DEVELOPING DATA-DRIVEN MODELS FOR ANOMALY DETECTION IN AUTOMOTIVE AND ADDITIVE MANUFACTURING APPLICATIONS

by
SHAWQI MOHAMMED OTHMAN FAREA

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

Sabancı University
July 2025

DEVELOPING DATA-DRIVEN MODELS FOR ANOMALY DETECTION IN AUTOMOTIVE AND ADDITIVE MANUFACTURING APPLICATIONS

Approved by:

Prof. Dr. MUSTAFA ÜNEL
(Thesis Advisor)

Prof. Dr. BAHATTİN KOÇ

Assist. Prof. Dr. MELİH TÜRKSEVEN

Assoc. Prof. Dr. ALİ FUAT ERGENÇ

Assist. Prof. Dr. ABDURRAHMAN ERAY BARAN

Date of Approval: 01/07/2025

ABSTRACT

DEVELOPING DATA-DRIVEN MODELS FOR ANOMALY DETECTION IN AUTOMOTIVE AND ADDITIVE MANUFACTURING APPLICATIONS

SHAWQI MOHAMMED OTHMAN FAREA

MECHATRONICS ENGINEERING PH.D. DISSERTATION, JULY 2025

Thesis Supervisor: Prof. Dr. MUSTAFA ÜNEL

Keywords: Anomaly Detection, Predictive Maintenance, Transformers, Explainable Artificial Intelligence (XAI), Air Pressure System (APS), Directed Energy Deposition (DED)

Anomaly detection is a fundamental yet inherently challenging task in machine learning and statistics, with wide-ranging applications spanning domains such as healthcare, manufacturing, automotive, and aerospace. Unlike conventional classification problems, anomaly detection must contend with intrinsic difficulties including class imbalance, anomaly heterogeneity, and the scarcity of labeled anomalies. Addressing these challenges requires thoughtfully designed, domain-aware frameworks capable of operating under limited supervision while maintaining robustness and interpretability. This thesis develops several data-driven anomaly detection frameworks, spanning supervised, semi-supervised, and unsupervised learning paradigms. In particular, an efficient semi-supervised framework built upon Transformer architectures is developed, effectively mitigating the inherent challenges of anomaly detection. In addition, the thesis adopts an interpretable framework grounded in Explainable Boosting Machine (EBM), offering transparency and domain-aligned insights without sacrificing performance. A domain-guided preprocessing pipeline is integrated into all frameworks to systematically incorporate expert knowledge, facilitate robust anomaly discrimination, and improve interpretability by aligning feature representations with meaningful physical phenomena.

Two real-world industrial applications were considered in this thesis: (1) failure detection in air pressure systems (APS) of heavy-duty vehicles using operational sensor data, and (2) defect detection in directed energy deposition (DED) using thermal imaging. The APS plays a vital role in ensuring the proper functioning of vehicle subsystems such as braking and suspension, where failures can pose significant safety risks and economic consequences. Meanwhile, DED, an effective additive manufacturing technology, offers a promising pathway for fabricating complex, large-scale components; however, it suffers from recurring in-situ defect formation, compromising part reliability and quality. The data-driven models yielded promising results in both applications. Remarkably, for APS failure detection, the semi-supervised transformer-based approach—although trained using only a small portion of non-anomalous data—led to strong predictive performance on par with the fully supervised models, attaining 91.4% accuracy and an F1 score of 0.79. In parallel, the interpretable EBM-based framework achieved similarly competitive performance (an F1 score of 0.80) while providing meaningful insights into feature contributions and potential root causes, corroborated by domain knowledge. For DED defect detection, semi-supervised models exhibited strong performance, with an accuracy and F1 score up to 95% and 0.88, respectively.

These findings demonstrate that combining domain-specific feature engineering with data-efficient learning paradigms enables effective anomaly detection across diverse settings. The thesis underscores the practical utility of semi-supervised learning—specifically for scenarios with limited anomaly labels—and highlights the growing importance of explainability, particularly in high-stakes applications, where transparent models such as EBM can provide actionable insights without sacrificing accuracy. The frameworks developed in this thesis are readily adaptable to other industrial contexts, depending on the nature of the underlying datasets (balanced vs imbalanced) and desirable characteristics (e.g., highly interpretable). Furthermore, they can be extended to incorporate multi-defect classification, closed-loop control integration, and real-time decision-making.

ÖZET

OTOMOTIV VE EKLEMELI IMALAT UYGULAMALARINDA ANOMALI TESPITI IÇIN VERI ODAKLI MODELLER GELIŞTIRME

SHAWQI MOHAMMED OTHMAN FAREA

MEKATRONİK MÜHENDİSLİĞİ DOKTORA TEZİ, TEMMUZ 2025

Tez Danışmanı: Prof. Dr. MUSTAFA ÜNEL

Anahtar Kelimeler: Anomali Tespiti, Öngörülü Bakım, Dönüştürücüler, Açıklanabilir Yapay Zekâ (XAI), Hava Basınç Sistemi (APS), Yönlendirilmiş Enerji Yığma (DED)

Anomali tespiti, sağlık, imalat, otomotiv ve havacılık gibi alanları içeren geniş kapsamlı uygulamalara sahip, makine öğrenmesi ve istatistik alanlarında temel fakat doğası gereği zorlu bir görevdir. Geleneksel sınıflandırma problemlerinin aksine, anomali tespiti sınıf dengesizliği, anomali heterojenliği ve etiketli anomalilerin kıtlığı gibi içsel zorluklarla mücadele etmelidir. Bu zorlukların ele alınması, sağlamlığı ve yorumlanabilirliği korurken sınırlı denetim altında çalışabilen, dikkatlice tasarlanmış, alana özgü yaklaşımlar gerektirmektedir. Bu tez, denetimli, yarı-denetimli ve denetimsiz öğrenme paradigmalarını kapsayan çeşitli veri odaklı anomali tespiti yaklaşımları geliştirmektedir. Özellikle, anomali tespitinin içsel zorluklarını etkili bir şekilde azaltan Dönüştürücü mimarileri üzerine kurulu verimli bir yarı-denetimli yaklaşım geliştirilmektedir. Ek olarak, tez, performanstan ödün vermeden şeffaflık ve alan hizalı içgörüler sunan Explainable Boosting Machine'ne (EBM) dayalı yorumlanabilir bir yaklaşım benimsemektedir. Alan rehberliğinde bir ön işleme hattı, uzman bilgisini sistematik olarak dahil etmek, sağlam anomali ayrımını kolaylaştırmak ve özellik gösterimlerini anlamlı fiziksel olgularla hizalayarak yorumlanabilirliği iyileştirmek için tüm yaklaşımlara entegre edilmiştir.

Bu tezde iki gerçek dünya endüstriyel uygulaması ele alındı: (1) operasyonel sensör verileri kullanılarak ağır vasıta araçların hava basınç sistemlerinde (APS) arıza tespiti ve (2) termal görüntüleme kullanılarak yönlendirilmiş enerji birikiminde

(DED) arıza tespitidir. APS, arızaların önemli güvenlik riskleri ve ekonomik sonuçlar doğurabileceği frenleme ve süspansiyon gibi araç alt sistemlerinin düzgün çalışmasını sağlamada hayati bir rol oynamaktadır. Öte yandan, eklemeli imalatta etkili bir teknoloji olan DED, karmaşık, büyük ölçekli bileşenler üretmek için umut verici bir yol sunmaktadır; ancak, parça güvenilirliğini ve kalitesini tehlikeye atan tekrarlı arıza oluşumlarından etkilenmektedir. Veri odaklı modeller her iki uygulamada da umut verici sonuçlar vermiştir. Dikkat çekici bir şekilde, APS arıza tespiti için, yalnızca küçük bir anormal olmayan veri bölümü kullanılarak eğitilmiş olmasına rağmen, yarı-denetimli Dönüştürücü tabanlı yaklaşım, tam olarak denetimli modellerle aynı seviyede güçlü bir tahmin performansına yol açtı ve %91,4 doğruluk ve 0,79'luk bir F1 puanı elde etmiştir. Buna paralel olarak, yorumlanabilir EBM tabanlı yaklaşım, alan bilgisiyle doğrulanan özellik katkıları ve olası temel nedenler hakkında anlamlı içgörüler sağlarken benzer şekilde rekabetçi bir performansa (0,80'lik bir F1 puanı) ulaşmıştır. DED kusur tespiti için, yarı-denetimli modeller sırasıyla %95'e ve 0,88'e kadar doğruluk ve F1 puanı ile güçlü bir performans sergilemiştir.

Bu bulgular, alan-spesifik öznitelik mühendisliğini veri açısından verimli öğrenme paradigmalarıyla birleştirmenin farklı senaryolarda etkili anomali tespitini mümkün kıldığını göstermektedir. Tez, özellikle sınırlı anomali etiketlerine sahip senaryolar için yarı-denetimli öğrenmenin pratik faydasını ve özellikle yüksek riskli uygulamalarda açıklanabilirliğin artan önemini vurgulamaktadır. Ayrıca, EBM gibi şeffaf modellerin doğruluktan ödün vermeden eyleme geçirilebilir içgörüler sağlayabileceğini ortaya koymaktadır. Bu tezde geliştirilen yaklaşımlar, temel veri setlerinin doğasına (dengeli veya dengesiz) ve istenen özelliklere (örneğin, son derece yorumlanabilir) bağlı olarak diğer endüstriyel uygulamalara kolayca uyarlanabilir. Dahası, çoklu kusur sınıflandırmasını, kapalı çevrim kontrol entegrasyonunu ve gerçek zamanlı karar vermeyi içerecek şekilde genişletilebilirler.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude to my advisor, Prof. Dr. Mustafa Ünel, for his exceptional supervision, guidance, and unwavering support throughout my PhD journey. His insightful advice, constructive feedback, and continuous encouragement have been instrumental in shaping my research perspective and academic development. It has been a great privilege and honor to be his student and to conduct my research under his distinguished mentorship.

I am also sincerely thankful to the members of my dissertation committee for their valuable time, insightful feedback, and valuable suggestions, which have significantly contributed to the quality and depth of this dissertation.

Special thanks are due to TÜBİTAK, Ford Otosan, and the Sabancı University Integrated Manufacturing Research and Application Center (SU-IMC) for their support and for providing the data essential to this research. I am particularly grateful to Prof. Dr. Bahattin Koç, the TÜBİTAK project coordinator, for his guidance and support throughout the project.

I would like to thank my colleagues for their collaboration and valuable discussions. I am especially grateful to Mehmet Emin Mumcuoğlu, with whom I had the opportunity to jointly work on the "System Health/Anomaly Prediction Based on Connected Vehicle Data" project supported by Ford Otosan. I also thank Hamidreza Javidrad, my colleague on the "New Generation 3D Printer Manufacturing Technologies Platform" (TÜBİTAK 1004 Project, Grant No: 20G050), for his collaboration. I am also thankful to Ufuk İ. Özdek and Yiğit K. Tonkaz for their collaboration on the defect detection studies. My sincere appreciation goes to all members of the Control, Vision, and Robotics (CVR) research group, particularly Naida Fetic, Muhammed Zemzemoğlu, and Asmaa Loulou. I am also thankful to my dear friend, Radhwan Saleh, for all his ongoing support and our many meaningful research discussions.

More importantly, I am deeply indebted to my parents and family members for their unconditional love, patience, and unwavering support throughout my life. Your belief in me has always been a source of strength and motivation.

Last but not least, I extend my heartfelt thanks to my wife for her wholehearted support, patience, and sacrifices throughout the course of this long and challenging journey. Thank you for being my constant companion and source of inspiration.

Dedicated to my dear parents, wife and daughter...

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Abbreviation	Definition
AE	Autoencoders
AI	Artificial Intelligence
ANNs	Artificial Neural Networks
APS	Air Pressure System
AUC	Area Under ROC Curve
CAD	Computer Aided Design
CAN	Controller Area Network
CCD	Charge Coupled Device
CNC	Computer Numerical Control
CNNs	Convolutional Neural Networks
COSMO	Consensus Self-Organizing Model
CT	Computed Tomography
DBSCAN	Density Based Spatial Clustering of Applications with Noise
DCC	Dual Control Charting
DED	Directed Energy Deposition
DL	Deep Learning
DMD	Directed Metal Deposition
E-APU	Electronic Air Pressure Unit
EBM	Explainable Boosting Machine
ECG	Electrocardiogram
ECUs	Electronic Control Units
GAMs	Generalized Additive Models
GAN	Generative Adversarial Networks
GLCM	Gray-Level Co-Occurrence Matrix
GMs	Gaussian Models
GMM	Gaussian Mixture Models
GPUs	Graphics Processing Units
GRU	Gated Recurrent Unit
HDVs	Heavy Duty Vehicles

Abbreviation	Definition
HVs	Healthy Vehicles' Data
iForest	Isolation Forest
IN718	Inconel 718
IQR	Interquartile Range
KDE	Kernel Density Estimation
KL	Kullback Leibler
KNN	K Nearest Neighbors
LENS	Laser Engineered Net Shaping
LIME	Local Interpretable Model-agnostic Explanations
LMD	Laser Metal Deposition
LOF	Local Outlier Factor
LSTM	Long Short Term Memory
ML	Machine Learning
MLPs	Multilayer Perceptrons
PBF	Powder Bed Fusion
PCA	Principal Component Analysis
PdM	Predictive Maintenance
PHM	Prognostics Health Management
Q1	First Quartile
Q3	Third Quartile
RBF	Radial Basis Function
RF	Random Forest
RNNs	Recurrent Neural Networks
ROC	Receiver Operating Characteristic
RoI	Region of Interest
SHAP	Shapley Additive Explanations
SMOTE	Synthetic Minority Over-Sampling Technique
SOM	Self Organizing Maps
STCAR	Spatial Temporal Conditional Autoregressive
SVDD	Support Vector Data Description
SVM	Support Vector Machine
t-SNE	t-distributed Stochastic Neighbor Embedding
Ti64	Ti-6Al-4V
XAI	Explainable Artificial Intelligence
XCT	X-ray Computed Tomography
XGBoost	Extreme Gradient Boosting

1. INTRODUCTION

Anomaly detection is a fundamental problem in machine learning (ML) and statistics, with broad applicability across various domains, including additive manufacturing, automotive systems, aerospace engineering, finance, and healthcare. Depending on the specific application, it is also referred to as fault detection, outlier detection, or novelty detection. Fundamentally, anomaly detection involves identifying data instances that deviate significantly from the expected norm (i.e., the majority of the dataset). These abnormal instances—commonly termed as anomalies, outliers, or novelties—often signal critical events such as faults, failures, or threats. Therefore, timely detection and intervention are crucial for initiating proactive measures and minimizing their adverse impact.

Numerous real-world problems can be framed as anomaly detection tasks. Examples include identifying diseases in the healthcare sector, detecting fraudulent transactions in finance, uncovering cyber intrusions in cybersecurity, and identifying defects or faults in industrial and cyber-physical systems. Moreover, anomaly detection serves as a cornerstone of predictive maintenance (PdM) as well as prognostics and health management (PHM), both of which are aimed at minimizing unplanned downtime and costly breakdowns through early detection of potential failures. Consequently, robust anomaly detection techniques play a pivotal role in ensuring system reliability, operational safety, and decision-making across diverse domains.

The prime objective of anomaly detection is to reliably differentiate between normal and anomalous behavior, despite the inherent challenges posed by the data complexity, dimensionality, and imbalance. Unlike conventional classification tasks, where labeled data is abundantly available for all classes, anomaly detection typically involves sparse and heterogeneous anomalies embedded within vast volumes of normal data. Thus, anomaly detection models must be designed to effectively handle the rarity and variability of anomalous data while maintaining robustness against noise and uncertainty. For high-sensitivity applications, incorporating a degree of interpretability is also essential to support trustworthy decision-making.

1.1 Generic Overview of Anomaly Detection

An overview of a generic anomaly detection framework is shown in Fig. 1.1. The framework comprises three sequential stages: (i) representation learning, (ii) anomaly scoring, and (iii) thresholding. Ideally, these stages are fully integrated, such that the learned representations are optimized for anomaly scoring, and the thresholding step effectively distinguishes between normal and anomalous data points. These stages are explained as follows:

- **Representation Learning:** Also referred to as feature learning, this stage aims to transform raw data into suitable representations for subsequent anomaly detection. However, it varies depending on the nature of the anomaly detection approach. In shallow anomaly detection methods, representation learning often acts as an identity function, where the feature and input spaces are identical. In contrast, kernel-based methods leverage kernel functions to transform the input data into a higher-dimensional feature space, thereby enhancing the separation of normal and anomalous data. For deep anomaly detection models, representation learning is achieved through hierarchical layers of neural networks, enabling the extraction of complex, high-level feature representations.

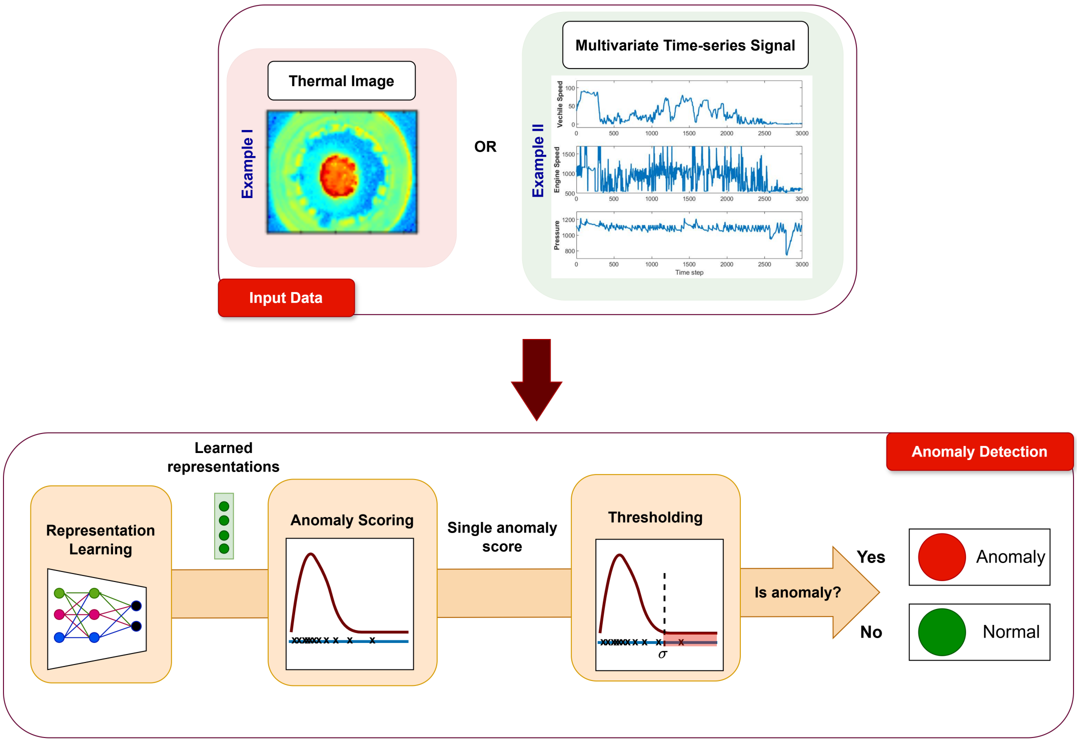


Figure 1.1 Generic overview of anomaly detection

- **Anomaly Scoring:** The second stage involves assigning a quantitative score to each data instance, reflecting its likelihood of being an anomaly. Higher scores indicate a greater degree of deviation from normal patterns. Anomaly scoring can be achieved using various approaches, including distance-based measures, density-based measures, or output layers in neural networks.

- **Thresholding:** The final stage determines whether a given data instance is classified as normal or anomalous by comparing its anomaly score against a threshold. Therefore, the outcome of this step is a binary label. In certain anomaly detection frameworks, such as supervised learning and one-class classification, the scoring and thresholding steps are often combined into a single unified decision function, further streamlining the anomaly detection process.

1.2 Intrinsic Challenges of Anomaly Detection

Anomaly detection presents several intrinsic challenges, driven by the diverse nature, heterogeneity, and rarity of anomalies. Anomalies are typically categorized into three distinct types as follows:

- **Point anomalies:** These are global anomalies that are considered abnormal relative to the entire dataset. A point anomaly is an individual data point that deviates significantly from the expected normal pattern. For example, a recorded temperature of 50 °C in Istanbul would be classified as a point anomaly, as it is highly unusual given the general climate of the region.

- **Contextual (or conditional) anomalies:** These are local anomalies in the sense that they are considered abnormal with respect to their neighbors in some spatial or temporal context. They are more complex than point anomalies and arise in context-dependent data, such as time series or spatial data. For example, a temperature of 30 °C in Istanbul during the winter season would be a contextual anomaly, as 30 °C is expected in summer but highly unusual in winter. In this case, the context (time) is crucial in identifying the anomaly.

- **Group (or collective) anomalies:** They are a set of data points that collectively exhibit abnormal behavior relative to the rest of the dataset. However, the individual points within the group may not be anomalous in isolation. For instance, a prolonged sequence of near-zero values in an electrocardiogram (ECG) signal may indicate a critical abnormality, even though each individual value may not be anomalous when considered independently.

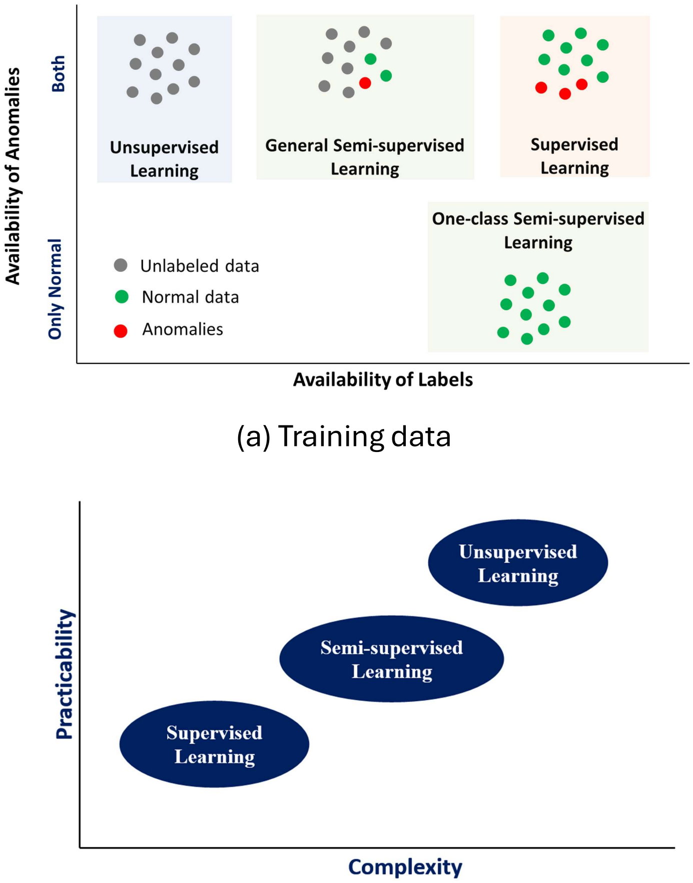
In literature, the majority of studies have predominantly focused on detecting point anomalies, the simplest and most straightforward anomaly type. However, some recent studies (e.g., see Mumcuoglu et al. (2024a)) have employed recurrent neural network architectures to address more complex anomalies that are typically inherent in sequential data. Despite these efforts, effectively identifying such complex anomalies remains relatively underexplored, underscoring the need for more advanced and targeted approaches. For instance, leveraging more advanced sequential architectures, such as attention-based models, holds promise for enhancing both detection performance and computational efficiency.

Traditionally, anomaly detection has been framed as a supervised binary classification problem, assuming that normal data belong to one class and anomalies to another. However, this approach is often impractical in real-world scenarios due to two fundamental characteristics of anomalies: heterogeneity and rarity. Anomalies can originate from multiple distinct classes, making it prohibitively costly, if not infeasible, to obtain sufficient labeled data for each anomaly class. Consequently, the supervised approach is only viable in cases where the dataset is relatively balanced and limited to a single class of anomalies.

To address the aforementioned challenges, anomaly detection is frequently framed as an unsupervised or semi-supervised[1] learning problem, rather than relying solely on supervised approaches. These three learning paradigms are illustrated in Fig. 1.2(a). In the general semi-supervised setting, the training dataset comprises a combination of unlabeled data along with a limited number of labeled instances representing normal and/or anomalous data. A specific case of semi-supervised learning is the well-known one-class learning paradigm, in which the training data exclusively consists of normal instances, as depicted in Fig. 1.2(a). This formulation is particularly effective in scenarios where labeled anomalies are scarce or infeasible to obtain.

Fig. 1.2(b) further illustrates the trade-off between complexity and practicability across the three learning paradigms, with unsupervised learning offering the highest practicality but also posing the greatest modeling complexity. Despite addressing labeling challenges, traditional unsupervised methods struggle to effectively handle high-dimensional and complex data—such as images, text, and multivariate time series—due to the curse of dimensionality. In contrast, deep learning has proven to be highly effective in managing such high-dimensional data, leveraging hierarchical network layers to extract informative representations. Consequently, deep learning architectures can be employed for anomaly detection, where anomaly scoring is applied based on the learned representations. Alternatively, another approach

[1] In literature, semi-supervised learning is sometimes called unsupervised learning as well.



(a) Training data

(b) Practicability vs Complexity

Figure 1.2 Comparison of the learning paradigms

to mitigate the curse of dimensionality involves integrating domain knowledge to extract key features from unstructured data, thus reducing dimensional complexity while retaining essential information.

Another significant challenge in deep anomaly detection lies in the data type dependency of most deep learning architectures. For instance, multilayer perceptrons (MLPs) are generally tailored to structured, tabular data; convolutional neural networks (CNNs) are typically employed for image data; and recurrent architectures are more suited to sequential data such as time series and videos. Consequently, developing a universal deep anomaly detection algorithm capable of effectively handling diverse data types remains a complex task.

Additionally, deep anomaly detection approaches—particularly those based on unsupervised learning—are often prone to high false alarm rates, where normal instances are incorrectly classified as anomalies. One potential solution to mitigate false alarms involves enhancing the expressiveness of the learned representations, enabling the model to more accurately capture the underlying patterns in the data.

Lastly, a substantial portion of the anomaly detection literature has been evaluated using benchmark datasets (e.g., MNIST) originally designed for classification tasks rather than anomaly detection. Therefore, there is a clear need for real-world anomaly detection datasets that accurately reflect the complexities and characteristics of real-world anomaly detection scenarios, enabling more reliable evaluation of proposed anomaly detection algorithms.

1.3 Anomaly Detection in Industrial Applications

Two important real-world industrial applications of anomaly detection are failure detection in air pressure systems of heavy-duty vehicles (HDVs) and defect detection in directed energy deposition, a well-known additive manufacturing process. These applications require further investigation as each one of them poses distinct and complex challenges.

1.3.1 Failure Detection in HDVs

Heavy-duty vehicles (HDVs) are an integral part of the logistics and transportation sectors, where their reliable operation is essential for maintaining industrial efficiency. Operating under demanding conditions, HDVs are susceptible to mechanical failures caused by factors including suboptimal driving practices, inadequate maintenance planning, and ineffective anomaly detection. These failures can result in significant repercussions, ranging from unwanted operational disruptions to serious safety issues. From a manufacturer's perspective, vehicle breakdowns incur considerable costs associated with repairs and warranty claims. For customers, each instance of vehicle downtime results in extended periods of inactivity and increased operational costs, compounding the overall economic burden. Such operational disruptions not only interrupt business processes but also adversely affect profitability, emphasizing the need for minimizing downtime and maximizing vehicle availability.

The air pressure system (APS) is an important subsystem in HDVs, responsible for maintaining adequate air pressure for the braking and suspension systems. A substantial number of roadside breakdowns in HDVs can be attributed to APS-related malfunctions, leading to costly interventions and significant customer dissatisfaction. Mechanical issues, sensor malfunctions, and component failures within the APS can result in excessive load and fatigue, ultimately compromising the integrity of the

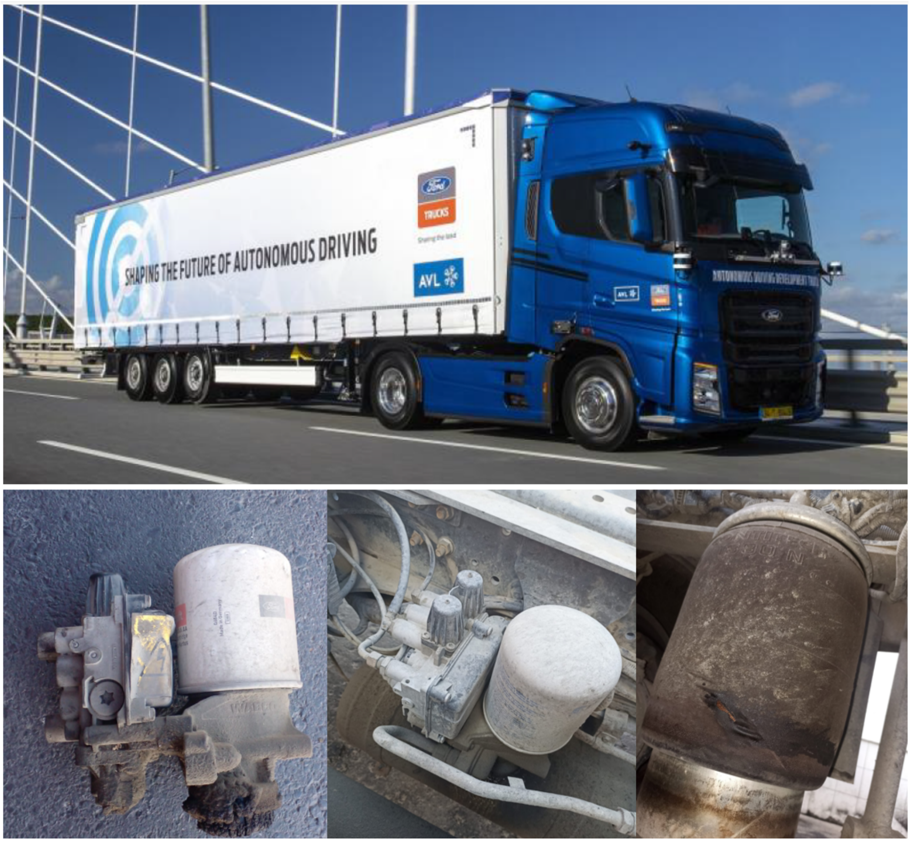


Figure 1.3 Overview of APS failures in HDVs. The top image shows a typical HDV, and the bottom displays three examples of failed E-APUs (images were adapted from Mumcuoglu et al. (2024b) and Aydemir (2024)).

overall system. Some samples of failed electronic air pressure units (E-APUs) are shown in Fig. 1.3. However, addressing these challenges requires a comprehensive approach that includes robust maintenance planning and advanced anomaly detection systems. By integrating these strategies, businesses can effectively mitigate the risks of mechanical failures, thereby leading to a more efficient and resilient transportation ecosystem. Particularly, early detection of APS failures is paramount to avert vehicle stranding during operation and minimize the need for costly roadside assistance.

1.3.2 Defect Detection in Additive Manufacturing

Additive manufacturing, also referred to as 3D printing, is a transformative manufacturing technology that fabricates complex structures through a layer-by-layer deposition process guided by three-dimensional digital models. This approach offers unparalleled design flexibility and enables high levels of customization. Additive manufacturing encompasses a diverse array of processes, including directed energy deposition (DED) and powder bed fusion (PBF), among others, each employing distinct mechanisms for material deposition and consolidation to fabricate high-precision components. Unlike traditional subtractive manufacturing methods, which remove material to shape a part, additive manufacturing directly fabricates components by sequentially adding material, typically in the form of metal powders, polymers, or composites. This approach enables the production of complex geometries and lightweight structures that would be challenging or impossible to achieve with conventional techniques. Additive manufacturing has gained significant traction across various industries—including aerospace, automotive, and healthcare—owing to its potential to reduce material waste, shorten production times, and enable on-demand manufacturing. As the field continues to evolve, research efforts are increasingly focused on enhancing process efficiency, improving material properties, reducing in-situ defects, and ensuring consistent quality in additively manufactured parts.

As one of the promising additive manufacturing processes, DED offers the ability to fabricate dense metal components with precise functional geometries and enhanced mechanical properties. Distinguished by its high deposition rates and optimized material utilization, DED presents a cost-effective solution particularly for applications such as prototyping, repairing, and modifying metal parts (Wolff et al., 2019). Moreover, it is known for its ability to process a wide range of materials, including metals, alloys, and metal matrix composites, with the added capability of multi-material deposition and large-scale structures manufacturing (Dong et al., 2023). These compelling characteristics enhance its applicability across a broad range of industries where precision, durability, and material efficiency are paramount.

The process involves the simultaneous feeding of a feedstock material—typically in the form of powder or wire—into a melt pool created by a focused high-energy heat source such as a laser, electron beam, or plasma arc (refer to Fig. 1.4). As the feedstock material is continuously fed into the localized melt pool through a nozzle system, it melts and subsequently solidifies upon subsequent cooling, forming a strong bond with the underlying substrate or previously deposited layers. The deposition head—typically mounted on a multi-axis robotic manipulator—follows a predefined toolpath derived from computer-aided design (CAD) data, enabling the fabrication of complex geometries with high dimensional precision and design flexibility. Achieving consistent build quality relies heavily on the precise tuning of key process parameters, including the energy source and material feed rate. This tight control is essential for ensuring melt pool stability, which directly influences the microstructural integrity and mechanical properties of the fabricated part.

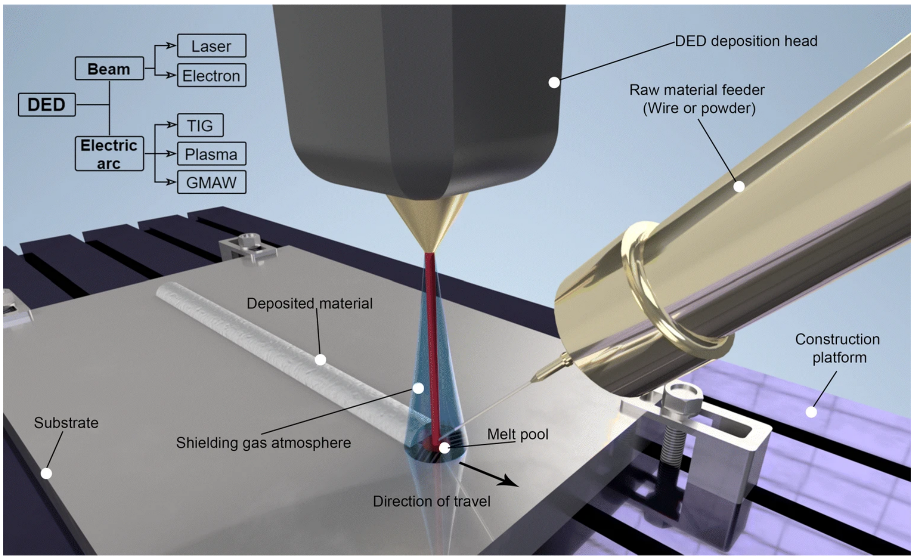


Figure 1.4 Overview of the DED process (Dávila et al., 2020)

Nonetheless, the primary limitation of DED processes arises from the in-situ formation of defects, including lack of fusion, porosity, cracking, and surface roughness. Some of these defects are shown in Fig. 1.5. Such defects can significantly compromise the mechanical properties and overall quality of the final components (Zhu et al., 2021). The formation of defects is predominantly attributed to factors such as feedstock quality (e.g., impurities or porosity in the feedstock), suboptimal process parameters (e.g., laser power and scan speed), as well as the high thermal gradients and rapid cooling rates inherent in the DED processes (Svetlizky et al., 2021).

1.4 Motivation

This thesis addresses the anomaly detection problem by developing different data-driven frameworks and demonstrating their effectiveness across two important industrial applications. Accordingly, the main motivation and objectives of the thesis can be summarized under these two anomaly detection applications as follows:

Failure Detection in Heavy-duty Vehicles: Early detection of APS failures is a particularly challenging task that requires extensive domain expertise. In practice, such failures are often identified through manual inspections during routine maintenance or in response to customer complaints, typically when air leaks in the APS are detected. However, this approach is reactive and prone to oversight, as air leakages may go unnoticed or misdiagnosed. Additionally, it is not uncommon for maintenance teams to replace fully functional units as a preventive measure or to

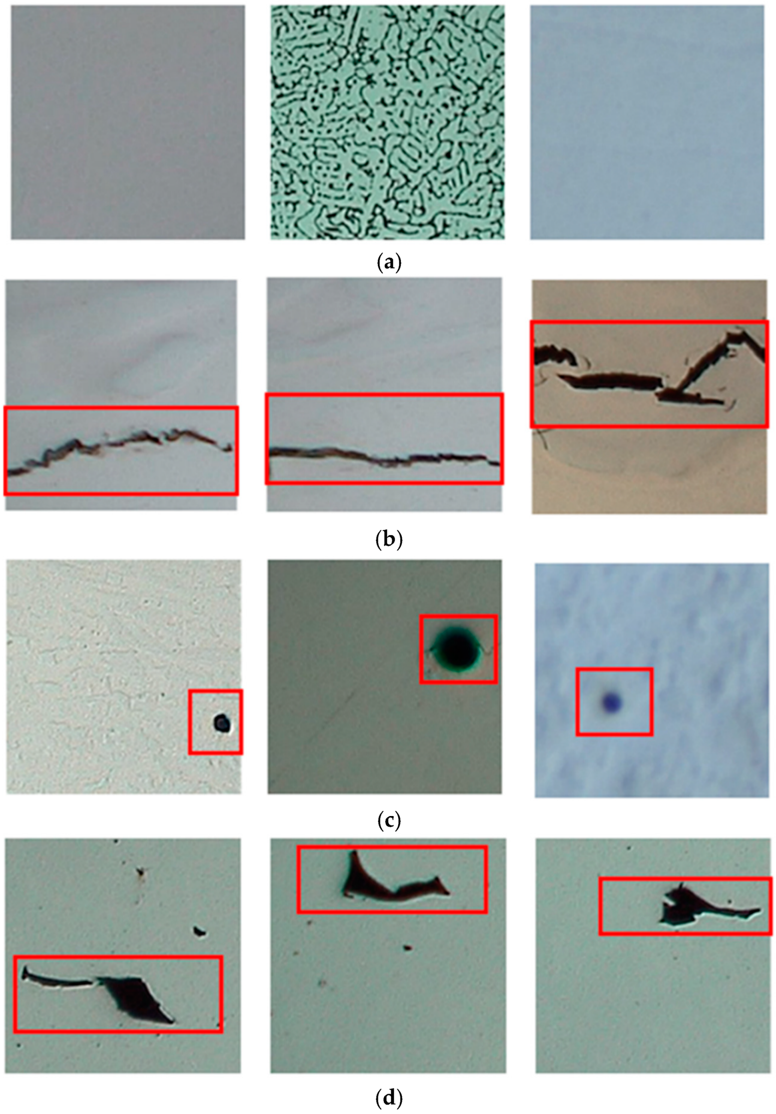


Figure 1.5 Examples of defects in DED processes: (a) defect-free samples, (b) cracks, (c) porosities, and (d) lack of fusion (Cui et al., 2020)

maintain customer satisfaction, further exacerbating operational costs. Addressing the complexities and risks associated with APS failures in HDVs necessitates the development of advanced anomaly detection systems. However, despite its critical importance, research on detecting APS failures using operational driving data remains limited. Given the uncertain nature of such failures and the scarcity of labeled data in the automotive sector, the adoption of modern ML techniques, particularly semi-supervised approaches, presents a promising direction. In addition, recent advancements in sequential deep models have demonstrated substantial success in anomaly detection across various domains, particularly in fault detection tasks (Khalid Fahmi et al., 2024; Maldonado-Correa et al., 2024). Leveraging these architectures to address the specific challenges of APS failure detection could be especially beneficial, enhancing the reliability of HDVs and advancing the field of intelligent transportation systems. Such a framework would not only mitigate the risks of unplanned vehicle breakdowns but also establish a benchmark for research and development in predictive maintenance for commercial vehicles.

Defect Detection in DED Processes: The underlying physics of DED processes is inherently complex, characterized by the dynamic interaction between the deposition material and the energy source, coupled with rapid heat transfer phenomena during melting and subsequent solidification (García-Moreno, 2019). These complexities present significant challenges in developing robust physics-based models capable of accurately predicting the relationships between process parameters and defect formation. Consequently, defect identification has traditionally relied on post-manufacturing inspection techniques, including metallurgical analysis and X-ray computed tomography (CT). While these techniques effectively identify and characterize defects, they are costly, time-consuming, and unsuitable for real-time detection, making them less practical for high-throughput production environments where immediacy and operational efficiency are crucial. An effective alternative to post-process inspection involves real-time monitoring of the melt pool, sometimes coupled with feedback control and in-situ process optimization (Svetlizky et al., 2021). To this end, thermal cameras and other temperature sensors have been extensively employed to monitor the thermal distribution and geometry of the melt pool, both of which are critical indicators of defect formation (Tian et al., 2021). The close correlation between defect generation and melt pool characteristics creates an opportunity for predictive modeling utilizing advanced computational techniques, including machine learning and deep learning. These data-driven techniques have substantial efficacy in analyzing complex thermal patterns and enabling real-time adjustments to process parameters, thereby enhancing defect detection and control. Thus, such data-driven frameworks establish a foundation for the development of more efficient and robust DED manufacturing systems.

1.5 Main Contributions of the Thesis

As stated in the previous section, this thesis develops different data-driven anomaly detection frameworks while demonstrating their effectiveness across the two industrial applications: (i) failure detection in APS systems through time series sensor data and (ii) defect detection in DED processes using thermal imaging. The proposed frameworks leverage a range of data-driven models encompassing supervised, semi-supervised, and unsupervised learning paradigms, tailored to address domain-specific challenges. The primary contributions of this work are as follows:

Problem I: Failure Detection in APS Systems

- The APS failure detection is framed as a semi-supervised anomaly detection problem, leveraging the robustness of this approach to address the inherent challenges associated with failure detection. Building on this formulation, we develop a framework based on two distinct transformer architectures, capitalizing on the demonstrated effectiveness of transformers in various domains, including anomaly detection.

- An explainable artificial intelligence-enhanced framework is developed for APS failure detection, aiming at improving the transparency and reliability of model predictions in this high-stakes application. The proposed framework hinges on two explainable artificial intelligence (XAI) models, namely Explainable Boosting Machine (EBM) and Shapley Additive Explanations (SHAP). The framework provides interpretable and actionable explanations for the final predictions while preserving competitive performance comparable to that of black-box models, effectively balancing predictive accuracy with interpretability.

- To effectively manage APS-related hierarchical operational data, several essential preprocessing steps are developed and integrated into the failure detection framework. Specifically, the framework includes the extraction of strategically engineered features, guided by domain expertise, to act as key indicators for identifying APS failures.

- The thesis work involved the curation of an APS-related dataset consisting of 30 days of time series operational data from two distinct groups of HDVs: 30 anomalous vehicles with documented APS failures that necessitated component replacement, and 110 vehicles classified as healthy based on their clean maintenance records.

Problem II: Defect Detection in DED Processes

- The formulation of the defect detection task is strategically aligned with the characteristics of the respective datasets. The problem is framed as semi-supervised anomaly detection, given the high class imbalance of the corresponding first dataset due to the scarcity of anomalies. Conversely, in the other defect detection dataset characterized by a nearly balanced class distribution and a single defect class, the defect detection task is framed as a supervised classification problem, leveraging the availability of sufficient labeled data from both classes.

- As an essential part of the defect detection framework, a comprehensive feature extraction strategy is meticulously developed to enhance defect detection in thermal images. It incorporates key geometric and thermal distribution features of melt pools, alongside spatial-temporal encodings to capture the sequential and spatial dynamics of the DED process, equipping the models with critical contextual information for more accurate and process-aware defect identification.

- Given the high cost and inherent challenges associated with collecting labeled data for defect detection in DED processes, this work makes a significant contribution by providing a relatively large and diverse dataset, compared to the datasets utilized in recent studies (e.g., Park et al. (2025)). Unlike most existing datasets, which predominantly feature simple geometries such as thin walls and cuboids, the collected dataset focuses on a more complex structure—multi-track, multi-layer hollow cylinders. These geometries present a more challenging detection scenario, as porosities near the inner surface are particularly critical, posing a heightened risk of crack initiation and reduced fatigue life (Ahn, 2021; Kim et al., 2024).

1.6 Thesis Structure

The remainder of this thesis is structured as follows:

- **Chapter 2:** provides a comprehensive review of the existing literature on anomaly detection, encompassing statistical methods, conventional ML techniques, and state-of-the-art deep learning models. In addition, it reviews data-driven approaches specifically developed for failure detection in APS systems and defect detection in DED processes.

- **Chapter 3:** presents the proposed anomaly detection frameworks while focusing on the underlying data-driven models, spanning supervised, semi-supervised, and unsupervised learning paradigms. These models include also XAI techniques to enhance the interpretability of anomaly detection systems.

- **Chapter 4:** introduces the two industrial application domains: the automotive APS system and the additive manufacturing-based DED process. A detailed description of each system is provided, followed by a thorough explanation of the data preparation and preprocessing strategies applied to the respective datasets.

- **Chapter 5:** presents the experimental results of the proposed anomaly detection frameworks, along with detailed discussions. It covers failure detection in APS systems and defect detection in DED processes.
- **Chapter 6:** provides the concluding remarks and future research directions in data-driven anomaly detection and explainable AI in industrial systems.

1.7 Publications

The work of this thesis resulted in the following publications:

- **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.
- Mumcuoglu, M. E., **Farea, S. M.**, Unel, M., Mise, S., Unsal, S., Cevik, E., Yilmaz, M., & Koprubasi, K. (2024). Detecting APS failures using LSTM-AE and anomaly transformer enhanced with human expert analysis. *Engineering Failure Analysis, 165*, 108811.
- **Farea, S. M.**, Javidrad, H., Unel, M., & Koc, B. (2025). In-situ defect detection in directed energy deposition using thermal imaging and machine learning. *Progress in Additive Manufacturing.* **(Minor Revision)**
- **Farea, S. M.**, Mumcuoglu, M. E., Unel, M., Mise, S., Unsal, S., Cevik, E., Yilmaz, M., & Koprubasi, K. (2024). Prediction of failures in air pressure system: A semi-supervised framework based on transformers. In *2024 IEEE 22nd International Conference on Industrial Informatics (INDIN)*, (pp. 1–5).
- **Farea, S. M.**, Unel, M., & Koc, B. (2024). Defect prediction in directed energy deposition using an ensemble of clustering models. In *2024 IEEE 22nd International Conference on Industrial Informatics (INDIN)*, (pp. 1–6).
- Mumcuoglu, M. E., **Farea, S. M.**, Unel, M., Mise, S., Unsal, S., Cevik, E., Yilmaz, M., & Koprubasi, K. (2024). Air pressure system failures detection using LSTM-autoencoder. In *2024 IEEE International Workshop on Metrology for Automotive (MetroAutomotive)*, (pp. 82–87).
- Ozdek, U. I., Tonkaz, Y. K., **Farea, S. M.**, & Unel, M. (2025). Semi-supervised anomaly detection in directed energy deposition using thermal im-

ages. In *22nd International Conference on Informatics in Control, Automation and Robotics (ICINCO)*. **(Under Review)**

- Ayyildizli, B., Balota, B., Tatari, K., **Farea, S. M.**, & Unel, M. (2025). Anomaly detection in directed energy deposition: A comparative study of supervised and unsupervised machine learning. In *22nd International Conference on Informatics in Control, Automation and Robotics (ICINCO)*. **(Under Review)**

2. LITERATURE REVIEW

This chapter provides a foundational review of existing approaches to anomaly detection, spanning statistical methods, machine learning (ML), and deep learning (DL) techniques. It further contextualizes the scope of this thesis by reviewing existing research in the two industrial applications: failure detection in air pressure systems (APS) and defect detection in directed energy deposition (DED) processes. By examining the literature landscape, this chapter establishes the motivation and background for the anomaly detection frameworks developed in the next chapter.

2.1 Anomaly Detection

Anomaly detection has widespread applications across various domains, leading to the publication of numerous review studies. For example, Chandola et al. (2009) reviewed shallow anomaly detection methods, while Pang et al. (2021) and Landauer et al. (2023) focused on DL-based anomaly detection techniques. A broader review, covering both shallow and deep anomaly detection approaches, was conducted by Ruff et al. (2021). Additionally, some reviews focused on specific data types, such as time series anomaly detection (Braei & Wagner, 2020; Zamanzadeh Darban et al., 2024), while others have explored anomaly detection applications in specialized domains, e.g., wireless sensor networks (Xie et al., 2011). This extensive anomaly detection literature can be divided into three broad categories: statistical, classical ML, and DL techniques. In this section, we will summarize the key algorithms in each category, with a special emphasis on DL techniques.

2.1.1 Statistical Techniques

Statistical techniques are based on the assumption that normal data falls into high-probability regions while abnormal data belongs to low-probability regions. Therefore, statistical techniques, in general, attempt to estimate the data-generating distribution. Based on that estimated distribution, data points with low probability are deemed anomalies. The main techniques in this category are Gaussian models, boxplots, histograms, and kernel density estimation.

Gaussian Models (GMs): GMs assume that the underlying distribution is Gaussian. Then, the distance (e.g., Euclidean or Mahalanobis distance) between the data point and sample mean is used as the anomaly score that any data point that is more than a predefined threshold far from the sample mean is labeled as an anomaly. This threshold can be defined based on some validation data. As an example, a GM was proposed by Zamouche et al. (2023) to detect the anomalies among the basic safety messages coming from connected vehicles.

Boxplots: As another statistical techniques, boxplots calculate the first quartile (Q1), second quartile (median), and third quartile (Q3), with the difference between Q1 and Q3 termed as the interquartile range (IQR). Then, anomalies are the data points that are $1.5 \times IQR$ higher than Q3 or $1.5 \times IQR$ lower than Q1. This simple technique was utilized by Sajid et al. (2021) to detect defects in concrete plates.

Histogram-based Approaches: Some histogram-based approaches were also used in the literature for anomaly detection, one of which is Kind et al. (2009). After reconstructing the histogram of the data, the anomaly score for any data point can be deduced from the height of the bin in which that data point falls.

Kernel Density Estimation (KDE): KDE is a more sophisticated statistical anomaly detection approach. This method estimates the underlying density function of the data through a linear combination of kernel functions, each of which is centered at one data point. Next, any data point that lies in the low-probability regions is declared as an anomaly. A tricky difficulty associated with KDE is the choice of its bandwidth. To remedy this problem, Zhang et al. (2018) proposed an adaptive KDE where the bandwidth for each point depended on its distance from its k nearest neighbors. Lang et al. (2022) utilized KDE for anomaly detection in semiconductor fabrication processes.

Limitations of Statistical Techniques: The major drawback to GMs, though a simple and intuitive approach, is the fact that the underlying distribution of most practical, complex data is not Gaussian. In contrast, boxplot-based, histogram-based, and KDE-based anomaly detection techniques are non-parametric techniques; that is, they do not assume any specific distribution, and that makes them more practical than the parametric GMs. However, the main disadvantage of all statistical approaches is their inefficiency with high-dimensional data.

2.1.2 Classical ML Techniques

The algorithms in this category can be further divided into four subcategories: distance-based, clustering-based, classification-based, and reconstruction-based techniques.

2.1.2.1 Distance-based Techniques

The distance-based techniques are relatively straightforward and depend on the assumption that the normal data points constitute dense neighborhoods in contrast to anomalies, which are far away from their closest neighbors.

The simplest distance-based technique is the k-nearest neighbors (KNN) algorithm. In this algorithm, the distance between the data point and its k-th nearest neighbor is defined as its anomaly score. Al Samara et al. (2023) used this technique to detect outliers in wireless sensor networks. As another well-known distance-based anomaly score, the local outlier factor (LOF) measures the local density of a data point with respect to the local densities of its k nearest neighbors; this local density of a data point is inversely proportional to the distance to its k-th nearest neighbor. Accordingly, such an anomaly measure incorporates information about the local neighborhood of the data point. Pokrajac et al. (2007) applied this anomaly score to video anomaly detection. Although KNN and LOF are easy-to-implement unsupervised methods, they suffer from high computational complexity as they involve finding the k-th nearest neighbors for each data point.

Isolation forest (iForest), which is an ensemble of binary trees (Liu et al., 2008), and its variant (Hariri et al., 2019) are considered as distance-based anomaly detection algorithms as well. iForest tries to isolate all the data points in the hope that anomalies will be easily isolated. Thus, anomalies normally lie near the root node of the binary trees on average. For each data point, its anomaly score is inversely proportional to the path length from the root node to the terminating node.

2.1.2.2 Clustering-based Techniques

Clustering is a well-established unsupervised subfield. In the anomaly detection literature, many studies leveraged different clustering algorithms – e.g., K-means, Gaussian mixture models (GMM), self-organizing maps (SOM), and density-based spatial clustering of applications with noise (DBSCAN). These anomaly detection techniques are based upon the assumption that anomalies are not clustered together as opposed to normal data, which normally belongs to dense clusters.

For instance, Wu et al. (2023) applied GMM to detect anomalous sounds from industrial machines whilst Ramadas et al. (2003) used instead SOM for detecting network intrusions. Further, Ijaz et al. (2020) detected outliers in a cervical cancer data set using DBSCAN. One key disadvantage of these algorithms is that they are originally optimized for clustering, not anomaly detection. Moreover, anomalies in some applications tend to form their clusters, which violates the underlying assumption.

2.1.2.3 Classification-based Techniques

One-class Support Vector Machine (One-class SVM) and Support Vector Data Description (SVDD) have been widely used in the literature for anomaly detection. While one-class SVM searches in the kernel space for the hyperplane separating the data from the origin with the maximum margin, SVDD searches for the hypersphere with the minimum radius that encloses the data in the kernel space. Yet, both algorithms are equivalent in the case of the Gaussian kernel. These one-class classification algorithms hinge upon the assumption that normal data comes from one class; hence, the learned boundary is used to separate the normal data from anomalies. Hejazi & Singh (2013) applied one-class SVM to detect fraudulent transactions in credit cards. SVDD, on the other hand, was implemented by Zhang et al. (2016) to detect anomalous objects and behaviors in videos.

The limitation of classification techniques is their sensitivity to the existence of anomalies in the training data. To overcome this limitation, they are trained in a one-class semi-supervised fashion where the training data includes only normal data.

2.1.2.4 Reconstruction-based Techniques

Principal component analysis (PCA) and its variants, like kernel PCA and robust PCA, are the key shallow reconstruction-based algorithms. Their underlying assumption is that normal data, contrary to anomalies, can be efficiently reconstructed from a reduced low-dimensional space. In accordance with that assumption, the reconstruction error between a data point and its reconstructed version is considered as its anomaly score. For instance, Jablonski et al. (2015) utilized PCA for anomaly detection in hyperspectral images. Nonetheless, the main limitation of using PCA and its variants for anomaly detection is that they are optimized for dimension reduction rather than anomaly detection.

2.1.2.5 Limitations of Classical ML Techniques

ML anomaly detection techniques are inefficient in dealing with high-dimensional data, including images and time series. That is due to their shallow nature and inherent dependence on handcrafted features. The common solution for that limitation is deep learning, which has been proven to deal efficiently with high-dimensional data without any prior feature extraction step.

2.1.3 DL Techniques

Recently, many DL approaches have been proposed for anomaly detection. These approaches can be categorized into four subcategories: deep feature extraction-based, reconstruction-based, one-class classification-based, and self-supervised learning-based techniques.

2.1.3.1 Deep Feature Extraction-based Techniques

Since shallow anomaly detection methods fail to deal with high-dimensional data, the naïve extension is by using deep learning solely for feature extraction, on top of which a shallow anomaly detection method is used to produce anomaly scores for the data points.

Andrews et al. (2016) combined transfer learning with one-class SVM for anomaly detection in the benchmark MNIST dataset (Lecun et al., 1998). One-class SVM was trained on the features extracted from a pre-trained VGG. Similarly, a pre-trained ResNet-50 was used by Pang et al. (2020) as a feature extractor; then, a fully connected single-hidden-layer neural network was trained based on these features to produce the anomaly score. In contrast, one-class SVM was trained by Ribeiro et al. (2020) using autoencoder-extracted features from video data.

The main drawback of these methods is that representation learning is completely independent from anomaly scoring. As a result, the learned representations might not be optimized for anomaly detection; in other words, these representations might not be sufficiently expressive to enable the discrimination between normal and abnormal data.

2.1.3.2 Reconstruction-based Techniques

Autoencoders (AE) and generative adversarial networks (GAN) are the key techniques in this subcategory. They are normally trained using only normal data (i.e., one-class semi-supervised learning) so that AE (or GAN) learns the regularities of normal data. During the test stage, the reconstruction error is used as the anomaly score. Similar to PCA-based techniques, the underlying assumption is that normal data is easy to reconstruct from the latent space, whereas anomalies are poorly reconstructed as they are not seen during training.

Bergmann et al. (2019) applied a convolutional AE, together with other anomaly detection methods, to the MVTec anomaly detection dataset. This dataset is recognized as an anomaly detection benchmark comprising defect-free images and some images with different defects. Semi-supervised learning was adopted in this work, where the training data consisted of only defect-free images. Tsai & Jen (2021) utilized a regularized convolutional AE to detect anomalies (i.e., defects) among images of liquid crystal displays and printed circuit boards.

The previous works dealt with image data, for which a convolutional architecture was adopted as the underlying AE architecture. For sequence data, on the other hand, a recurrent architecture is normally proposed so as to capture the intrinsic temporal dependence in sequence data. Zhang et al. (2019) proposed a convolutional recurrent AE to detect anomalies in multivariate time series collected from a power plant. First, signature matrices were generated from the multivariate time series. Next, these matrices were fed into the convolutional AE, and an attention-based convolutional long short-term memory (LSTM) network was applied in the latent space to capture the temporal dependencies. The reconstruction errors between the original signature matrices and their reconstructed signature matrices were computed to provide anomaly scores. Similarly, Park et al. (2018) built a variational AE based on LSTM, a recurrent neural network (RNN) variant, to model time dependencies in multivariate time series. On the other hand, Su et al. (2019) combined a variational AE with another RNN variant—gate recurrent unit (GRU)— to detect anomalies in two aerospace-related multivariate time series and a third multivariate time series collected from a server machine. Kieu et al. (2019) used the median of the reconstruction errors of recurrent AE ensembles as the anomaly score. In contrast, Lu et al. (2017) integrated a denoising AE with an RNN and applied their proposed method to different sequential datasets.

Another research direction in AE-based methods is through imposing a probabilistic model on AE's latent space. Abati et al. (2019) estimated the probability density of the latent representations. Their overall training objective was to minimize the reconstruction error while maximizing the log-likelihood of the latent representations. At test, the anomaly score was computed as the combination of the reconstruction error and negative log-likelihood. Likewise, Zong et al. (2018) implemented a GMM to estimate the density of the latent representations, where the parameters of the GMM were learned through a separate neural network. Therefore, the parameters of AE and GMM were learned simultaneously in an end-to-end fashion. The anomaly score was calculated from the sample energy. Unlike the previous unsupervised methods, Zhou et al. (2021) followed a supervised paradigm and implemented a binary classifier on the hidden representations of an LSTM-based AE.

As one of the earliest works in GAN-based anomaly detection, Schlegl et al. (2017) proposed a GAN for anomaly detection in tomography images of the retina. However, at test time, they implemented a backpropagation-based optimization to map the test image into the latent space of the trained generator. Likewise, Deecke et al. (2019) used a GAN for anomaly detection in image data while conducting an iterative gradient-based search for a latent variable in the latent space at test time. Unlike Schlegl et al. (2017), only the reconstruction error between the test image and its reconstructed version was used as the anomaly score. One obvious shortcoming of the aforesaid studies is the computationally expensive optimization during testing. As fast alternatives, Zenati et al. (2019,1) adopted a bidirectional GAN. The bidirectional GAN involves an additional encoder, along with the generator and discriminator, to perform the mapping from the input space into the latent space. Similar to bidirectional GAN, Schlegl et al. (2019) proposed a two-step approach; the first step involved training a GAN while the second step involved training an additional encoder based on the already trained generator.

Additionally, RNN-based GANs were suggested in the literature to process sequence data. Li et al. (2019) proposed an LSTM-based GAN (i.e., LSTM was used as the underlying architecture for the generator and discriminator) for anomaly detection in time series. At test time, a combination of the discriminator-produced probability of the test sample and the reconstruction error between the test sample and its generator-reconstructed version was used as the anomaly score. However, testing was computationally expensive since it involved an optimization problem to search for the closest latent variable to the test sample. Apart from the previous GAN-based methods, which were based on the reconstruction error, Liang et al. (2021) proposed a probabilistic GAN-based method in which the discriminator-produced probabilities of signature matrices were used as their anomaly scores.

The limitation of reconstruction-based approaches is that they are not directly optimized for anomaly detection. While AE-based methods are optimized for dimension reduction, GAN-based methods are optimized for data generation. GANs are also renowned for their efficiency in estimating the distribution of high-dimensional, complex data. However, they suffer from instability during training due to vanishing gradients, which leads to their failure to converge. Some researchers suggested using Wasserstein GAN instead, claiming it was more stable than standard GAN.

2.1.3.3 One-class Classification-based Techniques

One-class classification is one of the most efficient anomaly detection approaches. As the prime work in this direction, Ruff et al. (2018) proposed deep SVDD, a deep learning extension of standard SVDD, for unsupervised anomaly detection. Deep SVDD trains a deep neural network to transform the majority of the network outputs into a hypersphere with minimum volume (i.e., deep SVDD solves the optimization problem of standard SVDD using deep neural networks). This hypersphere can be considered as the discriminative boundary between normal and abnormal data. The same research team extended deep SVDD into the semi-supervised setting where some portions of labeled data, both normal and abnormal data, were available in addition to the unlabeled data (Ruff et al., 2019,2). Besides the objective of the unsupervised deep SVDD, the semi-supervised version aimed to correctly discriminate the labeled anomalies from the labeled normal data as well.

2.1.3.4 Self-supervised Learning-based Techniques

Many researchers have proposed self-supervised approaches for anomaly detection. Given unlabeled data, self-supervision involves creating an auxiliary classification/prediction task with corresponding pseudo-labels in the hope of learning expressive representations for anomaly detection through this auxiliary task. Accordingly, normal data is expected to be consistent with this auxiliary task compared to anomalies, which are most likely to be inconsistent with the task.

For static image data, the auxiliary task can be multiclass classification of different geometric transformations (e.g., flipping and rotation) which have already been applied to the data. Each transformation is considered a pseudo-class. Golan & El-Yaniv (2018) trained a deep neural network to perform this auxiliary classification using only normal images. At inference, the test image underwent the same transformations, and its anomaly score was induced from the softmax probabilities of the trained neural network. Wang et al. (2019) adopted a similar approach but in an unsupervised fashion. Although both normal and abnormal data were included during training, the model favored the normal data, as the majority, over anomalies during training. In contrast, the auxiliary task used by Li et al. (2021) was a binary classification between normal images and synthetic abnormal images. These synthetic images were generated by pasting randomly cut patches into random locations in those normal images. Yi & Yoon (2020) combined a deep SVDD with self-supervision to improve its efficiency for patch-based anomaly detection. As the auxiliary task, they trained a deep network to predict the relative position of a random patch with respect to one of its eight neighbor patches in a 3-by-3 grid. Nonetheless, the above-mentioned geometric transformations are appropriate only for image data. Thus, Bergman & Hoshen (2020) generalized these transformations into affine transformations, which are applicable even for non-image data.

In sequential data, the auxiliary task is normally defined as a prediction of a future sample based on past samples. Contrary to normal data, which is assumed to be easy to predict, anomalies are assumed to be unpredictable. Liu et al. (2018) followed this approach for anomaly detection in video data; they implemented U-net to predict a future frame from past frames, and the anomaly score was calculated based on the discrepancy between the predicted frame and its ground truth. In a similar manner, Munir et al. (2019) used convolutional neural networks (CNN) to predict the next time stamp of a time series based on the past time stamps in a predefined horizon. Unlike these studies, Ren et al. (2019) conducted a binary classification to discriminate anomalies from normal data using a CNN trained on fast Fourier transform-based features extracted from time series. However, the anomalies were synthetically created while the real unlabeled data was considered normal data.

Like reconstruction-based techniques, the limitation of self-supervised learning-based methods is that the learned representations are optimized for auxiliary tasks rather than anomaly detection. To sum up, a summary of the presented anomaly detection literature is provided in Table 2.1.

Table 2.1 Summary of anomaly detection literature

Category	Subcategory	Algorithm	Learning setting	Data type	References
Statistical	Parametric	GMs	Unsupervised	Time series	Zamouche et al. (2023)
	Non-parametric	Boxplot	Unsupervised	Time series	Sajid et al. (2021)
		KDE	Unsupervised	Time series	Zhang et al. (2018)
ML	Distance-based	KNN	Unsupervised	Time series	Al Samara et al. (2023)
		LOF	Unsupervised	Video	Pokrajac et al. (2007)
		iForest	Unsupervised	Tabular	Hariri et al. (2019); Liu et al. (2008)
	Clustering-based	GMM	Unsupervised	Time series	Wu et al. (2023)
		SOM	Unsupervised	Tabular	Ramadas et al. (2003)
		DBSCAN	Unsupervised	Tabular	Ijaz et al. (2020)
	Classification-based	One-class SVM	Unsupervised	Tabular	Hejazi & Singh (2013)
		SVDD	Unsupervised	Video	Zhang et al. (2016)
	Reconstruction-based	PCA	Unsupervised	Image	Jablonski et al. (2015)
DL	Feature extraction-based	-	Semi-supervised	Image	Jablonski et al. (2015); Pang et al. (2020)
		-	Unsupervised	Video	Andrews et al. (2016)
	Reconstruction-based	AE	Semi-supervised	Image	Bergmann et al. (2019); Lu et al. (2017); Napoletano et al. (2021); Ribeiro et al. (2020)
			Semi-supervised	Video	Audibert et al. (2020)
			Semi-supervised	Time series	Tsai & Jen (2021)
			Unsupervised	Time series	Kieu et al. (2019); Park et al. (2018); Su et al. (2019); Zhang et al. (2019)
			Unsupervised	Tabular	Abati et al. (2019)
			Supervised	Tabular	Zong et al. (2018)
		GAN	Semi-supervised	Image	Deecke et al. (2019); Schlegl et al. (2017); Zenati et al. (2019,1); Zhou et al. (2021)
			Semi-supervised	Time series	Li et al. (2019); Schlegl et al. (2019)
	Classification-based	Deep SVDD	Unsupervised	Image	Liang et al. (2021)
			Semi-supervised	Image	Ruff et al. (2018,2)
	Self-supervised learning-based	Classification	Semi-supervised	Image	Li et al. (2021); Ruff et al. (2019); Wang et al. (2019)
			Unsupervised	Image	Golan & El-Yaniv (2018)
		Prediction	Unsupervised	Video	Bergman & Hoshen (2020)
			Unsupervised	Time series	Liu et al. (2018)

2.2 APS Failure Detection

The detection of failures in the air pressure system (APS) has been widely explored through a range of approaches, including traditional methods, ML, and DL approaches. This section categorizes the existing research in this direction into these three methodological domains and presents a comparative evaluation of their respective strengths and limitations.

2.2.1 Traditional Techniques

Traditional approaches for detecting APS failures primarily utilize rule-based systems, expert-driven methodologies, and statistical techniques. While these methods can be effective, particularly when supported by well-established domain expertise, they often lack the adaptability and predictive capabilities offered by ML and DL models.

For instance, Fan et al. (2015a) introduced the Consensus Self-Organizing Model (COSMO) for APS failure detection in a fleet of Volvo buses. They later enhanced this approach by integrating COSMO with expert-driven techniques (Fan et al., 2015b) and incorporating echo state networks, a recurrent neural network (RNN) variant, to improve predictive accuracy and enhance failure detection (Fan et al., 2016). In a different study, Nowaczyk et al. (2013) proposed a fuzzy rule-based model for detecting APS failures in Volvo trucks, benchmarking its performance against ML algorithms such as decision trees and random forests.

2.2.2 ML Techniques

Unlike traditional approaches, ML techniques excel at uncovering intricate, complex patterns within data, enabling them to make accurate predictions and classifications. As a result, both classical ML algorithms (e.g., k-nearest neighbors) and ensemble-based methods (e.g., Random Forest) have been extensively employed in APS failure detection.

A significant portion of research in this category has focused on APS failure detection using the publicly available Scania Trucks dataset (Scania CV AB, 2016). Various ML classifiers have been applied, including k-nearest neighbors (KNN) (Costa &

Nascimento, 2016; Ozan et al., 2016; Rafsunjani et al., 2019), Naive Bayes (Rahman & Sumathy, 2024), logistic regression (Hussain et al., 2024; Muideen et al., 2023; Rahman & Sumathy, 2024), Support Vector Machine (SVM) (Costa & Nascimento, 2016; Rafsunjani et al., 2019; Selvi et al., 2022; Syed et al., 2020), XGBoost (Cerqueira et al., 2016; Hussain et al., 2024; Lokesh et al., 2020), and Random Forest (RF) (Cerqueira et al., 2016; Gondek et al., 2016; Jose & Gopakumar, 2019; Ranasinghe & Parlikad, 2019). Nevertheless, this dataset presents challenges due to severe class imbalance and missing values. To mitigate the class imbalance issue, researchers have employed rebalancing techniques such as the Synthetic Minority Over-sampling Technique (SMOTE). Additionally, missing data has been addressed using various imputation strategies, including median imputation and KNN-based methods.

Despite these efforts, a significant limitation shared by studies utilizing the Scania Trucks dataset lies in the anonymization of feature names, implemented to protect proprietary information. While this preserves data confidentiality, it substantially hinders model interpretability by obscuring the semantic meaning of individual features. As a result, it becomes challenging to assess the relevance of specific inputs in the model's decision-making process or to validate their contributions using domain knowledge—an essential aspect for gaining trust in safety-critical applications.

Beyond the Scania dataset, Prytz et al. (2013,1) explored APS failure prediction in Volvo trucks, applying decision trees, RF, and KNN classifiers. To handle class imbalance in their dataset, they also leveraged the SMOTE technique. Meanwhile, Panda & Singh (2023) investigated APS failure detection in medium-duty vehicles, utilizing decision tree classifiers, both with and without boosting, on a dataset consisting of diagnostic trouble codes in addition to onboard operational data.

2.2.3 DL Techniques

Various DL techniques have also been applied to APS failure detection, capitalizing on their proven ability to automatically extract hierarchical and complex features from raw data. In contrast, traditional ML methods typically depend on manual feature engineering, which is often labor-intensive and potentially less effective. As a result, DL approaches offer significant advantages when analyzing large-scale, high-dimensional datasets.

Rengasamy et al. (2020) investigated the use of DL classifiers based on different deep neural network architectures for APS failure detection in the Scania Trucks dataset. The deep architectures include multilayer perceptron (MLP), CNN, and RNN. In contrast, Fan et al. (2016) applied echo state networks, a rapid variant of RNNs, for identifying APS failures in Volvo bus fleets.

2.2.4 Limitations of APS Failure Detection Literature

The main limitations in the current body of research on APS failure detection can be summarized as follows:

- **Problem formulation:** A key limitation is the prevalent formulation of the problem as a supervised classification task. This approach presumes that there exists only a single class of APS failures. However, APS failures are inherently heterogeneous, as failures in different components can lead to distinct failure types. For example, a malfunctioning pressure sensor results in a failure mode that differs from one caused by a mechanical valve breakdown. Due to this heterogeneity, the failure instances within the training dataset do not comprehensively represent all possible failure classes. A more suitable approach is to frame APS failure detection as a semi-supervised learning problem. In this paradigm, models are trained exclusively on normal (i.e., healthy) data, enabling them to learn the normal behavior and distribution. During inference, these models reconstruct normal data with high accuracy while struggling to do the same for anomalous data. The resulting reconstruction errors serve as anomaly scores, effectively identifying failures. Furthermore, by relying solely on normal data for training, this approach inherently addresses the issue of class imbalance, a common challenge in failure detection tasks.

- **Model explainability:** Another significant limitation is the heavy reliance on black-box ML/DL models. Although these models often deliver strong predictive performance, their lack of interpretability poses a major challenge. The opaque nature of these models makes it difficult for manufacturers and end-users to comprehend the reasoning behind their predictions, potentially diminishing confidence in the obtained predictions.

  Conversely, explainable artificial intelligence (XAI) has gained traction in various automotive applications, offering enhanced transparency and interpretability. For example, in the context of autonomous vehicle localization, Charroud et al. (2023) applied explainable AI techniques, including SmoothGrad and VarGrad, to generate gradient-based explanations for their deep learning model. Li et al. (2023) utilized Shapley Additive Explanations (SHAP), a well-known XAI technique, to enhance the interpretability of their lane change detection model. On the other hand, Mohanty & Roy (2023) leveraged SHAP to gain insights into the key factors influencing energy consumption at electric vehicle charging stations. Despite the increasing adoption of XAI in the automotive field, its integration into APS failure detection remains notably limited. As highlighted in Table 2.2, there is a noticeable gap in the literature when it comes to leveraging XAI techniques to improve model explainability in this domain.

Ahmad Khan et al. (2024) explored the use of XAI techniques, specifically LIME (Local Interpretable Model-agnostic Explanations) (Ribeiro et al., 2016,1) and SHAP (Lundberg & Lee, 2017), to explain the predictions of black-box models applied to APS failure detection in the Scania Trucks dataset. However, due to proprietary constraints, the dataset's features are anonymized, significantly limiting the interpretability and reliability of the generated explanations. Without knowledge of the actual feature semantics, understanding the influence of specific inputs on model decisions becomes challenging, ultimately reducing the practical value of these explanations.

2.3 DED Defect Detection

A wide range of statistical, ML, and DL approaches have been explored in the literature for defect detection in DED processes. A substantial number of these studies have focused on detecting porosities in laser-based DED systems, leveraging thermal images of melt pools as a primary data source. Additionally, most research efforts have been directed toward Ti-6Al-4V components.

Statistical Approaches: Among statistical methods, the Spatial-Temporal Conditional Autoregressive (STCAR) model was employed by Guo et al. (2020), while Dual Control Charting (DCC), in combination with multilinear principal component analysis, was utilized by Khanzadeh et al. (2018) to predict porosity in Ti-6Al-4V-deposited parts based on thermal melt pool images.

Compared to statistical techniques, ML and DL methods have demonstrated superior performance across various domains, including defect detection. Their ability to process large-scale, high-dimensional image data gives them a significant advantage over statistical approaches. As a result, a substantial portion of existing research has focused on leveraging ML and DL techniques for defect detection in DED processes.

Table 2.2 Summary of APS failure detection literature

Category	Learning method	Learning scheme	Dataset	Reference
Traditional methods	COSMO	-	Volvo buses	Fan et al. (2015a)
	Expert-based approach	-	Volvo buses	Fan et al. (2015b)
	Fuzzy rule-based model	-	Volvo trucks	Nowaczyk et al. (2013)
ML methods	KNN	Supervised	Scania trucks Volvo trucks	Costa & Nascimento (2016); Ozan et al. (2016); Rafsunjani et al. (2019) Prytz et al. (2013,1)
	Logistic regression	Supervised	Scania trucks	Hussain et al. (2024); Muideen et al. (2023); Rahman & Sumathy (2024)
	Naive Bayes	Supervised	Scania trucks	Rahman & Sumathy (2024)
	Decision trees	Supervised	Medium-duty trucks Volvo trucks	Panda & Singh (2023) Prytz et al. (2013)
	SVM	Supervised	Scania trucks	Costa & Nascimento (2016); Rafsunjani et al. (2019); Selvi et al. (2022); Syed et al. (2020)
	XGBoost	Supervised	Scania trucks	Cerqueira et al. (2016); Hussain et al. (2024); Lokesh et al. (2020)
	RF	Supervised	Scania trucks Volvo trucks	Cerqueira et al. (2016); Gondek et al. (2016); Jose & Gopakumar (2019); Ranasinghe & Parlikad (2019) Prytz et al. (2013,1)
DL methods	MLP	Supervised	Scania trucks	Rengasamy et al. (2020)
	CNN	Supervised	Scania trucks	Rengasamy et al. (2020)
	RNN	Supervised	Scania trucks	Rengasamy et al. (2020)
XAI methods	SHAP and LIME	Supervised	Scania trucks	Ahmad Khan et al. (2024)

ML Approaches: For example, Khanzadeh et al. (2018) applied multiple ML models—including linear and quadratic discriminant analysis, decision trees, KNN, and SVM—for in-situ porosity detection through thermal imaging of the melt pool. Similarly, Shin et al. (2023) utilized artificial neural network (ANN), KNN, and SVM classifiers to identify pores in single-layer 316L steel specimens. Their approach integrated thermal data from a pyrometer with melt pool images captured by a charge-coupled device (CCD) camera to enhance accuracy. On the other hand, Chen & Moon (2024) employed multiple ML classifiers, including SVM, Random Forest, and XGBoost, for defect detection in wall structures using data collected from a CCD camera and a microphone. Additionally, Assad et al. (2024) implemented logistic regression, KNN, SVM, and MLP models to identify process instabilities in thin 316L steel walls based on high-speed camera data. On a different front, Gaja & Liou (2018) developed supervised classifiers using logistic regression and ANN to detect porosities and cracks through acoustic signals obtained from acoustic emission sensors mounted on the fabricated part.

DL Approaches: DL techniques have also been widely utilized to address defect detection in additive manufacturing. Various studies have explored CNNs and hybrid deep learning models for this purpose. For instance, Zhang et al. (2019) developed a compact CNN-based classifier to detect porosity in sponge titanium parts using melt pool images captured by a high-speed digital camera. Similarly, Cui et al. (2020) proposed a CNN classifier for identifying multiple defect types, including cracks, porosity, and lack of fusion. Tian et al. (2020) introduced a VGG16-based classifier for porosity detection in thin-wall structures manufactured via Laser Engineered Net Shaping (LENS). Their approach utilized thermal images of the melt pool acquired through a built-in pyrometer. Expanding on CNN applications, Patil et al. (2023) evaluated multiple CNN architectures—AlexNet, VGG16, GoogleNet, and ResNet—to detect rough texture and voids in Inconel 625 structures, including horizontal and vertical wall structures as well as cuboid components. In another study, Tian et al. (2021) integrated RNNs with CNN classifiers to detect porosity in titanium samples by fusing thermal data from a pyrometer and a thermal camera. Beyond image-based defect detection, Chen et al. (2023) employed a CNN-based classifier to identify pores and cracks in wall structures using acoustic signal analysis. Additionally, Dong et al. (2023) addressed issues related to size drift and dimensional inconsistencies by implementing a ResNet18-based model to predict deposited layer sizes. Their approach incorporated process parameters, melt pool images, and temperature data to enhance predictive accuracy.

Semi-supervised/Unsupervised Approaches: Beyond the supervised approaches discussed earlier, semi-supervised and unsupervised techniques have also been explored in the literature for defect detection in DED processes. These methods offer advantages in scenarios where labeled data is limited or unavailable. For instance, Zheng et al. (2024) introduced a semi-supervised framework utilizing a convolutional autoencoder followed by a classification head to detect porosity in 30CrNi2MoVA specimens based on melt pool images. In the realm of unsupervised learning, Gaja & Liou (2017) applied K-means clustering in combination with PCA for dimensionality reduction to identify cracks and porosity in titanium- and steel-based alloys using acoustic signals. Similarly, García-Moreno (2019) employed SOM, another unsupervised clustering algorithm, to detect porosity in aluminum-based manufactured parts. Building on the use of SOM, Khanzadeh et al. (2016,1,1) developed a porosity detection methodology that analyzed thermal images of melt pools in titanium alloy-based single-track thin walls. More recently, Farea et al. (2024) proposed an ensemble approach using DBSCAN (Density-Based Spatial Clustering of Applications with Noise) to identify defects in Inconel 718 samples based on thermal melt pool images. These semi-supervised and unsupervised strategies demonstrate the potential of clustering and feature-learning techniques in defect detection, particularly when labeled data is scarce. An overall summary of all studies reviewed in this section is presented in Table 2.3.

Table 2.3 Summary of DED defect detection literature (S: supervised learning, SS: semi-supervised learning, US: unsupervised learning)

Study	**Heat source**	**Feedstock type**	**Feedstock material**	**Sensor**	**Defect type**	**Learning scheme**	**Technique**
Guo et al. (2020)	Laser	Powder	Ti-6Al-4V	Pyrometer	Porosity	Statistical	STCAR
Khanzadeh et al. (2018)	Laser	Powder	Ti-6Al-4V	Pyrometer	Porosity	Statistical	DCC
Khanzadeh et al. (2018)	Laser	Powder	Ti-6Al-4V	Pyrometer	Porosity	S	ML classifiers
Gaja & Liou (2018)	Laser	Powder	Ti-6Al-4V H13 steel	AE sensors *	Cracks Porosity	S	ML classifiers
Shin et al. (2023)	Laser	Powder	316L steel	Pyrometer CCD camera	Porosity	S	ML classifiers
Chen & Moon (2024)	Laser	Powder	MS C300	Microphone CCD camera	Cracks Porosity	S	ML classifiers
Assad et al. (2024)	Laser	Wire	316L steel	High-speed camera	Process instabilities	S	ML classifiers
Zhang et al. (2019)	Laser	Powder	Titanium	High-speed camera	Porosity	S	DL classifier
Cui et al. (2020)	Laser	Powder	Stainless steel Ti-6Al-4V AlCoCrFeNi Inconel 718	Pyrometer	Cracks Porosity Lack of fusion	S	DL classifier
Tian et al. (2020)	Laser	Powder	Ti-6Al-4V	Pyrometer	Porosity	S	DL classifier
Tian et al. (2021)	Laser	Powder	Ti-6Al-4V	Pyrometer IR camera	Porosity	S	DL classifiers
Patil et al. (2023)	Laser	Powder	Inconel 625	Camera	Porosity Rough texture	S	DL classifier
Chen et al. (2023)	Laser	Powder	C300 steel	Microphone	Porosity Cracks	S	DL classifier
Dong et al. (2023)	Laser	Powder	CoCrNi	Pyrometer	Layer size	S	DL classifier
Zheng et al. (2024)	Laser	Powder	30CrNi2MoVA	CCD camera	Porosity	SS	DL classifier
Gaja & Liou (2017)	Laser	Powder	Ti-6Al-4V H13 steel	AE sensors *	Cracks Porosity	US	ML clustering
Khanzadeh et al. (2016)	Laser	Powder	Ti-6Al-4V	Pyrometer	Porosity	US	ML clustering
Khanzadeh et al. (2017)	Laser	Powder	Ti-6Al-4V	Pyrometer	Porosity	US	ML clustering
Khanzadeh et al. (2019)	Laser	Powder	Ti-6Al-4V	Pyrometer	Porosity	US	ML clustering
García-Moreno (2019)	Laser	Powder	Al-5083	Camera	Porosity	US	ML clustering

* In this table, AE sensors refer to acoustic emission sensors

3. METHODOLOGY

This chapter presents the methodological foundation of the proposed anomaly detection frameworks. It begins with a high-level overview of a generic anomaly detection framework, outlining its principal components and flow. It then delves into the core modeling module, which, in this thesis, involves a range of black-box machine learning (ML) techniques as well as explainable AI methods, leading to multiple frameworks. These techniques encompass supervised, semi-supervised, and unsupervised paradigms, each tailored to specific anomaly detection scenarios.

3.1 Overview of Anomaly Detection Framework

An overview of the proposed anomaly detection framework is presented in Fig. 3.1. This thesis focuses on two distinct application domains: automotive systems and additive manufacturing. In the context of automotive systems, anomalies are detected using time series data acquired from various onboard sensors, as will be elaborated in the following chapter. Conversely, anomaly detection in additive manufacturing is performed through thermal images of the melt pool.

The general framework ideally comprises three sequential stages: (i) data preparation and preprocessing, (ii) anomaly detection, and (iii) interpretability. The data preparation and preprocessing stage involves essential steps such as missing value imputation, segmentation, and feature extraction—each tailored to the characteristics and requirements of the specific dataset. Detailed procedures for this stage are presented in the subsequent chapter. The second stage entails the employment of data-driven, black-box models for anomaly detection. These models span supervised, semi-supervised, and unsupervised learning paradigms, selected according to the nature of the detection task and the availability of labeled data. Finally, the third stage introduces interpretability into the pipeline through both ante-hoc and post-hoc explainable AI (XAI) techniques, enhancing transparency and supporting trust in the model predictions. Ante-hoc models, such as Explainable Boosting

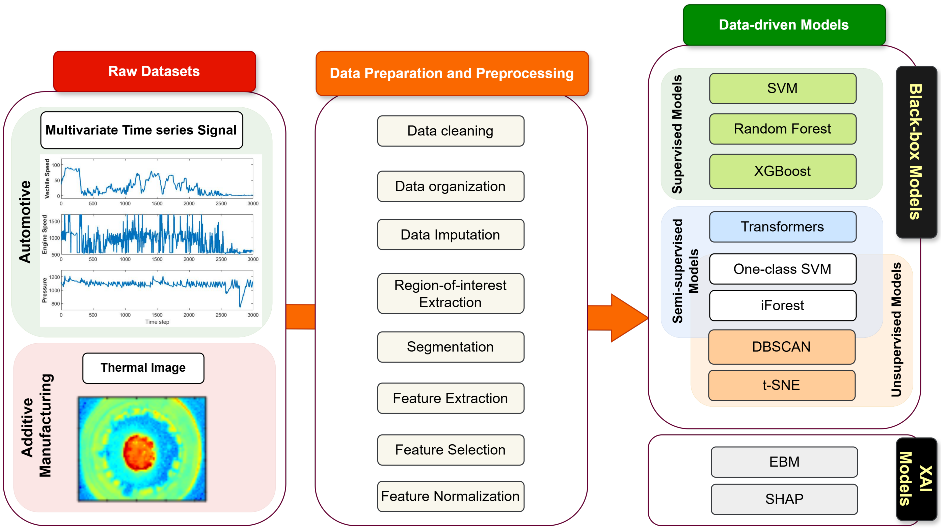


Figure 3.1 Overview of anomaly detection framework

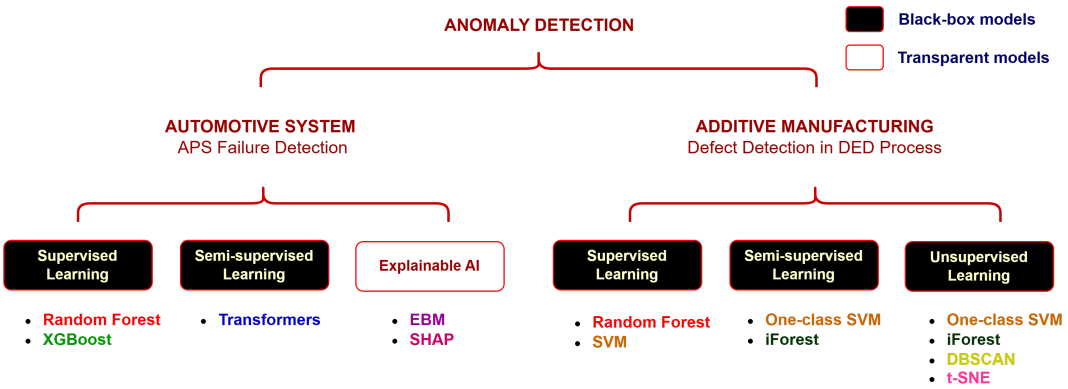


Figure 3.2 Data-driven models across the two industrial applications

Machines (EBM), integrate detection and interpretation simultaneously, whereas post-hoc methods provide explanations after model training[1]. Fig. 3.2 illustrates the full spectrum of the data-driven models utilized in this thesis, indicating their learning scheme and specific application to each of the two industrial problems. The following sections provide an in-depth discussion of these data-driven techniques.

3.2 ML Models for Anomaly Detection

As illustrated in Fig. 3.1, the ML models employed for anomaly detection span across three primary learning paradigms: supervised, semi-supervised, and unsupervised

[1] It is also noted that some of the anomaly detection frameworks proposed in this thesis do not include an interpretability module.

learning. The supervised models include Support Vector Machine (SVM), Random Forest, and XGBoost. Although EBM is also a supervised model, it is discussed separately in Section 3.3 due to its nature as a transparent, glass-box model. The semi-supervised category encompasses transformer-based architectures, as well as models like one-class SVM and Isolation Forest (iForest), which—depending on the training configuration—can also operate in fully unsupervised settings. In addition to these, the unsupervised models include DBSCAN (Density-Based Spatial Clustering of Applications with Noise) and t-SNE (t-distributed Stochastic Neighbor Embedding).

3.2.1 Supervised Approaches

A straightforward approach to anomaly detection frames the problem as a supervised binary classification task. However, this formulation is only appropriate when the dataset is relatively balanced and includes one or more well-represented anomaly classes. In this thesis, the supervised models employed for anomaly detection include a kernel machine (i.e., SVM) as well as tree ensemble methods—Random Forest and XGBoost. These models are widely recognized for their effectiveness in supervised learning, particularly in the context of structured data, and demonstrated strong performance in fault and anomaly detection in various domains (Khan et al., 2023; Noshad et al., 2019; Zhang et al., 2018).

3.2.1.1 SVM

The SVM algorithm is designed to identify an optimal hyperplane that separates data points from different classes with the greatest possible margin. This hyperplane serves as a decision boundary, partitioning the feature space into distinct regions associated with each class label. For datasets that are linearly separable, SVM seeks the hyperplane that maximizes the margin, which is the distance between the hyperplane and its closest data points (known as support vectors) from each class.

In cases of linearly non-separable datasets, SVM leverages the kernel trick to map the original input features into a higher-dimensional space where a linear separation becomes feasible. This transformation is performed implicitly, avoiding the computational burden of explicitly calculating the high-dimensional feature coordinates. Through the use of kernel functions, SVM is able to model complex, non-linear relationships within the data efficiently. In this work, the Radial Basis Function (RBF) kernel, also referred to as the Gaussian kernel, is employed. The RBF kernel is mathematically defined as follows:

$$K(x_i, x_j) = exp(-\frac{||x_i - x_j||^2}{2\sigma^2}) \tag{3.1}$$

where x_i and x_j represent the i-th and j-th input instances, respectively. The parameter σ defines the kernel scale, and $|| \cdot ||$ denotes the Euclidean norm.

In practical applications, achieving perfect separation between data points is rarely feasible due to factors such as noise and overlapping class distributions. To accommodate this, SVM introduces the concept of a soft margin, which permits a limited number of misclassifications. The balance between minimizing classification errors and maximizing the margin is governed by the regularization parameter C. Higher values of C prioritize the accurate classification of training instances, potentially resulting in a narrower margin, whereas lower values of C allow for a broader margin by tolerating more classification errors.

The effectiveness of an SVM model is highly sensitive to the selection of its key hyperparameters, notably the choice of kernel function, the regularization parameter C, as well as kernel-specific parameters (such as the kernel scale in the case of the RBF kernel). Careful tuning of these hyperparameters is essential to strike an appropriate balance between model complexity and generalization performance.

3.2.1.2 Random Forest

Random Forest is widely recognized for its robustness to noisy data and its strong resistance to overfitting, demonstrating high effectiveness across a broad range of applications (Parmar et al., 2019). Random Forest is an ensemble learning method that builds upon the principle of bootstrap aggregating (bagging) to improve the predictive performance of individual decision trees, which are inherently prone to overfitting the training data. During the training phase, the algorithm operates by constructing a collection of decision trees—often referred to as weak learners. Each tree is trained on a different bootstrap sample, created by randomly sampling with replacement from the original training dataset, with the sample size matching that of the original data. Additionally, at each node within a tree, only a randomly selected subset of features is considered for splitting, rather than the full feature set. This dual source of randomness—both in the data used to build each tree and in the feature selection process—reduces the correlation among individual trees. As a result, Random Forest effectively mitigates overfitting while significantly enhancing the ensemble's ability to generalize to unseen data.

The final prediction of the Random Forest model is obtained by aggregating the outputs of all individual trees in the ensemble through a majority voting mechanism. In this process, the class label predicted by the majority of trees is selected as the overall prediction, as formalized by the following equation:

$$\hat{y}_i = mode\{h_j(x_i)|j = 1,2,...,m\} \tag{3.2}$$

Here, $\hat{y}_i$ refers to the predicted class of the observation x_i, $h_j(x_i)$ is the j-th decision tree, and m is the number of weak learners (trees) in the ensemble.

This aggregation strategy significantly reduces variance and enhances the overall robustness of the model (Breiman, 2001; Parmar et al., 2019), enabling Random Forest to maintain strong performance even in the presence of noisy features. Additionally, since the individual decision trees are trained independently, the Random Forest algorithm naturally supports parallel processing, leading to a substantial acceleration in training time. These characteristics make Random Forest particularly well-suited for handling large-scale datasets efficiently.

One of the notable strengths of Random Forest is its intrinsic capability to quantify and rank the importance of input features based on their predictive performance. By leveraging measures such as the reduction in Gini impurity or the decrease in model accuracy upon feature permutation, Random Forest assigns an importance score to each feature, reflecting its relative contribution to model performance. This capability not only improves the model interpretability but also facilitates efficient feature selection—enabling practitioners to prioritize influential features, eliminate noise, and optimize computational efficiency without compromising predictive power. Such interpretability is particularly valuable in high-dimensional datasets, where identifying key features is essential for both optimizing model performance and deriving meaningful domain-specific insights.

3.2.1.3 Extreme Gradient Boosting (XGBoost)

XGBoost (Chen & Guestrin, 2016) is a highly efficient and scalable implementation of the gradient boosting framework (Friedman, 2001), introduced to address the limitations of traditional boosting algorithms in terms of speed, scalability, and model performance. Widely adopted in both industry and academia, XGBoost has demonstrated exceptional accuracy and robustness across a wide range of ML tasks, including classification, regression, and ranking. It has been widely adopted in various fields, including additive manufacturing, automotive, and bioinformatics.

At its core, XGBoost constructs an additive model in a forward manner, where new weak learners (i.e., decision trees) are sequentially added to adjust the residual errors made by existing trees. In other words, each boosting iteration fits a new tree to the residuals of the current model's predictions. At the t-th iteration, the model attempts to minimize a regularized cost function that combines a convex loss function measuring prediction accuracy and a regularization term to control model complexity:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i\,,\, \hat{y}_i^{(t-1)} + f_t(\mathbf{x}_i)\right) + \Omega(f_t) \tag{3.3}$$

where y_i represents the ground truth label of the i-th observation, and $\hat{y}_i^{(t-1)}$ denotes its predicted value at iteration $t-1$. In addition, l is the loss function (e.g., logistic loss for classification) while $\Omega(f_t)$ is the regularization term penalizing the complexity of the newly added tree f_t. This regularized objective function improves the generalization ability of the model by reducing overfitting, which is especially important when working with noisy or high-dimensional datasets.

Like Random Forest, XGBoost, as an ensemble-based algorithm, offers the ability to quantify and rank feature importance based on their contribution to the model's predictive performance. This built-in mechanism enhances the interpretability of XGBoost by providing insights into which input features most significantly influence the model's decisions. Therefore, XGBoost can be deemed a supervised learning baseline with a proven capability to capture complex, non-linear relationships while offering a degree of interpretability through feature importance measures.

3.2.2 Semi-supervised Approaches

The proposed semi-supervised anomaly detection approaches encompass three distinct methodologies: reconstruction-based transformer architectures, distance-based iForest, and classification-based one-class SVM models. In all these methods, the training phase is conducted exclusively on normal (non-anomaly) data, enabling the models to learn the underlying patterns of healthy system behavior and detect deviations indicative of anomalies.

It is important to note that both iForest and one-class SVM can also be applied in an unsupervised setting, wherein the models are trained on datasets containing both normal and anomalous instances. Within the context of this thesis, both iForest and one-class SVM were applied under both learning paradigms, facilitating a systematic comparison of their performance across different settings and providing insights into their robustness in the absence of labeled anomaly data.

3.2.2.1 Transformer-based Models

Originally introduced by Vaswani et al. (2017), transformers have become state-of-the-art models across several domains, including natural language processing (Brown et al., 2020), speech recognition (Kim et al., 2022), and computer vision (Liu et al., 2021). Recently, transformer-based architectures have also shown great promise in anomaly detection (Tuli et al., 2022; Xu et al., 2021), and in related applications like fault detection (Maldonado-Correa et al., 2024).

Architecturally, transformers are built upon attention mechanisms organized within an encoder-decoder framework. Their attention layers enable them to capture global dependencies by dynamically focusing on different parts of input sequences, regardless of sequence length. In contrast to recurrent neural networks (RNNs) and their improved variants, transformers are capable of modeling long-range temporal relationships without suffering from the vanishing gradient problem—a well-known limitation of RNN-based architectures. Additionally, transformers can process all elements of an input sequence simultaneously, offering a key advantage in parallelization. This parallelism significantly accelerates both training and inference, particularly when leveraging modern hardware like graphics processing units (GPUs) (Vaswani et al., 2017). These preferred characteristics are largely attributed to the so-called multi-head attention mechanism, serving as the core part of transformer architectures. In the following, we delve more into this intriguing mechanism.

Multi-Head Attention Mechanism. The mechanism begins by transforming the positionally encoded input sequence $X \in \mathbb{R}^{N \times d}$—with N d-dimensional tokens—into three distinct representations: the value (V), key (K), and query (Q) matrices. This transformation is accomplished via linear projections using three separate learnable weight matrices, $W^V \in \mathbb{R}^{d \times d_k}$, $W^K \in \mathbb{R}^{d \times d_k}$, and $W^Q \in \mathbb{R}^{d \times d_k}$, as follows:

$$V = XW^V, \quad K = XW^K, \quad Q = XW^Q \tag{3.4}$$

where d_k represents the dimensionality of the embedding space.

The attention mechanism comprises multiple parallel attention heads, each following the same computational structure but employing independent sets of weight matrices. This design enables the model to project the input sequence into diverse representation subspaces, allowing it to capture different types of contextual relationships and dependencies among the tokens.

Within each attention head, the output is a convex combination (i.e., a weighted sum) of the value (V) vectors, where the combination weights reflect the relative importance of other tokens in the sequence with respect to a given token. These weights are computed by applying the softmax function to the scaled dot product of the query and key matrices, as depicted in Fig. 3.3 and detailed in Eq. (3.5). The resulting attention scores emphasize the most relevant parts of the input sequence (the rows of the value matrix V) by assigning higher weights to tokens[2] deemed more contextually relevant. This mechanism enables the model to selectively aggregate information, thereby producing latent representations that effectively capture the underlying dependencies and interactions among subsequences within the input sequence, regardless of its length.

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}}) \cdot V \tag{3.5}$$

Finally, the outputs from all attention heads are concatenated and projected through another linear transformation to form the final output of the multi-head attention sub-layer. This aggregation allows the model to integrate various aspects of inter-token dependencies, thereby producing a richer and more expressive representation of the input sequence.

In this work, transformer-based models were employed for detecting APS failures, incorporating two distinct architectures. The first model is the vanilla architecture adapted for semi-supervised anomaly detection in time series data. In contrast, the second model is built upon a more sophisticated architecture, TranAD, introduced by Tuli et al. (2022) for time series anomaly detection. This model is distinguished by its impressive performance across a wide range of time series anomaly detection benchmarks. The detailed description of both models is provided below.

- **First architecture:** The architecture of the first transformer-based model is illustrated in Fig. 3.5. It consists of a single-layer encoder and a single-layer decoder. Given that the input sequences were normalized to the [0,1] range, a sigmoid activation function was applied at the output layer, following

[2] Here, a token refers to a single data instance constructed from features extracted via a sliding window.

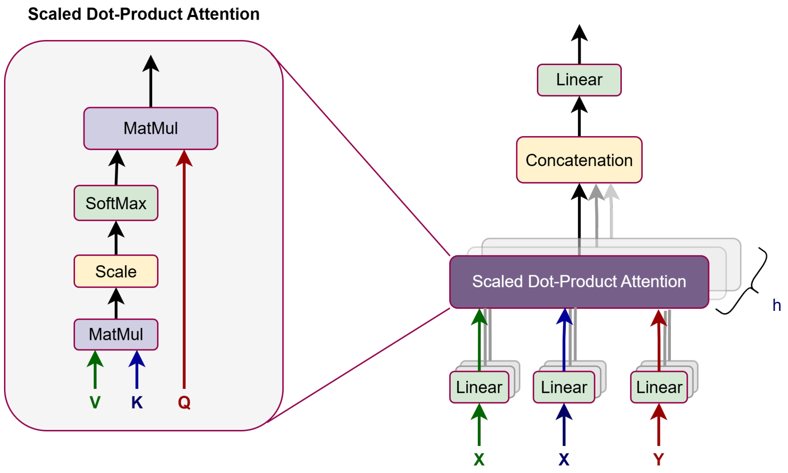


Figure 3.3 Multi-head attention mechanism. For the self-attention case, X and Y are the same, while they are different for the cross-attention case.

the recommendation by Tuli et al. (2022), as it typically leads to improved performance. The overall objective of the model is to reconstruct the input windows using latent representations learned through the multi-head attention mechanisms in both the encoder and decoder. Each input sequence is formed using 10 consecutive samples, where each sample comprises the statistical features derived from a sliding window, as will be detailed in Section 4.1.2. The overall loss function, serving as the anomaly score, is defined according to the following:

$$Loss = ||\hat{S}_t - S_t|| \tag{3.6}$$

where S_t refers to the input sequence and $\hat{S}_t$ denotes its reconstructed counterpart.

- **Second architecture:** The architecture of the TranAD model is illustrated in Fig. 3.5. It employs a transformer-based framework to reconstruct input sequences by leveraging attention-driven feature representations. Its core structure comprises an encoder and dual decoders operating in parallel. The encoder extracts global attention-based representations from the input sequence, capturing its contextual information. Meanwhile, the two decoders independently attempt to reconstruct subsequences of the original input based on these representations. In the current work, both decoders aim to reconstruct the current window using encoder-extracted representations from the past ten timestamps. To enhance sensitivity to subtle deviations indicative of anomalies, TranAD utilizes a two-stage adversarial training framework. During the initial phase, both decoders work to reconstruct the current window with max-

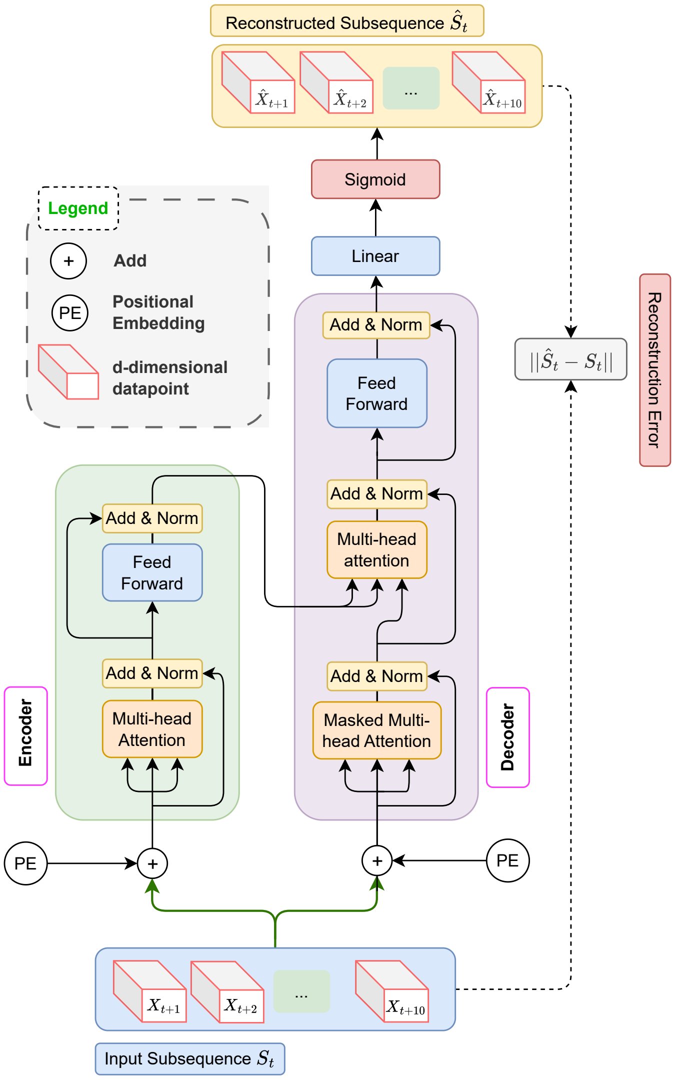


Figure 3.4 Vanilla transformer architecture (Farea et al., 2024). The input of the model is a sequence of 10 d-dimensional data points (d is 11 in this work).

imum accuracy. In the subsequent phase, the second decoder tries to reconstruct the current window while conditioning on the reconstruction error of the first decoder from the first phase. This reconstruction error is referred to as the *focus score* in Fig. 3.5. Such an adversarial mechanism directs the model's attention to short-term trends and magnifies poorly reconstructed subsequences, thereby improving the detection of anomalies. The composite loss function used for training the model is formalized as follows:

$$Loss = \frac{1}{n} \cdot ||O_1 - W||_2 + (1 - \frac{1}{n}) \cdot ||\hat{O}_2 - W||_2 \tag{3.7}$$

where O_1 and $\hat{O}_2$ represent the outputs of Decoder 1 in the initial phase and Decoder 2 in the subsequent phase, n refers to the training epoch, while W denotes the input window at the current timestamp.

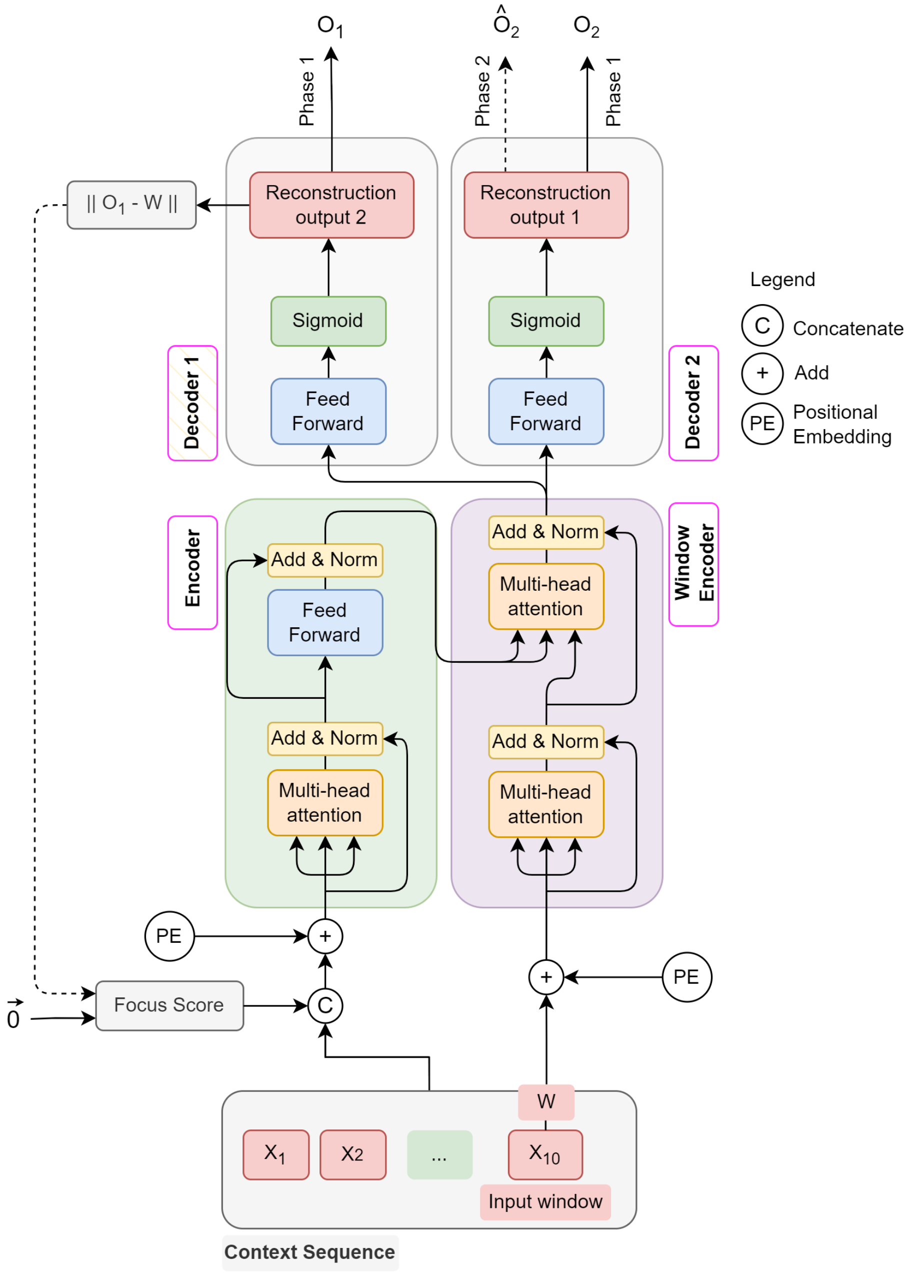


Figure 3.5 TranAD architecture (dotted lines refer to Phase 2) (Mumcuoglu et al., 2024b)

Since both transformer-based models are trained in a semi-supervised learning paradigm, they are expected to reconstruct normal windows with higher accuracy compared to anomalous ones. Consequently, the reconstruction errors of the input windows will serve as their respective anomaly scores. Nevertheless, from a practical standpoint, it is essential to derive a single anomaly score that reflects the status of the APS in each vehicle. To achieve this, for a given vehicle, the overall anomaly score is computed as the median of the anomaly scores across all windows associated with that vehicle. The median is preferred over the mean because of its known robustness to noisy scores and outliers.

3.2.2.2 Isolation Forest

Isolation Forest (iForest), introduced by Liu et al. (2008), is a tree-based ensemble algorithm designed for anomaly detection. Unlike classification-based or clustering-based anomaly detection approaches, iForest takes a fundamentally different perspective by isolating anomalies rather than profiling normal data. The underlying insight is that anomalies are few in number and different in characteristics. Thus, they are easier to isolate than normal instances. iForest is renowned for its computational efficiency, scalability to high-dimensional datasets, and its minimal reliance on hyperparameter tuning. Coupled with its unsupervised nature, these strengths have established iForest as a popular and robust baseline for anomaly detection across different domains, including cybersecurity, finance, and manufacturing.

The algorithm works by recursively partitioning the data space using randomly selected features and corresponding split values. A collection of binary trees, known as isolation trees (iTrees), is constructed, where each tree is grown by randomly selecting a feature and a random split value between the minimum and maximum values of that feature. This process continues until all instances are isolated or a predefined maximum tree depth is reached.

The path length of a data instance is defined as the number of edges traversed from the root node to the terminating node. Since anomalies tend to be isolated closer to the root of the tree due to their sparsity and distinctiveness of their feature values, they generally have shorter average path lengths across the ensemble of trees. Accordingly, the anomaly score for a data point x is computed based on the average path length $h(x)$, and is defined as the following:

$$S(x,n) = exp\,(\,-\frac{h(x)}{c(n)} \times ln2\,) \tag{3.8}$$

Here, $h(x)$ is the average path length of x over all grown trees, n is the number of data instances in the dataset, and $c(n)$ is the average path length of unsuccessful searches in a Binary Search Tree, approximated as

$$c(n) = 2H(n-1) - \frac{2(n-1)}{n} \tag{3.9}$$

where $H(i)$ is the i-th harmonic number.

According to Eq. (3.8), anomaly scores lie in the range (0,1], where values close to 1 indicate a high likelihood of anomaly. A user-specified threshold can be applied to classify data instances as normal or anomalous.

3.2.2.3 One-class SVM

One-class Support Vector Machine (one-class SVM) is a well-established unsupervised learning algorithm designed for anomaly detection. Unlike conventional binary or multiclass SVMs, which aim to separate labeled data into distinct categories, one-class SVM attempts to learn the decision boundary (a hyperplane) that separates the majority of the data from the origin with the maximum margin. The primary objective is to identify whether new instances deviate significantly from this distribution.

Mathematically, one-class SVM aims to find a decision boundary that optimally encloses the majority of the training data, possibly in a high-dimensional feature space. This is achieved by mapping the d-dimensional input data $X_i \in \mathbb{R}^d$ to a higher-dimensional feature space via a kernel function and then finding a hyperplane that maximally separates the origin from this transformed data. The decision function for a new data instance x is expressed as follows:

$$f(x) = \text{sign}(\sum_{i=1}^{n} \alpha_i K(x_i, x) - b) \tag{3.10}$$

Here, α_i is the Lagrange multiplier, x_i represents the i-th support vector, and b is the bias term. $K(x_i, x)$ is the kernel defining the dot product between the i-th support vector and the test data point x in the high-dimensional kernel (feature) space. As with the supervised SVM model, the RBF kernel was utilized for one-class SVM as well, owing to its capability to capture complex, non-linear relationships within the data. The RBF kernel is defined as follows:

$$K(x_i, x_j) = \exp(-\gamma ||x_i - x_j||^2) \tag{3.11}$$

where γ is the kernel width parameter, which controls the scale of the influence of support vectors.

In addition to γ, the parameter ν serves as a regularization hyperparameter that defines an upper bound on the fraction of training errors and a lower bound on the fraction of support vectors. Proper tuning of both γ and ν is critical to achieving an optimal balance between sensitivity to anomalous data instances and generalization to unseen data.

A new data instance x is considered either normal or an anomaly according to the decision function in Eq. (3.10). If $f(x) = 1$, then that data instance is deemed normal. Otherwise, the data instance is considered anomalous since $f(x) = -1$.

3.2.3 Unsupervised Approaches

This section introduces two unsupervised learning algorithms, namely DBSCAN and t-SNE, both of which were employed for preliminary defect detection in DED processes. DBSCAN functions as a density-based clustering algorithm to identify patterns and anomalies in the feature space, while t-SNE serves as a nonlinear dimensionality reduction technique for visualizing high-dimensional data in a lower-dimensional space.

3.2.3.1 DBSCAN

An unsupervised clustering approach based on DBSCAN was adopted for preliminary defect detection in DED processes. This choice of learning paradigm was necessitated by the nature of the associated dataset, which lacks ground-truth labels and thus precludes supervised and semi-supervised learning. The adopted methodology utilizes an ensemble of DBSCAN models to improve robustness and mitigate the sensitivity of the algorithm to hyperparameter selection. DBSCAN is a robust, density-based clustering algorithm that excels at identifying arbitrary-shaped clusters and detecting noise or outliers—making it highly suitable for anomaly detection tasks. Compared to other ML clustering algorithms such as K-means and self-organizing maps (SOMs), DBSCAN does not require prior specification of the number of clusters. Instead, it groups data into one or more high-density clusters while labeling low-density, unclustered points as outliers, which in this context correspond to potential defects. The DBSCAN algorithm relies on two key hyperparameters:

- **Eps (ϵ):** defines the radius of the neighborhood around each point. It is also known as the neighborhood search radius, serving as the key hyperparameter.
- **MinPts:** indicates the minimum number of points required within this radius to form a dense cluster. It specifies the minimum number of neighboring points for a core point.

A detailed explanation of the algorithm and guidelines for hyperparameter selection can be found in Ester et al. (1996). According to these two hyperparameters, the data points can be categorized into the following:

- **Core point:** a data point that has at least MinPts neighboring points (including itself) within its ϵ-neighborhood. Core points reside in the dense interior of a cluster and play a pivotal role in defining and expanding cluster structures, as they can directly reach other core points as well as border points.

- **Border point:** a data point that falls within the ϵ-neighborhood of a core point but does not itself have enough neighboring points (i.e., fewer than MinPts) to qualify as a core point. Border points typically lie on the outer edges of clusters and are reachable from core points but do not contribute to expanding the cluster.
- **Noise point:** any other data point that is neither a border point nor a core point. Such points do not belong to any cluster and are considered outliers or anomalies. They are located in sparse regions of the data space with insufficient nearby points to be considered part of a cluster. In the context of anomaly detection, these noise points are of particular interest, as they often correspond to rare, defective, or abnormal patterns that deviate substantially from the general data distribution.

These three DBSCAN point categories are pictorially illustrated in Fig. 3.6 for visual clarity. The clustering performance of DBSCAN is highly sensitive to its two key hyperparameters: Eps and MinPts. Varying these parameters results in different clustering behaviors. In this work, four distinct values were selected for each of these hyperparameters (as will be discussed in Section 5.2.2.2), while using the Euclidean distance as the distance measure. This configuration yielded an ensemble of 16 DBSCAN models, each representing a unique parameter combination. To improve the robustness of anomaly detection in the absence of ground-truth labels, a majority voting strategy was employed. Under this ensemble scheme, a data point (i.e., an image) is classified as an outlier only if at least 9 out of the 16 models identify it as such. This majority voting approach mitigates the uncertainties inherent in working with unlabeled datasets and enhances the reliability of the clustering outcome. The underlying rationale is that the likelihood of multiple independent (or weakly dependent) models incorrectly labeling the same data point is relatively low; thus, consensus among the majority provides a more confident approximation of the ground-truth label.

3.2.3.2 t-SNE

t-SNE (Van der Maaten & Hinton, 2008) is a nonlinear dimensionality reduction technique widely used for exploratory data analysis and high-dimensional data visualization. It is particularly effective at mapping complex high-dimensional data into

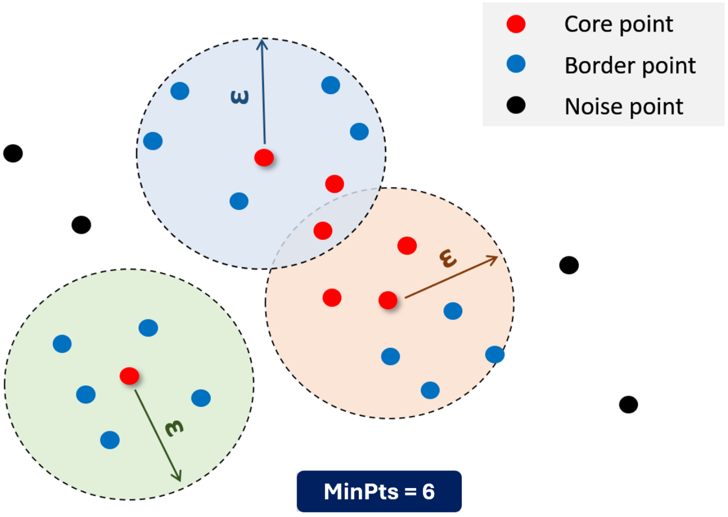


Figure 3.6 Overview of DBSCAN clustering. Data points are categorized into core, border, and noise (outliers) points.

a lower-dimensional space (typically a two-dimensional or three-dimensional space), while preserving the local structure of the original data space. The technique is an enhancement of Stochastic Neighbor Embedding (Hinton & Roweis, 2002), offering a more stable optimization process and improved visualization quality

Unlike linear methods such as PCA, which preserve global variance, t-SNE focuses on maintaining pairwise similarities between data points based on their probability distributions. In the high-dimensional space, t-SNE models the similarity between two points using a Gaussian distribution centered at each one. In the low-dimensional embedding, these similarities are modeled using a Student's t-distribution with one degree of freedom (a Cauchy distribution), which allows t-SNE to better handle the so-called "crowding problem" and to separate clusters more effectively.

The core idea behind t-SNE is to minimize the Kullback-Leibler (KL) divergence between the joint probability distributions of the high-dimensional data and their corresponding low-dimensional embeddings. This is achieved through an iterative optimization process using gradient descent.

In this thesis, t-SNE was employed as a postprocessing tool to support the interpretation of clustering results obtained through the DBSCAN algorithm. Feature vectors extracted from thermal images were projected into a two-dimensional space using t-SNE, thereby enabling a qualitative assessment of cluster separability and anomaly distribution. The provided visualization is assumed to approximate the underlying cluster topology in the original feature space. Consequently, t-SNE plays a complementary role in validating the clustering structure identified by DBSCAN and offers valuable insights into the latent organization of the unlabeled dataset.

3.3 Explainable AI for Anomaly Detection

Explainable AI (XAI) plays a critical role in enhancing the interpretability and trustworthiness of anomaly detection systems. By integrating XAI into the modeling pipeline, these systems not only identify anomalous behavior but also provide insights into the contributing factors of such events. This dual capability supports regulatory compliance and promotes safer, more transparent decision-making—an increasingly important consideration, especially in light of transparency mandates introduced by the EU AI Act in 2024 (European Commission, 2021).

XAI methods are classified into two categories: post-hoc and ante-hoc techniques. Post-hoc methods are applied after training to interpret the outputs of complex, opaque (black-box) models. Prominent examples include LIME (Local Interpretable Model-Agnostic Explanations) (Ribeiro et al., 2016,1) and SHAP (Shapley Additive Explanations) (Lundberg & Lee, 2017), which offer local and global interpretability by attributing feature contributions to individual predictions. In contrast, ante-hoc methods are inherently interpretable models that generate explanations by design. Classical models such as linear regression and decision trees fall into this category, though they often sacrifice predictive performance for transparency.

A notable exception is Explainable Boosting Machine (EBM), which achieves a strong balance between accuracy and interpretability. EBMs are inherently explainable models that leverage generalized additive models with modern enhancements, often rivaling the predictive power of black-box approaches like random forests and gradient-boosted trees (Nori et al., 2019). In this thesis, explainable frameworks incorporating EBM and SHAP were developed to enhance model transparency and enable interpretable anomaly detection across a real-world industrial application.

3.3.1 Explainable Boosting Machine (EBM)

EBM is an interpretable, glass-box model grounded in the framework of generalized additive models (GAMs). In GAMs, the prediction is formulated as a linear combination of functions applied individually to each feature, along with an intercept term. Analogous to linear regression, each feature function represents its independent contribution to the prediction. However, unlike linear regression, GAMs allow these functions to be non-linear, enabling the model to capture complex relationships between features and the target variable without sacrificing interpretability.

Building upon the classical GAM framework, EBM introduces key enhancements, notably the ability to model pairwise interactions between features. Furthermore, EBM leverages advanced ML techniques such as bagging and gradient boosting to learn the shape of each feature function. Specifically, each feature function is modeled through an iterative gradient boosting process, where a sequence of shallow decision trees is trained using only that individual feature and the residuals from the previous trees (see Fig. 3.7). To maintain invariance to the ordering of features, a very low learning rate is employed throughout training. In parallel, EBM automatically identifies and learns important pairwise feature interactions. Ultimately, the model's prediction is computed as a linear combination of the learned individual feature functions and pairwise interaction terms, formulated as follows:

$$g(E[y]) = a_0 + \sum_i f_i(x_i) + \sum_{i<j} f_{ij}(x_i, x_j) \tag{3.12}$$

In this formulation, a_0 represents the intercept term, x_i denotes the i-th feature, and $f_i(x_i)$ corresponds to the feature function learned for feature x_i. Similarly, $f_{ij}(x_i, x_j)$ captures the pairwise interaction function between features x_i and x_j. Both the individual feature functions and the interaction functions are constructed using an ensemble of weak learners, specifically shallow decision trees. The model further incorporates a link function $g(.)$, which adapts the EBM framework to various predictive tasks, such as regression or classification. In classification problems, the inverse link function $g^{-1}(.)$ typically corresponds to a sigmoid function for binary classification or a softmax function for multi-class settings. For regression tasks, $g^{-1}(.)$ simplifies to the identity function.

The relative importance and contribution of any feature x_i to the model's final prediction can be directly interpreted from the plot of its associated feature function $f_i(x_i)$. Likewise, the impact of pairwise feature interactions can be visualized through a two-dimensional heatmap representing the interaction function $f_{ij}(x_i, x_j)$. These visualizations facilitate the generation of both global and local model explanations in a straightforward and transparent manner.

Compared to widely used post-hoc explainability techniques such as SHAP and LIME, EBM offers several compelling advantages. EBM is inherently an end-to-end, glass-box model that achieves predictive performance comparable to advanced black-box algorithms, including random forests and gradient-boosting machines. Furthermore, it is distinguished by its computational efficiency at inference time, requiring only straightforward additive operations and table lookups within the learned feature and interaction functions (Nori et al., 2019). In contrast, techniques like LIME and

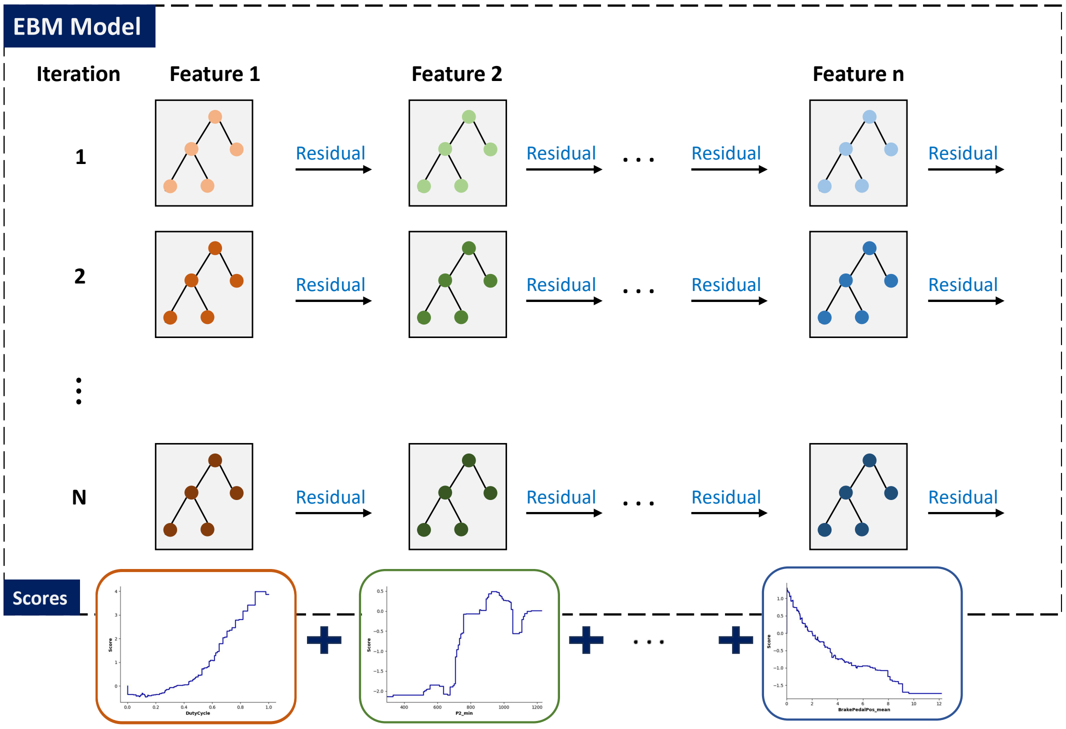


Figure 3.7 EBM architecture (Farea et al., 2025). It consists of shallow decision trees, each of which is trained on a single feature and the residual from the previous decision tree.

SHAP, while powerful, are often associated with substantial computational overhead due to the need for extensive perturbations and sampling-based approximations (Das et al., 2020). Motivated by these strengths, we adopt EBM as the backbone of our XAI-enabled framework for APS failure detection, enabling both accurate anomaly identification and transparent, real-time explanations of model predictions.

3.3.2 Shapley Additive Explanations (SHAP)

SHAP (Lundberg & Lee, 2017) represents a unified interpretability framework for interpreting the predictions of any (trained) black-box ML model. Rooted in game theory, SHAP offers a principled approach to model interpretability by constructing a simplified, inherently explainable surrogate model, known as an explanation model, that serves as an interpretable approximation of the original black-box model.

At the core of SHAP is the concept of additive feature attribution models, where the model output is expressed as a linear combination of the input features. For each individual prediction, SHAP computes an importance score—known as a SHAP value—for every input feature. These SHAP values quantify the contribution of each feature toward the final model output, offering a clear and consistent explanation of the prediction. Formally, for a prediction $f(x)$ of a data instance x, SHAP approximates the output as follows:

$$f(x) \approx g(x) = \phi_0 + \sum_{j=1}^{d} \phi_j \tag{3.13}$$

where $f(x)$ is the prediction of the original black-box model, $g(x)$ is the output of the simplified explanation model, and d is the number of input features. ϕ_0 is the expected value of the model output over the training distribution (the baseline prediction), and ϕ_j represents the contribution (or SHAP value) of the j-th input feature towards the model's prediction $f(x)$.

To compute SHAP values for a specific instance, the algorithm evaluates the marginal contribution of each feature by considering all possible combinations (coalitions) of features, assessing how the inclusion or exclusion of a given feature influences the model prediction. Although the exact computation of Shapley values is computationally expensive for models with many features (exponential in the number of features), various approximation methods, such as KernelSHAP and TreeSHAP, have been developed to make SHAP scalable to real-world datasets and complex models.

Overall, SHAP offers a robust, mathematically grounded approach for attributing model predictions to individual input features, thereby enhancing model transparency, supporting feature importance analysis, and enabling practitioners to build greater trust in ML models, particularly in high-stakes or regulated domains.

4. INDUSTERIAL APPLICATIONS

This chapter introduces the two real-world industrial applications: the air pressure system (APS) in heavy-duty vehicles (HDVs) and directed energy deposition (DED), an additive manufacturing process. For each application, a brief overview of the system is provided first to establish the operational context for that industrial system. This is followed by a description of the dataset(s) associated with each application, along with a detailed account of data preparation and preprocessing procedures, each tailored to the specific characteristics and requirements of the respective domain.

4.1 Air Pressure System (APS) in HDVs

The APS plays a key role in delivering pressured air to various HDV subsystems such as braking and suspension systems, with the electronic air processing unit (E-APU) serving as its central component. It combines an electronically controlled air drying mechanism with a multi-circuit valve, which allocates air to various vehicle circuits. The system consists of two main service braking circuits and an additional circuit for the parking brake and trailers, each equipped with a pressure sensor to continuously monitor pressure levels. The E-APU transmits pressure data along with a binary signal indicating the air compressor's status to the vehicle's electronic control units (ECUs) via the controller area network (CAN) bus. The internal design of E-APU ensures that a failure in one circuit does not compromise the entire braking system. A schematic representation of the APS, highlighting the central role of the E-APU, is provided in Fig. 4.1.

The E-APU electronically monitors the vehicle and engine status, facilitating an optimized compressor operation cycle. Air compression is reduced during high engine load conditions and increased during engine overrun phases to enhance fuel efficiency. It also schedules regeneration cycles to preserve air purity and quality, removing moisture and contaminants that could lead to corrosion, wear, or system malfunctions. In cases of faults—such as mechanical failure, overpressure, or fail-

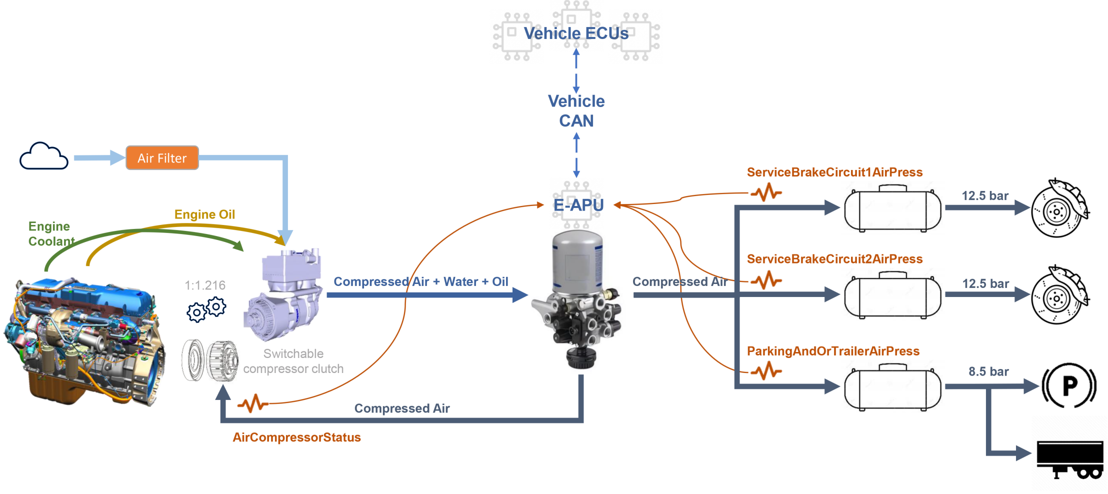


Figure 4.1 Overview of the air pressure system (Mumcuoglu et al., 2024b)

safe mode activation—the E-APU communicates the degraded operational status via the CAN bus.

Failures in the APS can stem from design flaws, manufacturing defects, or adverse operational conditions. Common issues include component wear caused by manufacturing defects, system overuse, or contaminated air supply. Moisture in the air supply, often due to air dryer malfunctions, can lead to corrosion in the compressor, valves, and air tanks. Additionally, harsh operating environments, such as extreme temperatures, and poor driving habits significantly contribute to the failure of APS components.

4.1.1 Dataset Description

The APS dataset comprises the operational data of cloud-connected Ford trucks. It was collected from these FMAX-branded trucks, operating across Turkey as well as other European countries. These vehicles were meticulously selected by experts to reflect a variety of conditions. They encompass different seasons, multiple times throughout the year, and a wide spectrum of mileages, providing a comprehensive representation of real-world usage scenarios. The dataset contains time series driving signals recorded over 30-day periods from two groups of vehicles. The first group includes 30 anomalous vehicles having experienced E-APU failures with subsequent E-APU replacements. In contrast, the other group consists of 110 healthy vehicles with a clean maintenance history. For anomalous vehicles, the dataset includes run-to-failure data, which captures daily driving records leading to the failure. On the other hand, for healthy vehicles, 30-day historical data sequences were selected from various periods throughout the year. In total, the dataset comprises 3,550 daily driving records from 140 vehicles. Further details are provided in Table 4.1.

Table 4.1 Description of the APS dataset

Details	**Healthy subset**	**Anomalous subset**
No. of vehicles	110	30
No. of daily records	2,779	771
No. of daily records per vehicle [min-max]	[15-30]	[17-30]
Avg. No. of datapoints per record	32,988	32,988
No. of drive cycles	18,556	5,552
No. of relevant signals	9	9
Nominal sampling rate	1 Hz	1 Hz

The raw dataset includes an extensive set of time series signals, capturing the overall vehicle dynamics, operational state, and certain environmental conditions. However, many of these signals are not directly relevant to the braking system. To focus on the APS, only nine APS-specific signals were selected, along with a DateTime signal that records the timestamp of each entry. All other signals were discarded. Table 4.2 provides a detailed list of these nine signals.

4.1.2 Data Preparation and Preprocessing

Data preparation and preprocessing constitute a crucial phase in data analytics pipelines, involving key steps such as data cleaning, segmentation, interpolation, and feature extraction. These steps are essential for maintaining data quality and supporting the subsequent task of model development. Such a preprocessing pipeline is particularly important in complex, hierarchical datasets, such as the operational data collected from HDVs.

The APS dataset consists of 140 vehicles in total, each contributing 15 to 30 daily driving records. Every daily record contains multiple drive cycles of varying durations. To structure the data effectively, the daily driving records were first divided into individual drive cycles where the temporal gap between any consecutive cycles is at least five minutes. This is because the absence of data logging for more than five minutes simply indicates that the vehicle was off.

Table 4.2 APS-specific signals ("On change" indicates that data logging is triggered only when the signal value changes.)

No.	**Signal**	**Sampling period** [seconds]
1	Air compressor status	on change
2	Brake pedal position	on change
3	Service brake circuit 1 air pressure	1
4	Service brake circuit 2 air pressure	1
5	Parking and/or trailer air pressure	1
6	Engine speed	1
7	Vehicle speed	1
8	Total traveled distance	10
9	Engine total hours of operation	300

The dataset suffers from missing data due to connectivity issues or inconsistent sampling rates across different signals. To address the missing data in each signal, we applied data imputation tailored to the characteristics of that signal. Based on the imputed data, moving statistics—such as mean, minimum, and standard deviation—were computed using sliding windows. This sliding window-based downsampling helps reduce the volume of raw data and enhance the extraction of informative features, e.g., duty cycle. Furthermore, it helps smooth out noisy signals, particularly when using moving averages. A visual representation of the preprocessing pipeline is provided in Fig. 4.2, illustrating how each data point in the preprocessed dataset corresponds to a set of features derived from sliding window-based moving statistics.

Key features were extracted through the application of sliding windows. These features were carefully designed based on expert knowledge, ensuring they effectively capture APS failures. They were meticulously handcrafted to serve as a reliable indicator of system anomalies. The extracted features are listed in Table 4.3, and they are as follows:

Duty Cycle: The duty cycle quantifies the air compressor's operational duration within the sliding window, expressed as a percentage of the total window duration. It serves as a crucial indicator of APS failures, as malfunctioning vehicles typically exhibit higher duty cycles compared to healthy ones. This increase occurs because faulty APS components disrupt the air supply, forcing the air compressor to operate for extended periods to compensate.

Air compressor on/off count: It measures the number of times the air compressor switches on and off within the respective sliding window. Elevated values of this frequency-based feature often indicate a fault in the APS, such as air leakage, which

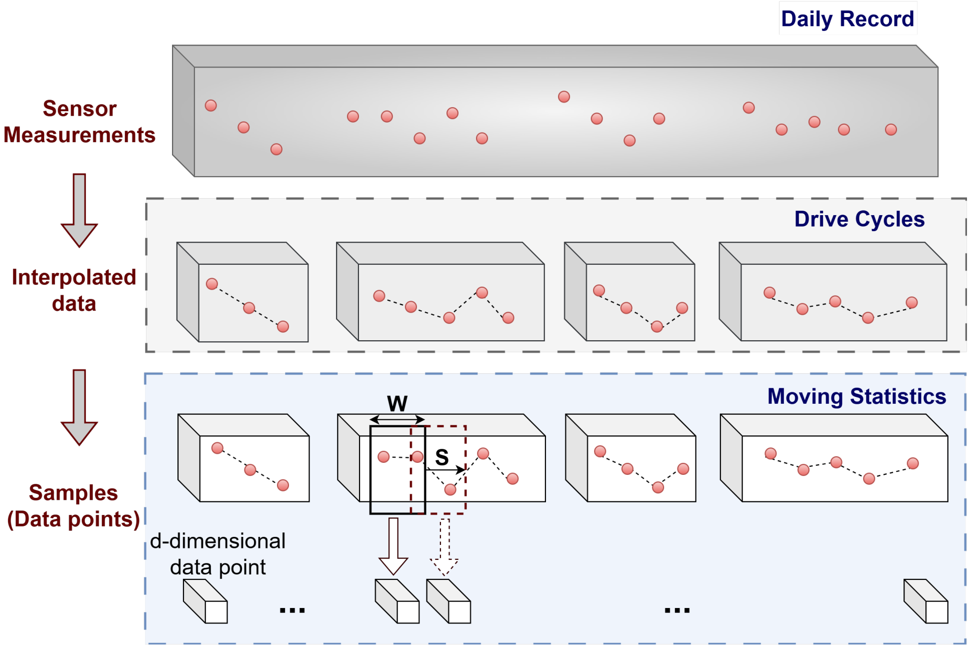


Figure 4.2 Data preparation and preprocessing pipeline (Farea et al., 2025). Firstly, the daily driving records are segmented into drive cycles, followed by the application of data imputation, and the extraction of moving statistics through sliding windows (**w** is the window length and **s** is the window shift)

prevents the system from maintaining optimal pressure levels across the circuits. As a result, the compressor operates more frequently to compensate for pressure loss. However, high values of this feature are not always a sign of failure. The air compressor also cycles on and off more frequently during the regeneration process, which ensures air quality by removing moisture and contaminants from the supply. Although this is a normal function, it occurs infrequently. Another normal scenario where frequent cycling is expected is during braking events, where pressure demands fluctuate rapidly. In summary, persistently high values of this feature across multiple sliding windows, without braking events, are considered anomalous and may signal an underlying APS failure.

Minimum and standard deviation of pressures: The minimum values and standard deviations of Service Brake Circuit 2 Air Pressure (P2) and Parking and/or Trailer Air Pressure (P3) were also included as key features. Since Service Brake Circuits 1 and 2 operate in parallel and are highly correlated, only Service Brake Circuit 2 was considered. Pressure irregularities often signal APS failures. For instance, air leakage causes lower minimum pressures and higher variability, as the system struggles to maintain stable pressure levels. Conversely, a faulty relief valve may lead to excessive pressurization, resulting in abnormally high minimum pressure values. Due to their sensitivity to such failures, minimum pressure values and standard deviations serve as key indicators of potential APS malfunctions.

Table 4.3 List of the features extracted from the APS dataset (using sliding windows)

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

Mean of Brake Pedal Position: Since the braking system relies on pressurized air from the APS, braking activity significantly influences APS behavior. Thus, the average brake pedal position serves as a crucial indicator for detecting APS failures. In both anomalous and healthy vehicles, an increased brake pedal position mean often correlates with higher duty cycles and more frequent compressor activations. This is expected during intensive braking when the demand for pressurized air is naturally higher. However, persistently elevated brake pedal engagement may signal aggressive driving habits, which can strain the APS and contribute to the wear of the overall vehicle system.

Mean and standard deviation of the engine speed: These statistical measures provide insights into the vehicle's dynamic behavior. For example, engine speed averages can indicate idling periods, while standard deviation serves as a key indicator for identifying irregular patterns such as aggressive driving or fluctuating vehicle loads. When such driving behaviors persist over extended periods, they can accelerate wear and contribute to failures across multiple vehicle subsystems, including the APS.

Standard deviation of the vehicle speed: It serves as a key indicator of variable driving conditions and behavioral patterns, much like its counterpart in engine speed. However, while engine speed variability primarily reflects internal vehicle dynamics, fluctuations in vehicle speed are more directly influenced by external factors. These include road conditions, traffic flow, route characteristics, and driver behavior—such as instances of aggressive driving, including rapid acceleration, harsh braking, and erratic speed changes.

4.2 Directed Energy Deposition (DED)

DED has emerged as a highly promising additive manufacturing technology, offering the ability to fabricate dense metal components with precise functional geometries and enhanced mechanical properties. Distinguished by its high deposition rates and optimized material utilization, DED presents a cost-effective solution for applications such as prototyping, repairing, and modifying metal parts (Wolff et al., 2019). Furthermore, its capability to support multi-material fabrication and manufacture large-scale structures (Dong et al., 2023) enhances its applicability across industries such as aerospace, automotive, and healthcare.

DED processes utilize computer-aided design (CAD) models to precisely guide a high-energy heat source, which melts and deposits material in a track-by-track, layer-by-layer fashion to incrementally build components. Typically integrated into a multi-axis computer numerical control (CNC) system, the heat source—commonly a laser, plasma arc, or electron beam—enables controlled deposition of feedstock, which is introduced in either powder or wire form. While DED can accommodate a variety of materials, including metals, ceramics, and composites, it is predominantly used for metal additive manufacturing, leading to its alternative designation as Directed Metal Deposition (DMD). When employing a laser-based heat source, it is also referred to as laser metal deposition (LMD), laser cladding, or Laser-Engineered Net Shaping (LENS). The process's versatility in handling diverse feedstocks and its precise control over deposition make it particularly well-suited for fabricating intricate geometries, repairing damaged metal parts, and enhancing component performance.

4.2.1 Dataset Description

This thesis addresses defect detection in DED processes using three distinct datasets. The first two datasets, hereinafter designated as DED-IN718 and DED-IN718-U, were curated as part of the thesis work and are based on the Inconel 718 (IN718) alloy. In contrast, the third dataset, hereinafter referred to as DED-Ti64, features the Ti-6Al-4V (Ti64) alloy. It was introduced by Zamiela et al. (2023) and is publicly available. The names of the datasets reflect the respective feedstock materials employed during fabrication. While DED-IN718 and DED-Ti64 are supervised datasets that include ground-truth labels, DED-IN718-U is an unsupervised dataset, distinguished by the absence of such labels.

4.2.1.1 IN718-based Datasets

The samples were fabricated using a LASERTEC 65 DED hybrid machine, which integrates a five-axis CNC milling system with a laser-based powder deposition unit, as depicted in Fig. 4.3. The deposition system utilizes the COAX14 coaxial nozzle to precisely deliver metallic powder from the hoppers to the substrate via a controlled feeding mechanism. The laser source is a fiber laser operating at a wavelength of 1064 nm, with a maximum power output of 2500 W. A 15 mm-thick C45 steel plate served as the substrate, providing ample space to accommodate all deposited samples without interference. To maintain an inert processing environment, argon gas was employed as both the carrier and shielding gas, preventing oxidation and ensuring optimal material properties.

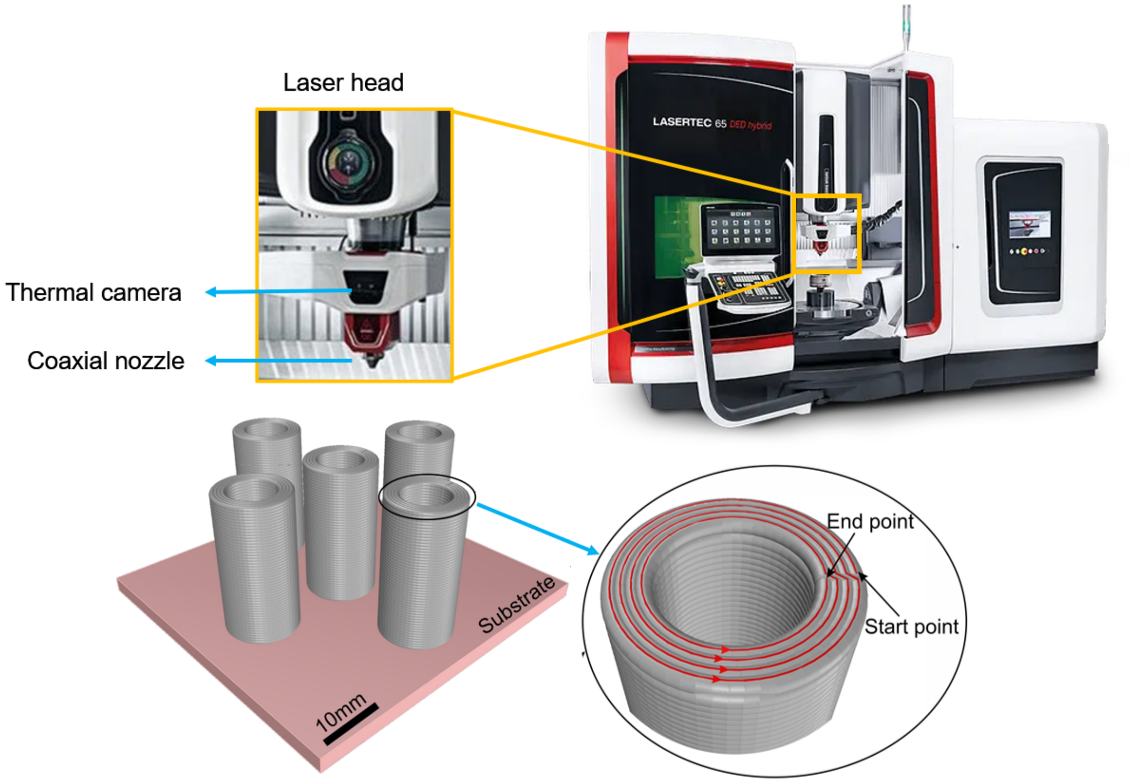


Figure 4.3 DED machine (DMG MORI, 2024) and sample configuration for DED-IN718 datsaset

The datasets consist of thermal images of the melt pool, captured by a thermal camera embedded within the deposition nozzle. This camera continuously monitors the melt pool, recording spatial temperature distributions that are crucial for analyzing defect formation mechanisms. The feedstock material is IN718, a nickel-chromium superalloy renowned for its exceptional mechanical strength, thermal stability, and corrosion resistance. These properties make IN718 a preferred choice

Table 4.4 Chemical composition (in percentage) of the IN718 alloy-based powder

Material	Ni	Fe	Cr	Ti	Al	Nb	Mo
Inconel 718	Balance	18	19	1	0.5	5	6

Table 4.5 DED process variables for the IN718-based datasets

Process Variable	**DED-IN718**	**DED-IN718-U**
Laser power (W)	400 - 1000	1500 - 2500
Laser spot diameter (mm)	1.6	3
Scan speed (mm/min)	{500, 700, 900}	750 - 1350
Powder flow rate (g/min)	12	15
Shield gas (l/min)	5	5
Carrier gas flow rate (l/min)	3	6
Layer thickness (mm)	0.97 ∼ 1.1	0.45

in high-performance applications, including aerospace engines and nuclear reactors (Ma et al., 2015; Zhang et al., 2021). In this research, the powder particles range in diameter from 70 to 120 μm, with their elemental composition detailed in Table 4.4.

For the DED-IN718 dataset, five multi-layer, multi-track cylindrical samples were fabricated using varying laser power and scan speed settings, to investigate the effects of different process parameters. The complete set of process parameters is outlined in Table 4.5. The laser power was initially set at 1000 W and gradually reduced to 400 W throughout the build direction, with a 25 W decrement every three successive layers. Each layer was constructed using a single continuous deposition track, which circled the cylindrical structure four times, as illustrated in Fig. 4.3. Due to the variations in process parameters, layer thickness was adjusted accordingly to ensure successful deposition. This diverse selection of parameter configurations was deliberately implemented to enhance the dataset's comprehensiveness, capturing the dynamic behavior of the deposition process across different operating conditions.

The DED-IN718 dataset comprises a total of 7,490 thermal images, each captured at a standard resolution of 164 × 218 pixels. In contrast, the DED-IN718-U dataset consists of 2,894 thermal images in total, each with a standard resolution of 164 × 218 pixels. Nonetheless, some of these images correspond to inter-layer transition periods when the laser was inactive, resulting in frames without a visible melt pool. After filtering out these melt pool-free images, the DED-IN718 dataset contains 4,889 valid thermal images, while the DED-IN718-U dataset includes 2,295 valid thermal images, as detailed in Table 4.6. Fig. 4.4 presents sample thermal images, where the heated melt pools can be observed at the center of each frame, highlighting the region of active material deposition.

Table 4.6 Overview of the IN718-based datasets

Data set	**DED-IN718**	**DED-IN718-U**
Collected images	7490	2894
Invalid images	2601	599
Valid images	4889	2295

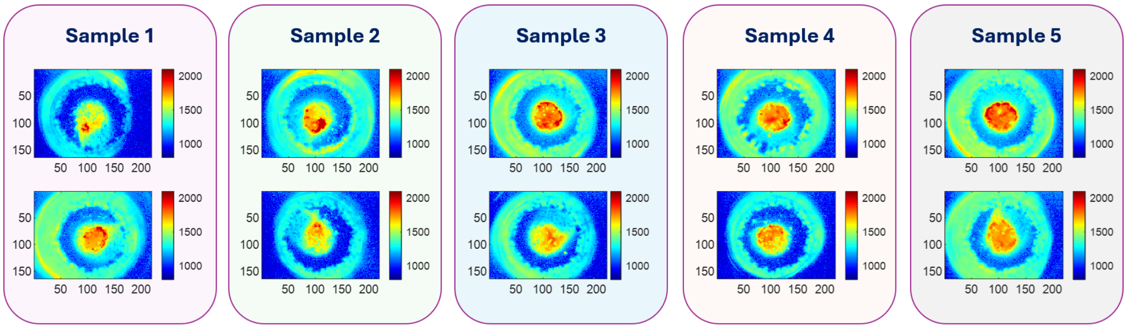
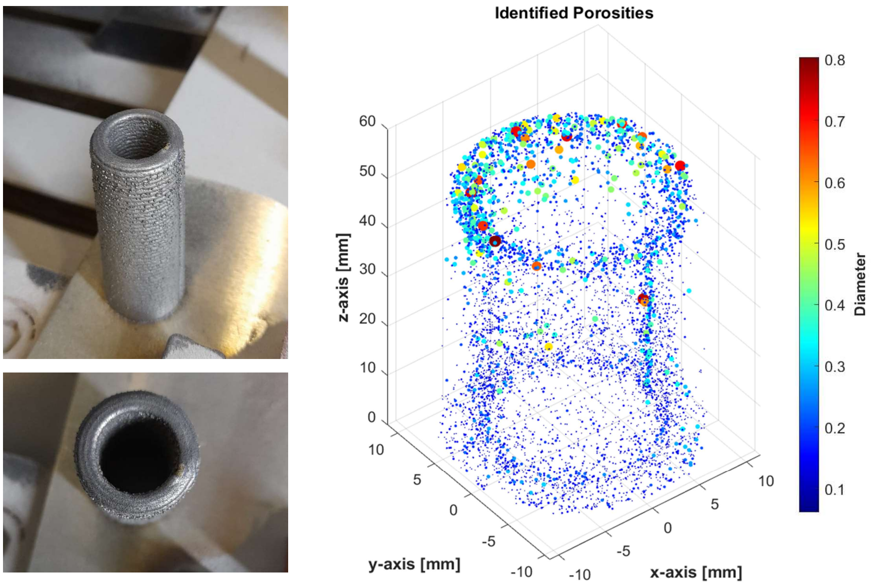


Figure 4.4 Representative thermal images from the DED-IN718 dataset. Two thermal images are shown for each deposited sample.

For the DED-IN718 dataset, following deposition, the characterization of porosities in all fabricated samples was conducted using a Phoenix V micro-computed tomography (μCT) scanner equipped with a 300 kV micro-focus X-ray tube. To ensure adequate penetration and high-resolution defect detection, the scans were performed at 240 kV and 220 A. The voxel size was set to 10 μm, optimized based on the sample dimensions. A full 360° rotation consisted of 2,500 projection images, determined according to the manufacturer's guidelines and voxel size considerations. The machine sensitivity was configured to 2×2, with a frame resolution of 2300 × 2300 pixels. Each image was captured with an exposure time of 334 ms. To mitigate the beam-hardening effect, a 500 μm copper filter was applied, allowing for higher-energy X-ray penetration. For comprehensive porosity characterization, VGStudio MAX (Volume Graphics, 2023) was used to process the μCT scan data, with a minimum defect detection threshold of 10 μm. According to the analysis, 2,268 thermal images (46.4%) were associated with porosities, whilst the remaining 2,621 images (53.6%) were classified as defect-free.

One of the fabricated samples is shown in Fig. 4.5. In the same figure, the internal porosities, detected via X-ray μCT, are displayed as well. The identified porosities vary in size, ranging from 60 μm to 0.8 mm in diameter. Conversely, Fig. 4.6 displays an example thermal image of the melt pool, alongside its respective temperature distribution. The temperature histogram exhibits a multimodal profile, indicating distinct thermal zones. The first peak, observed around 1180 °C, corresponds to the



(a) Deposited sample

(b) Porosities distribution

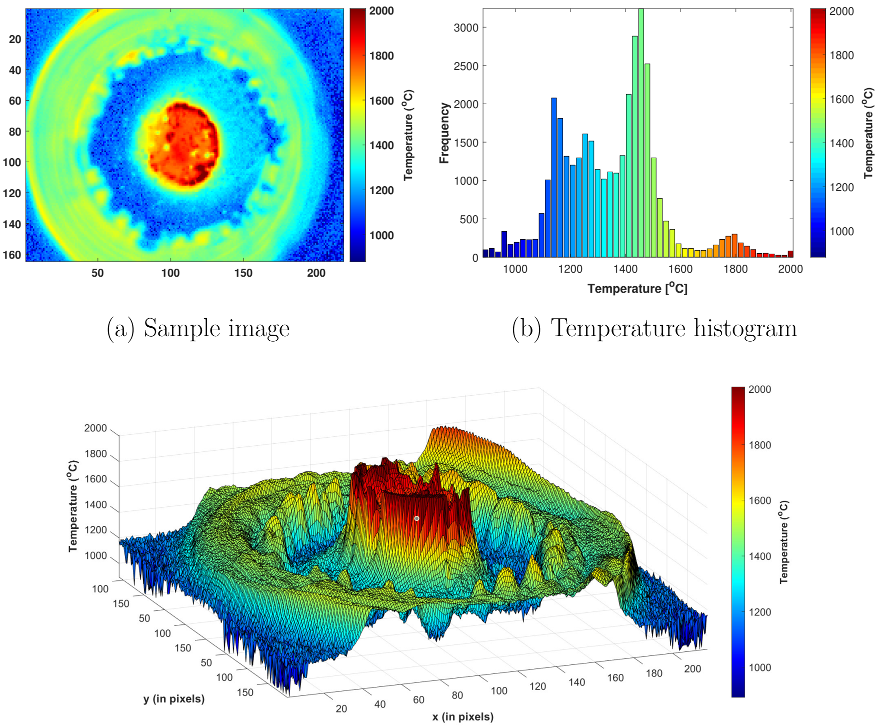
Figure 4.5 One sample with its X-ray CT-identified porosities. The porosities were color-coded according to their diameters.

background temperature. The second peak, approximately 1450 °C, represents the heat-affected zone, including the nozzle tip and the melt pool boundary. Finally, the third peak, near 1800 °C, signifies the core region of the melt pool.

4.2.1.2 DED-Ti64 Dataset

The DED-Ti64 dataset, introduced by Zamiela et al. (2023), consists of in-situ thermal images and post-process porosity annotations for Ti-6Al-4V thin-walled structures manufactured using the OPTOMEC Laser Engineered Net Shaping (LENS™) 750 DED system. Thermal imaging was performed using a Stratonics dual-wavelength pyrometer, capturing top-down melt pool and heat-affected zone views at temperatures exceeding 1660 ^{o}C. The original pyrometer images (752×480 pixels) were cropped to 200×200 pixels centered around regions above 1000 ^{o}C to isolate the melt pool. Post-fabrication, internal porosity was characterized using a Nikon XT H225 X-ray computed tomography (XCT) system, while the MyVGL Studio MAX DefX algorithm was applied to the volumetric XCT data to quantify and localize porosity within the fabricated structures.

The dataset includes a total of 1,564 thermal images in CSV format, each with a resolution of 200 × 200 pixels (some representative thermal images are shown



(a) Sample image

(b) Temperature histogram

(c) 3D temperature distribution

Figure 4.6 Sample thermal image and its temperature distribution. The temperature distribution is viewed as both a 2D histogram and a 3D plot.

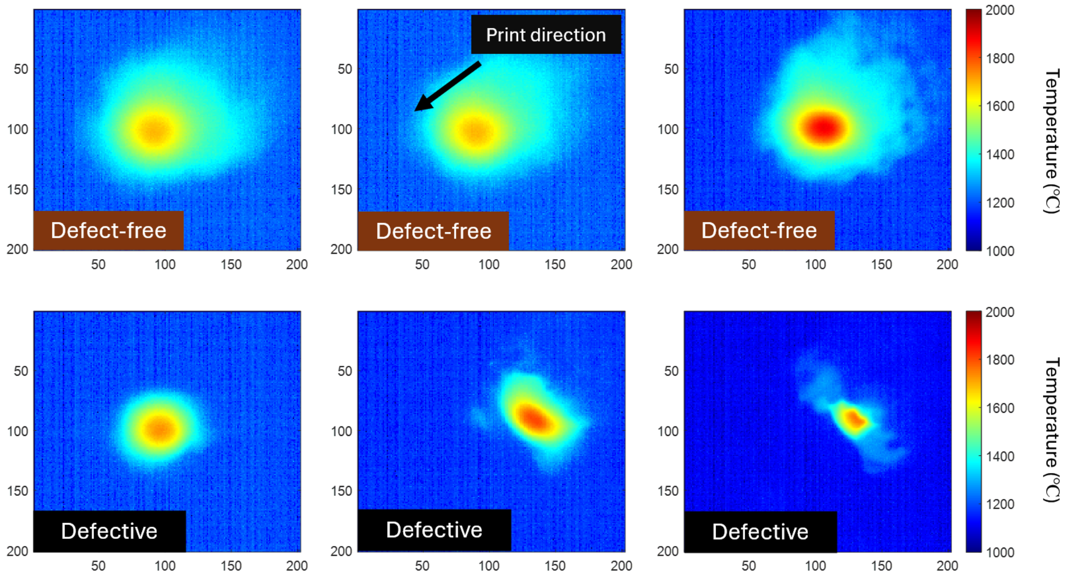


Figure 4.7 Representative thermal images from the DED-Ti64 dataset

in Fig. 4.7). Each image is annotated with a binary label indicating the presence (1) or absence (0) of porosity, based on the XCT analysis. The dataset is notably imbalanced, comprising 1,493 non-defective samples and 71 defective ones—resulting in a class imbalance of approximately 4.5%. In addition to the binary labels, each thermal frame is supplemented with metadata including timestamp, frame number, spatial coordinates, melt pool features, and porosity size where applicable (ranging from 0.05 mm to 0.98 mm). This dataset provides a valuable benchmark for defect detection in metal additive manufacturing, particularly in the context of highly imbalanced and high-temperature sensor data.

4.2.2 Data Preparation and Preprocessing

The main objective of data preparation and preprocessing is to convert the raw thermal images into compact feature representations that capture the essential characteristics of the melt pools. These features serve as a concise and informative abstraction of the thermal dynamics, facilitating efficient and structured modeling. The proposed preprocessing pipeline effectively transforms unstructured image data into a structured tabular form, thereby enabling the application of static ML models such as tree-based ensembles and kernel-based methods, which perform optimally on well-defined feature spaces.

In its raw form, the collected DED-IN718 and DED-IN718-U data represent the radiation intensities emitted from the melt pool, rather than direct temperature measurements. To obtain accurate thermal readings, these intensities are first converted into temperature values using a calibration file provided by the manufacturer. This calibration file contains a lookup table that maps each recorded radiation intensity to its corresponding temperature, ensuring precise thermal characterization of the melt pool.

An overview of the proposed preprocessing pipeline for the three datasets is shown in Fig. 4.8. The pipeline includes region-of-interest (RoI) extraction, segmentation, feature extraction, feature selection, and finally normalization. More details about these steps are provided in the following subsections.

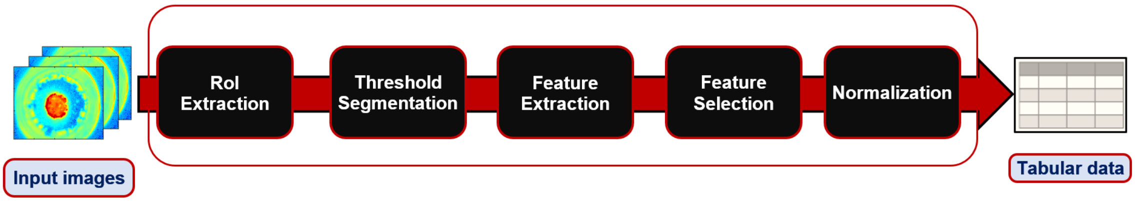


Figure 4.8 Proposed preprocessing pipeline for defect detection

4.2.2.1 RoI Extraction

For the DED-Ti64 dataset, the raw thermal images acquired via the dual-wavelength pyrometer had a resolution of 752 × 480 pixels. However, each image was cropped to 200 × 200 pixels to isolate the region of interest, which focused on the melt pool and surrounding heat-affected zone. In contrast, for the DED-IN718 dataset, the region of interest was isolated in each thermal image through the application of circular masking. This masking process preserved only the pixels within a circular area centrally aligned with the image while setting all external pixels to zero, as illustrated in Fig. 4.9(a). By focusing on the central region, this image processing method ensured that the melt pool, which is consistently positioned at the center of thermal images, remained in the analysis while excluding background noise and the nozzle tip. As shown in Fig. 4.9, the nozzle tip frequently exhibits temperature levels comparable to the melt pool; therefore, the exclusion of the nozzle tip is critical to prevent interference in subsequent analyses. Accordingly, the RoI extraction step proposed here plays a fundamental role in enhancing the accuracy and efficiency of defect detection by concentrating only on the most relevant parts of thermal images and discarding the irrelevant image regions.

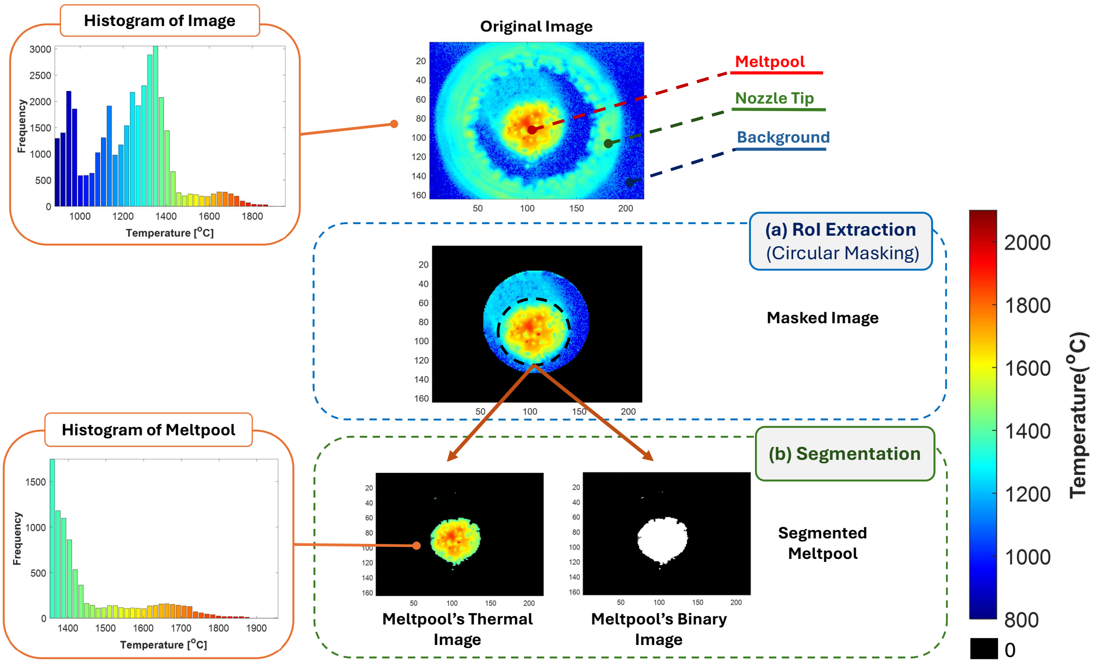


Figure 4.9 Thermal image preprocessing: (a) Region-of-interest (RoI) extraction via circular masking, and (b) segmentation of RoI into the segmented binary and thermal images of the melt pool (histograms of the input image and melt pool are shown on the left)

4.2.2.2 Threshold Segmentation

Following RoI extraction, the melt pool was segmented using a threshold-based segmentation approach. For the DED-based datasets, a prespecified temperature threshold around 1330 °C, corresponding to the approximate melting point of IN718 alloys (ESPI Metals, 2024), was employed to isolate the melt pool region. Similarly, for the DED-Ti64 dataset, a threshold of 1640 °C was applied, reflecting the estimated melting point of Ti-6Al-4V alloys. The largest connected region within the image exhibiting temperatures at or above this threshold was identified and segmented. The resulting segmentation output, illustrated in Fig. 4.9(b), includes both binary and thermal versions of the melt pool. This dual segmentation strategy serves two key purposes: The binary segmentation facilitates the extraction of melt pool geometry and shape-related features. On the other hand, the thermal segmentation provides the foundation for histogram-based color and texture feature extraction, as elaborated in the next subsection. By leveraging both binary and thermal representations, this methodology ensures a comprehensive characterization of the melt pool, focusing exclusively on its critical attributes while eliminating irrelevant background noise. This approach enhances the precision of feature extraction, enabling an accurate assessment of melt pool dynamics and potential process anomalies.

4.2.2.3 Feature Extraction

IN718-based Datasets: This work utilizes two distinct categories of features: melt pool-related features and process-related features. The melt pool features, extracted directly from each thermal image, are further classified into three subgroups: shape, color, and texture. These features capture critical aspects of the melt pool's geometry and thermal distribution. Meanwhile, the process-related features provide context and insights into the process conditions that influence melt pool behavior. A comprehensive list of all extracted features for the DED-IN718 and DED-IN718-U datasets is presented in Table 4.7, and they are explained in detail as follows:

(a) Shape Features: They encapsulate the geometric characteristics of the melt pool. These features serve as critical indicators of potential defects, as irregularities in melt pool geometry often correlate with underlying defect formation during the deposition process. The extracted shape features from the melt pool include:

- **Area:** represents the area of the melt pool region. It is determined as the total number of pixels in that region.

Table 4.7 List of extracted features from the IN718-based datasets (features marked with * are extracted from both DED-IN718 and DED-IN718-U, while the remaining features are exclusive to the DED-IN718 dataset)

Melt pool-related features			**Process-related features**
Shape features	**Color features**	**Texture features**	
Area *	Mean *	Contrast *	Sample ID
Perimeter	Standard deviation *	Correlation *	Scan speed
Major axis length *	Skewness *	Energy *	Laser power
Minor axis length	Kurtosis *	Homogeneity *	Cylindrical coordinates (ρ, θ, z)
Equivalent diameter	Entropy		
Circularity *	Mode		
Eccentricity *	Median (Q2)		
Solidity	1st quartile (Q1)		
Orientation	3rd quartile (Q3)		

- **Hole area:** is the total number of pixels within the melt pool region that have temperature values less than the predefined threshold temperature (1330 °C).

- **Perimeter:** represents the total length of the melt pool's boundary.

- **Major/Minor axis length:** is the length (in pixels) of the major/minor axis of the equivalent ellipse. The equivalent ellipse is defined as an ellipse whose second moments are equal to those of the melt pool region.

- **Equivalent diameter:** is the diameter of the equivalent circle. The equivalent circle is defined as a circle whose area is equal to the melt pool's area.

- **Circularity:** represents the roundness of the melt pool region and it is calculated according to the following formula:

  $$Circularity = \frac{4\pi \times Area}{(Perimeter)^2} \tag{4.1}$$

  where it is between 0 and 1, with 1 corresponding to a perfect circle.

- **Eccentricity:** is the eccentricity of the equivalent ellipse and it is calculated according to the following formula:

  $$Eccentricity = \sqrt{1 - \frac{b^2}{a^2}} \tag{4.2}$$

  Here, a and b represent the lengths of the major and minor axes. The eccentricity value ranges from 0 to 1, where 0 indicates a perfect circle and 1 represents a degenerate ellipse, or a line.

- **Orientation:** is the angle (in degrees) between the major axis of the equivalent ellipse and the x-axis. It ranges from -90^o to 90^o.

- **Solidity:** it is defined according to the following formula:

  $$Solidity = \frac{A1}{A2} \tag{4.3}$$

  where A1 is the area of the melt pool and A2 is the area of the smallest convex polygon that completely encloses the melt pool, both measured in pixels.

(b) Color features: They are derived from the temperature histogram of the melt pool (refer to Fig. 4.9), providing insights into its temperature distribution. These features are instrumental in identifying anomalous images that may indicate the presence of defects. The extracted color features are outlined in Table 4.7. These features include various statistical measures, such as mode, median, and first and third quartiles. The other used statistics are defined as follows:

$$Mean\,(\mu) = \frac{1}{n} \sum_{(i,j)\in R} I(i,j) \tag{4.4}$$

$$Std.\,(S) = \sqrt{\frac{1}{n-1} \sum_{(i,j)\in R} [I(i,j)-\mu]^2} \tag{4.5}$$

$$Skewness = \frac{1}{n\,S^3} \sum_{(i,j)\in R} [I(i,j)-\mu]^3 \tag{4.6}$$

$$Kurtosis = \frac{1}{n\,S^4} \sum_{(i,j)\in R} [I(i,j)-\mu]^4 \tag{4.7}$$

$$Entropy = -\sum_i p_i\, log_2\, p_i \tag{4.8}$$

where R denotes the extracted region containing the melt pool with a total of n pixels, $I(i,j)$ represents the temperature value at the (i,j) pixel, and p_i refers to the relative frequency of the i-th bin within the histogram of R.

(c) Texture Features: The texture features are extracted from the melt pool region using the Gray-Level Co-occurrence Matrix (GLCM). This matrix analyzes the spatial relationships between pixel pairs in the segmented thermal version of the melt pool (see Fig. 4.9). It is constructed by evaluating how frequently pairs of pixels with specific temperature values occur at a given distance and orientation relative to each other. This approach captures how pixel values vary in relation to their neighbors, thereby providing texture information on the melt pool region. The spatial configuration of these pixel pairs is governed by two user-defined parameters: the distance between the pixels and the orientation angle at which they are examined. The following features are extracted from the formed GLCM:

- **Energy:** is the sum of the squared elements of the GLCM matrix according to the following formula:

  $$Energy = \sum_{i,j} [M(i,j)]^2 \tag{4.9}$$

  where $M(i,j)$ represent the (i,j) element in the GLCM matrix.

- **Correlation:** indicates the joint probability occurrence of the specified pixel pairs and is computed according to the following formula:

  $$Correlation = \sum_{i,j} \frac{(i-\mu_i)(j-\mu_j) \times M(i,j)}{\sigma_i \, \sigma_j} \tag{4.10}$$

  where μ_i represent the mean of the i-th row with standard deviation σ_i, and μ_j is the mean of the j-th column with standard deviation σ_j.

- **Contrast:** quantifies the local variation in the GLCM matrix, and is calculated as follows:

  $$Contrast = \sum_{i,j} |i-j|^2 \times M(i,j) \tag{4.11}$$

- **Homogeneity:** measures the similarity between the distributions of the GLCM diagonal and the GLCM elements. It is calculated according to the following formula:

  $$Homogeneity = \sum_{i,j} \frac{M(i,j)}{1+|i-j|} \tag{4.12}$$

(d) Process-related Features: In addition to the abovementioned features directly associated with the melt pool, process-related features are also incorporated to provide context for the thermal images—linking them to the respective process parameters, the fabricated sample, and spatio-temporal information inside each sample. The process parameters include laser power and scan speed, whilst spatio-temporal information is determined based on the cylindrical coordinates of the fabricated sample in hand. The cylindrical coordinates reflect the sample's cylindrical geometry more accurately as opposed to the Cartesian coordinates (refer to Fig. 4.5(a)). For each thermal image, the process-related features offer critical contextual insight complementing the melt pool-related features which capture the fundamental characteristics of the melt pool's thermal distribution and geometry. The process-related features establish a connection between each thermal image and its manufacturing environment by incorporating spatial, temporal, and process parameter information. This holistic approach is essential for understanding defect formation. For example, defects are more likely to occur in upper layers and inner tracks than in lower layers and outer tracks, due to the cumulative effects of vertical and lateral heat buildup (see Fig. 4.5(b)). By integrating these complementary feature sets, the model gains a more holistic understanding of the manufacturing process, enhancing its ability to detect defects with greater accuracy.

DED-Ti64 Dataset: As with the DED-IN718 dataset, key shape and color features are extracted from the DED-Ti64 dataset. In addition to these core features, several supplementary shape, color, and texture features are extracted as well. Notably, gradient-based texture features are computed to quantify the spatial temperature transitions within the thermal images, capturing the rate of change across the melt pool in both horizontal and vertical directions. A comprehensive list of all extracted features for the DED-Ti64 dataset is presented in Table 4.8.

4.2.2.4 Feature Selection

The extracted features may exhibit dependencies, with some features expected to be correlated. To address this, a Pearson correlation test was performed as a preprocessing step to assess the correlation among these features. This test evaluates the linear pairwise correlation between each feature pair. It is applied to all feature pairs to construct the correlation matrix, which is defined as follows:

$$C(i,j) = \frac{Cov(f_i, f_j)}{\sigma_i \, \sigma_j} \tag{4.13}$$

where $C(i,j)$ is the correlation coefficient between the i-th and j-th features. $Cov(f_i, f_j)$ represents the covariance between the i-th feature, with standard deviation σ_i, and the j-th feature, with standard deviation σ_j. The correlation coefficient values range from -1 to 1, where the magnitude indicates the strength of the correlation, and the sign indicates the direction of the relationship.

Table 4.8 List of the features extracted from the DED-Ti64 dataset

Shape features	**Color features**	**Texture features**
Area	Mean	Gradient mean
Perimeter	Standard deviation	Gradient maximum
Maximum	Variance	Gradient standard deviation
Minimum	Skewness	
Major axis length	Kurtosis	
Minor axis length	Median (Q2)	
Orientation	1st quartile (Q1)	
Circularity	3rd quartile (Q3)	
Eccentricity	Interquartile range (IQR)	
Aspect ratio		
Centroid coordinates		
Peak temp coordinates		

4.2.2.5 Normalization

The input features vary in scale, which causes features with higher scales to dominate those with lower scales. To address this, normalization is performed by subtracting the mean and dividing by the standard deviation of each feature, as follows:

$$f_i^{'} = \frac{f_i - \mu_i}{\sigma_i} \tag{4.14}$$

where f_i is the i-th feature with mean μ_i and standard deviation σ_i while $f_i^{'}$ is its normalized version.

This normalization method, commonly referred to as Z-score normalization or standardization (see Cabello-Solorzano et al. (2023)), transforms all features into a consistent scale, ensuring a mean of 0 and a standard deviation of 1. By standardizing the features, the influence of each feature during training is balanced, which typically leads to notable improvements in both the accuracy and numerical stability of machine learning models.

5. EXPERIMENTAL RESULTS

This chapter presents the experimental validation of data-driven anomaly detection frameworks applied to the two real-world industrial applications described in the previous chapter. Initially, the results for failure detection in the air pressure system (APS) are detailed and examined. Subsequently, the chapter discusses and analyzes the defect detection results within directed energy deposition (DED) processes.

5.1 Detection of APS Failures

This thesis explores various frameworks for APS failure detection, spanning supervised, semi-supervised, and explainable approaches. This section presents and thoroughly discusses the results obtained from these frameworks, with particular emphasis on the most promising ones—the EBM-based and transformer-based frameworks. Through detailed comparative evaluation, their strengths, limitations, and practical applicability in APS failure detection are critically examined.

5.1.1 Supervised Learning

The supervised models utilized in this thesis for APS failure detection include the black-box models, namely Random Forest and XGBoost, as well as the explainable AI (XAI) models: EBM and SHAP. EBM functions as an ante-hoc XAI model, generating both predictions and interpretable explanations simultaneously. In contrast, SHAP is a post-hoc XAI model that provides interpretability by being trained on top of existing black-box models to explain their predictions. In this work, SHAP was applied to fit an explainer to Random Forest, enabling a detailed interpretation of its predictions and serving as a baseline for assessing the explainability of the EBM model.

To ensure a comprehensive evaluation, the methodology for assessing model performance is first outlined. This is followed by a detailed presentation and analysis of the classification results for each supervised model. Finally, the interpretability outcomes generated by the XAI models are presented and thoroughly discussed.

5.1.1.1 Evaluation Approach

The performance of the supervised models was evaluated using a stratified five-fold cross-validation scheme, as illustrated in Fig. 5.1. Unlike holdout validation, which assesses the model using only a subset of the dataset, this approach ensures that the entire dataset is tested, thereby providing a more comprehensive evaluation of model performance. The APS dataset exhibited a notable class imbalance, with 79% of the data representing healthy vehicles (majority class) and 21% representing anomalous vehicles (minority class). To address this imbalance, stratified sampling was employed within the five-fold cross-validation, ensuring that each fold maintained the original class distribution. This approach prevents the minority class from being underrepresented in any fold, thus preserving the integrity of the evaluation process.

In the cross-validation process, the dataset was divided into five stratified folds, and each of the supervised models was trained five times, accordingly. At each run, one fold served as the testing set, whilst the remaining four folds were used for training. As depicted in Fig. 5.1, each testing fold received classification probabilities from the respective trained model. Specifically, $F_i \in \mathbb{R}^{\frac{N}{5} \times d}$ and $y_i \in \mathbb{R}^{\frac{N}{5}}$ represent the i-th fold and its associated model-provided probabilities where N denotes the total size of the dataset and d is the corresponding dimensionality (the number of features). $V_i \in \mathbb{R}^{n_i \times d}$ refers to the set of n_i data instances from the i-th vehicle, with their corresponding probabilities denoted by $\hat{v}_i \in \mathbb{R}^{n_i}$ and the vehicle-level anomaly score $s_i \in \mathbb{R}$. Since each vehicle consists of multiple observations, each has an individual classification probability; the probabilities of the testing folds were combined and then grouped by vehicle. While the classification probability for each observation serves as its anomaly score, a vehicle-level anomaly score is needed to provide a more meaningful assessment. To this end, the anomaly score for each vehicle was defined as the median of the classification probabilities across its observations, effectively capturing the overall likelihood of anomalous behavior for that vehicle and its APS. Finally, a one-dimensional grid search was conducted to identify the optimal probability threshold that maximized the F1 score. This optimized threshold was then applied to classify each vehicle as either Healthy or Anomaly, based on its aggregated anomaly score.

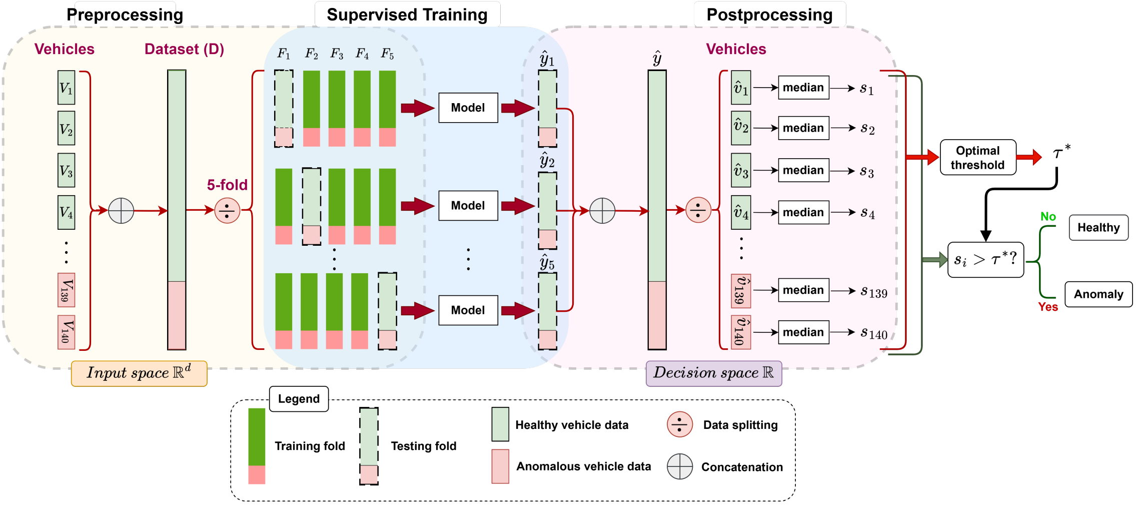


Figure 5.1 Validation scheme for evaluating supervised models (Farea et al., 2025)

5.1.1.2 Classification Results

Evaluation Metrics: The classification performance of the supervised models is summarized in Table 5.1. All three models were trained using the default values of their hyperparameters, adhering to the procedure outlined by Nori et al. (2019). Overall, the models achieved comparable results, with an accuracy of 91.4% and an F1 score of approximately 0.80. Notably, despite being a glass-box model, EBM demonstrated performance comparable to the more complex black-box models, Random Forest and XGBoost. These findings underscore EBM's ability to maintain high predictive accuracy in detecting APS failures while simultaneously providing interpretable explanations for its predictions. Unlike conventional glass-box models, which often trade off accuracy for interpretability, EBM effectively balances both, delivering robust performance without sacrificing transparency.

A statistical analysis was conducted to determine whether significant performance differences exist among the three supervised models—Random Forest, XGBoost, and EBM. The analysis was performed using the Kruskal-Wallis H test (Kruskal & Wallis, 1952), a non-parametric alternative to ANOVA that does not require assumptions of normality or homogeneity of variance, making it well-suited for the current setting. Each model was evaluated over five independent runs, and the statistical test was applied to each performance metric. The resulting p-values, summarized in Table 5.2, were above 0.1 for all evaluation metrics, indicating no statistically significant differences in performance among the models. This finding reinforces the comparable predictive capabilities of the glass-box EBM relative to the black-box models, demonstrating EBM's potential as a viable, interpretable alternative that maintains competitive performance without sacrificing transparency.

Table 5.1 Supervised learning results for APS failure detection

Supervised model	**Precision**	**Recall**	**F1 score**	**Accuracy**	**AUC**
Random Forest	0.82	0.77	0.79	91.4%	**0.892**
XGBoost	0.80	**0.80**	**0.80**	91.4%	0.886
EBM	0.80	**0.80**	**0.80**	91.4%	0.881

Table 5.2 Kruskal–Wallis H test results assessing statistical differences in performance among the supervised models

Metric	Precision	Recall	F1 score	Accuracy	AUC
p-value	0.651	0.183	0.32	0.494	0.108

Confusion Matrices: Fig. 5.2 presents the confusion matrices for the three supervised models. All models demonstrated comparable performance in terms of true positives, true negatives, false positives, and false negatives, with EBM and XGBoost producing identical results. Notably, both EBM and XGBoost exhibited a well-balanced performance between false alarm (false positive) and miss (false negative) rates. These findings again emphasize the strong performance of EBM, especially in comparison to the black-box baseline models.

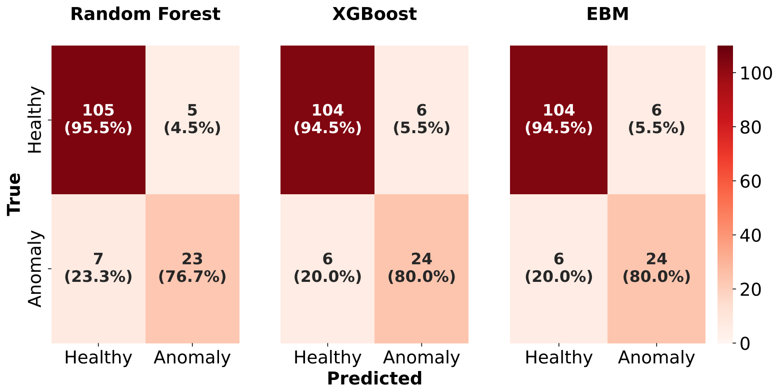


Figure 5.2 Confusion matrices of the supervised models applied to the APS dataset (with the row-normalized percentages displayed)

5.1.1.3 Explainability Results

Both XAI models, EBM and SHAP, generate two levels of explanations: global and local. Global explanations offer a comprehensive overview of the model's behavior across the entire dataset, illustrating the relative importance of input features and, in the case of EBM, their pairwise interactions as well. This perspective provides a broader understanding of how various input features collectively influence model predictions. Conversely, local explanations focus on the model's reasoning behind a specific prediction for a particular instance (e.g., a single vehicle), delivering a more detailed, instance-specific interpretation of the model's decision-making process.

Global Explanations: Fig. 5.3 illustrates the feature importances provided by EBM, ranked in descending order based on their mean absolute scores (or contributions). Similarly, the global explanations generated by SHAP are presented in Fig. 5.4 utilizing the beeswarm and heatmap plots, which display the ranked importance of input features in descending order. Overall, the feature importance rankings produced by the two independent models are closely aligned, demonstrating consistency between EBM and SHAP in identifying key predictive features. Furthermore, the identified ranking aligns well with domain knowledge, highlighting the most influential features such as AC_on/off_count, BrakePedalPos_mean, DutyCycle, P2_min, and P3_min. This ranking is consistent with domain expertise, as described in Section 4.1. For example, air leakage, a common APS issue, is typically characterized by increased AC_on/off_count and DutyCycle values, along with decreased P2_min and P3_min values. However, these patterns may also occur during episodes of heavy braking, underscoring the importance of BrakePedalPos_mean in providing contextual information. Accordingly, BrakePedalPos_mean is ranked among the three most important features, reflecting its critical role in distinguishing between normal and anomalous behavior under varying operational conditions.

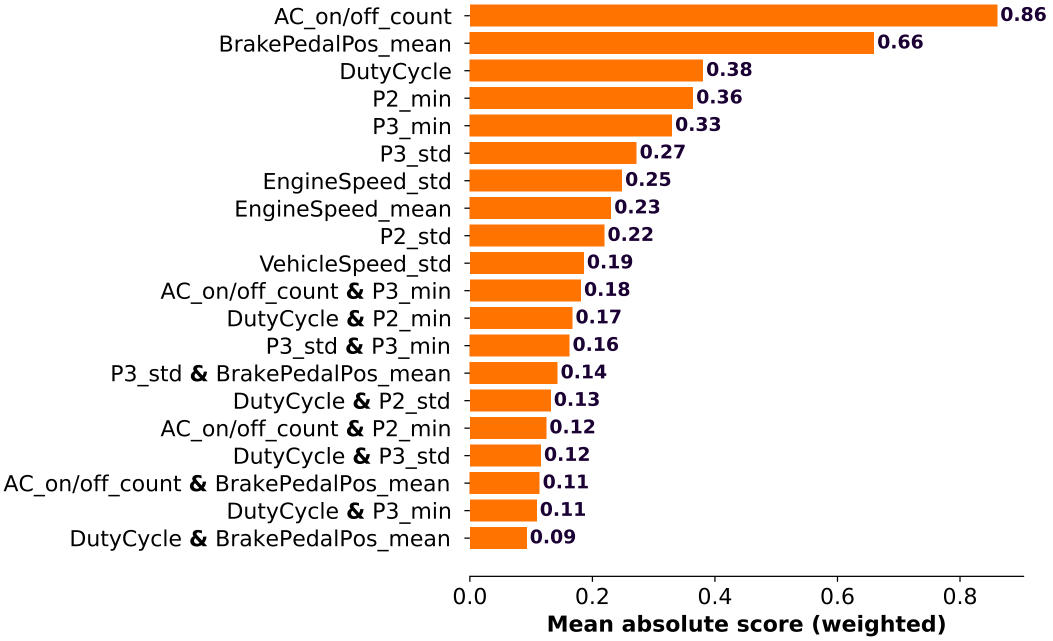
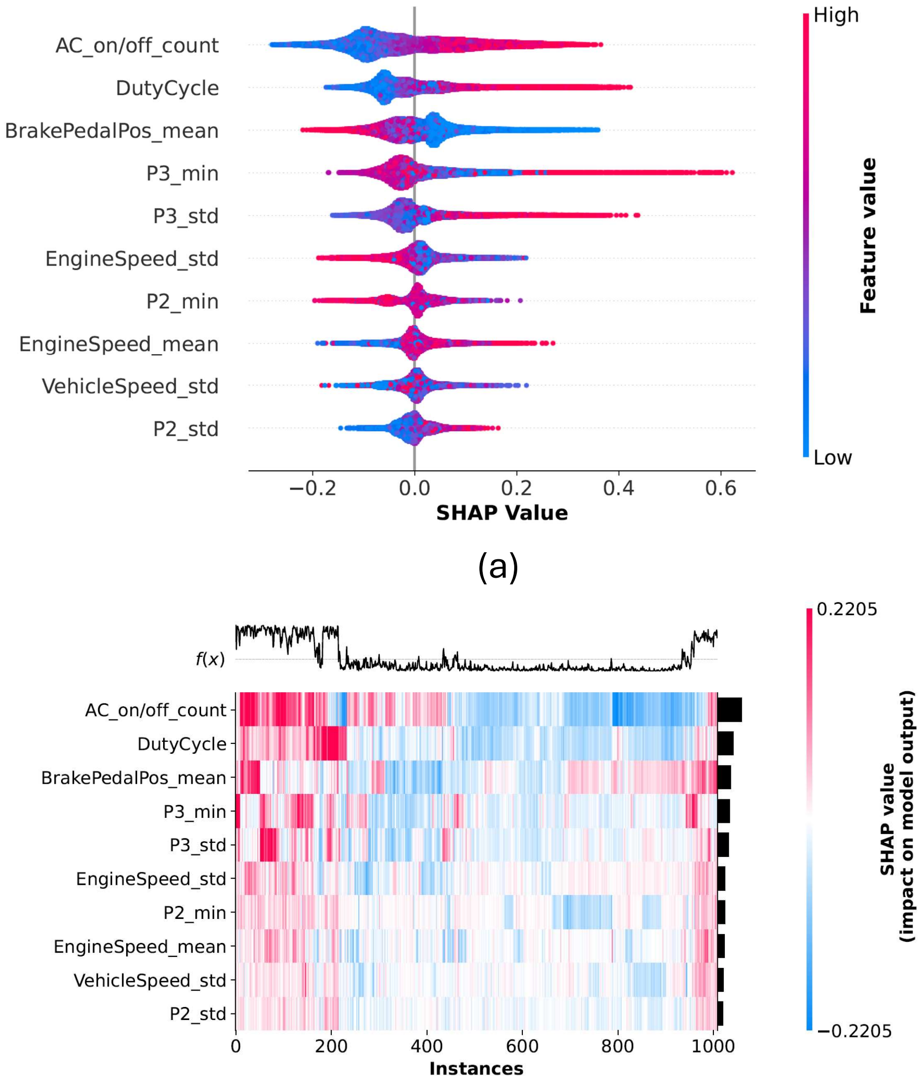


Figure 5.3 EBM-provided feature importance (Farea et al., 2025). Features are ranked in descending order according to importance.



(a)

(b)

Figure 5.4 SHAP-provided global explanations (Farea et al., 2025): (a). beeswarm plot showing the ranking of feature importance and the distribution of SHAP values for each feature across the entire dataset, and (b). heatmap plot showing feature importance rankings based on only 1000 data instances, with the classification probability log-odds function $f(x)$ shown at the top

The per-feature explanations generated by EBM for the five key features are presented in Fig. 5.5. Each feature is depicted with two graphs: the upper graph plots the feature's contribution scores against its value range, while the lower graph displays the feature's histogram across the dataset, with low-density regions excluded to minimize the impact of potential outliers. According to Eq. (3.12), positive contribution scores drive predictions towards the 'Anomaly' class, whereas negative scores indicate a tendency toward the 'Healthy' class.

For SHAP, the per-feature explanations are embedded in the beeswarm plot (see Fig. 5.4(a)), which offers a detailed visualization of feature importance and the

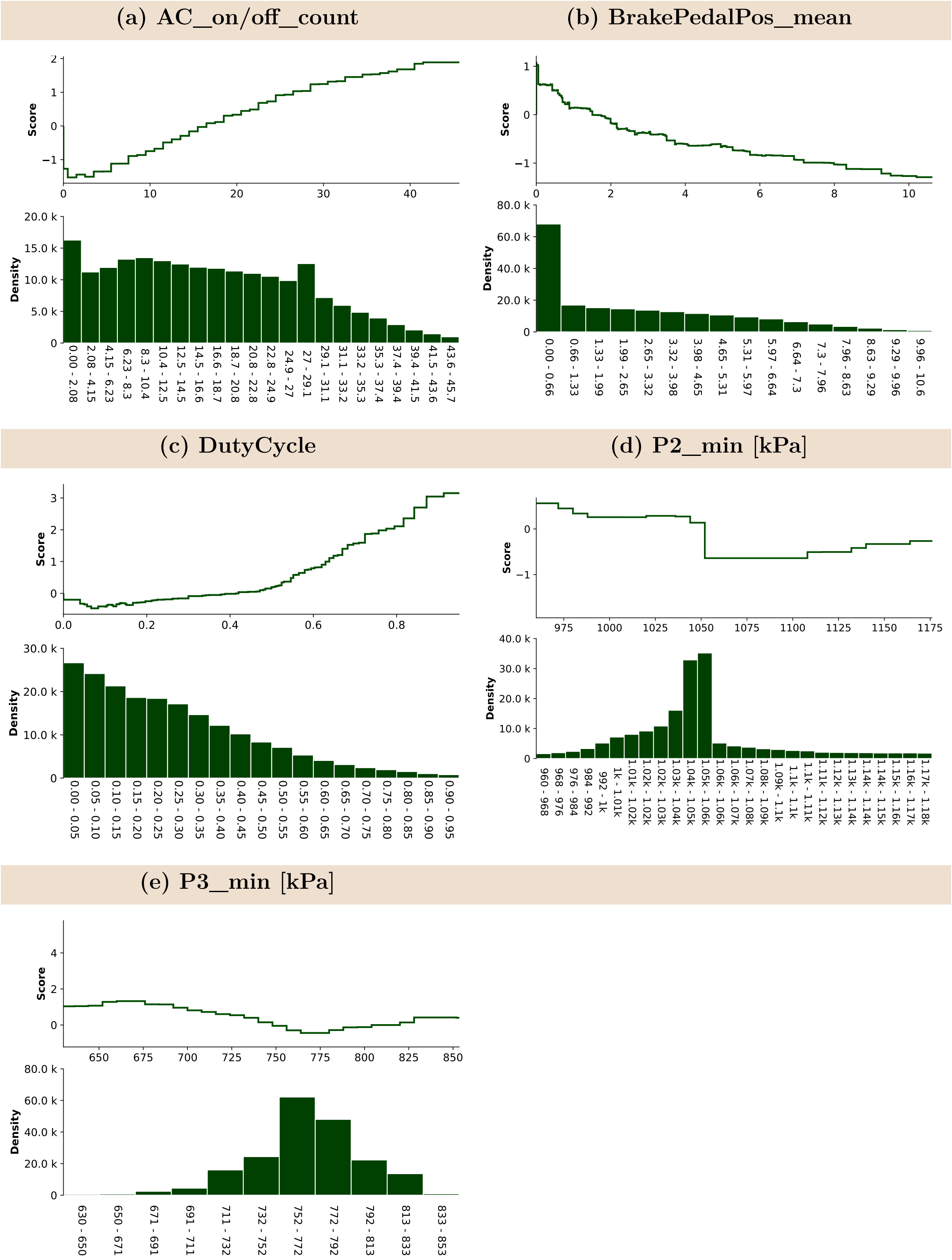


Figure 5.5 EBM explanations for the key features in the APS dataset (Farea et al., 2025). For each feature, its contributions (scores) across different ranges of its values are shown as a line plot, accompanied by a bar plot displaying the feature's distribution throughout the dataset, while excluding low-density regions.

distribution of SHAP values relative to feature values. This plot provides a concise yet information-rich summary of how the important features impact the model's final predictions. Each point represents a single data instance, with its position on the x-axis indicating the SHAP value (i.e., the feature's contribution score to the prediction) and its color denoting the feature value.

For both EBM and SHAP, AC_on/off_count and DutyCycle exhibit increasing trends in contribution scores as their values rise, signifying that higher values of these features are key indicators of potential APS failures, particularly during light braking periods. In contrast, BrakePedalPos_mean demonstrates a decreasing trend in contribution scores, with notably negative scores at higher values. This pattern counterbalances the high positive scores of AC_on/off_count and DutyCycle during heavy braking episodes, where increased APS activity is expected to maintain sufficient air pressure for braking. For P2_min and P3_min, the scores reveal dual behavior. Low values are associated with positive scores, indicating a potential air leakage issue due to insufficient minimum pressure. Conversely, very high values also receive positive scores, suggesting a possible overpressurization scenario, potentially caused by a malfunctioning relief valve. Overall, the distribution of contribution scores—both EBM scores and SHAP values—demonstrates a remarkable alignment with domain knowledge, reinforcing the credibility of the model's interpretability and its relevance to real-world APS failure detection scenarios.

Local Explanations: The local explainability of both XAI models was further assessed and compared for a healthy vehicle and an anomalous vehicle, as illustrated in Fig. 5.6. These local explanations provide valuable insights into the model's decision-making process, illustrating the specific factors that influenced its predictions for individual samples (i.e., vehicles). Both models generated consistent explanations, effectively aligning in their interpretation of key features for each vehicle. For the healthy vehicle, both EBM and SHAP attributed its classification primarily to the low values of AC_on/off_count and DutyCycle, coupled with normal values of P2_min and P3_min, indicating stable APS operation. Conversely, for the anomalous (faulty) vehicle, both models identified high values of AC_on/off_count, DutyCycle, and P3_std as significant indicators of anomalous behavior, aligning with known patterns of APS failure. This consistency between EBM and SHAP underscores the robustness of their interpretability frameworks in accurately identifying key failure indicators in APS data.

Further Investigation of EBM Local Explainability: The EBM-generated local explanations were further analyzed by examining additional six vehicles: two healthy vehicles correctly classified by EBM (true negatives) in Fig. 5.7, two anomalous vehicles correctly classified by EBM (true positives) in Fig. 5.8, and one anomalous vehicle misclassified by EBM as healthy (false negative) and one healthy vehicle misclassified by EBM as an anomaly (false positive) in Fig. 5.9. For each sample

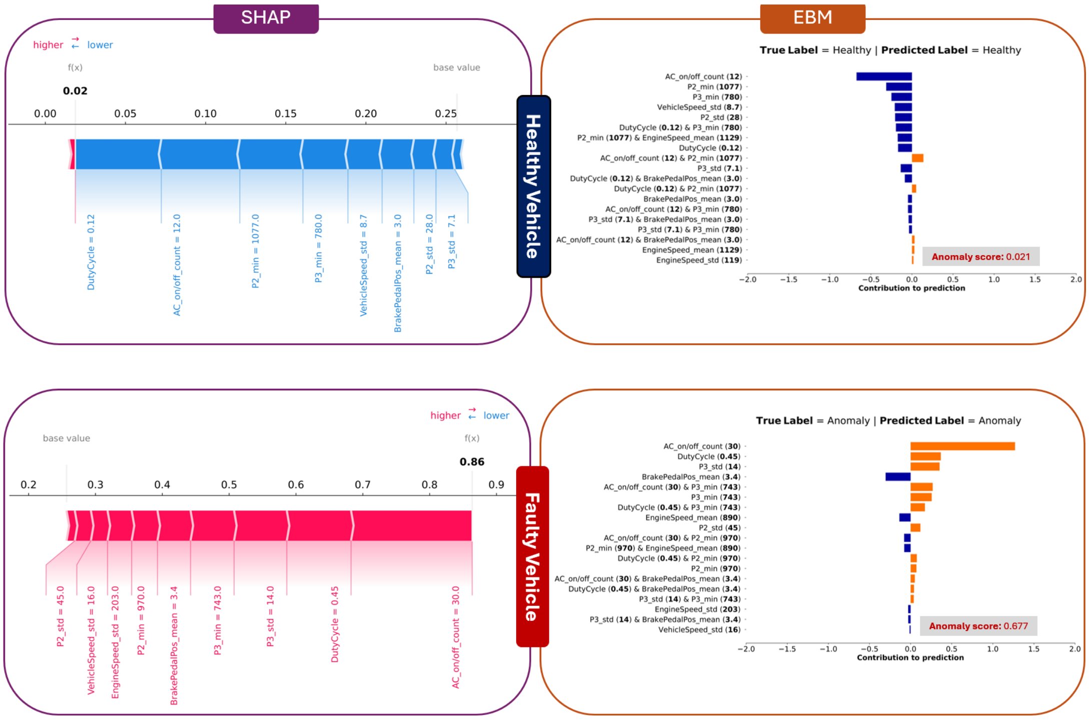


Figure 5.6 Comparison of local explanations generated by EBM and SHAP for a healthy vehicle and an anomalous (faulty) vehicle (Farea et al., 2025)

vehicle, the upper section illustrates the local explanations, with the per-vehicle average feature values indicated in parentheses. In contrast, the lower section presents time series plots of the key indicative features: DutyCycle, AC_on/off_count, and minimum pressures. These time series plots facilitate a detailed examination of the temporal evolution of each feature, providing valuable domain knowledge-based insights. The anomaly score—calculated as the median probability of all windows (i.e., data points) associated with that vehicle—is also provided. This score quantifies the EBM model's confidence in classifying the vehicle as anomalous. Additionally, the anomaly flags for each indicative feature are displayed above their respective time plots. The flag values range from zero to two, depending on the severity of the anomalous behavior exhibited by that feature. These EBM-generated local explanations for the six sample vehicles are as follows:

- EBM accurately classified Sample 1 as healthy, as shown in Fig. 5.7, primarily due to the low values of AC_on/off_count and DutyCycle, both of which align with typical APS behavior in non-anomalous conditions. P2_min and P3_min maintained normal values, further supporting the "healthy" classification. This assessment is corroborated by the time plots of DutyCycle, AC_on/off_count, and minimum pressures, which exhibit normal patterns with zero flags for each feature, indicating the absence of anomalous behavior.

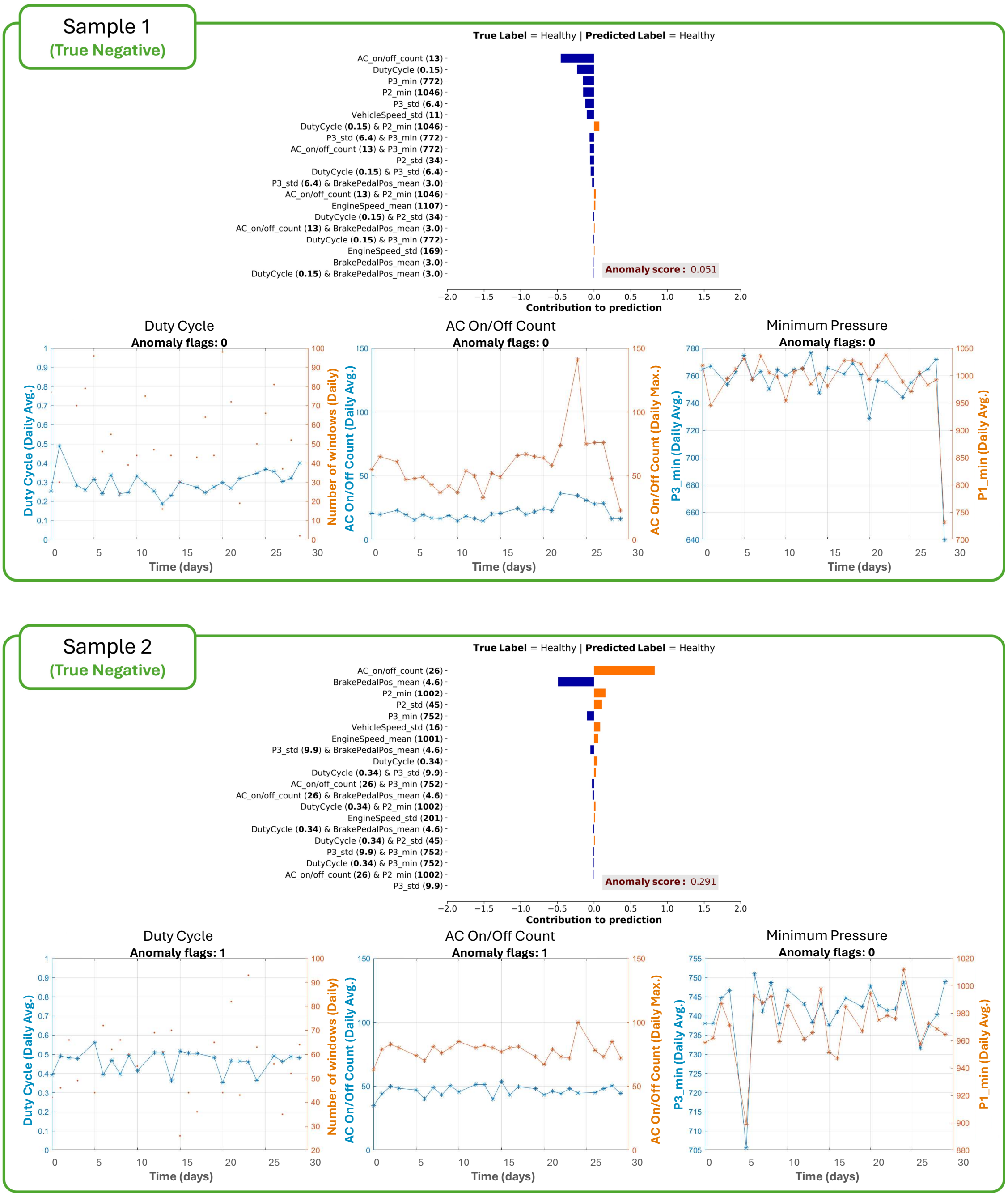


Figure 5.7 EBM-generated local explanations for correctly classified healthy vehicles (Farea et al., 2025). For each sample, the upper plot presents the EBM-provided explanations, with the vehicle-wise average values of the features indicated in parentheses. The lower plot depicts the daily average time series of the key features.

- Similarly, the EBM model correctly classified Sample 2 as healthy, despite its relatively high AC_on/off_count value. This classification is justified by the elevated BrakePedalPos_mean, as highlighted in the corresponding local explanation plot. In this context, the high AC_on/off_count value is interpreted as normal behavior during heavy braking periods, demonstrating the model's capability to contextualize feature values based on operational conditions.

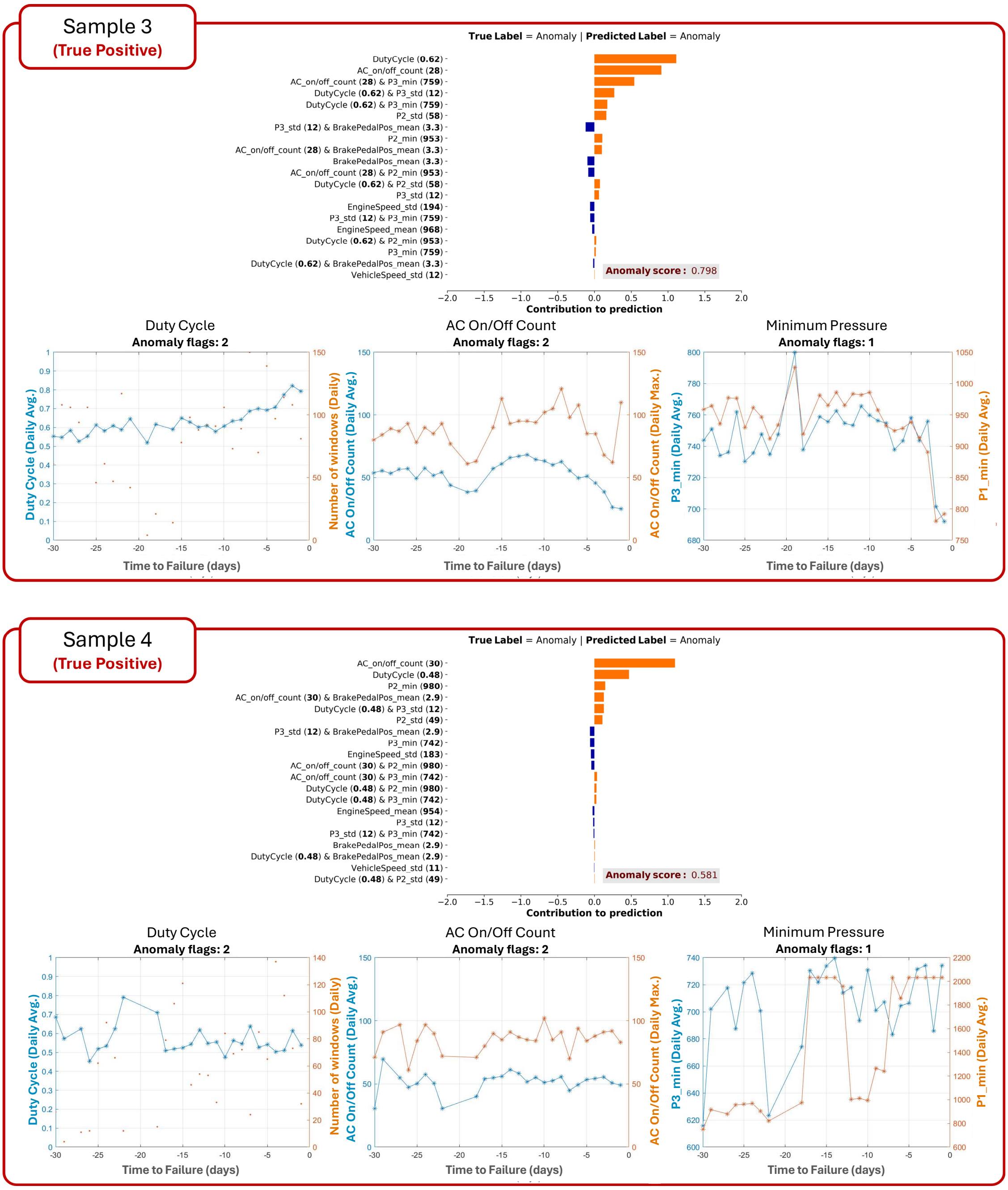


Figure 5.8 EBM-generated local explanations of correctly classified anomalous vehicles (Farea et al., 2025)

- Samples 3 and 4, both anomalous vehicles, were accurately classified as anomalous based on their high values of DutyCycle and AC_on/off_count, as indicated in Fig. 5.8. Upon closer examination of the time plots for these features, their values remained elevated throughout the final month before failure, with a notable increasing trend, particularly in the DutyCycle plot for Sample 3. The EBM-generated explanations effectively highlight the extended and frequent operation of the air compressor as a potential root cause of APS failure,

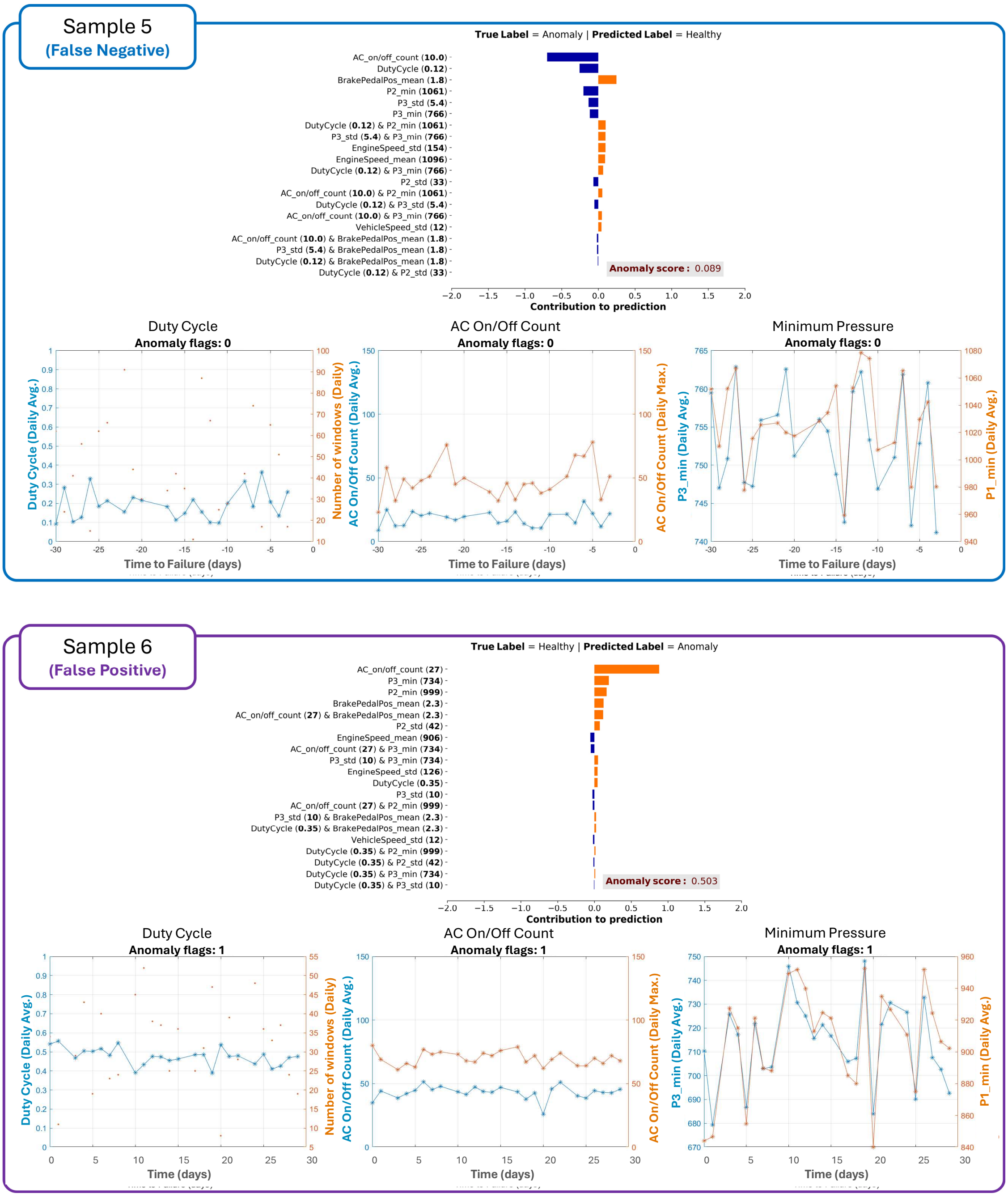


Figure 5.9 EBM-generated local explanations of misclassified vehicles (Farea et al., 2025)

emphasizing how sustained high values of DutyCycle and AC_on/off_count serve as early indicators of impending malfunction.

- Sample 5, a vehicle with a reported APS failure, was misclassified as healthy. It has a low median probability of being an anomaly (0.089) as shown in Fig. 5.9. The model's decision was driven by the low values of DutyCycle and AC_on/off_count, alongside the normal values of P2_min and P3_min. The corresponding time plots for these features confirm this assessment, displaying no significant anomalies or trends, and all features received zero flags. The absence of discernible abnormal behavior during the final month before failure suggests that the root cause of APS failure might have been subtle and not easily detectable through these key features. Another plausible explanation is that the E-APU was replaced as a preventive measure despite being functionally intact, though such cases are considered rare.

- Lastly, Sample 6, a healthy vehicle, was incorrectly classified by the EBM model as anomalous, attributing its decision to the elevated AC_on/off_count and the low values of P2_min and P3_min, as depicted in Fig. 5.9. The time plots of these three features reveal some degree of anomalous behavior, with each receiving one flag, indicating slight deviations from expected patterns.

5.1.2 Semi-supervised Learning

As stated in Chapter 3, two transformer-based architectures were employed as semi-supervised models for APS failure detection: a vanilla transformer-based model and the more advanced TranAD architecture. The vanilla transformer-based model was applied to a reduced version of the APS dataset, consisting of all the 30 anomalous vehicles and 77 of the 110 healthy vehicles. In contrast, for TranAD, the complete APS dataset was considered, allowing for a more comprehensive evaluation. For the vanilla transformer-based model, 70% of the healthy data in the reduced dataset was used for training, with the remaining 30% of the healthy data reserved for validation. Meanwhile, TranAD was trained using varying proportions of the healthy data, as detailed in Table 5.3. In each case, the remaining proportion of the healthy data served as the validation set. During inference, both models were evaluated on their respective full datasets, incorporating all healthy and anomalous vehicles to assess their generalization capabilities. For both models, the maximum number of training epochs was selected as 20, while all the other hyperparameters were set as the default values specified by Tuli et al. (2022).

Evaluation Metrics: The results summarized in Table 5.3 highlight distinct performance patterns between the two models. Despite its relatively simple architecture, the vanilla transformer model achieved promising results, with an accuracy exceeding 85% and an F1 score of 0.76. Notably, the model exhibited a high recall of 0.83, indicating its effectiveness in detecting 83% of anomalous vehicles. However, its relatively low precision suggests a higher false alarm rate, indicating potential

Table 5.3 Results of transformer-based models (HVs denoting healthy vehicles' data)

Model	Training Data	Precision	Recall	F1 Score	Accuracy
Vanilla transformer	70% of HVs	0.69	**0.83**	0.76	85.1%
TranAD model	20% of HVs	**0.82**	0.77	**0.79**	**91.4%**
	40% of HVs	0.81	0.73	0.77	90.7%
	60% of HVs	0.73	0.80	0.76	89.3%
	80% of HVs	0.79	0.77	0.78	90.7%

over-sensitivity to anomalies. In contrast, the TranAD model, a more sophisticated architecture, achieved superior performance, with an accuracy of 91.4% and an F1 score of 0.79. It also demonstrated higher precision, effectively reducing false alarms while maintaining strong recall. Remarkably, TranAD achieved its best performance using only 20% of the healthy data for training, which highlights its data efficiency and generalization capability under limited training data conditions. This impressive performance can be attributed to its effective attention mechanisms and adversarial training strategy, which enable the model to extract informative representations and accurately detect anomalies even with reduced training data. This finding underscores TranAD's robustness and adaptability, reinforcing its potential as a reliable anomaly detection framework in data-constrained scenarios.

Confusion Matrices: The false positives and false negatives for both models are highlighted in their corresponding confusion matrices in Fig. 5.10. For TranAD, the matrix reflects the scenario in which only 20% of the healthy data was used for training. The vanilla transformer model exhibits a relatively high rate of false positives, indicating a tendency to incorrectly classify healthy instances as anomalous. In contrast, TranAD maintains a more balanced performance, with a manageable number of false positives and false negatives, effectively controlling the false alarm rate while still capturing anomalous instances.

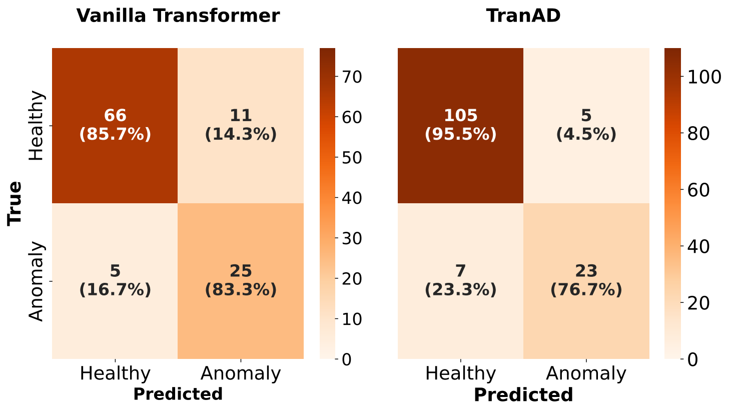


Figure 5.10 Confusion matrices for the transformer-based models

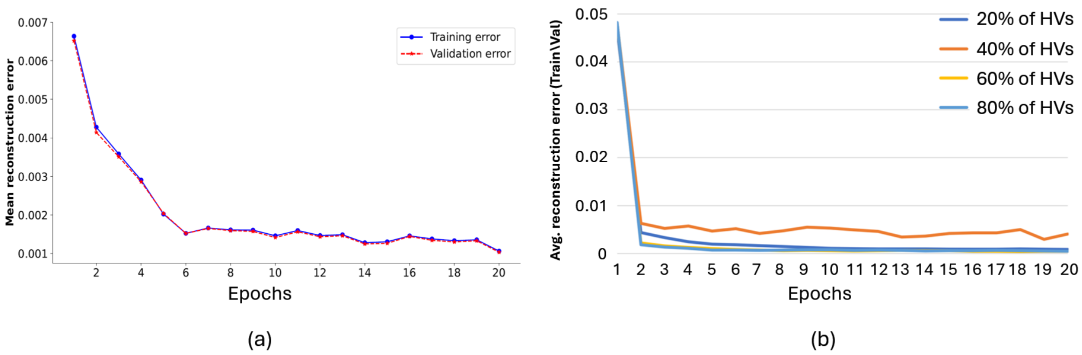


Figure 5.11 Learning curves for the semi-supervised models: (a) vanilla transformer-based model (Farea et al., 2024) (b) TranAD model (HVs: healthy vehicles' data) (Mumcuoglu et al., 2024b)

Learning Curves: The learning curves for both models are shown in Fig. 5.11. For TranAD, the displayed curves represent the averaged training and validation learning curves across multiple experiments, with different proportions of the healthy data utilized for training. TranAD demonstrates a substantially faster learning rate than the vanilla transformer model, rapidly achieving accurate reconstruction of input windows within the first two epochs. In contrast, the vanilla transformer requires over six epochs to reach a comparable level of accuracy. It is noteworthy that TranAD achieves this rapid convergence using only 20% of the healthy data.

Further Investigation into TranAD Performance: Lastly, Fig. 5.12 presents a comparison between a sample driving section from a healthy vehicle and a corresponding section from an anomalous vehicle, aiming to diagnose the root causes of the anomaly as identified by the TranAD model. The displayed driving sections include the true duty cycle, compressor on/off count, and minimum pressure signals, alongside their TranAD-reconstructed counterparts. Beneath each signal, the respective reconstruction error is plotted, highlighting discrepancies between the actual and reconstructed signals. The TranAD model accurately reconstructs the signals in the healthy driving section, indicating a close match between the original and reconstructed data. In contrast, it struggles to reconstruct certain segments of the anomalous driving section accurately. Such poorly reconstructed areas are highlighted in red within Fig. 5.12. These discrepancies, marked as high reconstruction errors, are indicative of potential anomaly sources. A closer examination of these segments reveals that the three signals indeed exhibit abnormal patterns compared to the typical behavior observed in the healthy section. Moreover, it is noteworthy that TranAD demonstrates a heightened sensitivity to short anomalous segments, particularly within the minimum pressure signal. This ability to detect even short-duration anomalies underscores its effectiveness in detecting subtle, localized anomalies within the data.

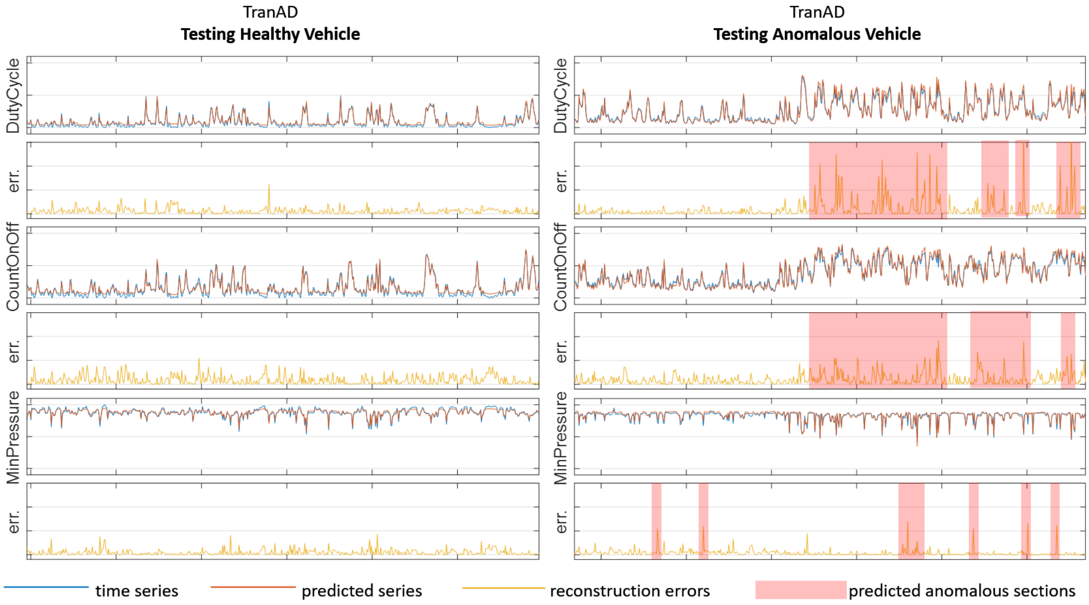


Figure 5.12 True vs TranAD-reconstructed driving sections from a healthy vehicle and an anomalous vehicle (Mumcuoglu et al., 2024b)

5.1.3 Overall Comparison

Table 5.4 provides a comprehensive comparison of the data-driven models applied to APS failure detection. Overall, the supervised models and the TranAD model demonstrated strong and comparable performance, effectively addressing the detection task. However, the primary emphasis is placed on EBM, as an interpretable model, and TranAD, as a semi-supervised architecture, while the remaining models serve as baseline references for evaluating their performance.

EBM's performance is particularly noteworthy, with several key strengths highlighted as follows:

- **Strong Performance:** Despite being a glass-box interpretable model, EBM delivered performance comparable to advanced black-box models such as Random Forest and XGBoost, effectively balancing interpretability and classification accuracy. This capability to offer clear, interpretable insights into its decision-making process while achieving black-box-level accuracy highlights the practical value of EBM in real-world applications, particularly in high-stakes settings where model explainability is crucial.

- **Explanation Diversity and Computational Efficiency:** The explanations generated by both EBM and SHAP, a widely recognized XAI baseline, align well with domain expertise. However, EBM demonstrates clear advantages over SHAP in terms of explanation diversity and computational effi-

Table 5.4 Comparison of data-driven models for APS failure detection: The best result for each metric is shown in bold, and values comparable to the best are underlined.

Learning paradigm	Model	Precision	Recall	F1 Score	Accuracy	AUC
Supervised	Random Forest	**0.82**	0.77	0.79	**91.4%**	**0.892**
	XGBoost	0.80	0.80	**0.80**	**91.4%**	0.886
	EBM	0.80	0.80	**0.80**	**91.4%**	0.881
Semi-supervised	Vanilla transformer	0.69	**0.83**	0.76	85.1%	0.891
	TranAD model	**0.82**	0.77	0.79	**91.4%**	0.890

ciency. While SHAP focuses exclusively on individual feature attributions, EBM extends beyond single features to incorporate pairwise feature interactions, resulting in richer, more comprehensive explanations. In terms of computational efficiency[1], EBM is significantly faster than SHAP, as shown in Table 5.5. The time complexity per observation for EBM is approximately 1 millisecond, while SHAP requires around 0.7 seconds per observation, representing a substantial increase in computational cost. This notable difference positions EBM as a more practical choice for large-scale or real-time applications, effectively maintaining interpretability and accuracy while significantly reducing computational overhead.

On the other hand, the performance of TranAD demonstrated several significant strengths, as outlined below:

- **Data Efficiency and Rapid Convergence:**TranAD exhibited exceptional data efficiency, achieving robust performance using only 20% of the healthy data for training, a fraction of the training data used by the other models. Additionally, it demonstrated fast convergence during training, effectively learning representative patterns in a relatively short training period.
- **Effective Semi-Supervised Learning:** Despite its semi-supervised formulation, TranAD delivered performance comparable to fully supervised models (see Table 5.4). This outcome is particularly significant given that the semi-supervised approach is inherently more challenging, as it is trained without labeled anomalies. Nevertheless, this formulation is more practical in real-world scenarios, where labeled anomalous data is typically scarce. Through the semi-supervised formulation, TranAD effectively addressed critical anomaly detection challenges such as class imbalance and anomaly heterogeneity.

[1] The experiments were conducted on a Windows 11 system with an Intel Core i7-10510U CPU (1.80 GHz base clock) and 16 GB of RAM, using the following implementations: InterpretML (0.5.0) for EBM, Scikit-learn (1.2.2) for Random Forest, and SHAP (0.46.0) for SHAP TreeExplainer.

Table 5.5 Time complexity of EBM and SHAP (measured in seconds per observation)

Model	Time complexity (s)
(Random Forest + SHAP)	(3.995×10^{-4} + 6.749×10^{-1})
EBM	1.161×10^{-3}

- **Temporal Sequence Modeling:** As a sequential architecture, TranAD effectively captures temporal dependencies across input windows, a capability that static models like Random Forest and XGBoost lack. This temporal modeling capacity is particularly advantageous in APS failure detection, where anomalies often manifest as subtle, time-dependent patterns (e.g., conditional or group anomalies).

Overall, TranAD's combination of data efficiency, effective semi-supervised learning, and temporal modeling capabilities underscores its robustness and suitability for APS failure detection, particularly in data-constrained and sequential data settings.

Main Limitations of EBM and TranAD-based Frameworks: Unlike TranAD, EBM faces limitations in its static modeling nature, which restricts its capacity to capture temporal dependencies inherent in time series data. Additionally, its reliance on fully supervised training necessitates labeled data for both normal and anomalous classes. In contrast, the primary limitation of the semi-supervised TranAD-based framework lies in the challenge of providing meaningful interpretations for its predictions, primarily due to the absence of labeled anomalous instances during training. Furthermore, such semi-supervised models are not explicitly optimized to distinguish between normal and anomalous data; rather, they are trained to reconstruct the training data—typically normal instances—as discussed in Section 2.1.3.2.

5.1.4 Ablation Study: Investigation of Window Parameters

This subsection investigates the impact of the sliding window parameters— window length and shift—on the APS failure detection performance of EBM and TranAD. For EBM, three window lengths (5, 10, and 20 minutes) were examined, with each tested under two shift configurations: a full window-length shift and a 50% overlap (i.e., half-window shift)[2]. The corresponding results are summarized in Table 5.6.

[2] The results presented in Section 5.1.1 for all supervised models, including EBM, are based on a nominal window length of 10 minutes and a window shift of 5 minutes.

Table 5.6 Effect of window size and shift on the performance of EBM

Window parameters		**Evaluation metrics**				
Size	**Shift**	**Precision**	**Recall**	**F1 score**	**Accuracy**	**AUC**
5 min	2.5 min	**0.846**	0.733	0.786	**91.4%**	0.881
	5 min	0.774	0.800	0.787	90.7%	0.885
10 min	5 min	0.800	0.800	**0.800**	**91.4%**	0.881
	10 min	0.767	0.767	0.767	90.0%	**0.889**
20 min	10 min	0.686	0.800	0.738	87.9%	0.884
	20 min	0.667	**0.867**	0.754	87.9%	0.879

Among the tested configurations, a 10-minute window combined with a 5-minute shift yielded the highest F1 score, indicating the best trade-off between precision and recall. This outcome suggests that the chosen window length is sufficiently long to extract informative features—such as the DutyCycle and AC_on/off_count—without introducing excessive temporal smoothing. The use of overlapping windows (50% shift) improved performance for the 10-minute window, but had minimal effect on the shorter (5-minute) and longer (20-minute) windows.

In the case of the TranAD model, the analysis compared two window lengths: the default 20-minute setting (as reported in Section 5.1.2) and a shorter 10-minute window. In both scenarios, the window shift was fixed at 10 minutes. The resulting F1 scores across different training data proportions are shown in Fig. 5.13. On average, the 20-minute window outperformed the shorter alternative, especially when only 20% of the healthy data was available for training. This finding aligns with domain-specific insights: longer input sequences lead to more meaningful and robust values for the key indicators, such as the duty cycle and compressor activation patterns, which are crucial for accurately modeling normal and anomalous behavior.

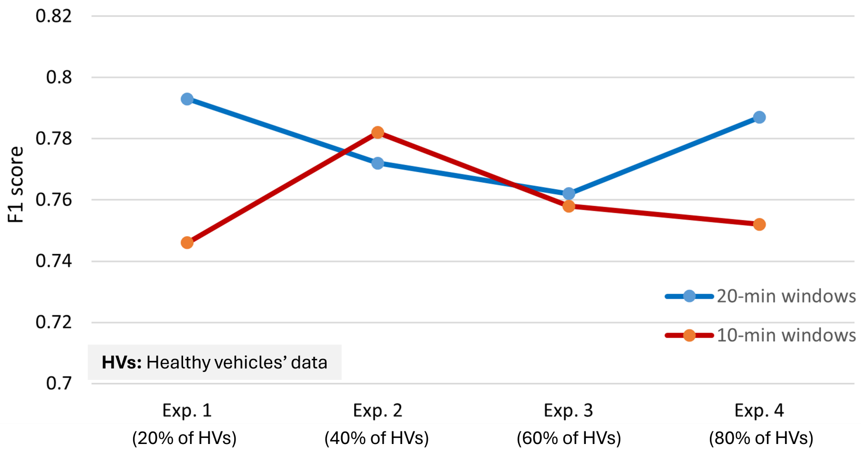


Figure 5.13 Effect of the input window length on the performance of TranAD

5.2 Detection of DED Defects

In the context of defect detection in DED processes, this thesis incorporates three distinct datasets. Two of them, referred to as DED-IN718 and DED-IN718-U, were curated and prepared as part of this research, while the third dataset, referred to as DED-Ti64, is publicly available.

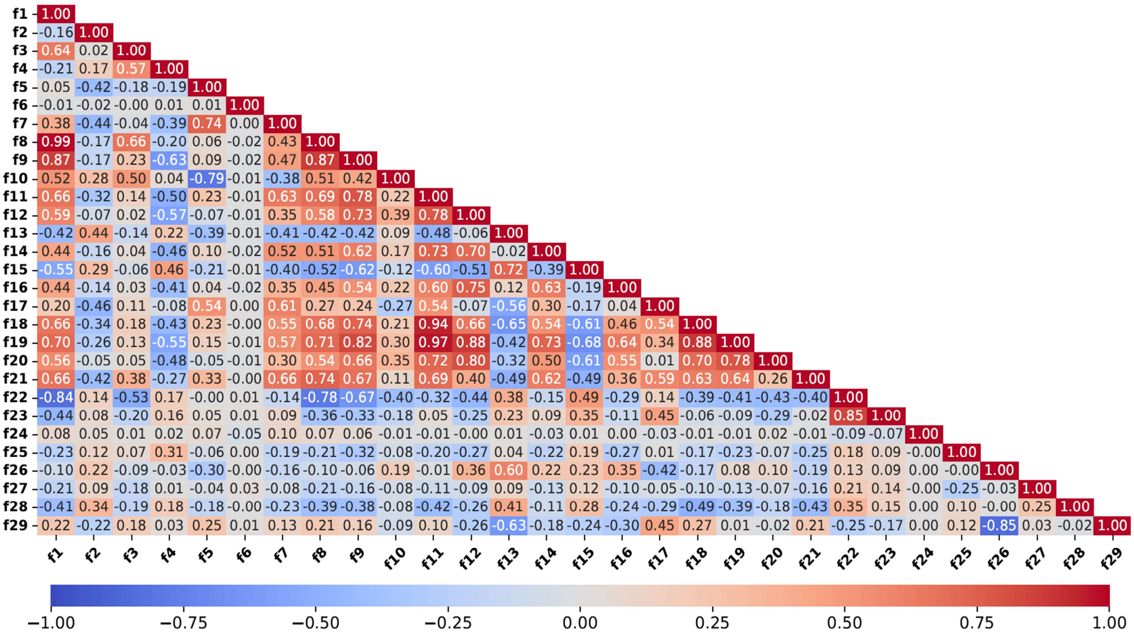
For the DED-IN718 dataset, the defect detection framework was developed using supervised models, given its balanced class distribution and single-class anomalies. In contrast, the framework for the DED-Ti64 dataset was based on one-class SVM and iForest, implemented in semi-supervised and unsupervised settings, aligning with the dataset's imbalanced class distribution. Lastly, a fully unsupervised approach was adopted for the DED-IN718 dataset due to the lack of ground-truth labels.

5.2.1 Supervised Learning

This section starts by presenting the feature correlation analysis for the DED-IN718 dataset, followed by a comprehensive evaluation of the supervised models, Random Forest and SVM, applied to the same dataset.

Feature Correlation Analysis: As detailed in Section 4.2, a comprehensive set of shape, color, and texture features was extracted from each thermal image in the DED-IN718 dataset. These features, combined with process-related features, provide a comprehensive summary of both the melt pool characteristics and the spatio-temporal context of each image. Consequently, it is expected that normal (defect-free) images will exhibit distinct feature patterns compared to anomalous (defective) images based on this feature set.

However, the extracted features may contain redundancies due to interdependencies, potentially compromising model performance and computational efficiency. To address this, a Pearson correlation analysis was conducted to assess the linear relationships between feature pairs. The resulting correlation matrix, presented in Fig. 5.14, provides the pairwise correlation coefficients. Features with a correlation coefficient exceeding 0.7 in magnitude were considered highly correlated, and only one feature from each highly correlated pair was retained. Through this analysis, the initial set of 29 features was significantly reduced to 17, effectively eliminating 12 redundant features while maintaining the most informative ones. This feature



	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10	f11	f12	f13	f14	f15	f16	f17	f18	f19	f20	f21	f22	f23	f24	f25	f26	f27	f28	f29
f1	1.00																												
f2	-0.16	1.00																											
f3	0.64	0.02	1.00																										
f4	-0.21	0.17	0.57	1.00																									
f5	0.05	-0.42	-0.18	-0.19	1.00																								
f6	-0.01	-0.02	-0.00	0.01	0.01	1.00																							
f7	0.38	-0.44	-0.04	-0.39	0.74	0.00	1.00																						
f8	0.99	-0.17	0.66	-0.20	0.06	-0.02	0.43	1.00																					
f9	0.87	-0.17	0.23	-0.63	0.09	-0.02	0.47	0.87	1.00																				
f10	0.52	0.28	0.50	0.04	-0.79	-0.01	-0.38	0.51	0.42	1.00																			
f11	0.66	-0.32	0.14	-0.50	0.23	-0.01	0.63	0.69	0.78	0.22	1.00																		
f12	0.59	-0.07	0.02	-0.57	-0.07	-0.01	0.35	0.58	0.73	0.39	0.78	1.00																	
f13	-0.42	0.44	-0.14	0.22	-0.39	-0.01	-0.41	-0.42	-0.42	0.09	-0.48	-0.06	1.00																
f14	0.44	-0.16	0.04	-0.46	0.10	-0.02	0.52	0.51	0.62	0.17	0.73	0.70	-0.02	1.00															
f15	-0.55	0.29	-0.06	0.46	-0.21	-0.01	-0.40	-0.52	-0.62	-0.12	-0.60	-0.51	0.72	-0.39	1.00														
f16	0.44	-0.14	0.03	-0.41	0.04	-0.02	0.35	0.45	0.54	0.22	0.60	0.75	0.12	0.63	-0.19	1.00													
f17	0.20	-0.46	0.11	-0.08	0.54	0.00	0.61	0.27	0.24	-0.27	0.54	-0.07	-0.56	0.30	-0.17	0.04	1.00												
f18	0.66	-0.34	0.18	-0.43	0.23	-0.00	0.55	0.68	0.74	0.21	0.94	0.66	-0.65	0.54	-0.61	0.46	0.54	1.00											
f19	0.70	-0.26	0.13	-0.55	0.15	-0.01	0.57	0.71	0.82	0.30	0.97	0.88	-0.42	0.73	-0.68	0.64	0.34	0.88	1.00										
f20	0.56	-0.05	0.05	-0.48	-0.05	-0.01	0.30	0.54	0.66	0.35	0.72	0.80	-0.32	0.50	-0.61	0.55	0.01	0.70	0.78	1.00									
f21	0.66	-0.42	0.38	-0.27	0.33	-0.00	0.66	0.74	0.67	0.11	0.69	0.40	-0.49	0.62	-0.49	0.36	0.59	0.63	0.64	0.26	1.00								
f22	-0.84	0.14	-0.53	0.17	-0.00	0.01	-0.14	-0.78	-0.67	-0.40	-0.32	-0.44	0.38	-0.15	0.49	-0.29	0.14	-0.39	-0.41	-0.43	-0.40	1.00							
f23	-0.44	0.08	-0.20	0.16	0.05	0.01	0.09	-0.36	-0.33	-0.18	0.05	-0.25	0.23	0.09	0.35	-0.11	0.45	-0.06	-0.09	-0.29	-0.02	0.85	1.00						
f24	0.08	0.05	0.01	0.02	0.07	-0.05	0.10	0.07	0.06	-0.01	-0.01	-0.00	0.01	-0.03	0.01	0.00	-0.03	-0.01	-0.01	0.02	-0.01	-0.09	-0.07	1.00					
f25	-0.23	0.12	0.07	0.31	-0.06	0.00	-0.19	-0.21	-0.32	-0.08	-0.20	-0.27	0.04	-0.22	0.19	-0.27	0.01	-0.17	-0.23	-0.07	-0.25	0.18	0.09	-0.00	1.00				
f26	-0.10	0.22	-0.09	-0.03	-0.30	-0.00	-0.16	-0.10	-0.06	0.19	-0.01	0.36	0.60	0.22	0.23	0.35	-0.42	-0.17	0.08	0.10	-0.19	0.13	0.09	0.00	-0.00	1.00			
f27	-0.21	0.09	-0.18	0.01	-0.04	0.03	-0.08	-0.21	-0.16	-0.08	-0.11	-0.09	0.09	-0.13	0.12	-0.10	-0.05	-0.10	-0.13	-0.07	-0.16	0.21	0.14	-0.00	-0.25	-0.03	1.00		
f28	-0.41	0.34	-0.19	0.18	-0.18	-0.00	-0.23	-0.39	-0.38	-0.08	-0.42	-0.26	0.41	-0.11	0.28	-0.24	-0.29	-0.49	-0.39	-0.18	-0.43	0.35	0.15	0.00	0.10	-0.00	0.25	1.00	
f29	0.22	-0.22	0.18	0.03	0.25	0.01	0.13	0.21	0.16	-0.09	0.10	-0.26	-0.63	-0.18	-0.24	-0.30	0.45	0.27	0.01	-0.02	0.21	-0.25	-0.17	0.00	0.12	-0.85	0.03	-0.02	1.00

f1	Area	**f2**	HoleArea	**f3**	MajorAxisLength	**f4**	Eccentricity	**f5**	Circularity
f6	Orientation	**f7**	Solidity	**f8**	EquivalentDiameter	**f9**	MinorAxisLength	**f10**	Perimeter
f11	Mean	**f12**	StandardDeviation	**f13**	Skewness	**f14**	Entropy	**f15**	Kurtosis
f16	Mode	**f17**	Q1	**f18**	Q2	**f19**	Q3	**f20**	Contrast
f21	Correlation	**f22**	Energy	**f23**	Homogeneity	**f24**	Theta	**f25**	Rho
f26	Z	**f27**	SampleID	**f28**	ScanSpeed	**f29**	LaserPower		

Figure 5.14 Pearson pairwise correlation between extracted features (redundant features to be excluded due to high correlation with other features are marked in red)

selection process obviously reduces dimensionality and enhances the model's efficiency, facilitating more robust and computationally efficient learning. The final set of selected features, along with their statistical characteristics, is presented in Table 5.7. Given the varying value ranges of the selected features, feature normalization was subsequently applied to standardize the data, ensuring that all features contribute equally during the classification process. This step is crucial to maintain model accuracy and stability.

Classification Results: The five-fold cross-validation results for both Random Forest (RF) and SVM are presented in Table 5.8, with the corresponding hyperparameter configurations detailed in Table 5.9. Following a series of experiments, the selected hyperparameters were identified as those providing the best overall performance. For the Random Forest model, the default hyperparameter settings were maintained, as adjustments to these values did not lead to notable performance enhancements.

The evaluation metrics, summarized in Table 5.8, include two scenarios: (i) using the complete set of 29 features and (ii) using only the 17 selected features.

Table 5.7 Selected features for the DED-IN718 dataset and their statistics

Feature category	Feature name	mean	std	min	max
Shape Features	(1) Area (in pixels)	3491	469.6	0	5731
	(2) HoleArea (in pixels)	16.04	15.34	0	187
	(3) Major Axis Length (in pixels)	74.88	6.050	0	94.28
	(4) Eccentricity *	0.5509	0.1255	0	0.9016
	(5) Circularity *	0.5395	0.1050	0	1.969
	(6) Orientation (in degrees)	1.990	51.66	-89.95	89.97
Color Features	(7) Mean (in ^{o}C)	1594	51.81	0	1761
	(8) Skewness *	0.04952	0.3952	-0.8334	1.747
	(9) Mode (in ^{o}C)	1813	200.1	0	2007
	(10) 1^{st} Quartile (in ^{o}C)	1462	33.93	0	1569
Texture Features	(11) Correlation *	0.9007	0.03596	-1.0158E-4	0.9410
	(12) Homogeneity *	0.9195	0.01579	0	0.9998
Process-related Features	(13) Theta (in degrees)	180	101.8	25.71	334.3
	(14) Rho (in mm)	8.585	0.9644	7.1	10.5
	(15) Z (in mm)	27.37	14.41	1.56	52.02
	(16) Sample ID	2.918	1.337	1	5
	(17) Scan Speed (in mm/min)	627.3	143.2	500	900

* These are dimensionless quantities.

Table 5.8 Results of the supervised classifiers on the DED-IN718 dataset

Model	Feature Set	Precision	Recall	F1 Score	AUC	Accuracy
Random Forest	All features	0.802	0.857	0.829	0.903	83.6%
	Selected features	**0.812**	**0.858**	**0.834**	**0.910**	**84.2%**
SVM	All features	0.804	0.836	0.820	0.890	83.0%
	Selected features	0.796	**0.858**	0.826	0.896	83.2%

The corresponding receiver operating characteristic (ROC) curves for both models under these scenarios are illustrated in Fig. 5.15, where the area under the ROC curve (AUC) remains consistently high for both models, with values around 0.9 (see Table 5.8). These high AUC values indicate the supervised classifiers' strong discriminative ability in distinguishing defect-free from defective thermal images, highlighting their effectiveness in binary classification tasks. Notably, reducing the feature set did not diminish model performance; instead, it led to a slight improvement, especially for Random Forest. This enhancement underscores the effectiveness of feature selection, which eliminated redundant features while preserving the informative ones, thereby improving model accuracy and computational efficiency.

When comparing the performance of the two classifiers, the Random Forest model consistently outperformed the SVM model across all evaluation metrics. Random

Table 5.9 Hyperparameters of the supervised classifiers (n denotes the training data size, k is the total number of input features, and f is the number of features to randomly select at each split)

Random Forest		**SVM**	
Hyperparameter	Value	Hyperparameter	Value
- No. of learners (m)	100	- Kernel function	Gaussian
- Split criterion	Gini Impurity	- Kernel scale	3.85
- No. of features (f)	$\sqrt{k}$	- Regularization parameter (C)	1
- Max. number of splits	$n-1$		

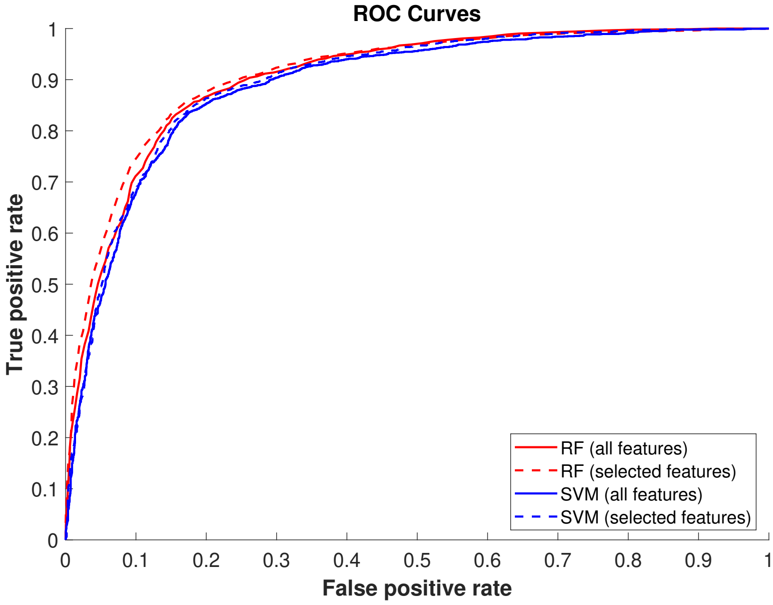


Figure 5.15 ROC curves for the supervised classifiers on the DED-IN718 dataset

Forest achieved the highest accuracy, approximately 84%, along with an F1 score exceeding 0.83, attaining a well-balanced performance between precision and recall. In terms of precision, Random Forest demonstrated a value of 0.81, signifying that 81% of the thermal images classified as anomalous are actual anomalies with internal defects. Additionally, it achieved a recall of approximately 0.86, effectively detecting 86% of the actual anomalous images. This balanced performance between precision and recall highlights Random Forest's strong capability in identifying defects while minimizing false positives, making it a more reliable classifier for this application of defect detection based on thermal imaging data.

The confusion matrices for both classifiers are depicted in Fig. 5.16. When utilizing the complete set of features, SVM generates fewer false positives than Random Forest, but at the cost of higher false negatives, suggesting that it is more conservative

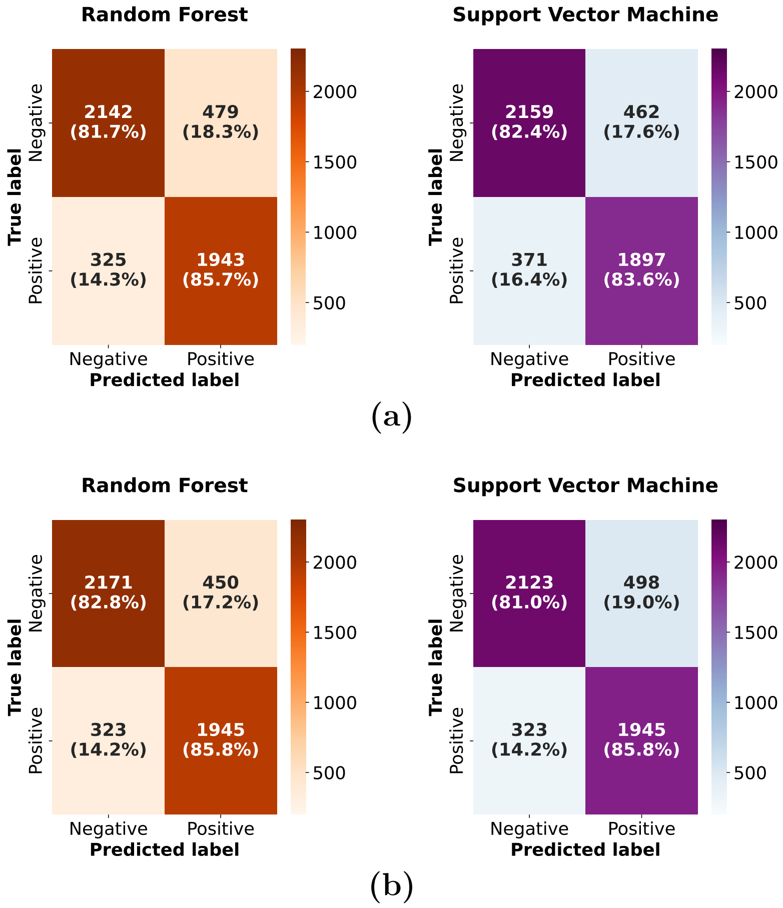
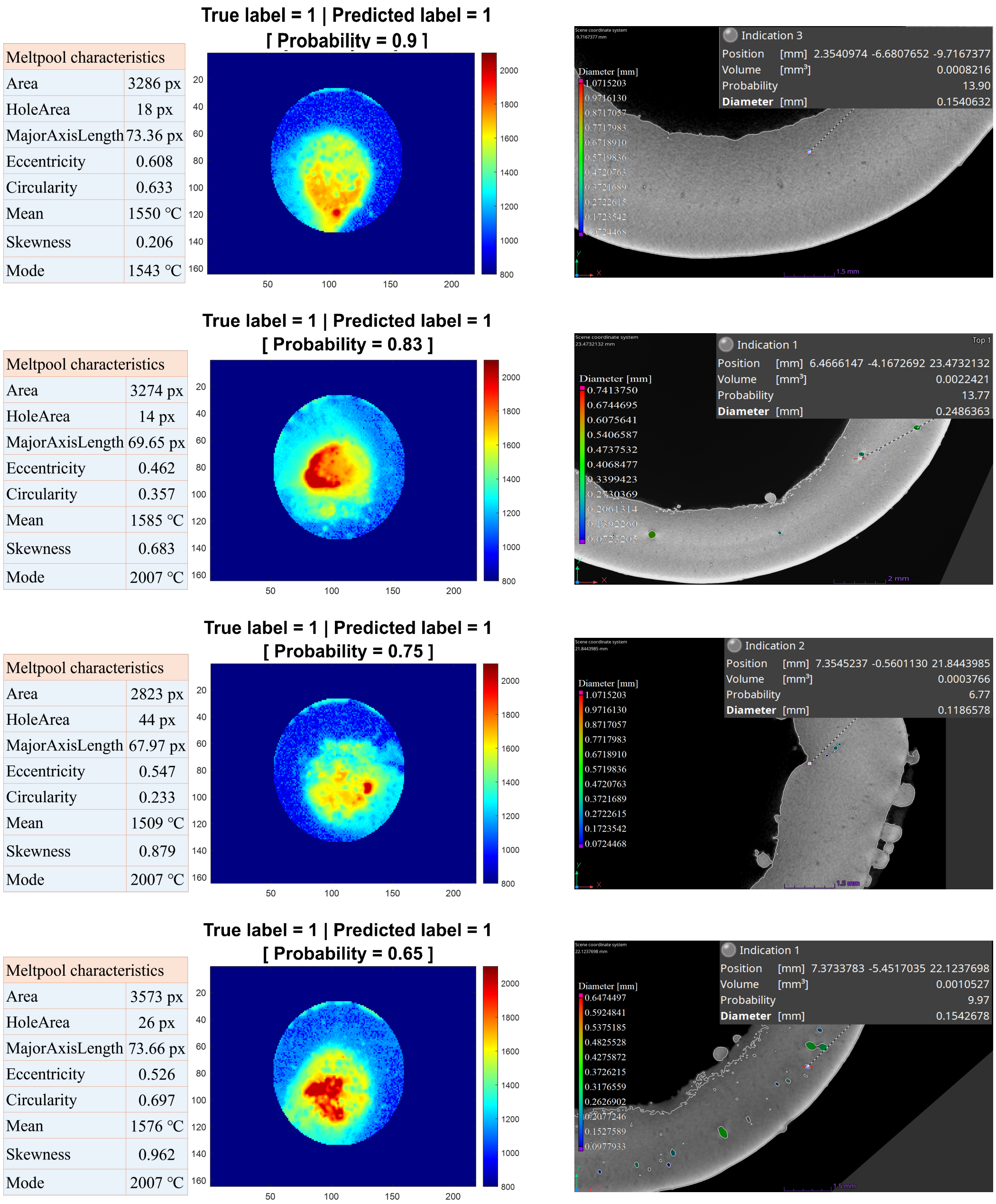


Figure 5.16 Confusion matrices of the supervised classifiers applied to the DED-IN718 dataset: (a) using all features and (b) using the selected features as the input features

in predicting defects. However, when the selected feature set is employed, Random Forest shows substantial improvement, with reductions in both false positives and false negatives, underscoring its robustness and enhanced generalization ability with a more concise, independent feature set. In contrast, SVM exhibits a notable reduction in false negatives with the selected features; however, this gain is counterbalanced by an increase in false positives, indicating a trade-off in its classification strategy. Overall, Random Forest consistently outperforms SVM, particularly in reducing false positives, which is critical for maintaining high defect detection accuracy and reducing unnecessary false alarms.

Fig. 5.17 illustrates representative true positive cases, comprising thermal images which Random Forest successfully classified as anomalous, along with key melt pool characteristics. Corresponding X-ray images are also provided, where CT-identified porosities are annotated with relevant characteristics, including spatial coordinates and diameters. The thermal images exhibit nonuniform temperature distributions, frequently marked by overheating or underheating and accompanied by irregular



Meltpool characteristics	
Area	3286 px
HoleArea	18 px
MajorAxisLength	73.36 px
Eccentricity	0.608
Circularity	0.633
Mean	1550 °C
Skewness	0.206
Mode	1543 °C

Indication 3
Position [mm] 2.3540974 -6.6807652 -9.7167377
Volume [mm³] 0.0008216
Probability 13.90
Diameter [mm] 0.1540632

Meltpool characteristics	
Area	3274 px
HoleArea	14 px
MajorAxisLength	69.65 px
Eccentricity	0.462
Circularity	0.357
Mean	1585 °C
Skewness	0.683
Mode	2007 °C

Indication 1
Position [mm] 6.4666147 -4.1672692 23.4732132
Volume [mm³] 0.0022421
Probability 13.77
Diameter [mm] 0.2486363

Meltpool characteristics	
Area	2823 px
HoleArea	44 px
MajorAxisLength	67.97 px
Eccentricity	0.547
Circularity	0.233
Mean	1509 °C
Skewness	0.879
Mode	2007 °C

Indication 2
Position [mm] 7.3545237 -0.5601130 21.8443985
Volume [mm³] 0.0003766
Probability 6.77
Diameter [mm] 0.1186578

Meltpool characteristics	
Area	3573 px
HoleArea	26 px
MajorAxisLength	73.66 px
Eccentricity	0.526
Circularity	0.697
Mean	1576 °C
Skewness	0.962
Mode	2007 °C

Indication 1
Position [mm] 7.3733783 -5.4517035 22.1237698
Volume [mm³] 0.0010527
Probability 9.97
Diameter [mm] 0.1542678

Figure 5.17 Examples of anomalous thermal images in the DED-IN718 dataset: (a) thermal images, each annotated with its ground truth and predicted labels (where 1 refers to the defective class), along with the corresponding classification probability provided by Random Forest; and (b) the corresponding X-ray images, where CT-identified porosities are highlighted and accompanied by their key characteristics.

melt pool geometries. Insufficient fusion resulting from underheating is a predominant factor contributing to porosity formation. In contrast, overheating introduces potential complications, such as increased surface roughness and compromised mechanical integrity (Ranjan et al., 2023). Additionally, overheating can exacerbate pore generation through gas entrapment, driven by pronounced temperature differentials between the melt pool and surrounding powder particles (Zhao et al., 2021).

5.2.2 Semi-supervised and Unsupervised Learning

iForest and one-class SVM were applied to the DED-Ti64 dataset under both semi-supervised and unsupervised settings. In contrast, only an unsupervised approach using DBSCAN was applied to the DED-IN718-U dataset due to the absence of ground-truth labels. This section first presents the results for the DED-Ti64 dataset, followed by the results for the DED-IN718-U dataset.

5.2.2.1 DED-Ti64 Dataset

Feature Correlation Analysis: The correlation matrix presented in Fig. 5.18 delivers valuable insights into the relationships among features in the DED-Ti64 dataset, facilitating a more informed selection of features for anomaly detection. Similar to the DED-IN718 dataset, by leveraging this correlation analysis, the feature set can be refined to eliminate redundancy while preserving key information, ultimately enhancing model efficiency and effectiveness in detecting defects.

Certain features exhibit strong correlations, notably the temperature mean, median, and first quartile (Q1), with correlation coefficients of 0.98 or higher. Retaining only one representative from such highly correlated groups can significantly reduce model complexity while preserving essential information. A similar relationship is evident between variance and standard deviation, where their strong association justifies the exclusion of one. This pattern also extends to gradient-based features, such as gradient mean and gradient standard deviation. Additionally, geometric features—such as major axis length, minor axis length, area, and perimeter—demonstrate consistently high positive correlations among each other, further supporting the case for dimensionality reduction without sacrificing critical information.

Accordingly, filtering out redundant features resulted in a substantial reduction in model complexity, effectively achieving significant dimensionality reduction while also enhancing model accuracy. As shown in Table 5.10, models trained on the reduced feature set outperformed those using the complete set of features.

Performance Evaluation: The evaluation metrics for iForest and one-class SVM across different feature sets and learning paradigms are presented in Table 5.10. To ensure a comprehensive assessment, both models were evaluated under semi-supervised and unsupervised learning settings using both the complete and the selected feature sets.

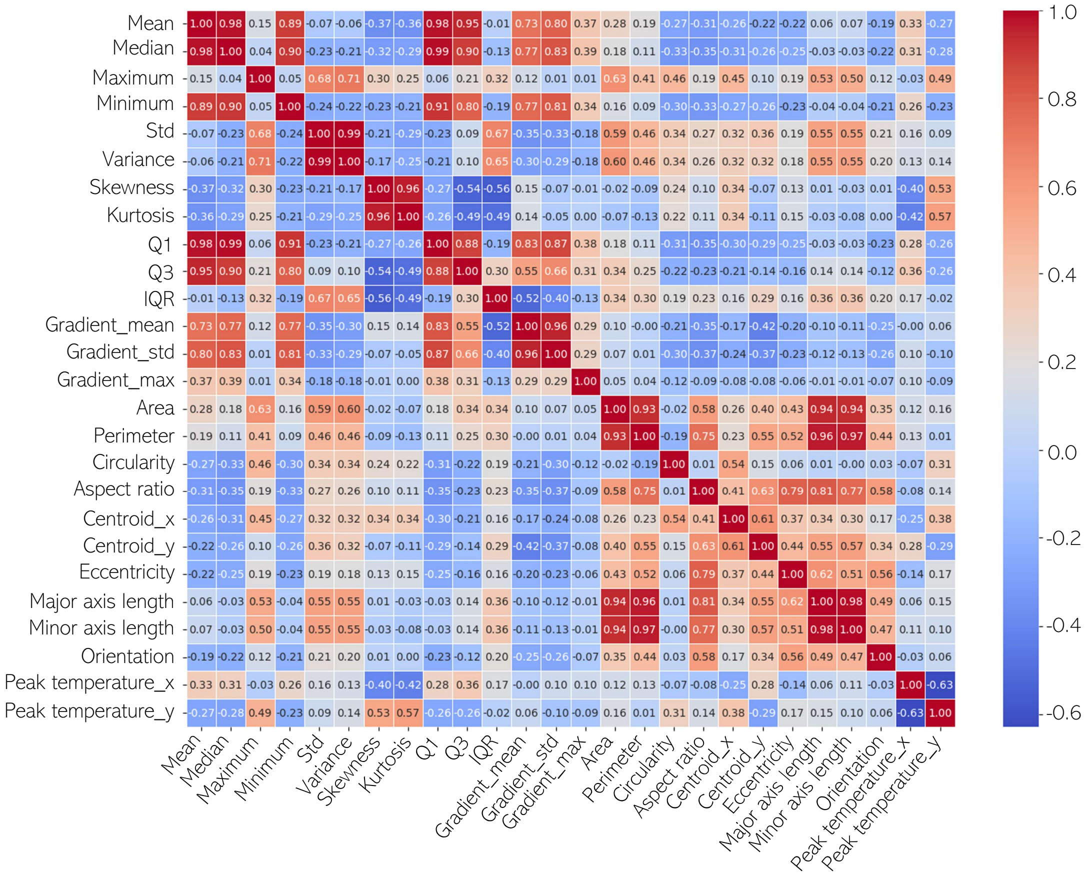


Figure 5.18 Feature correlation matrix for the DED-Ti64 dataset

Table 5.10 Results of iForest and one-class SVM on the DED-Ti64 dataset

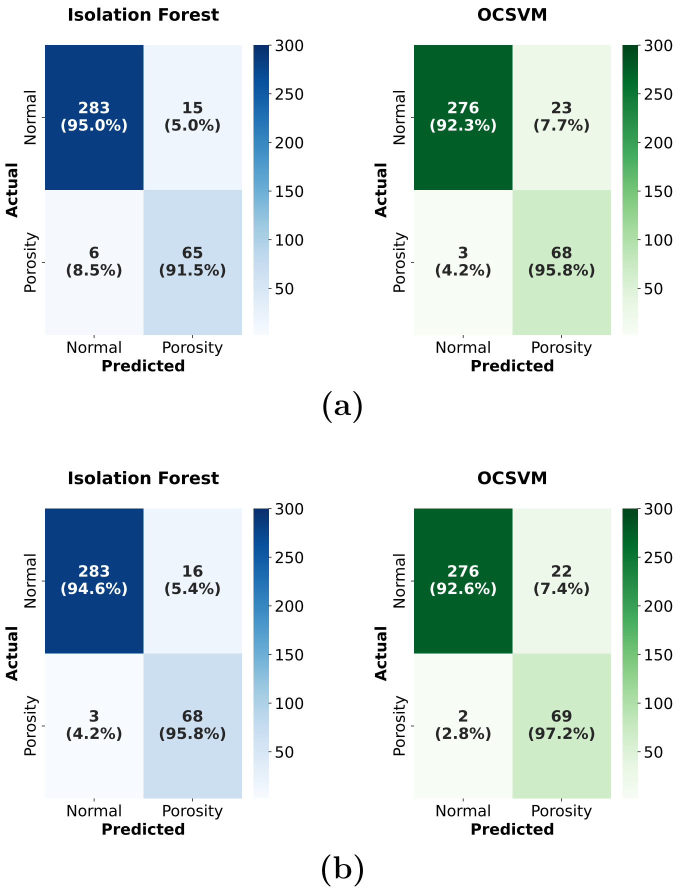
Feature Set	**Training**	**iForest**				**One-class SVM**			
		Precision	**Recall**	**F1**	**Accuracy**	**Precision**	**Recall**	**F1**	**Accuracy**
All Features	Semi-Supervised	0.81	0.91	0.86	94%	0.75	0.96	0.84	93%
	Unsupervised	0.62	0.68	0.64	96%	0.39	0.67	0.49	93%
Selected Features	Semi-Supervised	**0.82**	0.96	**0.88**	95%	0.76	**0.97**	0.85	93%
	Unsupervised	0.74	0.84	0.78	**98%**	0.44	0.77	0.56	94%

Overall, iForest consistently outperformed one-class SVM in both learning paradigms. The best results for both models were achieved using the selected features in the semi-supervised setting, underscoring the effectiveness of semi-supervised learning in detecting anomalies for these models. Specifically, iForest achieved a precision of 0.82, a recall of 0.96, and an F1 score of 0.88, indicating strong detection capabilities. In comparison, the optimal performance of one-class SVM was characterized by a precision of 0.76, a recall of 0.97, and an F1 score of 0.85, demonstrating a relatively strong recall but a slightly lower precision.

Under the unsupervised learning setting, iForest maintained relatively stable performance, with an F1 score of 0.78, highlighting its robustness to variations in data distribution. In contrast, one-class SVM's performance deteriorated significantly, with the precision and F1 score dropping to 0.44 and 0.56, respectively. This finding indicates the sensitivity of one-class SVM to training data impurities and its limited generalization capacity in unsupervised scenarios.

The confusion matrices in Fig. 5.19 provide a detailed visual representation of the classification performance of iForest and one-class SVM under the semi-supervised learning paradigm, offering valuable insights into each model's effectiveness in identifying normal and defective samples. The key observations are as follows:

- **Effect of Learning Paradigm:** The choice of training paradigm (semi-supervised vs. unsupervised) had a pronounced impact on model performance: Under the semi-supervised setting, one-class SVM achieved an F1 score of 0.85, with corresponding precision 0.76 and recall 0.97, effectively capturing anomalous instances while maintaining a relatively controlled false positive rate (see Fig. 5.19). Nonetheless, in the unsupervised setting, its performance declined sharply, with the F1 score dropping to 0.56, alongside a decrease in



(a)

(b)

Figure 5.19 Confusion matrices for iForest and one-class SVM applied to the DED-Ti64 dataset (semi-supervised setting): (a) using all features and (b) using the selected features

precision to 0.44 and recall to 0.77. This pronounced decline underscores the model's susceptibility to data contamination, as the inclusion of both normal and anomalous data during training compromises its ability to effectively distinguish between classes, leading to significant misclassifications. For iForest, the semi-supervised approach yielded a precision of 0.82, a recall of 0.96, and an F1 score of 0.88, reflecting a balanced trade-off between defect detection and false positive minimization (see Fig. 5.19). However, under the unsupervised setting, iForest's F1 score decreased to 0.78, with a corresponding decrease in precision to 0.74 and recall to 0.84. Although iForest exhibited greater robustness than one-class SVM, the increase in false positives underscores the impact of training data purity on model reliability. Overall, the semi-supervised training scheme demonstrated clear advantages, significantly enhancing both precision and recall for both models.

- **Effect of Feature Selection:** Feature selection played a prime role in optimizing model performance, particularly in distinguishing normal from anomalous instances. For iForest, the use of the selected feature set reduced the number of missed anomalies from 6 to 3, resulting in an increase in recall from 0.91 to 0.96, while maintaining a comparable false alarm rate of approximately 5% (see Fig. 5.19). This represents a clear improvement over the model trained on the full feature set (refer to Fig. 5.19(a)), where a higher number of missed anomalies was observed. Similarly, one-class SVM's performance improved slightly under the selected feature set. This pattern suggests that eliminating redundant features enabled the model to focus on more informative features, effectively enhancing its capacity to distinguish between normal and defective instances.

- **Comparison Between iForest and one-class SVM (when using the selected features):** iForest achieved a high recall of 0.96, correctly identifying 68 out of 71 defective instances, underscoring its strong sensitivity to anomalies (see Fig. 5.19(b)). However, it misclassified 16 normal instances as anomalies, leading to a precision of 0.82. Accordingly, this performance results in an F1 score of 0.88, effectively balancing defect detection and false positive reduction. In contrast, one-class SVM achieved a slightly higher recall of 0.97, detecting 69 out of 71 anomalous instances. However, its precision dropped to 0.76, misclassifying 22 normal samples as anomalies, indicating a higher rate of false positives. Consequently, one-class SVM has an F1 score of 0.85, reflecting a trade-off between higher defect detection and increased false alarms. These results suggest that iForest maintains a better balance between precision and recall, making it more reliable in practical DED applications where minimizing false positives is critical. Conversely, one-class SVM prioritizes recall at the expense of higher false positive rates, potentially increasing costs associated with unnecessary defect identification.

5.2.2.2 DED-IN718-U Dataset

Feature Correlation Analysis: The distributions of the extracted features are shown in Fig. 5.20, several of which exhibit approximately Gaussian behavior (e.g., mean and energy). Based on these distributions, images with feature values lying in the distribution tails are likely to correspond to defective instances, thereby enabling spatial differentiation between normal and anomalous images in the feature space. The corresponding summary statistics for the extracted features are provided in Table 5.11. Nonetheless, the initial set of 12 extracted features may exhibit redundancy due to potential linear dependencies. To identify and address such correlations, a Pearson correlation analysis was performed to assess the pairwise linear relationships among the features. The resulting correlation matrix, shown in Fig. 5.21, reveals that area, homogeneity, and contrast are highly correlated, with absolute correlation coefficients exceeding 0.9. This strong linear dependence suggests that these features carry overlapping information. Consequently, to reduce redundancy while retaining discriminative power, contrast and homogeneity were excluded.

Defect Detection Results: A DBSCAN ensemble was employed for defect detection, comprising 16 distinct models. This ensemble was constructed by systematically varying the two key DBSCAN hyperparameters: MinPts and Eps. Following the guidance of Ester et al. (1996), MinPts—representing the minimum number of points required to form a dense region—should be set to at least $k+1$, with k denoting the number of features. Given that the dataset dimensionality (k) is 10, four sufficiently spaced MinPts values were selected: 11, 16, 21, 26, to encourage diversity among the models. Eps, which defines the neighborhood radius for a point, was selected based on the analysis of the k-distance graph, a widely adopted technique in DBSCAN hyperparameter tuning (Ester et al., 1996; Schubert et al., 2017). This graph, illustrated in Fig. 5.22, depicts the sorted distances of each data point to its k-th nearest neighbor for various MinPts values. From the inflection region of these curves, four Eps values were identified: {1.0, 1.5, 2.0, 2.5}, spanning the informative range of the k-distance graph. Combining these four MinPts values with the four selected Eps values yielded a diverse ensemble of 16 DBSCAN configurations. This ensemble structure enhances robustness and generalization by incorporating variability across different clustering granularities and density assumptions.

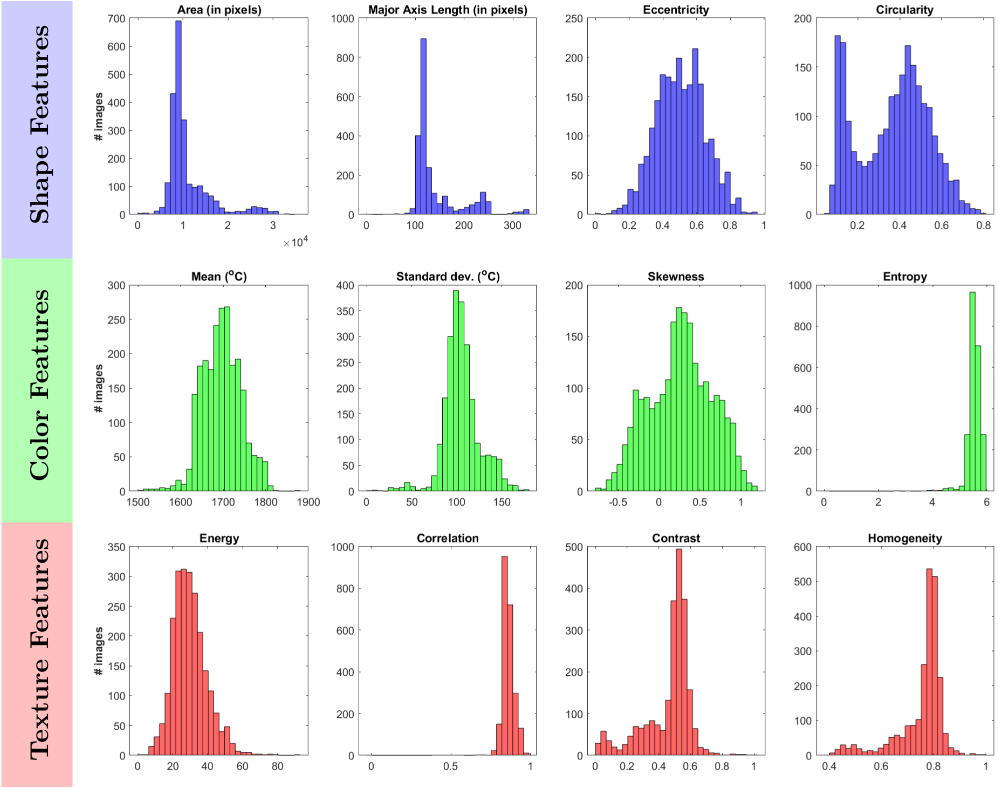
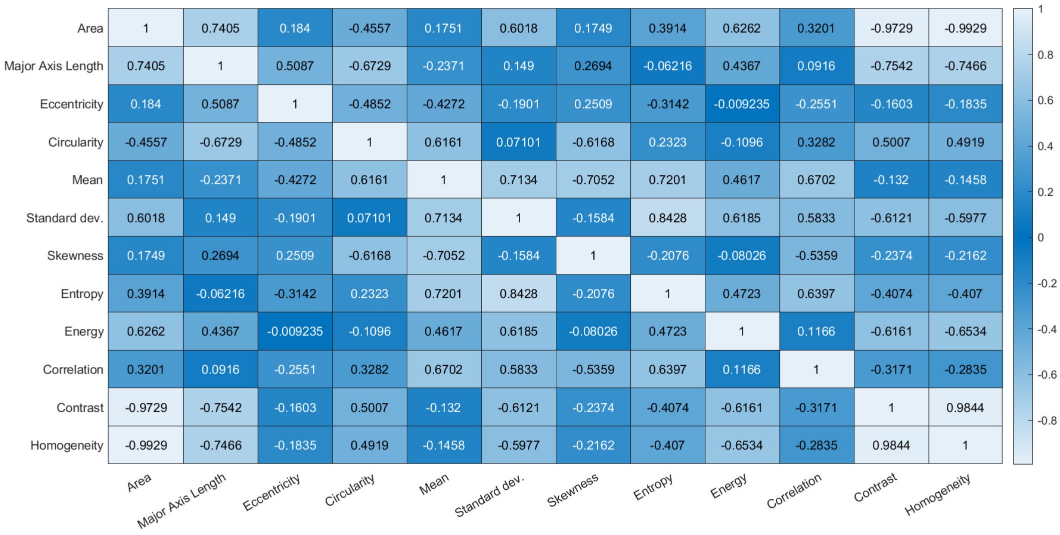


Figure 5.20 Distributions of the extracted features from the DED-IN718-U dataset

Table 5.11 Statistics of features extracted from the DED-IN718-U dataset

Features	**mean**	**std**	**min**	**max**
Area (in pixels)	11059	4908	427	35752
Major axis length (in pixels)	142	50.2	37.9	331
Eccentricity	0.506	0.145	0.076	0.946
Circularity	0.367	0.169	0.044	0.811
Mean (in ^{o}C)	1697	46	1515	1879
Standard dev. (in ^{o}C)	105	20.3	9.49	179
Skewness	0.263	0.384	-0.729	1.19
Entropy	5.53	0.248	2.64	5.94
Energy	29.9	9.80	5.62	91.0
Correlation	0.859	0.039	0.406	0.962
Contrast	0.460	0.149	0.011	0.963
Homogeneity	0.753	0.091	0.388	0.984

The clustering results are provided in Table 5.12. The DBSCAN ensemble detected two structured clusters alongside some anomalies (outliers) in the feature space. A significant majority of the thermal images (2228 instances) were grouped into Cluster 1, while a smaller subset of 25 images was assigned to Cluster 2. The remaining 42 images did not belong to either cluster and were consequently labeled



	Area	Major Axis Length	Eccentricity	Circularity	Mean	Standard dev.	Skewness	Entropy	Energy	Correlation	Contrast	Homogeneity
Area	1	0.7405	0.184	-0.4557	0.1751	0.6018	0.1749	0.3914	0.6262	0.3201	-0.9729	-0.9929
Major Axis Length	0.7405	1	0.5087	-0.6729	-0.2371	0.149	0.2694	-0.06216	0.4367	0.0916	-0.7542	-0.7466
Eccentricity	0.184	0.5087	1	-0.4852	-0.4272	-0.1901	0.2509	-0.3142	-0.009235	-0.2551	-0.1603	-0.1835
Circularity	-0.4557	-0.6729	-0.4852	1	0.6161	0.07101	-0.6168	0.2323	-0.1096	0.3282	0.5007	0.4919
Mean	0.1751	-0.2371	-0.4272	0.6161	1	0.7134	-0.7052	0.7201	0.4617	0.6702	-0.132	-0.1458
Standard dev.	0.6018	0.149	-0.1901	0.07101	0.7134	1	-0.1584	0.8428	0.6185	0.5833	-0.6121	-0.5977
Skewness	0.1749	0.2694	0.2509	-0.6168	-0.7052	-0.1584	1	-0.2076	-0.08026	-0.5359	-0.2374	-0.2162
Entropy	0.3914	-0.06216	-0.3142	0.2323	0.7201	0.8428	-0.2076	1	0.4723	0.6397	-0.4074	-0.407
Energy	0.6262	0.4367	-0.009235	-0.1096	0.4617	0.6185	-0.08026	0.4723	1	0.1166	-0.6161	-0.6534
Correlation	0.3201	0.0916	-0.2551	0.3282	0.6702	0.5833	-0.5359	0.6397	0.1166	1	-0.3171	-0.2835
Contrast	-0.9729	-0.7542	-0.1603	0.5007	-0.132	-0.6121	-0.2374	-0.4074	-0.6161	-0.3171	1	0.9844
Homogeneity	-0.9929	-0.7466	-0.1835	0.4919	-0.1458	-0.5977	-0.2162	-0.407	-0.6534	-0.2835	0.9844	1

Figure 5.21 Pairwise correlation between features extracted from the DED-IN718-U dataset

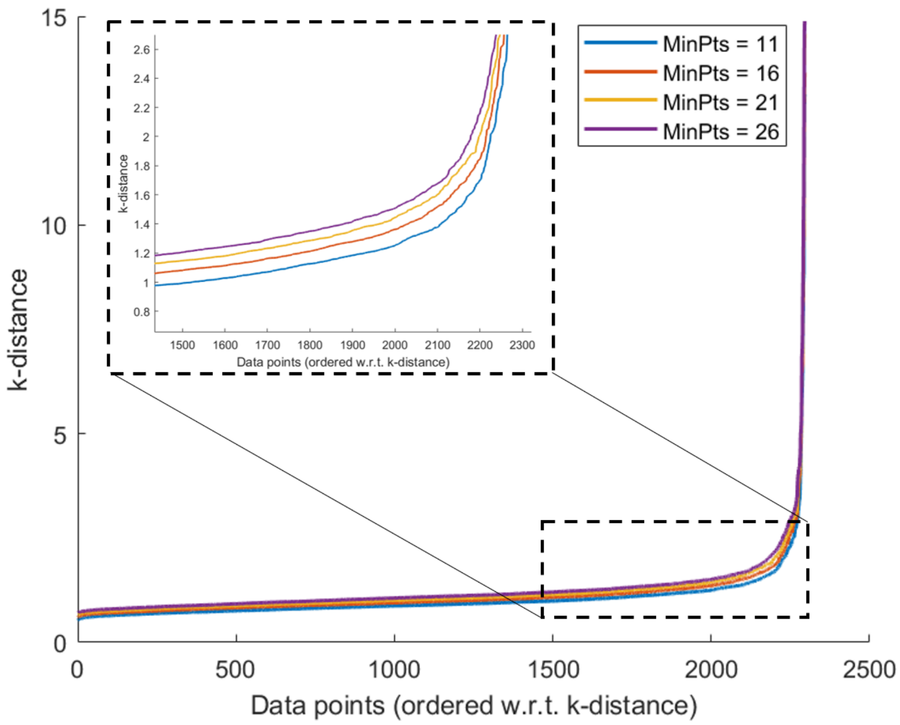


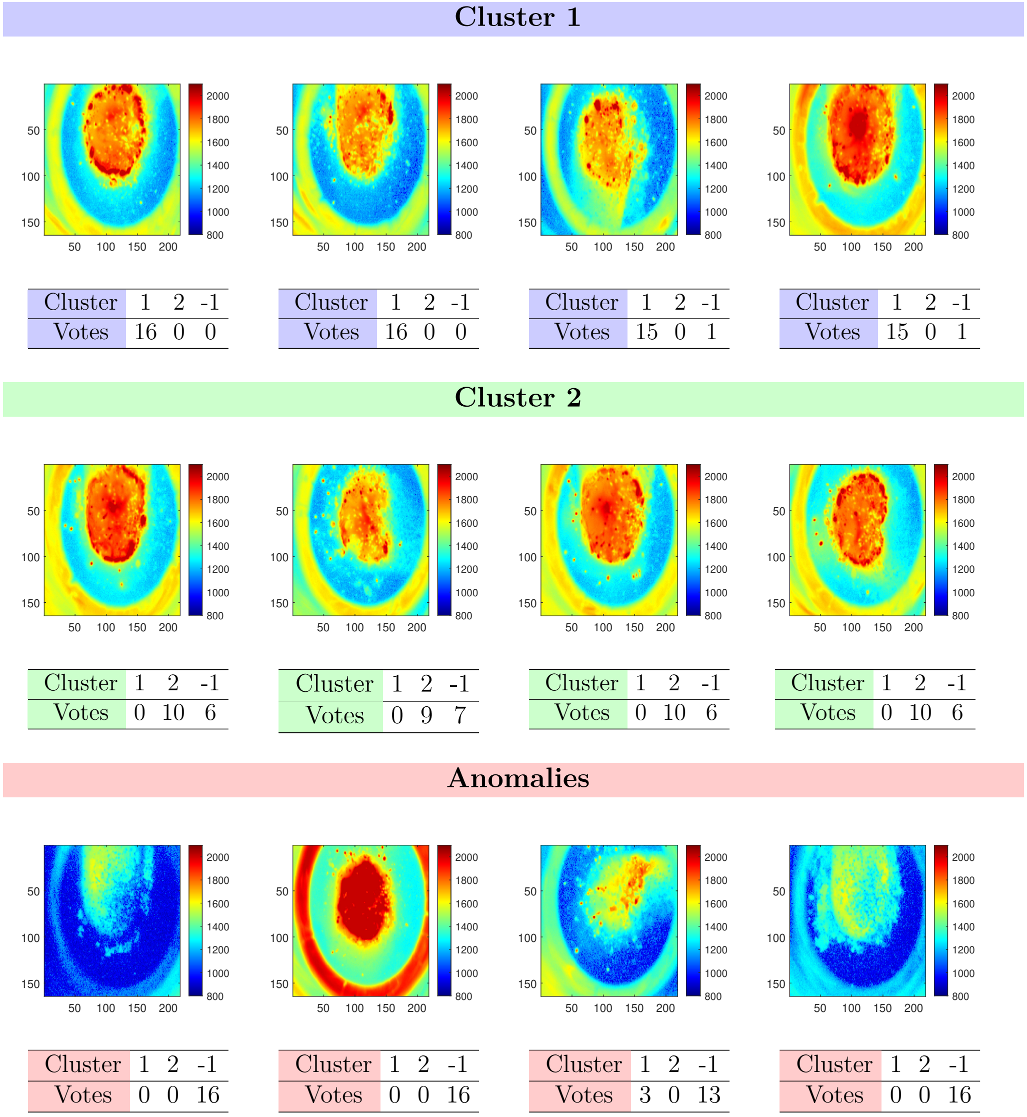
Figure 5.22 K-distance of data points in the DED-IN718-U dataset (ordered according to distance to k-nearest neighbors)

as anomalies by the DBSCAN ensemble. Given the dominance of Cluster 1 in terms of population, it is presumed to represent the baseline class—namely, defect-free or normal images. In contrast, the relatively small size of Cluster 2 suggests it may correspond to group anomalies instead of normal images. As such, Cluster 2 warrants additional investigation, which is addressed in the subsequent analysis.

Table 5.12 Results of the DBSCAN ensemble

Cluster	Cluster 1	Cluster 2	Anomalies
No. of images	2228	25	42

Postprocessing Analysis and Visualization: To visually assess the clustering results, Fig. 5.23 presents representative images randomly selected from each of the two identified clusters, as well as from the detected anomalies. Beneath each image, the number of DBSCAN models in the ensemble that assigned the corresponding label is provided. Upon inspection, the anomalous samples appear distinct from



Cluster 1

Cluster	1	2	-1
Votes	16	0	0

Cluster	1	2	-1
Votes	16	0	0

Cluster	1	2	-1
Votes	15	0	1

Cluster	1	2	-1
Votes	15	0	1

Cluster 2

Cluster	1	2	-1
Votes	0	10	6

Cluster	1	2	-1
Votes	0	9	7

Cluster	1	2	-1
Votes	0	10	6

Cluster	1	2	-1
Votes	0	10	6

Anomalies

Cluster	1	2	-1
Votes	0	0	16

Cluster	1	2	-1
Votes	0	0	16

Cluster	1	2	-1
Votes	3	0	13

Cluster	1	2	-1
Votes	0	0	16

Figure 5.23 Representative images from both identified clusters and the detected anomalies (Farea et al., 2024). The number of votes received from the DBSCAN ensemble for each cluster is indicated below each image, where a value of -1 denotes classification as an anomaly by the ensemble.

those in Cluster 1 (defect-free class). The anomalies exhibit either overheating (e.g., the second image in the anomalies row of Fig. 5.23) or underheating accompanied by visibly distorted melt pool shapes (e.g., the remaining images in the same row). Underheating often results in lack of fusion and contributes to porosity formation, whereas overheating can degrade surface quality and lead to diminished mechanical properties (Ranjan et al., 2023). The sample images from Cluster 2 reveal melt pools with either slight geometric distortions or moderate signs of overheating. Notably, these samples were classified as anomalous by at least six models within the ensemble. This consistent labeling suggests that Cluster 2 likely corresponds to a set of group anomalies—anomalous instances that form a coherent cluster.

As a postprocessing analysis, the clustering structure of the thermal images was further examined using a dimensionality reduction technique. The high-dimensional feature representations of the images were projected into a two-dimensional space via t-SNE, as shown in Fig. 5.24. This nonlinear technique preserves local structures in the data and offers an intuitive visualization of the clustering behavior. As depicted in Fig. 5.24, Cluster 1 and Cluster 2 exhibit clear separation, while the detected anomalies tend to reside along the boundaries of both clusters. This spatial configuration aligns with the earlier visual inspection of representative images in Fig. 5.23, reinforcing the interpretation that Cluster 1 corresponds to normal samples, Cluster 2 to group anomalies, and the remaining isolated points to individual (point) anomalies. Together, these observations support the effectiveness of the clustering approach in capturing distinct thermal behaviors within the dataset.

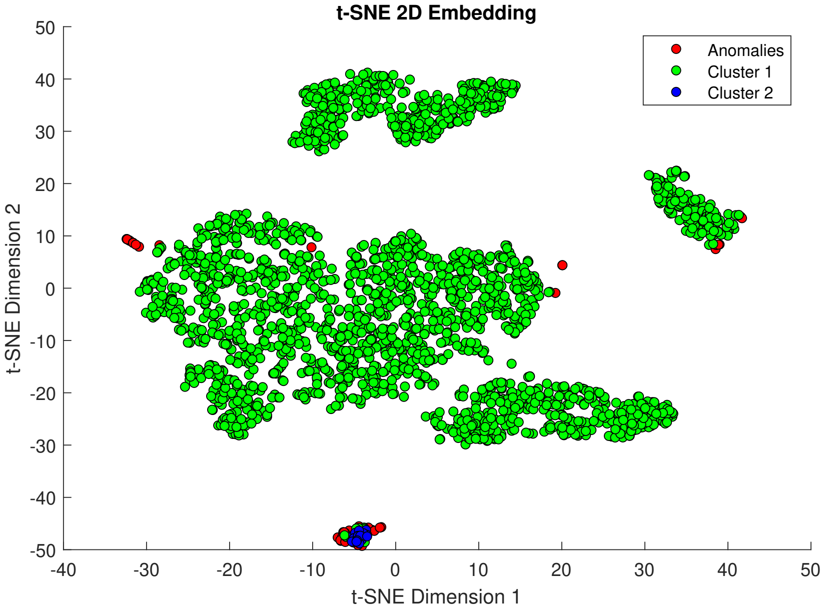


Figure 5.24 t-SNE visualization of the selected features in the DED-IN718-U dataset (the identified clusters and anomalies are displayed) (Farea et al., 2024)

6. CONCLUSION

This thesis addresses the critical challenge of anomaly detection by developing and evaluating a range of data-driven frameworks tailored to two distinct industrial applications: failure detection in air pressure systems (APS) of heavy-duty vehicles using operational data and defect detection in directed energy deposition (DED) processes through thermal imaging. The proposed frameworks encompass supervised, semi-supervised, and unsupervised learning paradigms, alongside explainable models utilized to enhance the transparency and interpretability of the anomaly detection process. Additionally, a domain knowledge-driven preprocessing pipeline—including steps such as data imputation, segmentation, and sliding window-based downsampling—was implemented to facilitate the extraction of informative features specific to each application. Subsequently, data-driven models were constructed to effectively distinguish between anomalous and normal instances based on these meticulously engineered features.

For APS failure detection, key features such as duty cycle, air compressor operating frequency, and minimum pressures across various braking circuits were identified as indicative of system health, according to domain expertise. The modeling phase involved both fully supervised models, including Random Forest and XGBoost, and semi-supervised transformer-based architectures, notably TranAD. Despite being trained in a one-class semi-supervised fashion using only normal data, TranAD achieved performance comparable to the fully supervised models, demonstrating remarkable data efficiency and rapid learning convergence. Specifically, when trained on a reduced dataset (only 20% of healthy data), TranAD maintained strong predictive performance with an accuracy of 91.4%, an F1 score of 0.79, and precision and recall values of 0.82 and 0.77, respectively. However, the primary limitation of the semi-supervised approach was its lack of interpretability, as the model was trained to reconstruct input sequences without providing transparent explanations for its predictions.

To address this limitation, an interpretable framework based on Explainable Boosting Machine (EBM) was employed for APS failure detection. EBM, a glass-box interpretable model, offered comprehensive insights into the decision-making process through feature importance rankings and local explanations for individual predictions. The explanations provided by EBM closely aligned with domain knowledge and were also corroborated by the explanations generated by SHAP (Shapley Additive Explanations), a robust XAI baseline. Despite its inherently interpretable nature, EBM delivered predictive performance comparable to black-box models, achieving an accuracy of 91.4% and F1 score of 0.80. Thus, it effectively balanced transparency and predictive accuracy, offering critical interpretability without sacrificing model performance. Additionally, EBM facilitated the identification of potential root causes of APS failures and highlighted specific features contributing to misclassified instances. These findings underscore the practical value of EBM in many anomaly detection applications where it generates valuable explanations about the decision-making process while maintaining strong predictive performance. However, EBM's reliance on fully labeled data restricts its applicability in contexts where anomaly data is sparse or prohibitively expensive to collect. Moreover, unlike transformer-based architectures, its static modeling nature constrains its ability to capture temporal dependencies inherent in sequential data, presenting a potential area for future improvement.

In addition, future work may focus on enhancing the explainability of semi-supervised models for APS failure detection to foster human-in-the-loop decision-making. Integrating post-hoc XAI techniques tailored for semi-supervised settings could provide actionable insights into model predictions, enabling domain experts to validate and refine model outputs effectively. They can even intervene in the modeling loop, e.g., by redesigning the input features or relabeling some of the data. Additionally, the proposed frameworks can be extended to other predictive maintenance applications across sectors such as automotive and aerospace, demonstrating their adaptability to diverse operational contexts. Furthermore, exploring the integration of remaining useful life estimation for APS components presents a promising avenue for future research.

This thesis also addressed defect detection in DED-manufactured IN718 and Ti–6Al–4V parts using thermal imaging, with each dataset presenting distinct challenges. The supervised datasets include single-type anomalies (i.e., porosities); however, the supervised IN718 (DED-IN718) dataset featured balanced class distributions, whereas the Ti–6Al–4V (DED-Ti64) dataset was characterized by significant class imbalance.

To effectively capture defect-related patterns, key features summarizing the geometry and thermal distribution of the melt pool were extracted from thermal images. Additionally, for the DED-IN718 dataset, spatio-temporal context was integrated to capture positional and temporal relationships among thermal images within the same deposited part, thereby enriching the feature set for subsequent modeling.

For the DED-IN718 dataset, supervised learning models—Random Forest and Support Vector Machine (SVM)—were employed to classify thermal images as normal or anomalous based on the extracted features. Given the balanced nature of the dataset, the supervised classifiers demonstrated robust performance, achieving an AUC of approximately 0.9, an F1 score of 0.83, and an accuracy of 84%. Notably, Random Forest exhibited superior robustness in reducing false positives, underscoring its effectiveness in detecting porosities in DED processes.

In contrast, the DED-Ti64 dataset required approaches capable of handling class imbalance. Accordingly, Isolation Forest (iForest) and one-class SVM were applied using both semi-supervised and unsupervised training paradigms. Optimal performance was achieved in the semi-supervised setting, yielding an F1 score of 0.88 and an accuracy of 95%. Across both training paradigms, iForest consistently outperformed one-class SVM, which showed strong sensitivity to data contamination during training. These results underscore the robustness of iForest's ensemble-based approach in detecting anomalies within imbalanced datasets, even in the presence of noisy or contaminated training data.

The current approach for defect detection leveraged expert-engineered features, effectively incorporating domain knowledge through tailored preprocessing steps. This methodology enabled the extraction of key indicators of porosity formation in DED processes. Future research could explore semi-supervised or self-supervised deep learning architectures to automate feature extraction while retaining a degree of interpretability. Extending this framework to other additive manufacturing technologies, such as powder bed fusion, would further demonstrate its generalizability and practical relevance in industrial settings. In addition, integrating real-time defect detection with closed-loop control systems could enable immediate corrective actions, enhancing process reliability. Future work could also broaden the scope beyond porosity detection to include the classification of multiple defect types—such as cracks and lack of fusion—enhancing the diagnostic capabilities of anomaly detection systems for safety-critical applications.

BIBLIOGRAPHY

Abati, D., Porrello, A., Calderara, S., & Cucchiara, R. (2019). Latent space autoregression for novelty detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (pp. 481–490).

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.

Ahn, D. G. (2021). Directed energy deposition (DED) process: State of the art. *International Journal of Precision Engineering and Manufacturing-Green Technology*, *8*(2), 703–742.

Al Samara, M., Bennis, I., Abouaissa, A., & Lorenz, P. (2023). Complete outlier detection and classification framework for WSNs based on OPTICS. *Journal of Network and Computer Applications*, *211*, 103563.

Andrews, J., Tanay, T., Morton, E. J., & Griffin, L. D. (2016). Transfer representation-learning for anomaly detection. In *Proceedings of the 33rd International Conference on Machine Learning*.

Assad, A., Bevans, B. D., Potter, W., Rao, P., Cormier, D., Deschamps, F., Hamilton, J. D., & Rivero, I. V. (2024). Process mapping and anomaly detection in laser wire directed energy deposition additive manufacturing using in-situ imaging and process-aware machine learning. *Materials & Design*, *245*, 113281.

Audibert, J., Michiardi, P., Guyard, F., Marti, S., & Zuluaga, M. A. (2020). USAD: Unsupervised anomaly detection on multivariate time series. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, (pp. 3395–3404).

Aydemir, E. (2024). *LiDAR based ground plane estimation, hybrid visual-LiDAR odometry and navigation of autonomous trucks*. PhD thesis, Sabanci University.

Bergman, L. & Hoshen, Y. (2020). Classification-based anomaly detection for general data. arXiv preprint. Retrieved from `https://arxiv.org/abs/2005.02359`.

Bergmann, P., Fauser, M., Sattlegger, D., & Steger, C. (2019). MVTec AD–A comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (pp. 9592–9600).

Braei, M. & Wagner, S. (2020). Anomaly detection in univariate time-series: A survey on the state-of-the-art. arXiv preprint. Retrieved from `https://arxiv.org/abs/2004.00433`.

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

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Cabello-Solorzano, K., Ortigosa de Araujo, I., Peña, M., Correia, L., & J. Tallón-Ballesteros, A. (2023). The impact of data normalization on the accuracy of machine learning algorithms: A comparative analysis. In *International*

Conference on Soft Computing Models in Industrial and Environmental Applications, (pp. 344–353).

Cerqueira, V., Pinto, F., Sá, C., & Soares, C. (2016). Combining boosted trees with metafeature engineering for predictive maintenance. In *Advances in Intelligent Data Analysis XV*, (pp. 393–397).

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys (CSUR)*, *41*(3), 1–58.

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.

Chen, L. & Moon, S. K. (2024). In-situ defect detection in laser-directed energy deposition with machine learning and multi-sensor fusion. *Journal of Mechanical Science and Technology*, *38*(9), 4477–4484.

Chen, L., Yao, X., Tan, C., He, W., Su, J., Weng, F., Chew, Y., Ng, N. P. H., & Moon, S. K. (2023). In-situ crack and keyhole pore detection in laser directed energy deposition through acoustic signal and deep learning. *Additive Manufacturing*, *69*, 103547.

Chen, T. & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (pp. 785–794).

Costa, C. F. & Nascimento, M. A. (2016). IDA 2016 industrial challenge: Using machine learning for predicting failures. In *Advances in Intelligent Data Analysis XV*, (pp. 381–386).

Cui, W., Zhang, Y., Zhang, X., Li, L., & Liou, F. (2020). Metal additive manufacturing parts inspection using convolutional neural network. *Applied Sciences*, *10*(2), 545.

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).

Dávila, J. L., Neto, P. I., Noritomi, P. Y., Coelho, R. T., & da Silva, J. V. L. (2020). Hybrid manufacturing: A review of the synergy between directed energy deposition and subtractive processes. *The International Journal of Advanced Manufacturing Technology*, *110*, 3377–3390.

Deecke, L., Vandermeulen, R., Ruff, L., Mandt, S., & Kloft, M. (2019). Image anomaly detection with generative adversarial networks. In *Machine Learning and Knowledge Discovery in Databases*, (pp. 3–17).

DMG MORI (2024). Lasertec 65 DED Hybrid. `https://en.dmgmori.com/products/machines/additive-manufacturing/powder-nozzle/lasertec-65-ded-hybrid`. Accessed: 19-12-2024.

Dong, F., Kong, L., Wang, H., Chen, Y., & Liang, X. (2023). Cross-section geometry prediction for laser metal deposition layer-based on multi-mode convolutional neural network and multi-sensor data fusion. *Journal of Manufacturing Processes*, *108*, 791–803.

ESPI Metals (2024). Inconel 718. `https://www.espimetals.com/index.php/technical-data/91-inconel-718`. Accessed: 12-12-2024.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, volume 96, (pp. 226–231).

European Commission (2021). Proposal for a regulation laying down harmonised rules on artificial intelligence. `https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence`. Accessed: 9-1-2023.

Fan, Y., Nowaczyk, S., Antonelo, E. A., et al. (2016). Predicting air compressor failures with echo state networks. In *PHM Society European Conference*, volume 3.

Fan, Y., Nowaczyk, S., & Rögnvaldsson, T. (2015a). Evaluation of self-organized approach for predicting compressor faults in a city bus fleet. *Procedia Computer Science*, *53*, 447–456.

Fan, Y., Nowaczyk, S., & Rögnvaldsson, T. S. (2015b). Incorporating expert knowledge into a self-organized approach for predicting compressor faults in a city bus fleet. In *SCAI*, (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., Cevik, E., Yilmaz, M., & Koprubasi, K. (2024). Prediction of failures in air pressure system: A semi-supervised framework based on transformers. In *2024 IEEE 22nd International Conference on Industrial Informatics (INDIN)*, (pp. 1–5).

Farea, S. M., Unel, M., & Koc, B. (2024). Defect prediction in directed energy deposition using an ensemble of clustering models. In *2024 IEEE 22nd International Conference on Industrial Informatics (INDIN)*, (pp. 1–6).

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189–1232.

Gaja, H. & Liou, F. (2017). Defects monitoring of laser metal deposition using acoustic emission sensor. *The International Journal of Advanced Manufacturing Technology*, *90*, 561–574.

Gaja, H. & Liou, F. (2018). Defect classification of laser metal deposition using logistic regression and artificial neural networks for pattern recognition. *The International Journal of Advanced Manufacturing Technology*, *94*, 315–326.

García-Moreno, A.-I. (2019). Automatic quantification of porosity using an intelligent classifier. *The International Journal of Advanced Manufacturing Technology*, *105*, 1883–1899.

Golan, I. & El-Yaniv, R. (2018). Deep anomaly detection using geometric transformations. *Advances in Neural Information Processing Systems*, *31*.

Gondek, C., Hafner, D., & Sampson, O. R. (2016). Prediction of failures in the air pressure system of Scania trucks using a random forest and feature engineering. In *Advances in Intelligent Data Analysis XV*, (pp. 398–402).

Guo, S., Guo, W. G., & Bain, L. (2020). Hierarchical spatial-temporal modeling and monitoring of melt pool evolution in laser-based additive manufacturing. *IISE Transactions*, *52*(9), 977–997.

Hariri, S., Kind, M. C., & Brunner, R. J. (2019). Extended isolation forest. *IEEE Transactions on Knowledge and Data Engineering*, *33*(4), 1479–1489.

Hejazi, M. & Singh, Y. P. (2013). One-class support vector machines approach to anomaly detection. *Applied Artificial Intelligence*, *27*(5), 351–366.

Hinton, G. E. & Roweis, S. (2002). Stochastic neighbor embedding. *Advances in Neural Information Processing Systems*, *15*.

Hussain, S. A., Prasad V, P., Kodali, R., Rapaka, L., & Sanki, P. K. (2024). Predicting and categorizing air pressure system failures in Scania trucks using machine learning. *Journal of Electronic Materials*, *53*, 3603–3613.

Ijaz, M. F., Attique, M., & Son, Y. (2020). Data-driven cervical cancer prediction model with outlier detection and over-sampling methods. *Sensors*, *20*(10), 2809.

Jablonski, J. A., Bihl, T. J., & Bauer, K. W. (2015). Principal component reconstruction error for hyperspectral anomaly detection. *IEEE Geoscience and Remote Sensing Letters*, *12*(8), 1725–1729.

Jose, C. & Gopakumar, G. (2019). An improved random forest algorithm for classification in an imbalanced dataset. In *2019 URSI Asia-Pacific Radio Science Conference (AP-RASC)*, (pp. 1–4).

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.

Khan, P. W., Yeun, C. Y., & Byun, Y. C. (2023). Fault detection of wind turbines using SCADA data and genetic algorithm-based ensemble learning. *Engineering Failure Analysis*, *148*, 107209.

Khanzadeh, M., Bian, L., Shamsaei, N., & Thompson, S. M. (2016). Porosity detection of laser based additive manufacturing using melt pool morphology clustering. In *2016 International Solid Freeform Fabrication Symposium*.

Khanzadeh, M., Chowdhury, S., Bian, L., & Tschopp, M. A. (2017). A methodology for predicting porosity from thermal imaging of melt pools in additive manufacturing thin wall sections. In *International Manufacturing Science and Engineering Conference*, volume 50732, (pp. V002T01A044).

Khanzadeh, M., Chowdhury, S., Marufuzzaman, M., Tschopp, M. A., & Bian, L. (2018). Porosity prediction: Supervised-learning of thermal history for direct laser deposition. *Journal of Manufacturing Systems*, *47*, 69–82.

Khanzadeh, M., Chowdhury, S., Tschopp, M. A., Doude, H. R., Marufuzzaman, M., & Bian, L. (2019). In-situ monitoring of melt pool images for porosity prediction in directed energy deposition processes. *IISE Transactions*, *51*(5), 437–455.

Khanzadeh, M., Tian, W., Yadollahi, A., Doude, H. R., Tschopp, M. A., & Bian, L. (2018). Dual process monitoring of metal-based additive manufacturing using tensor decomposition of thermal image streams. *Additive Manufacturing*, *23*, 443–456.

Kieu, T., Yang, B., Guo, C., & Jensen, C. S. (2019). Outlier detection for time series with recurrent autoencoder ensembles. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI-19)*, (pp. 2725–2732).

Kim, H., Seo, J., Chung Baek, A. M., Shin, W. Y., Jeon, H., Moon, S. K., Kim, H., Kim, N., & Jung, I. D. (2024). Direct energy deposition for smart micro reactor. *Virtual and Physical Prototyping*, *19*(1), e2411024.

Kim, S., Gholami, A., Shaw, A., Lee, N., Mangalam, K., Malik, J., Mahoney, M. W., & Keutzer, K. (2022). Squeezeformer: An efficient transformer for automatic speech recognition. *Advances in Neural Information Processing Systems*, *35*, 9361–9373.

Kind, A., Stoecklin, M. P., & Dimitropoulos, X. (2009). Histogram-based traffic anomaly detection. *IEEE Transactions on Network and Service Management*, *6*(2), 110–121.

Kruskal, W. H. & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, *47*(260), 583–621.

Landauer, M., Onder, S., Skopik, F., & Wurzenberger, M. (2023). Deep learning for anomaly detection in log data: A survey. *Machine Learning with Applications*, *12*, 100470.

Lang, C. I., Sun, F.-K., Lawler, B., Dillon, J., Al Dujaili, A., Ruth, J., Cardillo, P., Alfred, P., Bowers, A., Mckiernan, A., et al. (2022). One class process anomaly detection using kernel density estimation methods. *IEEE Transactions on Semiconductor Manufacturing*, *35*(3), 457–469.

Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

Li, C.-L., Sohn, K., Yoon, J., & Pfister, T. (2021). Cutpaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (pp. 9664–9674).

Li, D., Chen, D., Goh, J., & kiong Ng, S. (2019). Anomaly detection with generative adversarial networks for multivariate time series. arXiv preprint. Retrieved from `https://arxiv.org/abs/1809.04758`.

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.

Liang, H., Song, L., Wang, J., Guo, L., Li, X., & Liang, J. (2021). Robust unsupervised anomaly detection via multi-time scale DCGANs with forgetting mechanism for industrial multivariate time series. *Neurocomputing*, *423*, 444–462.

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, (pp. 413–422).

Liu, W., Luo, W., Lian, D., & Gao, S. (2018). Future frame prediction for anomaly detection–a new baseline. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 6536–6545).

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (pp. 10012–10022).

Lokesh, Y., Nikhil, K. S. S., Kumar, E. V., & Mohan, B. G. K. (2020). Truck APS failure detection using machine learning. In *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, (pp. 307–310).

Lu, W., Cheng, Y., Xiao, C., Chang, S., Huang, S., Liang, B., & Huang, T. (2017). Unsupervised sequential outlier detection with deep architectures. *IEEE Transactions on Image Processing*, *26*(9), 4321–4330.

Lundberg, S. M. & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30* (pp. 4765–4774). Curran Associates, Inc.

Ma, M., Wang, Z., & Zeng, X. (2015). Effect of energy input on microstructural evolution of direct laser fabricated IN718 alloy. *Materials Characterization*, *106*, 420–427.

Maldonado-Correa, J., Torres-Cabrera, J., Martín-Martínez, S., Artigao, E., &

Gómez-Lázaro, E. (2024). Wind turbine fault detection based on the transformer model using SCADA data. *Engineering Failure Analysis*, *162*, 108354.

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).

Muideen, A. A., Lee, C. K. M., Chan, J., Pang, B., & Alaka, H. (2023). Broad embedded logistic regression classifier for prediction of air pressure systems failure. *Mathematics*, *11*(4), 1014.

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).

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.

Munir, M., Siddiqui, S. A., Dengel, A., & Ahmed, S. (2019). DeepAnT: A deep learning approach for unsupervised anomaly detection in time series. *IEEE Access*, *7*, 1991–2005.

Napoletano, P., Piccoli, F., & Schettini, R. (2021). Semi-supervised anomaly detection for visual quality inspection. *Expert Systems with Applications*, *183*, 115275.

Nori, H., Jenkins, S., Koch, P., & Caruana, R. (2019). Interpretml: A unified framework for machine learning interpretability. arXiv preprint. Retrieved from `https://arxiv.org/abs/1909.09223`.

Noshad, Z., Javaid, N., Saba, T., Wadud, Z., Saleem, M. Q., Alzahrani, M. E., & Sheta, O. E. (2019). Fault detection in wireless sensor networks through the random forest classifier. *Sensors*, *19*(7), 1568.

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*, (pp. 205–214).

Ozan, E. C., Riabchenko, E., Kiranyaz, S., & Gabbouj, M. (2016). An optimized K-NN approach for classification on imbalanced datasets with missing data. In *Advances in Intelligent Data Analysis XV*, (pp. 387–392).

Panda, C. & Singh, T. R. (2023). ML-based vehicle downtime reduction: A case of air compressor failure detection. *Engineering Applications of Artificial Intelligence*, *122*, 106031.

Pang, G., Shen, C., Cao, L., & Hengel, A. V. D. (2021). Deep learning for anomaly detection: A review. *ACM Computing Surveys (CSUR)*, *54*(2), 1–38.

Pang, G., Yan, C., Shen, C., Hengel, A. v. d., & Bai, X. (2020). Self-trained deep ordinal regression for end-to-end video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (pp. 12173–12182).

Park, D., Hoshi, Y., & Kemp, C. C. (2018). A multimodal anomaly detector for robot-assisted feeding using an LSTM-based variational autoencoder. *IEEE Robotics and Automation Letters*, *3*(3), 1544–1551.

Park, H., Kang, Y. S., Choi, S.-K., & Park, H. W. (2025). Quality evaluation modeling of a DED-processed metallic deposition based on ResNet-50 with few training data. *Journal of Intelligent Manufacturing*, *36*(4), 2677–2693.

Parmar, A., Katariya, R., & Patel, V. (2019). A review on random forest: An ensemble classifier. In *International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018*, (pp. 758–763).

Patil, D. B., Nigam, A., Mohapatra, S., & Nikam, S. (2023). A deep learning approach to classify and detect defects in the components manufactured by laser directed energy deposition process. *Machines*, *11*(9).

Pokrajac, D., Lazarevic, A., & Latecki, L. J. (2007). Incremental local outlier detection for data streams. In *2007 IEEE Symposium on Computational Intelligence and Data Mining*, (pp. 504–515).

Prytz, R., Nowaczyk, S., Rögnvaldsson, T., & Byttner, S. (2013). Analysis of truck compressor failures based on logged vehicle data. In *9th International Conference on Data Mining*.

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.

Rafsunjani, S., Safa, R. S., Al Imran, A., Rahim, M. S., & Nandi, D. (2019). An empirical comparison of missing value imputation techniques on APS failure prediction. *International Journal of Information Technology and Computer Science*, *2*, 21–29.

Rahman, M. S. & Sumathy, V. (2024). Forecasting failure-prone air pressure systems (FFAPS) in vehicles using machine learning. *Automatika*, *65*(1), 1–13.

Ramadas, M., Ostermann, S., & Tjaden, B. (2003). Detecting anomalous network traffic with self-organizing maps. In *International Workshop on Recent Advances in Intrusion Detection*, (pp. 36–54).

Ranasinghe, G. D. & Parlikad, A. K. (2019). Generating real-valued failure data for prognostics under the conditions of limited data availability. In *2019 IEEE International Conference on Prognostics and Health Management (ICPHM)*, (pp. 1–8).

Ranjan, R., Chen, Z., Ayas, C., Langelaar, M., & Van Keulen, F. (2023). Overheating control in additive manufacturing using a 3D topology optimization method and experimental validation. *Additive Manufacturing*, *61*, 103339.

Ren, H., Xu, B., Wang, Y., Yi, C., Huang, C., Kou, X., Xing, T., Yang, M., Tong, J., & Zhang, Q. (2019). Time-series anomaly detection service at Microsoft. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, (pp. 3009–3017).

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).

Ribeiro, M., Gutoski, M., Lazzaretti, A. E., & Lopes, H. S. (2020). One-class classification in images and videos using a convolutional autoencoder with compact embedding. *IEEE Access*, *8*, 86520–86535.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (pp. 1135–1144).

Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Ruff, L., Kauffmann, J. R., Vandermeulen, R. A., Montavon, G., Samek, W., Kloft, M., Dietterich, T. G., & Müller, K.-R. (2021). A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, *109*(5), 756–795.

Ruff, L., Vandermeulen, R., Goernitz, N., Deecke, L., Siddiqui, S. A., Binder, A., Müller, E., & Kloft, M. (2018). Deep one-class classification. In *International Conference on Machine Learning*, (pp. 4393–4402).

Ruff, L., Vandermeulen, R. A., Gornitz, N., Binder, A., Muller, E., & Kloft, M. (2019). Deep support vector data description for unsupervised and semi-supervised anomaly detection. In *Proceedings of the ICML 2019 Workshop on Uncertainty and Robustness in Deep Learning*, (pp. 9–15).

Ruff, L., Vandermeulen, R. A., Görnitz, N., Binder, A., Müller, E., Müller, K.-R., & Kloft, M. (2020). Deep semi-supervised anomaly detection. arXiv preprint. Retrieved from `https://arxiv.org/abs/1906.02694`.

Sajid, S., Taras, A., & Chouinard, L. (2021). Defect detection in concrete plates with impulse-response test and statistical pattern recognition. *Mechanical Systems and Signal Processing*, *161*, 107948.

Scania CV AB (2016). APS Failure at Scania Trucks. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C51S51.

Schlegl, T., Seeböck, P., Waldstein, S. M., Langs, G., & Schmidt-Erfurth, U. (2019). f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks. *Medical Image Analysis*, *54*, 30–44.

Schlegl, T., Seeböck, P., Waldstein, S. M., Schmidt-Erfurth, U., & Langs, G. (2017). Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International Conference on Information Processing in Medical Imaging*, volume 10265, (pp. 146–157).

Schubert, E., Sander, J., Ester, M., Kriegel, H. P., & Xu, X. (2017). DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN. *ACM Transactions on Database Systems (TODS)*, *42*(3), 1–21.

Selvi, K. T., Praveena, N., Pratheksha, K., Ragunanthan, S., & Thamilselvan, R. (2022). Air pressure system failure prediction and classification in Scania trucks using machine learning. In *2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS)*, (pp. 220–227).

Shin, H., Lee, J., Choi, S.-K., & Lee, S. W. (2023). Development of multi-defect diagnosis algorithm for the directed energy deposition (DED) process with in situ melt-pool monitoring. *The International Journal of Advanced Manufacturing Technology*, *125*(1), 357–368.

Su, Y., Zhao, Y., Niu, C., Liu, R., Sun, W., & Pei, D. (2019). Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, (pp. 2828–2837).

Svetlizky, D., Das, M., Zheng, B., Vyatskikh, A. L., Bose, S., Bandyopadhyay, A., Schoenung, J. M., Lavernia, E. J., & Eliaz, N. (2021). Directed energy deposition (DED) additive manufacturing: Physical characteristics, defects, challenges and applications. *Materials Today*, *49*, 271–295.

Syed, M. N., Hassan, M. R., Ahmad, I., Hassan, M. M., & De Albuquerque, V. H. C.

(2020). A novel linear classifier for class imbalance data arising in failure-prone air pressure systems. *IEEE Access*, *9*, 4211–4222.

Tian, Q., Guo, S., Guo, Y., et al. (2020). A physics-driven deep learning model for process-porosity causal relationship and porosity prediction with interpretability in laser metal deposition. *CIRP Annals*, *69*(1), 205–208.

Tian, Q., Guo, S., Melder, E., Bian, L., & Guo, W. G. (2021). Deep learning-based data fusion method for in situ porosity detection in laser-based additive manufacturing. *Journal of Manufacturing Science and Engineering*, *143*(4), 041011.

Tsai, D.-M. & Jen, P.-H. (2021). Autoencoder-based anomaly detection for surface defect inspection. *Advanced Engineering Informatics*, *48*, 101272.

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.

Van der Maaten, L. & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*(11).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Wang, S., Zeng, Y., Liu, X., Zhu, E., Yin, J., Xu, C., & Kloft, M. (2019). Effective end-to-end unsupervised outlier detection via inlier priority of discriminative network. *Advances in Neural Information Processing Systems*, *32*.

Wolff, S. J., Wu, H., Parab, N., Zhao, C., Ehmann, K. F., Sun, T., & Cao, J. (2019). In-situ high-speed X-ray imaging of piezo-driven directed energy deposition additive manufacturing. *Scientific Reports*, *9*(1), 962.

Wu, J., Yang, F., & Hu, W. (2023). Unsupervised anomalous sound detection for industrial monitoring based on ArcFace classifier and Gaussian mixture model. *Applied Acoustics*, *203*, 109188.

Xie, M., Han, S., Tian, B., & Parvin, S. (2011). Anomaly detection in wireless sensor networks: A survey. *Journal of Network and computer Applications*, *34*(4), 1302–1325.

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

Yi, J. & Yoon, S. (2020). Patch SVDD: Patch-level SVDD for anomaly detection and segmentation. In *Proceedings of the Asian conference on computer vision*.

Zamanzadeh Darban, Z., Webb, G. I., Pan, S., Aggarwal, C., & Salehi, M. (2024). Deep learning for time series anomaly detection: A survey. *ACM Computing Surveys*, *57*(1), 1–42.

Zamiela, C., Tian, W., Guo, S., & Bian, L. (2023). Thermal-porosity characterization data of additively manufactured Ti–6Al–4V thin-walled structure via laser engineered net shaping. *Data in Brief*, *51*, 109722.

Zamouche, D., Aissani, S., Omar, M., & Mohammedi, M. (2023). Highly efficient approach for discordant BSMs detection in connected vehicles environment. *Wireless Networks*, *29*(1), 189–207.

Zenati, H., Foo, C. S., Lecouat, B., Manek, G., & Chandrasekhar, V. R. (2019). Efficient GAN-based anomaly detection. arXiv preprint. Retrieved from `https://arxiv.org/abs/1802.06222`.

Zenati, H., Romain, M., Foo, C.-S., Lecouat, B., & Chandrasekhar, V. (2018). Adversarially learned anomaly detection. In *2018 IEEE International Conference on Data Mining (ICDM)*, (pp. 727–736).

Zhang, B., Liu, S., & Shin, Y. C. (2019). In-process monitoring of porosity during laser additive manufacturing process. *Additive Manufacturing*, *28*, 497–505.

Zhang, C., Song, D., Chen, Y., Feng, X., Lumezanu, C., Cheng, W., Ni, J., Zong, B., Chen, H., & Chawla, N. V. (2019). A deep neural network for unsupervised anomaly detection and diagnosis in multivariate time series data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, (pp. 1409–1416).

Zhang, D., Qian, L., Mao, B., Huang, C., Huang, B., & Si, Y. (2018). A data-driven design for fault detection of wind turbines using random forests and XGboost. *IEEE Access*, *6*, 21020–21031.

Zhang, L., Lin, J., & Karim, R. (2018). Adaptive kernel density-based anomaly detection for nonlinear systems. *Knowledge-Based Systems*, *139*, 50–63.

Zhang, S., Lin, X., Wang, L., Yu, X., Hu, Y., Yang, H., Lei, L., & Huang, W. (2021). Strengthening mechanisms in selective laser-melted Inconel 718 superalloy. *Materials Science and Engineering: A*, *812*, 141145.

Zhang, Y., Lu, H., Zhang, L., & Ruan, X. (2016). Combining motion and appearance cues for anomaly detection. *Pattern Recognition*, *51*, 443–452.

Zhao, T., Wang, Y., Xu, T., Bakir, M., Cai, W., Wang, M., Dahmen, M., Zheng, Q., Wei, X., Hong, C., et al. (2021). Some factors affecting porosity in directed energy deposition of AlMgScZr-alloys. *Optics & Laser Technology*, *143*, 107337.

Zheng, F., Xie, L., Bai, Q., Zhu, Y., Yin, M., Zhang, Y., & Niu, K. (2024). Semi-supervised learning for laser directed energy deposition monitoring via co-axial dynamic imaging. *Additive Manufacturing*, *97*, 104628.

Zhou, X., Hu, Y., Liang, W., Ma, J., & Jin, Q. (2021). Variational LSTM enhanced anomaly detection for industrial big data. *IEEE Transactions on Industrial Informatics*, *17*(5), 3469–3477.

Zhu, K., Fuh, J. Y. H., & Lin, X. (2021). Metal-based additive manufacturing condition monitoring: A review on machine learning based approaches. *IEEE/ASME Transactions on Mechatronics*, *27*(5), 2495–2510.

Zong, B., Song, Q., Min, M. R., Cheng, W., Lumezanu, C., Cho, D., & Chen, H. (2018). Deep autoencoding Gaussian mixture model for unsupervised anomaly detection. In *International Conference on Learning Representations*.