9	PCCM_Slope	Road slope	-
10	DStgy_dmRdcAgAct	AdBlue dosing rate	-
11	EngOilTemp1	Engine oil temperature	°C
12	EngCoolantTemp	Engine coolant temperature	°C
13	GCVW	Gross comb. vehicle weight	tons
14	EngTotalFuelUsed	Total fuel consumed	litres

Among the many factors that shape fuel use—rear-axle weight, average speed, driver behaviour, cruise-control usage, tyre pressures, road profile, and seasonal variation—gross vehicle weight and road slope dominate. Figure 1.3 (Chapter 1) shows how increases in either metric markedly raise both the average and the spread of consumption. Capturing GCVW and slope signals provides a necessary baseline to control for road and loading effects, enabling more accurate detection of driving or vehicle anomalies.

3.2 APS Failure Dataset

This section includes a description of the APS architecture and the operational data acquisition process used to collect thirty-day driving records from vehicles with and without E-APU failures. These time-series data form the basis for designing the anomaly detection framework and human-expert analysis described in Chapter 5.

3.2.1 Air Pressure System of HDVs

The E-APU serves as the central component of the APS, supplying pressurized air to the braking and suspension systems in HDVs. It features an electronically controlled air dryer integrated system with a multi-circuit valve arrangement. This valve system distributes pressurized air to various vehicle circuits, each equipped with pressure sensors to monitor system conditions. The E-APU is designed to ensure that a failure in one circuit does not compromise the functionality of the entire braking system. Figure 3.3 illustrates a schematic representation of the APS configuration with the E-APU centrally located.

The E-APU continuously monitors vehicle and engine conditions electronically, facilitating an optimized compressor operation cycle. Air compression is reduced during periods of high engine load and increased during engine overrun phases to maximize fuel efficiency. Additionally, the E-APU maintains outlet pressure within specified limits to ensure reliable braking system performance. It also coordinates regeneration processes to preserve the cleanliness and quality of the air supply, eliminating moisture and contaminants that could cause corrosion, component wear, or system

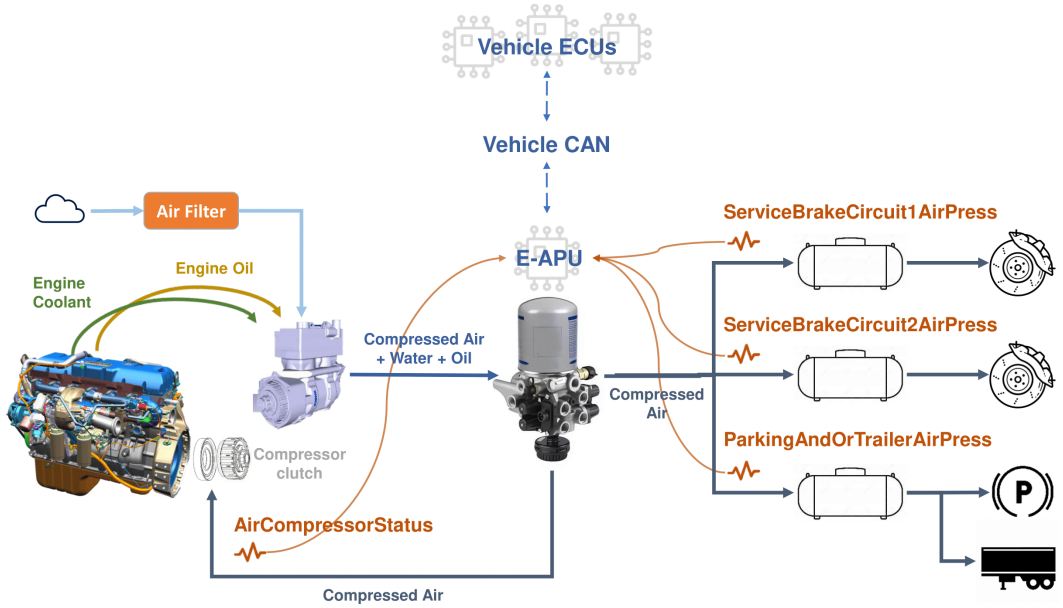

Figure 3.3 Schematic overview of the E-APU in the HDV Air Pressure Systems (Mumcuoglu et al., 2024b).

malfunctions. In the event of faults such as mechanical breakdowns, excessive pressure conditions, or activation of fail-safe mode, the E-APU transmits the operational status via the CAN bus.

The root causes of E-APU failures can include design flaws, manufacturing defects, or harsh operating conditions. A common root cause is component wear due to poor manufacturing, overuse of the system, or contaminated air intake. Additionally, the presence of moisture in the air supply due to air dryer malfunction results in corrosion affecting the compressor, valves, and air storage tanks. Severe operating conditions, including extreme ambient temperatures, and poor driving practices also represent significant contributing factors to failures across various E-APU components.

3.2.2 Data Acquisition

For the APS failure detection application, an operational dataset was constructed from F-MAX Trucks deployed across Turkey and Europe. These vehicles are incorporated within Ford Otosan's extensive connectivity framework, where operational vehicle data are recorded in their cloud infrastructure, enabling both real-time and retrospective monitoring capabilities essential for health monitoring and anomaly detection applications. This dataset contains time-series driving signals gathered

over 30-day intervals from two distinct vehicle groups: 30 vehicles with anomalous behavior that experienced E-APU failures necessitating component replacement, and 110 vehicles with normal operation maintaining clean maintenance histories. For vehicles exhibiting anomalous behavior, the dataset encompasses daily driving records from the period immediately before failure occurrence, referred to as run-to-failure data. Conversely, for vehicles with normal operation, historical 30-day data sequences from different periods throughout the year were chosen. The complete dataset encompasses 3550 driving records, with each record representing one day's operational data, collected from a total of 140 distinct vehicles. Comprehensive data details are shown in Table 3.4, while APU-related signals are specified in Table 3.5.

Table 3.4 Data description.

	Healthy	Anomaly
Number of vehicles	110	30
Number of daily records	2,779	771
Number of files per vehicle (min-max)	15-30	17-30
Number of drive cycles	18,556	5,552
Avg. number of data points per record	32,988	
Number of signals	9	
Nominal sampling rate	1-5 Hz	

Table 3.5 APS Related Signals

#	Signal name	Sampling period
1	Air compressor status	on change
2	Brake pedal position	on change
3	Engine speed	1 sec.
4	Service brake circuit 1 air pressure	1 sec.
5	Service brake circuit 2 air pressure	1 sec.
6	Parking and/or trailer air pressure	1 sec.
7	Tachograph vehicle speed	1 sec.
8	Vehicle total traveled distance	10 sec.
9	Engine total hours of operation	300 sec.

4. DETECTING ANOMALOUS FUEL CONSUMPTION IN HEAVY-DUTY VEHICLES

This chapter presents a comprehensive approach to detecting anomalous FC in HDVs through intelligent classification models that account for the complex interplay between vehicle load and road conditions. Existing threshold-based approaches often fail to consider the significant impact of vehicle weight and road slope on fuel efficiency, leading to inaccurate anomaly identification. A classification framework is developed that combines bagged decision trees with a novel quartile-based labeling methodology. The approach progressively refines FC thresholds by incorporating weight normalization and multi-level slope segmentation, enabling more accurate identification of both high FC patterns and true outliers. Time-series vehicle telemetry is transformed into meaningful features through sliding window analysis, enabling accurate distinction between normal operational variations and genuine FC anomalies that require intervention.

4.1 Load and Slope-Aware Fuel Consumption Classification Framework

Two machine learning models are developed to characterize FC behavior in HDV driving data: the high FC model and the outlier FC model. Both models employ an ensemble of bagged decision trees for FC classification. As detailed in Section 2.2.1, bagging (Bootstrap Aggregation) with decision trees forms the foundation of Random Forest algorithms, which enhance bagging by introducing feature randomness alongside sample randomness. The bagged decision tree approach is selected for its interpretability and robust performance with the available dataset characteristics. Specifically, this method enables comprehensive feature importance analysis, which provides valuable insights into the most significant predictors of FC anomalies, as will be presented in the results section.

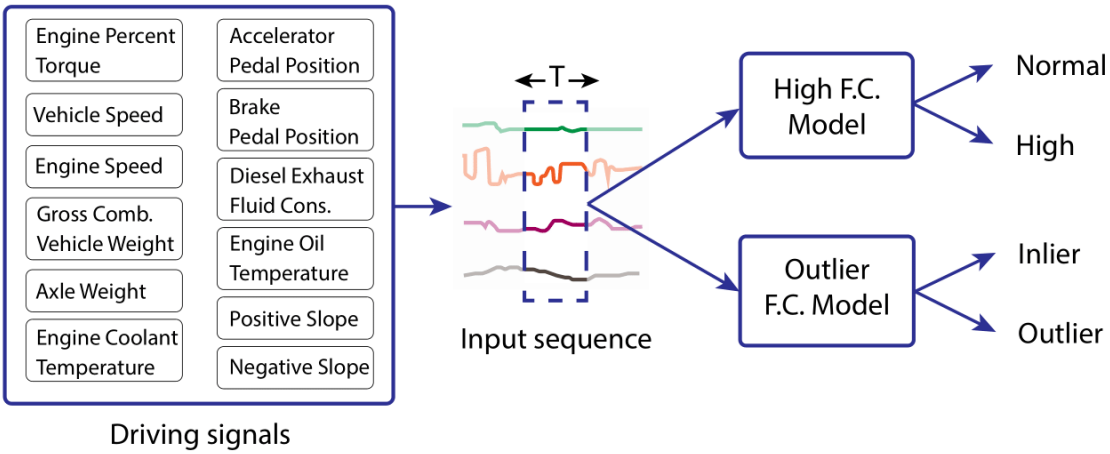

Figure 4.1 FC classification system overview (Mumcuoglu et al., 2023).

From an engineering validation perspective, the OOB feature importance estimation provides a robust framework to verify that the model captures physically meaningful relationships in HDV fuel consumption. The OOB importance scores allow engineers to confirm that identified predictors align with established domain knowledge of fuel efficiency factors, such as vehicle load, road gradient, and engine operating conditions. This validation approach ensures model interpretability and enables systematic feature selection for practical implementation in commercial vehicle fleet monitoring systems.

The training data preparation involves a sliding window approach to transform time-series vehicle signals into static feature vectors. Simple statistical measures—mean, standard deviation, minimum, and maximum—are calculated for selected signals within each sliding window. However, special consideration is required for road slope signal due to their bidirectional nature. Road slope signals typically contain both positive and negative components corresponding to uphill and downhill grades. When these signals are averaged over conventional time windows (5-10 minutes), the positive and negative slope values tend to cancel each other out, resulting in significant loss of critical gradient information that directly impacts FC patterns.

To preserve this essential slope information, the original road slope signal is decomposed into separate positive and negative slope components. The positive slope component is extracted as:

$$positive\_slope = \begin{cases} slope & \text{if } slope > 0 \\ 0 & \text{otherwise} \end{cases} \tag{4.1}$$

Similarly, the negative slope component is extracted as:

$$negative\_slope = \begin{cases} slope & \text{if } slope < 0 \\ 0 & \text{otherwise} \end{cases} \tag{4.2}$$

The decomposition is illustrated in Figure 4.2, where the uphill and downhill portions of a sample slope trace are shaded green and red, respectively, with the corresponding positive- and negative-slope signals plotted beneath.

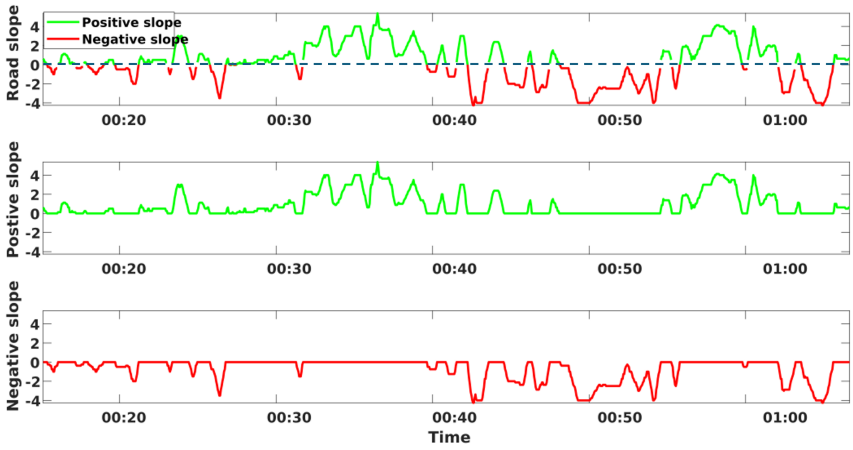

Figure 4.2 Separation of road gradient into positive and negative segments (Farea et al., 2023).

This separation enables the feature extraction process to capture the distinct FC dynamics associated with ascending and descending road segments. The decomposed slope signals, along with other vehicle parameters, are processed using the sliding window aggregation method described earlier.

For each window, four summary statistics—mean, standard deviation, minimum, and maximum—are computed for every signal, including the slope components. The resulting feature vectors are then fed into both classification models (Figure 4.1), providing a compact yet informative representation of each driving segment for ensemble learning.

4.2 Weight-Normalized Quartile Labeling with Multi-Level Slope Segmentation

To characterize FC levels as high or normal and establish discrete calibration thresholds, the baseline approach (Gong et al., 2021) utilized quartiles of FC per 100 km. A FC level is labeled as high if it exceeds the overall upper quartile (representing the top 25% of FC levels), while remaining levels are labeled as normal FC (Method 1). However, reasonable FC threshold determination requires consideration of vehicle weight and road slope influences. To address this limitation, a progressive refinement methodology is implemented through the following steps:

- Initially, FC thresholds are obtained using weight-normalized FC quartiles (Method 2).
- Subsequently, the dataset is segmented into 4 equally-distributed slope levels based on average slope intervals, with weight-normalized FC quartiles computed for each segment (Method 3).
- Finally, the dataset is partitioned into 16 equally-distributed slope levels based on average slope intervals, with weight-normalized FC quartiles calculated for each partition (Method 4).

At each refinement step, FC levels are labeled as high or normal based on the obtained thresholds.

For each sample, the weight-normalized average FC per 100 km, denoted as WAFC, is computed across sliding windows using the following formula:

$$WAFC = \frac{\Delta FC}{\Delta D \times GCVW} \times 100 \tag{4.3}$$

where ΔFC denotes the total fuel consumed in liters, ΔD represents the total distance covered in kilometers within the window, and $GCVW$ corresponds to the gross combination weight of the vehicle during the data collection period.

The training samples for both FC classification models are labeled according to WAFC quartile distributions. These quartiles are computed across the 16 slope levels established in the slope segmentation methodology (Method 4). The labeling criteria differ between the two models:

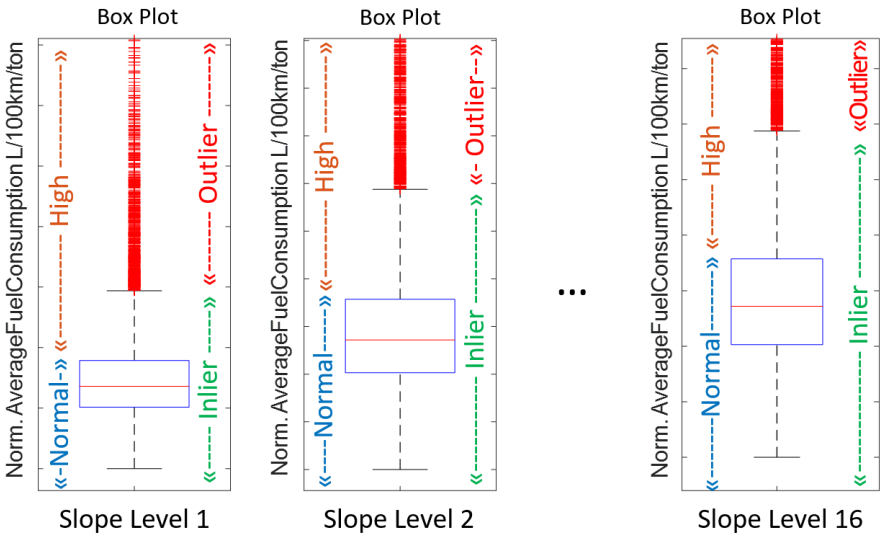

Figure 4.3 Quartile-based labeling scheme for high FC and outlier FC models (Mumcuoglu et al., 2023).

- **High FC Model**: Samples with WAFC exceeding Q_3 within their respective slope levels are designated as high FC, while remaining samples are classified as normal FC.
- **Outlier FC Model**: Samples with WAFC surpassing $Q_3 + 1.5 \times (Q_3 - Q_1)$ within their respective slope levels are marked as outlier FC, with all other samples categorized as inlier FC.

Here, Q_1 and Q_3 represent the first and third quartiles of the boxplot distribution. The complete labeling strategy is illustrated in Figure 4.3.

Figure 4.4(a) highlights the limitations of the labeling method proposed in (Gong et al., 2021), which identifies trips as having high FC simply if they exceed a certain torque threshold. This approach overlooks scenarios where higher torque is justified by increased load, causing mislabeling. Introducing weight-normalized FC quartiles results in a more balanced labeling across torque-weight space (Figure 4.4(b)). However, as illustrated in Figure 4.5(b), this approach still struggles to account for the effects of varying slope conditions. Trips requiring higher torque on steep slopes continue to be incorrectly flagged as high FC. To address this issue, separate weight-normalized FC quartiles were defined for four discrete slope intervals, producing more uniformly labeled data across varying slopes (Figure 4.5(c)). Ultimately, refining the slope intervals further to sixteen distinct levels achieved a high degree of uniformity in labeling, ensuring accurate representation of FC across diverse loading and slope conditions (Figure 4.4(d), Figure 4.5(d)).

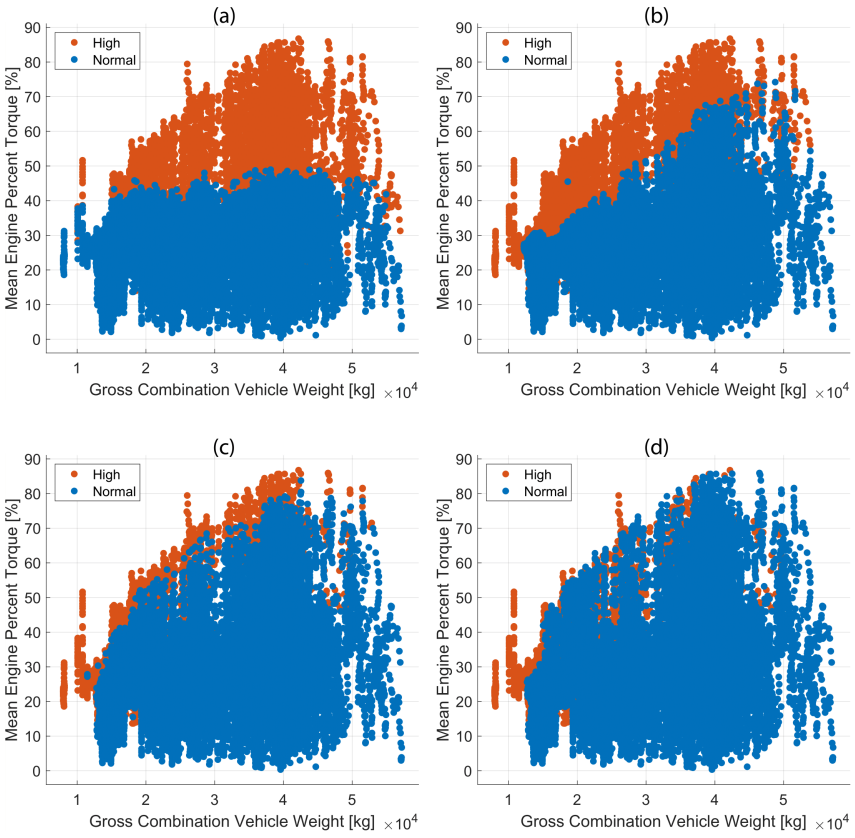

Figure 4.4 Engine torque versus vehicle weight under four different labeling strategies: method 1 (a), method 2 (b), method 3 (c), and method 4 (d) (Mumcuoglu et al., 2023).

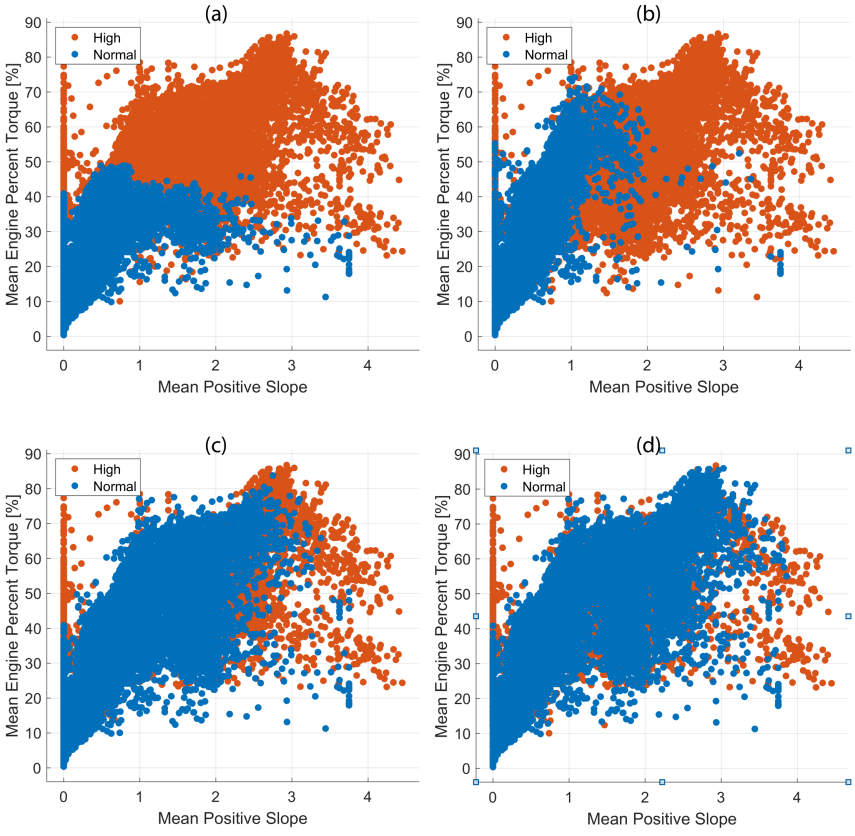

Figure 4.5 Engine torque versus road slope under four alternative labeling approaches: method 1 (a), method 2 (b), method 3 (c), and method 4 (d) (Mumcuoglu et al., 2023).

4.3 Interactive Dashboard for Fleet-Level Fuel Consumption Monitoring

An interactive MATLAB-based dashboard was developed to provide fleet managers with practical insights into FC behavior at fleet, vehicle, and individual trip levels. The dashboard integrates the predictions from the load- and slope-aware FC classification models and presents them through three complementary visualization screens, guiding users from an overall fleet analysis to detailed trip diagnostics.

Fleet Overview (Bar Plots Tab)
The first visualization (Figure 4.6a) ranks vehicles based on the percentage of their travel time flagged as *High FC* or *Anomaly FC*. High FC classifications are indicated with blue bars, signifying consistently excessive fuel use, while Anomaly FC events—represented by orange bars—highlight sporadic abnormal consumption. Fleet managers can quickly identify vehicles that consistently deviate from expected fuel-efficiency patterns, thus prioritizing them for detailed investigation or maintenance.

Distribution Analysis (Histograms Tab)
The second screen (Figure 4.6b) provides histograms showing how frequently vehicles are classified within the High FC and Anomaly FC categories. This view offers a clear depiction of the distribution and variation of fuel-consumption behaviors across the fleet. Adjustable histogram bins allow managers to explore and identify the fleet's overall consumption patterns and to pinpoint potential outlier vehicles.

Trip-Level Inspection (Time Series Tab)
The third visualization screen (Figure 4.7) enables detailed inspection of individual vehicle trips. After selecting a specific vehicle and trip record, each 10-minute trip segment is plotted on an interactive map using distinct markers: blue triangles indicate normal consumption, red triangles indicate high FC, and red crosses denote anomalous consumption segments. A supplementary information panel provides essential trip metrics such as duration, average FC, and distance traveled, assisting in the in-depth analysis and root-cause investigation.

In addition to monitoring capabilities, this dashboard lays the foundation for future enhancements, including route optimization studies. By correlating fuel-consumption classifications with driver behavior, road topology, and other contextual data, fleet managers could potentially optimize driving practices and routes to achieve minimal FC and improved operational efficiency.

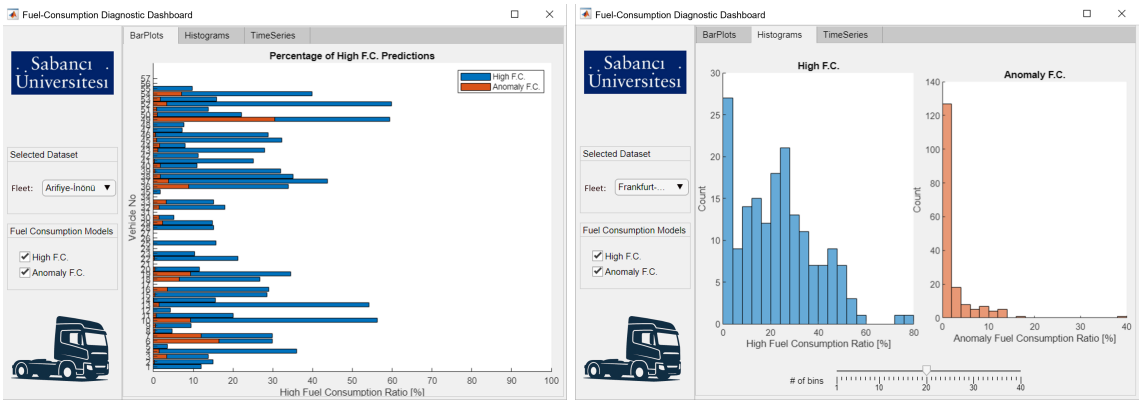

Figure 4.6 Fleet-level dashboard views: (a) bar chart summarizing high and anomalous FC ratios for each truck; (b) histograms illustrating the fleet-wide distribution of FC classes.

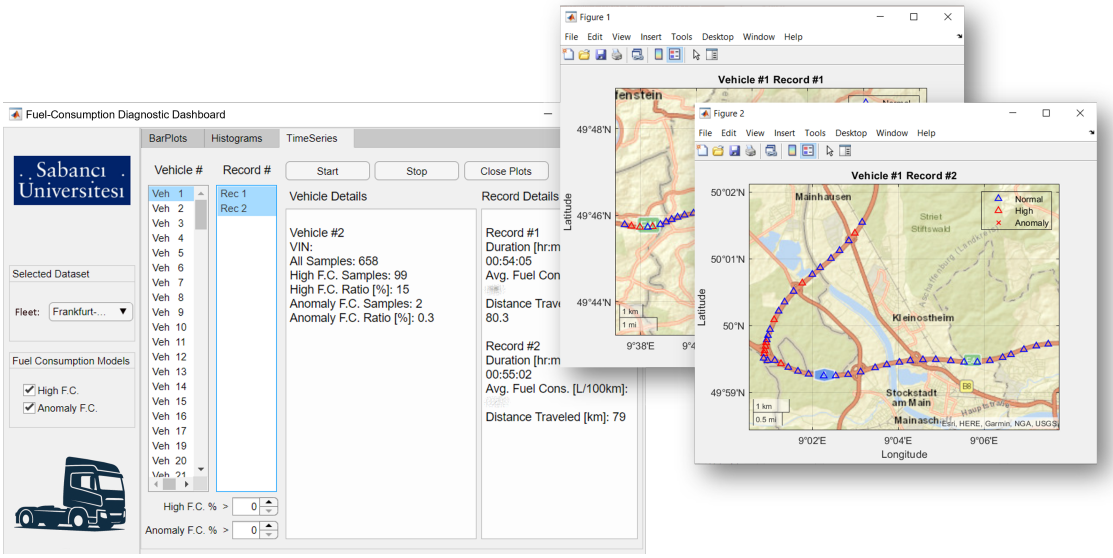

Figure 4.7 Animated trip-level dashboard visualizing a selected truck's route, with each 10-minute segment color-coded by fuel-consumption classification.

5. AIR PRESSURE SYSTEM FAILURE DETECTION IN HEAVY-DUTY VEHICLES

This chapter addresses the detection of APS failures in HDVs through a hybrid approach that combines ML techniques with domain expertise. The proposed methodology develops a comprehensive failure detection framework, beginning with a baseline approach that integrates LSTM autoencoders with Human Expert Analysis for reliable failure identification. This baseline is enhanced through XAI modules, including EBM and an innovative LLM-based agentic framework that decomposes diagnostic reasoning into specialized AI agents. This multi-tiered approach transforms raw operational data from multiple APS sensors into fully interpretable insights for system failure prevention.

5.1 Data Processing and Feature Extraction

Effective data preprocessing is crucial for extracting meaningful patterns from APS operational data and developing reliable data-driven models. The APS dataset presents several challenges including signals with varying sampling rates and types (Table 3.5), extended stationary periods where vehicles remain operationally active but motionless, and intermittent data gaps caused by logging errors or connectivity issues. These characteristics necessitate a systematic preprocessing approach to isolate genuine operational periods and ensure data quality for subsequent analysis.

The preprocessing workflow begins by segmenting daily driving records into discrete drive cycles based on temporal continuity, where consecutive data points are separated by no more than 5 minutes. Within each drive cycle, signals are interpolated to achieve uniform sampling frequencies according to their individual characteristics. Subsequently, feature extraction is performed using sliding windows of 20 minutes with 10-minute shifts, computing moving statistics including mean, standard devia-

tion, and minimum values. This approach reduces computational complexity while preserving essential temporal patterns such as duty cycle variations within each window period.

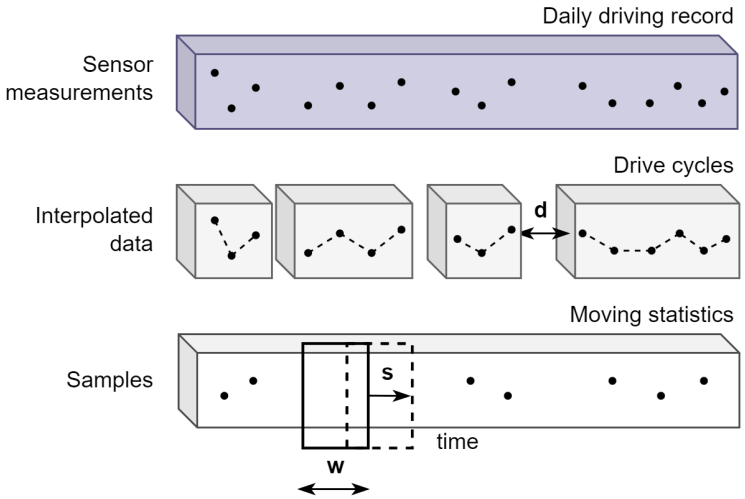

Figure 5.1 Data preprocessing workflow: segmenting daily driving records into drive cycles, applying data interpolation and sampling via moving statistics.

The final preprocessing step applies vehicle and engine speed thresholds to eliminate stationary or non-operational segments, ensuring that only meaningful driving data contributes to the analysis. Each resulting sample point represents aggregated information from a 20-minute operational window. The complete preprocessing workflow is illustrated in Figure 5.1, where w' and s' denote the sliding window length and shift respectively, while 'd' represents the minimum time gap threshold for drive cycle formation.

Following the preprocessing stage, a set of handcrafted features was extracted from each sliding window. These features were carefully selected based on expert domain knowledge, aiming to capture pressure fluctuations, compressor behavior, and vehicle dynamics that are indicative of early-stage APS failures. The same feature set is used as input to the traditinal ML methods evaluated in this study, including the LSTM-AE and the EBM, and is also the foundation for expert assessments described in the HEA section.

The list of extracted features is summarized in Table 5.1.

Table 5.1 Extracted features computed over each sliding window.

No.	Feature
1	Duty cycle (*DutyCycle*)
2	Air compressor on/off count (*AC_on/off_count*)
3	Minimum pressure of service brake circuit 2 (*P2_min*)
4	Standard deviation of service brake circuit 2 pressure (*P2_std*)
5	Minimum pressure of service brake circuit 3 (*P3_min*)
6	Standard deviation of service brake circuit 3 pressure (*P3_std*)
7	Mean brake pedal position (*BrakePedalPos_mean*)
8	Mean engine speed (*EngineSpeed_mean*)
9	Standard deviation of engine speed (*EngineSpeed_std*)
10	Standard deviation of vehicle speed (*VehicleSpeed_std*)

5.2 Baseline APS Failure Detection Methods

This section outlines a two-tier APS failure-detection pipeline. First, a semi-supervised LSTM auto-encoder learns the normal pressure dynamics of healthy vehicles. Its anomaly scores are then cross-checked with structured Human Expert Analysis to filter false alarms and relate deviations to physical failure modes.

5.2.1 Design of an LSTM Autoencoder for Failure Detection

Autoencoders are widely adopted for anomaly detection, especially in semi-supervised contexts, due to their effectiveness in modeling normal system behavior. They consist of two core components: an encoder, which compresses the input data into representative features, and a decoder, which reconstructs the original input from these features. In semi-supervised anomaly detection, autoencoders are trained exclusively on normal operational data, allowing the model to reconstruct normal sequences accurately, reflected by a low reconstruction error. Conversely, anomalous data will yield higher reconstruction errors, signaling deviations from learned normal patterns.

LSTM networks are frequently employed in autoencoders due to their capability to

capture dynamic temporal patterns in sequences. As a specialized form of RNNs, LSTMs effectively model both short-term fluctuations and long-term dependencies, making LSTM-AE particularly suited for multivariate time-series anomaly detection.

For detecting APS failures in HDVs, we propose a semi-supervised LSTM-AE model (illustrated in Figure 5.2). The model processes operational data sequences, using two LSTM layers with dropout regularization in the encoder to extract representative features. The decoder symmetrically mirrors this structure with two LSTM layers to reconstruct the original data from the encoded features. An overcomplete autoencoder architecture, featuring encoding dimensions greater than the input size, is adopted due to its superior capacity for modeling intricate underlying processes compared to undercomplete autoencoders with smaller encoding dimensions (Ranjan, 2020). The resulting reconstruction errors provide vehicle-specific anomaly scores, enabling early detection of APS failures and proactive maintenance interventions.

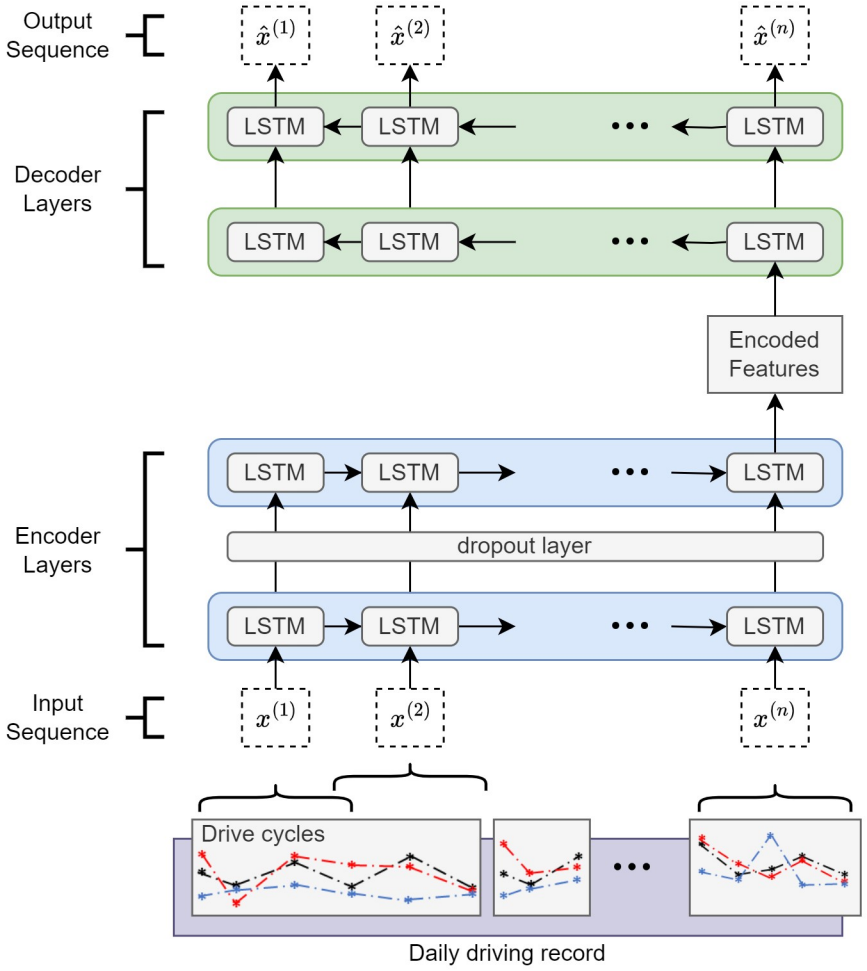

Figure 5.2 LSTM-based autoencoder architecture for APS anomaly detection (Mumcuoglu et al., 2024b).

To effectively capture the underlying temporal dynamics and relationships among

selected APS signals, LSTM networks (Hochreiter & Schmidhuber, 1997) are utilized. A standard LSTM architecture consists of three primary layers: an *input layer*, a *recurrent hidden layer*, and an *output layer*. Unlike classical neural networks, LSTMs possess internal memory that enables them to retain information across long sequences. This memory capability is governed by three specialized gates: the *forget*, *input* (update), and *output* gates. Each gate is implemented as an independent neural network with matching dimensions and sigmoid activation functions.

During training, sequential multivariate time-series data are provided in the form $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t, \ldots]$, where $\mathbf{x}_t$ denotes the multivariate input vector at timestamp t. These inputs are processed through the LSTM units to learn both short- and long-term dependencies critical for detecting subtle temporal anomalies.

5.2.2 Human Expert Analysis (HEA)

Four critical signals provide direct insights into the health of the APS: the compressor status signal and the pressure signals from the three primary brake circuits (illustrated in Figure 5.3). Any potential APS failure typically appears as deviations in these operational signals. To effectively capture these anomaly patterns, we propose monitoring the following three derived indicators:

Duty cycle: The duty cycle, defined as the ratio of the air compressor's active operation time to the total operational period, serves as a key indicator of APS health. Typically, a healthy HDV exhibits a lower and more stable duty cycle, whereas anomalies cause the compressor to operate more frequently or continuously. When the APS fails to maintain the required air pressure or quality, the compressor is forced into excessive operation, resulting in notably higher duty cycle values.

Compressor on/off count: Apart from maintaining system pressure, the APU also ensures air quality within the APS through a cyclic "regeneration" process, periodically releasing and replenishing small quantities of air. Although this cycling is generally normal, abnormal frequency or irregular patterns in compressor state changes might signal early-stage APS issues. Thus, we propose tracking the frequency of compressor state transitions within a sliding time window as a novel indicator to detect unusual fluctuations associated with potential APS anomalies.

Minimum pressure: Air leakage is among the primary concerns in APS systems, causing significant operational stress. Although the minimum pressure levels in

brake circuits are strongly influenced by brake usage, persistent air leakage produces consistently lower pressure levels, clearly distinguishable from healthy operational baselines. Consequently, monitoring minimum brake circuit pressures throughout driving periods serves as an additional indicator for identifying anomalous APS conditions.

Let the compressor state be represented by the binary sequence $S = [s_1, s_2, s_3, \ldots, s_n]$, where $s_i \in \{0, 1\}$ denotes the compressor's off (0) or on (1) status at sample i. For each sample index k, a sliding window w_k of fixed length W is evaluated.

The duty cycle within w_k is computed as

$$DC_k = \frac{1}{W} \sum_{w_k} s_i, \tag{5.1}$$

giving the fraction of time the compressor is active.

The total number of on/off transitions in the same window is

$$C_k = \sum_{w_k} |s_{i+1} - s_i|, \tag{5.2}$$

where k identifies the current position of the sliding window.

Let the pressure signal from a brake circuit be $P = [p_1, p_2, p_3, \ldots, p_n]$. The minimum pressure within window w_k is

$$\mathrm{MinP}_k = \min_{w_k} \{p_i\}. \tag{5.3}$$

The set $\{DC_k, C_k, \mathrm{MinP}_k\}$ constitutes the proposed indicators for characterising APS behaviour over time.

To investigate APS failure behaviour, daily averages of duty cycle, compressor on/off count, and minimum brake-circuit pressure are plotted for representative vehicles in Figure 5.3. Panel (a) shows a healthy vehicle whose signals remain stable within nominal ranges: the duty cycle remains consistently low, compressor switching frequency is modest, and minimum pressure stays at safe levels. In contrast, panels (b) and (c), corresponding to two vehicles that eventually experienced E-APU failure, illustrate three distinct anomaly patterns: (i) gradual performance drift, exemplified by steadily increasing duty cycles; (ii) isolated spikes indicating brief, severe pressure disruptions; and (iii) persistent off-nominal signal levels, such as consistently low minimum pressure or elevated compressor switching. These differences highlight early indicators of impending failure, yet significant overlap between normal

and faulty signals persists, complicating reliable discrimination and underscoring the necessity of advanced feature analysis (see Figure 5.4).

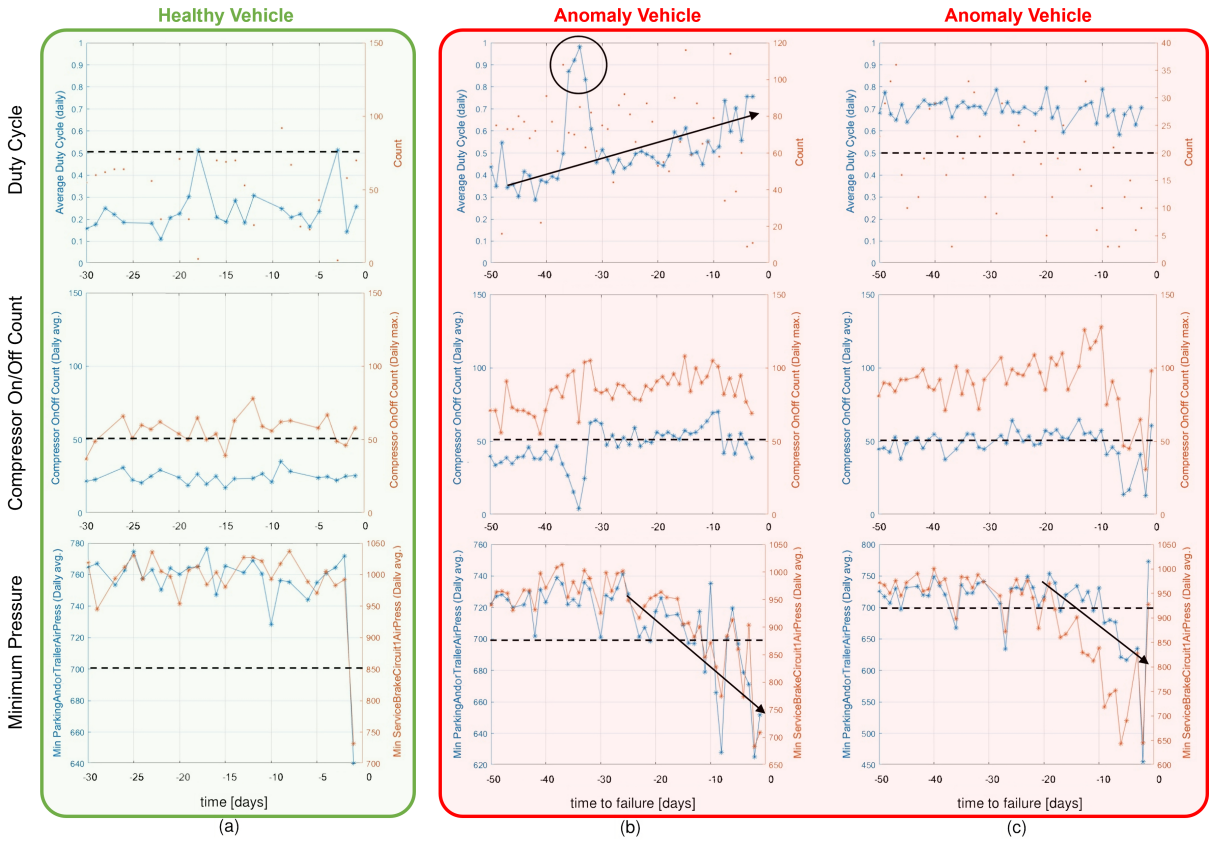

Figure 5.3 Temporal trends in duty cycle, compressor switching frequency, and minimum pressure levels, derived from data of healthy (a) and faulty (b–c) vehicles (Mumcuoglu et al., 2024b).

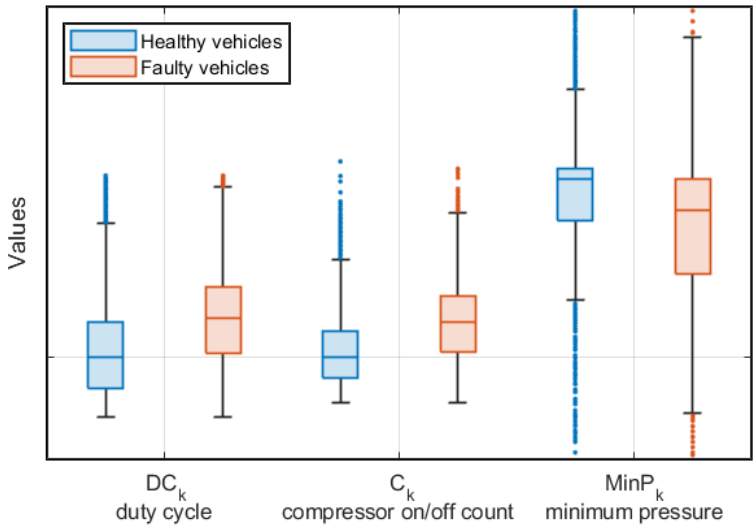

Figure 5.4 Box plots of the proposed indicators, showing distinctions between healthy and faulty vehicles, with overlaps highlighting the challenge of anomaly detection (Mumcuoglu et al., 2024b).

Based on the preceding analysis, a manual labeling procedure was developed in collaboration with data analytics specialists and brake system experts. Initially, acceptable daily-average ranges were defined for the duty cycle, compressor on/off count, and minimum pressure, derived from data of carefully selected healthy vehicles. Each vehicle's APS failure risk was then evaluated by experts who reviewed deviations from these established limits and identified anomalous patterns within these three features. Risk scoring involves assigning each feature a grade: 0 (no visible anomaly), 1 (potential anomaly that warrants monitoring), or 2 (clear anomaly present). Vehicles receive a cumulative score of up to 6 flags across the three features, and this total determines whether they are labeled as normal or anomalous.

Enhanced Human Expert Analysis (HEA+): While the three core indicators provide valuable insights into APS health, distinguishing between genuine anomalies and operational variations remains challenging. High duty cycle values, for instance, may result from legitimate heavy braking during demanding driving conditions rather than system deterioration. To address this limitation, HEA+ incorporates brake usage patterns as a contextual indicator to enhance anomaly discrimination.

The average brake pedal position within window w_k is computed as

$$\text{AvgBrake}_k = \frac{1}{W} \sum_{w_k} b_i, \tag{5.4}$$

where b_i represents the brake pedal position at sample i.

By correlating duty cycle patterns with brake usage, HEA+ enables experts to differentiate between duty cycle elevations caused by system faults versus those attributed to operational demands. Figure 5.5 illustrates this enhanced analysis through four representative cases: (a) shows elevated duty cycle patterns with minimal brake usage, indicating potential system anomalies requiring attention; (b) demonstrates high duty cycle values coinciding with intensive braking, suggesting the elevation may be operationally justified; (c) presents moderate duty cycle levels that align closely with brake patterns, indicating normal system response; and (d) exhibits consistently low duty cycle values regardless of brake usage, confirming healthy system operation. This contextual analysis significantly reduces false positives by filtering out duty cycle anomalies that correlate with legitimate operational demands.

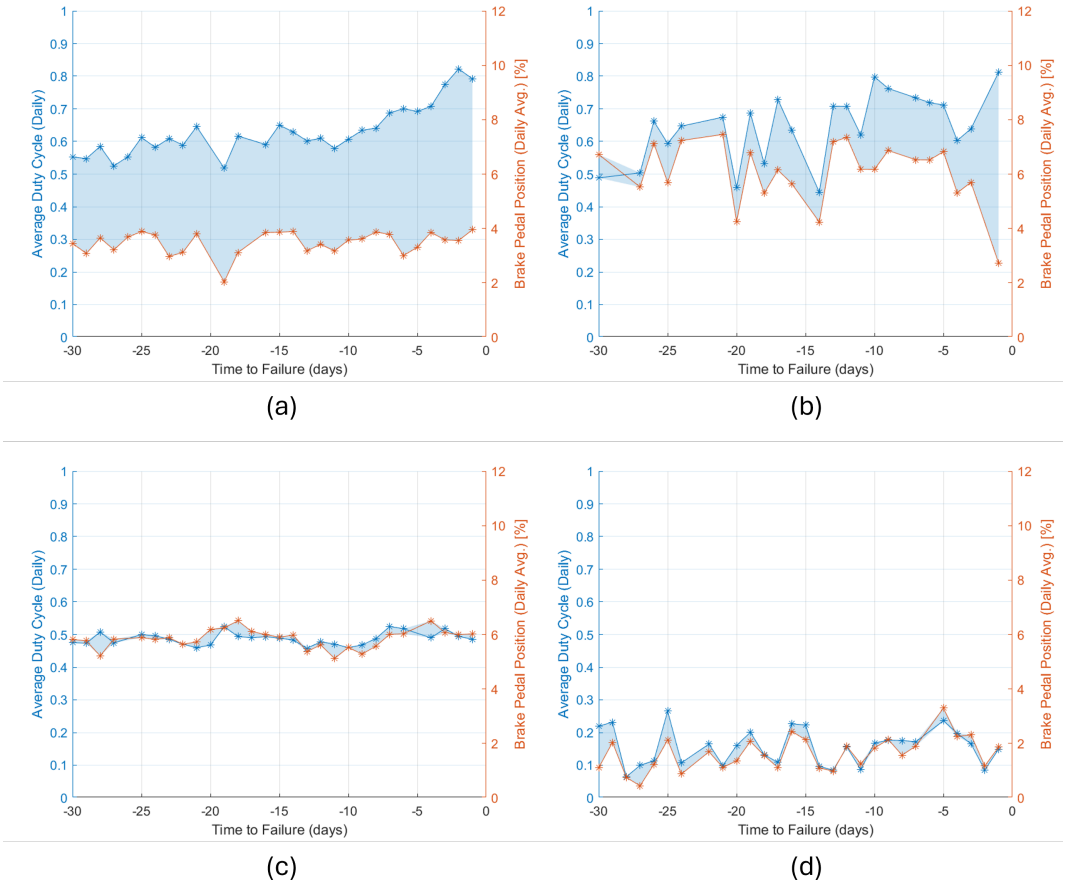

Figure 5.5 Enhanced expert analysis (HEA+) incorporating brake usage patterns: (a) elevated duty cycle with low brake usage suggests system anomaly; (b) high duty cycle correlating with heavy braking indicates operational demand; (c) moderate duty cycle matching brake patterns shows normal response; (d) consistently low duty cycle confirms healthy operation.

5.2.3 Predictive Maintenance Protocol for Anomalous Vehicle Detection

For vehicle anomaly detection using HEA, a threshold for the number of anomaly flags is determined by maximizing the F1 score. Similarly, when directly applying the LSTM-AE model, its anomaly score threshold is optimized via grid search, tuned to achieve the best F1 performance. The combined approach, integrating the ML model with HEA, requires simultaneous optimization of two thresholds: the ML-derived anomaly score and the number of HEA flags. Grid search optimization is again employed to identify threshold combinations yielding optimal performance in fault detection. The proposed predictive-maintenance protocol combining the ML model with HEA involves four main steps:

Step 1: Historical driving data are collected from the fleet, followed by preprocessing steps necessary for both ML-based analyses and expert review.

Step 2: The developed ML model calculates anomaly scores, identifying vehicles likely exhibiting anomalies.

Step 3: Human experts then examine the engineered features—duty cycle, compressor on/off count, and minimum pressure trends—for vehicles flagged as potentially anomalous.

Step 4: Vehicles that exceed both the ML anomaly-score threshold and the expert-derived anomaly-flag threshold are prioritized for immediate maintenance.

5.3 Explainable AI Modules

5.3.1 Explainable Boosting Machine

The EBM is an interpretable machine learning model extending GAM, as previously described in Section 2.3.4. It uses shallow, gradient-boosted decision trees trained iteratively on the handcrafted APS features listed in Section 5.1. Pairwise feature interactions, when relevant, are identified and similarly modeled through separate shallow decision trees.

EBM inherently provides transparent local explanations by visualizing how each feature or interaction contributes individually to a model's predictions. This intrinsic interpretability allows direct root-cause identification, making EBM suitable either as a complementary or alternative method to manual Human Expert Analysis. Implementation specifics, including hyperparameter tuning and detailed evaluation, are presented in the Experimental Results Chapter.

5.3.2 LLM-based Agentic Pattern Analysis

To further enhance expert analysis capabilities while maintaining interpretability, we developed an agentic framework that decomposes the diagnostic process into

specialized AI agents, each focusing on specific aspects of APS behavior. We implemented this framework using Google's Gemini 2.0 Flash (Google DeepMind, 2024) for rapid inference, providing a 1M-token context window, native tool-calling capabilities, and a one-month free API trial for experimental validation. Preliminary third-party evaluations confirm strong reasoning improvements over 1.5-Flash (Al-Hayeket al. , 2025; Balestri, 2025).

The framework comprises four specialized agents operating in a hierarchical structure. Three signal-specific agents analyze the core indicators independently: the Duty Cycle Agent evaluates duty cycle patterns in conjunction with brake usage (following HEA+ methodology), the Switching Pattern Agent examines compressor on/off count anomalies, and the Pressure Agent monitors minimum pressure deviations. Each agent receives daily-averaged signal values in text format and applies domain-specific pattern recognition to identify anomalous behaviors, temporal trends, and potential failure modes.

The agent prompts were carefully engineered to encapsulate expert knowledge and diagnostic reasoning patterns observed in traditional HEA processes. Each signal-specific agent incorporates domain-specific thresholds and operational ranges: duty cycle analysis considers values above 0.5 as potentially anomalous when sustained over 3-4 consecutive days without corresponding brake usage elevation; switching pattern analysis monitors compressor state changes above 50 cycles per day; and pressure analysis flags sustained drops below 700 kPa. The agents apply pattern recognition logic that distinguishes between isolated anomalies and persistent degradation patterns, mirroring expert diagnostic reasoning. The complete prompt specifications for all agents are provided in Appendix B.

The Decision Agent synthesizes outputs from all three signal-specific agents to render final diagnostic decisions. This agent implements a hierarchical weighting system where duty cycle serves as the primary health indicator, while compressor switching and minimum pressure function as secondary confirmatory signals. The decision logic follows expert prioritization: duty cycle anomalies alone can indicate system deterioration (probability 60-80%), while secondary indicators require corroboration with primary signals to suggest critical failures (combined anomalies yield 80-95% fault probability).

Let A_{DC}, A_{SW}, and A_{MP} represent the anomaly assessments from the Duty Cycle, Switching Pattern, and Pressure agents, respectively. The Decision Agent computes

the final classification through structured reasoning:

$$\text{Classification} = \begin{cases} \text{Faulty} & \text{if } A_{DC} = \text{True and sustained patterns} \\ \text{Faulty} & \text{if } A_{DC} = \text{True and } (A_{SW} \text{ or } A_{MP}) = \text{True} \\ \text{Healthy} & \text{otherwise} \end{cases} \tag{5.5}$$

Each agent outputs structured JSON responses containing anomaly flags, pattern descriptions, and diagnostic rationale. The complete prompt engineering approach ensures consistent analysis while preserving the interpretability and root cause identification capabilities essential for maintenance decision-making. This agentic framework effectively transfers human expert knowledge into a scalable, consistent diagnostic system that maintains expert-level diagnostic reasoning.

6. EXPERIMENTAL RESULTS

This chapter presents the evaluation results of the two anomaly-detection pipelines developed in this thesis. In Section 6.1, the load- and slope-aware quartile-labeling scheme combined with bagged decision trees is assessed for detecting excessive fuel consumption events. In Section 6.2, the LSTM autoencoder, both individually and in combination with Human Expert Analysis, is evaluated for early detection of air-pressure-system failures. For each task, experimental setup, tuning procedures, and prediction results are presented, emphasizing the practical value of these methods for fleet-scale predictive maintenance.

6.1 FC Anomaly Detection

6.1.1 Model Configuration & Feature Selection

Within the developed fuel-consumption classification system, both the High Fuel Consumption Model and the Outlier Fuel Consumption Model are implemented as ensembles of 30 bagged decision trees. Each tree is grown with a minimum leaf size of 8, and training continues up to 37,244 iterations to ensure convergence. The ensemble architecture—built on bootstrap sampling and majority voting—is described in the preliminaries and illustrated in Figure 2.2 (Section 2).

As illustrated in Figure 6.1, *percent torque* is the most influential feature, followed by *vehicle speed*, *road slope*, and *vehicle weight*. This ranking aligns closely with established vehicle-dynamics knowledge, highlighting these variables' significant impact on fuel consumption. Based on this analysis, the top twelve features were selected

as predictors. Four statistical measures (mean, standard deviation, minimum, and maximum) were calculated for each selected predictor over sliding windows, resulting in a total of 48 input features (12 predictors × 4 statistics).

Sliding window lengths of 5 and 10 minutes were examined, and their results are compared in the following subsections.

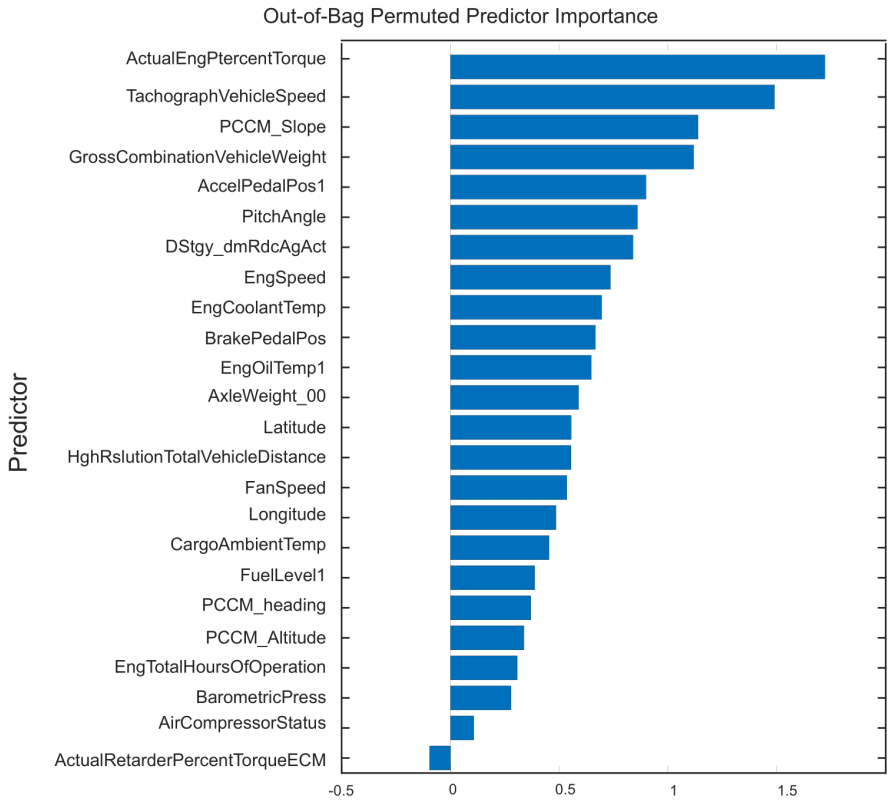

Figure 6.1 Feature importance ranking for fuel consumption prediction (Mumcuoglu et al., 2023).

6.1.2 High Fuel Consumption Model

Method Selection and Comparison

Table 6.1 reports the classification accuracy obtained with 5-minute and 10-minute sliding windows under four proposed labeling schemes, evaluated on Dataset A (Arifiye-İnönü route). When the data are labeled only by quartiles of average fuel consumption, following the approach of (Gong et al., 2021), the model achieves 95.4% and 96.4% accuracy for the 5 and 10-minute windows, respectively—well above the 86.6% reported in the original study. Because this method ignores vehicle weight and road slope, the resulting task is relatively easy.

Incorporating these two factors increases the complexity of the classification task, thereby lowering accuracy. Specifically, when employing weight-normalized fuel consumption quartiles with a finer slope segmentation (16 slope levels), the classification accuracy reduces to 88.5% for the 5-min window and 92.2% for the 10-min window, reflecting a more challenging yet realistic scenario.

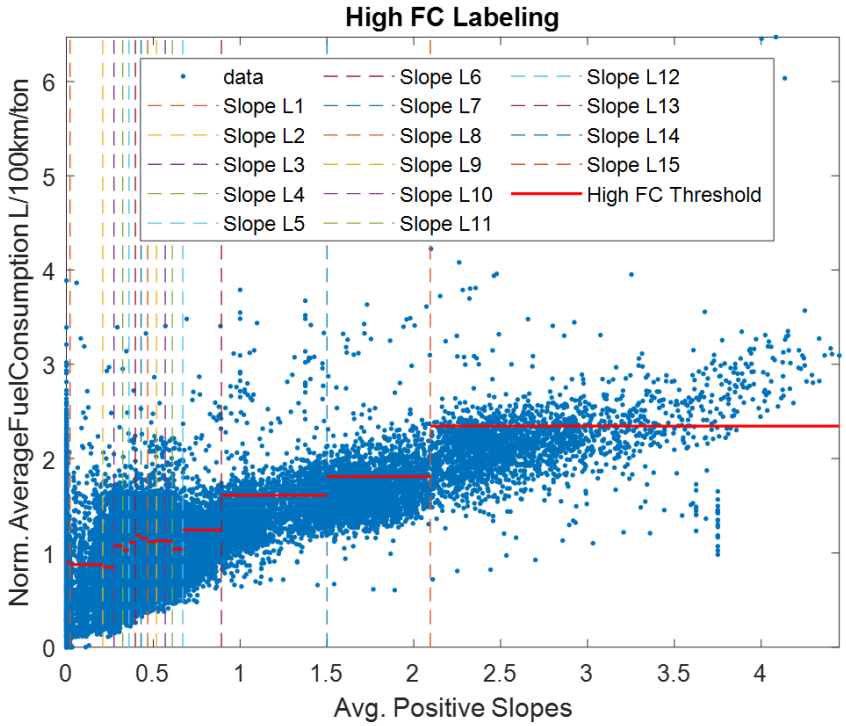

Figure 6.2 Weight-normalized average FC thresholds for high FC labeling model (Mumcuoglu et al., 2023).

The thresholds for each slope band are shown in Figure 6.2, which guides the model in distinguishing excessive fuel consumption from torque demands induced by load and grade. Across all four methods, the 10-min window consistently outperforms the 5-min window, indicating that a longer averaging period yields smoother, more reliable FC estimates. Accordingly, the 10-min setting with weight-normalized FC quartiles and 16 slope levels (Method 4) is selected and used in all subsequent analyses, including the outlier-FC study.

Table 6.1 High-FC results by method on Dataset A (Arifiye–İnönü route).

Labeling Method	Classification Accuracy [%]	
	5 min window	10 min window
FC Quartiles	95.4	96.4
Weight-Norm. FC Quartiles	92.2	96.1
Weight-Norm. FC Quartiles 4 Slope Levels	91.0	94.2
***Weight-Norm. FC Quartiles 16 Slope Levels**	88.5	**92.2**

Final Model Evaluation

Following the method selection, the proposed approach was validated across both datasets to assess its generalizability. Table 6.2 presents the final classification results for both the Arifiye-İnönü Dataset (Türkiye) and the Frankfurt-Würzburg Dataset (Germany). The model demonstrates consistent performance across different geographical and operational contexts, with average accuracies of 92.2% and 90.6% for the Turkish and German datasets, respectively. The slight performance difference between datasets can be attributed to variations in road characteristics, traffic patterns, and operational conditions between the two routes. The confusion matrices for both datasets and their combination are shown in Figure 6.3.

Table 6.2 High FC model validation results across datasets.

	5 min window				10 min window			
Dataset	run 1	run 2	run 3	avg.	run 1	run 2	run 3	avg.
Arifiye–İnönü	88.5	88.5	88.6	88.5	92.2	92.3	92.2	**92.2**
Frankfurt–Würzburg	87.6	87.6	87.6	87.6	90.7	90.6	90.6	**90.6**
All Combined	89.5	89.3	89.3	89.4	92.5	92.5	92.5	**92.5**

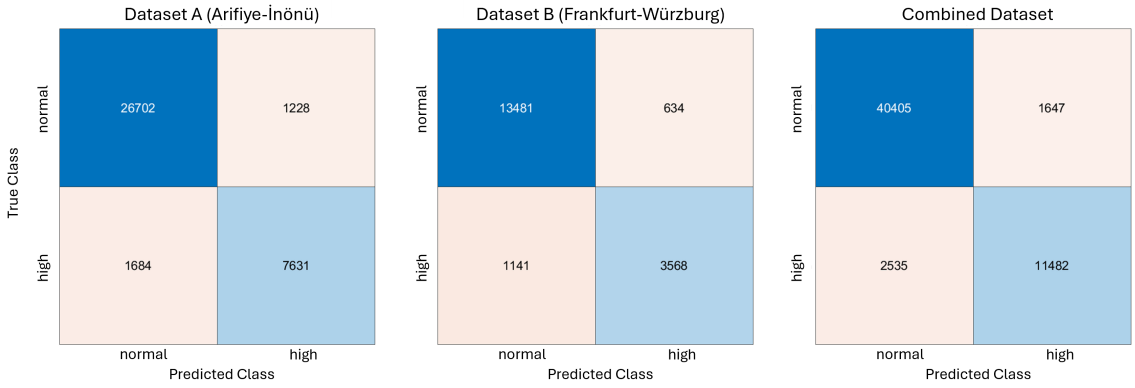

Figure 6.3 High FC model classification results.

6.1.3 Outlier Fuel Consumption Model

Figure 6.4 illustrates the weight-normalized average FC thresholds used to label outliers, with separate thresholds defined for each of the sixteen slope intervals. Since outlier cases represent only a small fraction of the datasets (approximately 1.8% for Dataset A, 2% for Dataset B, and 1.7% for the combined dataset, as detailed in Table 6.3), accuracy alone is insufficient for evaluating model performance due to class imbalance. Instead, performance metrics including precision, recall, and their harmonic mean, the F1 score, are utilized (see Appendix A).

Applying the proposed 16-level slope-aware labeling approach across both datasets, the classification results (Table 6.4) show F1 scores of 0.78 for Dataset A, 0.71 for Dataset B, and 0.70 for the combined dataset. These results indicate that the developed model effectively identifies genuine cases of anomalously high fuel consumption while limiting false alarms across different operational contexts. The confusion matrices for both datasets and their combination are shown in Figure 6.5.

Table 6.3 Outlier data statistics across datasets

Dataset	# of samp.	# of Outl.	Inlier Ratio	Outlier Ratio
Arifiye-İnönü Dataset	37,245	674	98.2%	1.8%
Frankfurt-Würzburg Dataset	18,824	380	98.0%	2.0%
All Combined	56,069	1,054	98.3%	1.7%

Table 6.4 Outlier FC Model Classification Results

Dataset	Accuracy	Precision	Recall	F1 Score
Arifiye-İnönü Dataset	99.25%	0.83	0.74	**0.78**
Frankfurt-Würzburg Dataset	98.91%	0.76	0.67	**0.71**
All Combined	99.06%	0.78	0.63	**0.70**

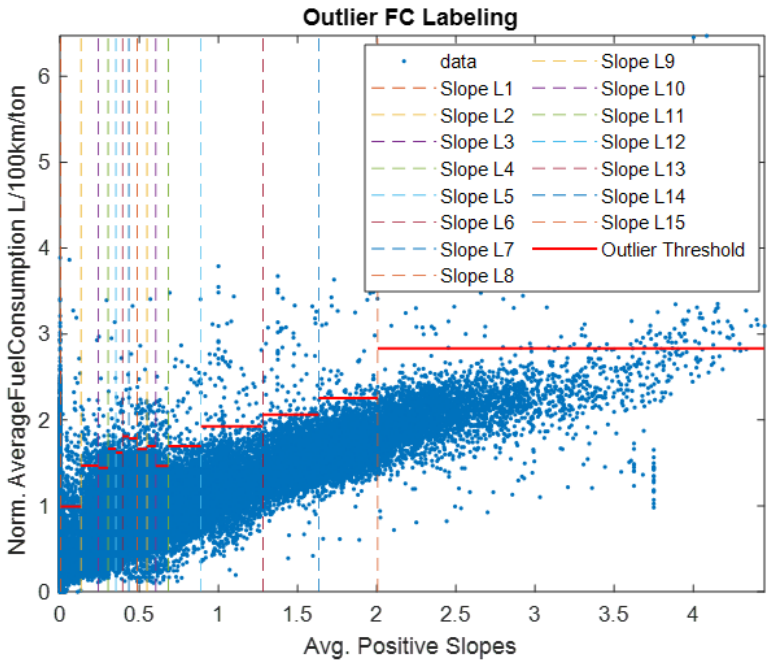

Figure 6.4 Weight-normalized average FC thresholds for outlier FC labeling model.

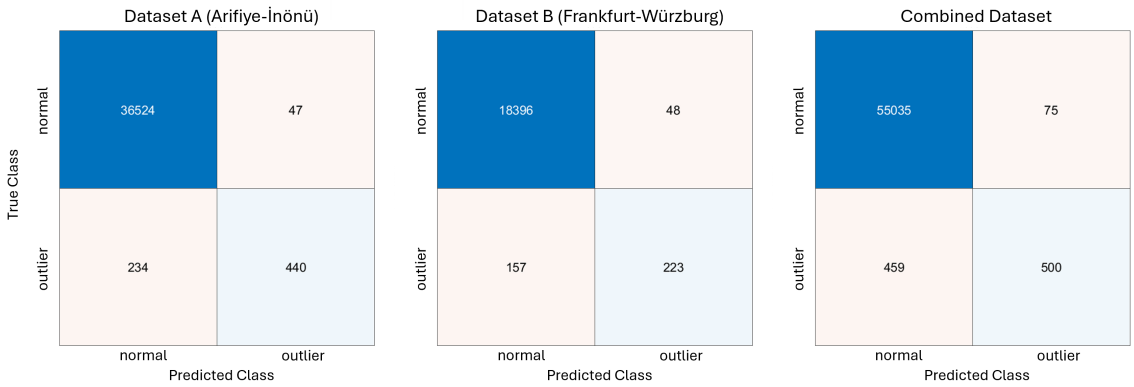

Figure 6.5 Outlier FC model classification results.

6.1.4 Example Fleet Analysis and Vehicle Comparison

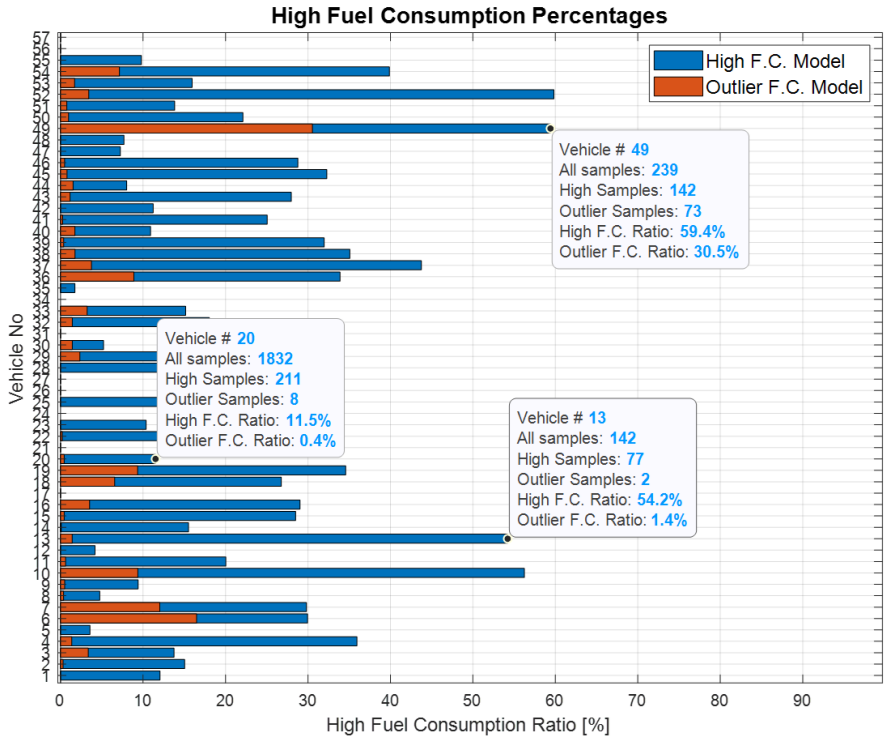

Figure 6.6 Evaluation of 57 HDVs by the proposed FC classification system (Mumcuoglu et al., 2023).

The following analysis is conducted on Dataset A (Arifiye-İnönü route) as a sample demonstration of the classification system's practical application. Figure 6.6 presents an example fleet analysis produced by the FC classification system. Several vehicles clearly stand out, either consistently flagged for high FC (e.g., units 52, 49, 10, 13) or frequently identified as FC outliers (e.g., units 49, 6, 7). Since these classifications already account for variations in vehicle load and road slope, the repeated flags likely reflect anomalous driving behaviors or potential mechanical

faults, highlighting specific vehicles that should be prioritized for further inspection or maintenance.

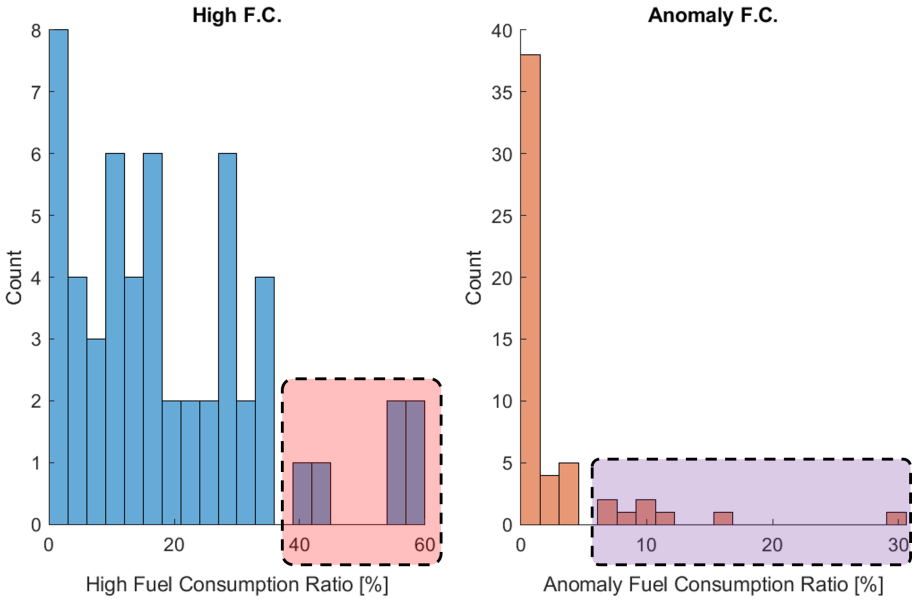

Figure 6.7 Fleet-wide distribution of high FC and anomaly FC ratios showing vehicles with elevated fuel consumption patterns.

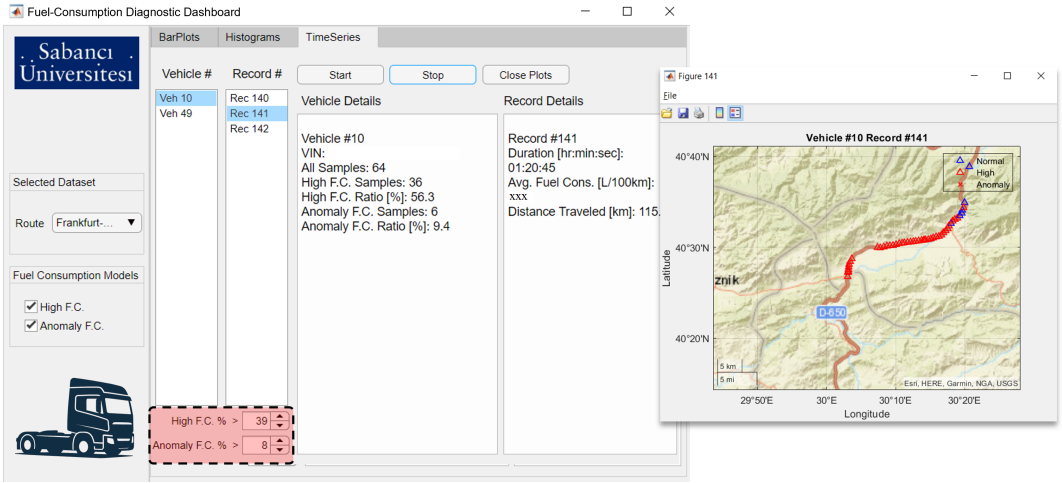

Figure 6.8 Fuel consumption diagnostic dashboard interface demonstrating vehicle filtering and identification capabilities.

To better understand the fleet's FC patterns, Figure 6.7 displays histograms showing the distribution of high FC and anomaly FC ratios across the fleet. The analysis reveals that a significant portion of vehicles exhibit high FC ratios below 38-39%, while anomaly FC percentages of 7-8% clearly stand outside the typical fleet performance range. These distributions demonstrate the system's ability to identify vehicles that deviate substantially from normal operational patterns. The practical application of these findings is illustrated in Figure 6.8, which shows the fuel

consumption diagnostic dashboard where fleet managers can filter vehicles based on these high FC and anomalous FC thresholds. In this example, vehicles 10 and 49 are prominently identified as outliers, enabling targeted investigation and maintenance scheduling for these specific units.

6.2 APS Failure Detection Results

This section presents experimental results for the comprehensive APS failure detection framework developed in Chapter 5. The evaluation encompasses baseline methods (LSTM-AE, HEA, and HEA+), explainable AI modules (EBM and LLM-based agentic analysis). Performance is assessed across the collected fleet dataset using standard anomaly detection metrics, with emphasis on interpretability and practical deployment considerations.

6.2.1 Data Division and Evaluation Protocol

The experimental evaluation employs distinct protocols tailored to each method's learning paradigm. For the semi-supervised LSTM-AE, four experiments vary the proportion of HV data allocated for training, with AV data remaining consistent across all tests. The detailed configuration is presented in Table 6.5.

Table 6.5 Experimental configuration for semi-supervised APS failure detection model evaluation

	Training	Validation	Testing
Experiment 1	20% of HVs	80% of HVs	All HVs and AVs
Experiment 2	40% of HVs	60% of HVs	All HVs and AVs
Experiment 3	60% of HVs	40% of HVs	All HVs and AVs
Experiment 4	80% of HVs	20% of HVs	All HVs and AVs

The domain knowledge-based methods (HEA, HEA+, and LLM-based agentic analysis) operate unsupervised, requiring no data division as they rely entirely on expert-defined thresholds and pattern recognition. For the supervised EBM method, 5-fold

cross-validation is employed to ensure robust performance estimation, the procedure for fold generation and evaluation is detailed in Appendix A.

Given the substantial class imbalance in the APS dataset, performance evaluation employs precision, recall, F_1 score, accuracy, and AUC. These metrics are comprehensively defined in Appendix A. Throughout this analysis, anomalous vehicles are treated as the positive class, while healthy vehicles represent the negative class.

6.2.2 Baseline Method Results

6.2.2.1 Human Expert Analysis (HEA) Results

Following the methodology described in Section 5.2, human experts performed unsupervised pattern analysis on the three core indicators (duty cycle, compressor on/off count, and minimum pressure) across all vehicles. Each indicator receives a score of 0 (no anomaly), 1 (potential anomaly), or 2 (clear anomaly), resulting in a maximum of 6 flags per vehicle. Figure 6.9(a) illustrates the distribution of anomaly flags, demonstrating clear separation between healthy and anomalous vehicles. The HEA

Figure 6.9 HEA performance analysis: (a) distribution of anomaly flags by vehicle class, showing clear separation between healthy and anomalous vehicles; (b) precision-recall trade-off across different flag thresholds, with optimal F1 performance at 3 flags.

method achieves robust detection capabilities, with 90% of failed vehicles receiving at least one anomaly flag. Figure 6.9(b) shows the precision-recall trade-off across different threshold settings. At the 4-flag threshold, HEA attains perfect precision

(1.00) but with reduced recall. The optimal F1 score of 0.63 occurs at the 3-flag threshold, balancing precision (0.76) and recall (0.53) with 86.4% overall accuracy.

Enhanced Human Expert Analysis (HEA+) Results:

The enhanced HEA+ method incorporates brake usage patterns to reduce false positives in duty cycle assessments. The scoring system employs weighted flags: duty cycle (weight 1.0), compressor on/off count (weight 0.5), and minimum pressure (weight 0.5), resulting in a maximum of 4 flags per vehicle. This weighting reflects the relative importance and reliability of each indicator when contextualized with brake usage patterns.

Figure 6.10(a) shows the refined flag distribution, demonstrating improved discrimination between vehicle classes compared to the baseline HEA method. The precision-recall analysis in Figure 6.10(b) reveals enhanced performance characteristics, with the optimal threshold at 1.5 flags achieving precision of 0.79, recall of 0.63, and F1 score of 0.70. This represents a significant improvement over the baseline HEA method, with enhanced precision while maintaining reasonable recall.

Table 6.6 Comparative performance of the baseline HEA and the enhanced HEA$^+$ models at their respective optimal flag-count thresholds.

Method	Threshold	TN	FN	FP	TP	Precision	Recall	F1	Accuracy
HEA	≥ 3 flags	105	14	5	16	0.76	0.53	0.63	86.4%
HEA$^+$	≥ 1.5 flags	105	11	5	19	0.79	0.63	0.70	88.6%

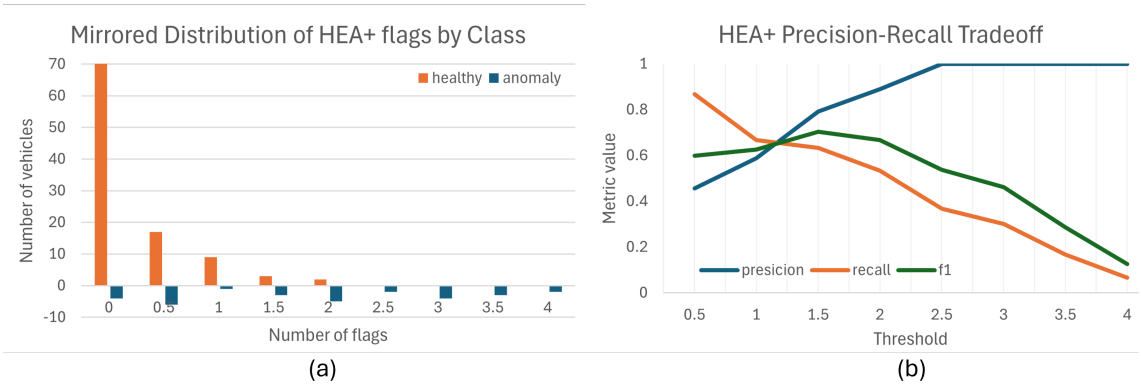

Figure 6.10 HEA+ performance analysis: (a) distribution of weighted anomaly flags incorporating brake usage context, showing improved class separation; (b) precision-recall trade-off demonstrating enhanced performance over baseline HEA, with optimal F1 score of 0.70 at 1.5 flags threshold.

6.2.2.2 LSTM Autoencoder Results

The LSTM autoencoder's performance was evaluated across four experimental configurations with varying proportions of HV training data. Figure 6.11 presents the averaged learning curves, demonstrating consistent convergence across all experiments. Even with minimal training data (20% HVs), the model achieves reconstruction errors below 0.015, with faster convergence observed as training data increases. Window length analysis reveals that 20-minute windows outperform 10-minute alternatives, achieving the highest F1 score of 0.75 with 80% HV training data (Figure 6.12). This configuration represents the optimal balance between data requirements and detection performance.

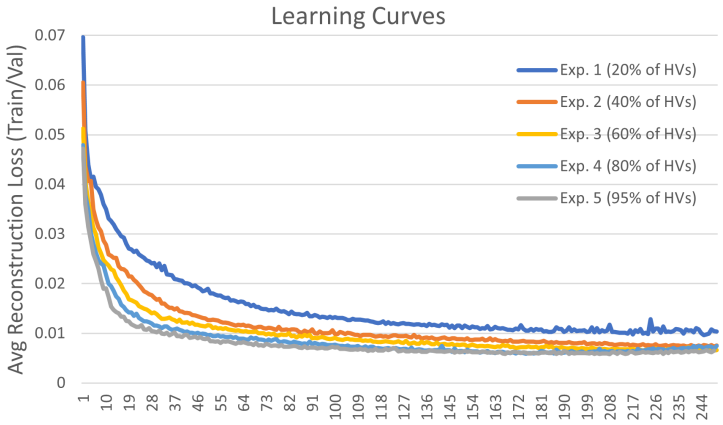

Figure 6.11 LSTM autoencoder learning curves (Mumcuoglu et al., 2024a).

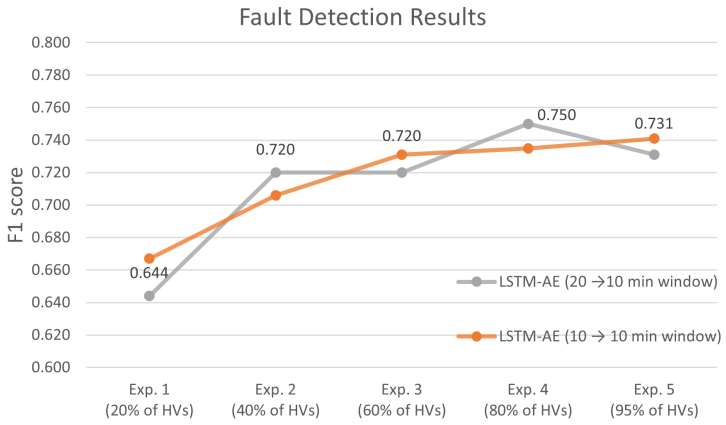

Figure 6.12 Impact of window length on LSTM-AE performance (Mumcuoglu et al., 2024a).

Table 6.7 summarizes the quantitative performance across all experimental configurations. The model exhibits progressive improvement with increased training data, achieving perfect precision (1.00) and optimal F1 score (0.75) at 80% HV utilization. Beyond this point, additional training data yields diminishing returns, suggesting an optimal data sufficiency threshold.

Table 6.7 LSTM autoencoder performance across different training data proportions using 20-minute windows.

Model	Training Data	Precision	Recall	F1	Accuracy
LSTM-AE	20% of HVs	0.66	0.63	0.64	85.0%
	40% of HVs	0.90	0.60	0.72	90.0%
	60% of HVs	0.90	0.60	0.72	90.0%
	80% of HVs*	**1.00**	**0.60**	**0.75**	**91.4%**
	95% of HVs	0.86	0.63	0.73	90.0%

The optimal model configuration (80% HV training) correctly identified all healthy vehicles while detecting 18 of 30 anomalous vehicles, as illustrated in the confusion matrix (Figure 6.13). This performance demonstrates strong reliability in healthy vehicle classification with reasonable anomaly detection capabilities.

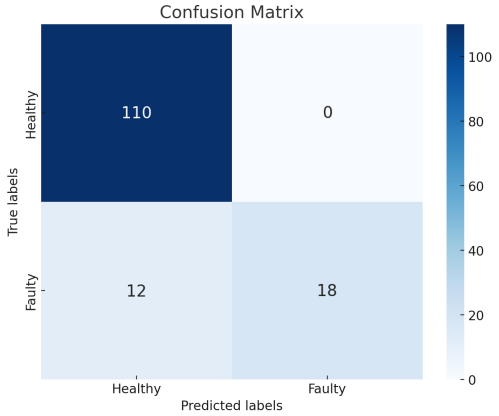

Figure 6.13 Confusion matrix for the optimal LSTM autoencoder configuration (80% HV training data), showing perfect healthy vehicle classification and 60% anomaly detection rate (Mumcuoglu et al., 2024a).

Figure 6.14 provides interpretability insights through reconstruction error analysis of representative driving sections. The model accurately reconstructs signals from healthy vehicles while exhibiting significant reconstruction errors for anomalous sections. The highlighted regions of poor reconstruction correspond to abnormal patterns in duty cycle, compressor switching behavior, and pressure dynamics, providing direct indication of potential failure modes.

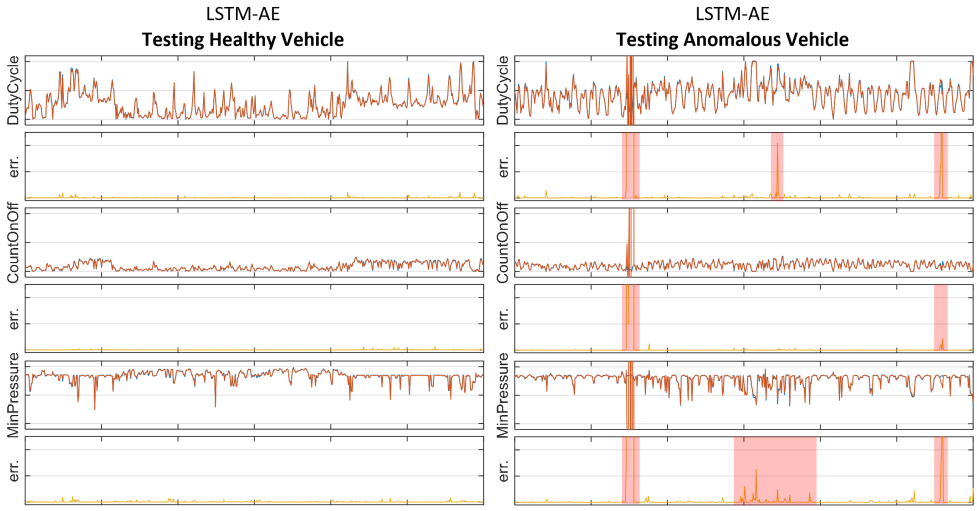

Figure 6.14 Reconstruction error analysis comparing healthy and anomalous driving sections. Poor reconstruction regions (highlighted) correspond to abnormal APS behavior patterns, enabling failure mode identification.

6.2.3 Explainable AI Module Results

6.2.3.1 EBM Performance

The EBM model was evaluated using the stratified five-fold cross-validation protocol detailed in Appendix A. Following the methodology established in Chapter 5, vehicle-level anomaly scores were computed as the median of observation-level classification probabilities. The optimal classification threshold was determined through grid search optimization to maximize F1 score performance.

The EBM achieved robust classification performance across all evaluation metrics. With an optimal threshold of 0.31, the model attained a precision of 0.80, recall of 0.80, and F1 score of 0.80, demonstrating balanced performance between false positive and false negative rates. The overall classification accuracy reached 91.4%, indicating strong discriminative capability across both healthy and anomalous vehicle categories.

The AUC of 0.88 demonstrates consistent performance across various threshold settings, indicating robust classification boundaries between vehicle classes. Figure 6.15

presents the ROC curve alongside threshold-dependent performance metrics, illustrating the model's stability across different operating points.

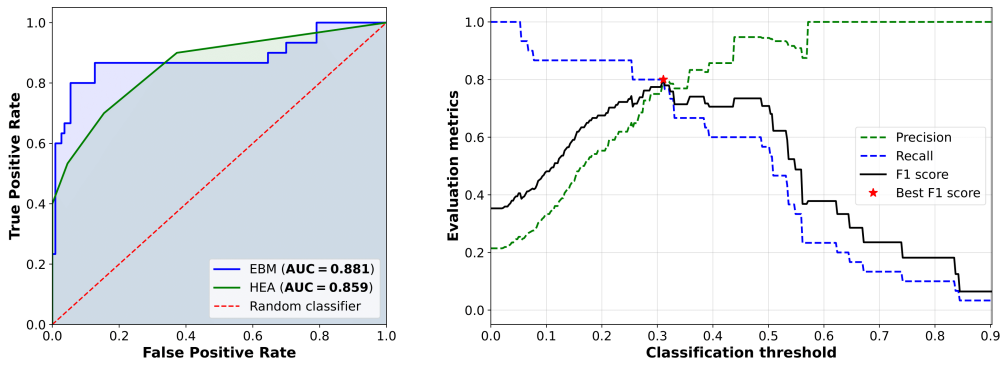

Figure 6.15 EBM performance characteristics: ROC curve (left) demonstrating AUC of 0.88, and threshold-dependent evolution of precision, recall, and F1 score (right) with optimal performance at threshold 0.31 (Farea et al., 2025).

The model successfully identified 24 of 30 anomalous vehicles while correctly classifying 104 of 110 healthy vehicles. This performance represents a balanced approach to the safety-critical nature of APS failure detection, where both missed failures and false alarms carry significant operational consequences. The consistent performance across cross-validation folds indicates model robustness and generalizability to unseen vehicle data.

Table 6.8 summarizes the quantitative performance metrics, demonstrating the EBM's effectiveness as a supervised learning approach for APS failure detection when sufficient labeled training data is available.

Table 6.8 EBM classification performance using five-fold cross-validation.

Model	Threshold	Precision	Recall	F1 Score	Accuracy	AUC
EBM	0.31	0.80	0.80	0.80	91.4%	0.88

6.2.3.2 EBM Model Interpretability Analysis

The EBM framework provides both global feature importance rankings and local decision explanations for individual vehicle classifications. The global feature importance analysis reveals that AC_on/off_count emerges as the most influential predictor, followed by BrakePedalPos_mean, DutyCycle, P2_min, and P3_min.

This ranking aligns closely with domain expertise, where air compressor cycling frequency and duty cycle patterns serve as primary indicators of APS degradation, while brake usage context provides essential interpretation for distinguishing normal operational stress from anomalous behavior. Figure 6.16 demonstrates the mean absolute contribution scores for all model features, confirming the significance of these core APS indicators in the classification decision process.

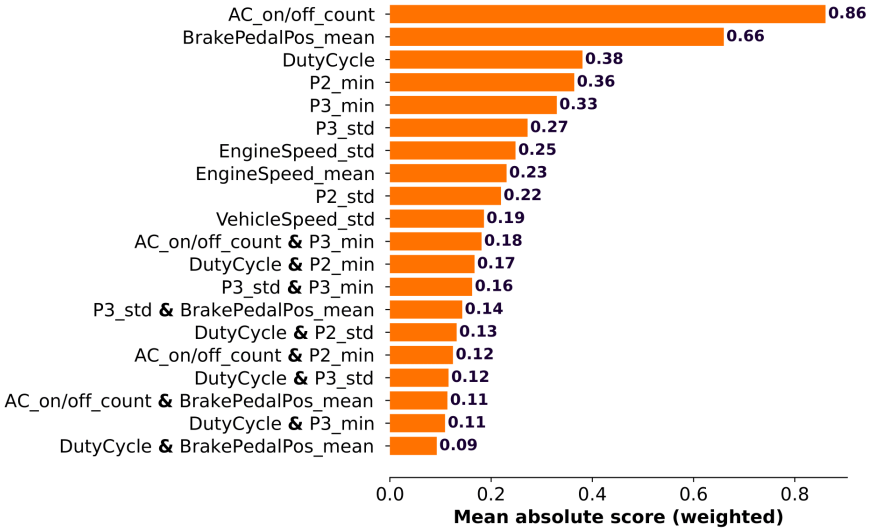

Figure 6.16 EBM global feature importance ranking showing the contribution of input features and their interactions to classification decisions (Farea et al., 2025).

Local explanation analysis provides insight into individual classification decisions through representative cases. Figure 6.17 illustrates correctly classified samples where Sample 1 (true negative) demonstrates healthy classification driven by low AC_on/off_count and DutyCycle values with normal pressure readings. Sample 2 (true positive) shows faulty classification based on elevated compressor cycling and duty cycle patterns, with time series analysis revealing progressive deterioration in the final month before failure, indicating extended compressor operation as a key failure precursor.

Figure 6.17 EBM local explanations for correctly classified vehicles: (top) true negative sample showing healthy operational patterns, (bottom) true positive sample demonstrating progressive APS degradation indicators.

Misclassification analysis reveals the model's limitations and edge cases. Figure 6.18 presents Sample 3 (false negative), where a failed vehicle was misclassified as healthy due to normal compressor behavior and pressure values during the monitoring period, suggesting either subtle failure modes not captured by key indicators or potential preventive replacement scenarios. Sample 4 (false positive) shows a healthy vehicle misclassified as faulty due to elevated compressor cycling and reduced minimum pressures, highlighting the challenge of distinguishing temporary operational stress from genuine system degradation.

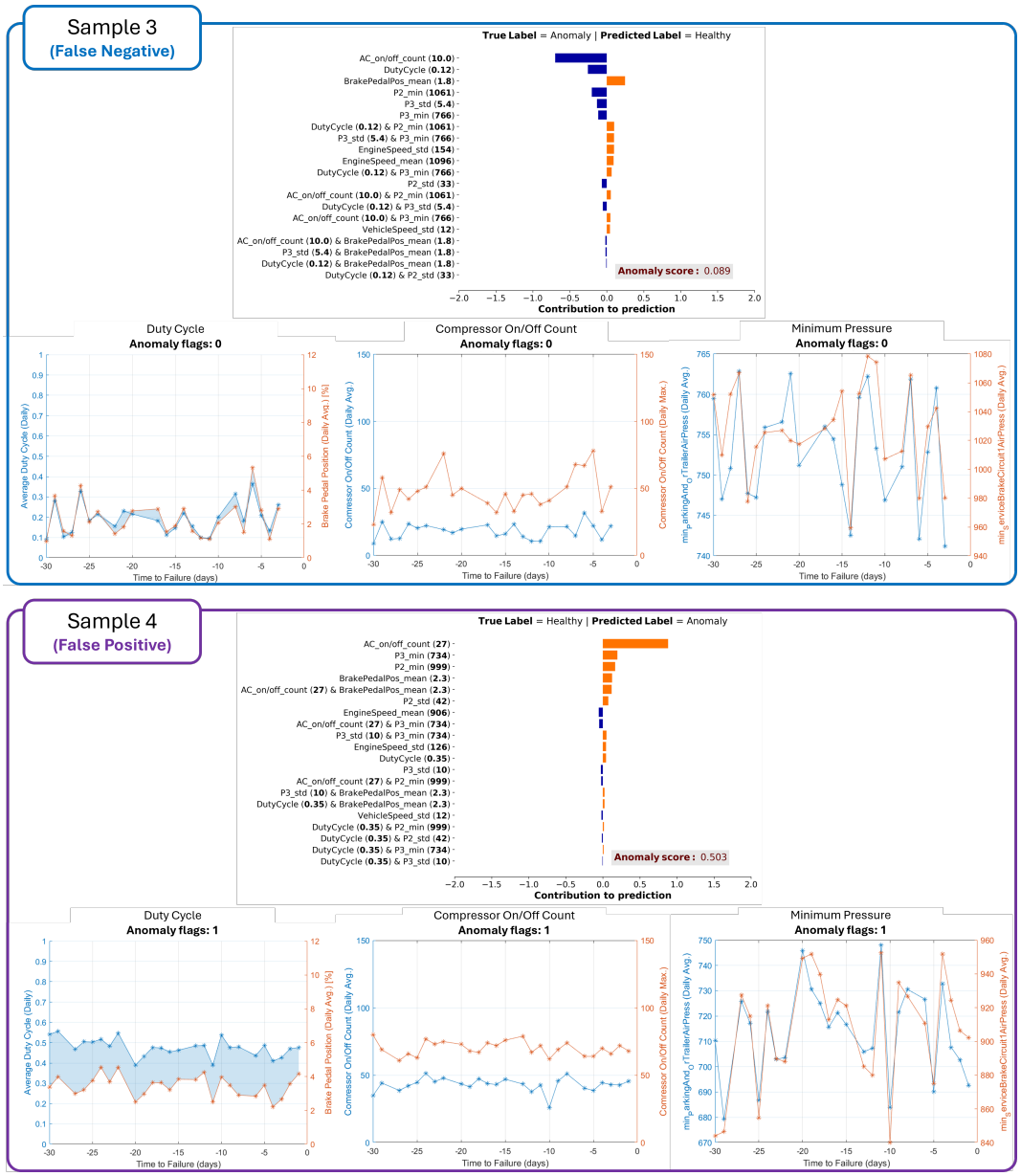

Figure 6.18 EBM local explanations for misclassified vehicles: (top) false negative showing normal patterns despite actual failure, (bottom) false positive indicating temporary operational anomalies in healthy vehicle.

6.2.3.3 Results of the LLM-based Agentic Framework

The LLM-based agentic framework, detailed comprehensively in Chapter 5, provides interpretable anomaly detection through multi-agent pattern analysis. In its best-performing run (Table 6.9), the framework achieved a precision of 0.82, recall of 0.60, F1 score of 0.69, and an overall accuracy of 89%. Although the quantitative performance is competitive, the method's primary strength lies in its exceptional

interpretability and the detailed natural-language explanations generated for each detected anomaly.

Performance evaluations across multiple independent runs revealed variability, with accuracy ranging from 0.86 to 0.89 and F1 scores between 0.62 and 0.69 (Table 6.10). This variability underscores the inherent stochasticity in LLM-driven decision-making, reflecting a trade-off between reproducibility and the nuanced reasoning capability enabled by the agentic approach.

In the optimal run, the framework successfully identified 18 out of 30 anomalous vehicles and accurately classified 106 of 110 healthy vehicles (Table 6.9). While recall was somewhat lower than traditional anomaly detection methods, this limitation was effectively offset by the framework's rich contextual explanations. These detailed justifications allow domain experts to comprehend and verify the reasoning behind each anomaly classification, thus enhancing practical applicability and expert trust.

Table 6.9 LLM-based agentic framework: best-run confusion matrix and performance metrics

TN	**FN**	**FP**	**TP**	**Precision**	**Recall**	**F1**	**Accuracy**
106	12	4	18	0.82	0.60	0.69	0.89

Table 6.10 Performance variability across five independent runs

Metric	**Min**	**Max**	**Mean ± SD**	**Range**
Accuracy	0.86	0.89	0.88 ± 0.01	0.03
F1 score	0.62	0.69	0.66 ± 0.03	0.07

The framework thus serves as a robust and interpretable alternative to traditional human-expert-based anomaly detection approaches. Developed through domain knowledge transfer and structured prompt engineering, the agentic framework demonstrates performance comparable to the HEA baseline method in terms of both F1 score and accuracy. These results underscore the effectiveness of systematically encoding domain knowledge within an LLM-based anomaly detection solution.

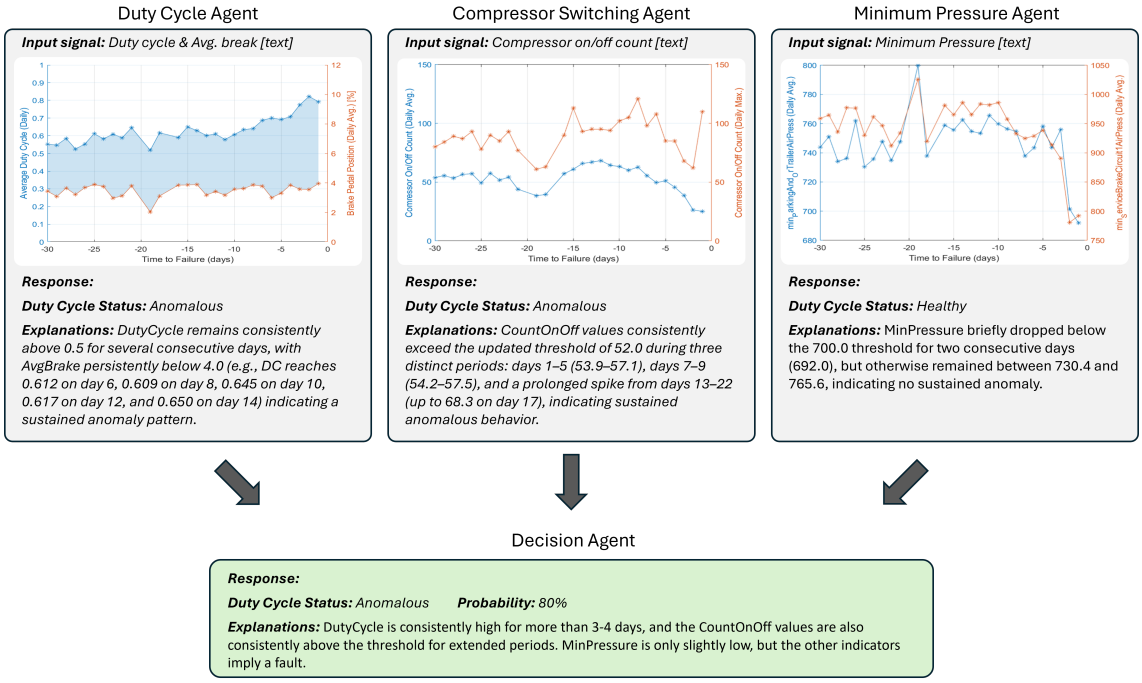

Figure 6.19 Output of the LLM-based agentic framework demonstrating multi-agent analysis of an anomalous vehicle. Each specialized agent provides detailed natural-language explanations of its APS indicator assessment, with the decision agent integrating these insights into a comprehensive anomaly classification.

Figure 6.19 further illustrates the framework's explanatory capabilities through a representative analysis of an anomalous vehicle. Here, each specialized agent contributes detailed, context-aware reasoning for its assigned APS indicator. Specifically, the duty cycle agent identifies sustained high readings surpassing operational thresholds, the compressor switching agent detects frequent and excessive cycling patterns, and the minimum pressure agent contextualizes pressure fluctuations relative to vehicle operation. The decision agent synthesizes these individual analyses into an integrative and interpretable anomaly determination.

Ultimately, the interpretability afforded by this agentic approach represents its most significant advantage, providing intuitive, human-readable justifications that align closely with expert reasoning. Unlike conventional ML models, which typically require separate explanatory techniques, the LLM-based system inherently generates clear and contextual explanations. This feature significantly enhances the ease of integration into existing maintenance workflows and supports expert validation and adoption in practical anomaly detection scenarios.

6.2.4 Integrated Methods and Comprehensive Analysis

Table 6.11 summarizes the optimal performance of each learning paradigm for APS failure detection. The supervised EBM achieved the most balanced results, with an F1 score of 0.80, effectively balancing precision and recall (both at 0.80). Conversely, the semi-supervised LSTM-AE model demonstrated exceptional precision (1.00) but at the cost of lower recall (0.60). Among the unsupervised methods, the LLM agentic approach provided competitive precision (0.82) with a recall comparable to expert-based methods.

Table 6.11 Summary of APS failure–detection performance by learning paradigm

Paradigm	Method	Optimal setting*	Precision	Recall	F1	Accuracy
Unsupervised**	HEA	≥ 3 flags	0.76	0.53	0.63	86.4%
	HEA$^+$	≥ 1.5 flags	0.79	0.63	0.70	88.6%
	LLM agentic	3+1 agents (Gemini 2.0 Flash)	0.82	0.60	0.69	89.0%
Semi-supervised	LSTM-AE	80 % HV train, 20-min window	1.00	0.60	0.75	91.4%
Supervised	EBM	threshold 0.31	0.80	0.80	0.80	91.4%

* Configuration that maximises F_1 score for each method.

** Unsupervised methods rely solely on expert logic and require no labelled training data.

Hybrid Approach Performance

To evaluate the effectiveness of integrating LSTM-AE with interpretable methods, we systematically assessed four configurations across varying proportions of HVs in the training set. Figure 6.20 illustrates F1-score progression for the standalone LSTM-AE and hybrid methods.

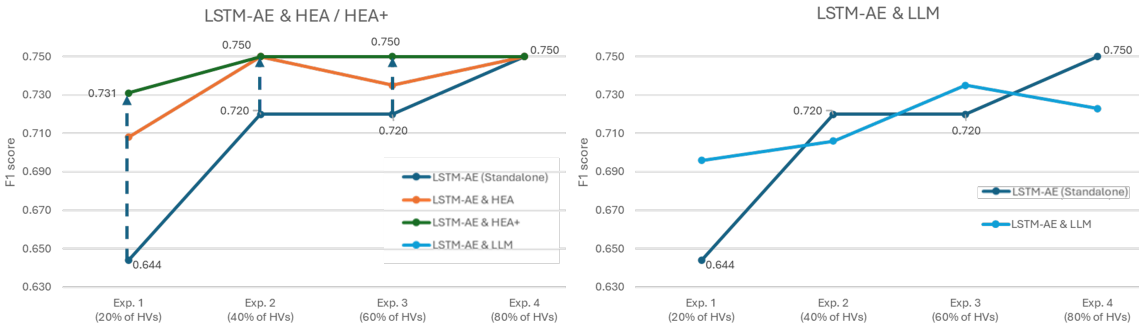

Figure 6.20 Performance of hybrid LSTM-AE approaches across different experiments. Left: LSTM-AE with HEA/HEA+. Right: LSTM-AE with LLM integration.

The results reveal that integrating domain-expert analysis substantially enhances

baseline LSTM-AE performance. The standalone LSTM-AE exhibited notable sensitivity to training data composition, improving from an F1 of 0.644 (20% HVs) to 0.750 (80% HVs). In contrast, the LSTM-AE combined with HEA consistently improved anomaly detection performance, delivering stable F1 scores ranging from 0.731 to 0.750 across experiments. Importantly, the enhanced expert analysis (HEA+) approach achieved its peak performance (F1 = 0.750) even with a limited amount of healthy vehicle data (40% HVs), underscoring the value of incorporating brake usage patterns for context-aware anomaly detection.

Meanwhile, the LSTM-AE + LLM hybrid demonstrated competitive results, achieving F1 scores between 0.694 and 0.736. Although the integration of LLM agents did not outperform HEA+ in terms of absolute accuracy, it notably offered significant improvements in interpretability, explainability, and scalability. Unlike HEA/HEA+ methods that rely heavily on manual expert calibration, the LLM-based approach automates diagnostic reasoning, thus simplifying deployment across diverse vehicle fleets without extensive domain-expert involvement.

These insights underline the trade-offs between incremental performance gains and operational scalability, positioning the LLM hybrid as particularly valuable in large-scale predictive maintenance scenarios where expert resources are constrained.

Correlation Analysis of Interpretable Models

To quantitatively assess consistency among the interpretable models (HEA+, EBM, and LLM-based agents), Pearson correlation coefficients were computed between their respective APS indicators. Table 6.12 summarizes key correlations ($|\rho| \geq 0.60$), demonstrating robust alignment across domain knowledge-driven and data-driven approaches.

Duty cycle exhibited strong correlation ($\rho = 0.72$ with LLM, $\rho = 0.76$ with EBM), confirming its critical role across all interpretability paradigms. Compressor on/off count similarly showed significant correlations ($\rho = 0.75$ with LLM, $\rho = 0.70$ with EBM), reflecting the effective transferability of frequency-based expert indicators into learned frameworks. The minimum pressure (P3) indicator demonstrated moderately strong yet consistent correlations ($\rho = 0.69$ with LLM, $\rho = 0.63$ with EBM), suggesting nuanced variations in how each model interprets dynamic sensor behavior.

Importantly, the global expert-defined indicator (HEA+ total flags) strongly correlated with both LLM predictions ($\rho = 0.77$) and EBM probabilities ($\rho = 0.79$), providing empirical validation that the data-driven models effectively internalized composite expert logic. These findings highlight strong intuitive consistency, re-

inforcing confidence that all interpretable models align closely with maintenance engineers' domain knowledge and operational priorities.

Table 6.12 Key Pearson correlations (ρ) between HEA+ features and their counterparts in the LLM and EBM models. Values above 0.70 are set in bold.

Feature	ρ(HEA+, LLM Agents)	ρ(HEA+, EBM)
Duty cycle	**0.72**	**0.76**
Compressor on/off count	**0.75**	0.70
Minimum pressure (P3)	0.69	0.63
Global indicator: ρ(HEA+ Total Flags, LLM Predict.) = **0.77**, ρ(HEA+ Total Flags, EBM Prob.) = **0.79**		

Expert analysis and interpretability are critical elements of predictive maintenance applications, essential not only for reliable detection but also for accurate root cause identification and minimizing false positives. As demonstrated by strong correlations among HEA+, EBM, and LLM-based indicators, interpretable models effectively internalize domain-expert logic, highlighting their potential to serve as scalable alternatives to manual expert analyses. In particular, EBM and LLM-based approaches exhibit consistent behavior with expert-defined methods, positioning them as valuable tools that enhance scalability without compromising interpretability or diagnostic quality.

Furthermore, employing semi-supervised models like LSTM autoencoders brings additional advantages by addressing the intrinsic challenges posed by the heterogeneous nature of APS faults. Unlike supervised models, which may struggle with the diverse and unpredictable range of potential fault conditions, semi-supervised learning frameworks adeptly adapt to unknown fault types by modeling normal system behavior exclusively from healthy data. This approach not only circumvents issues related to class imbalance—common in fault detection scenarios—but also provides robust baseline anomaly scores that effectively differentiate between normal and anomalous system conditions.

Overall, this integrated approach—leveraging interpretable models combined with semi-supervised anomaly detection—provides a comprehensive predictive maintenance solution. It balances interpretability, adaptability, and scalability, paving the way for effective real-world implementation in large-scale vehicle fleets while ensuring consistent reliability in APS fault detection and prevention.

7. CONCLUSION

This thesis has developed robust anomaly detection frameworks designed specifically for HDV applications, significantly advancing predictive maintenance strategies. Two primary vehicle system challenges were addressed: excessive fuel consumption and APS failures.

For FC anomalies, a novel quartile-based labeling method, sensitive to vehicle load and road slope, was introduced. The developed Bagged Decision Tree models effectively classify FC anomalies with high accuracy (up to 92.2%) across diverse geographical contexts. The complementary interactive dashboard provides fleet managers with actionable insights, enabling proactive interventions.

In APS failure detection, semi-supervised LSTM Autoencoders demonstrated strong predictive capability (F1: 0.75) with perfect precision in capturing subtle temporal anomalies. Integrating human expert analysis significantly reduced false positives and enhanced overall model performance. Additionally, the EBM achieved an optimal balance between accuracy (91.4%, F1: 0.80) and interpretability, augmented by a LLM-based diagnostic framework offering expert-level interpretability.

A notable contribution of this research is the emphasis on explainability, merging advanced machine learning with domain expertise. These explainable AI-driven frameworks substantially enhance trust, interpretability, and practical applicability, crucial for fleet management and maintenance operations.

Future research can extend these methods toward real-time detection, multi-fault classification, and integration into automated, closed-loop predictive maintenance systems, paving the way for more resilient and sustainable fleet management practices.

BIBLIOGRAPHY

Ahmad Khan, M., Khan, M., Dawood, H., Dawood, H., & Daud, A. (2024). Secure Explainable-AI approach for brake faults prediction in heavy transport. *IEEE Access*, *12*, 114940–114950.

Al-Hayek, M. et al. (2025). Evaluating gemini in an arena for learning. *arXiv preprint arXiv:2505.24477*.

Balestri, R. (2025). Gender and content bias in large language models: A case study on google gemini 2.0 flash experimental. *arXiv preprint arXiv:2503.16534*.

Barbado, A. & Corcho, Ó. (2022). Interpretable machine learning models for predicting and explaining vehicle fuel consumption anomalies. *Engineering Applications of Artificial Intelligence*, *115*, 105222.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and Regression Trees*. Boca Raton, FL: CRC Press.

Charroud, A., El Moutaouakil, K., Palade, V., & Yahyaouy, A. (2023). XDLL: Explained deep learning LiDAR-based localization and mapping method for self-driving vehicles. *Electronics*, *12*(3), 567.

Das, S., Agarwal, N., Venugopal, D., Sheldon, F. T., & Shiva, S. (2020). Taxonomy and survey of interpretable machine learning method. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, (pp. 670–677). IEEE.

Davari, N., Pashami, S., Veloso, B., Nowaczyk, S., Fan, Y., Pereira, P. M., Ribeiro, R. P., & Gama, J. (2022). A fault detection framework based on lstm autoencoder: A case study for volvo bus data set. In *International Symposium on Intelligent Data Analysis*, (pp. 39–52). Springer.

Elhafsi, A., Sinha, R., Agia, C., Schmerling, E., Nesnas, I. A. D., & Pavone, M. (2023). Semantic anomaly detection with large language models. *Autonomous Robots*, *47*, 1035–1055.

European Commission (2024). Questions and answers: Revised co emission standards for heavy-duty vehicles. [Accessed: 26-May-2025].

European Environment Agency (2022). Reducing greenhouse gas emissions from heavy-duty vehicles in europe. [Accessed: 26-May-2025].

European Parliament (2018). Co2 emission standards for heavy-duty vehicles. [Accessed: 26-May-2025].

Fan, Y., Hilber, P., & Rögnvaldsson, T. (2015). Incorporating expert knowledge into a self-organized approach for predicting compressor faults in a city bus fleet. In *Proceedings of the 28th International Workshop on Machine Learning*, volume 278, (pp. 58–67).

Farea, S. M., Mumcuoglu, M. E., & Unel, M. (2025). An explainable AI approach for detecting failures in air pressure systems. *Engineering Failure Analysis*, *173*, 109441.

Farea, S. M., Mumcuoglu, M. E., Unel, M., Mise, S., Unsal, S., Yilmaz, M., & Koprubasi, K. (2023). Towards driving-independent prediction of fuel consumption in heavy-duty trucks. In *Proceedings of the 2023 AEIT International Conference on Electrical and Electronic Technologies for Automotive (AEIT AUTOMOTIVE)*, Modena, Italy. IEEE.

Geglio, A., Hedayati, E., Tascillo, M., Anderson, D., Barker, J., & Havens, T. C. (2022). Deep convolutional autoencoder for assessment of drive-cycle anomalies in connected vehicle sensor data. In *2022 IEEE Symposium Series on Computational Intelligence (SSCI)*, (pp. 743–749).

Gong, J., Shang, J., Li, L., Zhang, C., He, J., & Ma, J. (2021). A comparative study on fuel consumption prediction methods of heavy-duty diesel trucks considering 21 influencing factors. *Energies*, *14*(23), 8106.

Google DeepMind (2024). Google introduces gemini 2.0: A new ai model for the agentic era. `https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/`. Accessed 23 Jun. 2025.

Guo, X., Zhang, Q., Jiang, J., Peng, M., Hao, Y., Yang, L., & Zhu, M. (2024). R2T-LLM: Towards responsible and reliable traffic flow prediction with large language models. *arXiv preprint arXiv:2404.02937*.

Hochreiter, S. & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Jafari, S. & Byun, Y. C. (2024). Accurate remaining useful life estimation of lithium-ion batteries in electric vehicles based on a measurable feature-based approach with explainable AI. *The Journal of Supercomputing*, *80*(4), 4707–4732.

Jung, D. (2020). Data-driven open-set fault classification of residual data using bayesian filtering. *IEEE Transactions on Control Systems Technology*, *28*(5), 2045–2052.

Jung, D. H. & Sundström, C. (2019). A combined data-driven and model-based residual selection algorithm for fault detection and isolation. *IEEE Transactions on Control Systems Technology*, *27*(2), 616–630.

Kang, J., Kim, C.-S., Kang, J. W., & Gwak, J. (2021). Anomaly detection of the brake operating unit on metro vehicles using a one-class lstm autoencoder. *Applied Sciences*, *11*(19), 9290.

Khalid Fahmi, A.-T. W., Reza Kashyzadeh, K., & Ghorbani, S. (2024). Fault detection in the gas turbine of the kirkuk power plant: An anomaly detection approach using dlstm-autoencoder. *Engineering Failure Analysis*, *160*, 108213.

Killeen, P., Ding, B., Kiringa, I., & Yeap, T. (2019). Iot-based predictive maintenance for fleet management. *Procedia Computer Science*, *151*, 607–613.

Kolekar, S., Gite, S., Pradhan, B., & Alamri, A. (2022). Explainable AI in scene understanding for autonomous vehicles in unstructured traffic environments on Indian roads using the inception U-Net model with Grad-CAM visualization. *sensors*, *22*(24), 9677.

Konstantinou, C., Fafoutellis, P., Mantouka, E. G., Chalkiadakis, C., Fortsakis, P., & Vlahogianni, E. I. (2023). Effects of driving behavior on fuel consumption with explainable gradient boosting decision trees. In *2023 8th International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*, (pp. 1–6). IEEE.

Kramer, M. A. (1991). Nonlinear principal component analysis using autoassociative neural networks. *AIChE journal*, *37*(2), 233–243.

Li, M., Wang, Y., Sun, H., Cui, Z., Huang, Y., & Chen, H. (2023). Explaining a machine-learning lane change model with maximum entropy Shapley values. *IEEE Transactions on Intelligent Vehicles*, *8*(6), 3620–3628.

Lin, C., Cheng, Y., Wang, X., Yuan, J., & Wang, G. (2023). Transformer-based dual-channel self-attention for uuv autonomous collision avoidance. *IEEE Trans-*

actions on Intelligent Vehicles, *8*(3), 2319–2331.

Loh, W.-Y. (2002). Regression trees with unbiased variable selection and interaction detection. *Statistica Sinica*, *12*(2), 361–386.

Mao, J., Qian, Y., Ye, J., Zhao, H., & Wang, Y. (2023). GPT-Driver: Learning to drive with gpt. *arXiv preprint arXiv:2312.01835*.

Masello, L., Castignani, G., Sheehan, B., Guillen, M., & Murphy, F. (2023). Using contextual data to predict risky driving events: A novel methodology from explainable artificial intelligence. *Accident Analysis & Prevention*, *184*, 106997.

Min, H., Fang, Y., Wu, X., Lei, X., Chen, S., Teixeira, R., Zhu, B., Zhao, X., & Xu, Z. (2023). A fault diagnosis framework for autonomous vehicles with sensor self-diagnosis. *Expert Systems with Applications*, *224*, 120002.

Mohanty, P. K. & Roy, D. S. (2023). Analyzing the factors influencing energy consumption at electric vehicle charging stations with Shapley additive explanations. In *2023 International Conference on Microwave, Optical, and Communication Engineering (ICMOCE)*, (pp. 1–5). IEEE.

Mumcuoglu, M. E., Farea, S. M., Unel, M., Mise, S., Unsal, S., Cevik, E., Yilmaz, M., & Koprubasi, K. (2024a). Air Pressure System Failures Detection Using LSTM-Autoencoder. In *2024 IEEE International Workshop on Metrology for Automotive (MetroAutomotive)*, (pp. 82–87). IEEE.

Mumcuoglu, M. E., Farea, S. M., Unel, M., Mise, S., Unsal, S., Cevik, E., Yilmaz, M., & Koprubasi, K. (2024b). Detecting APS failures using LSTM-AE and anomaly transformer enhanced with human expert analysis. *Engineering Failure Analysis*, *165*, 108811.

Mumcuoglu, M. E., Farea, S. M., Unel, M., Mise, S., Unsal, S., Yilmaz, M., & Koprubasi, K. (2023). Fuel consumption classification for heavy-duty vehicles: A novel approach to identifying driver behavior and system anomalies. In *Proceedings of the 2023 AEIT International Conference on Electrical and Electronic Technologies for Automotive (AEIT AUTOMOTIVE)*, Modena, Italy. IEEE.

Nori, H., Jenkins, S., Koch, P., & Caruana, R. (2019). InterpretML: A unified framework for machine learning interpretability. *arXiv preprint arXiv:1909.09223*.

Nowaczyk, S., Prytz, R., Rögnvaldsson, T., & Byttner, S. (2013). Towards a machine learning algorithm for predicting truck compressor failures using logged vehicle data. In *12th Scandinavian Conference on Artificial Intelligence, Aalborg, Denmark, November 20–22, 2013*, (pp. 205–214). IOS Press.

Peng, M., Guo, X., Chen, X., Zhu, M., Chen, K., Hao, Y., Yang, L., Wang, X., & Wang, Y. (2024). LC-LLM: Explainable lane-change intention and trajectory predictions with large language models. *arXiv preprint arXiv:2404.06527*.

Prytz, R., Nowaczyk, S., Rögnvaldsson, T., & Byttner, S. (2015). Predicting the need for vehicle compressor repairs using maintenance records and logged vehicle data. *Engineering Applications of Artificial Intelligence*, *41*, 139–150.

Ranjan, C. (2020). *Understanding deep learning: Application in rare event prediction*. Connaissance Publishing Atlanta, GA, USA.

Rengasamy, D., Jafari, M., Rothwell, B., Chen, X., & Figueredo, G. P. (2020). Deep learning with dynamically weighted loss function for sensor-based prognostics and health management. *Sensors*, *20*(3).

Routray, A., Rajaguru, A., & Singh, S. (2010). Data reduction and clustering techniques for fault detection and diagnosis in automotives. In *2010 IEEE*

International Conference on Automation Science and Engineering, (pp. 326–331).

Sang, J., Zhang, J., Guo, T., Zhou, D., Chen, M., & Tai, X. (2020). Detection of incipient faults in emu braking system based on data domain description and variable control limit. *Neurocomputing*, *383*, 348–358.

Saravanarajan, V. S., Chen, R.-C., Hsieh, C.-H., & Chen, L.-S. (2023). Improving semantic segmentation under hazy weather for autonomous vehicles using explainable artificial intelligence and adaptive dehazing approach. *IEEE Access*, *11*, 38194–38207.

Sima, C., Renz, K., Chitta, K., Chen, L., Zhang, H., Xie, C., Luo, P., Geiger, A., & Li, H. (2023). DriveLM: Driving with graph visual question answering. *arXiv preprint arXiv:2305.18203*.

Tagawa, T., Tadokoro, Y., & Aoki, Y. (2014). Structured denoising autoencoder for fault detection and analysis. In *Proceedings of the 2014 International Conference on Machine Learning*, (pp. 96–111).

Theissler, A. (2017). Detecting known and unknown faults in automotive systems using ensemble-based anomaly detection. *Knowledge-Based Systems*, *123*, 163–173.

Tuli, S., Casale, G., & Jennings, N. R. (2022). Tranad: deep transformer networks for anomaly detection in multivariate time series data. *Proceedings of the VLDB Endowment*, *15*(6), 1201–1214.

Wang, D., Ruan, P., Xu, D., Xie, W., Chen, X., & Li, H. (2023). Tranad: A deep transformer model for fault diagnosis of lithium batteries. In *2023 International Conference on Smart Electrical Grid and Renewable Energy (SEGRE)*, (pp. 133–139).

Wang, L., Zhang, X., Zeng, W., Liu, W., Yang, L., Li, J., & Liu, H. (2023). Global perception-based robust parking space detection using a low-cost camera. *IEEE Transactions on Intelligent Vehicles*, *8*(2), 1439–1448.

Wang, M., Pang, A., Kan, Y., Pun, M.-O., Chen, C.-S., & Huang, B. (2024). LLM-Assisted Light: Leveraging large language model capabilities for human-mimetic traffic signal control in complex urban environments. *arXiv preprint arXiv:2404.12249*.

Wang, T.-H., Maalouf, A., Xiao, W., Ban, Y., Amini, A., Rosman, G., Karaman, S., & Rus, D. (2023). Drive Anywhere: Generalizable end-to-end autonomous driving with multi-modal foundation models. *arXiv preprint arXiv:2311.08370*.

Wei, D., Gao, T., Jia, Z., Cai, C., Hou, C., Jia, P., Liu, F., Zhan, K., Fan, J., Zhao, Y., & Wang, Y. (2024). BEV-CLIP: Multi-modal bird's-eye-view retrieval for complex scenes in autonomous driving. *arXiv preprint arXiv:2403.14215*.

Wolf, P., Mrowca, A., Nguyen, T. V., Baker, B. R., & Günnemann, S. (2018). Pre-ignition detection using deep neural networks: A step towards data-driven automotive diagnostics.

Wong, P. K., Zhong, J., Yang, Z., & Vong, C. M. (2016). Sparse bayesian extreme learning committee machine for engine simultaneous fault diagnosis. *Neurocomputing*, *174*, 331–343.

Xu, J., Wu, H., Wang, J., & Long, M. (2021). Anomaly transformer: Time series anomaly detection with association discrepancy. In *International Conference on Learning Representations*.

Xu, X., Wang, H., Liang, Y., Yu, P. S., Zhao, Y., & Shu, K. (2025). Can multimodal llms perform time series anomaly detection? *arXiv preprint arXiv:2502.17812*.

Xu, Z., Zhang, Y., Xie, E., Zhao, Z., Guo, Y., Wong, K.-M. K., Li, Z., & Zhao, H. (2023). DriveGPT4: Interpretable end-to-end autonomous driving via large language model. *arXiv preprint arXiv:2310.09655*.

Yu, Z., Zhu, M., Chen, K., Chu, X., & Wang, X. (2023). Lf-net: A learning-based frenet planning approach for urban autonomous driving. *IEEE Transactions on Intelligent Vehicles*, 1–14.

Yuan, J., Sun, S., Omeiza, D., Zhao, B., Newman, P., Kunze, L., & Gadd, M. (2024). RAG-Driver: Generalisable driving explanations with retrieval-augmented in-context learning in multi-modal large language model. *arXiv preprint arXiv:2403.10008*.

Zhang, J., Wang, Y., Jiang, B., He, H., Huang, S., Wang, C., Zhang, Y., Han, X., Guo, D., He, G., & Ouyang, M. (2023). Realistic fault detection of li-ion battery via dynamical deep learning. *Nature Communications*, *14*(1), 5940.

Zhao, J., Feng, X., Wang, J., Lian, Y., Ouyang, M., & Burke, A. F. (2023). Battery fault diagnosis and failure prognosis for electric vehicles using spatio-temporal transformer networks. *Applied Energy*, *352*, 121949.

Zhong, J. H., Wong, P. K., & Yang, Z. X. (2018). Fault diagnosis of rotating machinery based on multiple probabilistic classifiers. *Mechanical Systems and Signal Processing*, *108*, 99–114.

Zhou, Y., Fu, C., Jiang, X., Yu, Q., & Liu, H. (2024). Who might encounter hard-braking while speeding? Analysis for regular speeders using low-frequency taxi trajectories on arterial roads and explainable AI. *Accident Analysis & Prevention*, *195*, 107382.

Zhou, Z. & Yu, R. (2024). Can llms understand time series anomalies? *arXiv preprint arXiv:2410.05440*.

APPENDIX A

Classification Performance Evaluation

This appendix summarises the evaluation metrics and validation procedures used throughout the experimental chapters. All metric formulas rely on the confusion-matrix counts of TP, FP, TN, and FN. In the datasets analysed here, the positive class corresponds to anomalous vehicles, while the negative class denotes healthy ones.

Precision

Precision indicates how many of the segments predicted as anomalous were indeed anomalous:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{A.1}$$

Recall

Recall (sensitivity) shows how many of the real anomalies the model detects:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{A.2}$$

$\mathbf{F_1}$ score

The F_1 score is the harmonic mean of precision and recall:

$$F_1 = \frac{2\,\text{Precision}\,\text{Recall}}{\text{Precision} + \text{Recall}} \tag{A.3}$$

Accuracy

Accuracy reports the proportion of all predictions that are correct:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{A.4}$$

Area Under the ROC Curve (AUC)

AUC summarises performance across all discrimination thresholds by integrating the TPR over the FPR:

$$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR})\, d\,\text{FPR} \tag{A.5}$$

K-fold cross-validation

K-fold cross-validation provides an estimate of model generalisation performance by partitioning the dataset and evaluating on held-out data.

Let $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ be the labelled dataset. It is partitioned into K disjoint folds $\{\mathcal{F}_1, \ldots, \mathcal{F}_K\}$ of approximately equal size, preserving the original class proportions within every fold.

For each $k \in \{1, \ldots, K\}$, a model $\mathcal{M}_k$ is trained on $K-1$ folds and evaluated on the held-out fold $\mathcal{F}_k$:

$$\mathcal{M}_k = \text{train}\Big(\mathcal{D} \setminus \mathcal{F}_k\Big) \tag{A.6}$$

$$s_k = \text{metric}\Big(\mathcal{M}_k, \mathcal{F}_k\Big) \tag{A.7}$$

where s_k is any evaluation metric (Precision, Recall, F_1, etc.) computed on fold $\mathcal{F}_k$.

The overall score for metric s is the arithmetic mean across folds:

$$\overline{s} = \frac{1}{K} \sum_{k=1}^{K} s_k \tag{A.8}$$

The variability is quantified by the sample standard deviation:

$$\sigma_s = \sqrt{\frac{1}{K-1} \sum_{k=1}^{K} \Big(s_k - \overline{s}\Big)^2} \tag{A.9}$$

APPENDIX B

LLM-based Agentic Framework Prompts

This appendix presents the complete prompt engineering specifications for the four-agent diagnostic framework described in Chapter 5. The prompts are designed to transfer expert knowledge into structured AI agents capable of analyzing APS failure patterns with interpretable reasoning.

Duty Cycle Agent Prompt

The Duty Cycle Agent analyzes duty cycle patterns in conjunction with average brake usage, implementing the HEA+ methodology to distinguish between operational demands and system anomalies.

Listing B.1 Duty Cycle Agent prompt

```
You are an AI specialized in analyzing DutyCycle (DC) and AvgBrake
    signals using pattern analysis.

**Context and Domain Knowledge**:
1. **DutyCycle (DC)** reflects the air compressor's workload and is a key
    indicator of system health.
- Reference threshold for DutyCycle: {dc_threshold} (Typical range of DC
    is 00.8)
- Values consistently above the threshold for 34 days indicate an
    anomaly, especially if AvgBrake is not similarly elevated.
2. AvgBrake (typical range 010; values 4 considered high) indicates
    average brake usage.
  If DC and AvgBrake both trend high, it's likely normal due to driving
      conditions (traffic, slopes, etc.).
```

```
3. Look for multi-day DC patterns that are high while AvgBrake remains
    low.

**Task**:
Given daily DutyCycle and AvgBrake data, identify:
- **No Anomaly** or **Anomaly** (34 consecutive days with
    DC>{dc_threshold} but not matching AvgBrake).
- Key Patterns: List 1-2 notable trends or violations with DC values and
    their occurrence times for further analysis.

**Output only JSON** in this format (no extra text):
{
  "Anomalies": ["description of anomalies, if any"],
  "IsAnomalous": true/false
}
```

Compressor Switching Agent Prompt

The Switching Pattern Agent examines compressor on/off count patterns to identify system instability and abnormal cycling behavior.

Listing B.2 Compressor Switching Agent prompt

```
You are an AI specialized in analyzing CountOnOff (CO) signals to detect
    compressor state changes using pattern analysis.

**Context and Domain Knowledge**:
1. **CountOnOff** represents how many times the compressor turns on/off
    in a day.
2. For your analysis, consider:
- Reference threshold for CountOnOff: {co_threshold} (Typical range of CO
    is 570).
- Sustained high values above the threshold over consecutive days (34 or
    more) may indicate system instability or frequent cycling.
3. One-day spikes might not be a fault if they are isolated.

**Task**:
Given daily CountOnOff data:
- Identify **No Anomaly** or **Anomaly** (34 days above {co_threshold}).
- Key Patterns: List 1-2 notable trends or violations with CO values and
    their occurrence times for further analysis.

**Output only JSON** in this format (no extra text):
```

```
{
  "Anomalies": ["description of anomalies, if any"],
  "IsAnomalous": true/false
}
```

Minimum Pressure Agent Prompt

The Pressure Agent monitors minimum pressure deviations to detect air leakage and pressure system degradation.

Listing B.3 Minimum Pressure Agent prompt

```
You are an AI specialized in analyzing the MinPressure (MP) signal using
    pattern analysis.

**Context and Domain Knowledge**:
1. **MinPressure** tracks the lowest observed pressure for the day.
2. For your analysis, consider:
- Reference threshold for MinPressure: {mp_threshold} (typical range of
    MP is 690-810).
- A sustained drop below the thresholds for 34 days might indicate
    leakage.
3. Single-day dips are less concerning if not repeated.

**Task**:
Given daily MinPressure data:
- Identify **No Anomaly** or **Anomaly** (34 days below ~{mp_threshold}).
- Key Patterns: List 1-2 notable trends or violations with MP values and
    their occurrence times for further analysis.

**Output only JSON** in this format (no extra text):
{
  "Anomalies": ["description of anomalies, if any"],
  "IsAnomalous": true/false
}
```

Decision Agent Prompt

The Decision Agent synthesizes analyses from the three signal-specific agents to render final diagnostic decisions with probability assessments.

Listing B.4 Decision Agent prompt

```
You are an AI that makes the final vehicle health classification based on
    analyses of three key indicators:
- DutyCycle & AvgBrake Analysis
- CountOnOff Analysis
- MinPressure Analysis

**Context and Domain Knowledge**:
1. Reference thresholds used in the analyses:
   - DutyCycle: {dc_threshold} (typical range 0-0.8)
   - CountOnOff: {co_threshold} (typical range 5-70)
   - MinPressure: {mp_threshold} (typical range 690-810)

2. Classification Guidelines:
   - Predict **Faulty** when strong failure patterns appear or failures
       appear across multiple indicators
   - The DutyCycle analysis is the PRIMARY indicator of system health
   - CountOnOff and MinPressure are SECONDARY indicators
   - CountOnOff alone or MinPressure alone DO NOT imply a fault!!!
   - Otherwise, predict **Healthy**

**Guidance for Probability**:
- If only DutyCycle is anomalous with clear patterns: 60-80%
- If only one secondary indicator is anomalous but not severe: 10-30%
- If DutyCycle plus one or more secondary indicators show sustained
    anomalies: 80-95%
- If all three indicators show strong anomalies: 95%+

**Output JSON only, in this format:**
{
  "Classification Result": "Healthy" or "Faulty",
  "Probability of Faulty": <number>,
  "AI Explanations": "brief rationale if Faulty, else None"
}
```