Bridging the Gap: Enhancing ML Transparency with Explainable Artificial Intelligence

PhD Thesis

Omnia Amin

Department of Electronic and Electrical Engineering
University of Strathclyde, Glasgow

Friday 20th February, 2026

Abstract

Transparency in Machine Learning (ML) and Artificial Intelligence (AI) models is increasingly important for reliable decision-making in safety-critical industries. This thesis examines the role of Explainable Artificial Intelligence (XAI) methods in revealing the reasoning behind model outputs, in contrast to traditional black-box ML approaches.

Through a structured review and assessment of XAI tools, this research develops a methodology to support practitioners in applying explainability techniques in a practical context. The methodology is designed to bridge the gap between ML experts and non-specialists by adapting complex XAI outputs into formats that are more intuitive and accessible for domain experts, thereby enhancing interpretability and supporting informed decision-making.

A framework is introduced and demonstrated through a real-world case study in condition monitoring. The framework combines textual and visual explanations, presenting ML-generated insights in ways that can be more readily understood without reducing analytical value.

Validation of the framework is carried out using established principles and engineering knowledge to assess the consistency and usefulness of the transformed explanations. This focus on both accessibility and reliability aims to encourage wider adoption of XAI in engineering practice, addressing the need for clearer and more dependable ML insights in high-stakes environments.

Contents

List of Figures

List of Tables

List of Acronyms

AAKR Auto-Associative Kernel Regression

AD Anomaly Detection

AI Artificial Intelligence

ALE Accumulated Local Effects

ANCHOR Anchors

ANN Artificial Neural Network

ANRC Advanced Nuclear Research Centre

ARIMA Autoregressive Integrated Moving Average

ASV Asymmetric Shapley Values

BEP Best Efficiency Point

BERT Bidirectional Encoder Representations from Transformers

BHp Brake Horsepower

BRL Bayesian Rule Lists

CBM Condition-Based Maintenance

CCGT Combined Cycle Gas Turbine

CDSS Clinical Decision Support Systems

CEM Contrastive Explanation Method

CFX Counterfactual Explanations Framework

CLIP Contrastive Language–Image Pre-training

CM Condition Monitoring

CNN Convolutional Neural Network

ConvNet Convolutional Neural Network

DALEX Descriptive mAchine Learning EXplanations

DCG Dynamic Convolutional Graph

DCG-CNN Dynamic Convolutional Graph Convolutional Neural Network

DeepLIFT Deep Learning Important FeaTures

DiCE Diverse Counterfactual Explanations

DL Deep Learning

DNN Deep Neural Networks

DP Differential Privacy

DR-CNN Deep Residual Convolutional Neural Network

EBMs Explainable Boosting Machines

EDA Exploratory Data Analysis

ES Expert Systems

FACE Feasible and Actionable Counterfactual Explanations

FL Federated Learning

GA Genetic Algorithm

GAMs Generalized Additive Models

GAN Generative Adversarial Network

GDPR General Data Protection Regulation

GLocalX GLObal to loCAL eXplainer

GPT Generative Pre-trained Transformer

GPT3 Generative Pre-trained Transformer 3

GRAD-CAM Gradient-weighted Class Activation Mapping

GPAI General-Purpose AI

HMM Hidden Markov Models

HSE Health and Safety Executive

ICE Individual Conditional Expectation

ICO Information Commissioner's Office

InTrees Interpretable Trees

IoT Internet of Things

ISO International Organisation for Standardization

IVHM Integrated Vehicle Health Management

KBS Knowledge-Based Systems

kNN k-Nearest Neighbor

LIME Local Interpretable Model-Agnostic Explanations

LIT Language Interpretability Tool

LLMs Large Language Models

LRP Layer-wise Relevance Propagation

LRI Logical Rule Induction

LSTM Long Short-Term Memory

ML Machine Learning

MKML-ISOMAP Multi-Kernel Metric Learning Embedded Isometric Mapping

MMD-CRITIC Maximum Mean Discrepancy Criticism

MRI Magnetic Resonance Imaging

MSE Mean Squared Error

MUSE Model Understanding through Subspace Explanations

NAM Neural Additive Models

NIST National Institute of Standards and Technology

NLP Natural Language Processing

NLU Natural Language Understanding

NPL National Physical Laboratory

OECD Organisation for Economic Co-operation and Development

PaLM Pathways Language Model

PCA Principal Component Analysis

PdM Predictive Maintenance

PD Partial Dependence

PIMP Permutation Feature Importance

PROTODASH Prototypical Explanations via Faster Search

PV Photovoltaic

RNN Recurrent Neural Network

RM Reactive Maintenance

RMF Risk Management Framework

RMSE Root Mean Square Error

ROPE RObust Post-hoc Explanations

RULEFIT Rule-based Model Fitting

SBRL Scalable Bayesian Rule Lists

SCALABLE-PRL Scalable Probabilistic Rule Learning

SHAP SHapley Additive exPlanations

SimpleGrad Simple Gradient-Based Method

SM Scheduled Maintenance

SmoothGrad Smooth Gradient

SVM Support Vector Machine

T5 Text-To-Text Transfer Transformer

TCAV Testing with Concept Activation Vectors

TraCE Training Calibration-based Explainers

TPU Tensor Processing Units

TSM Trust Scores

T-SNE T-distributed Stochastic Neighbor Embedding

UNESCO United Nations Educational, Scientific and Cultural Organisation

VGG Visual Geometry Group

VGGNet Visual Geometry Group Network

XAI Explainable Artificial Intelligence

XGBoost Extreme Gradient Boosting

XRAI eXplainable Region-based Activation Integration

List of Symbols

Symbol	Definition
General Notation	
n, N	Number of observations in the dataset (scalars)
$\binom{n}{2}$	Total number of unordered pairs from n elements
$\mathbf{X}, \mathbf{Y}$	Ranking vectors used in Kendall's τ
x_i, y_i	Scalar elements of vectors $\mathbf{X}$ and $\mathbf{Y}$
$\bar{y}$	Mean of the observed values y_i
$\hat{y}_i$	Predicted value for the i-th observation
Error Metrics	
MSE	Mean Squared Error
$RMSE$	Root Mean Squared Error
R^2	Coefficient of determination
Kendall's Tau	
τ	Kendall's rank correlation coefficient
C	Number of concordant pairs
D	Number of discordant pairs
Pump Affinity Laws	
H_1, H_2	Total head at initial/final point
N_1, N_2	Initial and final speed/condition
P_1, P_2	Power consumption at initial/final speed
Q_1, Q_2	Flow rate at initial/final speed

Shapley Values

N	Set of all features (in Shapley value formulation)
S	Subset of features not including feature i
$\mathcal{S}_i$	Randomly chosen subset of features excluding i (Monte Carlo)
ϕ_i	Shapley value for feature i
$\hat{\phi}_i$	Estimated Shapley value for feature i
M	Number of Monte Carlo samples
$\mathbf{x}$	Feature vector
x_i^m	m-th sampled value of feature i (Monte Carlo)
$v(S)$	Model prediction using only features in subset S
$v(S \cup \{i\})$	Model prediction when feature i is added to subset S

Preface/Acknowledgements

I would like to express my deepest gratitude to my supervisory team for their unwavering support and invaluable guidance throughout my research journey. I am really thankful to Prof. Stephen McArthur for his insightful direction and for always encouraging me to explore and push the boundaries of my research.

Special thanks to Blair, my primary supervisor, whose persistent efforts and constructive feedback have been instrumental in shaping my research and bringing out the best in me.

I am also grateful to Bruce, who, despite his incredibly busy schedule, found time to offer his valuable advice and support, contributing significantly to my research and personal growth.

I sincerely express gratitude for the support provided by the UK's National Physical Laboratory (NPL) and the Advanced Nuclear Research Centre (ANRC)at the University of Strathclyde, Glasgow. Special thanks go to Dr. Valerie Livina for her continuous support and guidance throughout this work. Additionally, heartfelt appreciation is extended to Dr. Michael Chrubasik for serving as a valuable mentor during the industrial placement period.

My heartfelt appreciation extends to my family, whose love, encouragement, and sacrifices have been my constant source of strength and inspiration. To my parents, Zeinab and Izzeldin, words fall short to express my gratitude. Without your unwavering support and endless sacrifices, I would not be where I am today.

To my sisters, Israa, Abia, Allaa, and Omayma, along with their families, thank you for being my pillars of strength. Your continuous support and encouragement, especially during the challenging times our country faced, have been my beacon of hope, urging me to persevere and excel.

To my brothers, Yousif and Kamil, and their families, thank you for your enduring support and for being there for me every step of my journey. Your belief in my abilities has been a constant source of motivation.

My deepest love and appreciation go to my sons, Abdallah and Izzeldin. You have been my rock through this journey. Despite your young age, your understanding, patience, and words of encouragement, "Mom is the best," have been the source of my strength and determination. You are the reason I aspire to be better every day.

I am also immensely grateful to my colleagues for being supportive listeners and for enduring my weekly moaning sessions. Their encouragement and companionship have been invaluable throughout this journey.

Last but certainly not least, I extend my profound gratitude to my husband, Mohamed. Your kind words, unwavering support, and belief in me have been my guiding light. Your presence and encouragement have made this journey not only possible but also a memorable and enriching experience.

To each one of you who has been a part of my journey, thank you from the bottom of my heart. Your support has shaped me into the researcher and individual I am today, and for that, I am eternally grateful.

Chapter 1

Introduction

Developing AI and ML models that exhibit strong performance yet lack transparency in revealing their predictive mechanisms falls short of meeting the transparency demands in high-stakes industry applications. Stakeholders in such domains are increasingly demanding greater clarity regarding how and why ML models generate predictions. In sectors characterised by high-stakes concerns, such as the nuclear sector, entrusting black-box ML models raises inherent risks. This underscores the imperative of understanding the inner workings of AI and ML models, as well as the logic behind their predictions, to ensure their seamless integration across high-stakes industry applications.

Addressing this necessitates dedicated research into designing and deploying explainability tools to explain the decision-making pathways of models. Additionally, it involves investigating how these outcomes can be interpreted and deployed by subject matter experts without data science backgrounds, fostering trust in these critical decisions.

There is a need to apply XAI in condition monitoring to ensure that AI's maintenance predictions are transparent and understandable, fostering trust and enabling more effective decision-making. Providing explanations to domain experts in condition monitoring will help in preventing equipment failures, reducing downtime, and optimising maintenance tasks, leading to significant operational efficiencies and cost savings.

The next section will serve as an introduction to this research. It will start by defining the problem statement, outlining the different phases of the research, and highlighting the key contributions it aims to make.

1.1 Introduction to the Research

Recently, AI and ML models have found their way into various sectors of our society [5], starting from industries where errors have minimal impact, such as entertainment, and extending to critical domains like health and nuclear energy, blindly relying on black-box ML models for decision-making poses significant risks [6] [7].

A considerable number of ML models operate as black-boxes, concealing the decision-making process and mechanisms behind their predictions. AI and ML applications use black-box models because their outputs provide better performance than using white-box models; however, they are difficult to understand. In black-box models often stakeholders, including end-users, subject matter and decision-makers [8], cannot understand the underlying structure, and how and why predictions are made [9]. By understanding how ML models behave, stakeholders will be able to better detect errors, foresee possible bias learned by their models, and eventually be able to decide when and how they will deploy these ML models [7]. As an example, in the nuclear industry, stakeholders can't blindly trust predictions generated by these applications. Hence, they are demanding more transparency in order to trust these predictions when deploying these models in their industry [10]. XAI techniques are intended to enable stakeholders to understand how ML models work and why they generate their predictions [11]. As more complex ML models are proposed, developed, and deployed, the need for XAI tools to understand predictions made by these models also grows [12].

A new rule is to be enforced by the General Data Protection Regulation (GDPR): "a right to an explanation of all decisions made by automated or artificially intelligent algorithmic systems" [13], which means that there are additional requirements for the technology's deployment.

The increased usage of black-box models has been described in applications controlled by complex algorithms and the severe consequences of a lack of transparency, such as possible biases in decision-making and unfair competition [14] [15]. Therefore, blindly relying on black-box ML models to make decisions in safety-critical sectors is unacceptable, because serious safety-related consequences can occur [6]. Many well-known companies enforce explainability in AI-powered applications to pass trustworthiness checks [14]. XAI techniques are intended to provide credibility to stakeholders by understanding how ML models work and why they generate predictions [11].

Explainability and trust are different concepts, but are tied together in the literature to the extent that researchers assume a direct connection. For example, in [16], the authors developed a new explainability and trust hypothesis, where explainability is one approach to establishing stakeholders' trust. As a natural consequence of more complex ML models being proposed, developed, and deployed alongside new regulations, the need for XAI tools to understand predictions has also increased.

Despite the recent increase in interest in XAI, much of the current research and practice relies on the researcher's subjective judgment of what constitutes a "good" explanation rather than a rigorous and objective standard [17].

1.2 Problem Statement

As ML technologies become more common in various fields, a significant challenge has emerged: making these complex technologies understandable to people who aren't experts in ML, such as domain specialists and engineering professionals. These individuals play a crucial role in applying and evaluating ML models in real-world settings, but often find it difficult to interpret and trust the results these models produce. A large contributing factor is that most ML research and development have focused more on improving the technical aspects of models, like their accuracy, and less on explaining how these models work in a way that non-experts can easily understand.

This has led to a gap between the advanced nature of ML algorithms and the capacity of domain experts to make sense of them. This gap makes it challenging for these professionals to fully use ML technologies in their work, limiting the potential benefits of ML, such as making better decisions, improving efficiency, and fostering innovation.

Additionally, there's a lack of methods and tools that provide explanations suited to the needs of non-expert users, which is crucial for building trust and collaboration between

ML developers and domain experts.

Exploring how to make ML technologies easier for non-experts raises important questions. These questions guide the research's direction, focusing on developing XAI methods that bridge the gap between complex ML algorithms and their practical application by domain experts. Addressing these questions is essential for enhancing the usability and transparency of ML technologies across various sectors. The following research questions have been formulated:

1. **How can the most appropriate XAI tools be systematically identified and selected for different types of ML models and domain-specific applications?**

2. **What methods can be employed to structure XAI explanations so that clarity and interpretability are ensured for domain experts with limited ML expertise in engineering applications?**

3. **Can the explanations generated by XAI methods be validated, using objective metrics, to determine whether model insights are effectively conveyed to domain experts?**

1.3 Novelty of the Research and Contributions

The research presents a novel approach to bridging the knowledge gap between ML developers and domain experts, particularly those without a deep understanding of ML. Unlike much existing research that often overlooks the end-user's perspective, this work focuses on delivering user-friendly explanations tailored to the needs of engineering specialists and domain experts across the entire life cycle of ML models. The novelty lies in the development of a structured methodology that integrates explainability into the AI system's life cycle from conception to deployment. This methodology emphasises identifying crucial areas for explainability and selecting the right XAI tools and techniques that align with the specific needs and contexts of domain experts. Through this approach, the research aims to enhance the practicality and accessibility of AI and ML

technologies, ensuring they are effectively tailored to the needs of non-ML experts.
The literature on explainability taxonomies is diverse, with many prior works offering partial, overlapping, or domain-specific classifications. This creates a fragmented landscape that can make it challenging to compare approaches or identify their scope. Building on this body of work, the thesis proposes a structured taxonomy organised along five complementary dimensions: input data type, type of explanation, timing of application, explanation technique, and model specificity. Rather than imposing a rigid hierarchy, this framework aims to provide a clearer way of navigating existing XAI methods, highlighting their relationships and overlaps while accommodating the hybrid nature of modern approaches. The proposed taxonomy is informed by a critical review of earlier classification efforts and extends them by presenting categories side by side for clarity, supported by tables that map well-known XAI tools to their respective dimensions.
Another innovative aspect of this research is the emphasis on facilitating communication of explanations to domain experts, addressing the distinct gap between ML model developers and the intended end-users. By focusing on comprehensive explanations during both the development and implementation phases of ML models, the research aims to empower domain experts with the knowledge to make informed decisions and actively participate in the ML model life cycle. This effort promotes a more collaborative environment, fostering better outcomes and benefits for the fields and industries involved.

The contributions of this research extend beyond the provision of accessible explanations. A key advancement is the engagement in a verification process to ensure the accuracy of XAI tool explanations through alignment with domain-specific knowledge. This research introduces a collection of innovative algorithms designed to generate explanations in both text-based and visual formats, tailored to the comprehension levels of non-experts. These algorithms aim to clarify the inner workings of ML models and the XAI models used to explain the ML model, enabling individuals without a background in ML to grasp complex concepts more easily.

By tailoring explanations to the target audience and employing intuitive visualisations, the research effectively bridges the gap between complex ML concepts and the understanding of domain experts. This has led to the development of a framework that not only encompasses the proposed methodology but also its practical application and thorough evaluation within specific real-world contexts. Through detailed case studies and empirical results, the research demonstrates the benefits and enhanced performance achieved by implementing the proposed methodology. This evidence-based approach underscores the value and applicability of the framework, highlighting its potential to empower domain experts and facilitate meaningful collaboration with ML developers.

In summary, the contributions of this research lie in its ability to make AI and ML technologies more accessible and understandable to domain experts, thereby enhancing collaboration, transparency, trust, and accountability in the use of ML models.
To ensure coherence and traceability, Table 1.1 explicitly maps each research contribution to the corresponding research question (RQ1–RQ3). This alignment demonstrates how the thesis systematically addresses the identified research gaps and fulfils its overarching objective of bridging the divide between ML developers and domain experts in safety-critical engineering applications.

Contributions	**Related RQ(s)**	**Impact and Benefits**
Comprehensive taxonomy for selecting and implementing XAI tools.	RQ1	Establishes a structured basis for systematically identifying and selecting appropriate XAI tools according to data type, model characteristics, and domain-specific requirements, thereby enabling transparent and trustworthy AI deployment.
Structured methodology for designing explainable ML systems.	RQ1, RQ2	Provides a systematic process to guide practitioners in integrating explainability throughout the ML lifecycle, ensuring explanations are technically sound and aligned with domain expert needs in critical infrastructure contexts.
User-centred explanation design and provision of user-friendly explanations.	RQ2	Tailors explanation formats to domain experts with limited ML expertise, enhancing clarity, interpretability, and practical usability in engineering environments.
Adaptive and reusable XAI framework for engineering applications.	RQ1–RQ3	Demonstrates integrated tool selection, structured explanation design, and validation across real-world case studies, promoting transparency, accountability, and cross-domain applicability.
Validation of XAI explanations through alignment with domain expertise.	RQ3	Introduces objective validation mechanisms by cross-referencing explanations with established engineering knowledge, increasing reliability, robustness, and stakeholder trust.

Table 1.1: Contributions of the Research and Their Alignment with the Research Questions

1.4 Research Phases

Our study aims to make AI systems in industrial settings easier to understand. It's divided into seven phases, as shown in Table 1.2. Initially, the pressing need for XAI is identified, setting the stage for an in-depth investigation into the details of ML explainability. This exploration encompasses the critical evaluation of existing methodologies, the significance of explainability, and its challenges within the AI domain.

Exploring various XAI tools and understanding the challenges associated with each allowed for the development of a robust methodology.
Progressing to practical applications, three case studies are leveraged to illustrate the theoretical principles in action, highlighting the benefits of applying XAI tools in operational contexts. Central to the methodology is the development of a user-centric framework for XAI system design. This framework, initially employed in constructing foundational stages for explainability, undergoes rigorous application across three case studies, laying the groundwork for the subsequent creation of an innovative XAI framework.
This novel framework aims to make AI operations easier to understand for non-expert users in industrial environments, encouraging a deeper, more intuitive grasp of AI model predictions. Through a pivotal real-world condition monitoring case study, the framework's capability to render AI predictions both accessible and trustworthy is demonstrated.
The culmination of this research not only showcases the successful application of the framework but also sets the direction for future exploration. Broadening the framework's applicability to diverse engineering domains is envisaged, thereby enhancing the reliability and accessibility of AI systems across various industrial sectors. These research phases, from theoretical groundwork to practical application and future exploration, underscore our approach to integrating explainability into AI, as detailed in Table 1.2

1.5 Thesis Structure

This Thesis is organised to provide a thorough understanding of the research conducted. In Chapter 1, the foundation of the study is laid out by introducing the research problem and outlining the contributions. This chapter lays the foundation for the upcoming discussions.
Chapter 2 provides a comprehensive overview of explainability in the context of AI systems. It begins by defining explainability and outlining its growing necessity, especially as machine learning models become more complex and widely adopted. The chapter

Table 1.2: Research Phases

Stage	Description
1	Identifying research problem & objectives.
2	Deep dive into XAI techniques (SHAP, LIME, etc.).
3	Study the practical applications of ML in condition monitoring. Finding suitable case studies to facilitate the use of XAI.
4	Identify the methodology for crafting explainable AI systems (Factors every practitioner should contemplate when designing XAI for infrastructure).
5	The foundational stages, including data preprocessing, ML model selection, XAI tool application, and user-friendly explanation provision, have been applied to two case studies.
6	A novel framework has been devised from the foundational stages outlined in the fifth phase, ensuring comprehensive, user-friendly explanations whenever applied, catering to domain experts' needs.
7	The framework has been applied to real condition monitoring data and validated using a suitable metric against domain knowledge to demonstrate its reliability.

then discusses the evolution of interpretability, followed by an analysis of existing AI regulations and their implications for explainable AI (XAI), including a review of key regulatory frameworks and their specific requirements. Challenges such as trade-offs, constraints, and the needs of various explanation recipients are addressed, alongside considerations for non-expert users. The chapter further explores how explainability contributes to AI trustworthiness through validation techniques and discusses vulnerabilities such as adversarial attacks. Finally, it illustrates practical applications of explainability across both high-risk and lower-risk domains, highlighting the critical role of transparency in fields like healthcare, finance, and condition monitoring.

Chapter 3 provides a taxonomy of XAI techniques, categorising methods by input data types, explanation levels (global, cohort, local), model specificity (agnostic vs. specific), and timing (pre-model, during-model, post-hoc). With a focus on SHAP, detailing its theoretical foundations and comparing it with other XAI tools. The chapter also outlines the XAI classification range and addresses challenges for applying XAI in advanced AI models, highlighting SHAP's significance in the broader XAI landscape. Chapter 4 explores condition monitoring techniques, focusing on their application to centrifugal pumps and rotating machinery. Metrics such as performance, efficiency, and operation relative to the Best Efficiency Point (BEP) are examined, along with challenges associated with operating outside the BEP. Traditional condition monitoring methods are reviewed alongside ML-based approaches, including supervised, semi-supervised, unsupervised, and deep learning techniques. The role of XAI in enhancing interpretability and trust in ML-driven condition monitoring systems is also discussed, with a focus on its relevance to high-risk engineering and maintenance domains.

This chapter serves to establish foundational knowledge on condition monitoring, specifically for centrifugal pumps, by addressing the history, key concepts, and existing techniques in the field. Its inclusion is justified by the reliance on real operational data from thrust bearing wear for the primary case study, which necessitates an understanding of condition monitoring principles and their application to centrifugal pump systems.

Chapter 5 outlines the key considerations for designing explainable ML systems. It provides a blueprint for aligning user needs with appropriate XAI tools and describes foundational stages in developing an innovative XAI framework. The chapter applies these concepts through two case studies: a Combined Cycle Gas Turbine and a Gearbox in a Boiler Feed Pump. These case studies are conducted to demonstrate and facilitate the application of XAI methods and ensure the reproducibility of the methodology for the main case study presented in the subsequent chapter. This chapter bridges theoretical foundations with practical implementation, demonstrating XAI's real-world utility.

Chapter 6 combines the development of a novel framework for generating understandable explanations with its application to a real-world condition monitoring case study. The framework is designed to bridge the gap between complex ML outputs and the needs of non-expert users in industrial and engineering contexts. It emphasises the validation of explanations generated by XAI tools to ensure reliability and usability, and it evaluates the framework's effectiveness in practical settings.

The chapter also applies this framework to a thrust-bearing wear case study, demonstrating the proposed approach in a condition monitoring scenario. It details the data pre-processing, modelling, application of XAI tools, and generation of human-understandable explanations, including graphical representations. Furthermore, explanations are validated against domain knowledge, incorporating pump affinity laws and addressing uncertainty and limitations. This combined discussion provides a cohesive link between the theoretical development of the framework and its practical application, offering a comprehensive approach to enhancing explainability in high-stakes engineering applications.

Chapter 7 serves as the concluding chapter of this thesis, where findings are comprehensively discussed, and meaningful conclusions are drawn from the research endeavours. Additionally, future research directions are outlined, emphasising the potential for further advancements in the ever-evolving field of XAI. This concluding chapter not only brings closure to the present research but also extends an invitation to future exploration and development in the field of XAI.

1.6 Associated Publications

- Amin, O., Brown, B., Stephen, B., & McArthur, S.. (2022). A Case-study Led Investigation of XAI to Support Deployment of Prognostics in industry. In PHM Society European Conference, 7(1), 9–20. (See Chapter 5)

- Amin, O., Brown, B., Stephen, B., McArthur, S., & Livina, V. (2023). ML explanations by design: a case study explaining the predicted degradation of a roto-dynamic pump. Paper presented at NDT 2023 - 60th Annual British Conference on NDT, Northampton, United Kingdom (See Chapter 5).

- Amin O, Brown B, Stephen B, Livina V, McArthur S. (2025). Domain-adapted explainability for machine learning predictions of rotodynamic pump degradation in safety-critical industrial sectors. Proceedings of the Institution of Mechanical Engineers, Part A: Journal of Power and Energy, 240(1), 61–84. https://doi.org/10.1177/09576509251387765 (see Chapter 6).

- Blair, J., Amin, O., Brown, B., McArthur, S., Forbes, A., Stephen, B. (2024). The transfer learning of uncertainty quantification for industrial plant fault diagnosis system design. Data-Centric Engineering, 5. https://doi.org/10.1017/dce.2024.54 The contribution in the above journal involved developing and integrating the SHAP-based explainability framework into the proposed methodology. Specifically, SHAP was applied to the data pipeline configurations, and human understandable explanations of model outputs were derived. This facilitated the ranking and interpretation of pipeline components, enabling the identification of positive and negative performance drivers and enhancing the transparency of the diagnostic system.

1.7 Impact of COVID-19 on Research Progress

Life, academia, and research worldwide have been significantly affected by the COVID-19 pandemic. In our study, challenges were encountered due to restrictions on movement and in-person interactions. Adaptations had to be made to collaboration and communication methods, shifting them to virtual platforms. This acknowledgement reflects the broader impact of the pandemic on academia. Additionally, collaboration in the academic community was reduced by the pandemic, necessitating more flexible research strategies to overcome challenges.

Chapter 2

An Overview of Explainability

This chapter introduces various definitions of explainability and highlights the significance of explainability in the context of ML. It also explores state-of-the-art XAI tools, discussing their capabilities and limitations. Furthermore, the chapter identifies the advantages and disadvantages associated with these methods.

2.1 Definitions of Explainability

Explainability has been referred to in the literature by various terminologies such as "Interpretability" [14], "Understandability" [18], "Comprehensibility" [19], "Accountability" and "Responsible AI" [20]. Depending on the explanations provided for the target users, definitions of these terminologies might differ [21]. Generally, in ML literature, the term "Interpretability" is used more often than the term "Explainability" [22].

One of the specific definitions for XAI was given by [23], "XAI typically refers to post hoc analysis and techniques used to understand a previously trained model or its predictions". Another definition was given in [24]: "XAI will create a suite of ML techniques that enable human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners".

Explainability has been defined as the idea that a model's behaviours can be communicated using terminology that is easily understandable to individuals who are making

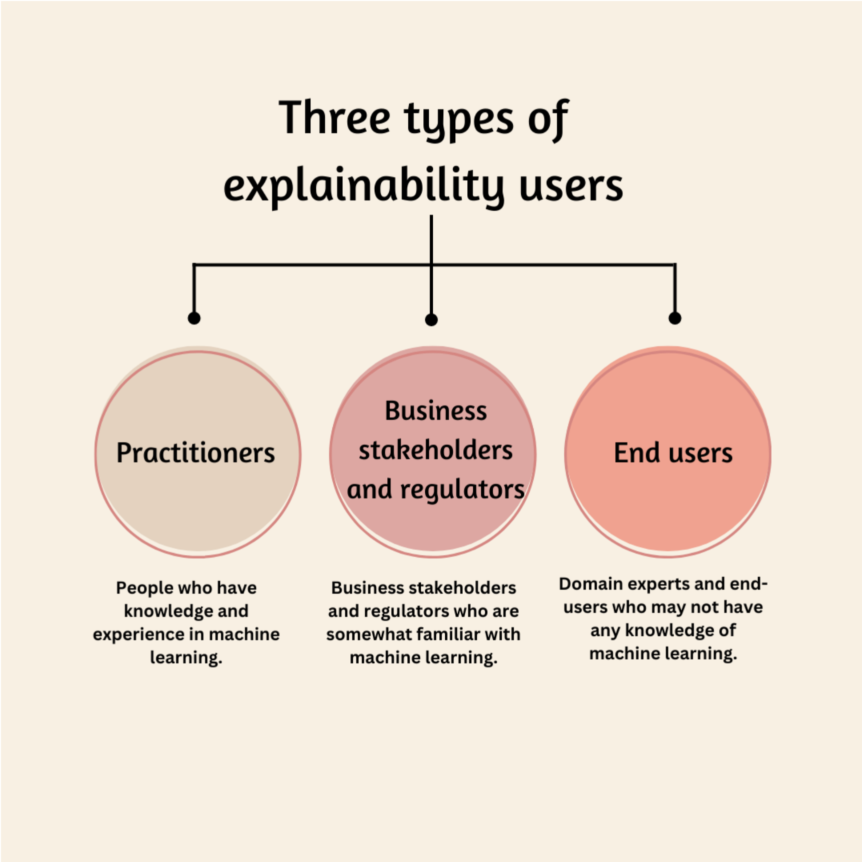


Figure 2.1: Types of users of explainability.

use of the predictions produced by the ML model and therefore has divided consumers of XAI into three groups of individuals who desire explanations with varying degrees of familiarity with ML, each group has its reasons and requirements shown in Figure 2.1 [25]: Practitioners who have enough ML expertise to build and design ML and AI models; business stakeholders and regulators; and end users who are domain experts with no ML expertise. In Chapter 4, an important aspect to consider is how to match explainability tools to meet the specific needs of each individual or stakeholder involved in the engineering industrial application have been discussed. This customisation is crucial to ensure the explanations provided by XAI tools are tailored to the knowledge, context, and requirements of the recipients.

XAI allows users to gain insight into why certain predictions are made. In [11], the Authors suggested that the system may or may not be explainable depending on the definition of explainability for that system. Explainability has been defined as making sure that predictions produced by ML models are explainable to non-experts [11].

Interpretable models refer directly to transparent models and are sometimes referred to as white-box models. A model can be defined as interpretable when it can provide accurate solutions to humans in an understandable way. In [26] and [27], both authors gave similar definitions for model interpretability as the degree to which humans can understand the logic behind generated predictions. Another interesting definition was given by [28]: "when you can no longer keep asking why". Achieving complete interpretability for a specific modelling problem can prove challenging. It is only when a clear explanation for all the model outputs generated in a task is obtained that complete interpretability can be claimed, a feat that is often elusive in practice [29]. Authors in [30] argued that, rather than using XAI tools to interpret black-box models, trying to build interpretable models, especially in high-stakes applications, is always the preferred solution.

2.2 The Need for Explainability

In some low-risk domains, in which there are no major consequences for incorrect or inaccurate predictions, having good accuracy in general is simply enough to trust the model. For these models, there is no need for providing explanations for the obtained predictions, e.g., adverts appearing while using Facebook, movie recommendations, etc. However, in high-risk domain tasks, where predictions can have a severe impact, knowing why a model has produced a certain output is important. It can help stakeholders (e.g. researchers, end users, regulators, business owners) to understand the task better and improve the prediction model by transferring the learned predictions into a broader knowledge base [31]. Stakeholders can then use this additional insight to choose whether to accept the outputs produced from the ML model. Figure 2.2 shows the different reasons why explainability is needed in AI.

Safety can be critically important for high-risk domain tasks, and safety insights can be an added value through the application of ML Interpretability. Interpretability enables ML models to go through testing, validation and debugging, hence increasing the safety

Figure 2.2: The need for Explainability

of their application.

The increased usage of black-box models that are controlled by “secret algorithms” and the serious consequences of a lack of transparency have been described by [14] [15]. These serious consequences can be possible human manipulation of the stock market, corrupt businesses, unfair competition, intentionally or unintentionally biased systems and many more.

A real-life example to demonstrate why interpretability is crucial: A predictive model used by Amazon to determine areas where they will apply “same day delivery” service was found to eliminate restricted minority areas from being in the program unintentionally, despite allowing neighbourhood areas to participate [32].

The need for interpretability is caused by formalising an incomplete problem -i.e., having small and insufficient training datasets, incomplete end-to-end testing, a biased system against specific classes or an incomplete goal, are all examples of incomplete problems [33].

2.3 Evolution of Machine Learning and Interpretability

Machine learning (ML) has advanced dramatically over the past decades, progressing from simple linear models to complex multilayered architectures that now underpin state-of-the-art systems in healthcare, finance, and natural language processing. These developments have brought unprecedented accuracy and versatility, but they have also introduced major challenges in interpretability.

The term machine learning is widely attributed to Arthur Samuel [34], who in 1959 described it as enabling computers to improve performance through experience, without explicit programming.
Figure 2.3 illustrates key milestones in the historical progression of ML, from early computational logic to the emergence of neural networks and modern applications [1].

The journey of machine learning has unfolded through a series of milestones that shaped today's AI systems. Early foundations were laid with the perceptron [35], one of the first neural network models, and the development of decision trees [36] and Hidden Markov Models (HMM), which became essential for tasks such as classification and speech recognition.

The 1980s marked a turning point with the backpropagation algorithm [37], which made it possible to train multilayer perceptrons and sparked renewed interest in neural networks. This was followed by support vector machines (SVMs) [38], which proved powerful for classification in high-dimensional spaces.

Breakthroughs in the late 1990s and early 2000s, including Long Short-Term Memory networks (LSTM) [39], Convolutional Neural Network (CNN) [40], and Random Forests [41], provided robust tools for sequential data, vision, and ensemble learning.

The deep learning revolution of the 2010s transformed the field. Authors in [42] demonstrated the strength of deep CNNs for large-scale image recognition, while Visual Geometry Group Network (VGGNet) [43] and Generative Adversarial Network (GAN) [44] pushed boundaries in vision and generative modelling. The introduction of the Transformer architecture [45] redefined natural language processing, enabling models such as BERT [46] and GPT-3 [47].

Most recently, AI has expanded into multimodal and generative models, including CLIP [48], DALL-E [49], ChatGPT [50], and open-source image generators like Stable Diffusion (2022). These advances illustrate the rapid progression from simple models to systems capable of generating text, images, and beyond, reshaping both research and industry.

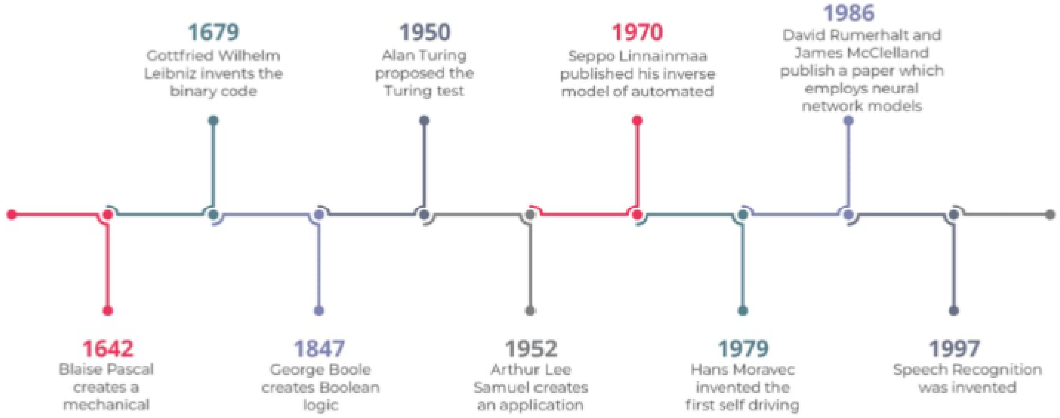


Figure 2.3: Key milestones in the history of machine learning [1].

The early era of ML was characterised by inherently interpretable models such as linear regression, logistic regression, and decision trees. These models were valued for their simplicity, efficiency, and transparency, as users could readily trace how inputs translated into outputs. Linear regression, for instance, directly mapped features to predictions [51], while decision trees decomposed decisions into intuitive, human-readable rules [52]. Such transparency aligned well with domains requiring clarity and accountability, including finance and policy-making. However, their inability to capture complex, non-linear patterns in high-dimensional data spurred the development of more sophisticated approaches.

As datasets became larger and more complex, the need for higher predictive power drove the rise of black-box algorithms. Ensemble methods, such as random forests and gradient boosting, aggregate multiple weak learners to achieve strong performance but at the expense of interpretability [53]. Similarly, deep neural networks excel at modelling non-linear relationships, yet their layered architecture and vast parameter space obscure the decision-making process [54]. The opacity of these models is particularly problematic in high-stakes applications such as healthcare, autonomous systems, and criminal justice, where understanding the rationale behind predictions is critical.

This shift towards opaque models has raised several interpretability challenges. Neural networks, for example, learn abstract representations that are difficult to map to real-world concepts [18], often producing outputs that appear non-intuitive to users [14]. Issues of bias and fairness further complicate trust, as models trained on unrepresentative datasets may yield discriminatory outcomes [55]. Regulatory frameworks, including the EU AI Act and the GDPR, have amplified these concerns by emphasising transparency and mandating a right to explanation for automated decisions [13].

To address these concerns, modern interpretability techniques have developed in parallel with advances in ML. Post-hoc methods such as SHAP [56] and LIME [57] provide model-agnostic explanations by highlighting feature contributions or approximating local decision boundaries. SHAP leverages cooperative game theory to assign consistent feature importance values, while LIME employs simpler surrogate models to approximate complex decision regions. At the model level, attention mechanisms have become integral to transformer-based architectures such as BERT [46] and Generative Pre-trained Transformer (GPT) [58], enabling models to selectively focus on relevant inputs. Visualising attention weights offers a degree of inherent interpretability by revealing which elements most influenced predictions [45].

In addition, efforts to design inherently interpretable architectures have gained momentum. Generalized Additive Models (GAMs) [59] and Explainable Boosting Machines (EBMs) [60] retain intuitive feature-level contributions while offering competitive accuracy, making them suitable for sensitive domains such as healthcare [61]. Meanwhile, the rise of generative models, including GANs and diffusion models, has created new interpretability challenges, prompting the development of advanced techniques to trace generative processes and identify influential components [62].

The push for interpretable AI is therefore not merely a technical concern but also a societal one [52]. Stakeholders, including domain experts, regulators, and end-users, demand transparency to trust and effectively use ML systems [33]. Techniques such as attention visualisation and SHAP exemplify efforts to demystify advanced models.

However, achieving interpretability without compromising performance remains a delicate balance. As ML continues to evolve, interpretability will remain central to its ethical deployment, ensuring that AI systems are not only accurate and efficient but also understandable, fair, and accountable [9].

2.4 AI Regulations and Their Implications for XAI

As AI systems become more integrated into critical areas like healthcare, finance, and law enforcement, regulators and policymakers around the world have recognised the need to respond to a wide range of ethical, legal, and social challenges [63]. These challenges include ensuring safety, preventing bias and discrimination, protecting data and privacy, upholding human rights, and managing the broader impact of automation on jobs and society [64] [65]. Among these concerns, the lack of transparency in complex AI models—especially those based on deep learning—has made explainability a key ethical and technical priority [20].

2.4.1 Overview of Key AI Regulatory Frameworks

Efforts to regulate artificial intelligence have gained momentum worldwide, with the European Union, the United Kingdom, and the United States each adopting distinct approaches. The European Union's AI Act, proposed in 2021, represents one of the first attempts to establish a comprehensive, legally binding framework for AI. It adopts a risk-based model, categorising AI applications according to the potential harm they pose and imposing stricter requirements on higher-risk systems [66].

In contrast, the United Kingdom has pursued a more decentralised and innovation-oriented strategy. Rather than enacting a single overarching law, the UK government has emphasised sector-specific guidance, encouraging regulators in areas such as healthcare, finance, and data protection to apply core principles—including accountability, transparency, and fairness—within their domains [67].

The United States has also advanced its own regulatory framework through the National Institute of Standards and Technology (NIST), which published the AI Risk

Management Framework (RMF) in 2023. The RMF is a voluntary set of guidelines designed to help organisations design, develop, and deploy AI systems that are safe, trustworthy, and socially responsible [68].

Although these initiatives differ in scope and emphasis, they share a common recognition of the challenges posed by opaque AI systems. Importantly, they frame explainability not as an isolated goal but as part of a broader regulatory agenda—one that seeks to align AI technologies with ethical principles, legal standards, and public values [67] [69] [66].

UK Government's National AI Strategy

The UK Government has adopted a cross-sector, outcomes-based framework for AI regulation that is guided by five core principles: safety, security and robustness, transparency and explainability, fairness, accountability and governance, and contestability and redress. Of these, transparency and explainability are directly aligned with the objectives of fostering interpretable and trustworthy AI systems. Rather than prescribing a single piece of legislation, the framework empowers sector-specific regulators to apply these principles within their existing legal mandates, with some expected to publish annual AI strategies. Although the approach has not yet been formalised into binding law, regulators are able to issue supplementary guidance, and there are clear indications that statutory development may follow—particularly in response to the emergence of General-Purpose AI (GPAI) models and gaps in cross-sector risk management [67].

The framework distinguishes between three categories of advanced AI: highly capable GPAI, highly capable narrow AI, and agentic AI. Large Language Models (LLMs) fall under the GPAI category, reflecting their broad applicability across domains [67]. Ahead of the UK AI Safety Summit, leading AI companies voluntarily committed to safety and transparency measures, including independent model evaluations in collaboration with the UK's newly established AI Safety Institute [70]. These voluntary initiatives strengthen sectoral oversight but also highlight the growing need for more specialised and enforceable regulatory instruments as AI capabilities continue to advance.

For the purposes of this research—focused on explainability in industrial condition monitoring—the most relevant regulatory authority is the Information Commissioner's Office (ICO), which oversees the use of AI in contexts involving personal data [71]. Depending on the sector, additional regulators such as the Health and Safety Executive (HSE) or Ofgem may also be directly relevant [72]. In such domains, regulatory considerations extend beyond compliance to encompass the practical requirement that AI models must be interpretable to human users, particularly when decisions affect safety, accountability, or obligations under the UK's Data Protection Act and the Equality Act [73].

The European Approach to AI Regulation

The European Union's approach to AI governance, set out in the AI Act introduced in April 2021, forms a cornerstone of the EU's wider Digital Strategy. The Act establishes a risk-based framework that classifies AI systems into four categories: minimal risk, limited risk, high risk, and unacceptable risk [74]. Systems deemed unacceptable—such as those enabling social scoring or manipulative behavioural targeting—are prohibited outright, with only narrow exceptions for law enforcement. High-risk AI systems, including those used in critical sectors such as healthcare, energy, and transportation or those affecting fundamental rights, face stringent legal obligations. These include mandatory risk management, data governance, human oversight, transparency, and post-market monitoring. In addition, all high-risk AI systems must be registered in a publicly accessible EU database [66].

Unlike the UK's voluntary and sector-led framework, the EU AI Act adopts a binding, horizontal structure that centralises enforcement through national market surveillance authorities and coordinated oversight at the EU level. Providers of high-risk AI are required to carry out conformity assessments and maintain technical documentation demonstrating compliance [75]. Generative AI systems, including LLMs, are subject to additional transparency rules, such as mandatory labelling of AI-generated content and the publication of training data summaries to mitigate risks of misuse and rights infringement [76, 77].

Whereas the UK framework articulates high-level principles—such as transparency and explainability—without codifying them into law, the EU AI Act creates enforceable obligations backed by penalties for non-compliance. The Act entered into force in 2024, with full applicability expected by 2026, although certain provisions will take effect sooner [78, 79]. This divergence underscores a broader contrast between the EU's precautionary, legally binding approach and the UK's more flexible, principle-based model of AI regulation [66, 67].

The US Approach to AI Regulation

The regulatory landscape for artificial intelligence in the United States is characterised by a principle-driven and decentralised approach, with a strong emphasis on fostering "trustworthy AI". Central to this effort is the National Institute of Standards and Technology (NIST)'s Artificial Intelligence Risk Management Framework (AI RMF 1.0), which identifies explainability and interpretability as key properties of trustworthy systems. Within the framework, explainability refers to the extent to which the internal mechanics of an AI system can be understood, while interpretability captures the ability to make sense of the system's outputs within their context of use [80]. These concepts are reinforced in federal directives such as Executive Order 13960, which requires that AI applications in the public sector be "sufficiently understandable" to stakeholders and that agencies provide meaningful disclosures about their use of AI to the public and oversight bodies [81].

The U.S. approach was further strengthened by Executive Order 14110, issued in October 2023, which directs federal agencies to appoint Chief AI Officers and to promote the secure, responsible, and transparent use of AI technologies [82]. While the United States has not enacted a comprehensive statutory framework comparable to the European Union's AI Act, it has relied on a combination of non-binding guidance, policy frameworks, and executive actions to steer AI governance. This flexible model enables rapid adaptation to technological advances, while embedding explainability and interpretability into federal standards, procurement practices, and oversight mechanisms. In doing so, the U.S. contributes to the broader international movement towards ensuring that AI systems are transparent, accountable, and aligned with societal values [80, 81, 82].

Global Approach to AI Regulation

Alongside regional initiatives, a number of global organisations have advanced frameworks to guide the responsible development and use of artificial intelligence. These efforts consistently emphasise transparency, accountability, and explainability as essential elements of trustworthy AI.

In 2019, the Organisation for Economic Co-operation and Development (OECD) released its Recommendation of the Council on AI, which set out principles to promote AI that is innovative, trustworthy, and aligned with human rights and democratic values [69]. Building on this, the United Nations Educational, Scientific and Cultural Organisation (UNESCO) introduced its Recommendation on the Ethics of AI in 2021—the first global framework dedicated to ethical AI use [83]. The UNESCO guidelines stress the importance of transparency, fairness, and accountability, and call for AI systems to respect human dignity while minimising risks, particularly for vulnerable groups. Explainability is highlighted as a means of fostering trust and enabling informed oversight.

In parallel, the International Organisation for Standardization (ISO), through the ISO/IEC JTC 1/SC 42 committee, has been developing global technical standards for AI. These standards address issues such as transparency, fairness, accountability, and data privacy, with the aim of providing practical benchmarks that help organisations implement safe and reliable AI systems in line with ethical principles [84].

Together, these international initiatives underscore a growing consensus: that explainability is not only a technical feature of AI but also a cornerstone of global efforts to ensure that AI development is consistent with ethical values, human rights, and public trust.

2.4.2 Explainability Requirements in Regulatory Frameworks

Growing interest in AI has raised concerns about the need for transparency and explainability, especially as regulations develop. Frameworks like the EU AI Act [66] and ISO 42001 [84] require or encourage explainability for high-risk AI to ensure user safety and trust. While explainable AI aids compliance and boosts consumer confidence, it also poses challenges by extending development time and potentially revealing proprietary details. To manage these demands, developers can adopt tiered explainability, aligned with model risk, and prioritise simplicity in model design and user explanations.

A key emphasis in IEEE 2894 [85] is the integration of human oversight and interpretability into AI systems. The standard recommends that explanation methods should exhibit properties such as fidelity, comprehensibility, and usefulness. It also outlines criteria for evaluating explanation effectiveness using metrics tied to the user's understanding, trust calibration, and decision-making improvement. These provisions reflect the broader regulatory ambition of ensuring that non-expert users can interpret, challenge, and act upon AI-generated outcomes. From a compliance perspective, the standard advises developers to balance explainability with system performance, protect proprietary algorithms, and consider tiered explanation strategies based on model risk levels. Collectively, these standards reinforce the principle that explainability is not merely a technical add-on but a core enabler of ethical, legal, and operational accountability in AI systems. Beyond explainability, these frameworks emphasise the importance of human oversight and interpretability. Regulators expect AI outputs to be accessible to non-expert users, enabling those impacted by AI-generated outcomes to review and understand the decision-making processes. This supports the broader regulatory goals of transparency and accountability, ensuring that human users can interpret and, if needed, challenge AI decisions. By upholding these standards, regulatory bodies aim to foster public trust and ensure AI systems operate ethically and within societal norms [84] [66] [83].

2.4.3 Obstacles and Constraints in XAI

The domain of XAI presents a range of benefits, yet it is not without its share of challenges and constraints. These encompass the inherent complexity of AI models, which can pose difficulties in offering comprehensible explanations. Furthermore, the absence of established standards for XAI hinders the ability to effectively compare and assess various XAI methods. Balancing accuracy, complexity, and interpretability presents another challenge in XAI, often requiring trade-offs. Ethical concerns also arise, with a need to ensure that the provided explanations do not inadvertently perpetuate biases or discrimination [86].

Regulatory frameworks such as the EU AI Act [66], the UK AI White Paper [87], and OECD AI Principles [88] place specific constraints on the design, implementation, and deployment of explainable AI (XAI) systems. These constraints aim to ensure that AI systems are trustworthy, transparent, and aligned with human oversight.

Firstly, regulatory guidelines require that AI models, especially those deployed in high-risk applications, must provide explanations that are understandable by affected stakeholders, not only technical experts [89]. This requirement directly impacts the choice of algorithms and explanation techniques: complex black-box models without adequate post-hoc explanation mechanisms may be non-compliant in regulated domains.

Secondly, frameworks stress the principles of accountability and contestability [90]. This implies that explanations must be sufficiently detailed and auditable, allowing end-users to challenge and contest AI-driven decisions. As a result, XAI systems must include mechanisms to log decisions and generate reproducible explanations for individual predictions.

Thirdly, some regulations establish constraints on data minimisation and fairness [13]. These impact the extent to which explanation systems can access and use sensitive data. XAI designers must balance explanation fidelity with privacy constraints, ensuring that explanations do not inadvertently expose protected attributes or personal data.

In summary, regulatory frameworks create explicit design constraints that shape XAI system architecture, favouring models that offer interpretability by design or robust post-hoc explanations, while ensuring explanations are accessible, auditable, and compliant with privacy and fairness principles.

The effectiveness of XAI in implementing a model hinges upon the intentions of those seeking explanations. However, the current presentation of explanations by explainability tools often remains incomprehensible to inexperienced users. There should be an intermediary phase post-application of these explanations before presenting them to users. Additionally, there exist challenges regarding the validation of XAI explanations against domain knowledge relevant to specific applications. Is there adequate domain expertise available to authenticate these explanations?. Furthermore, another obstacle lies in the usability of explainability tools, which often require expertise in ML, making them less accessible to non-experts [17].

2.5 Explainability Trade-offs

Although XAI offers several advantages, it does not come without inherent trade-offs. A comprehensive understanding of the constraints associated with various XAI methods becomes crucial for discerning the contexts where specific methods might excel or prove more precise than others. Figure 2.4 illustrates the trade-offs involved in explainability.

This section delves into the complex interplay among explainability, accuracy, security, privacy, and fairness within the realm of AI. Exploring the intricate balance between explainability and crucial facets like accuracy, security, privacy, and fairness is pivotal in understanding the multifaceted trade-offs within AI models. Presently, the conventional trade-off believed to exist between the accuracy and interpretability of predictive models appears to have been dismantled—or perhaps it never truly existed to begin with.

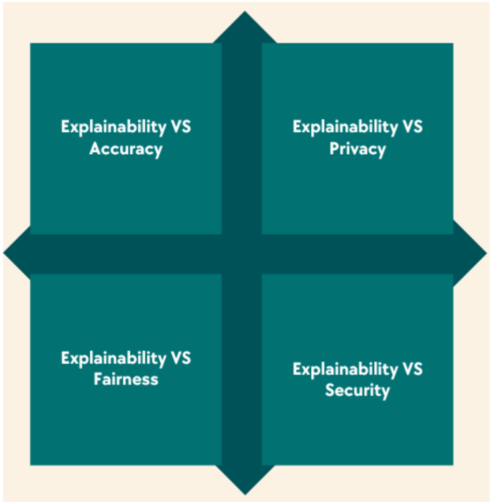


Figure 2.4: Explainaility Trade-Offs

Recent advancements in technology [91] have equipped us with the means to construct highly accurate and intricate modelling systems, leveraging diverse datasets and sophisticated ML algorithms. Simultaneously, these tools facilitate human comprehension and trust in these complex systems. In essence, the current technology allows us to achieve both accuracy and interpretability in our models, enabling us to enjoy the benefits of both without compromising either [23].

The increased emphasis on data privacy, enforced by regulatory frameworks such as GDPR, poses a significant challenge to explainability in real-world systems. The necessity for privacy can often restrict the effectiveness of explanations within these systems. The trade-off between the quality of explanations and model privacy remains complex, especially considering models are typically trained on a mix of private and non-private data. Without intervention, private data might unintentionally infiltrate model explanations, complicating the task of adjusting explanations to filter out private information and potentially leading to incomplete explanations that sacrifice accuracy. Recent advancements in research aim to address this trade-off by employing encryption and novel privacy-preserving ML techniques. However, these methods frequently introduce an added computational burden. Nevertheless, ongoing progress in computational capabilities remains instrumental in mitigating these associated costs [92].

In the fields of AI and ML, the correlation between explainability and security stands as a pivotal consideration. [93] explores the diverse ways in which XAI can significantly enhance cybersecurity, addressing a broad spectrum of aspects from user engagement to specialized applications like threat analysis and vulnerability reduction. As an increasing number of businesses adopt ML and XAI, there arises a parallel concern: the susceptibility of these systems to cyberattacks. This vulnerability extends not only to the AI models themselves but also to the interpretability mechanisms that make their decisions understandable. In the subsequent sections, the realm of adversarial attacks will be explored, with an emphasis on their relevance and potential impact. Moreover, the field of XAI presents ample opportunities for research and development, particularly in enhancing the security of the explanations provided by these AI systems, ensuring they are robust against manipulation and misinterpretation.

When examining the intertwined concepts of fairness and bias in various applications, fairness, in a broader sense, entails equitable treatment of similar entities, regardless of their affiliation with specific demographic groups. Though this concept seems intuitive, achieving a strict mathematical formulation demands certain theoretical assumptions. Hence, multiple definitions of fairness exist, tailored to meet the needs of specific applications [94]. Nevertheless, a common thread among these definitions emphasises the shared objective of minimising the deviation or statistical bias across individuals or particular groups of interest.

2.6 Exploring Explanation Recipients Content and Purpose

Different kinds of explanations shed light on various aspects of the model, covering details about the learned representations, strategies for making predictions, and the overall behaviour of the model. Depending on who will receive the explanations and their goals, it could be beneficial to emphasise a particular type of explanation.

In addition to considering the recipient and the type of information conveyed, it is essential to understand the purpose behind providing an explanation. Explanations serve diverse objectives beyond simply visualizing or validating predictions. For instance, explanations can support compliance with regulatory requirements or improve the practical utility of models in real-world applications, making them more aligned with stakeholders' needs [95]. Additionally, as explanations furnish insights into the significant aspects of the model, they could also be harnessed for tasks like model compression and pruning. While numerous other applications (model certification, legal compliance) of explanations can be envisioned, the specifics of these applications remain a subject for future investigation.

For example, when AI is used in medicine, these groups could be patients, doctors, and hospitals. The AI might give simple explanations to patients, like pointing out high blood sugar, while giving more detailed explanations to doctors, like showing unusual connections between different blood tests. Hospitals or organisations like the FDA might not need to know about individual patients, but instead, they might want to see general patterns that the AI has learned from looking at many patients' data [96].

Different types of condition monitoring insights illuminate various aspects of a system's operational health, encompassing details from direct sensor readings, and trend analyses, to predictive behaviours of the machinery. The specificity and depth of these insights should be tailored to the audience's needs and objectives, whether it's for immediate maintenance actions or strategic planning. The purpose behind providing such insights is as crucial as the information itself, extending beyond mere fault detection to enhancing system reliability and performance.

For example, in condition monitoring, real-time vibration data may inform a technician of an immediate issue requiring urgent maintenance, while long-term trend analysis could help engineers optimise operational parameters. Similarly, advanced predictive analytics might offer strategic insights into life-cycle management, aiding in decisions on equipment replacement or overhaul. The importance of these explanations also lies in their potential to refine monitoring strategies themselves, through methods such as anomaly detection algorithms that learn to identify potential failures more accurately over time. The nature of explanations varies by the recipient's role; Technicians receive immediate, actionable insights for urgent maintenance; engineers get detailed analyses for optimising operations and making strategic decisions; and management is provided with strategic insights focused on business impacts and financial implications. Tailoring explanations to the audience ensures informed decision-making across all levels of the organisation.

2.7 Explainability for Non-expert Users

Most XAI tools that are in use produce explanations that are aimed at users with expertise in ML [97]. There is a real challenge in presenting explanations and communicating them to stakeholders in a human-understandable context due to the knowledge gap between ML experts and non-ML experts. Research indicates that experts in the application domain tend to trust ML models when they are provided with user-friendly explanations that will enable them to understand the rationale of ML models. There are distinct explanations needed for stakeholders in different application domains.

Users of ML and AI applications have been divided into ML/AI experts, Domain experts and newcomers to the domain. Different groups of people might need different kinds of explanations. To solve this challenge many researchers have proposed different techniques [98].

An explainability framework has been proposed by [99] in which three levels of explanations are developed depending on users' need: Low-level explanations in which pre-model XAI methods are used to provide an initial understanding of the data, high-level explanations in which low-level explanations are used along expert knowledge to generate explanations, and co-created explanations in which users interact with the model to provide information incorporated later to generate tailored explanations according to users need.

An online experiment has been made on 199 participants in which he evaluated users' understanding of different XAI interfaces, The results of their experiment indicate that interactive explanations improve user's understanding of explanations [100].

In [101] three distinct groups of individuals who desire explanations with varying degrees of familiarity with ML have been identified, and each group has its own reasons and requirements shown in Figure 2.1: Practitioners who have enough ML expertise to build and design ML and AI models; business stakeholders and regulators; and end users who are domain experts with no ML expertise.

2.8 Towards AI Trustworthiness through Enhanced Explainability

As AI technology grows, it becomes increasingly important to make AI understandable and to align it with ethical principles to ensure trustworthiness and responsibility. This involves making sure that the way AI makes decisions is clear and can be checked for accuracy.

The EU AI Act [2] sets forth seven essential requirements to ensure the trustworthiness of AI systems as shown in Figure 2.5, encompassing fundamental rights and ethical principles. These obligations span various dimensions, including human oversight, technical robustness, privacy, transparency, fairness, societal and environmental impacts, and accountability. Stakeholders throughout the AI lifecycle, such as developers, deployers, end-users, and society at large, are tasked with adhering to these requirements, underscoring the collaborative effort needed to promote ethical and responsible AI development and deployment.

It's also critical to ensure that AI systems are strong enough to withstand malicious attempts to mislead them while maintaining transparency. Moreover, the foundation of responsible AI use must be based on ethical considerations, stressing the importance of AI acting in ways that are fair, respectful of privacy, and accountable. Through focusing on these dimensions, the pathway towards AI that is not only technologically advanced but also trusted and ethically sound is envisioned.

Figure 2.5: Seven Essential Requirements to ensure the Trustworthiness of AI systems [2]

2.8.1 Validation of Explanations

Validating the quality and reliability of XAI explanations remains an active and essential area of research, particularly for fostering trust and promoting the adoption of AI models in industrial applications. Validation is critical in safety-critical and high-stakes industries, where stakeholder confidence in model behaviour is paramount. Establishing robust validation frameworks ensures that XAI explanations are accurate, reliable, and aligned with the contextual domain knowledge that underpins decision-making processes [33], [20].

Collaboration with Domain Experts

Involving domain experts in the validation process is pivotal for aligning XAI-generated explanations with real-world domain knowledge and expertise. These experts provide critical feedback that verifies the relevance and correctness of explanations against nuanced domain-specific considerations [32] [26]. Their insights not only enhance the reliability of the explanations but also help bridge the gap between technical outputs and practical applicability. This alignment is vital for building end-user trust, particularly when deploying ML models in decision-making contexts that require a deep understanding of domain intricacies [102].

Leveraging Domain Knowledge Repositories

In cases where immediate access to domain experts is not feasible, leveraging established repositories of domain knowledge becomes a practical alternative. Cross-referencing XAI-generated explanations with documented insights, such as industry standards, guidelines, or historical case data, allows for an independent validation mechanism [20] [51]. This approach ensures consistency with well-established knowledge bases and provides a fallback mechanism to evaluate the reliability of explanations. The dual strategy of involving domain experts and utilising domain repositories strengthens the robustness of the validation process, fostering trust in the resulting insights.

Controlled and Real-World Dataset Validation

Another effective approach for validating XAI tools involves testing their explanations on controlled public datasets. This method allows researchers to evaluate the accuracy, interpretability, and consistency of the explanations in a well-defined environment where the ground truth is known [30] [52]. Subsequently, these tools are applied to real-world or industrial datasets to assess their performance under practical conditions. Studies have demonstrated that the transition from controlled settings to industrial applications highlights the generalizability and scalability of XAI explanations, reinforcing their reliability in diverse use cases [17].

Alternative Validation Strategies

Beyond expert validation and dataset-based testing, emerging methods focus on quantifying explanation quality using objective metrics. For example, techniques such as fidelity, stability, and consistency metrics measure how closely explanations reflect the underlying model logic and how robust they are to perturbations in the data [14], [103]. Additionally, user studies and feedback loops involving end-users serve as qualitative validation methods to gauge the comprehensibility and usability of explanations in real-world contexts [20].

Metrics for Explanation Validation

Beyond qualitative strategies, explanation quality can be quantified using complementary metrics that address different dimensions of interpretability [20] [104]:

Fidelity (Faithfulness to the Model)

Fidelity refers to the extent to which an explanation accurately reflects the true behaviour of the underlying model. It captures whether the explanation is consistent with the model's decision-making process, rather than merely being plausible to a human observer [57].

Several approaches exist to measure fidelity. One common proxy is local surrogate agreement, where a simple interpretable model (e.g., a linear regression or rule set) is fitted locally around a prediction to approximate the behaviour of the black-box model, as demonstrated in LIME [105]. Another approach involves perturbation-based tests, which examine whether altering or removing features identified as important by the explanation produces meaningful changes in the model's output [106] [107]. Building on this idea, Yeh et al. [108] introduced the infidelity metric, which quantifies the discrepancy between the output changes predicted by an explanation and the actual changes observed under perturbations; in this case, lower infidelity indicates higher fidelity.

In industrial condition monitoring, fidelity is particularly important because it underpins trust in automated decision support. High-fidelity explanations allow engineers to audit model decisions with confidence, ensuring, for example, that anomaly detections are attributed to domain-relevant signals such as vibration signatures, load variations, or temperature shifts, rather than to irrelevant noise [109].

Robustness/Stability

Robustness, or stability, refers to the degree to which explanations remain consistent when inputs are subject to small, benign perturbations [110]. In other words, similar inputs should yield similar explanations; if minor variations in sensor readings cause large shifts in the explanation, the reliability of both the explanation and the model is called into question.

Several metrics have been proposed to quantify robustness. These include max-sensitivity, which measures the maximum change in explanation resulting from small input perturbations, and rank stability, often evaluated using Kendall's τ correlation, where higher scores indicate more stable and trustworthy explanations [110]. Authors in [108] further formalised robustness as a diagnostic property, explicitly capturing the sensitivity of explanations to perturbations.

In industrial applications, robustness is especially critical. Sensor noise, calibration differences, and environmental fluctuations are common sources of benign variation. Explanations that remain stable under such conditions provide engineers with confidence in the system's outputs, supporting both reliable diagnosis and effective model debugging [24].

Comprehensibility/Clarity

Comprehensibility, or clarity, concerns the ease with which explanations can be understood by their intended audience. An explanation may be technically accurate but will fail its purpose if it is overly complex or opaque to non-specialist stakeholders [111].

Clarity can be evaluated using structural proxies such as explanation length, number of features, or the depth of a decision tree, with simpler forms generally being more interpretable [112]. Human-grounded studies provide an alternative: users may be asked to restate explanations, apply them to new cases, or show improved task performance when supported by explanations [113]. High levels of clarity reduce cognitive load and allow users to act confidently, whereas low clarity often results in confusion and mistrust.

In industrial systems, clarity is particularly important because domain experts, such as maintenance engineers, may not have expertise in machine learning. Explanations must therefore bridge this gap by presenting reasoning in plain language, supplemented with visual aids or schematic highlights, rather than exposing raw model internals [33]. Clear explanations in this context build trust, facilitate accountability, and support effective decision-making.

2.8.2 Completeness/Faithfulness

Completeness and faithfulness together address whether an explanation provides both the full scope of relevant causes and a faithful reflection of the model's internal reasoning. Completeness requires that no significant contributing factor is omitted, while faithfulness ensures that no spurious or misleading elements are introduced [112].

Several methods have been proposed to evaluate these properties. Integrated Gradients, for example, satisfy the completeness axiom by ensuring that the sum of attributions equals the difference between the model's output and a baseline [114]. Other approaches include feature addition or coverage tests, which check whether the highlighted features collectively reconstruct the model's prediction. Further diagnostic techniques include deletion/insertion curves [115], ROAR retraining to assess feature necessity [116], and sanity checks that confirm explanations genuinely depend on model parameters rather than artefacts of the training process [117].

In industrial applications, completeness and faithfulness are essential for safety and accountability. Engineers must be confident that explanations cover all critical factors influencing a prediction, while regulators and auditors require assurance that no misleading or extraneous elements are introduced. Together, these properties ensure explanations are both exhaustive and honest, enabling robust human oversight of AI-driven decisions [118].

In Summary, validating XAI explanations requires a multifaceted strategy that draws on complementary perspectives. Metrics such as fidelity, robustness, relevance, clarity, and completeness/faithfulness each capture different dimensions of explanation quality. Together with approaches involving expert collaboration, domain knowledge integration, controlled experimentation, and real-world testing, these metrics strengthen the credibility, transparency, and trustworthiness of XAI methods. Such a holistic validation process is essential for the reliable adoption of explainable AI in industrial and other high-stakes settings [112] [119].

2.9 Adversarial Attacks on Explainability

The increasing reliance on black-box ML models in critical decision-making processes has underscored the necessity for accurate and trustworthy explanations of their behaviour. However, recent apprehensions have emerged, casting doubt on the fidelity of such explanations, particularly regarding their ability to reveal the inherent biases embedded within these models. This discrepancy between expectation and reality presents a significant risk: the potential for explanations to misguide human users, leading them to place undue trust in these black-box models. Such misplaced trust is not merely a theoretical concern but a pressing issue that calls for a deeper understanding of the nature and implications of adversarial attacks on explanations. As this complex landscape is navigated, the integrity and transparency of ML models hang in the balance, urging a concerted effort to scrutinise and fortify the explanations relied upon to interpret the decisions of these powerful yet opaque systems [120]. Adversarial attacks on explanations can be done by altering:

- The ML model itself.
- The data.

This paper [121] highlights a critical vulnerability in model explanation methods, demonstrating how explanations can be manipulated through changes in the input data. It reveals that subtle input modifications, which do not significantly affect the model's output, can nonetheless alter explanation results, raising concerns about their trustworthiness and interpretability. The authors provide a theoretical foundation for this phenomenon, rooted in neural network geometry, and propose strategies to enhance the robustness of explanations against such manipulations.

Conversely, [122] introduces a defense strategy to counter adversarial attacks that exploit data manipulations. The paper proposes an approach that aggregates multiple explanation methods, thereby enhancing the neural network's resilience to such attacks. Remarkably, this robustness persists even when adversaries have full knowledge of the model's parameters and the explanation methods employed.

In a related vein, [123] examines how explanations can be manipulated by altering the model itself. The study investigates adversarial model manipulation, where interpretation methods are intentionally influenced without affecting the model's accuracy. By integrating interpretation results into the fine-tuning process, the authors demonstrate that widely-used interpreters, such as Layer-wise Relevance Propagation (LRP) [106], Gradient-weighted Class Activation Mapping (GRAD-CAM) [124], and Simple Gradient-Based Method (SimpleGrad) [125], can be deceived. Two types of deception—passive and active—are presented, and their efficacy is validated across different models and interpretation methods. The research underscores the importance of evaluating interpretation methods for their stability under adversarial model manipulation to ensure their reliability and robustness.

Similarly, [126] introduces a technique for generating misleading explanations, exposing the limitations of post hoc explanation methods that rely on input perturbations, such as LIME [57] and SHAP [56]. The authors propose a scaffolding technique that can effectively mask classifier biases, allowing adversaries to generate seemingly benign explanations of their choice.

Further, [127] provides an example of how minor perturbations to a model can distort explanations. This deliberate concealment of biases in closed-box models represents a significant vulnerability, among others, in explainable AI systems.

Finally, [96] identifies a fundamental issue with saliency-based explanation methods, noting that they often fail to maintain input invariance. The study shows that even minor input modifications can lead to substantial changes in the output and the perceived relevance of individual input features, thereby questioning the reliability of such explanation methods.

2.10 Practical Applications of Explainability

Explainability techniques have been broadly applied across various industries, significantly enhancing the transparency and trustworthiness of AI systems. These methods have transformed decision-making in high-risk domains such as healthcare, finance, and condition monitoring by providing clear insights into complex AI-driven predictions and recommendations [128] [129] [130]. In these critical sectors, explainability ensures that stakeholders can trust and understand AI outcomes, which is essential for regulatory compliance and ethical considerations. Below are some examples categorised by domain risk:

2.10.1 High-Risk Domains: Healthcare, Finance and Condition Monitoring

Healthcare

In healthcare, explainability is a critical requirement for AI-driven systems, given the high stakes associated with patient outcomes and clinical decision-making. A comprehensive assessment by [131] underscores the multi-dimensional implications of explainability in medical AI, examining the topic from technological, legal, medical, and patient-centred perspectives. The study highlights that although AI systems can outperform human experts in specific analytical tasks, their inherent opacity raises significant ethical, legal, and societal concerns. The authors analyse AI-based Clinical Decision Support Systems (CDSS), emphasising that explainability is essential for fostering trust among healthcare professionals and patients. From a technological standpoint, achieving explainability must be carefully balanced with development objectives, while the legal perspective underscores the importance of transparency in areas such as informed consent, certification, and liability. Both the medical and patient perspectives stress the need for clear and interpretable outputs to ensure that AI augments, rather than replaces, human judgment [131].

In healthcare, XAI addresses critical demands for transparency, fairness, and interpretability in AI-driven decision-making systems. The study reviews various XAI methods, including LRP [106], Local Interpretable Model-Agnostic Explanations (LIME) [57], SHapley Additive exPlanations (SHAP) [56], and GRAD-CAM [124], demonstrating their applications in medical diagnostics and treatment planning. The usability and reliability of these techniques are illustrated through case studies, such as XGBoost for tumour analysis and Training Calibration-based Explainers (TraCE) for medical imaging, showcasing their potential to render complex models more accessible to healthcare professionals. Overall, the study highlights XAI's transformative potential in healthcare by supporting clinicians with interpretable insights, ultimately enhancing patient care and decision-making [132].

However, Authors in [133] argue that current explainability methods may be insufficient for patient-level decision support. The authors advocate for rigorous internal and external validation of AI models as a more effective approach to ensuring reliability and trustworthiness in clinical settings. They caution against making explainability a strict prerequisite for the deployment of clinical AI systems.

Finance

In highly regulated fields like Finance, the adoption of AI requires stringent transparency to ensure decision traceability for third parties. This is where XAI becomes crucial. This paper [134] presents a systematic literature review of XAI in Finance, analysing 2,022 articles from leading journals in Finance, Information Systems, and Computer Science. From these, 60 relevant studies were identified and categorised based on the XAI methods used and their intended objectives. The review highlights that areas such as risk management, portfolio optimisation, and stock market applications are well-covered, while anti-money laundering remains underexplored. The findings also reveal a growing preference for post-hoc explainability over inherently transparent models in recent research. In this study [135], the Authors describe a bank's use of an AI-powered solution to assess loan applications, combining historical data, customer attributes (e.g., income, credit score, debt-to-income ratio), and business rules to improve decision-making. While AI enhances predictive accuracy, it introduces complexity that necessitates explainability. Loan officers require tools like SHAP [56] and LIME

[57] to validate decisions by understanding factors such as credit score or employment history that influenced approvals or denials. Meanwhile, loan applicants benefit from counterfactual explanations [25], which suggest actionable changes, such as increasing income or reducing debt, to achieve approval. By providing tailored explanations to stakeholders, explainable AI ensures transparency, accountability, and alignment with ethical and regulatory standards in AI-driven lending.

Authors in [136] emphasise that XAI is critical in finance, with different stakeholders requiring tailored insights. Developers and regulators focus on individual predictions for queries or complaints, while model checkers and management seek a broader understanding of governance. The World Economic Forum and De Nederlandsche Bank (DNB) stress the ethical need for transparency to maintain trust and prevent discrimination, particularly in a sector reliant on responsible data use. Explainability supports trust, accountability, and compliance in high-stakes applications like capital requirement calculations and stress testing.

Financial services rely on XAI for risk assessment, fraud detection, and compliance. PwC's collaboration with a major auto insurer serves as a prime example, where XAI was integrated into an ML model for claims estimation. PwC emphasised the model's interpretability, crucial for adoption by claims adjusters and stakeholders, who required insight into AI-driven estimates to make informed decisions. The system used XAI techniques like feature importance ranking and rule-based explanations, ensuring that all stakeholders could understand the model's rationale, which is vital in regulated, risk-sensitive financial environments [137].

Future developments in XAI in the finance sector, as identified in [138] include focusing on enhancing interpretability while maintaining model complexity, integrating XAI with deep learning, and improving methods like SHAP [56] and LIME [57]. Standardisation of tools, evaluation metrics, and regulatory alignment is critical for consistent application. Ethical considerations, including addressing bias, ensuring privacy, and adhering to regulations like GDPR, will be pivotal in fostering transparency and fairness in financial decision-making.

Condition Monitoring

ML has demonstrated significant effectiveness in condition monitoring, but the lack of transparency in model decision-making remains a persistent challenge.

In [139], the authors explore the use of SHAP [56] to enhance interpretability in fault detection for rotating machinery. By applying SHAP to Support Vector Machine (SVM) and k-Nearest Neighbor (kNN) models, the study identifies key features, such as the skewness and shape factor of vibration signals, as critical to diagnostic outcomes. The approach achieved high diagnostic accuracy, illustrating SHAP's potential to improve understanding and trust in ML-driven condition monitoring systems.

Similarly, [140] presents an XAI-based predictive maintenance strategy utilising LIME [57]. The study provides interpretable insights into ML model predictions, emphasising proactive maintenance strategies and enhanced spare parts management. By explaining key predictive factors, the approach addresses challenges in interpreting binary model outcomes, supporting more effective maintenance decisions and optimising resource allocation.

A study by [141] introduces a framework for monitoring hydraulic systems using multi-sensor data and Deep Neural Networks Deep Neural Networks (DNN). This framework predicts the operating states of system components while tackling the challenges of applying DNNs in complex, real-world applications. To improve transparency, the authors employed DeepSHAP to elucidate the contributions of individual sensors in the decision-making process. The results demonstrate the model's robustness and reliability, with DeepSHAP fostering trust and enabling broader adoption of AI in industrial settings.

For wind turbine performance evaluation, [142] leverages XAI to move beyond traditional univariate models. This study incorporates multiple input variables to better account for factors influencing power output. A sequential feature selection process minimises model error, while Shapley coefficients quantify each variable's contribution. Using real-world data, the study identifies novel variables related to rotor and blade pitch control, offering new perspectives for anomaly detection and condition monitoring in wind turbines.

In the maritime sector, [143] applies the isolation forest algorithm for anomaly detection in marine engine monitoring. By integrating SHAP, the study enhances transparency, effectively illustrating the factors driving anomaly detection results.

Further, [144] demonstrates the use of SHAP to improve the explainability of data-driven prognostic models, providing clearer insights into the decision-making processes underlying these models.

In the production domain, [145] investigates the integration of gradient-boosting machine learning models with XAI techniques. The authors utilise visualisation tools to showcase how these approaches can generate valuable insights into model predictions.

Finally, [146] proposes an explainable predictive maintenance framework for diagnosing multiple faults in rotating machinery. Their approach combines multi-sensor data fusion with interpretability techniques to provide actionable insights into the maintenance process.

2.10.2 Moderate Risk to low Risk Domains

LinkedIn has effectively leveraged XAI across various business sectors since 2018, including recruitment, sales, and machine learning engineering. In 2022, LinkedIn underscored the importance of XAI in implementing a ranking and recommendation ML system for its sales team. This system prioritised customer engagement by predicting churn and upselling potential. Despite its success, the AI team recognised that the sales teams required interpretable explanations for the predictions to foster trust and encourage adoption of the system [147].

In the retail sector, Authors in [148] used XAI to support needs-based customer segmentation for new product development. By applying SHAP to online product reviews, they derived feature-level importance values that revealed the non-linear relationships between product attributes and overall satisfaction. This approach enabled the identification of customer segments with unmet needs and provided actionable insights for designing new product concepts, outperforming traditional sentiment-based clustering methods.

In the agricultural sector, [149] explored the role of XAI in Leaf Disease Classification using deep learning models. The study applied Visual Geometry Group (VGG), GoogLeNet, and ResNet to the Five Leaves dataset. The ResNet model, augmented with an attention mechanism, achieved remarkable accuracy rates of 99.11%, 99.4%, and 99.89% across three experiments, demonstrating the potential of explainability to enhance performance and usability in agricultural applications.

In the domain of online advertising, [150] proposed a novel method utilising the decision paths of a tree-based ensemble classifier to offer actionable recommendations. This approach transforms true negative instances (low-quality ads) into positively predicted instances (high-quality ad suggestions) by adjusting feature values within the ad quality feature space. This technique enhances the relevance and performance of online ads while maintaining transparency in the process.

These examples highlight the practical applications of XAI across diverse industries. By integrating explainability into AI systems, organisations can improve understanding, trust, and management of ML-driven solutions. This emphasis on transparency and accountability ensures that AI systems align with user expectations and ethical standards across sectors, from sales and insurance to agriculture and advertising.

2.11 Conclusion

This chapter has provided a comprehensive examination of explainability in machine learning, positioning it as both a technical necessity and a societal imperative. It began by clarifying the conceptual foundations of explainability and interpretability, highlighting their varying definitions and stakeholder-dependent interpretations. The historical evolution of machine learning models demonstrated how increasing model complexity has amplified the need for structured explanation mechanisms, particularly in high-stakes domains.

The chapter further explored the regulatory landscape across the UK, EU, US, and international organisations, illustrating how explainability has become embedded within emerging AI governance frameworks. These regulatory developments reinforce that transparency, accountability, and human oversight are no longer optional attributes but formal requirements in many applications.

In addition, the discussion of validation strategies, explanation quality metrics (fidelity, robustness, clarity, completeness), trade-offs, and adversarial vulnerabilities emphasised that explainability is not a trivial add-on but a multi-dimensional design challenge. Effective explainability requires careful alignment between technical methods, stakeholder needs, domain knowledge, and regulatory constraints.

By synthesising theoretical foundations, regulatory considerations, validation methodologies, and practical applications across domains, this chapter establishes the conceptual and analytical groundwork for the thesis. It frames explainability as a structured, context-dependent process rather than a single technique, thereby motivating the development of a systematic taxonomy and domain-adapted framework presented in the subsequent chapters.

Chapter 3

Explainability Taxonomies

In this chapter, explainability methods are organised according to several complementary dimensions, forming a multi-dimensional taxonomy rather than a rigid or sequential structure. The framework shown in Figure 3.1 builds on widely recognised classification criteria from the XAI literature. These include: (i) the type of input data (e.g., structured or unstructured), (ii) the explanation scope (local, global, or cohort-level), (iii) the timing of explanation (pre-model, ante-hoc, or post-hoc), (iv) the explanation technique (e.g., example-based, rule-based, counterfactual, gradient-based, or feature-relevance), and (v) model specificity (model-agnostic or model-specific).

This taxonomy is inspired by and extends previous classification efforts in the literature [151, 32, 152]. While the figure presents these categories as separate branches for clarity, they are not mutually exclusive. In reality, many XAI methods span multiple categories. For instance, SHAP is a post-hoc, model-agnostic, feature-relevance method that works with both structured and image data, and provides both local and global explanations. The cross-cutting relationships between these dimensions are captured in Table 3.1 and the following tables, which link existing XAI tools to their corresponding categories. The goal of this taxonomy is not to impose a strict hierarchy, but to provide a structured way to navigate and understand the wide variety of XAI approaches.

3.1 Comparison of Existing XAI Taxonomies

Numerous taxonomic frameworks have been proposed in recent years to organise the expanding landscape of XAI. These general-purpose taxonomies aim to provide structured perspectives on the field by classifying methods across key dimensions such as the timing of explanation, model dependency, and the scope of interpretation. Despite differences in terminology and emphasis, many taxonomies converge on a set of foundational dimensions: when the explanation is produced (pre-model, ante-hoc, or post-hoc), how it applies in terms of scope (local, global, or group-level), and whether it is model-specific or model-agnostic [151].

One of the earliest and most cited taxonomies classifies XAI techniques according to three fundamental criteria: the complexity of interpretability, the scope of explanation, and the method's dependency on the underlying machine learning model [151]. This framework introduced the distinction between intrinsic and post-hoc explanations, local and global applicability, and specific versus agnostic dependencies. These dimensions continue to form the backbone of many subsequent taxonomies.

A complementary framework is presented in a comprehensive survey that classifies explanation methods based on the target of explanation—either the model or a specific outcome—and the mechanism by which the explanation is delivered [32]. This taxonomy distinguishes between outcome-based methods such as feature attribution, rule-based reasoning, counterfactuals, and visual saliency techniques, and model-level approaches including surrogate modelling and global feature importance. The inclusion of explanation formats and decision targets provides an additional level of interpretive granularity.

Another widely adopted taxonomy introduces a dual perspective that separates transparent, interpretable models from post-hoc explanatory tools [20]. Within this structure, intrinsically interpretable models include decision trees, linear models, and rule-based learners, while post-hoc methods are categorised by their underlying technique—feature relevance, visual explanation, textual generation, example-based reasoning, and architecture modification. This taxonomy is notable for incorporating explanation modality and for providing a specific subclassification tailored to deep learning systems.

A further perspective shifts the focus from technical characteristics to user-oriented and application-specific considerations [135]. This approach classifies methods based on their position within the machine learning lifecycle and their intended audience—be it developers, domain experts, regulators, or end-users. By mapping explanation techniques to stakeholder requirements, this taxonomy highlights the importance of context and use-case relevance in selecting appropriate interpretability tools.

The taxonomy proposed in [153] organises explainability approaches along multiple dimensions, beginning with the distinction between transparent models such as linear regression, decision trees and rule-based learners that are inherently interpretable, and opaque models such as random forests, SVMs and neural networks that require post-hoc interpretation. Post-hoc methods are further classified as model agnostic or model specific, and grouped into categories such as explanation by simplification, feature relevance explanation, local explanations and visual explanations. These categories are underpinned by core explainability principles, including rule-based learning, decision tree or prototype extraction, influence functions, sensitivity analysis, game theory-based attribution, interaction analysis, counterfactual reasoning and distillation, which provide the conceptual basis for various techniques. The taxonomy also links these principles to representative tools, such as SHAP [56] for game theory-inspired feature relevance, LIME [57] for local linear approximation, Anchors [154] for rule-based local explanations, Individual Conditional Expectation (ICE) [155] and Partial Dependence (PD) [156] for visual feature analysis, and Interpretable Trees (InTrees) [157] for rule extraction from ensembles, offering a comprehensive mapping from theoretical foundations to practical implementations.

The taxonomy proposed in [22] offers a broad conceptual framework for organising explainability methods based on model transparency and the nature of the explanation. It distinguishes between transparent models, which are inherently interpretable, and opaque models that require post-hoc interpretability. It further classifies XAI approaches as model-agnostic or model-specific, depending on whether the explanation relies on internal model details. Several explanation strategies are included, such as simplification through surrogate models, feature relevance estimation, visualisation techniques, and local approximations around individual predictions. This taxonomy emphasises the diversity of interpretability methods and provides a high-level structure to assess their applicability, generality, and interpretive mechanism.

Authors in [158] categorise existing taxonomic efforts in explainable AI into four methodological approaches, each based on a distinct organising principle. The functioning-based approach groups methods by how they operate—such as perturbation, structural analysis, or architectural modification—focusing on the mechanisms behind the explanation. The result-based approach classifies methods according to the type of output they produce, including feature importance scores, surrogate models, or representative examples, making it suitable for applications with specific explanatory needs. The conceptual approach constructs taxonomies using core dimensions such as explanation timing (ante-hoc vs. post-hoc), model applicability (agnostic vs. specific), and scope (local vs. global), and may include further dimensions like output format or user context. Finally, the mixed approach combines elements from the other three, typically beginning with conceptual dimensions and refining them using functional and result-oriented criteria, offering a flexible and comprehensive framework, particularly helpful for newcomers to the field.

3.2 Proposed Taxonomy

The taxonomy proposed in this thesis organises explainability methods along five complementary dimensions shown in Figure 3.1, offering a structured yet flexible framework to capture the diversity of XAI approaches. The first dimension, Input Data Type, distinguishes between structured data (e.g., tabular datasets) and unstructured data (e.g., images, text, and video), recognising that the nature of the input directly affects the choice of explanation technique. The second dimension, Type of Explanations, classifies methods according to the granularity of insight they provide, ranging from local (individual-level explanations) to global (overall model behaviour), to cohort-level (subgroup-focused explanations relevant for fairness or group analysis). The third dimension, Timing of Application, considers when explainability is introduced within the machine learning workflow, encompassing pre-model tools (e.g., data-centric interpretability strategies), ante-hoc methods (inherently interpretable models), and post-hoc techniques (methods applied after model training). The fourth dimension, Explanation Technique, captures the underlying mechanism used to generate explanations, including example-based, counterfactual, rule-based, gradient-based, and feature relevance methods. Lastly, the Model Specificity dimension differentiates between model-specific methods, which require access to internal model structures, and model-agnostic methods, which treat the model as a black box. Collectively, these five dimensions form a comprehensive and multidimensional taxonomy that accommodates the hybrid nature of modern XAI techniques while supporting both methodological development and practical evaluation. This taxonomy builds on and extends earlier classification work in the literature, presenting categories as separate branches for clarity while acknowledging that many XAI methods span multiple categories. The relationships across these dimensions are outlined in accompanying tables, which map existing XAI tools to their respective categories. The aim is not to enforce a strict hierarchy, but to offer a structured framework for navigating and understanding the diversity of XAI approaches.

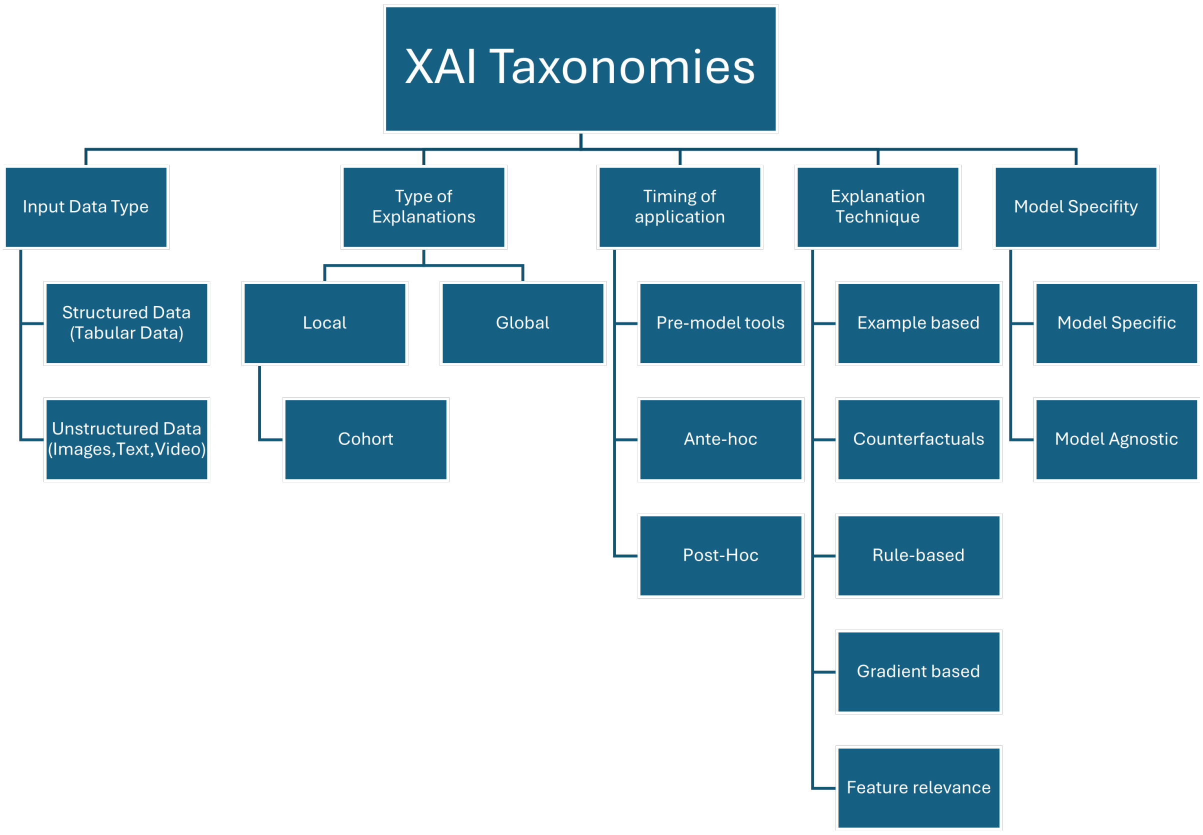


Figure 3.1: Explainability taxonomies illustrating the main dimensions. These include input data type, type of explanations (local, global, cohort), timing of application (pre-model, ante-hoc, post-hoc), explanation techniques (e.g., example-based, counterfactuals, rule-based, gradient-based, feature relevance), and model specificity (model-specific vs. model-agnostic)

3.2.1 Input Data Types

Explainability methods must be adapted to the structure and modality of input data, as varying data types demand tailored interpretability approaches. Broadly, input data can be categorised into structured and unstructured forms. Structured data—such as tabular representations—follows defined schemas and is prevalent in domains like healthcare, finance, and industrial monitoring. In contrast, unstructured data includes formats such as images, text, and video, which lack formal organisation and typically require more advanced preprocessing and interpretive techniques.

This subsection provides a systematic overview of explainability approaches categorised by input data type—specifically tabular, image, text, and video—within the context of the structured/unstructured distinction. For each modality, techniques are evaluated based on their underlying mechanisms, interpretability scope (e.g., local, global), and model specificity. This taxonomy supports the selection of context-appropriate explanation tools, ensuring interpretability that is both meaningful to stakeholders and applicable across varied machine learning applications.

Tabular Data

Tabular data, which is prevalent in industries like finance, healthcare, and logistics, often consists of structured data with multiple features influencing the model's predictions. Due to its structured nature, a range of explainability methods has been developed to cater to both global and local interpretability needs [25].

These methods can be broadly classified into feature importance, rule-based, prototype-based, and counterfactual explanation techniques [25]. Feature importance methods such as SHAP [56] and LIME [57] offer insights into how individual features contribute to model predictions. SHAP, in particular, provides a unified measure of feature importance applicable across various model types. Similarly, Descriptive mAchine Learning EXplanations (DALEX) [159] stands out as a versatile framework that supports a range of models, providing both global and local explanations. Its focus on creating unified visualisations enhances interpretability, making it especially useful for model debugging and exploratory data analysis. Neural Additive Models (NAMs) [160], another feature importance method, extend Generalised Additive Models (GAMs) by integrating neural networks, allowing them to provide accurate and interpretable additive feature effects for structured data.

Rule-based methods, including ANCHOR [154] and Rule-based Model Fitting [161], generate interpretable rules that help users understand model decisions by focusing on the conditions under which predictions are made. These methods can provide both local and global insights, depending on the context. Logical Rule Induction (LRI) [162] also falls under this category, offering ante-hoc explanations by using inductive logic to derive human-readable rules specific to the underlying data. Additionally, Scalable Probabilistic Rule Learning (SCALABLE-PRL) [163] uses probabilistic frameworks to efficiently create interpretable rules, addressing scalability challenges for larger datasets and providing global and local interpretability.

Prototype-based approaches like GLObal to loCAL eXplainer (GLocalX) [164] and Prototypical Explanations via Faster Search (PROTODASH) [165] utilise representative examples from the data to explain model behaviour, helping users relate model predictions to real-world instances. Maximum Mean Discrepancy Criticism (MMD-CRITIC) [166], another prototype-based method, employs the Maximum Mean Discrepancy (MMD) measure to select representative prototypes, enabling users to gain a global understanding of the model through clear, real-world instances. This approach is particularly suited for ante-hoc explanations, providing insights into the overall behaviour of the model.

Counterfactual explanation methods such as Contrastive Explanation Method (CEM) [167], Diverse Counterfactual Explanations (DiCE) [168], and Feasible and Actionable Counterfactual Explanations (FACE) [169] provide hypothetical scenarios that illustrate how slight changes in input features could alter the model's prediction, offering actionable insights for decision-making. Counterfactual Explanations Framework (CFX) [170] takes this further by optimising the feasibility and proximity of counterfactuals to real-world conditions, ensuring that the generated explanations are both practical and meaningful for tabular data.

The combination of these methods ensures comprehensive interpretability for models handling tabular data, addressing both the overall behaviour of the model and individual predictions. The following Table 3.1 presents an overview of key explainability methods for tabular data, categorised by the type of method, applicability to data, the stage at which explainability is provided (ante-hoc or post-hoc), the scope of the explanation (global or local), and their specificity. This taxonomy helps in selecting the appropriate method based on the use case and interpretability requirements.

Table 3.1: Overview of Explainability Methods for Tabular Data

Name	Type	Data Type	Scope	Specificity
SHAP [56]	Feature Importance	Any	Global/Local	Agnostic
LIME [57]	Feature Importance	Any	Local	Agnostic
LRP [106]	Feature Importance	Any	Local	Agnostic
DALEX [159]	Feature Importance	Any	Local/Global	Agnostic
Neural Additive Models (NAM) [160]	Feature Importance	Tabular	Local	Specific
Anchors (ANCHOR) [154]	Rule-Based	Tabular	Local/Global	Agnostic
LRI [162]	Rule-Based	Tabular	Local	Specific
Rule-based Model Fitting (RULEFIT) [161]	Rule-Based	Tabular	Global/Local	Specific
SCALABLE-PRL [163]	Rule-Based	Tabular	Global/Local	Specific
GLocalX [164]	Prototype-Based	Tabular	Local/Global	Agnostic
MMD-CRITIC [166]	Prototype-Based	Tabular	Global	Specific
PROTODASH [165]	Prototype-Based	Tabular	Local	Specific
CEM [167]	Example-based	Any	Local	Specific
DiCE [168]	Counterfactual	Any	Local	Specific
FACE [169]	Counterfactual	Any	Local	Specific
CFX [170]	Counterfactual	Tabular	Local	Specific

Images Data

In the context of computer vision tasks, working with image data often requires methods that can offer interpretable insights into model behaviour, particularly given the complexity and high dimensionality of visual inputs. Unlike other machine learning tasks, the significance of computer vision lies in how pixel-level features combine to form meaningful patterns like edges and textures. This necessitates specialised methods tailored for image data [25].

Key explainability techniques for image models, particularly those using CNN [171], include gradient-based backpropagation methods, perturbation-based approaches, and methods leveraging the model's internal states. These methods aim to visualise and interpret how models make decisions by highlighting influential regions in input images [25].

Gradient-based methods such as Integrated Gradients [172], Smooth Gradient [173], and Saliency Maps [125] provide insights into which pixel-level features are most influential for a given prediction. Integrated Gradients attribute a model's output to its input features by integrating gradients along a path from a baseline to the input, ensuring a principled explanation of feature importance. SmoothGrad improves upon saliency methods by reducing noise in explanations through gradient averaging, creating visually smooth and more interpretable results. Saliency Maps highlight the most important pixels by analysing the gradient of the output with respect to the input image, providing a basic yet effective visualisation.

Region-based backpropagation methods, like eXplainable Region-based Activation Integration (XRAI) [174], build upon gradient-based techniques by segmenting images into meaningful regions. This enables explanations to focus on interpretable areas rather than individual pixels, offering a higher-level understanding of model decisions.

Methods that leverage the internal state of models, such as GRAD-CAM [124], Guided Backpropagation [175], and Guided GRAD-CAM [124], are specifically tailored for CNNs. GRAD-CAM uses class-specific gradients to generate visual heatmaps, showing the regions most influential for a prediction. Guided Backpropagation modifies standard backpropagation to visualise positive gradients that significantly impact predictions. Guided GRAD-CAM combines the strengths of GRAD-CAM and Guided Backpropagation, offering more refined and interpretable visualisations.

Perturbation-based methods like LIME [57] and SHAP [56] are model-agnostic and provide explanations by analysing how changes in input data impact predictions. Although these methods are applicable to any data type, they are particularly useful for explaining black-box models in image tasks by identifying the key regions or pixels responsible for predictions.

Additional methods like Deep Learning Important FeaTures (DeepLIFT) [176] and SmoothGrad [173] enhance the interpretability of gradient-based approaches. DeepLIFT calculates contributions of input features by comparing activations to a reference, providing explanations that are both efficient and intuitive. SmoothGrad, as mentioned earlier, refines explanations by averaging noisy gradients, making the results more stable and interpretable.

Specificity is another crucial aspect of explainability methods. While tools like GRAD-CAM and Guided Backpropagation are specifically designed for CNNs, methods such as Integrated Gradients and SHAP are broadly applicable across various model types handling image data. This diversity of methods allows for a wide range of interpretability, catering to both specific and general-purpose needs.

The following Table 3.2 provides an overview of these explainability methods tailored for image data. It categorises the methods based on their underlying mechanism, the nature of the data they handle, their scope (local or global), and their specificity. This taxonomy serves as a concise reference for selecting the appropriate explainability techniques based on the nature of the task and the desired level of insight into the model's decision-making process.

Table 3.2: Overview of Explainability Methods for Image Data

Name	**Type**	**Data Type**	**Scope**	**Specificity**
Integrated Gradients [172]	Backpropagation	Images	Local	Specific
XRAI [174]	Region-Based Backpropagation	Images	Local	Specific to Images
GRAD-CAM [124]	Internal State (Class Activation)	Images	Local	Specific to CNNs
LIME [57]	Perturbation-Based	Any	Local	Agnostic
Guided Backpropagation [175]	Modified Backpropagation	Images	Local	Specific to CNNs
Guided GRAD-CAM [124]	Combined Method	Images	Local	Specific to CNNs
Saliency Maps [125]	Gradient-Based	Images	Local	Specific
SHAP [56]	Perturbation-based	Any	Global/ Local	Agnostic
DeepLIFT [176]	Backpropagation	Images	Local	Agnostic
SmoothGrad [173]	Gradient-Based	Images	Local	Specific

Text Data

Text data represents a primary form of unstructured data and differs fundamentally from structured data (e.g., numerical or tabular data) in both format and complexity. Unlike structured data, which is organised into fixed schemas, text data is inherently sequential, variable in length, and rich in semantics. It may appear in diverse formats—including plain text, documents, code, transcribed speech, and web content—and often requires extensive preprocessing (e.g., tokenisation, stop-word removal, and lemmatisation) before it can be meaningfully analysed by machine learning models [177]. Text can be encoded at various linguistic levels, such as characters, subwords, words, or sentences, and is typically transformed into numerical representations (e.g., word embeddings or token IDs) to enable computational processing. These characteristics make text data particularly challenging and demand specialised techniques for both modelling and explainability [178].

In this context, language models play a central role in modern deep learning applications, particularly in Natural Language Processing (NLP), which has advanced rapidly in recent years. NLP focuses on understanding human language and is essential in applications like machine translation, sentiment analysis, text summarisation, and question-answering. Models for these tasks rely on text data to capture language nuances, employing architectures such as Long Short-Term Memory (LSTM), attention mechanisms, and transformers. The transformer architecture, introduced by [45], revolutionised NLP by efficiently handling sequential data through the attention mechanism. The Transformer revolutionised NLP by using self-attention to process entire sequences simultaneously, improving efficiency and parallelisation compared to Recurrent Neural Network (RNN) and LSTMs. Its encoder-decoder structure generates contextual representations in the encoder and produces outputs in the decoder by attending to both encoder outputs and prior tokens. The self-attention mechanism calculates the relevance of words in a sequence through "queries", "keys", and "values", with multi-head attention enabling parallel computations to capture diverse relationships.

Following the transformer, Bidirectional Encoder Representations from Transformers (BERT) was introduced by [46] and quickly surpassed existing benchmarks in Natural Language Understanding (NLU) tasks, such as sentiment classification, textual entailment, and text similarity [46]. BERT employs a bidirectional approach, enabling it to understand the context of words by considering both left and right contexts simultaneously. Its pre-training consists of two key tasks: predicting masked words within a sentence and identifying whether one sentence logically follows another. Once pre-trained, BERT can be fine-tuned for various NLU applications, including sentiment analysis, text classification, and question answering, setting new benchmarks for performance.

Subsequent models, including Generative Pre-trained Transformer 3 (GPT3) by [58], and Text-To-Text Transfer Transformer (T5) by [179], further pushed the boundaries of NLP, showcasing capabilities in various complex tasks [58] [179]. GPT3 [58] demonstrated unprecedented capabilities in NLP. GPT-3 builds on the autoregressive nature of its predecessors (GPT and GPT-2) and significantly scales the model's size, featuring 175 billion parameters. This vast increase in model capacity enabled GPT-3 to handle a wide range of complex NLP tasks, including text generation, summarisation, translation, and even creative writing, without requiring task-specific fine-tuning. Instead, GPT-3 relies on prompt engineering to perform these tasks in a zero-shot, one-shot, or

few-shot learning paradigm, adapting to new tasks with minimal examples.
More recently, Authors in [180] introduced Pathways Language Model (PaLM), a 540-billion-parameter transformer model capable of mathematical reasoning, code generation, and even explaining jokes [180]. One of PaLM's key strengths lies in its integration of the Pathways system, a highly efficient infrastructure that enables the model to scale seamlessly across multiple Tensor Processing Units (TPU) v4 Pods. TPU Pods are essentially clusters of specialised hardware, TPU, designed specifically to handle the immense computational demands of training massive AI models. By connecting hundreds or even thousands of these TPUs, TPU Pods work together to distribute the workload, making the training process faster and more efficient.
Building on the introduction to the foundational concepts of language models, the following paragraphs outlines several XAI tools that have been applied to text data.
Integrated Gradients [172] is a gradient-based XAI method that attributes model predictions to input features by integrating gradients along a path from a baseline (e.g., zero embedding) to the actual input. This approach effectively highlights the contribution of individual tokens in text-based models. Building on this, Gradient x Input [176] simplifies attribution by multiplying the gradient of the model's output with respect to the input by the input itself, offering a quick way to identify influential words or tokens. Extending these ideas further, Layer Integrated Gradients [172] not only attributes relevance to input tokens but also to intermediate representations within the model, providing deeper insights into how information is processed through different layers. Additionally, tools like Language Interpretability Tool (LIT) complement these methods by offering interactive visualisations of model behaviour, enabling users to explore token importance, attention mechanisms, and layer-wise outputs in an intuitive manner [181]. These methods, among others, have been adapted to address the unique challenges of NLP, such as tokenisation and the handling of textual data.
The following Table 3.3 provides an overview of key explainability methods for text data, highlighting their type, data applicability, stage of explainability, scope, and specificity. This summary serves as a guide for selecting appropriate methods based on the specific requirements of text-based models, ranging from perturbation-based

techniques like LIME to gradient-based methods such as Gradient x Input [176] and Layer Integrated Gradients [172]. Additionally, tools like the LIT [181] offer interactive visualisations, enabling deeper insights into model behaviour.

Table 3.3: Overview of Explainability Methods for Text Data

Name	Type	Data Type	Scope	Specificity
SHAP [56]	Perturbation-based	Any	Global/Local	Agnostic
LIME [57]	Perturbation-based	Any	Local	Agnostic
Gradient x Input [176]	Gradient-Based	Text	Local	Specific to Differentiable Models
Grad L2-Norm [172]	Gradient-Based	Text	Local	Specific to Differentiable Models
Layer Integrated Gradients (LIG) [172]	Gradient-Based	Text	Local	Specific
Layer-Wise Relevance Propagation (LRP) [106]	Backpropagation	Text	Local	Agnostic
Attention visualisation [45]	Attention Mechanism	Text	Local	Specific to Transformer Models
Language Interpretability Tool (LIT) [181]	Interactive Tool	Text	Local/Global	Agnostic

Video Data

The rapid advancements in deep learning have expanded its applications from static image recognition to more complex spatio-temporal tasks, including video recognition. This progression has led to the development of models capable of analysing dynamic scenes, facilitating tasks such as human activity recognition and anomaly detection. The inherent complexity of video data, which encompasses both spatial and temporal elements, requires specialised explainability methods distinct from those used for static images [182].

XAI for video data seeks to clarify how these models arrive at their decisions by providing insights into the spatial and temporal focus of the model. Techniques like Saliency Tubes [183], video-adapted GRAD-CAM [124], and Layer-wise Relevance Propagation [106] have been specifically designed to address the challenges of video data. These methods enhance the understanding of model predictions and support the refinement of model performance by highlighting the influence of temporal dynamics and spatial features on decision-making.

Testing with Concept Activation Vectors (TCAV), introduced by [184], interprets neural networks by measuring the impact of human-understandable concepts on model predictions. Using user-provided examples, TCAV creates concept vectors and evaluates their influence through directional derivatives. TCAV can analyse high-level concepts like "motion patterns" or "object interactions" across frames, linking semantic ideas to model outputs. This makes it particularly useful in applications such as action recognition, bias detection in video datasets, and debugging temporal models. Its ability to connect abstract concepts to predictions fosters transparency in complex video-based neural networks.

The following Table 3.4 describes a range of explainability methods that have been applied or adapted for video data, outlining their type, scope, data specificity, and whether they are model-agnostic or tailored to specific architectures.

Table 3.4: Overview of Explainability Methods for Video Data

Name	Type	Data Type	Scope	Specificity
Saliency Tubes [183]	Saliency visualisation	Video	Local	Specific
Video-Adapted GRAD-CAM [124]	Internal State (Class Activation)	Video	Local	Specific to CNNs
Layer-Wise Relevance Propagation (LRP) [106]	Backpropagation	Video	Local	Agnostic
Temporal Concept Activation Vectors (TCAV) [184]	Concept-Based	Video	Global/Local	Agnostic
Spatio-Temporal Attention visualisation [185]	Attention Mechanism	Video	Local	Specific to Attention Models
SHAP [56]	Perturbation-based	Any	Global/Local	Agnostic

3.2.2 Type of Explanation provided: Global, Cohort and Local Explanations

Explainability tools can be classified based on the type of explanation they provide: global, cohort, or local. Understanding a model from these different perspectives is crucial for a comprehensive evaluation of its performance. Global explanations offer insights into the overall relationships between inputs and outputs across the entire dataset, though they may lack precision in specific instances. Cohort explanations focus on particular groups or segments of data, providing a middle ground between global and local perspectives. Local explanations, on the other hand, delve into individual data points or small groups, often yielding more precise insights due to the linear nature of model behaviour in these smaller regions. Combining these explanation types allows for a more thorough and nuanced analysis of machine learning models [20].

Local Explanations

Local explanations outline the reasoning behind a specific prediction made by an ML model, which enhances local fidelity. They focus on providing insight into segments of the relationship between the inputs and outputs of an ML model, such as certain groups of input records and their corresponding predictions, specific ranges of predictions and the input data associated with them, and even individual data points [186].

Unlike local methods, global methods may not provide an explanation for single observations but rather provide an overall description of black-box models [187]. These explanations are valuable for gaining an in-depth understanding of identifying problems and can be used to answer specific questions. For example, in a computer vision task, the focus might be on identifying the pixels that had the most significant influence on the classification [188].

Local explanations provide a better understanding of feature importance for particular predictions that do not follow the global explanations rules, e.g., in the health sector, for specific cases, some patients might have symptoms that play an important role in diagnosing this particular patient, while the same features have no important role on a global scale [189]. Novices in application domains usually prefer Local explanations over global explanations to understand the input-output relationship [98].

ICE plots [155] are local interpretability tools that illustrate how the prediction for each individual data instance changes as a specific feature varies while keeping other features constant. However, it should be noted that the average effect of a feature on the model's decision can be misleading and can obscure important interactions among variables. Therefore, it is a better approach when both ICE [190] and PD plots (global explainability tool) are used, as they complement each other in identifying the relationships between variables.

Global Explanations

Gaining insights into your trained model's performance on a global scale is often imperative, while also delving into specific localised aspects of your data or predictions to extract detailed information. Global assessments enable the comprehension of inputs and their overall relationship with the target prediction, yet these can sometimes yield approximations that might not hold true in all scenarios.

Global methods help users to logically understand the relationship between input variables and the predicted output. They help in forming an overall understanding of the

behaviour of the model. Users are able to understand all the different possible outcomes. Global explanations assess the magnitude of the impact or identify the features that have the most significant impact on the model's accuracy. Such global explanations are helpful in making informed decisions or supporting or dismissing the idea that a specific feature is substantial [51]. Global explanations primarily cater to business stakeholders and regulators and serve as a reference point for comparing the deployed model's performance with its original behaviour.
Both Accumulated Local Effects (ALE) plots and PD offer global interpretability. ALE addresses some limitations of PD plots, especially when features are correlated. ALE has an advantage over PD as it addresses the bias that can arise when the feature of interest is highly correlated with other features, which is not considered in PD [51].
There are subclasses to distinguish between global interpretability techniques [31] [186] [9]:

- Model Extraction: Using the black-box model, an interpretable model is extracted.
- Feature-based Methods: To determine features' importance.
- Transparent Model Design: To modify or redesign a Blackbox model to improve its interpretability.

Cohort Explanations

Cohort explanations occupy an intermediate position, offering insights into a specific subset of predictions generated by the model. In other words, they provide an understanding that is more focused and contextualised compared to global interpretations, yet broader in scope than local insights, which pertain to individual instances. This approach aids in grasping patterns and tendencies within distinct groups of predictions, enabling a more nuanced comprehension of the model's performance within specific scenarios or conditions.

3.2.3 Model Specificity: Model-Agnostic vs. Model-Specific Explanation Techniques

Another important way to classify explanation techniques depends on the model being used. These techniques can be model-agnostic or model-specific. Model-agnostic methods can be applied to any model, while model-specific methods are limited to specific models [153].

Model-Specific

Model-specific methods involve examining or utilising the internal components of a model. Model-specific interpretability methods generally use the model itself for interpretation, which can provide more accurate information.

Model-Agnostic

Model-agnostic methods examine the connection between the input and output of trained models without relying on their internal structure [191]. These methods are beneficial when no theory or understanding of what is happening within the model exists. Model-agnostic explanation methods typically only have access to the model's output and do not require any information about the model's internals, making them widely applicable and flexible. These methods have the benefit of not affecting the model's performance as they are separate from the internal functioning of the model by treating the ML models as black-box functions [23].
Popular examples of model-agnostic post-hoc explanations methods include LIME [57], LRP [106] and SHAP [56].
While model-agnostic interpretability techniques are convenient and have advantages, they often rely on substitute models or other estimates that can reduce the precision of the information they provide. In contrast, model-specific interpretability techniques use the model to be explained directly, resulting in more accurate measurements [192].

3.2.4 Explanation Technique

Explainability tools can also be classified by the specific techniques used to derive these explanations. Each technique provides a unique approach to interpreting model behaviour, depending on the underlying method of deriving insights. Common explanation techniques include feature relevance, model simplification, example-based explanations, counterfactuals, rule-based explanations, and gradient-based methods. These techniques offer various ways to interpret model decisions and provide insights tailored to different model architectures and application contexts. The following outlines key explanation techniques:

Feature Relevance/Importance

Feature relevance methods aim to attribute importance scores to individual features, providing insights into their influence on model predictions. LIME [57] and SHAP [56] are prominent examples of feature relevance methods, as both assign importance scores to features by assessing how changes in input data affect predictions. These techniques rely on perturbation to generate alternative versions of the input data, which allows them to analyse the relationship between inputs and outputs. By creating perturbed samples, LIME and SHAP can apply simple, interpretable models—such as linear models—to explain how the original, complex model arrives at its predictions.
LIME is a popular, model-agnostic explainability tool, applicable across various domains [105]. LIME works by perturbing a single data instance and, with the help of a surrogate model, examines the impact of these perturbations on the model's predictions. By focusing on local behaviour, LIME offers locally faithful explanations, giving insights into how the ML model behaves in the vicinity of the instance being explained. This flexibility makes LIME suitable for explaining any ML model, regardless of its underlying architecture.

However, LIME has limitations. Because it relies on a surrogate model to approximate the original model's behaviour, the quality of explanations is closely tied to how well the surrogate model fits. This reliance can lead to high computational costs and may introduce variability, potentially resulting in inconsistent explanations for the same input sample. Furthermore, LIME's perturbation process is random and can sometimes generate samples far from the original input data, which may produce misleading explanations [126].

To address these limitations, a modified version of LIME has been proposed [193]. Instead of random perturbations, this approach uses K-nearest neighbours to create input samples that remain closer to the original instance, aiming to produce more reliable explanations. However, this method presumes that test data follows the same distribution as training data, which may not always hold true.

Another extension, Submodular Pick LIME, has been developed to address the issue of providing broader insights beyond single-instance explanations. Rather than choosing instances randomly, Submodular Pick LIME carefully selects a diverse set of representative instances from the input dataset to offer non-redundant, multi-instance explanations. This approach provides a more comprehensive understanding of the model's behaviour as a whole [105].

RObust Post-hoc Explanations (ROPE), introduced by [7], represents an alternative to overcome the limitations of traditional post-hoc methods like LIME and SHAP. ROPE aims to provide consistent explanations by ensuring that explanations for similar predictions remain stable even when inputs undergo small adversarial perturbations. This approach achieves robustness by applying perturbations to the input distribution, rather than to individual data points, enhancing the reliability of explanations [194].

Gradient-Based Technique

Gradient-based techniques are explainability methods that utilise the gradient information of a model with respect to its input features to determine which features are most influential in the model's decision-making process. These methods analyse the sensitivity of the model's output to small changes in the input by calculating gradients. Generally, gradient-based explanations are structured to match the size of the input data, assigning a relevant value to each part of the input, such as individual pixels in an image or tokens in a text. This allows for a detailed, feature-level understanding of what drives the model's predictions [172].

Gradient-based attribution methods, such as saliency maps [125], which assign importance scores/attribution to each input feature and show which parts of the input image are most important. This is done by comparing removal scores of different regions of the image and finding the smallest region that causes a significant drop in the classification score [195]. Saliency methods (also as known as attribution maps or sensitivity maps) use different approaches (e.g., perturbation-based methods, and gradient-based methods) to show what features in the input strongly influence the model's prediction. Local approximation of the model, on the other hand, trains an interpretable model (learned from the main model) to locally represent the complex model's behaviour.

Another example of a gradient-based method is Integrated Gradients [172] is a gradient-based explainability technique that calculates feature attributions by integrating the gradients of the model's output with respect to the input features. This approach works by defining a baseline input (usually an input with minimal or neutral values) and measuring the model's sensitivity as the input progresses from this baseline to the actual input in small steps. By summing these gradients over the entire path, Integrated Gradients provide a comprehensive attribution score for each feature, indicating its contribution to the final prediction. This method is particularly effective in neural networks, where it offers insights into feature importance in a way that is both theoretically sound and visually interpretable. Integrated Gradients is often used in applications like image and text models, where understanding the impact of individual features or components (e.g., pixels or words) on the model's decision is essential for interpretability [172].

GRAD-CAM [124] is a visualisation technique that creates visual explanations by highlighting regions in an input image that are most influential to the model's decision. It achieves this by leveraging the gradients from a specific layer in the neural network, often the last convolutional layer, to identify the areas in the image that contribute most to the output prediction. This method is particularly valuable for interpreting CNN-based models, as it provides clear, spatial insights into how the model focuses on specific parts of the input to make its classification or recognition decisions [124].

Another example is Smooth Gradient (SmoothGrad) [173], which reduces noise in gradient-based saliency maps by adding Gaussian noise to the input multiple times and averaging the gradients. This approach sharpens visual explanations, making it easier to identify key features that influence the model's decision [173].

Rule-based Technique

Rule-based techniques in XAI provide interpretability by generating transparent, human-readable rules that describe the model's decision-making criteria. These methods create if-then rules or conditions that specify the combinations of features that lead to certain predictions, making them highly interpretable. Rule-based techniques typically work by breaking down complex model behaviour into sets of rules that can either represent the overall model or explain individual predictions [161]. RuleFit [161] is an example of a rule-based technique. It combines linear regression with rule-based interpretations to produce transparent, interpretable predictions. RuleFit extracts rules from ensemble models, such as random forests or boosted trees, by identifying patterns in the data as if-then rules. These rules are then incorporated as features in a sparse linear model, creating an interpretable structure where both simple linear terms and decision-based rules inform predictions [161].

Anchors [154] explain individual predictions of any black-box classification model by finding a decision rule that "anchors" the prediction. The resulting explanations are IF-THEN statements, which define regions in the feature space where the predictions are fixed to the class of the data point being explained. This ensures the classification remains the same, regardless of changes to other feature values that are not part of the anchor. The main drawback of Anchors is that it only supports textual and tabular data. It can generate explanations in tables and text, but it is limited to those types of data.

Another example is Bayesian Rule Lists (BRL) [196], an interpretable, probabilistic model that employs a Bayesian framework to create simple, rule-based lists for classification. BRL generates concise sets of if-then rules derived directly from data, making the resulting models intuitive and easy to follow. This approach is particularly valuable in high-stakes fields like healthcare, where interpretability and transparency are crucial

[197].

An additional example is Scalable Bayesian Rule Lists (SBRL) [198], an extension of BRL designed to handle larger datasets while maintaining interpretability and computational efficiency. SBRL addresses the limitations of standard BRL by introducing scalable sampling techniques, allowing it to generate rule-based models even when working with extensive data.

Model Understanding through Subspace Explanations (MUSE) [120] can also be considered as an example of rule-based techniques. MUSE aims to understand the behaviour of black-box models by providing explanations in specific subspaces that are defined by features of particular interest to users. At a high level, MUSE generates a set of If-then rules globally. In addition, based on the features of users' interest, it generates a separate set of rules [120]. One of the advantages of MUSE that differentiates it from other XAI tools is that MUSE produces tailored explanations by allowing users to specify features of interest, which narrows the search subspace [21].

Counterfactual Technique

Counterfactual explanations are a powerful approach within Explainable AI, helping users understand how small, hypothetical changes to input features could lead to different predictions from a model. By focusing on "what-if" scenarios, counterfactuals provide actionable insights into how individuals might alter features to achieve a desired outcome. For instance, in a loan application, a counterfactual explanation might reveal how adjustments in income or credit score could lead to loan approval, giving users specific guidance on factors influencing model predictions [26].

One example of a counterfactual explanation tool is DiCE. DiCE [168] generates a diverse set of counterfactual explanations for a given prediction, illustrating multiple ways that an outcome could change through small modifications in input features. This diversity allows users to explore different scenarios that could lead to alternative outcomes, making DiCE especially valuable in applications such as credit scoring or loan approval, where understanding various paths to a desired result is beneficial. Counterfactual tools often focus on creating realistic and feasible scenarios that provide users

with a clearer view of how specific changes in their data could affect outcomes [168].
Another example of a counterfactual explanation tool is FACE [169]. FACE generates counterfactual explanations that are both feasible—meaning they are realistic within the input space—and actionable, ensuring the suggested changes are practical for users. This feature makes FACE especially useful in contexts where the recommendations need to be achievable. By focusing on feasible and actionable recommendations, FACE supports users in understanding how minor adjustments in input features might lead to a favourable outcome, making it an ideal tool for applications requiring practical, real-world changes [169].
FACE and DiCE both provide counterfactual explanations, but with different focuses. FACE emphasises feasibility and actionability, generating realistic, achievable counterfactuals suited for practical applications like healthcare. In contrast, DiCE prioritises diversity, offering a range of alternative scenarios to explore various ways an outcome could change, which is valuable for understanding multiple factors in exploratory contexts like credit scoring.

Example-based explanations

Example-based explanation or explanation-by-example techniques involve selecting instances from testing and training datasets that the model predicts well and generates global or local explanations for the task model. Unlike other techniques, example-based techniques do not highlight the most important features, but instead, they create a simplified understanding of the black-box model [51].
There are several example-based explanation techniques, such as CEM [167], kNN [199], and Trust Scores (TSM) [200]. CEM [201] provides local explanations for predictions by comparing them to another prediction. CEM is based on the concepts of pertinent positives and negatives. Pertinent positives are the minimal set of features that can justify the prediction. In contrast, relevant negatives are the minimal set of features that must be absent from asserting the prediction, and both concepts provide a complete explanation.
KNN can be used as an example-based tool to provide local explanations for black-box

classifiers by identifying the most common among the k-closest data samples. TSM is classified as an explainability tool. It is a way to determine the confidence of a black-box classifier model in making a specific prediction. However, it does not explain the prediction but generates a trust score. It uses a modified version of the nearest neighbour classifier to calculate a trust score for a specific instance.

3.2.5 Timing of Application: Pre-model, Ante-Hoc and Post-Hoc Techniques

Explainability techniques can be classified based on when they are used within the modelling process, as illustrated in Figure 3.2. These techniques are commonly grouped into three categories: Pre-Model, Ante-hoc and post-hoc XAI tools. Post-hoc techniques are considered the most common because they offer flexibility and can be applied to existing trained models [14].

Pre-Model Techniques

Pre-model methods are done in the first stage of model development, after getting the data and before selecting the desired ML model appropriate for the problem statement. The main goal of using pre-model methods is to understand and describe the data used in the ML model. It is crucial for researchers to spend some time understanding their data. Pre-model methods include exploratory data analysis techniques [14].

Exploratory Data Analysis (EDA) is an approach that has different techniques starting from observing the data given to understanding its structure, picking important variables, odd values and outliers and proposing a good statistical model for it. Figure 3.3 shows some examples of EDA. Authors in [192] were the first to suggest the term "Exploratory Data Analysis" and show that using graphical methods (e.g., histograms and scatter plots) and quantitative techniques (e.g., interval estimation, measures of location or of scale and shapes of distributions) can be helpful to explore our data. Exploratory data analysis has several techniques that can be applied to the data, such as:

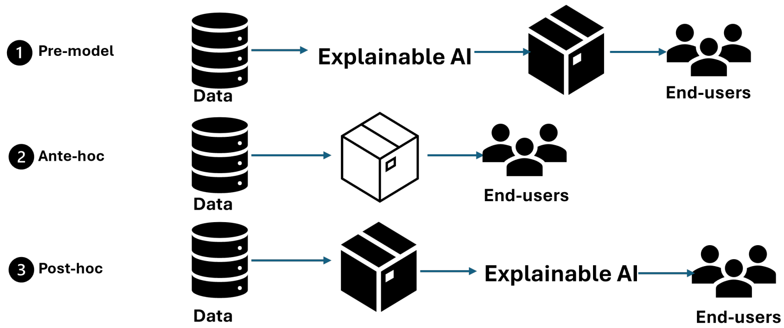


Figure 3.2: Pre-Model, Ante-Hoc, and Post-Hoc tools.

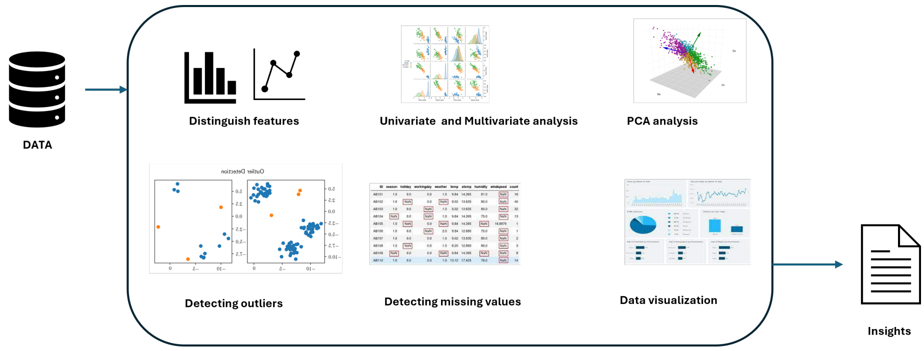


Figure 3.3: Examples of Exploratory data Analysis Techniques.

- **Principal Component Analysis (PCA)**: PCA [202] is a widely used technique in dimensionality reduction. The primary purpose of PCA is to decrease the number of variables under consideration while retaining as much variance from the original data as possible. It achieves this by transforming the data into a set of new, uncorrelated variables known as principal components, which sequentially capture the highest levels of variance in the data [203].

- **T-distributed Stochastic Neighbor Embedding (T-SNE)**: t-SNE [204], as described by [205], is a technique used primarily for visualizing high-dimensional data. It operates in an unsupervised manner, emphasising the preservation of local data structures by computing similarities between pairs of data points. Using a cost function, t-SNE assigns each data point a specific location in a two- or three-dimensional space, enabling users to visually explore complex datasets. While effective for capturing local relationships, t-SNE does not retain a fixed mapping, meaning it requires a re-computation when introducing new data points, as it does not inherently support updating based on new data.

Ante-Hoc Approach

Interpretability can be easily achieved by using a limited set of algorithms that produce interpretable models. Linear regression, logistic regression, and decision trees are examples of commonly used interpretable models [51].

One approach to achieving explainability is to use models that are naturally interpretable by humans, meaning their inner workings and decision-making processes can be easily understood without complex analysis. Using transparent models is one approach to achieve interpretability. In these models, humans can easily understand how inputs are mathematically mapped to outputs by having technical knowledge of the model itself and the algorithms used in the models [51].
Transparent Models can be simulatable (users can understand how the model functions), decomposable (users understand each component in the model), generative in which samples can be drawn from the model, and algorithmically transparent (users have prior knowledge of the algorithms behind the model) [187]. One of the drawbacks of this approach is that it is model-specific, and the number of representative models available to choose from is limited. Here are some examples of interpretable models:

- Linear regression model: Linear Regression is a model that predicts the target by taking the weighted sum of the feature inputs. The linear relationship between the inputs and the target makes it easy to interpret the results. Linear regression models are widely used by statisticians, computer scientists, and other professionals who deal with quantitative problems [206].

- Logistic regression: When dealing with classification problems where the objective is to determine the probability of data belonging to a particular class, linear regression is not a suitable approach. In such cases, logistic regression is a better alternative as it models the probability of a discrete outcome based on input variables. Unlike linear regression, which focuses on linear interpolation between data points [207], Logistic regression involves creating a model to predict the probability of a discrete outcome based on an input variable [208].

- Decision trees are another example of interpretable models. Decision trees facilitate prediction and classification while possessing inherent explainability. They function by recursively dividing data based on their attributes, grouping them into subsets in a hierarchical structure [52]. Predictions made by decision trees are inherently interpretable [207].

Post-Hoc Approach

Post-hoc explanation techniques are used to explain how models that are not interpretable by design work. These techniques are applied to a model after it has been developed to understand how predictions are made. The majority of pre-model and post-hoc explainability techniques are designed to be model-agnostic, meaning they can be applied across a broad range of different models.

Typically, post-hoc techniques involve using an external, simpler model that mimics the behaviour of the complex model in order to make it more transparent. Post-hoc interpretability is a different way of gaining insight from trained models. Although post-hoc interpretation may not provide a detailed understanding of how a model works, it can still provide valuable information for practitioners and users of ML [188] [209].

Currently, most black-box models are explained using a post-hoc approach. This approach is used for complex models in which humans cannot understand the underlying decision-making mechanism. One of the advantages of post-hoc approaches is that they do not affect the performance of a complex model as they treat the model as a black box [86].

Post-hoc approaches have several disadvantages; one of them is that post-hoc approaches have the ability to capture correlations but not causal relationships between input data and black-box predictions [52]. Post-hoc approaches might derive different explanations for the same prediction [120] [22] [210].

3.2.6 SHAP

A more detailed exposition of SHAP is provided here, given its central role in the explainability framework proposed and experimentally validated throughout this research. SHAP [56] is another example of a post-hoc approach based on perturbation. SHAP or Shapley Additive exPlanations tool computes feature importance by computing the contribution of each feature to the output obtained. These contributions are calculated using coalitional game theory, where each feature represents a player in a coalition [51]. SHAP tool increases transparency by producing SHAP values for each observation [51]. These SHAP values can be aggregated to provide global interpretability for ML models. SHAP is considered an optimal approach for providing interpretability since it is built on a theoretical foundation [56].

A Shapley value is computed through cooperative game theory, with Lloyd Shapley introducing this method in 1951. Grasping the fundamental concept of Shapley values is crucial to determining their suitability for your specific scenario and effectively leveraging them alongside your ML model.

Shapley values operate by analysing the impact of each individual feature on the model's predicted outcome, achieved by generating numerous predictions using subsets of the model's features and comparing the resulting predicted values [25].

SHAP uses Shapley values to assess the effect of individual features on the model's outcome. SHAP guarantees consistency and local accuracy by considering all possible predictions and input combinations. It comes in three variants: KernelSHAP, TreeSHAP and DeepSHAP as shown in Table 3.5 [51] [211]. Asymmetric Shapley Values (ASV) [212] is a variation of SHAP that considers the causal relationship between variables in the model explanation process. ASV values are asymmetric, unlike Shapley values, and they are often used in model fairness analysis because they can capture the indirect effects of a variable on the model.

SHAP is a powerful tool for interpreting complex machine learning models by providing insights into the contribution of each feature to individual predictions. It produces both local and global plots to visualise these contributions. For instance, the SHAP force plot illustrates the impact of each feature on a specific prediction, showcasing how the prediction changes from the base value to the final value, as shown in Figure 3.5. On the other hand, the SHAP summary plot as shown in Figure 3.4 offers a comprehensive overview of feature importance across multiple predictions, aiding in understanding the overall behaviour of the model. However, these plots can sometimes be challenging to interpret without proper context.

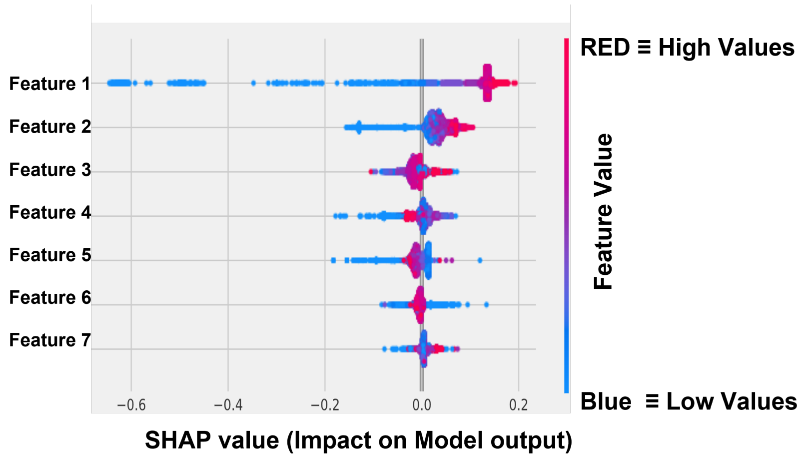


Figure 3.4: Example of SHAP summary plot.

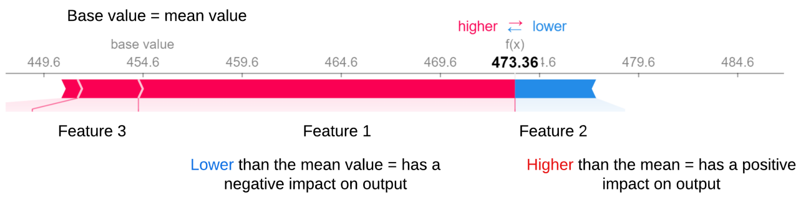


Figure 3.5: Example of SHAP force plot

Table 3.5: Comparison of SHAP Types

SHAP Type	**Description**	**Key Features**
KernelSHAP	A model-agnostic method for computing SHAP values applicable to any model, using weighted linear regression and a kernel-based data perturbation mechanism.	• Suitable for high-dimensional datasets • Compatible with any model • Flexible interpretability enhancement
TreeSHAP	Optimised for tree-based models (e.g., decision trees, random forests, gradient-boosted trees), leveraging the internal tree structure for efficient SHAP value computation.	• Captures non-linear and non-additive interactions • Scalable for large datasets and complex models
DeepSHAP	Integrates DeepLIFT with SHAP for deep neural networks, handling complex feature interactions through layer-wise relevance propagation.	• Designed for deep neural networks • Propagates SHAP values from output to input layers • Ensures consistency and fairness properties

3.2.7 Theoretical Background Behind SHAP

SHAP values are based on principles from cooperative game theory, specifically the Shapley value, which provides a fair allocation of "payouts" (or contributions) among players (features) within a coalition. In the context of machine learning, SHAP values measure each feature's contribution to a prediction by calculating the average contribution of that feature across all possible subsets of features.

The Shapley value for a feature i is denoted as ϕ_i and is given by:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} \left(v(S \cup \{i\}) - v(S)\right) \tag{3.1}$$

where:

- S: subset of features not including i,
- N: set of all features,
- $v(S)$: model prediction (or "value") when only features in subset S are considered,
- $v(S \cup \{i\}) - v(S)$: marginal contribution of feature i.

The equation uses factorial terms $|S|!$, $(|N| - |S| - 1)!$, and $|N|!$ to appropriately weigh each subset, accounting for different permutations in which feature i can be added to various subsets S.

Thus, the Shapley value ϕ_i for feature i represents the weighted average of its marginal contributions across all possible feature subsets, providing a fair measure of its influence on the model's prediction.

Computational Complexity and Approximation via Monte Carlo Sampling

Calculating the exact Shapley value requires evaluating all possible subsets of features, which becomes computationally infeasible as the number of features grows due to the exponential increase in subset combinations. To address this, Monte Carlo sampling [213] is often used as an approximation technique, estimating the Shapley values by averaging over a set of randomly selected subsets.

The Monte Carlo approximation for the Shapley value, $\hat{\phi}_i$, is given by:

$$\hat{\phi}_i = \frac{1}{M} \sum_{m=1}^{M} \left(v(\mathcal{S}_i \cup \{x_i^m\}) - v(\mathcal{S}_i)\right) \tag{3.2}$$

where:

- $\hat{\phi}_i$: estimated Shapley value for feature i,
- M: number of Monte Carlo samples,

- $\mathbf{x}$: feature vector, with x_i^m the m-th sample of feature i,
- $v(\mathcal{S}_i \cup \{x_i^m\})$: model prediction when feature i is included,
- $\mathcal{S}_i$: randomly chosen subset of features excluding i,
- $v(\mathcal{S}_i)$: prediction based on subset $\mathcal{S}_i$.

By using this approximation, we can estimate the contribution of feature i without computing every possible subset, making SHAP feasible for large models and datasets.

3.2.8 Comparative Analysis of SHAP and Other XAI Tools

In this section, the extensive review of the literature conducted in previous sections is leveraged to present a comprehensive comparison of various XAI tools. The focus of this comparison is to elucidate how SHAP, a prominent tool in the field of XAI, stands in relation to other tools in terms of capabilities, limitations, and model dependencies.

Understanding the multifaceted nature of XAI tools is crucial for practitioners and researchers who aim to implement the most effective methods for interpreting complex ML models. The comparison Table shown in Table A.1 in Appendix A is not just a mere list; it is a thoughtfully curated guide that evaluates each tool based on the type of explanations it provides, its unique features, the scope of explanations provided, and its dependencies on specific models. This comparative analysis is instrumental in shedding light on the strengths and weaknesses of each tool, including SHAP, within various operational contexts.

3.3 The XAI Classification Range

XAI is a wide-ranging field that encompasses a variety of techniques and definitions. The field itself is new and rapidly evolving, and as a result, there are many different approaches and interpretations of explainability. The concept of transparent models or interpretable models, which relates to understanding the internal mechanisms of an AI system and how it reaches decisions, sits at one end of the XAI spectrum as shown in

Figure 3.6. On the opposite end sits post-hoc explainability techniques, which concentrate on the ability to explain the decisions made by black-box systems. In between these two extremes, various methods and approaches fall under the XAI umbrella, such as those previously mentioned in this research. These different approaches have one common goal but different trade-offs [20].

The definition of XAI can change depending on the context, stakeholders involved, and application. The overall goal of XAI is to enhance the trust and accountability of AI systems by enabling a better understanding of how they work and why they make certain decisions [33] [187] [26] [28].

XAI tools are used to increase the trustworthiness and transparency of AI systems by providing additional information in the form of explanations that give insight into their decision-making or internal workings.

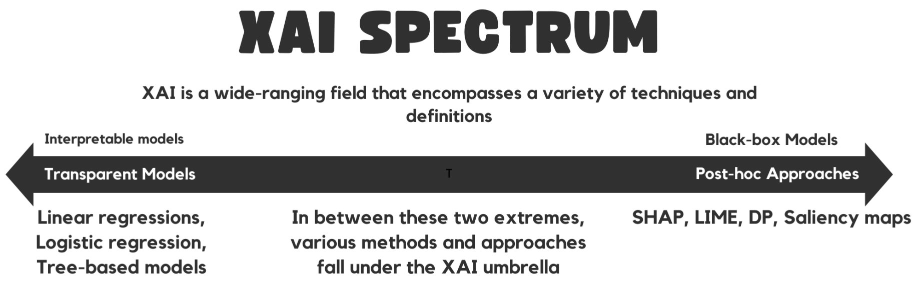


Figure 3.6: XAI Spectrum

3.4 A Note on XAI for Generative Models

Generative models are currently at the forefront of AI innovation, marking a transformative trend in the tech landscape. These models are reshaping industries by providing a new paradigm for data generation and analysis. Understanding the intricacies of generative models is pivotal. These models represent a sophisticated branch of statistical algorithms with the unique capability to create new datasets. They achieve this by meticulously estimating and replicating the probability distribution of an original dataset, thereby enabling the generation of data that mirrors the initial distribution. Unlike their discriminative counterparts, generative models excel in producing novel samples without the need for explicit labels, operating under the unsupervised learning paradigm. This independence from labelled data allows them to leverage existing datasets to their full potential [214].

Delving into the interpretability of generative models presents a relatively novel challenge in the field of XAI. Cutting-edge generative technologies like GPT-3, ChatGPT [50], DALL-e 2 [215], and stable diffusion methods, which produce text or images from given prompts, occasionally demand an explanation for their outputs. Whether these explanations are required hinges largely on the context and purpose for which these AI-generated samples are produced.

Among the plethora of generative models, variational autoencoders [216] and generative adversarial networks (GANs) [44] have garnered particular attention for their robustness and versatility. However, the landscape of generative models is diverse, including other significant models such as Gaussian mixture models, hidden Markov models, Boltzmann machines, and latent Dirichlet allocation. As these models are integrated into the broader spectrum of XAI and Cutting-Edge Models, a critical inquiry arises: How to ensure that the operation and outcomes of these advanced models remain interpretable and transparent?, Addressing this question is not only essential for maintaining trust and accountability in AI systems but also for harnessing the full potential of these cutting-edge models in various domains [214].

3.4.1 Attention Mechanisms in Generative Models

In many advanced generative models, particularly in transformer-based architectures like GPT-3 and DALL-E 2, attention mechanisms are fundamental.

Attention is a mechanism that enables deep learning models to prioritise certain parts of an input sequence during processing. This approach is inspired by human cognition—specifically, our tendency to selectively focus on specific elements within a scene, such as emphasising certain words that convey key information when reading [217].

Attention layers enable the model to selectively focus on specific parts of the input data while generating each output token or pixel. This ability to attend to relevant sections of data is essential not only for enhancing model performance but also for improving interpretability. By examining which parts of the input the model attends to, insights are gained into its decision-making process. For instance, visualising attention weights allows us to identify the words or image regions that were most influential in the model's output, making the generation process more transparent and justifiable [58].

Types of Attention in Generative Models

Self-Attention: Self-attention, or intra-attention, calculates dependencies within a single sequence [218]. Each token in the sequence generates a query, key, and value, and attends to all other tokens using an attention score. This score determines the relative importance of tokens in relation to one another, making self-attention effective for capturing long-range dependencies in tasks like text and sequence generation [217].

Scaled Dot-Product Attention: Scaled dot-product attention is a mechanism that helps models identify the most relevant parts of the input data by comparing "queries" (what the model is focusing on) with "keys" (representing the input) [219]. It assigns importance scores to each part of the input, scales these scores to ensure stability during training, and converts them into weights. These weights are then used to combine "values" (the actual information) into an output that highlights the most important details for the task at hand. This process allows the model to efficiently focus on the most critical information [45].

Multi-Head Attention: Multi-head attention extends scaled dot-product attention by projecting inputs into multiple subspaces, allowing the model to focus on different relationships within the data [58]. Outputs from each head are concatenated and linearly transformed, enhancing the model's representational power and its ability to learn syntactic and semantic dependencies [45].

Advantages of Attention Mechanisms

Attention mechanisms offer several key advantages that make them superior to traditional approaches like RNN and LSTM. One significant benefit is parallelisation, which allows attention mechanisms to process entire sequences simultaneously, greatly improving computational efficiency compared to sequential models [39]. Another advantage is their ability to capture long-range dependencies by connecting any two sequence positions in a constant number of operations. This capability effectively addresses the vanishing gradient problem encountered in traditional models, enabling better handling of long sequences [45]. Moreover, attention mechanisms enhance interpretability, as visualising attention weights provides valuable insights into which inputs influence specific outputs. This makes them particularly useful for tasks requiring explainability and debugging [217].

Key Contributions of the Transformer

The Transformer architecture, introduced by [45], highlights the remarkable power and efficiency of attention mechanisms in generative tasks. By replacing recurrent and convolutional layers with stacked self-attention and feed-forward layers, the Transformer simplifies the model architecture while enhancing scalability. It incorporates positional encodings to capture sequence order, allowing the model to process input tokens in parallel without requiring explicit recurrence. This design significantly reduces computational complexity and training time, enabling the Transformer to achieve state-of-the-art results in tasks such as machine translation.

3.4.2 Future Directions

Advancements in attention mechanisms continue to push the boundaries of generative models. One significant development is sparse attention, which reduces computational complexity by focusing on subsets of tokens. This approach makes it suitable for handling long sequences efficiently, as it lowers the computational load without sacrificing performance [220]. Another promising direction involves Hybrid models that combine attention mechanisms with convolutional or recurrent layers. By leveraging the strengths of different architectures, these models can be tailored for domain-specific applications, enhancing their versatility and effectiveness. Additionally, multimodal attention extends attention mechanisms to handle multimodal data, enabling models like DALL-E [49]and Contrastive Language–Image Pre-training (CLIP) [48] to process and generate content across different modalities, such as converting text prompts into images [48].

Attention mechanisms will remain central to the advancement of generative models, facilitating the development of more efficient, scalable, and interpretable AI systems. As research progresses, these mechanisms are expected to contribute significantly to the capabilities of AI, enabling it to tackle increasingly complex tasks across various domains.

3.4.3 Explainability Needs in Generative Models

Explainable AI techniques for generative AI provide explanations that clarify the outputs generated by the model, whether for specific inputs or in terms of overall model behaviour [62]. Historically, explanations have fulfilled various purposes to meet diverse needs, such as building trust in the model and aiding in debugging processes [102]. The discourse on generative models' explainability revolves around their necessity and feasibility. In low-risk domains like art, the need is minimal, though legal and privacy concerns can arise. In high-risk areas like healthcare, explainability is critical for understanding and correcting model errors. However, achieving clear explanations is challenging, particularly in complex models that learn from probability distributions [17].

Generative AI has introduced new demands for explainability as users increasingly interact with and control model outputs through prompt engineering and direct input, amplifying the need for understanding how to tailor and verify results. The ease of Generative AI access has made it pervasive, influencing diverse and sensitive applications, from mental health advice to education. Moreover, Generative AI presents unique security and safety risks, where outputs can inadvertently support harmful activities, making clear and reliable explanations essential for user trust and oversight

[62].

3.4.4 Addressing Privacy Risks in Generative Models

A pressing issue with advanced generative models is their risk of revealing private data. Addressing this involves the challenge of how XAI can clarify the workings of complex models, including those with advanced blending or transformation capabilities, in a way that's understandable. A primary step is defining what constitutes a clear explanation in this context, essential for developing XAI methods that ensure transparency and trust in these sophisticated models [17].

To mitigate privacy risks, various techniques have been proposed. Differential Privacy (DP) is one prominent approach that ensures individual data points cannot be inferred from model outputs. By adding controlled noise during training, DP can reduce the risk of data leakage while still allowing the model to learn useful patterns [221]. Differentially private training techniques have been successfully applied in generative models, but often with trade-offs in model accuracy or quality of generated outputs [222].

Another technique, Federated Learning (FL), enables training on decentralised data, reducing the risk of centralising sensitive information in a single model. In FL, data remains on users' devices, and only model updates are shared, limiting the risk of exposing personal data during the training process. While less common in current generative AI frameworks, federated approaches show potential for privacy-enhanced generative models, especially when applied to sensitive areas such as healthcare and finance [223].

3.5 Conclusion

This chapter presented a structured taxonomy for explainable AI that organises methods across five complementary dimensions: input data type, explanation scope (local, global, cohort), timing of application (pre-model, ante-hoc, post-hoc), explanation technique, and model specificity (model-agnostic vs. model-specific). Rather than imposing a rigid hierarchy, the taxonomy reflects the reality that many modern XAI methods span multiple categories, and that effective method selection depends on how these dimensions interact within a given application.

By reviewing and contrasting established taxonomies in the literature, the chapter positioned the proposed framework as both consistent with foundational classification criteria and extended in a way that better supports practical navigation of the XAI landscape. The modality-driven breakdown (tabular, image, text, and video) demonstrated that explainability strategies must be aligned with the structure of the input data and the requirements of the explanation recipient. The accompanying tables further operationalised the taxonomy by mapping representative tools to their scope and applicability, providing a practical reference for selecting explanation methods in real-world settings.

A focused discussion of SHAP was included due to its central role in this thesis, grounding its use in cooperative game theory while highlighting computational considerations and common approximation strategies. Finally, the chapter situates explainability methods along a broader XAI spectrum and briefly highlights emerging interpretability challenges in generative and transformer-based models, where attention mechanisms introduce both opportunities and limitations for transparency.

Overall, the taxonomy developed in this chapter provides a coherent organisational foundation for the remainder of the thesis. It supports systematic comparison of explainability techniques and motivates principled selection of tools for subsequent evaluation and deployment, particularly in application contexts where reliability, stakeholder trust, and accountability are critical.

Chapter 4

Condition Monitoring for Engineering Assets

In modern industrial settings, rotating machinery is indispensable, used across sectors including energy, manufacturing, and transportation [224]. Integral to these machines are rolling element bearings, chosen for their durability and low friction under various conditions [225]. Yet, these components are vulnerable to wear, installation issues, and maintenance lapses, making their upkeep critical to operational success.

Condition monitoring represents the natural evolution of predictive and proactive maintenance strategies. Initially developed to extend asset life, these strategies have become vital early failure detection tools. Predictive maintenance, in particular, enables timely interventions based on real-time data rather than reactive or scheduled checks, greatly reducing unscheduled downtimes across industries [226].
This chapter presents foundational concepts in condition monitoring that are intended to provide background for a subsequent case study on thrust bearing wear in pumps. Maintenance strategies and the rationale for adopting data-driven monitoring approaches are first reviewed. Common techniques and metrics used to assess the health of rotating machinery are then discussed. Following this, the role of modern ML techniques in enhancing fault detection and prognostics is explored, and the growing importance of XAI in this context is outlined. To conclude, a practical application of these condition monitoring principles is provided through an examination of centrifugal pump systems.

This includes a discussion of their performance characteristics, optimal operating conditions, and key monitoring metrics—particularly in relation to thrust bearing wear—as presented in the final section of this chapter.

4.1 Introduction

Industries allocate a significant amount of resources for the maintenance of equipment and facilities every year. Insufficient maintenance can result in substantial losses. The use of sophisticated science and technology for condition monitoring and fault diagnosis serves as an effective method to predict potential issues and minimise the expenses associated with machine malfunctions. This approach, known as fault diagnosis, has evolved significantly over the past three decades [227].
Condition monitoring is particularly valuable in pump systems, where components such as thrust bearings play a critical role in managing axial loads and minimising friction. Over time, these components can degrade due to factors like misalignment, unbalanced loads, and inadequate lubrication. By employing technologies such as vibration analysis, temperature monitoring, and acoustic emission analysis, maintenance teams can identify issues in their early stages, preventing severe wear and unexpected breakdowns. For example, increasing vibration levels may indicate issues with alignment or balancing, while elevated temperatures may suggest lubrication problems. Such insights enable maintenance teams to schedule timely interventions, thereby reducing the likelihood of major failures and contributing to asset longevity [228].

Condition monitoring technology has evolved significantly over the past three decades, now incorporating sophisticated science and data-driven methodologies to predict potential issues and minimise expenses associated with machine malfunctions. Advanced diagnostic technologies for engineering assets leverage real-time and historical data to detect deviations in operational performance and scrutinise key attributes of machinery, whether in use or not. This data-centric approach allows for an accurate assessment of the current condition of critical parts, anticipates emerging issues, and informs strategic maintenance plans, providing industries with a powerful tool to optimise asset reliability, therefore, encompasses a range of techniques designed to detect potential issues before they escalate, combining fault diagnosis with predictive maintenance strategies to support critical components in rotating machinery. This chapter will explore these monitoring techniques in depth, as they are applied to various types of machinery. The insights gained here will be further illustrated in Chapter 4, where the condition monitoring of thrust bearings in pumps will be discussed as a case study. This approach not only highlights the importance of proactive maintenance but also demonstrates how condition monitoring can mitigate risks, reduce costs, and enhance overall equipment reliability.

The choice of maintenance strategy depends on both the specific machinery in question and the resources available to the asset manager. It's essential to provide a clear rationale for selecting a particular strategy. For instance, investing substantial resources in condition-based maintenance for easily replaceable components would not be cost-effective. On the other hand, critical assets require careful consideration of their ongoing health based on usage patterns to prevent potentially catastrophic failures [229]. In addressing this issue, three conventional maintenance strategies are commonly employed [230]:

- Reactive Maintenance (RM): Takes place when a machine component experiences a failure, rendering it inoperable. In such cases, the malfunctioning machine component must be either replaced or repaired [231].
- Scheduled Maintenance (SM): A strategy involving maintenance performed at predetermined time intervals is known as scheduled maintenance. This approach encompasses activities such as inspections, adjustments, and planned shutdowns. Its implementation is contingent on assessing the probability of machine components failing within specified time intervals [232].
- Condition-Based Maintenance (CBM) or Predictive Maintenance (PdM): These strategies distinguish themselves from RM and SM by their data-driven nature, which aids operators in scheduling maintenance activities. In Condition-based maintenance, assets are continuously monitored, and maintenance is scheduled when they deviate from normal functioning. Conversely, predictive maintenance

employs proactive, data-driven maintenance methodologies to evaluate equipment conditions and determine the optimal timing for maintenance interventions [233].

Mechanical systems are frequently employed in industrial settings. When exposed to harsh environments, these widely used mechanical devices can deviate from their original state, leading to potential failures and ongoing degradation. Defects and malfunctions in mechanical parts can arise unpredictably, often leading to significant setbacks such as equipment breakdowns or production halts. Therefore, early detection of unforeseen faults and breakdowns in mechanical equipment is crucial. Implementing a condition monitoring process is beneficial in this context, as it facilitates the timely identification of issues, minimises system downtime and results in cost savings. Mechanical system faults can arise from various sources, including force imbalances, issues with the gearing system, bearing defects, undue stress on the shaft, or simply choosing the wrong machine for a specific task [234].

Consequently, a significant number of researchers have turned their attention to failure detection and diagnostic methods to monitor the health of mechanical systems. Various techniques, from visual inspections to sensor-based condition monitoring, have been devised.

These systems go beyond just enhancing equipment reliability; they also serve as the data foundation for the contemporary notion of "digital twins". These "digital twins" are digital replicas of actual machinery in industrial or aviation settings. These computer models undergo continuous refinement using data gathered from sensors and analysis systems, all integral components of the condition monitoring system [226].

Recently, there has been a surge of outstanding research in the field of mechanical equipment condition monitoring. This research [235] delves into an online method for feature learning, allowing for tailored automatic feature extraction per machine. This approach leverages the principles of sparse coding and dictionary learning, which notably compress raw sensor data. Such compression is particularly beneficial in systems such as Internet of Things (IoT) devices and wireless sensors, given their restricted processing and communication abilities. The study prominently highlights vibration monitoring in rolling element bearings, critical components in rotating equipment for load management and friction reduction. Monitoring their state is vital for ensuring the machinery's optimal performance and reliability.

Within the domain of condition monitoring using multi-sensor architectures, the technique of automated change detection has been employed, leveraging Autoregressive Integrated Moving Average (ARIMA) [236] regression and a multidimensional time series analytical approach. Authors of [237] incorporated graphical presentations of state configurations and frequency band chromograms for system monitoring. They captured vibrational signals from diverse mechanical elements and catalogued them in a specialised database. Before any further processing, this data was categorised into various subsets. It was then analysed in the time domain, frequency domain, and/or time-frequency domains. Following this, the power spectrum density of the data was determined, and mean spectral values were extracted.

One of the approaches of condition monitoring is CBM, or known as predictive maintenance, which is a proactive approach to equipment upkeep that prioritises maintenance tasks based on the actual condition of the machinery, rather than a predefined schedule or set intervals. This method is rooted in the principle of performing maintenance only when necessary, thus optimising resources and minimising unnecessary interventions. The CBM strategy is underpinned by three fundamental stages [238] as shown in Figure 4.1:

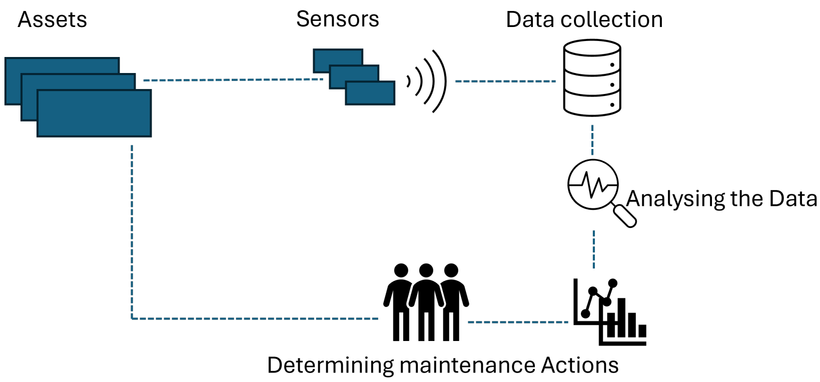


Figure 4.1: Diagram illustrating the stages of CBM, including sensor deployment, data collection, analysis, and determination of maintenance actions by domain experts

- Collecting Data: Data acquisition refers to the act of gathering and preserving relevant data from specific physical equipment for the implementation of CBM [239]. It's a pivotal component when rolling out a CBM strategy, especially for diagnosing and predicting machinery faults or failures. Failures often result from one or multiple machinery issues. Within the scope of CBM, the accumulated data falls broadly into two categories: event data and condition monitoring data. Event data encompasses details about incidents pertaining to the equipment, such as installation, malfunctions, major overhauls, and their respective causes, or actions taken, like minor fixes, preventative measures, or oil replacements. On the other hand, condition monitoring data captures metrics indicative of the machinery's health or operational status [238].

- Analysing the Data: The initial phase of data processing is data cleaning, essential for rectifying inaccuracies, especially in manually inputted event data. Both human errors and sensor malfunctions can introduce data inaccuracies, with the latter affecting condition monitoring data. While some issues can be resolved with manual review, the complexities of data cleaning won't be discussed in depth here [238]. Following the cleaning is data analysis, where different tools and models are employed based on the type of collected data.

- Determining Maintenance Actions: The final stage in a CBM programme revolves around maintenance decision-making. Effective decision support is pivotal for maintenance staff to determine the right interventions. Maintenance decision tools within CBM can be categorised into diagnostics and prognostics. Diagnostics is centred on detecting and identifying faults as they arise. In contrast, prognostics aims to foresee faults or failures before they manifest. Prognostics has an edge over diagnostics because it offers a chance to preemptively address or prepare for issues, potentially reducing unexpected maintenance costs. However, prognostics can't entirely supplant diagnostics. Some issues are unpredictable, and no predictive method, including prognostics, guarantees absolute accuracy. When predictions fall short, diagnostics serve as a backup, ensuring continuous maintenance support [238].

4.2 Condition Monitoring Techniques

There are a variety of Condition Monitoring (CM) tactics for fault detection and diagnosis, from systems based on heuristic qualities to those using AI techniques. Generally, condition monitoring methods are grouped into three main categories, as outlined by [240]:

- Data-driven techniques: Data-driven condition monitoring technology utilises diverse data types and sources gathered during equipment operation. By applying various data mining methods, this technology extracts valuable insights from the amassed data. As a result, it facilitates the monitoring of equipment conditions and aids in diagnosing any faults [241]. Considerable research effort has been dedicated to the advancement of data-driven methodologies in the realms of condition monitoring, detection, diagnostics, and prognostics. Artificial Neural Network (ANN), SVM, Genetic Algorithm (GA), and Auto-Associative Kernel Regression (AAKR) stand out as some of the extensively investigated approaches in this domain [242] [243].
- Knowledge Based Systems: Knowledge-Based Systems (KBS) are a subset of AI, often synonymous with Expert Systems (ES). At their core, these systems leverage human expertise and insights about a specific domain or system. They encapsulate this human understanding into a machine-readable format, which allows them to recognise, diagnose, and suggest remedies for problems that typically demand human intervention and specialised knowledge. Essentially, KBS aims to mimic the decision-making capabilities of a human expert, enabling automated systems to tackle complex issues using previously acquired human wisdom and experience [244].
- Model-based techniques: Model-based methods create a mathematical representation of a system using empirical data. If there's a consistent discrepancy between the output from the real system and its mathematical counterpart, it's taken as an indication of a fault [245]. A significant challenge of using a model-based approach is that the results from the mathematical model seldom align

with the real system, except in cases of very simple or basic systems. That's why applying model-based methods in real-world scenarios can be challenging. They are typically feasible only under certain conditions or within a controlled setting [246].

4.3 Condition Monitoring using ML

ML is a subset of AI focused on crafting algorithms that evolve by learning from data, advancing autonomously without human input [247]. As of 2020, a trend analysis in the PdM sector highlighted a shift towards solutions powered by ML [248]. This analysis underscored ML's pivotal role in catalysing improvements such as reduced downtimes, minimised maintenance costs, extended lifespan of spare parts, enhanced operator safety, augmented production, and confirmed repairs. ML encompasses supervised (e.g., classification, regression) and unsupervised learning (e.g., clustering, dimensionality reduction). Clustering is particularly valuable for grouping data without predefined classes [249].

4.3.1 Supervised ML for Condition Monitoring.

Supervised ML plays a pivotal role in the development of effective learning systems. It provides a structured framework for teaching machines to make decisions based on labelled training data. Supervised Learning techniques encompass two fundamental tasks: classification and regression. Regression pertains to predicting numerical value outputs from future input data, while classification involves predicting discrete value outputs, typically class labels [250].

To develop effective learning systems, it's often necessary to identify and focus on the most relevant features within a dataset, a process known as feature selection. Optimal feature selection can involve exhaustively exploring all possible combinations of features, which becomes computationally challenging, particularly for large datasets, as it falls into the Non-deterministic Polynomial-time (NP-hard) problem category. To address these computational complexities, researchers have developed greedy algorithms.

The advantages of feature selection are manifold, including the ability to mitigate the challenges posed by high-dimensional data, enhance algorithm speed, and improve generalisation capabilities. In essence, feature selection empowers the ML process to concentrate solely on critical features [249].

In [251], the research authors present a framework for sensor classification using multivariate time series sensor data. It transforms this data into coloured images and combines them into a larger image for classification via a Convolutional Neural Network (ConvNet).

The research in [252] underscores the relevance of manifold learning techniques, such as Multi-Kernel Metric Learning Embedded Isometric Mapping (MKML-ISOMAP), in the context of condition monitoring, particularly for high-dimensional sensor data analysis. By effectively reducing dimensionality and extracting meaningful features from sensor data, these methods hold great promise in enhancing the accuracy and efficiency of condition assessment in various applications, including online process monitoring and control, which is vital for ensuring the quality and reliability of machinery and systems.

4.3.2 Semi-Supervised ML for Condition Monitoring.

Semi-supervised methods position themselves in the middle of supervised and unsupervised AD techniques [253]. In the literature, semi-supervised learning methods are notably underutilised in machine condition monitoring. In [254], the authors introduce a multi-class semi-supervised learning framework, incorporating principal component analysis (PCA) and semi-supervised support vector machines. This approach entails extracting statistical features from vibration signals across time, frequency, and wavelet domains to capture fault-specific information. [255] described an approach is closely related to semi-supervised learning to bridge the gap between supervised and unsupervised learning by utilising both labelled and unlabelled data for training. Specifically, in the context of monitoring Photovoltaic (PV) production data for anomalies.

4.3.3 Unsupervised ML for Condition Monitoring.

Unsupervised ML algorithms or clustering algorithms primarily aim to identify similarities among feature vectors within data and subsequently group these similar vectors. These clustering algorithms revolve around the optimal arrangement of data points into clusters, which is a combinatorial task known to be NP-hard. To efficiently address this challenging combinatorial problem, various clustering algorithms have been developed, all sharing the common objective of limiting the exploration of numerous cluster combinations. Three primary types of clustering methods include hierarchical clustering, partition clustering, and spectral clustering. The choice of clustering approach may yield different cluster results, often influenced by the specific nature of the problem at hand, guiding the selection of an appropriate clustering methodology [249].

In [256], authors devised hybrid clustering methodologies that integrate evolutionary computing to achieve the global optima. To detect outliers effectively, they introduced a weighting system that considers both data point volume and density, identifying outliers as those with higher weights.

While [257] developed a novel clustering algorithm for the purpose of monitoring and analysing the behaviour patterns associated with ship handling. This algorithm relies on data collected from an array of sensors, which provide ship-related information that is then integrated into an identification system using trajectory data. To extract valuable insights from this sensor data, a sliding window algorithm is employed. Trajectories are segmented into sub-trajectories, and the behaviour patterns are uncovered through spectral clustering of these sub-trajectories. This method facilitates a deeper understanding of ship handling behaviour throughout various scenarios. Moreover, it holds the potential to enhance the efficiency of ML processes in ship route planning and collision avoidance decision-making. In the context of condition monitoring, this approach can be invaluable for analysing ship behaviour patterns, which can contribute to safety improvements, navigation optimisation, and better decision-making in maritime operations.

4.3.4 Deep Learning for Condition Monitoring.

Deep Learning (DL) represents an ML framework that evolved from conventional neural networks, with its development dating back to approximately 2006. In essence, deep learning relies on the utilisation of large-scale DNN, often described as neural networks with an intricate and deep structure [249].

In their work, Niu and Liu, along with their colleagues [258], introduced a Deep Residual Convolutional Neural Network (DR-CNN) that utilizes gray-scale images generated from multi-sensor data, specifically data collected from multiple 3-axis accelerometers. This innovative approach was designed to tackle the challenge of diagnosing bearing faults using multi-sensor data.

Authors in [259] introduced an innovative method, known as Dynamic Convolutional Graph Convolutional Neural Network (DCG-CNN), for monitoring the condition of gas sensors, particularly those sensing hydrogen gas. Their approach combines a CNN with a Dynamic Convolutional Graph (DCG). The DCG component incorporates the power of a CNN and a GAN, which generates new data samples with similar statistical characteristics to the existing data. This integration proves particularly effective when dealing with imbalanced data samples, enhancing the accuracy of defect detection.

4.4 XAI for Condition Monitoring

One of the prominent criticisms levelled against numerous ML techniques pertains to their inherent black-box nature. Obtaining a complete and precise mathematical description for many ML and deep learning approaches is a formidable challenge. This intrinsic opacity of ML methods presents a substantial limitation within the realm of maintenance. The core issue stems from the fact that relying solely on a single metric, such as classification accuracy, offers an inadequate representation when dealing with the complexities of most real-world tasks [57].

In predictive modelling, it's not just about knowing the prediction outcome; The aim is also to understand why a specific prediction was made. In some situations with low risks, just knowing the prediction results is enough, and detailed explanations may not be required. However, in more complex scenarios, having explanations for the model's decisions is crucial for gaining a better grasp of the data and the problem at hand. When the problem itself lacks clarity, the need for explainability becomes evident because relying solely on prediction outcomes may provide limited insight. The degree of explanation required depends on the complexity of the problem and the depth of understanding sought to be achieved [233].

In [260], the authors introduce a new approach called FaultDXAI. It uses fake data that mimics machine problems to help AI learn. They also use a method called Gradient-weighted Class Activation Mapping (GRAD-CAM [124]) to make AI results easier to understand. Their approach works well in diagnosing machine issues and can be used in different situations. This research shows how AI can be more understandable and reliable when dealing with industrial machines, which can help improve maintenance and reduce costs.

This research [261] highlights the importance of XAI within the context of Deep Neural Networks (DNNs) applied to Predictive Maintenance (PM), particularly in the Aerospace Integrated Vehicle Health Management (IVHM) domain. Its primary contribution is underscoring the necessity of XAI systems when utilising AI, like DNNs, to ensure that the results are clear and understandable for human experts.

This study [262] focuses on monitoring multisensor Magnetic Resonance Imaging (MRI) devices using anomaly detection techniques for complex multivariate time series data. It specifically investigates the detection of abnormalities in sensors during failed scans and presents a framework for MRI sensor condition monitoring with the aid of XAI and feature selection. The process involves pre-processing sensor properties, including normalisation, aggregation, and resampling, to generate statistical features. These features are then used as input for ML models, particularly the Isolation Forest algorithm, for Anomaly Detection (AD). Before applying the AD models, a feature selection step is employed, which utilises an Extreme Gradient Boosting (XGBoost) classifier to improve model performance by removing irrelevant sensor properties.

The concept of explainability, along with various methods to attain it, will be explored in the upcoming chapters. What explainability entails, the strategies employed to achieve it, and the drawbacks associated with each approach will be delved into. In essence, the meaning of explainability, the techniques used to realise it, and the limitations inherent to each of these methods will be thoroughly examined.

4.5 Condition Monitoring Techniques for Centrifugal Pump Systems and Rotating Machinery

The inclusion of Condition Monitoring Techniques for centrifugal pump systems and rotating machinery in this chapter is directly motivated by the focus of the main case study in this research, which examines centrifugal pump operations. These pumps are critical assets in numerous industrial sectors, enabling efficient fluid transport by converting kinetic energy into pressure [263]. In the case study, the proposed framework was applied to interpret machine learning model predictions related to pump degradation, providing actionable insights that contributed to optimised operation and extended service life. In particular, the analysis highlighted the role of operating pumps at their Best Efficiency Point (BEP) in reducing mechanical wear—especially on thrust bearings—and enhancing overall durability. This section, therefore, establishes the technical context for the case study. Condition monitoring for rotating machinery and centrifugal pump systems is crucial for predicting and preventing faults that could result in downtime or costly repairs [264]. Given the high demands on these systems, particularly in industrial applications, effective monitoring strategies are essential for maintaining reliability and performance [265]. This section focuses on monitoring techniques used in detecting early wear in pump components, specifically thrust bearings, by observing key indicators like vibration amplitude, temperature, and acoustic emissions [266].

4.5.1 Pump Performance and Efficiency

Centrifugal pump performance is typically described by a performance curve, a graphical representation that plots various operational parameters against flow rate [267]. Key parameters include total developed head (or pressure), brake horsepower (bhp), and efficiency. Efficiency represents the ratio of the pump's fluid power output to the shaft horsepower input. These parameters, especially efficiency, are critical for designing energy-efficient pumping systems [4].

4.5.2 Best Efficiency Point (BEP)

The BEP is the operating point where a pump functions at its peak efficiency, with minimised radial bearing loads and optimal energy use [267]. Operating close to the BEP is beneficial as it extends the lifespan of components, particularly bearings and seals, due to stable and balanced hydraulic flow. This point is associated with reduced vibrations and minimised wear, contributing to overall cost-efficiency and reliability [4]. Operating a rotodynamic pump at flow rates significantly above or below its Best Efficiency Point (BEP) can lead to adverse effects such as cavitation, vibration, impeller damage, suction and discharge recirculation, and accelerated wear of bearings and seals [268]. These risks are minimised by maintaining operation within the Preferred Operating Region (POR), defined as a range around the BEP where hydraulic efficiency and operational reliability are not substantially compromised [269]. A wider range, the Allowable Operating Region (AOR), may be acceptable under certain conditions, although service life is generally reduced compared with operation within the POR [269]. Figure 4.2 [4] shows the performance characteristics of a centrifugal pump, including head, Brake Horsepower (BHp), and efficiency across different flow rates. The BEP is highlighted, representing the optimal operating point where the pump achieves maximum efficiency. Operating near the BEP minimises energy loss and wear, while moving away from this point results in increased power demands and potential component stress [269].

4.5.3 Operating Outside the BEP

While centrifugal pumps are designed to work near the BEP, real-world conditions may cause operation away from this ideal point. Operating significantly away from the BEP can lead to several issues [4] [269]:

- **Accelerated Component Wear:** Operating far from the BEP increases stress

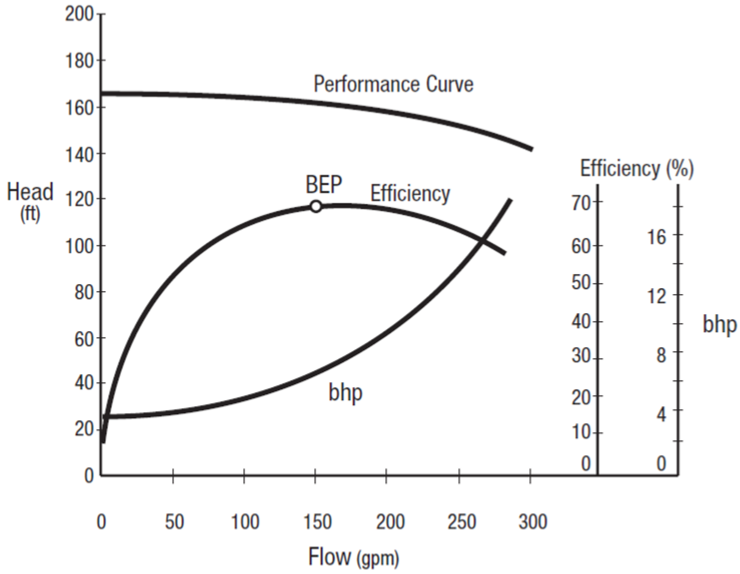


Figure 4.2: Centrifugal pump performance curves showing the best efficiency point (BEP).

on bearings and seals, leading to a shorter lifespan due to unbalanced hydraulic forces.

- **Increased Energy Consumption:** Efficiency decreases when operating away from the BEP, resulting in higher power consumption.
- **Reduced Service Life:** Prolonged operation outside the BEP can reduce the overall reliability and service life of the pump.

Maintaining operation near the BEP whenever possible helps in reducing operational costs and enhancing system reliability, especially in critical applications where uninterrupted service is crucial [270].

4.5.4 Common Monitoring Metrics for Pumps

Pumps are complex machines with various moving parts, including shafts, bearings, and impellers, all of which are subject to wear over time [271]. Monitoring these components can provide early indications of issues, allowing for timely maintenance [272]. Below are some common metrics used for monitoring pumps and identifying potential faults, especially in bearings [273]:

Vibration Analysis

Vibration monitoring is one of the most widely used techniques for detecting faults in rotating machinery, as it provides direct insight into the mechanical behaviour of components like bearings, shafts, and impellers [265]. Vibration analysis can reveal issues such as misalignment, unbalanced loads, bearing faults, and resonance [274]. It is particularly effective for identifying bearing wear and misalignment in centrifugal pumps, where increased vibration amplitude often signals mechanical degradation [273].

Temperature

Temperature monitoring is crucial in identifying early signs of overheating, which may result from poor lubrication, excessive load, or friction due to misalignment [275]. Elevated temperatures in thrust bearings can quickly lead to component failure if left unchecked [276]. According to [277], consistent temperature tracking in pumps helps in early identification of lubrication problems, enabling maintenance teams to take preventive action before serious wear or failure occurs.

Acoustic Emissions

Acoustic emission monitoring captures high-frequency sound waves emitted during the operation of rotating machinery. These emissions are typically generated by friction or impact events within the pump, such as the contact between worn surfaces in a bearing [278]. Acoustic emission techniques are highly sensitive and can detect small changes in component condition before they become detectable through vibration or temperature monitoring alone [279]. Acoustic emission is particularly useful for detecting bearing defects, as it can identify early-stage wear by capturing slight variations in noise emissions [279].

Pressure and Flow Rates

While not always directly linked to bearing wear, monitoring pressure and flow rates in pumps can provide indirect insights into the condition of internal components [280]. Anomalies in these parameters may indicate blockages, cavitation, or inefficiencies that, over time, can contribute to increased wear on bearings [281]. This type of monitoring complements vibration and temperature data, giving a more comprehensive view of pump health.

4.6 Conclusion

This chapter has established the technical and methodological foundations of condition monitoring for engineering assets, with particular emphasis on rotating machinery and centrifugal pump systems. It has reviewed traditional and data-driven maintenance strategies, outlined key condition monitoring techniques, and examined the integration of machine learning methods in diagnostics and prognostics.

By discussing supervised, semi-supervised, unsupervised, and deep learning approaches, the chapter has demonstrated how modern ML techniques enhance fault detection and predictive maintenance capabilities. Importantly, the limitations associated with black-box ML models were highlighted, reinforcing the need for XAI methods in safety-critical engineering contexts.

The discussion of centrifugal pump performance, BEP, and monitoring metrics such as vibration, temperature, acoustic emissions, and flow parameters provides the necessary engineering context for the thrust bearing wear case study presented in subsequent chapters. This technical grounding ensures that the later development and application of the proposed XAI framework are firmly anchored in domain knowledge.

Overall, this chapter contributes to the thesis by bridging classical condition monitoring principles with modern ML-based approaches, thereby establishing the engineering foundation upon which the proposed explainability framework is developed and validated.

Chapter 5

Navigating XAI in Condition Monitoring: Case Studies on Tool Selection and Application

The essence of this research lies in the application of XAI techniques within the engineering industry, with a particular emphasis on condition monitoring for machinery and critical assets. The central challenge addressed in this work is to generate domain-relevant, understandable explanations from ML models that can be readily interpreted by experts and effectively support the decision-making process.

Most engineering applications mandate a higher degree of transparency and accountability in decision-making processes [282]. The success and deployment of ML models in these contexts hinge on their ability to provide human-understandable explanations for stakeholders who may not possess expertise in ML. For centrifugal pump operators, understanding not just the predictions but also the rationale behind those predictions is pivotal in instilling confidence in ML models' reliability and effectiveness.

This research aims to apply post-hoc XAI methods to provide clear explanations for the predictions made by ML models. Many post-hoc XAI tools, while valuable for dissecting the inner workings of ML models, often present explanations that prove challenging for individuals lacking ML expertise to comprehend. These explanations frequently rely on complex visualisations, mathematical notations, or abstract concepts that can act as barriers to understanding for non-experts. This knowledge gap poses a significant hurdle to the widespread adoption of AI and ML technologies, particularly in safety-critical industries like the condition monitoring domain, where stakeholders, such as operators, require straightforward and accessible explanations to make informed decisions with confidence. Thus, the imperative lies in refining XAI techniques to bridge this communication gap, rendering the insights gleaned from ML models more interpretable and actionable for a broader spectrum of users.

Before exploring how complex explanations can be simplified for end-users, it is first necessary to address the challenge of selecting an appropriate XAI tool. Numerous tools exist, each tailored to particular applications and stakeholder needs, making a systematic selection process essential. To address this, a structured set of steps was proposed and developed in our earlier work [283], specifically to guide the choice of XAI methods in different contexts. These steps consider both the characteristics of the machine learning models being used and the requirements of the intended users, particularly non-experts. By applying this structured approach, the process of simplifying model explanations is more likely to achieve its aim of delivering clarity and usability in safety-critical engineering applications.

The ensuing chapters will delve into the methodology in greater detail, outlining the specific techniques, tools, and procedures employed to realise the objectives set forth in this research.

5.1 Designing Explainable ML: Key Considerations

During the exploration of XAI tools and their application to various case studies, it becomes evident that if you intend to design an explainable system, you must consider explainability from the beginning of the design process. However, incorporating explainability into existing systems is not only feasible but also practical. Even if these systems are complex and seemingly rigid, advancements in technology and methodology have made it possible to retrofit explainability features. This enhancement can significantly increase transparency and foster trust.

Selecting the appropriate XAI approach for your application can indeed be a daunting task due to the numerous options available. Certain factors have been identified that, when considered, will aid in the design of explainable ML systems. These essential factors are discussed below and illustrated in Figure 5.1. These factors should be taken into account at each stage of the design process to ensure the successful integration of explainability. These factors are:

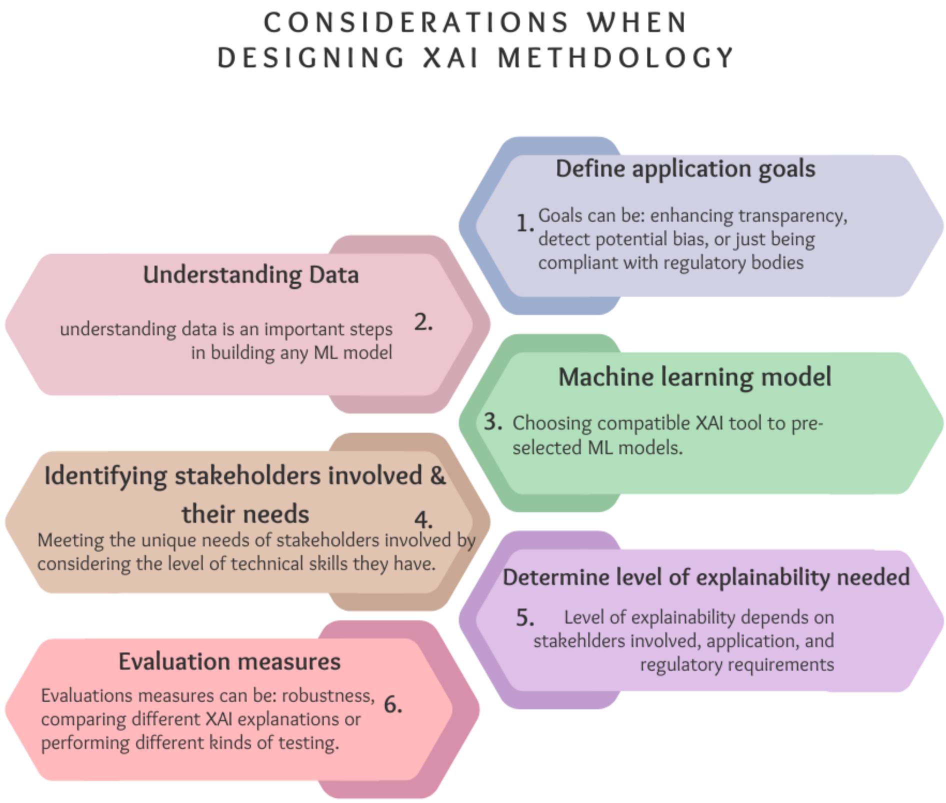


Figure 5.1: Considerations when designing XAI methodology.

5.1.1 Defining the Application Goal(s)

The integration of XAI into ML systems is a nuanced process that must be tailored to meet specific application goals. These goals typically include enhancing the transparency of ML models, ensuring compliance with regulatory standards, mitigating biases in decision-making, and making the workings of complex models understandable to non-expert users. Transparency is vital for sectors requiring high trust, while regulatory compliance ensures accountability and adherence to legal standards like the EU's

GDPR. Addressing biases is crucial for building ethical AI systems, and making models understandable to lay users enhances trust and usability.

Overall, the deliberate and goal-oriented application of XAI is about creating ML systems that are transparent, accountable, fair, and comprehensible, aligning with both ethical standards and business objectives, and ensuring that XAI strategies are perfectly in harmony with the intended outcomes of the ML application.

5.1.2 Understanding Data

Understanding the data is a pivotal aspect of designing an effective ML methodology, as the characteristics of the data fundamentally shape the decision-making process of the model. The way data is collected and categorised, whether as "hard" or "soft" data, also influences how the model interprets and responds to input. "Soft" data is derived from subjective sources, such as human experiences, impressions, and perceptions, whereas "hard" data consists of objective, quantifiable measurements [284]. This distinction is important when selecting XAI tools, as the explainability approach may vary depending on whether the data is objective or subjective in nature.

Additionally, as discussed in chapter 3, different XAI tools are designed to handle specific types of data. Recognising the data type—whether structured, such as tabular data, or unstructured, like text or images—is critical in choosing the most suitable XAI tool. For instance, tools like LIME and SHAP are effective for structured data, providing feature importance explanations that align well with tabular formats. In contrast, unstructured data, such as images, may require saliency maps or gradient-based methods like GRAD-CAM, which highlight regions of interest within an image to elucidate the model's focus [125].

Selecting the appropriate XAI tools based on data type is essential to ensure that the explanations generated are valid, meaningful, and aligned with the model's learning patterns. Furthermore, an understanding of the data also aids in identifying and mitigating potential biases within the ML model, contributing to a more robust and ethical XAI framework.

5.1.3 ML Model Selection

The compatibility of an XAI tool with the underlying ML model is a critical factor in ensuring the validity and relevance of the explanations it generates, as discussed in Chapter 2. This compatibility is deeply influenced by whether an XAI tool is model-specific or model-agnostic. Model-specific tools are tailored to provide explanations for a particular type of model, leveraging its internal mechanisms and structures, thus offering in-depth insights. In contrast, model-agnostic tools offer the flexibility to be used across various types of ML models, focusing on the input-output relationship rather than the internal workings of the model.

Choosing the appropriate XAI tool for the ML model involves a careful evaluation of the nature of the ML model, the type of explanations required, and the context in which the model operates. For instance, while model-specific tools might provide more granular insights for a particular model type, model-agnostic tools are invaluable for scenarios requiring a broader application across different model architectures.

5.1.4 Meeting the Unique Needs of Stakeholders Involved

Meeting the unique needs of stakeholders involved by considering the level of technical skills they have in the design process will help build an effective XAI methodology that is both explainable and trustworthy for all users involved. To attain explainability in AI systems, it is crucial to include human-centred design principles in the development process. This involves designing AI systems that consider the user's requirements and preferences through user research, user testing, and iterative design processes. As previously mentioned in Chapter 2. Different XAI tools require a level of expertise in AI. What is considered easy for practitioners might not be comprehensible for regulators or end users.

Explanations generated from the XAI tool should be tailored to the user's needs [17], some examples now follow. Business stakeholders pose broader, business-oriented inquiries when seeking an explanation; that's why they frequently favour a layman's explanation, aiming to obtain information that fosters confidence in the model's expected performance in the intended context. However, regulators aim to verify that the model complies with a particular set of standards. Domain experts have advanced knowledge of the environment in which the model operates, and usually, they have no ML background. They seek explanations as a tool to help in the decision-making process [92].

5.1.5 Determine the Level of Explainability Needed for the Application.

The degree of explainability an AI system requires is intricately linked to its specific use case, the stakeholders involved, and any existing regulatory mandates. As a foundational step in the journey towards crafting an explainable ML model, it is often advisable to start with simple and transparent models. These models offer a clear view into their decision-making processes, making them inherently more understandable. If these straightforward models satisfy the performance expectations of end-users, they are often preferred due to their inherent transparency and ease of explanation. The simplicity of such models generally aligns well with the goal of maintaining high levels of explainability, ensuring that stakeholders can easily comprehend how decisions are made [52].

However, the dynamics shift when the performance of these simple models falls short of the requirements. In scenarios where enhanced performance is paramount, more complex models may be necessary. These sophisticated models, while potentially offering improved accuracy or efficiency, often come at the cost of reduced transparency. It's in these situations that the appropriate XAI techniques become invaluable [20]. Selecting and applying the right XAI technique allows for the demystification of complex models, ensuring that their decisions can be interpreted and trusted by users. This process involves not just revealing how these models arrive at their conclusions, but also ensuring that the explanations are accessible and meaningful to all stakeholders involved, regardless of their technical expertise.

5.1.6 Determine the Evaluation Measures to Validate the XAI Explanations Generated

Assessing the quality of explanations provided by XAI is pivotal for bolstering stakeholder trust in ML models. The evaluation process is nuanced and often tailored to the specific requirements and contexts of the end users. A range of measures can be employed to ensure the robustness and relevance of the explanations generated by XAI tools.

Firstly, the robustness of explanations is a crucial measure of quality in XAI. Robustness refers to the consistency of explanations when faced with minor variations in input data or changes in model parameters. For instance, if a model slightly alters its prediction due to minor fluctuations in the data (such as slight noise or minor shifts in feature values), a robust explanation should still identify the same key features or reasoning paths with minimal variation. This consistency is essential because it signals that the XAI tool is reliably capturing the underlying logic of the model, rather than reacting to random fluctuations or noise in the input data [285].

Also, comparing explanations from various XAI tools can provide a comprehensive understanding of the model's decision-making process. By analysing the similarities and differences in the explanations provided by different tools, stakeholders can gain a multi-faceted view of the model's logic, further enhancing transparency and trust.

Involving domain experts in the evaluation process is another effective strategy. Experts in the specific field can provide valuable insights by assessing the plausibility and relevance of the explanations in the context of domain knowledge. Their feedback can highlight areas where the model's reasoning aligns or diverges from human expertise, guiding further refinement of the model or the explanations it generates.

When direct access to domain experts is limited, established repositories such as industry standards, guidelines, and historical case data can serve as independent validation mechanisms, ensuring that XAI-generated explanations remain consistent with well-documented knowledge bases [20, 51]. Complementing this, testing explanations on controlled public datasets provides an opportunity to assess their accuracy and consistency in environments where the ground truth is known [30, 52]. Applying the same tools to real-world or industrial datasets further demonstrates their scalability and practical relevance, reinforcing the robustness and trustworthiness of XAI explanations across diverse contexts [17].

Taken together, these measures offer a broad toolkit for evaluating the quality of XAI explanations. Depending on the context, they can be selectively applied to enhance understanding of the model's decision-making process, strengthen stakeholder trust, and ensure that outputs remain interpretable, reliable, and aligned with user expectations.

5.2 Connecting the Dots: Matching End Users' Needs with XAI Tools

One of the primary objectives of explainability is to actively engage stakeholders, ensuring their comprehension of the generated explanations and, in turn, fostering trust in the utilisation of ML models [211]. End users, stemming from diverse domains, exhibit distinct preferences in terms of how information should be conveyed to them.

Previous research publications [286] [287] shed light on the needs and levels of expertise among users, guiding the selection of appropriate explanations. It is of paramount importance to carefully choose the most suitable explanation for a particular model, given that it may need to be comprehensible to a broad spectrum of end-users, including those without technical backgrounds, data scientists, and domain experts, each possessing unique perspectives and potential regulatory requirements. The clarity of explanations is enhanced when they are tailored to the audience.

In Chapter 3, various explainability tools are mentioned, encompassing analytical, visual, rule-based, and textual explanations. The choice of an appropriate explanation type should depend on the specific end user(s). Customising the explanation format for end-users is essential, as different formats may be better suited to different individuals. For example, non-technical domain experts may find visual presentations and narrative-driven, rule-based explanations beneficial, while text-based explanations may be more suitable for individuals with limited prior knowledge. Overloading users with excessive information can lead to confusion and a diminished level of understanding and trust. The necessity for explanations and the type thereof vary based on contextual factors such as the user's identity, role, existing knowledge, and the application at hand.

For example, maintenance engineers require explanations that pinpoint specific signals driving anomaly detection, whereas plant managers may prioritise higher-level insights related to operational efficiency and safety compliance. This highlights the need for explanations to address the distinct needs, roles, and objectives of different stakeholders within condition monitoring [26].

5.3 Foundational Key Stages: Building the Blueprint for an Innovative Framework

Based on the considerations identified in this chapter, when designing an XAI methodology and selecting the appropriate XAI tools, several key stages have been identified as shown in Figure 5.2. These stages serve as the foundation for a more innovative framework that will be introduced in the following chapters. The application of these stages to two different case studies, which will be discussed in the next chapter, demonstrates their practical relevance and adaptability to various real-world scenarios. Analysing how these stages were implemented in different contexts can provide valuable insights into the effectiveness and versatility of the XAI methodology and tools. These stages likely form a structured approach to conducting case studies in the context of XAI:

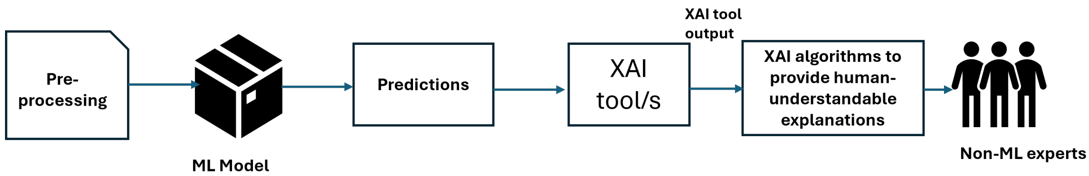


Figure 5.2: Foundational key stages

- Data pre-processing: The First stage revolves around the choice of an appropriate ML prediction model. Many ML models are often regarded as black-box models, making it challenging to comprehend their inner workings and how they translate inputs into outputs. Consequently, the development and deployment of XAI techniques are necessary to elucidate the rationale behind the model's predictions, enabling stakeholders in the industry to gain insights into the functioning of these ML models.

- Prediction models: The second stage revolves around the choice of an appropriate ML prediction model. Many ML models are often regarded as black-box models, making it challenging to comprehend their inner workings and how they translate inputs into outputs. Consequently, the development and deployment of XAI techniques are necessary to elucidate the rationale behind the model's predictions, enabling stakeholders in the industry to gain insights into the functioning of these ML models.

- Applying XAI tools: In this stage, an additional step is introduced wherein XAI tools are applied to provide insights into the functioning of machine learning ML models and the rationale behind their predictions. XAI tools, such as SHAP, typically offer explanations that are either local, global, or a combination of both. However, these explanations are often presented in complex formats—such as intricate plots—that may present challenges for individuals who lack expertise in ML. This complexity highlights the necessity of subsequent stages that aim to clarify and reformat these explanations to make them more accessible and comprehensible for a broader audience.

  The selection of an appropriate XAI tool should be guided by several factors. Primarily, the type of data involved—whether tabular, image, text, or video—determines which XAI tool is most suitable. Additionally, the format preferred by stakeholders, such as visual, textual, or alternative presentations, plays a crucial role in determining the tool and explanation format. Furthermore, the choice of ML model impacts the selection process, as some models require specific (model-dependent) XAI tools, while others are more compatible with model-agnostic tools. Finally, the requirements and preferences of stakeholders should be considered to ensure that explanations are effectively communicated and tailored to their needs.

- Generating human-understandable explanations: Adding this stage emphasises the significance of making explanations more accessible to a wider audience, including domain experts and end-users who may not have a background in ML. It underscores the recognition that effective communication of explanations is a critical aspect of XAI. During this stage, an additional processing step becomes imperative to transform XAI explanations into user-friendly, human-understandable explanations. This processing step is essential to bridge the gap between complex XAI outputs and the needs of non-experts, ensuring that the insights generated are presented in a format that is accessible and comprehensible to a broader audience. The algorithms used to transform XAI plots generated from the SHAP tool into text-based explanations will be revisited in more detail in Chapter 6 when the developed framework is introduced.

5.4 Case Studies

The first case study adopts an open-access dataset, chosen for its transparency and reproducibility. This initial step aims to facilitate the smooth integration of XAI tools into the realm of condition monitoring. By using open-access data, the groundwork is laid for subsequent investigations, bridging the gap between theory and practical application.

As progress is made through the following case studies, the goal is to refine the approach and select the most effective XAI tools that will aid in understanding and improving condition monitoring processes. As the case studies themselves increase in complexity, valuable insights and knowledge are aimed to be provided for practitioners looking to use XAI in their condition monitoring applications.

5.5 Case study 1: Combined Cycle Gas Turbine

In this case study, a publicly accessible dataset encompassing operational measurements derived from a Combined Cycle Gas Turbine (CCGT) generator was employed [288], as depicted in Figure 5.3 [3], illustrating the fundamental principles of a combined power cycle. The choice of utilising an open-source dataset for our initial case study serves a twofold purpose. Firstly, it streamlines the application of XAI tools, rendering our work accessible and reproducible for subsequent case studies. Secondly, the selected CCGT dataset spans a duration of six years, from 2006 to 2011, and has previously served as a foundation for demonstrating ML models aimed at predicting power output based on environmental factors, as evidenced in prior research by [289] and [290].

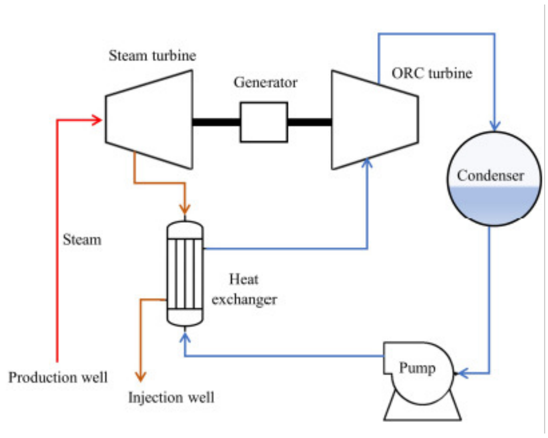


Figure 5.3: Basics of combined power cycle gas turbine [3]

This dataset comprises an array of operational measurements originating from various components, including the turbine, generator, and control valves. These measurements encapsulate vital information pertaining to the CCGT generator's performance and operational dynamics, forming the basis for our exploration into the realm of XAI and its application in enhancing the interpretability of ML models for this specific domain. The dataset comprises four critical operational parameters:

- Ambient Pressure (AP).
- Exhaust Vacuum (V).
- Ambient Temperature (AT).
- Relative Humidity (RH).

These parameters collectively serve as input variables in predicting the net hourly electrical energy output (PE) of the power plant. In the course of this case study, the discernible and explainable relationship between environmental conditions, encapsulated by these parameters, and the resultant power output has been successfully demonstrated. This explanation is a crucial part of our study, illuminating the complex relationship between environmental factors and the plant's energy generation process.

5.6 Pre-processing Stage

This stage encompasses an essential process of understanding the dataset, a crucial step prior to utilising the data in prediction models. Initially, the data set undergoes scrutiny to identify null values and potential duplicates. Fortunately, in this case, no missing values or duplicates were detected, rendering a data set with 9568 non-null entries. The data distribution is represented in Table 5.1.

	Temp	**Vacuum**	**Pressure**	**Humidity**	**Power**
Count	9568	9568	9568	9568	9568
Mean	19.65123	54.3058	1013.259	73.30898	454.365
Std	7.452473	12.70789	5.938784	14.60027	17.06699
Min	1.81	25.36	992.89	25.56	420.26
25%	13.51	41.74	1009.1	63.3275	439.75
50%	20.345	52.08	1012.94	74.975	451.55
75%	25.72	66.54	1017.26	84.83	468.43
Max	37.11	81.56	1033.3	100.16	495.76

Table 5.1: Summary Statistics for Parameters

To delve deeper into the dataset's characteristics and its correlation with the target variable, a correlation matrix was constructed. Figure 5.4 illustrates the features with the most significant correlations to the output power (PE). Notably, the temperature and vacuum exhibit a substantial negative correlation with the output power, while the pressure and humidity demonstrate a comparatively weaker positive correlation.

Furthermore, bivariate distribution visualisations have been employed, including pair plots, to explore the interrelationships among the dataset's features, as depicted in Figure 5.5.

Based on this EDA, it has been observed that an increase in temperature is associated with a reduction in output power. This negative relationship has been substantiated by the correlation matrix between temperature and output power. Thus, the importance of temperature as a predictor variable negatively impacting power generation has been highlighted.

Similarly, a parallel negative association has been identified between vacuum levels and power output. An increase in vacuum levels is accompanied by a decrease in power output. This negative relationship further supports the critical role of the vacuum in the power generation process.

Conversely, both pressure and humidity have been found to be positively correlated with power output. An increase in either variable corresponds to an increase in power generation. These findings underscore the significance of pressure and humidity as factors contributing positively to power generation.

These insights are instrumental in the comprehension of our dataset's behaviour, shedding light on the intricate interplay of variables influencing power generation. They also establish a strong foundation for informed feature selection and subsequent model development in the subsequent stages of our analysis. These findings hold significant implications for the research, facilitating the design of more effective models and enabling informed decision-making in our pursuit of advancing our understanding of this domain

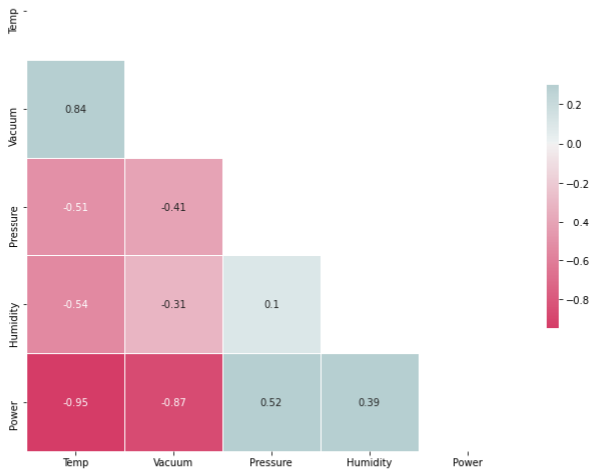


Figure 5.4: Correlation matrix for CCGT dataset

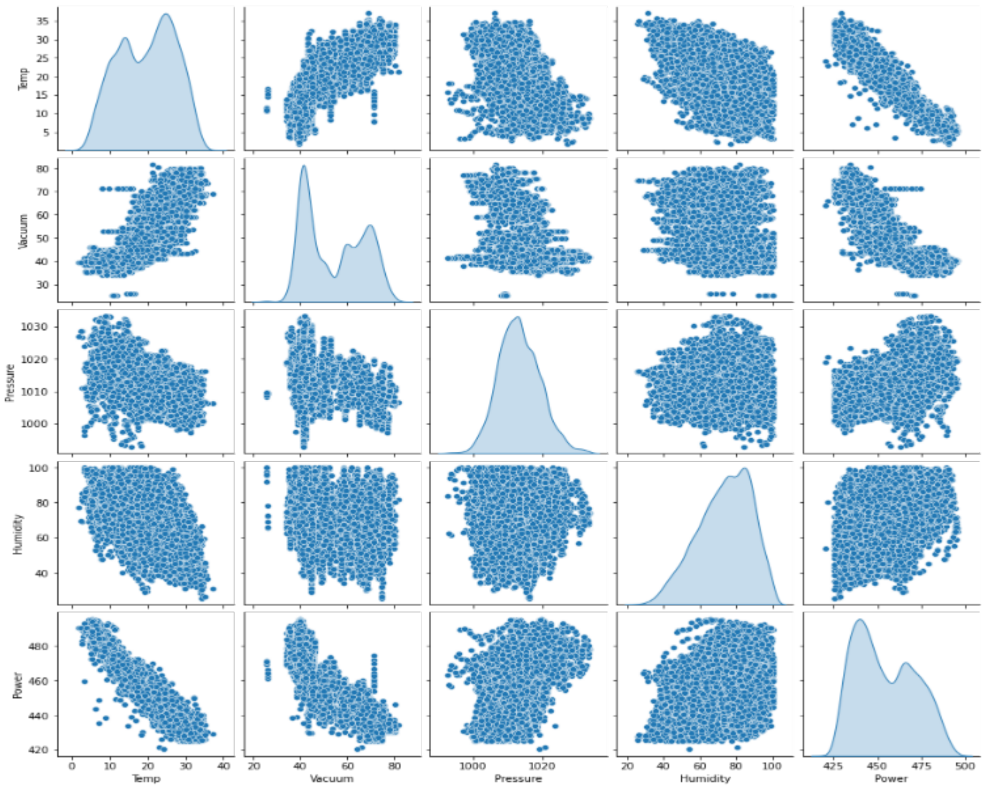


Figure 5.5: visualisation of the bivariate distribution of the dataset, showing how each feature correlates with each other and with the target variable (Power).

5.6.1 Shortcomings of Pre-Training Correlation Analysis

Pre-training correlation analysis highlights linear dependencies between features but falls short of capturing:

- Non-linear interactions between variables.
- Instance-specific contributions to predictions.
- Contextual interpretations relevant to individual cases or subsets of data.

This limitation is particularly evident in systems like CCGT, where interactions among operational parameters are often non-linear and dynamic.

5.7 Choosing Appropriate ML model Stage

Seven distinct regression models were employed to forecast output power, with their performances assessed using the Root Mean Square Error (RMSE) metric—a measure of the average prediction error by the models calculated as shown in Equation 5.1.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{5.1}$$

where:

- y_i represents the actual observed value,
- $\hat{y}_i$ is the predicted value generated by the model,
- n denotes the total number of observations.

These models encompassed linear regression, random forest, XGboost, GBoost, ruleFit, Decision trees, and KNN regression. In this case study, a range of machine learning models was chosen to capture diverse relationships within the data, particularly suited for regression tasks:

- **Linear Regression**: Serves as a baseline model to assess linear relationships within the data.

- **Random Forest and Decision Trees**: These models handle non-linear relationships effectively and include built-in feature interaction handling, making them both robust and interpretable.
- **Gradient Boosting Models (GBoost and XGBoost)**: Known for iteratively optimising performance, these models handle complex, high-dimensional data effectively.
- **RuleFit**: Combines decision rules with linear terms, balancing interpretability with flexibility.
- **K-Nearest neighbours (KNN)**: A non-parametric approach that captures local data patterns by considering proximity in feature space.

The following are the key hyperparameters chosen for each machine learning model used in this study. These hyperparameters have been selected based on common practices and prior research to balance model performance, prevent overfitting, and ensure computational efficiency:

- **Random Forest Regressor**:
  - `n_estimators`: set to 100 (Default value), the number of trees in the forest, typically chosen between 100 and 500 for balance between accuracy and computational efficiency.
  - `max_depth`: set to 10 Limited to prevent overfitting, typical values range from 10 to 30 [41].
  - `random_state`: Set to 42 for reproducibility.
- **Gradient Boosting Regressor**:
  - `learning_rate`: 0.1 (Default value).
  - `n_estimators`: 100 (Default value), the number of boosting stages, generally between 100 and 300 [156].
  - `random_state`: Set to 42 for reproducibility.

- **XGBoost Regressor**:
  - `learning_rate`: 0.3 (Default value), controlling the contribution of each tree, often set between 0.01 and 0.1 for optimal results.
  - `max_depth`: 6 with values from 3 to 10 as standard [291].
  - `random_state`: Set to 42 for reproducibility.
- **Decision Tree Regressor**:
  - `max_depth`: Default value (None), meaning nodes expand until all leaves contain fewer than `min_samples_split` samples. Often set between 5 and 15 to prevent overfitting.
  - `min_samples_split`: Controls node splitting, with values like 2 or 5 for balanced splits.
  - `random_state`: Set to 42 for reproducibility.
- **K-Nearest neighbours Regressor**:
  - `n_neighbours`: Set to 5, determining the number of nearest neighbours used for prediction. Optimal values generally range from 5 to 20, selected based on cross-validation.
  - `weights`: Default value `uniform`, meaning each neighbor is equally weighted [292].

The study commenced by inputting individual features, such as temperature, pressure, and humidity, into these models, gauging their performance independently as shown in Table 5.2. Notably, temperature emerged as the most influential feature, with the lowest RMSE of 5.16, indicating its significance in predicting output power. Subsequently, the investigation extended to combinations of two parameters as shown in Table 5.3, wherein sets including temperature consistently outperformed those with pressure or humidity. The most effective pairings were temperature with vacuum and temperature with humidity.

Model	AT	V	P	RH
Linear Regression	5.45	8.58	14.91	15.86
Random Forest	5.16	6.94	14.69	15.91
GBOOST	5.16	6.63	14.83	15.97
XGBOOST	5.30	5.44	15.19	16.50
RuleFit	5.14	6.10	14.63	15.87
Decision Trees	6.17	5.39	16.99	20.64
KNN Regression	5.24	5.32	14.43	15.69

Table 5.2: RMSE Values for Regression Models

Features	AT-V	AT-AP	AT-RH	V-AP	V-RH	AP-RH
Linear regression	5.00	5.43	4.98	7.87	8.12	13.13
Random forest	4.12	4.92	4.51	6.19	6.78	12.79
GBoost	4.08	4.82	4.71	6.00	6.32	12.75
XGBOOST	3.47	4.96	4.89	4.81	4.83	13.19
RuleFit	4.07	4.82	4.70	5.84	6.00	12.88
Decision Trees	4.88	6.73	6.49	5.34	6.37	17.59
KNN regression	4.22	4.71	4.46	6.23	7.13	12.99

Table 5.3: Regression Model Performance for using two Features

Finally, the analysis delved into three-parameter combinations as shown in Table 5.3, and the optimal results were achieved by employing temperature, vacuum, and pressure together. This outcome signifies the combined influence of these three factors in predicting output power.

5.8 Using XAI Tools

The principal objective of this study revolves around the application of XAI techniques to predict output power using the CCGT dataset, and subsequently, to render the resulting explanations in a manner that is easily understandable for human audiences. In pursuit of this goal, an exploratory journey through various XAI tools was embarked upon, aiming to assess their efficacy and applicability to our specific problem.

5.8.1 Capabilities of XAI Tools

XAI tools provide robust frameworks for understanding model behaviour, particularly in the context of tabular datasets like those used in the CCGT case study. Tabular

Features	AT-V-AP-RH	AT-V-AP	AT-V-RH	AT-AP-RH	V-AP-RH
Linear regression	4.43	4.82	4.72	4.64	7.52
Random forest	3.98	3.86	4.015	4.39	6.064
GBoost	3.55	3.74	3.94	4.29	5.73
XGBOOST	3.07	3.07	3.50	4.36	4.46
RuleFit	3.68	3.74	4.07	4.34	5.68
Decision Trees	4.54	4.67	4.84	5.81	5.90
KNN regression	3.88	3.80	4.25	4.50	6.33

Table 5.4: Regression Model Performance for Different Feature Combinations

data, characterised by rows representing individual instances and columns representing features, aligns well with the capabilities of certain XAI tools designed for structured data. The tools explored in this study include:

- **SHAP[56]:** SHAP offers both local and global explanations by computing Shapley values, derived from cooperative game theory. SHAP excels in identifying:
  - Feature importance rankings that are consistent across different model instances.
  - Interactions between features, providing a nuanced understanding of how multiple variables jointly influence predictions.
  - Instance-specific contributions, allowing users to understand the reasoning behind individual predictions.

  SHAP's suitability for tabular data stems from its ability to handle structured features effectively, producing interpretable outputs for domain experts.

- **LIME [57]:** generates instance-level explanations by approximating the original complex model with a simple surrogate model within a local neighbourhood of the instance being explained. LIME is particularly effective in scenarios requiring:
  - Quick, localised insights into a model's decision-making process.

- Understanding specific predictions where global explanations might not suffice.

- **Permutation Importance [51]:** This method assesses feature relevance by evaluating the change in model performance after shuffling the values of a specific feature. Its primary strengths include:

  - Quantifying the global impact of each feature on model predictions.
  - Providing an intuitive metric for stakeholders less familiar with complex explanation methods.

  Permutation Importance is highly compatible with tabular data, as the structured format allows for straightforward permutation of individual features.

- **PDP [156]:** PDP visualises the relationship between a feature and the predicted outcome while marginalising other features. This tool:

  - Provides global insights into how a specific feature influences predictions across the dataset.
  - Allows domain experts to interpret non-linear relationships between features and the target variable.

  PDP is particularly valuable in tabular datasets, as its visual representation facilitates intuitive understanding for users.

These tools, tailored for tabular data, offer a comprehensive suite of capabilities to enhance interpretability and usability in machine learning applications. Their ability to generate actionable insights makes them ideal for the CCGT case study, where the dataset's structured nature aligns well with their strengths.

The following section outlines the implementation of the aforementioned XAI tools, with a comparative analysis of their results to identify the most suitable option for selection.

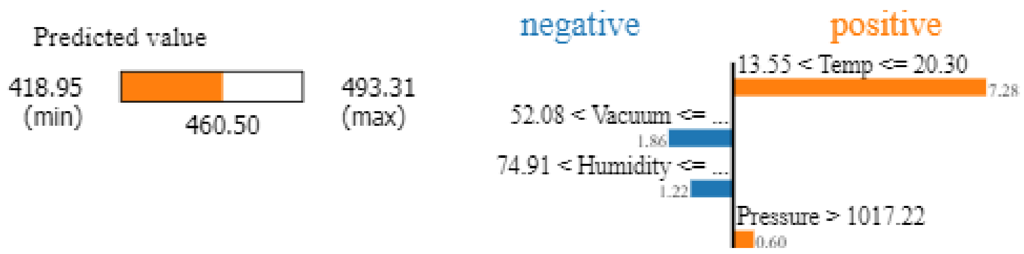
5.8.2 LIME

In this research, the utilisation of LIME [57] is being considered. LIME is employed to enable predictions made by any classifier or regressor to be explained faithfully. This is achieved through the creation of a local approximation of the model's behaviour for a specific data sample. For this analysis, the LIME Tabular Explainer was utilised in regression mode to interpret the relationships between individual input features—such as temperature, vacuum, pressure, and humidity—and the predicted power output of the CCGT dataset. This approach enables a detailed, instance-level explanation of model predictions by highlighting feature contributions in a tabular data structure, thus offering a comprehensive understanding of the model's behaviour in predicting continuous output values.

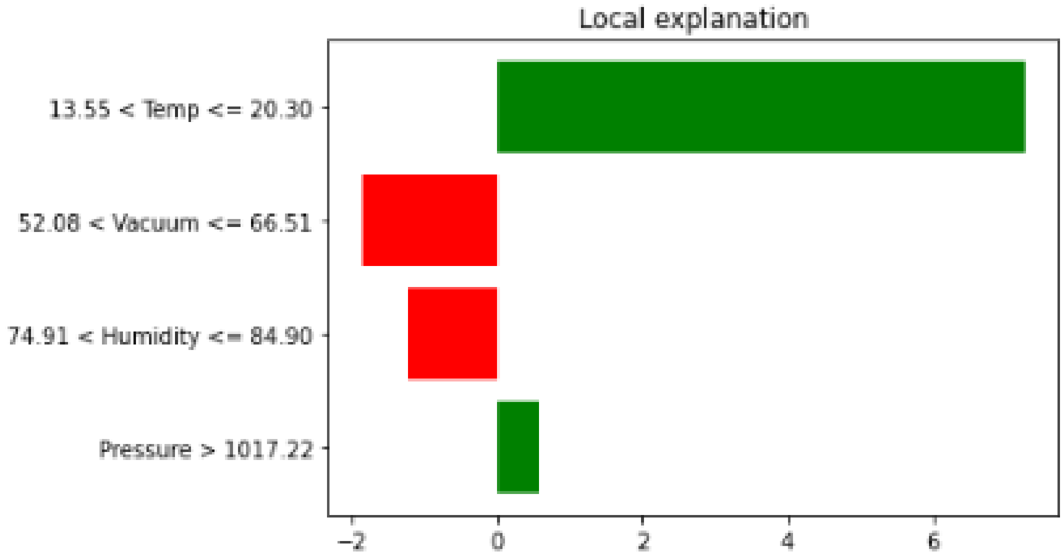
The feature values of an individual data point are systematically adjusted, and the resultant impact on the model's output is observed. In essence, LIME serves the role of an 'explainer' by illuminating the reasoning behind predictions for each individual data point. The outcome of LIME's process is a collection of explanations that succinctly portray the influence of each feature on the prediction for that specific sample, thereby providing a valuable form of local interpretability within the context of this research [57].

Figure 5.6 provides an explanatory visual for a specific instance, as outlined in Table 5.5. The interpretation is as follows: Within Figure 5.6, it can be discerned that the predicted power value stands at 460.50. Within this localised explanation, it becomes evident that both Temperature (Temp) and Pressure exert a positive influence on power, leading to an increase in the power value. Conversely, Vacuum and Humidity exhibit a negative impact on power, resulting in a decrease in the power value. This graphical representation aids in comprehending the factors influencing power prediction for the given instance.

Moving to Figure 5.6b, the most influential features are Temp, Vacuum, Humidity and lastly Pressure. LIME output presented below in Figure 5.6 can be explained through its three distinct sections, each offering insights into the local interpretation of specific predictions. It's important to emphasise that LIME provides a localised under-



(a) LIME explanation for a single instance.



(b) LIME explanation for a single instance.

Figure 5.6: LIME explanation for a single instance shown in Table 5.5

Feature	Value
Temp	15.45
Vacuum	61.27
Humidity	79.21
Pressure	1019.87

Table 5.5: Feature values for a single instance.

Feature binning	Contribution to Prediction
13.55 $<$ Temp $<=$ 20.30	7.275039
52.08 $<=$ Vacuum $\leq$ 66.51	-1.855047
74.91 $<$ Humidity $<=$ 84.90	-1.222377
Pressure $>$ 1017.22	0.596621

Table 5.6: Feature binning and corresponding contribution values for the selected instance.

standing of model predictions, encompassing both the direction (positive or negative) and magnitude (weight) of feature impact.

In the Predicted Value, Figure 5.6a displays the model's predicted median value for the output power of the particular instance shown in Table 5.5. Within this Figure, the prediction interval spans from a minimum of 418.95 to a maximum of 493.31.

Moving on to the Negative and Positive sections, insights into the features contributing to the increase or decrease of the predicted value are gained. Positive-weighted features have a positive impact, elevating the prediction value, whereas negative-weighted features have the opposite effect, diminishing it. LIME identifies these influential features for the prediction related to the instance in Table 5.5. For example, as illustrated in Figure 5.6b, the range "$13.55 < \text{Temp} <= 20.30$" is positively influencing the predicted value as shown in Table 5.6, while "$52.08 < \text{Vacuum} <= 66.51$" and "$74.91 < \text{Humidity} <= 84.90$" are exerting negative influences

Lastly, the Feature Value section provides the actual values of the top features for this instance. For instance, the Temp feature is recorded at 15.45, which falls within the range of greater than 13.55 and less than 20.30, aligning with a positive influence on the predicted value as per the model's interpretation.

It has been observed that LIME exhibits instability. This instability arises because LIME generates explanations by perturbing the input data and fitting a local surrogate model around the instance. The randomness inherent in the sampling process results in variations in the explanations provided for the same instance across repeated runs.

To evaluate the instability of LIME, an experiment was conducted by generating explanations for the same test instance 100 times. For each run, the feature rankings as determined by LIME were recorded. The variability in these rankings was assessed using Kendall's Tau, a statistical measure of rank correlation [293] [110].

Kendall's Tau (τ) is used as a measure of correlation between two ranked lists, quantifying the degree of similarity in their orderings. Given two ranking vectors $\mathbf{X} = (x_1, x_2, \ldots, x_n)$ and $\mathbf{Y} = (y_1, y_2, \ldots, y_n)$, Kendall's Tau is defined as shown in Equation 5.2:

$$\tau = \frac{C - D}{\binom{n}{2}} \tag{5.2}$$

where:

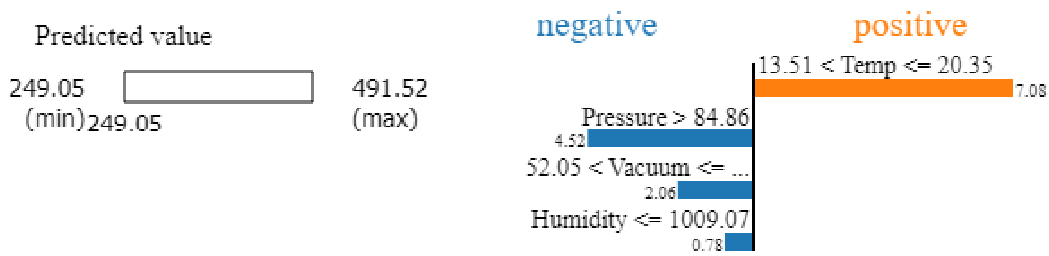
- C: number of concordant pairs, i.e., pairs (i, j) such that the ordering of x_i, x_j in $\mathbf{X}$ agrees with the ordering of y_i, y_j in $\mathbf{Y}$.
- D: number of discordant pairs, i.e., pairs (i, j) where orderings in $\mathbf{X}$ and $\mathbf{Y}$ disagree.
- $\binom{n}{2}$: total number of pairs, equal to $\frac{n(n-1)}{2}$, where n is the number of observations.

The value of τ ranges from -1 (perfectly reversed rankings) to $+1$ (perfect agreement).

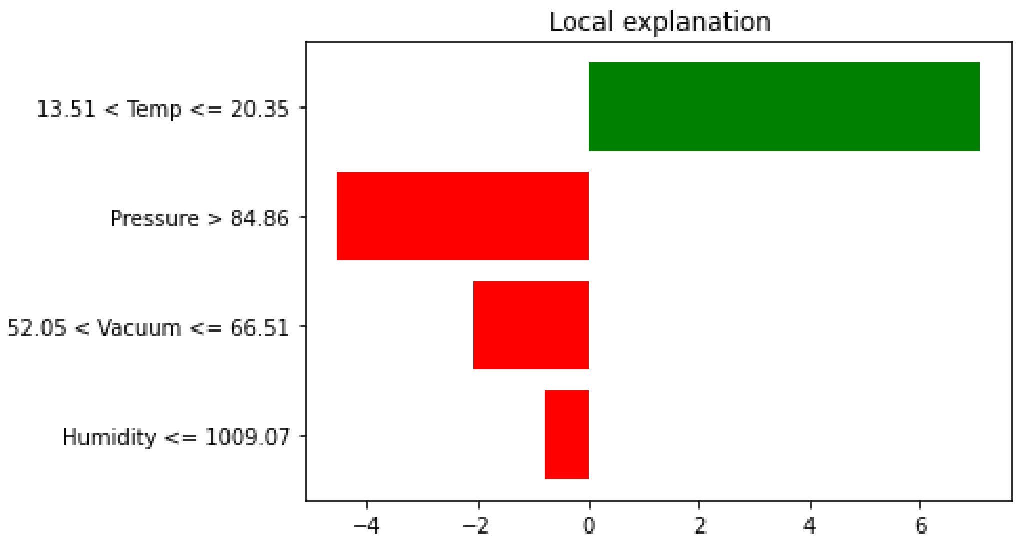
The experiment was conducted to evaluate the stability of LIME explanations. First, the LIME explainer was initialised using the training dataset to ensure consistent feature representations. A single test instance was then selected as the subject of repeated explanations. For each iteration, LIME generated feature importance rankings for the chosen test instance. These rankings, produced across 100 runs, were collected and stored in a data frame to facilitate comparison. Pairwise Kendall's Tau correlations were computed between the rankings from all runs to measure the degree of agreement among them. Finally, the average Kendall's Tau value was calculated as a quantitative indicator of the overall consistency of LIME's explanations across repeated runs. The average consistency of feature rankings across 100 runs, as measured by Kendall's Tau, was found to be:

$$\text{Average Consistency (Kendall's Tau)} = 0.67$$

This result indicates moderate agreement between rankings, confirming that LIME exhibits variability in its explanations. The inconsistency highlights the impact of randomness in LIME's perturbation-based approach, which causes different rankings of feature importance to be produced even for the same input instance. Here is an example illustrating the variability of LIME explanations. Figure 5.7 presents different explanations generated by LIME for the same instance across multiple runs, demonstrating that LIME can produce inconsistent results for identical input data.



(a) Different explanation generated by LIME for the same single instance shown in Table 5.5.

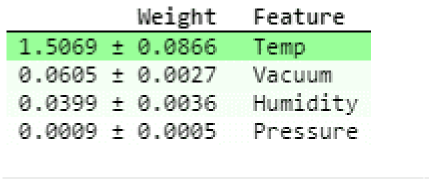


(b) LIME explanation for a single instance.

Figure 5.7: Different explanation generated by LIME for the same single instance shown in Table 5.5

5.8.3 Permutation Importance

Permutation feature importance is a method employed to assess the significance of individual features within an ML model. It operates by systematically permuting or shuffling the values of a specific feature while observing the impact on the model's performance metrics, such as accuracy or error rate. The rationale behind this approach lies in determining the extent to which altering a feature's values disrupts the model's predictive ability. If shuffling a feature leads to a notable degradation in model perfor-



Weight	Feature
1.5069 ± 0.0866	Temp
0.0605 ± 0.0027	Vacuum
0.0399 ± 0.0036	Humidity
0.0009 ± 0.0005	Pressure

Figure 5.8: Permutation importance scores

mance, it indicates that the feature plays a vital role in making accurate predictions. Conversely, if the model remains largely unaffected by the permutation, it suggests that the feature has lesser importance in the context of the model's decision-making process. In essence, permutation feature importance provides a quantitative measure of a feature's contribution to the model's overall performance [51].

Figure 5.8 shows the output. The initial number in each row, appearing before the plus and minus signs, quantifies the extent to which the model's performance deteriorated due to the random shuffling of a particular feature. It's important to emphasise that you have the flexibility to determine how many times a feature undergoes random shuffling as part of the process. Increasing the number of shuffles can enhance result stability, but it does come with the trade-off of increased computational demands. The value following the plus and minus sign serves to capture the randomness introduced through multiple shuffles. Essentially, this number quantifies the level of variability in the permutation importance calculation by conducting the process repeatedly with numerous shuffles. In simpler terms, the value after the $\pm$ sign measures how performance fluctuated from one reshuffling iteration to the next.

Permutation Feature Importance (PIMP) relies on the accuracy of the underlying model, making it less dependable when true outcome data is unavailable. In situations where the original model's predictions are not precise, there is a risk of incorrectly identifying certain features as important. Additionally, achieving feature independence can be challenging in real-world scenarios, as correlated features may dilute the importance assigned to each one. It's crucial to strike a balance in the number of permutations performed during feature importance computation, as too few can yield misleading results, while an excessive number may entail extensive processing times. In summary, while

PIMP serves as a valuable tool for model interpretability and feature importance assessment, it should be used judiciously, especially in cases where comprehensive knowledge of the feature set is lacking.

5.8.4 Partial Dependence

The partial dependence plot, often abbreviated as PDP or PD plot, provides a valuable visual representation that elucidates the isolated impact of one or two specific features on the predicted outcome generated by a ML model. This plot serves as a means to discern and understand the individual or joint influence of selected features on the model's predictions. By examining a PDP, you can gain insights into how variations in these chosen features correspond to changes in the predicted outcomes, offering a more comprehensive understanding of the model's behaviour and the significance of specific input variables [51]. Figure 5.9 shows a valuable representation of the impacts of each feature and also how two features like Temp, Vacuum and Temp and Humidity affect the output as shown in the rightmost bottom part of the figure. This plot provides valuable insights into the interplay between Temperature (Temp) and Humidity, shedding light on their combined impact on our predictions. Notably, it becomes evident that Humidity and Vacuum exert a comparatively weaker influence, as indicated by the relatively stable colour patterns along the y-axis and x-axis, respectively. In contrast, Temperature (Temp) emerges as a more dominant factor, as observed by the clear positive influence it exerts on the predicted power. As a consistent positive trend is noted in the predictions while ascending along the y-axis and moving towards the right on the x-axis in the rightmost plots, the substantial effect of Temp on the power output is underscored.

5.8.5 SHAP

In this section, SHAP is utilised to elucidate the decision-making process of our model. SHAP offers a powerful framework for interpreting model predictions by attributing each feature's contribution to the overall output.

Among the diverse XAI tools considered, SHAP emerged as a prominent choice. SHAP,

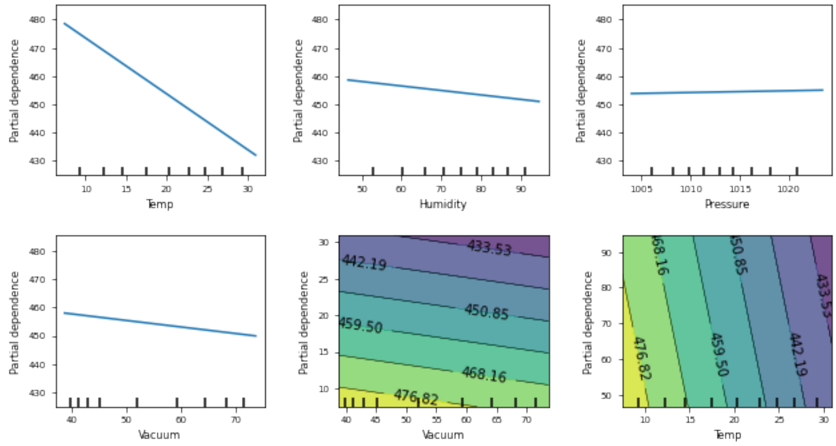


Figure 5.9: Partial dependence plots showing a valuable representation of the impacts of each feature and also how two features like Temp, Vacuum and Temp and Humidity affect the output as shown in the rightmost bottom part of the figure.

renowned in the field of post-hoc XAI, stood out for its ability to provide both local and global interpretability. It achieves this by generating explanations in the form of visualisations. However, it is essential to acknowledge that these visualisations, while rich in information, may at times appear intricate and less immediately intuitive, as elaborated earlier in this thesis.

In Figure 5.10, a local explanation is depicted for a single prediction, utilising the feature values displayed in Table 5.7 and employing a Random Forest model. The local interpretability plot presented in Figure 5.10 highlights the influence of various features on the prediction of output power. Notably, features like Humidity are shown in red, indicating their contribution to an increase in the output power prediction, while other factors such as Temperature, Pressure, and Vacuum are displayed in blue, signifying their role in decreasing the output power prediction. The visual size of the arrows in the plot corresponds to the magnitude of each feature's impact. According to this local explanation, the most influential features, in descending order of importance, are Temperature, Vacuum, Pressure and Humidity. Figure 5.11 provides an illustration of the SHAP local interpretability plot, demonstrating the impact of each feature on the output power prediction.

SHAP also possesses the capability of providing a global explanation, which is illus-

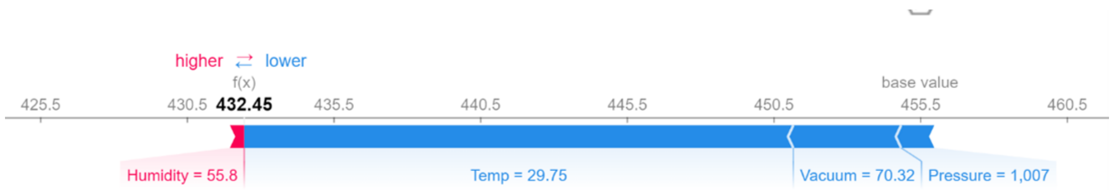


Figure 5.10: SHAP local interpretability, also known as SHAP force plot for a single instance, showing the most influential features Temp, Vacuum, Pressure and Humidity, respectively, for this instance.

trated in Figure 5.12 as the SHAP summary plot. This visualisation, often referred to as the SHAP summary plot, elucidates the most influential features contributing to the model's predictions. In the SHAP summary plot, each feature vector in the test dataset is represented as a coloured dot, with red indicating high values and blue indicating low values. This plot effectively communicates the impact of each feature on the predictive model.

As depicted in Figure 5.12, the most significant features for this model, in descending order of importance, are Temperature, Vacuum, Humidity, and Pressure. Additionally, SHAP summary plots reveal the correlation between each feature and the output power. For instance, elevated values of Temperature (indicated in red) exert a substantial negative impact on the output power, as reflected in the negative SHAP values along the X-axis. Similarly, higher Vacuum values negatively affect the output, while increased Pressure values positively influence the output power.

Features	**Values**
Temp	29.75
Vacuum	70.32
Pressure	1007.35
Humidity	55.80

Table 5.7: Feature vector chosen and explained by SHAP local interpretability plot.

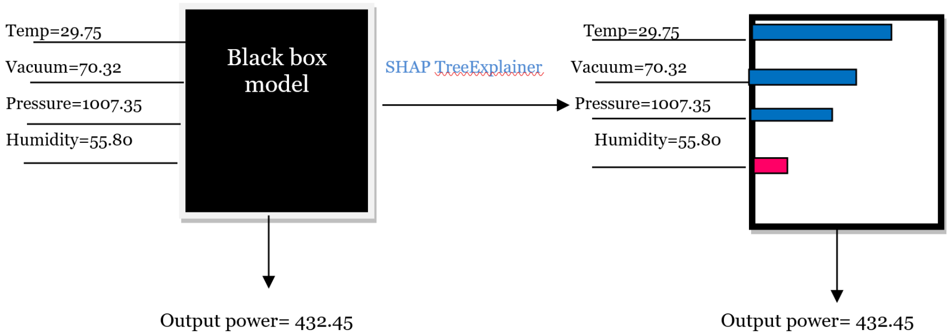


Figure 5.11: Impact of each feature value on the output.

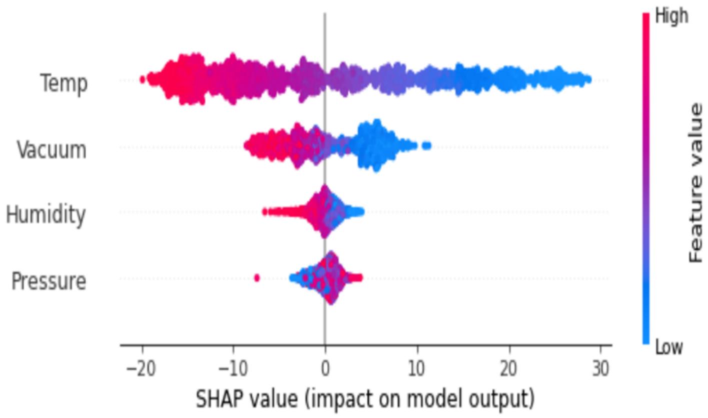


Figure 5.12: SHAP global interpretability plot, also known as SHAP summary plot, showing the most important feature on a global level,l which are Temp, Vacuum, Humidity and Pressure.

Justification for Choosing SHAP as the XAI Tool

A similar test was conducted using SHAP to evaluate its stability in generating explanations. As with LIME, SHAP was applied to the same single test instance across 100 runs. For each run, SHAP produced feature importance rankings, which were stored in a DataFrame for analysis.

Pairwise Kendall's Tau correlations were calculated between all rankings generated by SHAP across the 100 iterations. Unlike LIME, the results revealed a perfect agreement among the rankings, with an average Kendall's Tau value of 1.0. This indicates that SHAP explanations are stable, consistently providing identical feature rankings for the same instance across repeated runs. SHAP was ultimately selected based on several distinguishing criteria:

- **Data Type Compatibility and Variations [294]:** SHAP includes multiple variations tailored to different data types and models. For instance, Kernel SHAP is model-agnostic and suitable for general applications, Tree SHAP is optimised for tree-based models, and Deep SHAP works with deep learning models. This versatility across tabular, image, text, and time series data makes SHAP adaptable to diverse applications, demonstrating flexibility to evolve with application requirements.

- **Output Format:** Visual explanations were required to communicate complex model outputs intuitively. SHAP provides highly informative visualisations, capturing intricate relationships between features and predictions, which enhance stakeholder comprehension [294].

- **Local and Global Explanations:** SHAP's capability to provide both local and global explanations supports a comprehensive understanding of individual predictions while offering an overarching view of feature importance across the model. This dual functionality was less effective in other tools reviewed.

- **Theoretical Foundation:** Unlike some alternative tools, SHAP is grounded in solid theoretical principles based on cooperative game theory, offering a mathematically sound framework for feature attribution. This theoretical robustness instils confidence in SHAP's reliability [56].

- **Stability and Consistency:** In contrast to LIME, which occasionally produced inconsistent explanations for identical instances, SHAP consistently provided stable explanations, reinforcing its suitability for critical stakeholder applications [295] [296].

- **Post Hoc Flexibility:** As a post hoc tool, SHAP provides flexibility in model choice by offering explanations that are model-agnostic. This characteristic allowed for freedom in selecting the most suitable machine learning models for the data without constraining the interpretability framework.

- **Ease of Integration [56]:** SHAP's support for major machine learning libraries (such as Scikit-Learn, TensorFlow) facilitates straightforward implementation, which is valuable in environments where efficiency and integration into existing workflows are essential.

- **Versatility in Model Complexity [51]:** SHAP's applicability to both linear and non-linear models adds versatility, accommodating a range of model complexities. This flexibility allows SHAP to adapt if the analysis evolves to require different model types.

Based on these comparisons and criteria, SHAP emerged as the preferred XAI tool, delivering rich, stable, and theoretically grounded explanations suitable for the demands of this analysis.

5.9 Generating Human-understandable Explanations

Effectively communicating explanations to stakeholders in the application domain who lack expertise in ML is of paramount importance. Even for experienced engineers, comprehending SHAP plots can be a non-trivial task. Consequently, there arises a necessity to translate these intricate visualisations into an easily understandable format, thereby bridging the knowledge gap between ML experts and stakeholders. A suite of algorithms has been developed to fulfil the role of converting these visual plots into text-based explanations that are accessible to a wider audience.

In Figure 5.13, an exemplar of the text-based explanations generated is presented to convey both local and global insights derived from SHAP. Specifically, the text-based local explanations correspond to the SHAP force plot illustrated in Figure 5.10.

```
Features affecting the output power in order are

Temp
Vacuum
Pressure
Humidity

Impact of each feature value on output

Temp 29.75 High value contributes to a lower output power
Vacuum 70.32 High value contributes to a lower output power
Pressure 1007.35 Low value contributes to a lower output power
Humidity 55.8 Low value contributes to a higher output power
```

Figure 5.13: Text-based explanation corresponding to local explanation shown in Figure 5.10.

Moving forward to Figure 5.14, it depicts the most crucial global features influencing the prediction model.

In Figure 5.15 statistics for each feature are provided, encompassing the following aspects: the count of feature values that exert no influence on the output power; the count of values categorized as high, and their corresponding positive impact on the output, thereby increasing output power; and the count of values categorized as low, along with their associated positive or negative impact on the output power, affecting the prediction accordingly.

```
Most influential features on output power in order are
  Temp
  Vacuum
  Humidity
  Pressure
```

Figure 5.14: Most influential features according to SHAP summary plot.

```
Global interpretability for  Temp
No of points that have no impact on output= % 0.0
 % of points that have low values and positive impact on output = % 46.1337513061651
% of points that have high values and positive impact on output = % 0.0
% of points that have low values and Negative impact on output = % 1.4106583072100314
% of points that have high values and Negative impact on output = % 52.45559038662487
Global interpretability for  Vacuum
No of points that have no impact on output= % 0.0
 % of points that have low values and positive impact on output = % 51.67189132706375
% of points that have high values and positive impact on output = % 0.0
% of points that have low values and Negative impact on output = % 0.2612330198537095
% of points that have high values and Negative impact on output = % 48.06687565308255
Global interpretability for  Pressure
No of points that have no impact on output= % 0.0
 % of points that have low values and positive impact on output = % 0.0
% of points that have high values and positive impact on output = % 39.446185997910135
% of points that have low values and Negative impact on output = % 53.08254963427377
% of points that have high values and Negative impact on output = % 7.471264367816093
Global interpretability for  Humidity
No of points that have no impact on output= % 0.0
 % of points that have low values and positive impact on output = % 46.23824451410658
% of points that have high values and positive impact on output = % 0.8881922675026124
% of points that have low values and Negative impact on output = % 0.0
% of points that have high values and Negative impact on output = % 52.87356321839081
```

Figure 5.15: Additional Statistics derived for each feature.

5.10 Conclusions

In this case study, experiments have been conducted to evaluate the effectiveness of different explainability tools. The objective was to assess the practical application of various post-hoc XAI techniques, with SHAP being one of them. These tools were applied to multiple prediction models designed for forecasting the output power of a Combined Cycle Gas Turbine (CCGT) station. The primary focus of these experiments was to harness XAI algorithms in order to generate easily comprehensible explanations for the predictions made by these models.

The adoption of a post-hoc approach in this study has proven particularly advantageous. It affords the flexibility to select and employ different ML prediction models without being compelled to compromise their performance. Furthermore, the choice of SHAP techniques stands out due to its inherent advantages over other post-hoc XAI methods, making it a suitable candidate for this investigation.

The dataset utilised throughout this case study is the publicly available CCGT dataset. The transparency and accessibility of this dataset have enhanced the reproducibility and credibility of our findings.
In Table 5.8, a comparative analysis of various XAI tools has been presented, including SHAP, LIME, Permutation Importance, and Dependence Plot. The comparison highlights key attributes such as their theoretical foundations, local and global explanation capabilities, provision of visual explanations, and overall reliability.
This comparative study demonstrates our methodical approach in selecting XAI tools that not only reveal critical insights into feature influence but also support a wider understanding of model behaviour under varying environmental conditions. By choosing tools like SHAP, we aim to meet the dual goals of accuracy and interpretability, making it possible for practitioners to trust the model's insights while supporting replicable scientific findings.

Among these tools, SHAP emerges as the most comprehensive XAI solution, boasting a solid theoretical background, the ability to offer both local and global explanations, superior visual explanations, and a high level of reliability. In contrast, LIME provides local explanations but suffers from instability and limited global explanatory power, while Permutation Importance and Dependence Plot exhibit certain limitations in their theoretical foundations and capabilities.

Table 5.8: Comparison of XAI Tools with Additional Details

XAI Tool	**Background**	**Local**	**Global**	**Visual**	**Reliability / Limitations**
SHAP [56]	Solid	Yes	Yes	Diverse visual.	Reliable; suitable for both individual and feature-dependent analyses, does not require independence assumptions
LIME [57]	Limited	Yes	No	Yes	Instability under feature correlation; best for local explanations [293]
Permutation Importance [51]	Limited	No	Yes	No	Moderate reliability; less effective for correlated features or complex relationships [297]

Continued on next page

XAI Tool	**Background**	**Local**	**Global**	**Visual**	**Reliability / Limitations**
Dependence Plot [156]	Limited	No	No	Yes	Reliable but mainly for visual exploration; lacks theoretical consistency in complex cases [51]

The results of this endeavour underscore the remarkable utility of the SHAP tool in furnishing explanations at both local and global levels. This capability facilitates a profound understanding of the model's features and their respective impacts on the target variable.

It is worth noting that this case study has been designed as an open-access resource. Its primary objective has been to serve as a guiding example for the application of XAI tools, thereby paving the way for their seamless integration into various other case studies and research endeavours. By doing so, the aim is to foster the broader adoption and utilisation of XAI techniques in the scientific community, ultimately contributing to more interpretable and trustworthy ML models.

In conclusion, our thorough exploration and comparative analysis of various XAI tools have led us to a strong rationale for the preference of the SHAP tool as the optimal choice for our predictive modelling and explanation generation tasks. First and foremost, SHAP distinguishes itself with a robust theoretical foundation. Its underlying principles and methodologies are well-established within the field of XAI, providing a solid basis for the generation of meaningful explanations. This theoretical rigour instils confidence in the reliability and interpretability of the explanations it produces.

Furthermore, one key aspect that sets SHAP apart is its consistency. Unlike some other XAI tools that were evaluated, stable and coherent explanations were consistently provided by SHAP. This reliability is critical in ensuring that the explanations generated remain dependable across different runs and instances, reinforcing the trustworthiness of the insights derived.

Perhaps most importantly, our deep understanding of the theoretical underpinnings of SHAP has empowered us to transform its complex explanations into formats that are readily understandable to domain experts. This translation of SHAP's theoretical foundations into user-friendly explanations is of paramount importance. It bridges the gap between ML experts and stakeholders in the application domain, facilitating effective communication and decision-making.

In light of these considerations, SHAP emerges as not only a powerful and theoretically sound XAI tool but also as the preferred choice for our specific case study. Its consistency, combined with our ability to convey its insights in a comprehensible manner, solidifies our confidence in SHAP as the optimal tool to enhance the interpretability and utility of our predictive models.

5.11 Case Study 2: Gearbox in a Boiler Feed Pump

The dataset utilised in this case study captures operational data related to the gearbox system of the main boiler feed pump. This gearbox plays a crucial role in the overall functionality of the boiler feed pump, which is responsible for supplying pressurised water to the boiler in power generation facilities. The efficient and reliable operation of this gearbox is essential, as it directly impacts the pump's performance and, ultimately, the boiler's efficiency.

The dataset under examination encompasses diverse valve positions and corresponding RMS-vibration measurements. The primary aim of this case study is to employ XAI tools to analyse these measurements and discern whether there exists a correlation between the valve position (particularly the closing of the valve) and RMS vibration levels. The operational metrics utilised in this analysis are:

- **Stop-Valve Position:** This metric denotes the configuration of the stop-valve, representing a critical parameter in our analysis. Changes in the valve position are hypothesised to influence the vibration levels within the gearbox system.
- **RMS Vibration:** RMS vibration serves as a quantitative measure of vibration magnitude, providing insights into the mechanical behaviour of the system under varying operational conditions. High levels of vibration may indicate mechanical stress or instability, which could compromise the gearbox's performance.

The practical significance of this investigation lies in understanding how specific operational adjustments—such as altering the stop-valve position—can influence vibration characteristics within the gearbox. Excessive vibrations not only impact the mechanical efficiency of the gearbox but may also lead to component degradation, increased maintenance costs, and reduced lifespan of the main boiler feed pump system. Therefore, identifying patterns between stop-valve positioning and vibration metrics could inform maintenance schedules, optimise operational settings, and enhance the gearbox's longevity.

In summary, this dataset provides a rich source of operational data that, when analysed with XAI tools, can reveal critical insights into the factors influencing RMS vibration in the gearbox system of the main boiler feed pump. By identifying and understanding these relationships, the study contributes valuable knowledge to predictive maintenance and operational optimisation in mechanical systems involving gearboxes.

The following stages have been implemented on this dataset:

5.11.1 Pre-processing stage

In the initial stages of our analysis, data preprocessing was conducted to ensure the integrity and usability of the dataset. The process commenced with a comprehensive inspection of the data, revealing essential details such as data dimensions, data types, and the presence of missing values. Subsequently, data cleaning was undertaken to address any missing or duplicate data points. For data transformation, the appropriate formatting of timestamp data into datetime objects was ensured. Additionally, feature engineering was performed, extracting relevant temporal features from the timestamp, such as the hour of the day and day of the week, to capture potential time-dependent patterns. Descriptive statistics were computed, yielding key statistics for both stop-valve positions and RMS-vibration, as demonstrated in the Table 5.9. Correlation analysis was conducted to quantify relationships between the variables. The correlation matrix in Table 5.10 illustrates a strong positive correlation (0.8955) between "Stop-Valve Position" and "RMS-Vibration." This significant correlation indicates that changes in stop-valve position tend to be accompanied by corresponding variations in RMS-vibration. This insight will guide our further exploration of their relationship in

Statistic	Stop-Valve Position	RMS-Vibration
Count	44,143	44,143
Mean	85.121	0.897
Std	29.668	0.297
Min	0.165	0.000
25%	79.199	846
50%	10840	1.007
75%	100.859	1.036
Max	100.902	1.968

Table 5.9: Descriptive Statistics for Stop-Valve Position and RMS-Vibration

the subsequent analysis.

	Stop-Valve Position	RMS-Vibration
Stop-Valve Position	1.0000	0.8955
RMS-Vibration	0.8955	1.0000

Table 5.10: Correlation Matrix between Stop-Valve Position and RMS-Vibration

The histogram of RMS Vibration shown in Figure 5.16 reveals a distinctive bimodal distribution with two discernible components. The first component initiates at a value of 0, encompassing nearly 3845 instances with RMS Vibration values clustered around this low point. As values gradually increase from 0 to 0.7, this component represents a significant portion of the data characterised by lower RMS Vibration levels. In contrast, the second component commences at approximately 0.7, forming a bell-shaped distribution centred around a value near 1. This second component signifies a separate group of data points with higher RMS Vibration values.

The bimodal distribution in RMS Vibration data is significant because it identifies two distinct operational states in the gearbox system. The first component, with low vibration levels, likely reflects normal operation or idle states, while the second, with higher vibration, may indicate stress or altered conditions due to stop-valve positions. Recognising these two states aids in diagnosing potential issues, optimising maintenance, and preventing damage. Moreover, this insight supports the case study's goal of using XAI tools to develop predictive models that reliably highlight conditions leading to increased vibration, improving system performance and reliability.

The histogram depicting 'Stop-Valve Position' in Figure 5.17 unveils a multifaceted

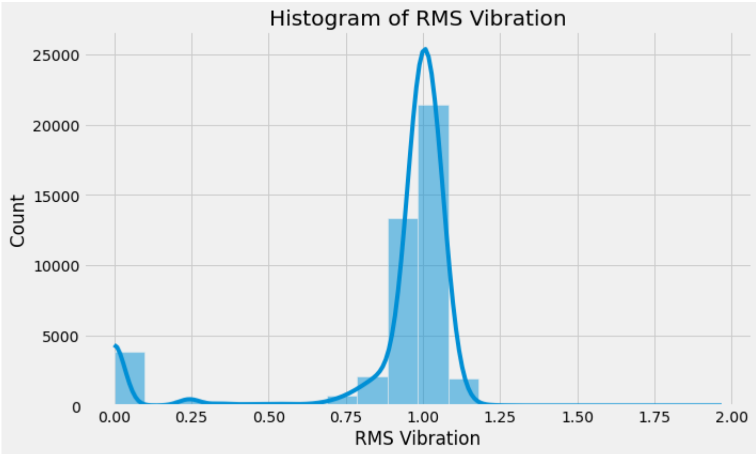


Figure 5.16: Histogram of RMS vibration revealing a distinctive bimodal distribution with two discernible components

distribution characterised by several distinctive features. To begin, the distribution initiates with a cluster of values around 0, indicating a notable presence of instances where the 'Stop-Valve Position' is at its lowest point. The frequency of occurrences then gradually diminishes as the values progress from 0 to around 70. At this juncture, the distribution transitions into a bell-shaped pattern, with the peak centred around this mid-range value. This central peak implies a significant concentration of data points with 'Stop-Valve Position' values near 70, signifying a mode within the data. Beyond this central peak, the distribution exhibits a secondary rise, creating a partial bell-shaped pattern again, and culminating with its highest peak at the maximum 'Stop-Valve Position' value of 100. These features collectively portray a nuanced distribution, with distinct modes and regions of concentration, offering valuable insights into the variability of 'Stop-Valve Position' within the dataset.

Before employing ML prediction models, the time-series dataset was transformed into a supervised learning problem, converting sequences into pairs of input and output sequences. The dataset format is depicted in Table 5.11 and Table 5.12. Following this transformation, operational measurements were utilised to forecast rms-vibration(t+1), encompassing Stop-valve-position(t-2), rms-vibration(t-2), stop-valve-position(t-1), rms-vibration(t-1), Stop-valve-position(t), rms-vibration(t), and Stop-valve-position(t+1).

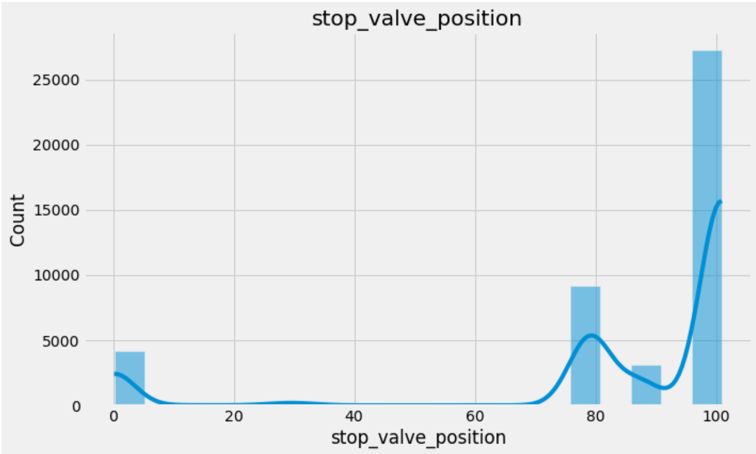


Figure 5.17: Histogram of Stop-Valve position unveils a multifaceted distribution characterised by several distinctive features.

stop_valve_position	**rms_vibration**
87.992363	1.035798
87.991951	1.006615
87.991325	1.006615

Table 5.11: Data before transforming

5.11.2 ML Model Selection

A comprehensive assessment was carried out regarding the selection of ML models for the prediction of rms-vibration. To explore the relationship between RMS Vibration and stop-valve position, three different models were employed: Linear Regression, XGBoost Regressor, and Random Forest Regressor. Each model was selected for its unique strengths and ability to capture different aspects of the relationship between the variables.

- **Linear Regression (Baseline Model)**:
  - Linear Regression is a straightforward, interpretable model, allowing for an assessment of linear relationships between stop-valve position and RMS vibration. This model serves as a baseline to compare against more complex models, helping determine if non-linear relationships are significant.
- **XGBoost Regressor**:

1 1.5

Table 5.12: Data after transforming

Stop Valve Position (t-2)	RMS Vibration (t-2)	Stop Valve Position (t-1)	RMS Vibration (t-1)	Stop Valve Position (t)	RMS Vibration (t)	Stop Valve Position (t+1)	RMS Vibration (t+1)
87.992363	1.035798	87.991951	1.006615	87.991325	1.006615	87.990913	1.035798
87.991951	1.006615	87.991325	1.006615	87.990913	1.035798	87.990501	1.035798
87.991325	1.006615	87.990913	1.035798	87.990501	1.035798	87.990082	1.035798

  - **Capturing Complex Non-linear Relationships**: XGBoost is well-suited to handle complex, non-linear data patterns. Given the potential non-linear interactions between stop-valve position and RMS Vibration, this model can capture nuanced dependencies that may not be detected by simpler models.

- **Random Forest Regressor**:
  - **Handling Non-linear and Complex Interactions**: Like XGBoost, Random Forest is adept at capturing non-linear relationships. The model's bagging approach reduces overfitting, which is crucial when dealing with potential noise or outliers in the data. This characteristic makes Random Forest particularly useful for capturing non-linear relationships without excessive model complexity.

The hyperparameters for each model are as follows:

- **Linear Regression**:
  - This model is initialised with default parameters.
- **XGBoost Regressor**:
  - initialised with default parameters provided by the `XGBRegressor` class.
- **Random Forest Regressor**:
  - `max_depth` = 6

 - `random_state` = 0
 - `n_estimators` = 100

Each model is trained on the training dataset, and predictions are made on the validation dataset. Finally, the output is averaged across the predictions of all three models to obtain the final result. Subsequently, an ensemble framework was employed, where these models were combined using an averaging technique. This ensemble model was crafted to harness the diversity offered by the individual models, with the primary aim of achieving an improvement in overall predictive accuracy.
The Mean Squared Error (MSE) metric has been used to evaluate the models' prediction accuracy by calculating the average of the squared differences between actual and predicted values, as shown in Equation 5.3

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{5.3}$$

where:

- n: number of observations,
- y_i: actual value of the i-th observation,
- $\hat{y}_i$: predicted value for the i-th observation.

The results, as presented in Table 5.13, clearly indicate the effectiveness of the ensemble model, as it outperforms the individual models with a lower Mean Squared Error (MSE) of 0.001345. This outcome underscores the potential of ensemble modelling in achieving superior predictive accuracy and hints at its promising applications in practical scenarios. Furthermore, it suggests possibilities for future research on the interpretability and generalisation capabilities of ensemble models, which could have far-reaching implications for enhancing decision-making processes across various operational contexts.

ML Model	**MSE**
Linear Regression	0.0014
Random Forest	0.0014
XGBoost	0.0015
Ensemble Model (Averaging)	0.001345

Table 5.13: Mean Squared Error (MSE) for Different ML Models

5.11.3 Applying XAI Technique

In our recent case study, a strategic decision was made to employ SHAP to gain insights into the inner workings of our ML model. This tool proved instrumental in shedding light on the predictions generated by our Random Forest model. In Figure 5.18, a local explanation is presented, focused on a single prediction, with the corresponding feature vector values displayed in Table 5.14. Notably, the local explainability plot in Figure 5.18 highlights the most influential feature for this specific prediction, which is the rms-vibration at time 't'. It's depicted in red, signifying a positive correlation with the output rms-vibration at time 't+1', indicating that an increase in the former leads to a corresponding increase in the latter. The SHAP local explanation further ranks the significance of features, revealing the following order of importance: rms-vibration(t), rms-vibration(t-1), stop-valve-position(t+1), stop-valve-position(t-2), stop-valve-position(t), and rms-vibration(t-2).

Feature	Value
Stop-valve-position(t-2)	10.9
rms-vibration(t-2)	0.8802
stop-valve-position(t-1)	100.8539
rms-vibration(t-1)	0.8802
Stop-valve-position(t)	100.85
rms-vibration(t)	0.8802
Stop-valve-position(t+1)	100.8547

Table 5.14: Feature Values for the local shap force plot in Figure 5.18.

In addition, SHAP provides a broader view of model explanations, as demonstrated in Figure 5.19, which presents the SHAP summary plot. This plot helps us understand the main factors influencing our model's predictions on a global scale. It displays SHAP values for each feature, showing how they impact the model's predictions. The most

Figure 5.18: SHAP force plot for a single instance in the dataset showing the most important features locally, which are rms_vibration(t), rms_vibration(t-1) coming next.

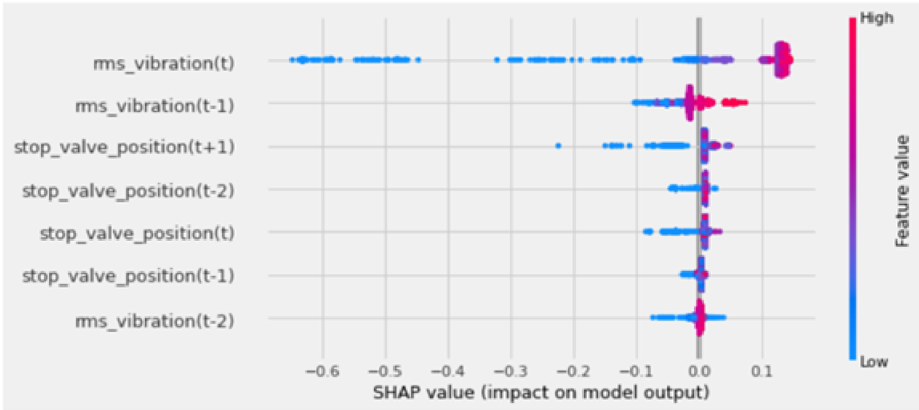


Figure 5.19: SHAP global interpretability plot, also known as SHAP summary plot, showing the most important features.

important features, ranked in descending order, include rms-vibration(t-1), stop-valve-position(t+1), stop-valve-position(t-2), stop-valve-position(t), stop-valve-position(t-1), and finally rms-vibration(t-2). The SHAP summary plot also visually represents the correlation between each feature and the output. For example, it clarifies that rms-vibration(t) has a positive correlation with the output (rms-vibration(t+1)), meaning that higher values of rms-vibration(t) correspond to higher values in the output, which is a similar pattern observed for rms-vibration(t-1) and stop-valve-position(t+1).

In conclusion, the utilisation of both global and local SHAP explanations has greatly enhanced our understanding of the ML models employed in predicting rms-vibration. The global SHAP summary plot, as depicted in Figure 5.19, provided a comprehensive overview of the key features influencing our model's predictions. It allowed us to discern the relative importance of each feature. Moreover, the local SHAP explanations, as illustrated in Figure 5.18, offered insights into individual predictions. Notably, they highlighted the immediate impact of features such as rms-vibration(t) on specific predictions. This dual approach of combining global and local explanations has not only increased the interpretability of our models but also provided valuable insights for optimising system performance and predictive accuracy in the context of rms-vibration forecasting.

5.11.4 Generating Human-understandable Explanations

Novel algorithms have been developed with the specific aim of generating human-understandable explanations. These novel algorithms will be explained in more detail in Chapter 6 when the developed framework has been introduced.

In this phase, these algorithms have been employed to automatically generate text-based explanations that are easily comprehensible. Figure 5.20 provides an example of text-based explanations aligned with the SHAP local explanations plot depicted in Figure 5.18. Initially, the locally generated text-based explanations highlight the crucial features for a specific instance. Subsequently, they depict the influence of each feature's value on the output, as depicted in the accompanying figure. Furthermore, Figure 5.21 showcases text-based explanations corresponding to the SHAP global interpretability plot. It highlights the globally significant features for our prediction model.

Figure 5.22 presents basic statistics for each feature, including the count of values having no impact on the output, the count of values considered high with a significant positive impact on the output (resulting in an increase), and the count of values deemed low with a notable positive/negative impact on the output, influencing the prediction to decrease or increase.

In conclusion, this case study encompassed the use of various prediction models, subject to performance comparisons. It was discerned that Random Forest outperformed the other ML models used. Feature importance analysis in the prediction model was conducted by harnessing the SHAP tool. However, it's worth noting that the XAI algorithms developed are still in their nascent stages. There exist opportunities for enhancements to yield more dependable and comprehensible explanations.

```
instance no 0  taken at time:  2014-09-20 17:10:00
Turbine speed at this time:
  TurbinespeedA=  4517.489258
  TurbinespeedB=  4521.881348
features affecting the rms-vibration in order are
  rms_vibration(t)
  rms_vibration(t-1)
  stop_valve_position(t+1)
  stop_valve_position(t-2)
  stop_valve_position(t)
  rms_vibration(t-2)
  stop_valve_position(t-1)
impact of each feature value on output
stop_valve_position(t-2)  100.8535309 High value contributes to higher rms-vibration
rms_vibration(t-2)  0.8802593 Low value contributes to lower rms-vibration
stop_valve_position(t-1)  100.8539429 High value contributes to higher rms-vibration
rms_vibration(t-1)  0.8802593 Low value contributes to lower rms-vibration
stop_valve_position(t)  100.8543549 High value contributes to higher rms-vibration
rms_vibration(t)  0.8802593 Low value contributes to higher rms-vibration
stop_valve_position(t+1)  100.8547668 High value contributes to higher rms-vibration
```

Figure 5.20: Automated text-based explanation generated using XAI algorithms.

```
Feature importance in Ascendng order
rms_vibration(t-2)
stop_valve_position(t-1)
stop_valve_position(t)
stop_valve_position(t-2)
stop_valve_position(t+1)
rms_vibration(t-1)
rms_vibration(t)
```

Figure 5.21: Text-based explanations corresponding to SHAP summary plot in Figure 5.19.

```
Global interpretability for  rms_vibration(t-2)
No of points that have no impact on output= % 0.0
 % of points that have low values and positive impact on output = % 0.8630648997621474
% of points that have high values and positive impact on output = % 81.97757390417941
% of points that have low values and Negative impact on output = % 16.71083927964662
% of points that have high values and Negative impact on output = % 0.44852191641182465
Global interpretability for  stop_valve_position(t-1)
No of points that have no impact on output= % 0.0
 % of points that have low values and positive impact on output = % 3.9755351681957185
% of points that have high values and positive impact on output = % 0.12232415902140673
% of points that have low values and Negative impact on output = % 0.0
% of points that have high values and Negative impact on output = % 95.90214067278288
Global interpretability for  rms_vibration(t-1)
No of points that have no impact on output= % 0.0
 % of points that have low values and positive impact on output = % 0.12232415902140673
% of points that have high values and positive impact on output = % 81.14169215086646
% of points that have low values and Negative impact on output = % 17.44478423377506
% of points that have high values and Negative impact on output = % 1.291199456337071
Global interpretability for  stop_valve_position(t)
No of points that have no impact on output= % 0.0
 % of points that have low values and positive impact on output = % 0.013591573224600747
% of points that have high values and positive impact on output = % 83.7037037037037
% of points that have low values and Negative impact on output = % 3.961943594971118
% of points that have high values and Negative impact on output = % 12.320761128100578
Global interpretability for  rms_vibration(t)
```

Figure 5.22: Simple Text-based statistics corresponding to SHAP summary plot 5.19.

5.12 Conclusion

This chapter demonstrated how explainable AI can be systematically integrated into condition monitoring workflows, with a focus on selecting and applying XAI tools that generate explanations meaningful to engineering stakeholders. A set of key considerations was first established—application goals, data characteristics, model selection, stakeholder needs, required explainability level, and evaluation criteria—highlighting that explainability should ideally be considered from the start of system design, while still remaining feasible to retrofit into existing pipelines.

Building on these considerations, the chapter presented a structured set of stages for operationalising XAI in practice: data pre-processing, prediction model development, application of XAI tools, and the generation of human-understandable explanations. This staged blueprint was then validated through two case studies that increased in complexity and realism.

In Case Study 1 (CCGT power output prediction), multiple regression models were evaluated and several post-hoc tools were compared, including LIME, permutation importance, partial dependence plots, and SHAP. The comparison illustrated that while LIME can offer intuitive instance-level explanations, it may produce inconsistent feature rankings across repeated runs due to its perturbation-based sampling. In contrast, SHAP provided stable and theoretically grounded explanations and supported both local and global interpretability, making it more suitable for safety-critical engineering contexts where explanation consistency is essential. The case study also demonstrated how SHAP outputs can be used to communicate both the dominant global drivers of power output and the feature-level reasoning behind individual predictions.

Case Study 2 extended the investigation to a gearbox condition monitoring setting, where time-series data were transformed into a supervised learning problem and multiple regression models were combined into an ensemble to improve predictive performance. SHAP was again employed to interpret the model locally and globally, revealing that recent vibration values and valve position dynamics were among the most influential factors driving the predicted vibration level. This case study confirmed that SHAP can remain effective even when feature representations reflect temporal dependencies and operational sequences, reinforcing its applicability to industrial monitoring scenarios.

Across both case studies, a key outcome emerged: producing explainability plots alone is insufficient for real-world deployment when end users lack ML expertise. Even highly informative visual explanations can be difficult to interpret without additional processing and contextualisation. Therefore, the chapter motivated the need for methods that translate complex XAI outputs into concise, stakeholder-oriented explanations. This requirement directly leads to Chapter 6, where the proposed framework and the algorithms for converting SHAP-based explanations into human-understandable text are introduced and evaluated.

Chapter 6

A Framework for user-friendly Explanations in Industry and Engineering Applications

6.1 Introduction

This chapter presents the proposed framework, developed as a direct extension of the key stages outlined in Chapter 5 and originally established in [298]. Its development builds upon the considerations discussed previously and extends the foundational stages introduced earlier.

This framework is purpose-built for implementation within industrial and engineering contexts. Its central objective is to cater to individuals who may not possess expertise in ML, ensuring accessibility and comprehensibility. This is achieved by providing explanations that are not only accessible but also readily understandable to domain experts. These explanations are thoughtfully presented in both human-readable text and visual formats, effectively bridging the divide between intricate ML models and domain experts. Thus, our framework not only reflects the foundational stages but also integrates the considerations laid out in Chapter 5, forming a holistic approach to designing XAI systems.

The framework is organised into five distinct yet interrelated explanatory stages, as illustrated in Figure 6.1. Each of these stages will be elaborated upon in the subsequent sections, offering a more comprehensive understanding and building upon the concepts introduced earlier.

In [299], authors focus on evaluating machine learning explanations through a comprehensive taxonomy across five key dimensions, aiming to compare and analyse different XAI methods systematically. It highlights the theoretical vs. practical discrepancies of these methods and introduces Explainability Fact Sheets for quick assessment. This approach is rooted in academic analysis, offering tools for a broad evaluation of explainable systems.

The framework in this research is specifically designed for industrial and engineering contexts, prioritising accessibility and comprehensibility for non-ML experts. It emphasises delivering explanations in formats that domain experts can easily understand, using human-readable text and visuals to demystify complex ML models. This framework aims to provide a practical, user-friendly approach to XAI, making it directly applicable in professional settings where ML expertise might be limited.

6.2 A Framework Proposing Understandable Explanations For Non-expert Users In Industrial and Engineering Applications

After implementing XAI tools in the case studies discussed in the previous chapter, the methodology was subsequently redefined and expanded. A comprehensive framework has been developed specifically tailored to engineering applications. Furthermore, this framework can be easily adapted to various other engineering and industrial contexts. The primary aim of this framework is to meet the requirements of individuals who are not necessarily experts in ML but are working within industrial and engineering domains.

The central goal of this framework is to provide accessible and understandable explanations, presented in both user-friendly text and graphical formats. By doing so, it bridges the gap between intricate ML models and domain experts. The first four

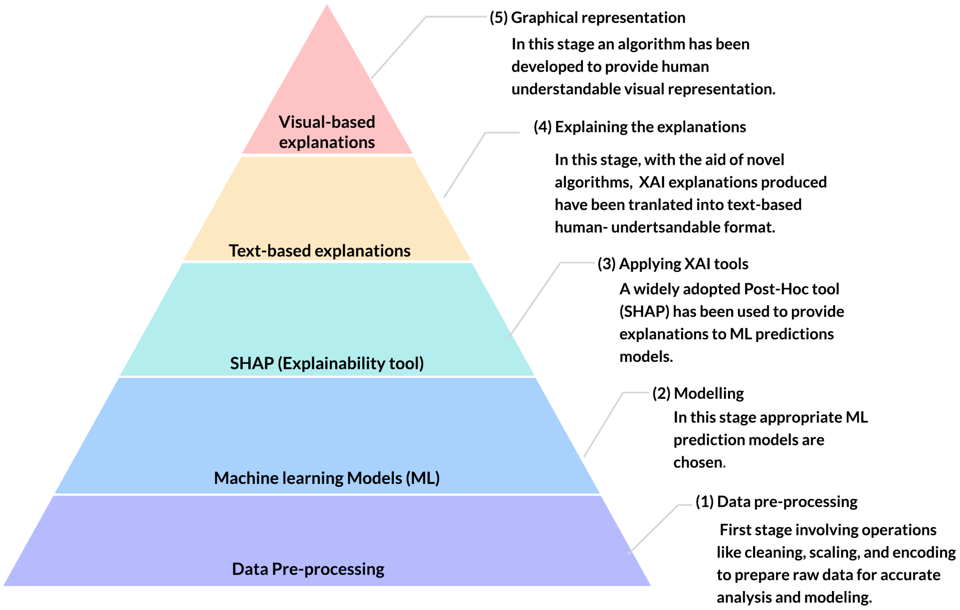


Figure 6.1: Proposed framework for understandable ML explanations for non-ML expert users.

stages of our framework have been previously established and are briefly defined here for context. Building upon these foundational elements, additional stages are introduced, expanding our framework to encompass a comprehensive and interconnected suite of explanatory strategies. This approach ensures a seamless integration of established principles with new advancements, creating a robust framework tailored to enhancing explainability within machine learning systems, as illustrated in Figure 6.1. Subsequently, each of these stages will be explored in detail.

1. Data Pre-processing: In the initial stage, the focus is on preparing the data for analysis. This involves evaluating the quality of the dataset and transforming it into a more comprehensible format. The goal is to make the data easier to understand, with clear definitions of its features. These preparatory steps are crucial as they lay the foundation for subsequent ML or analytic models.

2. Prediction models: Moving to the second stage, the task is to choose an appro-

priate ML prediction model. Many ML models are often considered "black-box," meaning that their inner workings are not easily interpretable. This opacity can pose challenges for understanding how these models arrive at their predictions. Hence, in this stage, it becomes imperative to incorporate XAI techniques that can elucidate the reasoning behind the model's predictions. This is essential for industry stakeholders seeking transparency and comprehension of the adopted ML models.

3. Application of XAI Tools: Building on the insights from [298], an additional layer is introduced to the explanation framework. Here, XAI tools are utilised to shed light on the inner workings of the ML models and to explain the rationale behind their predictions. For instance, SHAP is applied to generate dependable explanations. It provides both local explanations for individual predictions and a holistic global explanation of how the ML model operates. However, SHAP often produces intricate plots that can be daunting, especially for non-ML experts. Therefore, the subsequent stages aim to translate these complex explanations into a more user-friendly format.

4. Creation of Human-Understandable Explanations: Effectively communicating explanations to individuals in the application domain who may lack expertise in ML is a critical challenge. The visual plots generated by SHAP may be intricate and time-consuming to interpret. Consequently, there is a need to convert these plots into explanations that can be easily grasped by non-ML experts. Innovative algorithms have been developed to transform SHAP's local and global explanations into text-based explanations that are more accessible to a broader audience.
   In Figure 6.2 and Figure 6.3, the process of converting SHAP summary plots and SHAP local plots into text-based explanations is depicted.
   In Figure 6.2, the flowchart for converting SHAP summary plots to text-based explanations is illustrated. The process begins by aggregating the SHAP values for each feature. These aggregated SHAP values are then compared to determine the most important features. Finally, the features are ordered according

to their aggregated SHAP values, ensuring that the most influential features are highlighted first. This structured approach facilitates a clear and comprehensible transformation of SHAP plots into textual explanations.

In Figure 6.3, the algorithm focuses on transforming SHAP local plots into text-based explanations:

1. **Comparison of SHAP Values and Feature Values**: For each specific prediction, the algorithm examines the SHAP value of each feature.
2. **Determining Impact Based on SHAP Value**:
   - **Negative SHAP Value**:
     2.1. If the SHAP value is negative, the algorithm checks the feature value to determine its relationship to its mean value.
     2.2. If the feature value is less than or equal to its mean, it is considered to have a low value that negatively impacts the prediction or output value.
     2.3. Conversely, if the feature value is greater than its mean, it is identified as having a high value that negatively impacts the prediction.
   - **Positive SHAP Value**:
     2.1. If the SHAP value is positive, the algorithm again checks the feature value against its mean.
     2.2. If the feature value is less than its mean, it indicates a low value that positively impacts the prediction.
     2.3. If the feature value is greater than its mean, it indicates a high value that positively impacts the prediction.

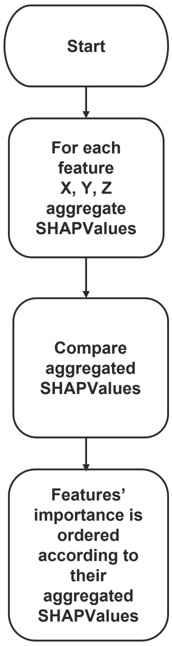


Figure 6.2: Flowchart to transform SHAP summary plot to text-based explanations

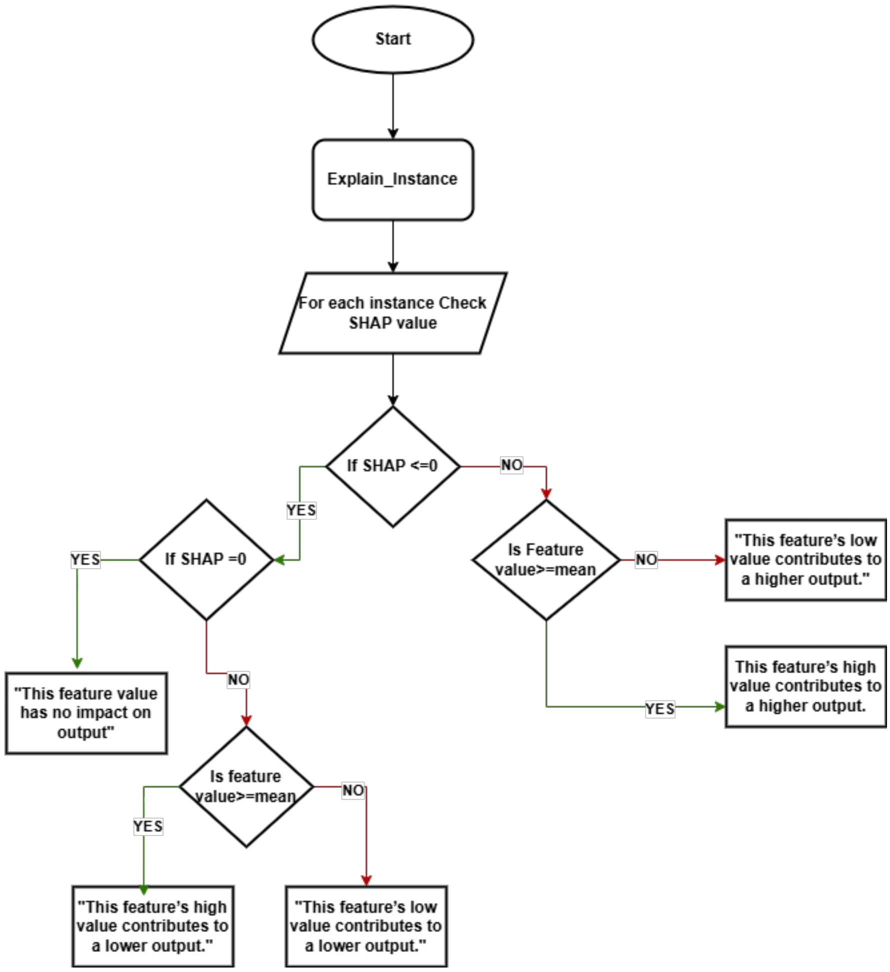


Figure 6.3: Flowchart to transform SHAP local plot to text-based explanations

5. Graphical Representation for Domain Experts: While many XAI approaches cater primarily to ML experts, this framework introduces an additional layer of explainability that specifically targets non-expert users. This stage focuses on providing graphical representations tailored to domain experts who are associated with a specific technical field or industry domain. In this context, a novel algorithm has been developed to convert SHAP plots into graphical representations that align with the expertise of non-ML users, particularly within the technical domain of the case study, such as the prediction of mechanical pump degradation. These graphical representations bridge the gap between complex ML models and domain experts, making AI more accessible and interpretable in specific application

areas. In Figure 6.4, the process of obtaining these graphical representations is depicted.

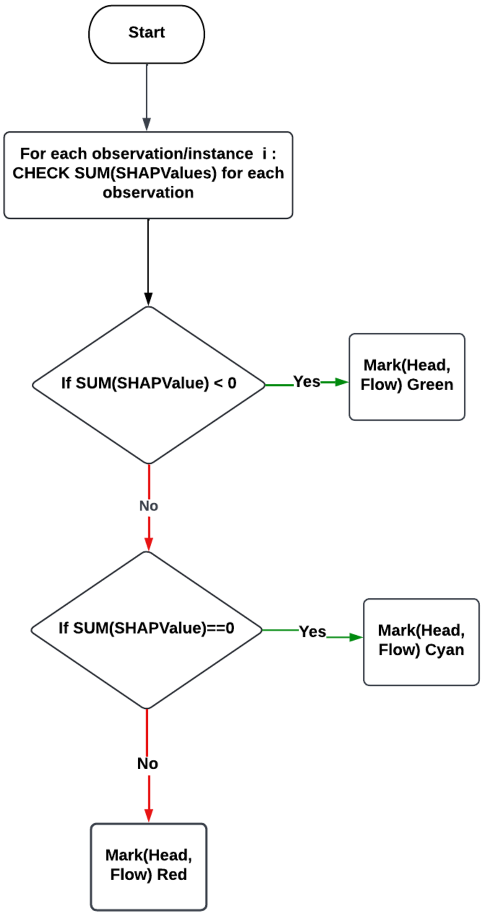


Figure 6.4: Flowchart to transform SHAP local plot to graphical representations

This framework will be applied to a real condition monitoring case study, where each stage will be visited and further explained. Additionally, it will be elucidated how visual-based explanations have been generated specifically for this case study and how an algorithm has been developed to accomplish this task. Furthermore, how this framework can be reapplied to other case studies to facilitate a broader understanding of its applicability will be outlined.

6.3 Validating Explanations Generated from XAI Tools

Many scholarly studies have commonly employed human participants to validate the explanations generated by XAI tools. For instance, in some cases, individuals were randomly chosen from the general public, devoid of any prior technical or domain-specific knowledge. These participants were instructed to interact with one or more explanatory tools and offer their feedback through the completion of questionnaires [300]. Additionally, domain experts were enlisted to provide informed assessments of the explanations produced by these methods, along with verifying the consistency of these explanations with domain-specific knowledge [301] .

In the realm of XAI research, it is a common practice to involve human participants in the validation process of AI-generated explanations. This approach allows researchers to gauge the effectiveness and comprehensibility of these explanations from the perspective of end-users or domain experts. Typically, individuals from the lay public, who lack technical or domain-specific knowledge, are randomly selected to interact with XAI tools and provide valuable feedback through various means, such as questionnaires. This human-in-the-loop evaluation process aids in assessing the clarity and utility of the generated explanations, contributing to the refinement and improvement of XAI techniques.

However, in this specific research domain, a unique challenge emerged. Unlike many XAI studies that benefit from access to domain experts for evaluating the quality of explanations, this research faced a scarcity of experts, as industrial condition monitoring specialists are few, widely dispersed, and typically unable to commit time to research validation. This scarcity posed a significant hurdle in our efforts to determine the quality and reliability of the explanations generated by our XAI methods.

To address this challenge, an innovative approach has been adopted. Instead of relying on domain experts, existing domain knowledge has been leveraged as a substitute for expert evaluations. This knowledge has been employed to validate the explanations generated by our XAI tools. By comparing the explanations to the established domain knowledge, the aim is to ascertain the accuracy and consistency of the AI-generated insights. This unconventional approach allowed for navigating the scarcity of domain experts and still ensure the credibility and validity of explanations generated within the context of the research domain.

In essence, while the utilisation of human assessments, especially by domain experts, is a prevailing practice in XAI research, our study demonstrates the adaptability and flexibility of approaches when such expertise is not readily accessible. Leveraging existing domain knowledge serves as a viable alternative, ensuring that the explanations remain reliable and valuable in domains where domain experts may be scarce or unavailable for evaluation.

6.4 Usability of XAI Framework

Assessing the usability of an XAI framework is a multifaceted process that involves several key elements. One of the primary ways to determine its usability is through user feedback and evaluation. This typically entails involving domain experts or end-users who interact with the framework and provide feedback on their experience. Their input helps gauge the framework's effectiveness in delivering comprehensible explanations and meeting their specific needs.

Additionally, usability testing often involves measuring the framework's efficiency and effectiveness in providing insights. This can include evaluating the time required for users to grasp the explanations, the accuracy of the information conveyed, and the ease of navigating the framework's interface. Usability metrics and benchmarks may also be employed to quantitatively assess the framework's performance and identify areas for improvement.

Furthermore, the adaptability and versatility of the framework in different use cases and scenarios play a crucial role in determining usability. If the framework can be readily customised and tailored to specific domains or applications without significant modifications, it signifies its usability and flexibility.

Ultimately, a usable XAI framework should not only produce understandable explanations but should also align with the unique requirements of its intended users and contexts. The combination of user feedback, efficiency metrics, and adaptability assessments collectively contribute to the comprehensive evaluation of the framework's usability, ensuring that it effectively serves its purpose in delivering AI-generated explanations to domain experts.

In the domain of XAI, the usability of a framework is a critical factor in its effectiveness and practical application. Within the realm of condition monitoring for pump systems, our framework is specifically crafted to provide XAI explanations accessible to engineers and specialists who may not have a background in machine learning. It aims to bridge the inner workings of ML models with the need for clear, actionable insights, ensuring that engineers and specialists without ML expertise can readily understand and apply these insights in their professional activities.

It is worth mentioning that the IEEE 2894-2024 Standard [85] outlines a set of assessment metrics to evaluate the quality of explainable AI systems. These include fidelity (how well the explanation reflects the model's true behaviour), interpretability (the ease with which users can understand the explanation), robustness (the stability of explanations under small input perturbations), and fairness (the absence of discriminatory bias in explanations). Although this standard was published toward the end of this research, the proposed framework aligns with some of these principles—most notably fidelity and interpretability—by validating explanations against established domain knowledge rather than relying on user feedback. This ensures consistency with expert expectations, supporting trustworthy and usable model explanations in practice.

This framework has been employed to analyse real condition monitoring data from pump systems operating under both normal and refuelling profiles. This practical application demonstrates the framework's adaptability and relevance in real-world scenarios, where accurate insights are crucial for informed decision-making.

Moreover, this research extends beyond a singular use case. The framework has been applied to a diverse range of case studies, each with its unique characteristics and requirements. This approach showcases the versatility and usability of the framework, as it can be customised and adapted to different scenarios by making minor adjustments to its stages. This flexibility allows the framework to align seamlessly with the specific demands of each case study, ensuring that it consistently delivers meaningful and understandable explanations to domain experts.

In essence, the usability of this XAI framework lies in its ability to provide actionable insights to domain experts in a wide array of applications, from pump condition monitoring to other domains. By successfully applying this framework to multiple case studies and tailoring it to the unique needs of each scenario, its practicality and effectiveness in empowering non-expert stakeholders with valuable AI-generated explanations have been demonstrated. This usability underscores the framework's potential to enhance transparency and decision-making across various industries and applications.

6.5 Applying XAI to Condition Monitoring: Explaining A Thrust Bearing Wear Predictive Model

The ability to comprehend the predictions of ML models is crucial, especially in industrial contexts where pivotal decisions hinge on these insights. The novel framework proposed in this thesis is designed to make ML outputs accessible to domain experts, fostering trust and enabling the practical application of these technologies in critical scenarios. A prime example of its real-world impact is demonstrated in our case study on centrifugal pump operations. These pumps are vital in many industries, optimising fluid movement by converting kinetic energy into pressure. Our framework helped clarify ML model predictions on pump degradation, leading to improved operations and prolonged pump life. It emphasises the importance of operating pumps at their BEP to reduce wear and extend durability, showcasing the framework's utility in enhancing industrial equipment's operational efficiency and lifespan.

Understanding the rationale behind the predictions made by ML models holds significant importance, particularly in industrial settings where critical decisions are at stake. This framework serves as a valuable tool for domain experts, enabling them to grasp the outputs generated by ML models, instil trust in these models, and facilitate their responsible and effective deployment in real-world scenarios. The real-world impact of this framework is made evident in our case study, where human-understandable

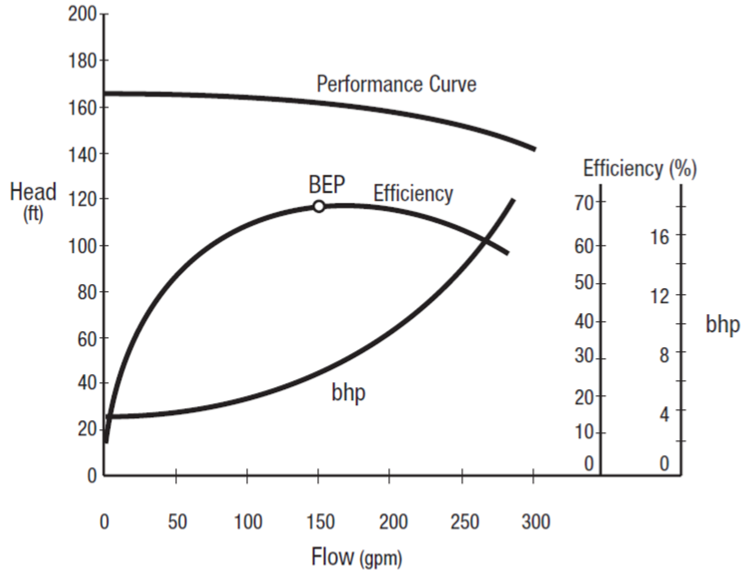


Figure 6.5: Centrifugal pump performance curves showing the best efficiency point (BEP) [4].

explanations were harnessed to elucidate a model's predictions concerning pump degradation. This application of the ML model resulted in enhanced pump operations and an extended pump lifespan.

Centrifugal pumps, integral to key processes across various industries, function by converting kinetic energy into pressure to move fluids. Their performance is described by a performance curve, as depicted in Figure 6.5 [4], which outlines the relationship between flow rates and pressure increase (Head). These pumps exhibit distinct physical attributes tailored to specific applications. During the pump design phase, engineers aim to minimise mechanical losses due to factors such as vibration, heating, or cavitation [4]. Figure 6.5 [4] highlights the significance of the Best Efficiency Point (BEP), where a pump operates converting energy into pressure increase in the most effective manner possible, e.g. by minimising vibration and associated wear. Deviating from the BEP can accelerate wear and lead to premature degradation [4]. This explanatory case study leveraged real condition monitoring data from feed-water pumps to predict thrust-bearing positions. Thrust-bearing movement is analogous to degradation, as it results from unwanted pump operation effects. As the thrust-bearing moves, the associated pads experience wear, emphasising the importance of minimising thrust-bearing movement to extend the pump's operational life. In the context of this case study and the associated analysis, the median thrust-bearing position is denoted as "median-TB" [302].

The ML model's objective is to predict the median-TB based on predetermined pump operational profiles for flow (referred to as "mean-flow") and head (referred to as "mean-head"). The ML model aims to identify the operational implications, i.e., predicted pump degradation, arising from proposed plant operating profiles and subsequently connect these findings with domain expertise. The prediction is presented in a clear and readily understandable context for the end-users of ML outputs. Two datasets were utilised:

1. Data collected during normal operation, representing operation near the BEP.
2. A profile associated with a power plant refuelling state, representing a pump operating away from the BEP.

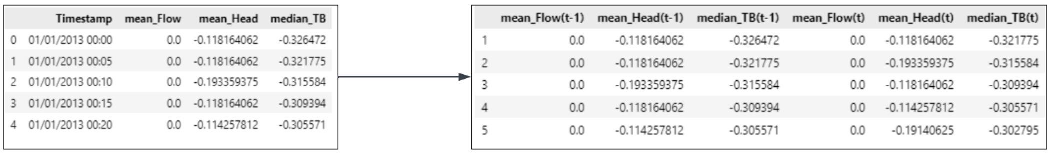
Now, the application of each stage of the proposed explainability framework in this case study will be elaborated on.

6.5.1 Data pre-processing

A thorough examination of the dataset was conducted to assess the quality and completeness, with special attention given to the identification of missing values. Instances of missing data were observed and subsequently addressed through the utilisation of the forward-filling method. This technique, specifically tailored to the temporal nature of the dataset, involved the replacement of missing values with the most recent available observations, thereby ensuring the chronological continuity of the dataset. By applying this method diligently, the dataset was rendered complete and consistent, allowing for meaningful insights to be extracted and informed predictions to be made. The process of handling missing values through forward filling played a crucial role in preserving the temporal structure and enhancing the overall reliability of the research findings, contributing significantly to the robustness of the research outcomes.

In scientific terms, stationarity in the context of time series data refers to the property where statistical properties such as mean and variance remain constant over time, thus ensuring the reliability of subsequent analyses and conclusions drawn from the dataset. The test presents the results as a test statistic and a p-value. A more negative test statistic suggests the data is stationary, indicating no unit root is present. Critical values for significance levels aid in interpreting this statistic. A low p-value (usually below 0.05) further supports stationarity, indicating consistent statistical properties over time. The stationarity of the dataset was assessed through the application of the Augmented Dickey-Fuller (ADF) test, a statistical examination method employed to ascertain the temporal stability of time series data. The results of this test indicated that the dataset exhibited stationarity, signifying that there were no significant variations or trends observed over time. As a consequence of this stationary attribute, no further processing or transformation was deemed necessary.

The datasets employed in this study consist of multivariate time series data encompassing flow and head measurements. These datasets play a pivotal role in forecasting thrust-bearing positions. To enhance the predictive modelling, the time series data underwent a restructuring process, converting them into a supervised learning framework that incorporates the temporal aspect. This transformation was achieved by incorporating lagged observations of input features to forecast thrust-bearing positions. The restructuring was executed through the utilisation of a shift function, aligning the data with the principles of supervised learning. In supervised learning, models learn from labelled training data, using this knowledge to predict outcomes for new, unseen data. By incorporating past observations as part of the input features, the temporal dynamics inherent in the time series data are preserved and utilised in the learning process, thereby enhancing the model's ability to make accurate predictions based on the sequence of data points. This approach simplifies the prediction task by removing complexities, as referenced in [303]. Figure 6.6 provides a visual representation of the data before and after this restructuring process. To predict median-TB(t), the following operational measurements are leveraged: mean-Flow(t-1), mean-Head(t-1), median-TB(t-1), mean-Flow(t), and mean-Head(t).



	Timestamp	mean_Flow	mean_Head	median_TB
0	01/01/2013 00:00	0.0	-0.118164062	-0.326472
1	01/01/2013 00:05	0.0	-0.118164062	-0.321775
2	01/01/2013 00:10	0.0	-0.193359375	-0.315584
3	01/01/2013 00:15	0.0	-0.118164062	-0.309394
4	01/01/2013 00:20	0.0	-0.114257812	-0.305571

Data before re-structuring

	mean_Flow(t-1)	mean_Head(t-1)	median_TB(t-1)	mean_Flow(t)	mean_Head(t)	median_TB(t)
1	0.0	-0.118164062	-0.326472	0.0	-0.118164062	-0.321775
2	0.0	-0.118164062	-0.321775	0.0	-0.193359375	-0.315584
3	0.0	-0.193359375	-0.315584	0.0	-0.118164062	-0.309394
4	0.0	-0.118164062	-0.309394	0.0	-0.114257812	-0.305571
5	0.0	-0.114257812	-0.305571	0.0	-0.19140625	-0.302795

Data after re-structuring

Figure 6.6: Example of data before and after restructuring.

6.5.2 Modelling

In the context of our explainability case study, three well-known and well-understood ML models were evaluated: linear regression, random forest, and XGboost. Linear regression offers simplicity and interpretability, making it ideal for initial analyses. Random Forest excels in handling complex data without intensive pre-processing and highlights influential features. XGBoost, known for its efficiency and flexibility, effectively balances performance and interpretability, even in complex scenarios. This selection spans from basic to advanced machine learning techniques, providing a broad perspective on model performance and challenges in explainability, catering to users with different levels of ML expertise. However, despite their simplicity, these models pose challenges in terms of interpretability, especially for individuals not well-versed in ML. These ML models have been applied to predict thrust-bearing positions. To gauge their performance, three key performance metrics have been employed, which can be found in Table 6.1 and Table 6.2. These metrics encompass the error magnitude (RMSE), quantifying the model's average prediction error; the R2 score, which assesses the proximity of calculated values to actual data; and the MSE, detailed in equations (6.1), (6.2), and (6.3).

Our findings revealed that, for the normal profile (as displayed in Table 6.1), the linear regression algorithm exhibited superior performance, consequently being selected as the predictive model for the normal profile dataset. Also, in the case of the refuelling profile (as illustrated in Table 6.2), the metrics for linear regression surpassed those of the random forest and XGboost regression models, boasting an R-squared value of 0.99 and an RMSE of 0.07. Therefore, linear regression was deemed the optimal predictive model for the refuelling profile dataset. (Refer to equations (6.1), (6.2) and (6.3) for additional details.)

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{6.1}$$

where:

- N: total number of observations in the dataset,
- y_i: actual value of the dependent variable for the i-th observation,
- $\hat{y}_i$: predicted value of the dependent variable for the i-th observation.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{6.2}$$

where:

- N: total number of observations in the dataset,
- y_i: actual value of the dependent variable for the i-th observation,
- $\hat{y}_i$: predicted value of the dependent variable for the i-th observation,
- MSE: the Mean Squared Error, as defined in Equation 6.1.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{6.3}$$

where:

- N: total number of observations,
- y_i: actual value of the dependent variable for the i-th observation,
- $\hat{y}_i$: predicted value of the dependent variable for the i-th observation,
- $\bar{y}$: the mean of the actual values y_i, calculated as $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$.

ML models	R2 Score	MSE	RMSE
Linear regression	0.99	0.0015	0.039
Random forest	0.77	0.085	0.29
XGBoost	0.76	0.09	0.3

Table 6.1: Comparison of ML models' performances for the normal profile.

The following are the key hyperparameters chosen for each ML model used in this study:

- **Linear Regression**:
  - This model is initialised with default parameters.
- **XGBoost Regressor**:
  - `n_estimators`: set to 100 (Default value), the number of boosting rounds, typically ranging between 100 and 500 [291].
  - `learning_rate`: set to 0.1 (Default value), controlling the contribution of each tree, often reduced in practice for improved generalisation.
  - `max_depth`: set to 6 (Default value), with values from 3 to 10 commonly applied depending on dataset complexity.
- **Random Forest Regressor**:
  - `n_estimators`: set to 100 (Explicit), the number of trees in the ensemble, balancing accuracy with computation time [41].
  - `max_depth`: set to 6 (Explicit), limiting tree growth to mitigate overfitting.
  - `random_state`: set to 0, ensuring reproducibility of results.

Data Split: 80/20 train–test ratio.

Motivation for Explaining Linear Regression

The issue at the core of the thrust bearing wear case study is not merely about explaining a transparent model like linear regression but about contextualising its explanations

ML models	**R2 Score**	**MSE**	**RMSE**
Linear regression	0.99	0.005	0.07
Random forest	0.75	0.13	0.13
XGBoost	0.72	0.15	0.39

Table 6.2: Comparison of ML models' performances for the refuelling profile.

for domain experts. While linear regression provides straightforward coefficients and predictions, these outputs often lack the necessary operational context to resonate with experts. The focus, therefore, is on translating these raw, mathematical outputs into insights that align with the specific language, knowledge, and decision-making processes used by experts in the domain [33].

This need for contextualization is reflected in broader discussions on explainability in complex systems. For instance, as emphasised in [304], effective explainability requires evaluating explanatory systems across dimensions such as usability, functional requirements, and user-centred validation. Linear regression outputs, while mathematically interpretable, often fail to meet usability standards, particularly in providing actionable and domain-specific insights.

A recent study in the biomedical domain demonstrated [305] that SHAP can be applied not only to complex black box models but also to inherently interpretable regression models, including linear and logistic regression. The analysis showed that even when transparency is already present through model coefficients, SHAP adds explanatory value by enabling visualisation, comparison, and contextualisation of feature effects. This evidence supports the view that SHAP is not restricted to opaque models but can also enrich understanding of linear models, which is directly relevant to the approach taken in this thesis.

Another line of work [306] has shown that SHAP can reveal important nuances even in simple models such as linear regression. In particular, it was demonstrated that when features are dependent, naive interpretation of coefficients may be misleading, while SHAP-based attribution helps to identify how these dependencies affect individual predictions. This illustrates that the value of SHAP is not limited to enhancing transparency in complex models, but also lies in uncovering insights that coefficients alone cannot provide, thereby reinforcing its usefulness even in linear modelling contexts.

The challenge becomes even more pressing when considering that simple models like linear regression can also exhibit information leakage if not contextualised properly [304]. This leakage can lead to misinterpretations, especially in domains where operational thresholds and domain-specific behaviours significantly influence decision-making. For the thrust-bearing wear study, leveraging an operational framework such as Explainability Fact Sheets could help align regression outputs with expert expectations and operational realities.

Moreover, it is crucial to identify the role of domain-specific sanity checks and the incorporation of common-sense knowledge bases to evaluate and justify system outputs [307]. This approach aligns closely with the challenge in the thrust-bearing wear case study: creating explanations that are understandable and actionable within the specific context of machinery wear and maintenance. For instance, presenting regression outputs alongside explanations grounded in known wear behaviours or operational thresholds can significantly enhance their relevance and trustworthiness [28] [57].

In addition, a technique such as SHAP can provide visual aids to complement numerical regression outputs. These visualisations can help bridge the gap between mathematical outputs and actionable insights, especially when coupled with domain-specific operational plots [304].

In summary, the motivation for introducing explanations in this case study is not to address the transparency of the linear regression model itself but to ensure that its outputs are contextualised and presented in a way that supports domain experts in their decision-making. This perspective aligns with the broader principle of explainability as a user-centred and domain-specific endeavour [28] [57] [304] [307].

6.5.3 XAI Tool Application

As previously discussed, the SHAP explainability tool is underpinned by a robust theoretical foundation, capable of furnishing both local and global explanations that are notably more dependable than those offered by alternative explainability tools [56].

The ensuing results stem from the application of SHAP techniques, elucidating the key determinants of model predictions. In Figure 6.7, the SHAP summary plot for normal profile data unveils the most influential features, namely median-TB(t-1), mean-Head(t-1), mean-Head(t), mean-Flow(t-1), and lastly mean-Flow(t).

Similarly, Figure 6.8 portrays the SHAP summary plot for refuelling profile data, highlighting the pivotal features shaping model predictions within the context of operational refuelling scenarios. These plots elucidate the SHAP values associated with each feature and their impact on the model's predictions. Furthermore, they present a global perspective by ranking the most crucial features, with median-TB(t-1), mean-Head(t-1), mean-Head(t), mean-Flow(t-1), and lastly mean-Flow(t) emerging as the most significant contributors, as illustrated in Figure 6.7 and Figure 6.8.

Analysing the SHAP summary plots for both profiles, it becomes evident that median-TB(t-1) exhibits a positive correlation with the model's output. An increase in median-TB(t-1) corresponds to an increase in the model's predictions.

For the normal profile, the SHAP force plot depicted in Figure 6.9 provides insights into the most critical local features. Notably, median-TB(t-1), highlighted in blue, emerges as the paramount feature, exhibiting a negative correlation with the model's output, signifying that an increase in median-TB(t-1) results in a decrease in thrust-bearing positions.

Similarly, for the refuelling profile, the SHAP force plot presented in Figure 6.9b unveils the salient local features pertaining to a specific instance, as described in the accompanying text-based explanation in Figure 6.10b. Here, median-TB(t-1) assumes the primary role as the most influential feature, denoted in blue, with a negative correlation, indicating a reduction in thrust-bearing positions as median-TB(t-1) increases. Additionally, mean-Head(t-1), also highlighted in blue, stands out as the second most significant feature, contributing to a decrease in the model's output for this particular prediction.

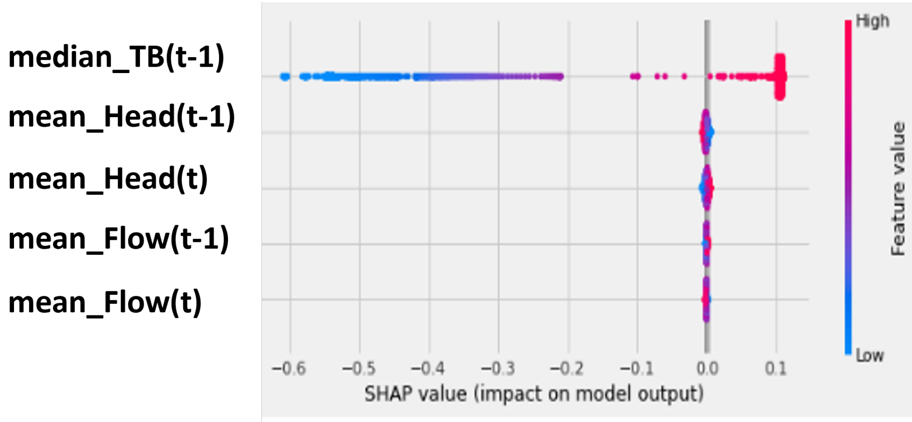


Figure 6.7: SHAP summary plot for the normal profile showing the most important features globally.

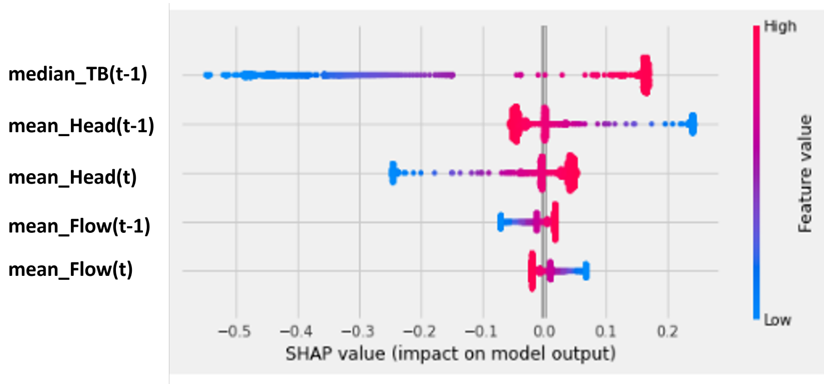
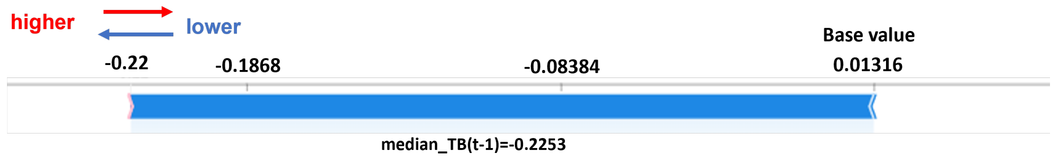


Figure 6.8: SHAP summary plot for the refuelling profile showing the most important features globally. showing the most important features globally.

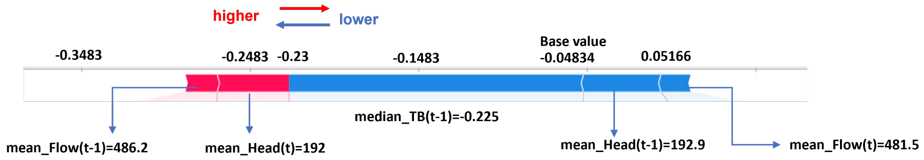
6.5.4 Generation of Human-understandable Explanations

The techniques elucidated in Section 6.2 have been employed to generate text-based explanations, a novel approach aimed at facilitating quicker and more accessible comprehension for individuals without expertise in ML. Figure 6.10a showcases text-based explanations corresponding to the normal profile, aligning with the SHAP force plot depicted in Figure 6.9a. Similarly, Figure 6.10b provides an exemplar of text-based explanations tailored for the refuelling profile, corresponding to the SHAP local explanations plot presented earlier in Figure 6.9b.

Furthermore, Figure 6.11a offers text-based explanations in line with the SHAP summary plot illustrated in Figure 6.7a, elucidating the most globally significant fea-



(a) SHAP force plot of an instance taken from the normal profile showing the most important features and their impact locally.

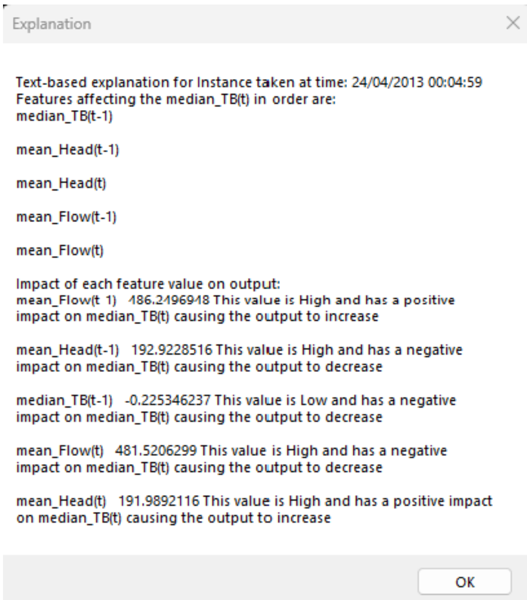


(b) SHAP force plot of an instance taken from the refuelling profile showing the most important features and their impact locally.

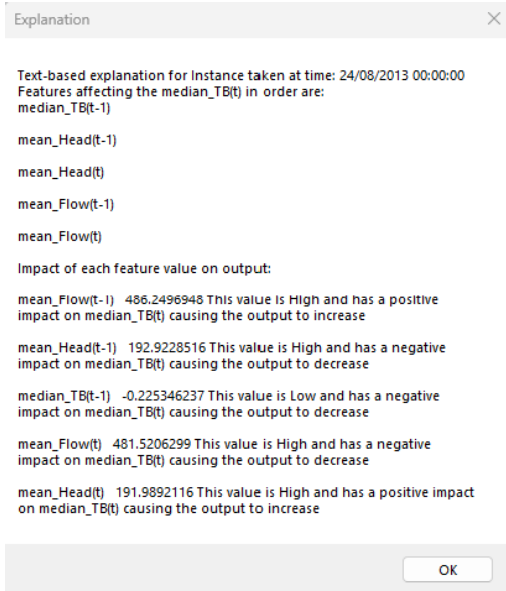
Figure 6.9: SHAP force plots of instances taken from normal and refuelling profiles.

tures within the normal profile dataset. In contrast, Figure 6.11b furnishes text-based explanations for the refuelling profile, aligning with the SHAP summary plot presented in Figure 6.7b, highlighting the pivotal features within the refuelling profile.

It is worth noting that these algorithms are not designed as standalone software applications but are intended to seamlessly integrate with existing software tools, as detailed in [302]. The referenced software tool operates by delineating the consequences of operational decisions using predictive features derived from domain expertise, specifically focusing on the relationship between pump set point positions within a pump curve (depicting the flow and head in efficiency terms) and the lateral movement of the pump shaft. This tool partitions the pump curve into sectors and captures the duration within these sectors—a crucial factor that affects wear. These sector-specific wear consequences, known to pump asset experts but not discernible in raw data, are unveiled through this tool. Importantly, this framework enhances the existing software's functionality by automatically predicting the consequences of operational decisions using raw data, thereby expanding its capabilities and facilitating the delivery of comprehensible explanations [302].

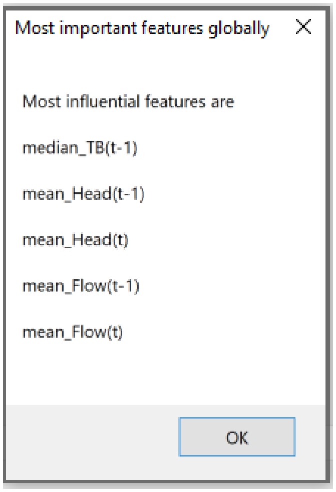


(a) Text-based explanations for SHAP force plot for normal profile shown in Figure 6.9a.

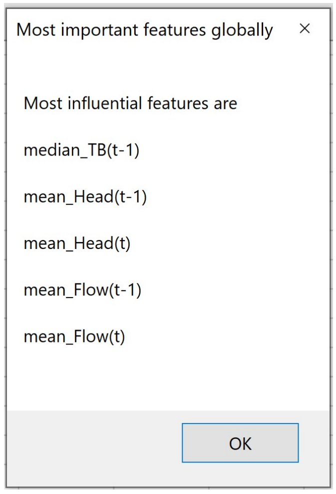


(b) Text-based explanations for SHAP force plot for refuelling profile shown in Figure 6.9b

Figure 6.10: Text-based explanations for SHAP force plot for normal and refuelling profiles.



(a) Text-based explanations for SHAP summary plot for normal profile shown in Figure 6.7



(b) Text-based explanations for SHAP summary plot for refuelling profile shown in Figure 6.8

Figure 6.11: Text-based explanations for SHAP force plot for refuelling and normal profile.

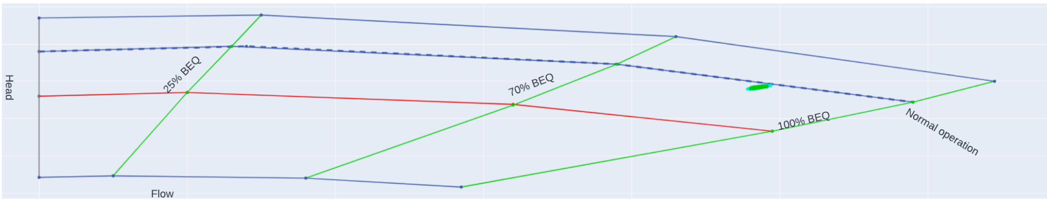
6.5.5 Graphical Representation of Explanations

This section introduces a novel algorithm specifically designed to generate visually intuitive representations that cater to non-ML experts. These graphical representations establish a causal connection with explanations generated by SHAP. The output of this algorithm superimposes a pump curve, with the primary objective of identifying safe operational regions.

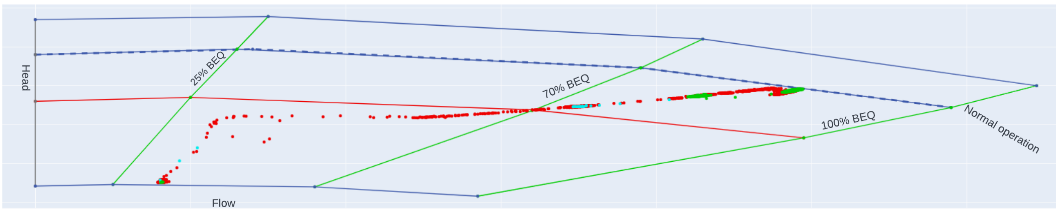
For each observation, the algorithm computes the sum of SHAP values across all features. If the resulting sum is negative, it indicates an expected decrease in output concerning the base value, which represents the average prediction within the training dataset. Conversely, if the sum is positive, it signifies an expected increase in output. To visually convey this information, the algorithm employs a colour scheme to annotate the corresponding values of flow and head. Specifically, it employs the colour red when the sum of SHAP values exceeds zero, signifying operational zones that deviate from the optimal efficiency point, resulting in reduced efficiency and increased thrust-bearing movement—undesirable outcomes that operators seek to avoid. In contrast, the colour green is employed when the sum of SHAP values is less than zero, indicating safe operational regions where the pump operates at peak efficiency with minimal degradation. Lastly, the colour cyan is used when the sum of SHAP values equals zero, denoting semi-safe operational areas (as illustrated in Figure 6.4).

The algorithm described in the previous chapter in Figure 6.4 has been applied to generate a graphical representation of the ML prediction that is familiar to a pump engineer with their related domain expertise. Figure 6.12a shows the operating state near the BEP overlaid by a schematic of the pump curve marked to show the areas of normal operations and Best Efficiency flow (BEQ). Green and cyan markers represent semi-safe and safe operational areas, respectively, as expected for data taken during normal operation, at which the pump has the highest efficiency, less thrust-bearing movement, and hence, less degradation.

In Figure 6.12b, the red markers in the graphical representation designate operating away from the BEP, which means that the pump does not operate under the optimal flow and head values and denotes operational values for head and flow that will lead to more thrust-bearing movement. Hence, more degradation operators aim to avoid. Cyan markers depict semi-safe operational areas. Upon comparing this graphical representation with the real pump curves for the assets, the graphical representation provides a causal link to the explanations provided. Hence, puts the explanations in context for a non-ML expert, helping them trust the ML algorithm's outputs more easily.



(a) Graphical representation for normal profile showing green and cyan markers denoting safe and semi-safe operational areas near the best efficiency point representing optimal conditions in terms of flow and head values, which result in higher efficiency rates and less degradation.



(b) Graphical representation for refuelling profile showing red markers operating away from the best efficiency point, resulting in less efficiency and more degradation, green and cyan markers represent operating towards the best efficiency point, which results in more efficiency and less degradation.

Figure 6.12: Graphical representations for normal and refuelling profiles overlaid onto a schematic of a pump curve.

6.6 Validating Explanations Generated from XAI Tools against Domain Knowledge

In the traditional approach to validation, experts well-versed in the domain would meticulously assess the explanations furnished by the SHAP explainability tool, ensuring their alignment with established principles. However, in the unique context of this case study, the availability of such domain experts was limited or nonexistent.

To surmount this challenge, an alternative path was sought by drawing upon existing and rigorously validated domain knowledge, particularly focusing on a set of principles known as the pump affinity laws. These laws are foundational within the realm of fluid mechanics and pump operation, providing precise insights into the inner workings of pumps and how variations in operational conditions influence their performance.

A pivotal revelation stemming from the pump affinity laws centres on the interplay between power consumption and pump efficiency. As delineated by these laws, an uptick in power consumption can serve as an indicator of inefficiency in the pump's operation. In practical terms, this suggests that when a pump consumes more power than anticipated given specific conditions, it may be functioning sub-optimally, potentially resulting in the wasteful expenditure of energy. Moreover, this inefficiency can contribute to heightened wear and tear on the pump's components, which is undesirable.

To validate the explanations generated by the SHAP tool, a thorough comparison has been conducted against the forecasts and insights derived from the pump affinity laws as denoted in Equations 6.4, 6.5 and 6.6 [308]. The assessment of whether the explanations are harmonised with the well-established principles has been conducted by scrutinising how alterations in input variables influenced the SHAP values and, subsequently, the model's predictions. In instances where the SHAP explanations indicated a notable increase in power consumption as a significant factor impacting predictions, it concurred with the tenets of domain knowledge, further affirming the validity of the generated explanations.

This methodology, rooted in a bedrock of extensively validated domain knowledge, served as a resilient mechanism for validation. It ensured that the explanations produced by the SHAP tool remained in harmony with established principles, even in the absence of direct input from domain experts. This validation framework bolstered the dependability and credibility of the generated explanations, thus fortifying their effectiveness in interpreting and comprehending the predictions of the ML model within the unique context of our case study.

$$\frac{H_2}{H_1} = \left(\frac{N_2}{N_1}\right)^2 \tag{6.4}$$

Where:

H_1 : Total head at the initial point, scalar.

H_2 : Total head at the final point, scalar.

N_1 : Initial speed or condition, scalar.

N_2 : Final speed or condition, scalar.

$$\frac{P_2}{P_1} = \left(\frac{N_2}{N_1}\right)^3 \tag{6.5}$$

Where:

P_1 : Power consumption at initial speed, scalar.

P_2 : Power consumption at final speed, scalar.

N_1 : Initial speed or condition.

N_2 : Final speed or condition.

$$\frac{Q_2}{Q_1} = \frac{N_2}{N_1} \tag{6.6}$$

Where:

Q_1 : Flow rate at initial speed, scalar.

Q_2 : Flow rate at final speed, scalar.

N_1 : Initial speed or condition.

N_2 : Final speed or condition.

6.6.1 Using Pump Affinity Laws to Predict the Performance of the Pump.

Utilising the invaluable insights provided by the pump affinity laws, a methodical process of categorising our datasets, encompassing measurements of head, flow, and thrust bearing movement, was embarked upon. As elucidated in the previous section, these principles served as our guiding compass. In particular, the pump affinity laws have been applied to discern patterns within the data that spanned head, flow, and thrust bearing movement measurements.

One pivotal observation that was gleaned from these laws was the correlation between power consumption and its potential consequences. According to the tenets outlined in equation 6.4, an increase in power consumption could signify a scenario where power is expended inefficiently, potentially leading to wastage. This inefficiency, in turn, has the potential to exacerbate wear and tear on the pump.

In essence, Pump affinity laws have been used to categorise the datasets, focusing on the measurements of head, flow, and thrust bearing movement as shown in Figure 6.13. This knowledge was leveraged to identify and interpret trends within the data, with a particular emphasis on power-related aspects. This proactive approach detected instances where elevated power consumption may indicate inefficiency and the potential for power wastage, which, as outlined in [309], can contribute to increased wear and tear in the system. The flowchart illustrates a step-by-step process for categorising a dataset based on power changes observed in a pump system. This categorisation is grounded in the principles of pump affinity laws and is aimed at identifying operational efficiency.

The process begins with the calculation of power changes across data points. This involves analysing variations in power consumption over time or between specific measurements. Once these individual power changes are determined, the next step is to compute the average power change for the dataset. This average serves as a benchmark to classify data points into distinct categories.

Each data point is then compared against the average power change. If the power change for a given data point exceeds the average, it is classified as "Red", indicating a scenario of potentially inefficient power usage. On the other hand, if the power change is less than or equal to the average, further evaluation is conducted. Specifically, if the power change is strictly less than the average, the data point is categorised as "Green", signifying efficient or normal operation. However, if the power change equals the average, the data point is classified as "Cyan", representing a neutral or balanced state.

This classification process leverages the pump affinity laws to systematically identify patterns in the data, focusing on power consumption changes. The categories—Red, Green, and Cyan—provide valuable insights into the operational efficiency of the pump system, enabling the identification of areas where power is either optimised or potentially wasted. This method not only aids in understanding pump performance but also serves as a foundation for further analysis and decision-making.

On the other hand, SHAP values generated by the SHAP explainability tool for the same dataset were utilised for a corresponding dataset categorisation. This categorisation process replicated the one predicated on the pump affinity laws, resulting in analogous categories represented by the colours red, green, and cyan. Consequently, two distinct yet comparable data frames were created, each categorised independently through different approaches.

Subsequently, a comparative analysis was conducted to evaluate the extent of alignment between the SHAP-derived categorisations and those obtained from domain knowledge (pump affinity laws). The objective was to determine how closely the SHAP-generated explanations reflected established principles, thereby assessing their reliability.

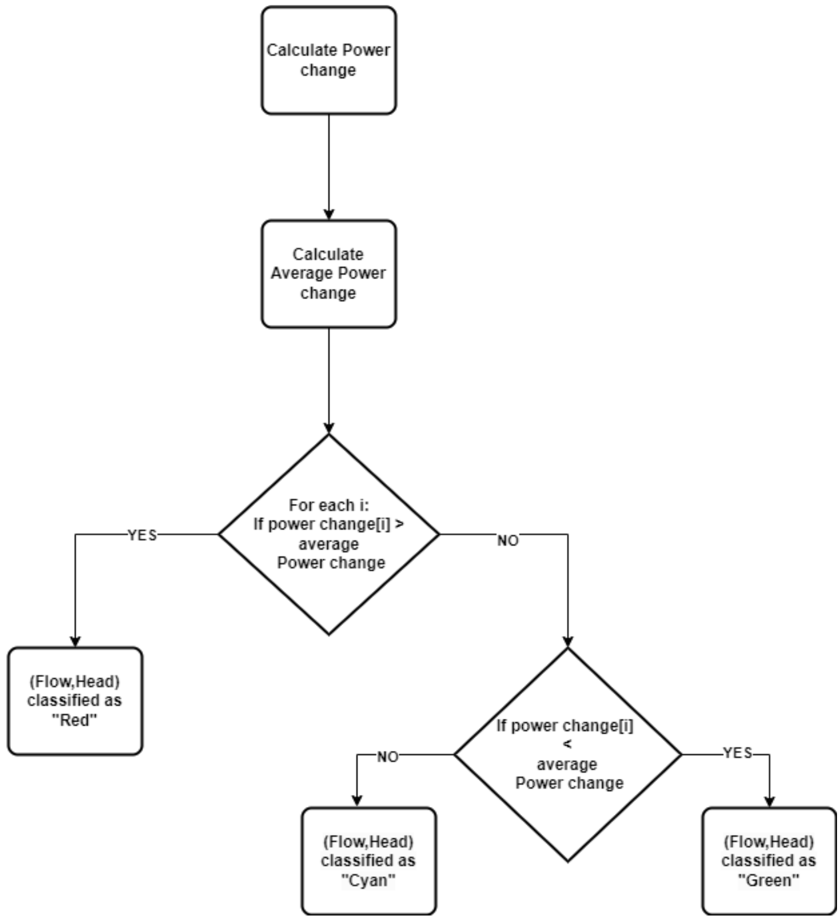


Figure 6.13: Using pump affinity laws to categorise the dataset.

To achieve this, the overall accuracy or percentage agreement was adopted as the primary measure of similarity. In this approach, each instance in the SHAP-derived dataset was compared with the corresponding instance in the domain knowledge-based dataset. A match was recorded when both datasets assigned the same category (Red, Green, Cyan), and the accuracy was then calculated as the proportion of matching instances relative to the total number of instances:

$$\text{Accuracy} = \frac{1}{N} \sum_{i=1}^{N} \delta(y_i^{\text{SHAP}},\, y_i^{\text{DK}}) \tag{6.7}$$

where:

- N is the total number of instances,

- y_i^{SHAP} is the category assigned by the SHAP-based method for instance i,
- y_i^{DK} is the category assigned by the domain knowledge-based method for instance i,
- $\delta(a, b)$ is an indicator function defined as:

$$\delta(a, b) = \begin{cases} 1, & \text{if } a = b \\ 0, & \text{otherwise} \end{cases}$$

This measure provides an intuitive and interpretable indication of agreement between the two categorisation methods, capturing the proportion of cases in which SHAP explanations and domain knowledge were consistent.

Similarity Measure	**Score**
Overall Accuracy (Percentage of Agreement)	0.712

Table 6.3: Agreement between SHAP-based and domain knowledge-based categorisations

The resulting accuracy score of 0.712 indicated a moderate-to-high level of agreement, reflecting substantial correspondence between the SHAP explanations and the classifications derived from the pump affinity laws, while also highlighting areas where the two approaches diverged.

6.6.2 How to Justify Similarity Measure

The recognition of the discrepancy between SHAP explanations and our findings underscores the need for further investigation to elucidate and justify this divergence. To address this, one plausible approach is to apply the same analytical framework to publicly available datasets that are well-documented with established domain knowledge. This comparative analysis would allow us to assess the viability and consistency of our approach in a broader context. Through examination of how our approach performs on datasets with established ground truth and domain principles, valuable insights into the robustness and generalizability of our methodology can be gained. This step is crucial in ensuring the reliability and credibility of these research findings and will contribute to a more comprehensive understanding of the applicability of SHAP explanations in various domains.

6.6.3 Addressing Uncertainty and Limitations

SHAP is a valuable tool for interpreting ML model predictions, but it is not without limitations. One significant limitation lies in its potential to contribute to discrepancies observed in case studies, such as those involving thrust bearing wear analysis. These discrepancies may arise due to SHAP's inherent assumption of equal attribution among features, which can oversimplify complex real-world relationships. This oversimplification is a primary contributing factor to the disconnect between domain knowledge and SHAP-based explanations in such scenarios.

In their study, [310] contend that SHAP values exhibit a noteworthy limitation stemming from their inherent disregard for the intricate causal structures inherent in the data. Specifically, this framework allocates equal attribution weight to all features within the model explanation, mandating that attributions be uniformly distributed across equally informative features.

The absence of a causal structure within the SHAP values framework can pose a significant concern, potentially resulting in erroneous decisions in specific use cases. This risk arises when attributing significance to a feature with substantial SHAP values, as these values may primarily reflect the influence of other interconnected features rather than the feature itself.

This limitation highlights a potential challenge in SHAP's ability to accurately reflect the underlying causal relationships within the data, particularly when certain features are intricately interdependent.

In this work, the validation of SHAP explanations within the context of real-world condition monitoring case studies has been undertaken, and this validation has been extended to encompass various domains of knowledge. The assessment of the similarity measure yielded a result of approximately 0.712. In this section, the factors contributing to this similarity measure are shown and elaborate on the reasons for its relatively modest value.

One notable aspect influencing the similarity measure is the inherent nature of the SHAP tool itself. SHAP explanations, while insightful and valuable, may introduce a level of dissimilarity due to their inherent properties and mechanisms. Understanding these inherent limitations is crucial in interpreting the similarity measure.

Furthermore, the quality of the data used in the case studies plays a pivotal role in determining the similarity measure. The data used in real-world condition monitoring scenarios is often characterised by its complexity and may contain various forms of noise, anomalies, and missing values. The "noisiness" of this data makes it significantly challenging to achieve high similarity measures. These data imperfections can result in inconsistencies in the SHAP explanations, leading to the observed lack of similarity.

One approach to overcoming the challenges of noisy data in real-world condition monitoring is through the generation of synthetic data using GAN [44]. GANs can create realistic synthetic samples that replicate the patterns found in original data while minimising noise and inconsistencies. By incorporating these high-quality synthetic samples into the dataset, GANs can help balance the data and reduce the influence of noise, leading to more robust and reliable SHAP-based explanations. This approach enhances the dataset's overall quality, supporting more accurate and consistent model interpretations, especially in complex, noisy environments.

To address sources of uncertainty and enhance the reliability of SHAP explanations, an analysis is conducted using the well-known Iris dataset. This dataset serves as a controlled environment, allowing for a focused examination of SHAP's effectiveness under ideal, well-structured conditions. By confirming SHAP's expected performance in this simplified context, the analysis establishes a baseline for its interpretability and reliability, helping to isolate SHAP's strengths and limitations before applying it to more complex, noisy datasets.

This initial benchmarking aligns with a large body of SHAP-related literature but adds a distinct perspective by explicitly demonstrating how SHAP explanations function within a low-uncertainty, well-understood dataset. Observing SHAP's performance in such a controlled setting provides valuable insight into its stability, allowing for a meaningful comparison when extending the analysis to real-world, condition monitoring data, where challenges like noise, feature interdependencies, and data imperfections are prevalent.

In this way, the analysis not only reaffirms SHAP's reliability in ideal conditions but also builds a foundation for exploring modifications (e.g., WeightedSHAP) and techniques that enhance its applicability and robustness in more complex scenarios, which is critical in industrial applications.

In summary, the similarity measure of 0.712 observed in the validation of SHAP explanations within real-world condition monitoring case studies reflects both the limitations of the SHAP framework and the challenges posed by noisy, real-world data. To improve the reliability of SHAP explanations, controlled datasets, such as the Iris dataset, have been used to establish a benchmark under ideal conditions, providing insight into SHAP's baseline performance.

Using IRIS data to Verify SHAP Explanations

To ensure credibility and reliability in the validation process, the Iris dataset was deliberately selected as the foundation for establishing a solid baseline. By using a well-understood, controlled dataset, the baseline allows for a clear evaluation of SHAP's performance under ideal conditions. This helps differentiate between limitations inherent to the SHAP framework and those introduced by data complexity. Such a baseline provides a critical reference point, enhancing the robustness of subsequent validations on more complex, real-world datasets, where noise and feature interdependencies pose greater challenges.

To illustrate the methodology applied to the IRIS dataset, Figure 6.14 showcases the approach undertaken. The Iris dataset holds a distinguished status in the field of data analysis and ML, recognised as a benchmark dataset with well-defined characteristics. It encompasses samples derived from three distinct species of iris flowers, and in total, it comprises 150 instances.

What lends added confidence to our validation effort is the wealth of established

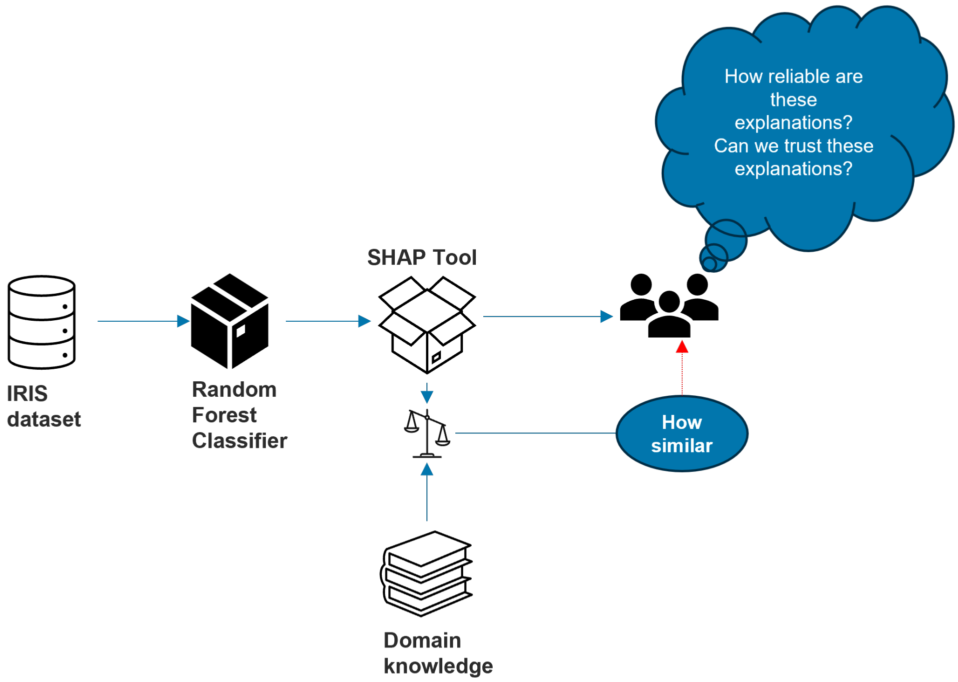


Figure 6.14: Methodology used to verify SHAP explanations using the IRIS dataset.

knowledge and research that has been conducted on the Iris dataset over the years. Numerous studies, as referenced in [311], have thoroughly explored and documented the characteristics and patterns within this dataset. This extensive body of prior research provides us with a well-structured foundation for our validation process.

Insights and results from these established sources are drawn upon to effectively benchmark our validation procedures against widely accepted standards. By this approach, the credibility of our validation is enhanced, and it is ensured that our results can be rigorously compared and validated within the broader context of the Iris dataset—a dataset renowned for its reliability and extensive analysis history.

Initially, as part of the data pre-processing phase aimed at achieving a deeper understanding and uncovering potential correlations, data visualisation was conducted. In this process, the data was visualised using the sepal width and sepal length attributes as shown in Figure 6.15. However, as shown in Figure 6.15a, it was observed that the separation between the different iris species might not have been as distinct as anticipated. This observation suggests that the influence of the sepal features alone in effectively distinguishing between the various iris species may not have been particularly pronounced.

In essence, despite the exploration of the sepal width and sepal length attributes as potential discriminators, it was inferred that other factors or features may have played a more substantial role in the accurate differentiation of iris species. The significance of considering a broader set of attributes or features in the analysis to ensure precise species classification is underscored by this insight.

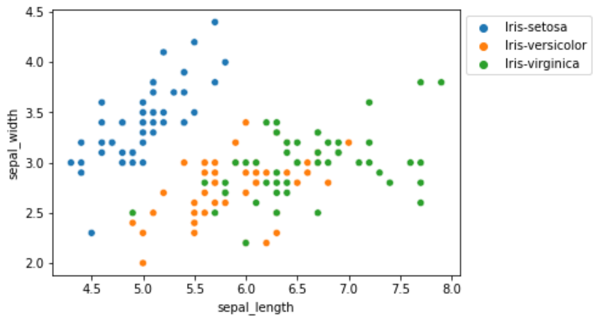
On the contrary, when visualising the data based on petal width and petal length as shown in Figure 6.15b, distinct clusters for each species can be clearly discerned. This serves as a compelling illustration that these two features possess a high level of discriminative power when it comes to accurately classifying the various types of iris flowers.

Following that, the Iris dataset was subjected to classification using a Random Forest classifier, where instances were classified into one of three classes: Setosa, Versicolor and Virginica. This ML algorithm, renowned for its versatility and robustness, was employed to discern and allocate each instance to its respective class based on the dataset's features.

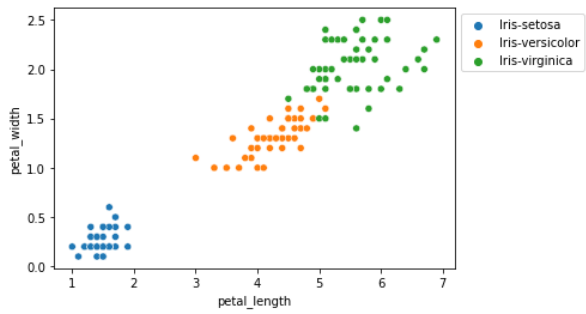
The Random Forest classifier works by building multiple decision trees, where each tree plays a role in making classification decisions. By combining the predictions from these individual trees, the classifier can deliver a robust and accurate classification result.

By employing this classifier on the Iris dataset, the aim was to leverage its capabilities in effectively distinguishing and classifying instances into the three distinct categories. This classification process relied on the underlying patterns and relationships within the dataset, allowing informed and precise decisions regarding class assignments to be made for each instance.

The feature importance analysis conducted using the Random Forest Classifier yielded valuable insights into the discriminative power of each feature within the Iris dataset. Among these features, petal length and petal width emerged as the most influential factors in classifying iris species, as shown in Table 6.4, both carrying substantial



(a) Scatter plot of Iris dataset based on sepal values.



(b) Scatter plot of Iris dataset based on petal value

Figure 6.15: Comparison of scatter plots based on sepal length and petal length in the Iris dataset. The plot (a) shows the distribution of data points based on sepal length, while plot (b) shows the distribution based on petal length. The significant separation of data points in plot (b) suggests that petal length may be a more important feature in distinguishing between Iris species.

importance values of 0.436130 and 0.436065, respectively. These high importance scores underscore the pivotal role of petal characteristics in the classifier's decision-making process.

In addition to petal attributes, sepal length was also identified as a relevant discriminative feature, albeit with a lower importance value of 0.106128. This indicates that variations in sepal length contribute significantly to species classification, although to a lesser extent compared to petal attributes.

Conversely, sepal width exhibited the lowest feature importance among the four attributes, with a score of 0.021678. While its influence is relatively minor in comparison to the other features, it still plays a role in the classification process.

Collectively, these feature importance values provide a clear hierarchy of the features' contributions to accurate species classification within the Iris dataset. The findings highlight the significance of petal characteristics, especially petal length and width, in differentiating iris species, offering valuable insights for further analysis and model interpretation.

Utilising the SHAP explainability tool in conjunction with the Random Forest classifier, a comprehensive analysis of feature importance was conducted. The outcome of this analysis is encapsulated in the SHAP summary plot in Figure6.16, as illustrated in

Feature	Importance
Petal Length (cm)	0.436130
Petal Width (cm)	0.436065
Sepal Length (cm)	0.106128
Sepal Width (cm)	0.021678

Table 6.4: Feature Importance by Random Forest Classifier

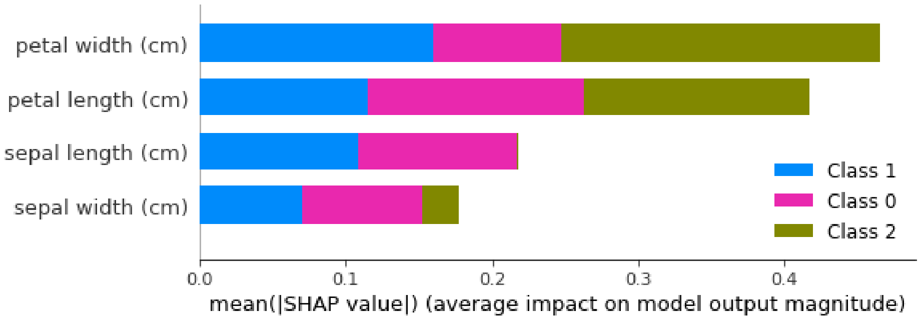


Figure 6.16: SHAP summary plot to show the most important features.

the accompanying Figure. This plot succinctly underscores the paramount significance of certain features, with petal width and petal length emerging as the most influential. Following closely in importance are sepal length and sepal width, in that order. Remarkably, the Random Forest classifier consistently and persistently identifies petal width and petal length as the preeminent features in the classification process, as shown in Table 6.4. This consistent recognition serves to underscore the pivotal role that these attributes play in effectively discerning between various iris species. In essence, these findings robustly reinforce the assertion that petal width and petal length are of paramount importance when it comes to classifying iris species, as consistently demonstrated by the Random Forest classifier.

SHAP, in addition to its role in feature importance analysis, offers valuable insights into the individual contributions of features to model predictions. In the context of this analysis, as exemplified in the provided figure, SHAP elucidates the specific contributions of each feature to predictions as shown in Figure 6.17. In this case, it unequivocally affirms the profound significance of petal characteristics in effectively distinguishing between various types of iris flowers.

Figure 6.17: SHAP force plot for one instance in the IRIS dataset showing the most important features and their impact.

The figure vividly illustrates how petal characteristics, notably petal width and petal length, exert a substantial influence on shaping the model's predictions. These insights reinforce the understanding that variations in petal attributes serve as key discriminators in accurately classifying different iris species. SHAP's ability to provide such granular insights enhances our comprehension of feature contributions within the predictive model and underscores the pivotal role of petal characteristics in the classification process.

After receiving SHAP explanations, domain knowledge rules were derived from established references, delineating feature thresholds within the context of various species. These rules encompassed feature thresholds pertinent to each species. Subsequently, SHAP value extraction was carried out by applying a trained Random Forest classification model to the Iris dataset, yielding SHAP values for a specific instance. The ensuing analysis of SHAP values for each feature within the given instance unveiled that positive SHAP values signified a feature's affirmative contribution to the prediction, whereas negative SHAP values denoted a detrimental contribution. A crucial step in this process involved comparing the effects of SHAP values on the prediction with the predefined domain knowledge rules. This comparison elucidated whether the impact of each feature, as indicated by the SHAP values, concurred with the predictions made by the established domain knowledge.

In order to establish a robust baseline for measuring the alignment between SHAP explanations and domain knowledge, the same methodology was applied to a controlled dataset. The objective of this approach was to determine a satisfactory level of similarity accuracy between SHAP explanations and domain knowledge, which not only enhances our methodological confidence but also furnishes a valuable benchmark for assessing the degree of disparity observed within the original industrial context.

In analysing SHAP values, it is crucial to interpret their implications accurately. Figure 6.18 shows how shap values are verified against rules extracted from domain knowledge taken from [311]. Positive SHAP values for specific features indicate that the presence of those features positively contributes to the likelihood of predicting the corresponding class. For instance, a positive SHAP value associated with the "petal width" feature in the "Setosa" class suggests that higher values of petal width increase the probability of the model predicting the class as "Setosa." Conversely, negative SHAP values for features imply a negative contribution to class prediction likelihood. If a negative SHAP value is linked to the "sepal length" feature in the "Virginica" class, it signifies that lower sepal length values elevate the probability of the model predicting the class as "Virginica." Additionally, the magnitude of SHAP values reflects the extent of a feature's impact on the prediction. Larger magnitudes indicate stronger influence, where a high positive SHAP value signifies a significant positive effect on class prediction, while a high negative value denotes a substantial negative impact. These insights are instrumental in comparing SHAP explanations with domain knowledge to assess the congruence of their findings.

In Figure 6.19, a sample output from verifying SHAP values against domain knowledge rules. For each data instance in our analysis, SHAP values are computed to understand the impact of individual features on the model's prediction. Specifically, the focus is placed on the two most critical features, petal length and petal width, which have consistently been found to be the most influential in our context.

Firstly, an assessment is made to determine whether the SHAP values for these features in a given instance have a significant positive effect on the predicted output. In other words, it is checked whether higher values of petal length and petal width contribute positively to the likelihood of the model predicting a specific class. This step assists in identifying which features are deemed crucial for the model's decision for that particular instance.

Once the impact of the SHAP values has been established, the next step involves comparing these findings with our domain knowledge rules, which define specific boundaries or thresholds for petal length and petal width based on established references. These rules encapsulate our understanding of how these features should behave in relation to the predicted class.

The crux of this analysis lies in determining whether the observed impact of petal length and petal width aligns with what our domain knowledge rules predict. If the SHAP values indicate that higher values of these features positively influence the prediction, and our domain knowledge rules are in agreement with this observation, it strengthens our confidence in both the model and the domain knowledge. Conversely, if there is a misalignment, it prompts further investigation into potential discrepancies or areas for improvement.

In summary, this analytical process involves the computation of SHAP values to assess the impact of crucial features, a comparison of these values with domain knowledge rules, and an evaluation of the degree of alignment between the two. This helps us validate the model's behaviour and gain insights into the consistency between data-driven predictions and established domain expertise.

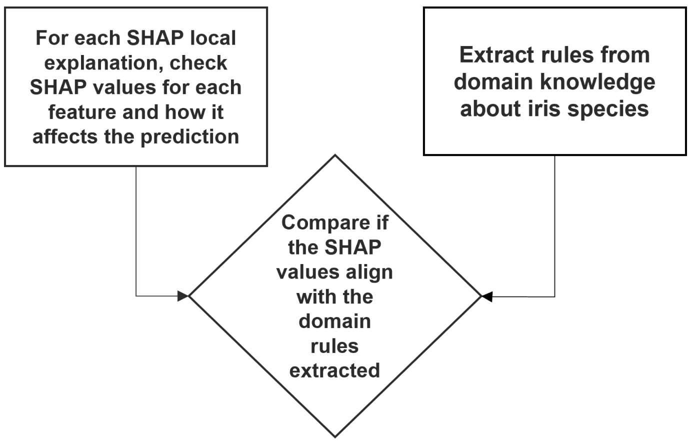


Figure 6.18: How to verify SHAP values against rules extracted from domain knowledge.

```
Instance 1:
Predicted Iris type: versicolor
Domain Knowledge: Petal Length: (2, 5)
Domain Knowledge: Petal Width: (0.5, 2)
SHAP Explanations: Petal Length SHAP Value: 0.3173
SHAP Explanations: Petal Width SHAP Value: 0.3003
Aligned with Domain Knowledge: Yes
---

Instance 2:
Predicted Iris type: setosa
Domain Knowledge: Petal Length: (1, 2)
Domain Knowledge: Petal Width: (0.1, 0.7)
SHAP Explanations: Petal Length SHAP Value: 0.3247
SHAP Explanations: Petal Width SHAP Value: 0.3336
Aligned with Domain Knowledge: Yes
---
```

Figure 6.19: Sample of output to verify SHAP values against domain knowledge rules.

In this approach, the comparison between SHAP explanations and domain knowledge was structured around a set of predetermined rules widely accepted within the domain. These rules serve as a clear representation of expert insights, specifying how various features should impact outcomes. Using these rules as a benchmark, each SHAP explanation was systematically compared against the domain knowledge to evaluate its accuracy.

The process entailed an examination of SHAP outputs to determine whether they adhered to or deviated from these predefined rules. A SHAP explanation was considered 'right' when it aligned with the expectations set forth by these rules, indicating a correct identification of feature impacts. Conversely, explanations were deemed 'not right' whenever there was a divergence from these established guidelines.

The similarity measure was thus calculated by tallying instances of agreement (where SHAP outputs conformed to the domain knowledge rules) versus instances of disagreement. In this analysis, the objective was to evaluate the alignment between the classifications generated by SHAP explanations and domain knowledge rules for the Iris dataset. A step-by-step process was followed:

Dataset and Model

The Iris dataset consists of 150 data points, each describing an Iris flower using features such as sepal length, sepal width, petal length, and petal width. The dataset includes three classes: Iris setosa, Iris versicolor, and Iris virginica. The dataset was divided into training and testing subsets:

- **Training set**: 80% of the data (120 data points) was used to train a Random Forest Classifier.
- **Test set**: 20% of the data (30 data points) was used to evaluate the model and conduct similarity analysis.

Domain Knowledge Rules

Domain knowledge was defined in terms of expected ranges for petal length and width for each class:

- **Iris setosa**: Petal length in $[1, 2]$, petal width in $[0.1, 0.7]$.
- **Iris versicolor**: Petal length in $[2, 5]$, petal width in $[0.5, 2]$.
- **Iris virginica**: Petal length in $[3, 7]$, petal width in $[1, 3]$.

Similarity Metric: Overall Accuracy (Percentage of Agreement)

To evaluate the alignment between SHAP-generated explanations and domain knowledge, overall accuracy (also referred to as percentage of agreement) was employed. This metric measures the proportion of instances where both categorisation methods produced the same class label. Formally, it is defined as:

$$\text{Accuracy} = \frac{\text{Number of matching instances}}{\text{Total number of instances}}$$

where:

- Matching instances: the number of data points for which SHAP and domain knowledge assigned the same class label,

- Total number of instances: the total number of data points evaluated.

Results

In the Iris test set of 30 data points, SHAP explanations were compared to domain knowledge rules on an instance-by-instance basis. The resulting overall accuracy (percentage of agreement) indicated a very high level of concordance:

Similarity Measure	**Score**
Overall Accuracy (percentage of agreement)	0.97

Table 6.5: Overall accuracy (percentage of agreement) between SHAP explanations and domain knowledge for the Iris dataset.

This high accuracy score of 0.97 demonstrates strong alignment between SHAP classifications and domain knowledge. Such a high similarity measure reinforces the validity of the SHAP explanations, showcasing their ability to accurately reflect the nuanced interplay of features as encapsulated by domain-specific rules. This methodical comparison underscores the robustness of our model in generating explanations that are not only interpretable but also deeply rooted in domain expertise.

6.7 Conclusions

Our approach was initially validated against existing domain knowledge due to the unavailability of domain experts. This validation yielded a similarity measure of 71.2%, indicating a moderate to high alignment between the SHAP explanations and the existing domain knowledge. The observed dissimilarity between SHAP explanations and domain knowledge could be attributed to several factors, with the complexities and 'noisiness' of real-world data being a primary consideration. Real-world data often comes with inherent challenges, such as noise, missing values, outliers, and biases, which can significantly impact the interpretability and reliability of ML models. These data imperfections can lead to a mismatch between the model's explanations and the human expert's expectations or understanding.

Furthermore, the SHAP method, while powerful, may have limitations in capturing and reflecting the multifaceted and nuanced nature of real-world data, potentially leading to discrepancies in explanations. Recognising and addressing these challenges is crucial in advancing the field of XAI, ensuring that the explanations generated are not only technically accurate but also meaningful and reliable in the complex and often imperfect context of real-world applications.

To understand and address the discrepancies observed, our framework was applied to the Iris dataset, an open-source dataset well-known in the ML community. Here, a 97% similarity measure was achieved, suggesting a high degree of alignment between SHAP explanations and the dataset's inherent characteristics.

The moderate to high similarity measure in our initial validation highlights the complexities involved in aligning machine-generated explanations with human domain knowledge. This discrepancy could stem from several factors, including the inherent limitations of the SHAP method, the complexity of the domain knowledge, or the nuances of the dataset used.

Moreover, the high similarity measure with the Iris dataset underscores the effectiveness of our framework in cases where the dataset characteristics are well-defined and less complex. However, it also brings to light the challenge of applying such techniques to more complex, real-world scenarios where domain knowledge is not as straightforward.

Moving forward, it is imperative to engage directly with domain experts to further validate and refine our explanations. Their insights would be invaluable in bridging the gap between ML explanations and practical, real-world applicability. Additionally, exploring other explanation methods and comparing their effectiveness against SHAP within our framework could provide deeper insights into the best practices for XAI.

In conclusion, our study demonstrates the potential of using SHAP-based explanations to make ML models more interpretable. However, it also highlights the challenges and complexities inherent in this endeavour, especially when dealing with intricate domain knowledge. Continuous collaboration with domain experts and further research into various explanation methods are essential steps towards achieving more reliable and practical AI interpretability.

Chapter 7

Conclusions and Future Work

7.1 Conclusions

Throughout the course of this research, an in-depth exploration of various explainability techniques has been conducted, going into great detail about how they function, their strengths, and their limitations. It's important to emphasise that in practical applications, explainability is not an isolated concept; rather, it should be seen as an integral part of a continually expanding set of tools for ML analysis.

This set of tools extends far beyond the scope of post hoc model analysis and offers a valuable feature known as XAI. XAI can seamlessly integrate into the complete ML life cycle, which encompasses a broad spectrum of activities. This integration encompasses everything from the initial stages of data collection and preprocessing, through the refinement and improvement of datasets, to model development, deployment, and the ongoing monitoring of ML systems.

In essence, XAI is a versatile and indispensable resource that can enhance the transparency, interpretability, and trustworthiness of the entire ML process, ensuring that it is not limited to a single phase but is an integral part of the end-to-end journey from data to actionable insights. This research began with the exploration of various XAI tools, with the primary objective of identifying the most robust and dependable ones to guide subsequent stages of research. During this initial phase, a set of considerations was meticulously crafted to serve as guiding principles when designing XAI systems.

As progress was made, the transition occurred into the practical phase of the research, where these considerations were put into action. A series of case studies was undertaken, spanning from open-source datasets to real-world applications. Throughout these endeavours, a recurring challenge was encountered - the XAI tools employed provided highly technical explanations, which, though informative, remained largely inaccessible to non-technical users. Recognising that the ultimate goal of XAI is to facilitate understanding of ML models, it was realised that presenting these technical explanations in their raw form would not suffice.

To bridge this gap, a mission was embarked upon to make these explanations more user-friendly. Novel algorithms were developed to transform the output of XAI tools, particularly SHAP, into text-based explanations that could be readily comprehended by anyone. This transformation marked a pivotal turning point in the journey.

Building on this foundation, a novel framework was then constructed, applicable across various engineering domains. This framework is aimed at providing non-expert users with a deeper understanding of ML applications within specific industrial domains. The framework focuses on generating auto-generated explanations in both text-based and visual formats. To achieve this, the widely recognised SHAP explainability tool was employed, which inherently provides both local and global explanations. It's worth noting that most currently available ML explanation frameworks offer only one level of explainability, whereas this framework introduces two levels. The first level offers text-based explanations, while the second level provides a graphical representation of these explanations.

Despite the limitation of not having immediate access to domain experts for validation, solid domain knowledge was harnessed to rigorously evaluate and validate these explanations. This research was guided by a relentless pursuit of transparency, understandability, and trustworthiness in XAI.

These explanations were further cross-referenced with domain knowledge to establish a degree of causation for ML predictions, catering to a wide audience. However, commitment to ensuring the trustworthiness and comprehensibility of these explanations led to another crucial step - validation.

One significant advantage of this XAI framework, designed to provide user-friendly explanations, is that it can easily be integrated into the V-cycle of ML software development. By aligning this framework with established software engineering standards, its practical applicability throughout the development process is demonstrated. It aids in feature selection, model development, testing, validation, deployment, and maintenance phases, thereby promoting transparency and accountability at every stage.

However, it's worth noting that even though SHAP offers more reliable explanations compared to many XAI tools, it is not immune to adversarial attacks. This vulnerability, where adversarial attacks can manipulate the local approximations of the black-box by perturbing the neighbourhood of the explained data point, is an area acknowledged as a potential concern for future work. Addressing these adversarial challenges is essential for enhancing the robustness of explanation techniques like SHAP and ensuring their reliability in various scenarios.

This framework represents a significant step toward providing user-friendly explanations, and ongoing work will continue to refine and improve it, addressing potential limitations and challenges in the field of XAI.

In conclusion, not only has exploration and harnessing of XAI tools been undertaken, but also paved the way for a more user-centric, accessible, and dependable approach to explaining ML models. This novel framework, born from this journey, stands as a testament to the commitment to making the complex world of ML understandable and trustworthy for all. Despite some explanation approaches offering theoretical assurances based on specific criteria, practical implementation may sometimes challenge these guarantees. In contrast, this approach has shown practical effectiveness, reaffirming its reliability and performance in real-world scenarios.

7.2 Outcome of Research

This research has produced several key outcomes and contributions, which are summarised below:

- **Comprehensive Exploration of Explainability Techniques:** This research provides an in-depth examination of various explainability techniques used in ML, including their functions, strengths, and limitations. It serves as a valuable resource for understanding the landscape of XAI tools.

- **Considerations for Designing XAI Systems:** The research establishes a set of considerations that serve as guiding principles for designing XAI systems. These principles serve as a foundation for implementing XAI effectively and ensuring transparency and trustworthiness.

- **Transformation of Technical Explanations:** To make XAI more accessible, the research develops novel algorithms capable of transforming highly technical explanations generated by XAI tools, such as SHAP, into user-friendly text-based explanations. This transformation enhances the comprehensibility of XAI outputs.

- **Development of a novel Framework:** A novel XAI framework is introduced, designed for use across various engineering domains. This framework generates auto-generated explanations in both text-based and visual-based formats, offering two levels of explainability. It facilitates a deeper understanding of ML applications for non-expert users.

- **Recognition of limitations:** The study recognises weaknesses in XAI tools, especially regarding causality and susceptibility to adversarial attacks. It emphasises the importance of tackling these weaknesses to improve the robustness and dependability of explanatory methods such as SHAP.

- **Practical Effectiveness:** The XAI framework developed in this research showcases practical effectiveness in real-world scenarios. It bridges the gap between complex ML models and non-technical users, reaffirming its reliability and performance.

- **Commitment to Transparency and Trustworthiness:** Throughout the research, there is a consistent commitment to transparency, understandability, and trustworthiness in XAI. Explanations are rigorously cross-referenced with domain knowledge to establish causation for ML predictions.

- **Future directions:** The research identifies future directions, including the need for further exploration of vulnerabilities in XAI tools and ongoing refinement of the developed framework. It emphasises the ongoing commitment to making XAI accessible and dependable.

7.3 Future Work

This research can be taken in several directions, one of which involves a comprehensive exploration into the vulnerabilities of SHAP and other perturbation-based XAI methods to adversarial attacks. This path is particularly intriguing as it delves deep into the intricate dynamics between data models and the adversarial elements that seek to exploit them. The journey begins with a thorough analysis aimed at understanding the underlying reasons these methods are susceptible to such manipulations. It's not just about identifying the vulnerabilities but also about dissecting the nature of the adversarial attacks, shedding light on the intrinsic properties of the data and the models that make them prone to these threats.

A pivotal aspect of this research direction is the development and implementation of robustness metrics. These metrics are crucial as they provide a means to measure the vulnerability of these explanation techniques quantitatively and offer a benchmark for tracking improvements over time. It's not just about understanding the problem but also about quantifying it and charting the progress in mitigating it.

Adding another dimension to this research is the involvement of domain experts in validating XAI explanations. This collaboration is crucial as these explanations are ultimately intended for these experts. Their insights can significantly enhance the relevance and utility of the explanations, ensuring they are not just technically sound but also practically meaningful and actionable. Domain experts can provide valuable feedback, helping to refine the explanations and tailor them more effectively to specific needs and contexts. Their involvement ensures that the developed solutions are grounded in real-world applicability, enhancing the interpretability and trustworthiness of AI systems in practice. While SHAP provides useful post-hoc attributions, its standard formulation gives equal weight to feature coalitions/permutations and does not encode causal structure. In datasets with strong interdependencies, this can yield explanations that diverge from domain knowledge [148].

To mitigate these limitations, future work will investigate WeightedSHAP [312], which relaxes the equal-coalition assumption by allowing unequal attribution weights informed by feature importance or prior knowledge, thereby better reflecting plausible causal mechanisms. Applying WeightedSHAP in complex industrial settings may improve the faithfulness and interpretability of explanations, particularly where precise assessments of feature importance are critical. Further enriching this research trajectory is the validation of XAI explanations against user requirements and standard regulations through the application of NLP techniques. This direction recognises the significance of aligning AI explanations with the linguistic and conceptual frameworks of the end-users, including adherence to prevailing regulatory standards. By harnessing NLP, the research aims to automate the process of evaluating explanations for compliance with user expectations and legal guidelines, thereby ensuring that the AI systems are not only transparent and interpretable but also aligned with ethical norms and regulatory requirements. This validation process is crucial in building trust and ensuring that AI systems are deployed responsibly and effectively in various domains.

Many XAI tools in contemporary applications primarily focus on identifying correlations within data, rather than establishing causality. This prevalent approach presents a notable drawback that warrants careful consideration and further exploration. In this research, this challenge was addressed by introducing a novel approach within the context of thrust bearing wear. Visual-based explanations were integrated into our framework, aiming to bridge the gap between correlation and causality in this specific case study.

Looking ahead to the future of XAI, it is imperative to move beyond mere correlation and focus on providing explanations that showcase cause-and-effect relationships [313]. To achieve this, XAI must incorporate causal relationships, offering domain experts a deeper understanding of the fundamental mechanisms guiding AI-driven decisions. One significant advantage of integrating causality into XAI is the potential to uncover and address biases effectively [314].

In addition to the aforementioned research directions, a third level of explainability is proposed by constructing knowledge graphs from text-based explanations generated by XAI methods. This innovative approach leverages natural language processing techniques to transform narrative explanations into structured, queryable knowledge graphs. By doing so, it opens up new avenues for users to interact with and understand AI decisions. Users can query these knowledge graphs to uncover relationships, probe the reasoning behind AI decisions, and gain insights that are not immediately apparent from linear explanations. This method not only enhances transparency but also empowers users, allowing them to engage with AI systems more effectively and ensuring that the explanations provided align closely with their specific information needs and cognitive styles.

Appendix A

Overview of Explainable AI Tools and Their Characteristics

Table A.1: Overview of Explainable AI Tools and Their Characteristics

Name	Year	Features	Explanation Technique	Scope	Input Type	Output Type
SHAP [56]	2017	Unified feature importance for global/local explanations	Perturbation-based	Global/ Local	Any	Visual, Numerical
LIME [57]	2016	Local explanation using perturbed samples with surrogate models	Perturbation-based	Local	Any	Visual
LRP [106]	2015	Relevance propagation across model layers	Backpropagation	Local	Text, Image, Tabular	Heatmap, Numerical
DALEX [159]	2018	Tool for model exploration and diagnostics	Feature Importance	Local/ Global	Tabular	Visual, Numerical
ANCHOR [154]	2018	Rule-based conditions for predictions	Rule-Based	Local/ Global	Text, Tabular	Rules, Visual

Continued on next page

Name	Year	Features	Explanation Technique	Scope	Input Type	Output Type
RuleFit [161]	2008	Combines regression with rule-based decision trees	Rule-Based	Global/ Local	Tabular	Rules, Coefficients
SBRL [198]	2017	Scalable rule-based lists for interpretability	Rule-Based	Global/ Local	Tabular	Decision Rules
GLocalX [164]	2019	Prototype-based instance selection explanations	Prototype-Based	Local/ Global	Tabular	Prototypes
MMD-CRITIC [166]	2016	Prototype-based method for representative samples	Prototype-Based	Global	Tabular	Representative Instances

Continued on next page

Name	Year	Features	Explanation Technique	Scope	Input Type	Output Type
PROTODASH [165]	2017	Fast prototypical instance selection	Prototype-Based	Local	Tabular	Representative Instances
CEM [167]	2018	Contrastive explanations with positives and negatives	Example-based	Local	Any	Contrastive Examples
DiCE [168]	2020	Diverse counterfactuals for multiple outcomes	Counterfactual	Local	Any	Diverse Counterfactuals
FACE [169]	2020	Feasible, actionable counterfactuals	Counterfactual	Local	Any	Actionable Counterfactuals
Integrated Gradients [172]	2017	Gradient integration for attribution	Gradient-Based	Local	Text, Image	Heatmap, Numerical

Continued on next page

Name	**Year**	**Features**	**Explanation Technique**	**Scope**	**Input Type**	**Output Type**
PM [41]	2001	Permutation-based feature importance	Feature Importance	Global	Tabular	Numerical
GRAD-CAM [124]	2017	Highlights influential regions using gradients	Gradient-Based	Local	Image, Video	Heatmap
SmoothGrad [173]	2017	Adds noise to reduce saliency map noise	Gradient-Based	Local	Image	Heatmap
Saliency Maps [125]	2013	Assigns pixel importance via gradients	Gradient-Based	Local	Image	Heatmap
Guided Backpropagation [175]	2014	Highlights key input pixels with backpropagation	Backpropagation	Local	Image	Heatmap

Continued on next page

Name	Year	Features	Explanation Technique	Scope	Input Type	Output Type
XRAI [174]	2019	Region-based saliency for images	Region-Based	Local	Image	Visual Regions
PD [156]	2001	Shows average feature effects on predictions	Feature Importance	Global	Tabular	Visual
ICE [155]	2015	Individual conditional expectations for features	Feature Importance	Local	Tabular	Visual
Explanation by Example [315]	2017	Representative examples for explanation	Example-Based	Local/ Global	Video, Tabular	Example Instances

Bibliography

[1] M. Klokova, “What you must know in history of machine learning,” https://medium.com/@marizombie/important-dates-in-history-of-machine-learning-d52d7f24b797, 2024, accessed: 18 March 2025.

[2] E. Commission, C. Directorate-General for Communications Networks, and Technology, *Ethics guidelines for trustworthy AI.* Publications Office, 2019.

[3] M. Kamran, “Chapter 2 - energy sources and technologies,” in *Fundamentals of Smart Grid Systems*, M. Kamran, Ed. Academic Press, 2023, pp. 23–69. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780323995603000107

[4] National Renewable Energy Laboratory (U.S.), *Improving Pumping System Performance: A Sourcebook for Industry, Second Edition (Revised).* Washington, D.C., United States: U.S. Department of Energy, Office of Energy Efficiency and Renewable Energy, 10 2012. [Online]. Available: https://www.osti.gov/biblio/1094887

[5] M. Berno, M. Canil, N. Chiarello, L. Piazzon, F. Berti, F. Ferrari, A. Zaupa, N. Ferro, M. Rossi, and G. A. Susto, “A machine learning-based approach for advanced monitoring of automated equipment for the entertainment industry,” in *2021 IEEE International Workshop on Metrology for Industry 4.0*, 2021, pp. 386–391.

[6] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (xai): Towards medical xai," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, pp. 4793–4813, 2021.

[7] H. Lakkaraju, N. Arsov, and O. Bastani, "Robust and stable black box explanations," in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research. PMLR, 2020, pp. 5628–5638.

[8] W. J, S. S, S. M, G. M, L. VX, D.-V. F, J. K, H. K, K. D, S. M, O. PN, T.-I. S, and G. A., "Do no harm: a roadmap for responsible machine learning for health care," *Nature Medicine*, vol. 25, no. 9, pp. 1337–1340, 2019.

[9] M. JA, J. Y, P. Z, and H. I., "Ai explainability: The technical and ethical dimensions," *Philosophical Transactions: Mathematical, Physical and Engineering Sciences*, 2021.

[10] A. D. Preece, D. Harborne, D. Braines, R. J. Tomsett, and S. Chakraborty, "Stakeholders in explainable ai," *ArXiv*, 2018, abs/1810.00184.

[11] M. A. Köhl, K. Baum, M. Langer, D. Oster, T. Speith, and D. Bohlender, "Explainability as a non-functional requirement," in *2019 IEEE 27th International Requirements Engineering Conference (RE)*, 2019.

[12] A. Vellido, J. D. Martín-Guerrero, and P. J. G. Lisboa, "Making machine learning models interpretable," in *The European Symposium on Artificial Neural Networks*, 2012. [Online]. Available: https://api.semanticscholar.org/CorpusID:9237797

[13] B. F. Goodman, "European union regulations on algorithmic decision making and a "right to explanation," *AI Magazine*, vol. 38, no. 3, p. 76, 2017.

[14] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, "Machine learning interpretability: A survey on methods and metrics," *Electronics*, vol. 8, no. 8, p. 832, 2019.

[15] F. PASQUALE, *The Black Box Society: The Secret Algorithms That Control Money and Information*. Harvard University Press, 2015. [Online]. Available: http://www.jstor.org/stable/j.ctt13x0hch

[16] L. Kästner, M. Langer, V. Lazar, A. Schomäcker, T. Speith, and S. Sterz, "On the relation of trust and explainability: Why to engineer for trustworthiness," in *2021 IEEE 29th International Requirements Engineering Conference Workshops (REW)*. IEEE, 2021, pp. 169–175.

[17] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J. M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez, and F. Herrera, "Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence," *Information Fusion*, vol. 99, p. 101805, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253523001148

[18] G. Montavon, W. Samek, and K. Müller, "Methods for interpreting and understanding deep neural networks," *Digital Signal Processing*, vol. 73, pp. 1–15, 2018.

[19] R. Michalski, J. Carbonell, and T. Mitchell, *A Theory and Methodology of Inductive Learning*. Berlin, Heidelberg: Springer, 1983.

[20] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities, and challenges towards responsible ai," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[21] G. Stiglic, P. Kocbek, N. Fijacko, M. Zitnik, K. Verbert, and L. Cilar Budler, "Interpretability of machine learning-based prediction models in healthcare," *WIREs Data Mining and Knowledge Discovery*, vol. 10, no. 5, p. e1379, 2020.

[22] P. P. Angelov, E. A. Soares, R. Jiang, N. I. Arnold, and P. M. Atkinson, "Explainable artificial intelligence: an analytical review," *WIREs Data Mining and Knowledge Discovery*, vol. 11, no. 5, p. e1424, 2021.

[23] P. Hall and N. Gill, *An introduction to machine learning interpretability*. O'Reilly Media, Incorporated, 2019.

[24] D. Gunning and D. Aha, "Darpa's explainable artificial intelligence (xai) program," *AI Magazine*, vol. 40, no. 2, pp. 44–58, 2019.

[25] M. Munn and D. Pitman, *Explainable AI for Practitioners.* " O'Reilly Media, Inc.", 2022.

[26] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence*, vol. 267, pp. 1–38, 2019.

[27] O. Biran and C. V. Cotton, "Explanation and justification in machine learning: a survey," in *Workshop on Explainable AI (XAI)*, 2017, p. 6.

[28] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. A. Specter, and L. Kagal, "Explaining explanations: An approach to evaluating interpretability of machine learning," *CoRR*, vol. abs/1806.00069, 2018. [Online]. Available: http://arxiv.org/abs/1806.00069

[29] A. A. Freitas, "Comprehensible classification models: a position paper," *SIGKDD Explorations Newsletter*, vol. 15, no. 1, p. 1–10, 2014.

[30] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, pp. 206–215, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:182656421

[31] X.-H. Li, C. C. Cao, Y. Shi, W. Bai, H. Gao, L. Qiu, C. Wang, Y. Gao, S. Zhang, X. Xue, and L. Chen, "A survey of data-driven and knowledge-aware explainable ai," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 1, pp. 29–49, 2020.

[32] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Comput. Surv.*, vol. 51, no. 5, aug 2018. [Online]. Available: https://doi.org/10.1145/3236009

[33] F. Doshi-Velez and B. Kim, “Towards a rigorous science of interpretable machine learning,” *arXiv: Machine Learning*, 2017. [Online]. Available: https://arxiv.org/abs/1702.08608

[34] A. L. Samuel, “Some studies in machine learning using the game of checkers,” *IBM Journal of Research and Development*, vol. 3, no. 3, pp. 210–229, 1959, first recorded use of the term "machine learning".

[35] F. Rosenblatt, “The perceptron: a probabilistic model for information storage and organization in the brain.” *Psychological review*, vol. 65 6, pp. 386–408, 1958.

[36] L. Breiman, J. Friedman, C. Stone, and R. Olshen, *Classification and Regression Trees.* Taylor & Francis, 1984. [Online]. Available: https://books.google.co.uk/books?id=JwQx-WOmSyQC

[37] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back-propagating errors,” *Nature*, vol. 323, no. 6088, pp. 533–536, 1986. [Online]. Available: https://doi.org/10.1038/323533a0

[38] C. Cortes and V. Vapnik, “Support-vector networks,” *Mach. Learn.*, vol. 20, no. 3, p. 273–297, Sep. 1995. [Online]. Available: https://doi.org/10.1023/A:1022627411411

[39] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997. [Online]. Available: https://www.bioinf.jku.at/publications/older/2604.pdf

[40] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[41] L. Breiman, “Random forests,” *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[42] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” *Commun. ACM*, vol. 60, no. 6, p. 84–90, May 2017. [Online]. Available: https://doi.org/10.1145/3065386

[43] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:14124313

[44] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, p. 139–144, Oct. 2020. [Online]. Available: https://doi.org/10.1145/3422622

[45] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 5998–6008.

[46] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2019. [Online]. Available: https://arxiv.org/abs/1810.04805

[47] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS '20. Red Hook, NY, USA: Curran Associates Inc., 2020.

[48] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:231591445

[49] A. Ramesh, M. Pavlov, G. Goh, S. Gray *et al.*, “Zero-shot text-to-image generation,” *arXiv preprint arXiv:2102.12092*, 2021.

[50] OpenAI, “Chatgpt: Optimizing language models for dialogue,” *OpenAI blog*, 2022.

[51] C. Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable.* Independently published, 2018. [Online]. Available: https://christophm.github.io/interpretable-ml-book/

[52] Z. C. Lipton, “The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery,” *Queue*, vol. 16, no. 3, p. 31–57, 2018.

[53] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[54] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[55] I. D. Raji and J. Buolamwini, “Saving face: Investigating the ethical concerns of facial recognition auditing,” in *Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society*, ser. AIES ’20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 145–151.

[56] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 4765–4774.

[57] M. T. Ribeiro, S. Singh, and C. Guestrin, “"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*, ser. KDD ’16. New York, NY, USA: Association for Computing Machinery, 2016, p. 1135–1144. [Online]. Available: https://doi.org/10.1145/2939672.2939778

[58] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[59] T. Hastie and R. Tibshirani, *Generalized Additive Models*. Chapman and Hall/CRC, 1990. [Online]. Available: https://www.crcpress.com/Generalized-Additive-Models/Hastie-Tibshirani/p/book/9780412343902

[60] H. Nori, S. Jenkins, P. Koch, and R. Caruana, "Interpretml: A unified framework for machine learning interpretability," *arXiv preprint arXiv:1909.09223*, 2019. [Online]. Available: https://arxiv.org/abs/1909.09223

[61] Y. Lou, R. Caruana, J. Gehrke, and G. Hooker, "Accurate intelligible models with pairwise interactions," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '13. New York, NY, USA: Association for Computing Machinery, 2013, p. 623–631. [Online]. Available: https://doi.org/10.1145/2487575.2487579

[62] J. Schneider, "Explainable generative ai (genxai): a survey, conceptualization, and research agenda," *Artificial Intelligence Review*, vol. 57, p. 289, 2024.

[63] C. Mennella, U. Maniscalco, G. De Pietro, and M. Esposito, "Ethical and regulatory challenges of ai technologies in healthcare: A narrative review," *Heliyon*, vol. 10, no. 4, p. e26297, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2405844024023284

[64] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of ai ethics guidelines," *Nature Machine Intelligence*, vol. 1, no. 9, pp. 389–399, 2019.

[65] L. Floridi and J. Cowls, *A Unified Framework of Five Principles for AI in Society*. John Wiley & Sons, Ltd, 2022, pp. 535–545.

[66] European Commission, "Proposal for a regulation laying down harmonised rules on artificial intelligence and amending certain union legislative acts," 2021, accessed: 2024-11-01. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206

[67] UK Government, "A pro-innovation approach to ai regulation: Government response," 2024, accessed: 2024-11-01. [Online]. Available: https://www.gov.uk/government/consultations/ai-regulation-a-pro-innovation-approach-policy-proposals/outcome/a-pro-innovation-approach-to-ai-regulation-government-response

[68] National Institute of Standards and Technology (NIST), "Artificial intelligence risk management framework (ai rmf 1.0)," NIST Framework Document, 2023. [Online]. Available: https://www.nist.gov/itl/ai-risk-management-framework

[69] Organisation for Economic Co-operation and Development, "Oecd principles on artificial intelligence," 2019, accessed: 2024-11-01. [Online]. Available: https://www.oecd.org/going-digital/ai/principles/

[70] UK Government, "Emerging processes for frontier ai safety," 2023, accessed: 2025-09-10. [Online]. Available: https://assets.publishing.service.gov.uk/media/653aabbd80884d000df71bdc/emerging-processes-frontier-ai-safety.pdf

[71] Information Commissioner's Office, "Explaining decisions made with ai: Guidance for organisations," 2020, accessed: 2025-09-10. [Online]. Available: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/explaining-decisions-made-with-artificial-intelligence/part-1-the-basics-of-explaining-ai/what-goes-into-an-explanation/

[72] Health and Safety Executive, "Artificial intelligence and health and safety," 2023, accessed: 2025-09-10. [Online]. Available: https://www.hse.gov.uk/news/hse-ai.htm

[73] Taylor Wessing, "The uk's proposed regulatory ai principles of transparency, explainability and fairness," 2023, accessed: 2025-09-10. [Online]. Available: https://www.taylorwessing.com/en/global-data-hub/2023/july---ai-and-data/the-uks-proposed-regulatory-ai-principles-of-transparency-explainability-and-fairness

[74] European Commission, "Regulatory framework proposal on artificial intelligence," 2021, accessed: 2025-09-10. [Online]. Available: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

[75] ISACA, "Understanding the eu ai act," 2024, accessed: 2025-09-10. [Online]. Available: https://www.isaca.org/resources/white-papers/2024/understanding-the-eu-ai-act

[76] Reuters, "Eu's new ai rules ignite battle over data transparency," 2024, accessed: 2025-09-10. [Online]. Available: https://www.reuters.com/technology/artificial-intelligence/eus-new-ai-rules-ignite-battle-over-data-transparency-2024-06-13

[77] ——, "Eu code of practice to help firms with ai rules will focus on copyright, safety," 2025, accessed: 2025-09-10. [Online]. Available: https://www.reuters.com/business/eu-code-practice-help-firms-with-ai-rules-will-focus-copyright-safety-2025-07-10

[78] T. LLP, "The eu ai act - ten key things to know," 2024, accessed: 2025-09-10. [Online]. Available: https://www.tlt.com/insights-and-events/insight/the-eu-ai-act---ten-key-things-to-know

[79] T. ML, "Eu ai act: how risk is classified," 2024, accessed: 2025-09-10. [Online]. Available: https://www.trail-ml.com/blog/eu-ai-act-how-risk-is-classified

[80] N. I. of Standards and Technology, "Artificial intelligence risk management framework (ai rmf 1.0)," U.S. Department of Commerce, Tech. Rep., 2023, nIST AI Risk Management Framework identifying explainability and interpretability as components of trustworthy AI. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

[81] D. J. Trump, "Executive order 13960: Promoting the use of trustworthy artificial intelligence in the federal government," https://www.federalregister.gov/documents/2020/12/08/2020-27065/promoting-the-use-of-trustworthy-artificial-intelligence-in-the-federal-government, 2020, establishes policy for federal AI to be transparent and understandable.

[82] J. R. Biden, "Executive order 14110: Safe, secure, and trustworthy development and use of artificial intelligence," https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligen 2023, establishes governance requirements for transparency and federal AI leadership.

[83] United Nations Educational, Scientific and Cultural Organization, "Recommendation on the ethics of artificial intelligence," 2021, accessed: 2024-11-01. [Online]. Available: https://unesdoc.unesco.org/ark:/48223/pf0000380455

[84] International Organization for Standardization, "Iso/iec jtc 1/sc 42 artificial intelligence standards," 2022, accessed: 2024-11-01. [Online]. Available: https://www.iso.org/committee/6794475.html

[85] I. S. Association, "Ieee guide for an architectural framework for explainable artificial intelligence," *IEEE Std 2894-2024*, pp. 1–55, 2024.

[86] F. K. Dosilovic, M. Brčić, and N. Hlupic, "Explainable artificial intelligence: A survey," *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pp. 0210–0215, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:49560076

[87] U. Government, "A pro-innovation approach to ai regulation," 2023, policy Paper.

[88] OECD, "Oecd principles on artificial intelligence," 2021, available at: https://www.oecd.org/going-digital/ai/principles/.

[89] N. A. Smuha, "The eu approach to ethics guidelines for trustworthy artificial intelligence," *Computer Law Review International*, vol. 20, no. 4, pp. 97–106, 2019.

[90] L. Floridi and J. Cowls, "A unified framework of five principles for ai in society," *Harvard Data Science Review*, vol. 1, no. 1, 2018.

[91] S. Tufail, H. Riggs, M. Tariq, and A. I. Sarwat, "Advancements and challenges in machine learning: A comprehensive review of models, libraries, applications, and algorithms," *Electronics*, vol. 12, no. 8, 2023. [Online]. Available: https://www.mdpi.com/2079-9292/12/8/1789

[92] U. Kamath and J. Liu, *Explainable artificial intelligence: An introduction to interpretable machine learning*. Springer, 2021.

[93] L. Vigano and D. Magazzeni, "Explainable security," in *2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*. Piscataway: IEEE, 2020, pp. 293–300.

[94] S. Mitchell, E. Potash, S. Barocas, A. D'Amour, and K. Lum, "Algorithmic fairness: Choices, assumptions, and definitions," *Annual Review of Statistics and Its Application*, vol. 8, no. 1, pp. 141–163, 2021. [Online]. Available: https://doi.org/10.1146/annurev-statistics-042720-125902

[95] M. Langer, D. Oster, T. Speith, H. Hermanns, L. Kästner, E. Schmidt, A. Sesing, and K. Baum, "What do we want from explainable artificial intelligence (xai)? – a stakeholder perspective on xai and a conceptual model guiding interdisciplinary xai research," *Artificial Intelligence*, vol. 296, p. 103473, 2021.

[96] W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Muller, *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, 1st ed. Springer Publishing Company, Incorporated, 2019.

[97] C. Bove, J. Aigrain, M. Lesot, C. Tijus, and M. Detyniecki, "Contextualization and exploration of local feature importance explanations to improve understanding and satisfaction of non-expert users," in *27th International Conference on Intelligent User Interfaces (IUI '22)*. New York, NY, USA: Association for Computing Machinery, 2022, p. 807–819.

[98] S. Mohseni, N. Zarei, and E. D. Ragan, "A survey of evaluation methods and measures for interpretable machine learning," *ArXiv*, 2018, abs/1811.11839.

[99] K. Martin, A. Liret, N. Wiratunga, G. Owusu, and M. Kern, "Developing a catalogue of explainability methods to support expert and non-expert users," in *Artificial Intelligence XXXVI: 39th SGAI International Conference on Artificial Intelligence*. Berlin, Heidelberg: Springer-Verlag, 2019, p. 309–324.

[100] H.-F. Cheng, R. Wang, Z. Zhang, F. O'Connell, T. Gray, F. M. Harper, and H. Zhu, "Explaining decision-making algorithms through ui: Strategies to help non-expert stakeholders," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2019, paper 559.

[101] S. Lo Piano, "Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward," *Humanities and Social Sciences Communications*, vol. 7, p. 9, 2020. [Online]. Available: https://doi.org/10.1057/s41599-020-0501-9

[102] J. S. Christian Meske, Enrico Bunde and M. Gersch, "Explainable artificial intelligence: Objectives, stakeholders, and future research opportunities," *Information Systems Management*, vol. 39, no. 1, pp. 53–63, 2022. [Online]. Available: https://doi.org/10.1080/10580530.2020.1849465

[103] G. Plumb *et al.*, "Model agnostic supervised local explanations," *Advances in Neural Information Processing Systems*, vol. 31, pp. 2515–2524, 2018.

[104] S. Mohseni, N. Zarei, and E. D. Ragan, "A multidisciplinary survey and framework for design and evaluation of explainable ai systems," *ACM Trans. Interact. Intell. Syst.*, vol. 11, no. 3–4, Sep. 2021. [Online]. Available: https://doi.org/10.1145/3387166

[105] M. T. Ribeiro, S. Singh, and C. Guestrin, "Model-agnostic interpretability of machine learning," in *6 ICML Workshop on Human Interpretability in Machine Learning*, New York, USA, 2016.

[106] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLOS ONE*, vol. 10, no. 7, p. e0130140, Jul. 2015.

[107] W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, and K.-R. Müller, "Evaluating the visualization of what a deep neural network has learned," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 11, pp. 2660–2673, 2017.

[108] C.-K. Yeh, C.-Y. Hsieh, A. S. Suggala, D. I. Inouye, and P. Ravikumar, *On the (in)fidelity and sensitivity of explanations.* Red Hook, NY, USA: Curran Associates Inc., 2019.

[109] Y. Liu, Y. Liu, J. McCrory, and X. Guo, "High-fidelity digital twin modelling for predictive maintenance state-of-the-art," *CIE51 Proceedings*, 2024.

[110] D. Alvarez-Melis and T. S. Jaakkola, "On the robustness of interpretability methods," in *Proceedings of the 35th International Conference on Machine Learning*, vol. 80. PMLR, 2018, pp. 2429–2438. [Online]. Available: https://arxiv.org/abs/1806.08049

[111] H. Lakkaraju, S. H. Bach, and J. Leskovec, "Interpretable decision sets: A joint framework for description and prediction," in *Proceedings of the 22nd ACM*

SIGKDD International Conference on Knowledge Discovery and Data Mining, ser. KDD '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 1675–1684. [Online]. Available: https://doi.org/10.1145/2939672.2939874

[112] J. Zhou, A. H. Gandomi, F. Chen, and A. Holzinger, "Evaluating the quality of machine learning explanations: A survey on methods and metrics," *Electronics*, vol. 10, no. 5, 2021. [Online]. Available: https://www.mdpi.com/2079-9292/10/5/593

[113] J. Jung, H. Lee, H. Jung, and H. Kim, "Essential properties and explanation effectiveness of explainable artificial intelligence in healthcare: A systematic review," *Heliyon*, vol. 9, no. 5, p. e16110, 2023. [Online]. Available: https://doi.org/10.1016/j.heliyon.2023.e16110

[114] M. Pawlicki, A. Pawlicka, F. Uccello, S. Szelest, S. D'Antonio, R. Kozik, and M. Choraś, "Evaluating the necessity of the multiple metrics for assessing explainable ai: A critical examination," *Neurocomputing*, vol. 602, p. 128282, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231224010531

[115] V. Petsiuk, A. Das, and K. Saenko, "Rise: Randomized input sampling for explanation of black-box models," *ArXiv*, vol. abs/1806.07421, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:49324724

[116] S. Hooker, D. Erhan, P.-J. Kindermans, and B. Kim, *A benchmark for interpretability methods in deep neural networks.* Red Hook, NY, USA: Curran Associates Inc., 2019.

[117] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, ser. NIPS'18. Red Hook, NY, USA: Curran Associates Inc., 2018, p. 9525–9536.

[118] S. Sutthithatip, S. Perinpanayagam, and S. Aslam, "(explainable) artificial intelligence in aerospace safety-critical systems," in *2022 IEEE Aerospace Conference,*

AERO 2022, ser. IEEE Aerospace Conference Proceedings. United States: IEEE Computer Society, Aug. 2022, publisher Copyright: © 2022 IEEE.; 2022 IEEE Aerospace Conference, AERO 2022 ; Conference date: 05-03-2022 Through 12-03-2022.

[119] G. Vilone and L. Longo, "Notions of explainability and evaluation approaches for explainable artificial intelligence," *Information Fusion*, vol. 76, pp. 89–106, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253521001093

[120] H. Lakkaraju, E. Kamar, R. Caruana, and J. Leskovec, "Faithful and customizable explanations of black box models," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society.* Honolulu, HI, USA: Association for Computing Machinery, 2019, p. 131–138.

[121] A.-K. Dombrowski, M. Alber, C. Anders, M. Ackermann, K.-R. Müller, and P. Kessel, "Explanations can be manipulated and geometry is to blame," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/bb836c01cdc9120a9c984c525e4b1a4a-Paper.pdf

[122] L. Rieger and L. Hansen, "A simple defense against adversarial attacks on heatmap explanations," in *Proceedings of 2020 Workshop on Human Interpretability in Machine Learning*, 2020, 5th Annual Workshop on Human Interpretability in Machine Learning, WHI 2020 ; Conference date: 17-07-2020 Through 17-07-2020.

[123] J. Heo, S. Joo, and T. Moon, "Fooling neural network interpretations via adversarial model manipulation," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019.

[Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/7fea637fd6d02b8f0adf6f7dc36aed93-Paper.pdf

[124] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626. [Online]. Available: https://doi.org/10.1109/ICCV.2017.74

[125] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *CoRR*, vol. abs/1312.6034, 2013. [Online]. Available: https://api.semanticscholar.org/CorpusID:1450294

[126] D. Slack, S. Hilgard, S. Singh, and H. Lakkaraju, "Reliable post hoc explanations: Modelling uncertainty in explainability," in *Proceedings of the 35th NeurIPS*, 2021.

[127] B. Dimanov, U. Bhatt, M. Jamnik, and A. Weller, "You shouldn't trust me: Learning models which conceal unfairness from multiple explanation methods," in *SafeAI@AAAI*, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:211838210

[128] J. Cação, J. Santos, and M. Antunes, "Explainable ai for industrial fault diagnosis: A systematic review," *Journal of Industrial Information Integration*, vol. 47, p. 100905, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2452414X25001281

[129] N. Bussmann, P. Giudici, D. Marinelli, and J. Papenbrock, "Explainable ai in fintech risk management," *Frontiers in Artificial Intelligence*, vol. 3, p. 26, 2020.

[130] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission," in *Proceedings of the 21th ACM SIGKDD International*

Conference on Knowledge Discovery and Data Mining, ser. KDD '15. New York, NY, USA: Association for Computing Machinery, 2015, p. 1721–1730. [Online]. Available: https://doi.org/10.1145/2783258.2788613

[131] J. Amann, A. Blasimme, E. Vayena, D. Frey, and V. Madai, "Explainability for artificial intelligence in healthcare: a multidisciplinary perspective," *BMC Medical Informatics and Decision Making*, vol. 20, p. 310, 2020. [Online]. Available: https://doi.org/10.1186/s12911-020-01332-6

[132] S. S Band, A. Yarahmadi, C.-C. Hsu, M. Biyari, M. Sookhak, R. Ameri, I. Dehzangi, A. T. Chronopoulos, and H.-W. Liang, "Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods," *Informatics in Medicine Unlocked*, vol. 40, p. 101286, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352914823001302

[133] M. Ghassemi, L. Oakden-Rayner, and A. L. Beam, "The false hope of current approaches to explainable artificial intelligence in health care," *The Lancet Digital Health*, vol. 3, no. 11, pp. e745–e750, 2021.

[134] P. Weber, K. V. Carl, and O. Hinz, "Applications of explainable artificial intelligence in finance—a systematic review of finance, information systems, and computer science literature," *Management Review Quarterly*, vol. 74, pp. 867–907, 2024. [Online]. Available: https://doi.org/10.1007/s11301-023-00320-0

[135] V. Arya, R. K. E. Bellamy, P.-Y. Chen, A. Dhurandhar, M. Hind, S. C. Hoffman, S. Houde, Q. V. Liao, R. Luss, A. Mojsilović, S. Mourad, P. Pedemonte, R. Raghavendra, J. Richards, P. Sattigeri, K. Shanmugam, M. Singh, K. R. Varshney, D. Wei, and Y. Zhang, "One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques," 2019. [Online]. Available: https://arxiv.org/abs/1909.03012

[136] P. Bracke, A. Datta, C. Jung, and S. Sen, "Machine learning explainability in finance: An application to default risk

analysis," Bank of England, Working Paper, 2019. [Online]. Available: https://www.bankofengland.co.uk/working-paper/2019/machine-learning-explainability-in-finance-an-application-to-default-risk-analysis

[137] PwC, "Insurance claims estimator uses ai for efficiency – case study," PwC case study, Ïnsurance claims estimator uses AI for efficiency", 2024. [Online]. Available: https://www.pwc.com/us/en/library/case-studies/auto-insurance-ai-analytics.html

[138] B. Kumar and T. Kumar, "Explainable ai in finance and investment banking: Techniques, applications, and future directions," *Journal of Scientific and Engineering Research*, vol. 9, no. 5, pp. 119–124, 2022, cODEN(USA): JSERBR. [Online]. Available: www.jsaer.com

[139] E. Brusa, L. Cibrario, C. Delprete, and L. G. Di Maggio, "Explainable ai for machine fault diagnosis: Understanding features' contribution in machine learning models for industrial condition monitoring," *Applied Sciences*, vol. 13, no. 4, 2023. [Online]. Available: https://www.mdpi.com/2076-3417/13/4/2038

[140] U. Dereci and G. Tuzkaya, "An explainable artificial intelligence model for predictive maintenance and spare parts optimization," *Supply Chain Analytics*, vol. 8, p. 100078, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2949863524000219

[141] A. T. Keleko, B. Kamsu-Foguem, R. H. Ngouna, and A. Tongne, "Health condition monitoring of a complex hydraulic system using deep neural network and deepshap explainable xai," *Advances in Engineering Software*, vol. 175, p. 103339, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S096599782200240X

[142] D. Astolfi, F. De Caro, and A. Vaccaro, "Condition monitoring of wind turbine systems by explainable artificial intelligence techniques," *Sensors*, vol. 23, no. 12, 2023. [Online]. Available: https://www.mdpi.com/1424-8220/23/12/5376

[143] D. Kim, G. Antariksa, M. P. Handayani, S. Lee, and J. Lee, "Explainable anomaly detection framework for maritime main engine sensor data," *Sensors*, vol. 21, no. 15, p. 5200, 2021.

[144] M. L. Baptista, K. Goebel, and E. M. Henriques, "Relation between prognostics predictor evaluation metrics and local interpretability shap values," *Artificial Intelligence*, vol. 306, p. 103667, 2022.

[145] B. Hrnjica and S. Softic, "Explainable ai in manufacturing: a predictive maintenance case study," in *IFIP International Conference on Advances in Production Management Systems*. Springer, 2020, pp. 66–73.

[146] S. Gawde, S. Patil, S. Kumar, P. Kamat, and K. Kotecha, "An explainable predictive maintenance strategy for multi-fault diagnosis of rotating machines using multi-sensor data fusion," *Decision Analytics Journal*, vol. 10, p. 100425, 2024.

[147] S. S. Jena, J. Yang, and F. Tan, "Unlocking sales growth: Account prioritization engine with explainable ai," *ArXiv*, vol. abs/2306.07464, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:259145105

[148] J. Joung and H. Kim, "Interpretable machine learning-based approach for customer segmentation for new product development from online product reviews," *International Journal of Information Management*, vol. 70, p. 102641, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0268401223000221

[149] K. Wei, B. Chen, J. Zhang, S. Fan, K. Wu, G. Liu, and D. Chen, "Explainable deep learning study for leaf disease classification," *Agronomy*, vol. 12, no. 5, p. 1035, 2022.

[150] G. Tolomei, F. Silvestri, A. Haines, and M. Lalmas, "Interpretable predictions of tree-based ensembles via actionable feature tweaking," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '17. New York, NY, USA:

Association for Computing Machinery, 2017, p. 465–474. [Online]. Available: https://doi.org/10.1145/3097983.3098039

[151] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (xai)," *IEEE Access*, vol. 6, pp. 52 138–52 160, 2018.

[152] G. Vilone and L. Longo, "Explainable artificial intelligence: a systematic review," 2020. [Online]. Available: https://arxiv.org/abs/2006.00093

[153] V. Belle and I. Papantonis, "Principles and practice of explainable machine learning," *Frontiers in Big Data*, vol. Volume 4 - 2021, 2021. [Online]. Available: https://www.frontiersin.org/journals/big-data/articles/10.3389/fdata.2021.688969

[154] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in *AAAI Conference on Artificial Intelligence*, 2018.

[155] A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin, "Peeking inside the black box: Visualising statistical learning with plots of individual conditional expectation," *Journal of Computational and Graphical Statistics*, vol. 24, no. 1, pp. 44–65, 2015.

[156] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[157] H. Deng, "Interpreting tree ensembles with intrees," *CoRR*, vol. abs/1408.5456, 2014. [Online]. Available: http://arxiv.org/abs/1408.5456

[158] T. Speith, "A review of taxonomies of explainable artificial intelligence (xai) methods," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2022, pp. 932–944.

[159] P. Biecek, "Dalex: explainers for complex predictive models in r," *J. Mach. Learn. Res.*, vol. 19, no. 1, p. 3245–3249, Jan. 2018.

[160] R. Agarwal, L. Melnick, N. Frosst, X. Zhang, B. Lengerich, R. Caruana, and G. E. Hinton, "Neural additive models: Interpretable machine

learning with neural nets," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 4699–4711. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2021/file/251bd0442dfcc53b5a761e050f8022b8-Paper.pdf

[161] J. H. Friedman and B. E. Popescu, "Predictive learning via rule ensembles," *The Annals of Applied Statistics*, vol. 2, no. 3, pp. 916–954, 2008.

[162] R. S. Michalski, *A theory and methodology of inductive learning.* Springer, 2000.

[163] Q. Wang, B. Li, F. Zhang, X. Lin, and X. Xu, "Scalable probabilistic rule learning for knowledge graph reasoning," in *Advances in Neural Information Processing Systems*, 2017, pp. 4585–4594.

[164] S. Teso and K. Kersting, "Why should i trust interactive learners? explaining interactive queries of classifiers to users," in *Advances in Neural Information Processing Systems*, 2019, pp. 1054–1064.

[165] K. S. Gurumoorthy, A. Dhurandhar, G. Cecchi, and C. C. Aggarwal, "Efficient data representation for scalable prototypical explanations," in *AISTATS*, 2017, pp. 532–541.

[166] B. Kim, R. Khanna, and O. O. Koyejo, "Mmd-critic: Unsupervised model criticism via maximum mean discrepancy," in *ICML*, 2016, pp. 2607–2616.

[167] A. Dhurandhar, P.-Y. Chen, R. Luss, C.-C. Tu, P.-Y. Ting, K. Shanmugam, and P. Das, "Explanations based on the missing: Towards contrastive explanations with pertinent negatives," in *NeurIPS*, 2018, pp. 592–603.

[168] R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," in *FAT*, 2020, pp. 607–617.

[169] R. Poyiadzi, K. Sokol, R. Santos-Rodriguez, T. De Bie, and P. Flach, "Face: Feasible and actionable counterfactual explanations," in *AISTATS*, 2020, pp. 1034–1042.

[170] A.-H. Karimi, J. von Kügelgen, B. Schölkopf, and I. Valera, "Model-agnostic counterfactual explanations for consequential decisions," in *AISTATS*, 2020, pp. 895–905.

[171] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[172] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *International Conference on Machine Learning*, 2017, pp. 3319–3328.

[173] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "Smoothgrad: removing noise by adding noise," in *arXiv preprint arXiv:1706.03825*, 2017. [Online]. Available: https://arxiv.org/abs/1706.03825

[174] A. Kapishnikov and et al., "Xrai: Better attributions through regions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4948–4957.

[175] J. T. Springenberg and et al., "Striving for simplicity: The all convolutional net," in *arXiv preprint arXiv:1412.6806*, 2014.

[176] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *International Conference on Machine Learning*, 2017, pp. 3145–3153.

[177] M. Gentzkow and J. M. Shapiro, "Text as data," *Journal of the Economics of Textual Data*, 2019, arXiv preprint arXiv:1903.08626.

[178] J. Camacho-Collados and M. T. Pilehvar, "On the role of text preprocessing in neural network architectures: An evaluation study on text categorization and sentiment analysis," *arXiv preprint arXiv:1707.01780*, 2017.

[179] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, no. 1, Jan. 2020.

[180] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, and N. Fiedel, "Palm: scaling language modeling with pathways," *J. Mach. Learn. Res.*, vol. 24, no. 1, Mar. 2024.

[181] I. Tenney, J. Wexler, J. Bastings, T. Bolukbasi, A. Coenen, S. Gehrmann, E. Jiang, M. Pushkarna, C. Radebaugh, E. Reif, and A. Yuan, "The language interpretability tool: Extensible, interactive visualizations and analysis for nlp models," in *Conference on Empirical Methods in Natural Language Processing*, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:225423556

[182] L. Hiley, D. Jiarong, W. Xue, and L. Yang, "Explainable video classification with spatio-temporal convolutions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2019.

[183] A. Stergiou and R. Poppe, "Saliency tubes: Visual explanations for spatio-temporal convolutions," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2019.

[184] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, and R. Sayres, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2018. [Online]. Available: https://arxiv.org/abs/1711.11279

[185] L. Meng, B. Zhao, B. Chang, G. Huang, W. Sun, F. Tung, and L. Sigal, "Interpretable spatio-temporal attention for video action recognition," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 2019, pp. 1513–1522.

[186] U. Bhatt, A. Xiang, S. Sharma, A. Weller, A. Taly, Y. Jia, J. Ghosh, R. Puri, J. M. F. Moura, and P. Eckersley, "Explainable machine learning in deployment," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. Barcelona, Spain: Association for Computing Machinery, 2020, p. 648–657.

[187] D. Doran, S. Schulz, and T. R. Besold, "What does explainable ai really mean? a new conceptualization of perspectives," in *Proceedings of the First International Workshop on Comprehensibility and Explanation in AI and ML*, Bari, Italy, 2017.

[188] A. Holzinger, R. Goebel, R. Fong, T. Moon, K.-R. Müller, and W. Samek, "Xxai - beyond explainable artificial intelligence," in *XxAI - Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers*. Berlin, Heidelberg: Springer-Verlag, 2020, p. 3–10. [Online]. Available: https://doi.org/10.1007/978-3-031-04083-2_1

[189] L. Kopitar, P. Kokol, and G. Stiglic, "Hybrid visualization-based framework for depressive state detection and characterization of atypical patients," *Journal of Biomedical Informatics*, vol. 147, p. 104535, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1532046423002563

[190] D. W. Apley and J. Zhu, "Visualizing the effects of predictor variables in black box supervised learning models," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 82, no. 4, pp. 1059–1086, 06 2020. [Online]. Available: https://doi.org/10.1111/rssb.12377

[191] M. T. Ribeiro, S. Singh, and C. Guestrin, “Anchors: High-precision model-agnostic explanations,” in *AAAI Conference on Artificial Intelligence*, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:3366554

[192] J. W. Tukey, *Exploratory data analysis.* Reading, Mass.: Addison-Wesley Pub. Co., 1977.

[193] J. Chakraborty, K. Peng, and T. Menzies, “Making fair ml software using trustworthy explanation,” in *2020 35th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, 2020, pp. 1229–1233.

[194] H. Namkoong and J. C. Duchi, “Stochastic gradient methods for distributionally robust optimization with f-divergences,” *Advances in neural information processing systems*, vol. 29, 2016.

[195] P. Dabkowski and Y. Gal, “Real time image saliency for black box classifiers,” 2017.

[196] T. Wang, C. Rudin, F. Velez-Doshi, Y. Liu, E. Klampfl, and P. MacNeille, “Bayesian rule sets for interpretable classification,” in *2016 IEEE 16th International Conference on Data Mining (ICDM)*, 2016, pp. 1269–1274.

[197] B. Letham, C. Rudin, T. H. McCormick, and D. Madigan, “Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model,” *The Annals of Applied Statistics*, vol. 9, no. 3, pp. 1350–1371, 2015.

[198] H. Yang, C. Rudin, and M. Seltzer, “Scalable Bayesian rule lists,” in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. PMLR, 06–11 Aug 2017, pp. 3921–3930. [Online]. Available: https://proceedings.mlr.press/v70/yang17h.html

[199] T. Cover and P. Hart, “Nearest neighbor pattern classification,” *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.

[200] H. Jiang, B. Kim, M. Y. Guan, M. Gupta, and P. Liang, "To trust or not to trust a classifier," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 31, 2018. [Online]. Available: https://papers.nips.cc/paper_files/paper/2018/hash/7720bf5fadba5c34410689e8f1bb0e4c-Abstract.html

[201] S. Wachter, B. D. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the GDPR," *CoRR*, vol. abs/1711.00399, 2017. [Online]. Available: http://arxiv.org/abs/1711.00399

[202] K. P. F.R.S., "Liii. on lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901.

[203] I. Jolliffe, *Principal Component Analysis.* Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 1094–1096. [Online]. Available: https://doi.org/10.1007/978-3-642-04898-2_455

[204] G. E. Hinton and S. Roweis, "Stochastic neighbor embedding," in *Advances in Neural Information Processing Systems*, S. Becker, S. Thrun, and K. Obermayer, Eds., vol. 15. MIT Press, 2002. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2002/file/6150ccc6069bea6b5716254057a194ef-Paper.pdf

[205] L. van der Maaten and G. Hinton, "Viualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 11 2008.

[206] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. Springer, 2009.

[207] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable ai for trees," *Nature Machine Intelligence*, vol. 2, pp. 56–67, 2020. [Online]. Available: https://doi.org/10.1038/s42256-019-0138-9

[208] M. Mehta, V. Palade, and I. Chatterjee, *Explainable AI: Foundations, Methodologies and Applications.* Springer Cham, 11 2022.

[209] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, and K.-R. Müller, *Layer-Wise Relevance Propagation: An Overview.* Springer, Cham, 09 2019, pp. 193–209.

[210] A. Ghorbani, J. Wexler, J. Y. Zou, and B. Kim, "Towards automatic concept-based explanations," *Advances in neural information processing systems*, vol. 32, 2019.

[211] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable ai: A review of machine learning interpretability methods," *Entropy*, vol. 23, no. 1, 2021. [Online]. Available: https://www.mdpi.com/1099-4300/23/1/18

[212] A. Jacovi, S. Swayamdipta, S. Ravfogel, Y. Elazar, Y. Choi, and Y. Goldberg, "Contrastive explanations for model interpretability," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, Eds. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 1597–1611. [Online]. Available: https://aclanthology.org/2021.emnlp-main.120

[213] E. Strumbelj and I. Kononenko, "An efficient explanation of individual classifications using game theory," *Journal of Machine Learning Research*, vol. 11, no. 1, pp. 1–18, 2010.

[214] H. GM, M. K. Gourisaria, M. Pandey, and S. S. Rautaray, "A comprehensive survey and analysis of generative models in machine learning," *Computer Science Review*, vol. 38, p. 100285, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1574013720303853

[215] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, p. 3, 2022.

[216] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[217] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *CoRR*, vol. abs/1409.0473, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:11212020

[218] Z. Lin, M. Feng, C. N. dos Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio, "A structured self-attentive sentence embedding," *CoRR*, vol. abs/1703.03130, 2017. [Online]. Available: http://arxiv.org/abs/1703.03130

[219] S. Chaudhari, G. Polatkan, R. Ramanath, and V. Mithal, "An attentive survey of attention models," *ACM Transactions on Intelligent Systems and Technology*, vol. 12, no. 5, pp. 1–32, 2021.

[220] R. Child, S. Gray, A. Radford, and I. Sutskever, "Generating long sequences with sparse transformers," *arXiv preprint arXiv:1904.10509*, 2019. [Online]. Available: https://arxiv.org/abs/1904.10509

[221] C. Dwork, "Differential privacy: A survey of results," in *Theory and Applications of Models of Computation*, 2008. [Online]. Available: https://api.semanticscholar.org/CorpusID:2887752

[222] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 308–318. [Online]. Available: https://doi.org/10.1145/2976749.2978318

[223] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *International Conference on Artificial Intelligence and Statistics*, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:14955348

[224] C. Wu, P. Jiang, C. Ding, F. Feng, and T. Chen, "Intelligent fault diagnosis of rotating machinery based on one-dimensional convolutional neural network," *Computers in Industry*, vol. 108, pp. 53–61, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016636151830513X

[225] R. Rejith, D. Kesavan, P. Chakravarthy, and S. Narayana Murty, “Bearings for aerospace applications,” *Tribology International*, vol. 181, p. 108312, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0301679X23000993

[226] J. C. A. Jauregui Correa and A. A. Lozano Guzman, “Chapter eight - condition monitoring,” in *Mechanical Vibrations and Condition Monitoring*, J. C. A. Jauregui Correa and A. A. Lozano Guzman, Eds. Academic Press, 2020, pp. 147–168. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128197967000081

[227] X. K. Wu, “The fault diagnosis based on information fusion theory and its application in internal combustion engine,” Ph.D. thesis, Wuhan University of Technology, Wuhan, 1998.

[228] A. Almasi, “Centrifugal pumps: Design and application,” *Journal of Mechanical Design*, vol. 137, no. 2, p. 021401, 2015.

[229] J. Costello, *Machine Learning Techniques for the Health Monitoring of Rotating Machinery in Nuclear Power Plants.* University of Strathclyde, 2019. [Online]. Available: https://books.google.co.uk/books?id=Fn4ZzgEACAAJ

[230] Y. Sun, T. Zhao, Z. Zou, Y. Chen, and H. Zhang, “Imbalanced data fault diagnosis of hydrogen sensors using deep convolutional generative adversarial network with convolutional neural network,” *Review of Scientific Instruments*, vol. 92, p. 095007, 09 2021.

[231] N. M. Paz and W. E. Leigh, “Maintenance scheduling: Issues, results and research needs,” *International Journal of Operations & Production Management*, vol. 14, pp. 47–69, 1994. [Online]. Available: https://api.semanticscholar.org/CorpusID:154542821

[232] D. Vanzile and I. Otis, *Measuring and Controlling Machine Performance.* New York, NY: John Wiley, 1992.

[233] J. Sharma, M. Mittal, and G. Soni, "Condition-based maintenance using machine learning and role of interpretability: a review," *International Journal of Systems Assurance Engineering and Management*, 2022. [Online]. Available: https://doi.org/10.1007/s13198-022-01843-7

[234] A. K. Choudhary and D. Ahmad Khan, *Introduction to Conditioning Monitoring of Mechanical Systems*. Singapore: Springer Singapore, 2020, pp. 205–230. [Online]. Available: https://doi.org/10.1007/978-981-15-1532-3_9

[235] S. M. del Campo and F. Sandin, "Online feature learning for condition monitoring of rotating machinery," *Engineering Applications of Artificial Intelligence*, vol. 64, pp. 187–196, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0952197617301318

[236] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time Series Analysis: Forecasting and Control*, 5th ed. John Wiley & Sons, 2015.

[237] A. Bustos, H. Rubio, C. Castejon, and J. C. Garcia-Prada, "Condition monitoring of critical mechanical elements through graphical representation of state configurations and chromogram of bands of frequency," *Measurement*, vol. 135, pp. 71–82, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0263224118310832

[238] A. K. Jardine, D. Lin, and D. Banjevic, "A review on machinery diagnostics and prognostics implementing condition-based maintenance," *Mechanical Systems and Signal Processing*, vol. 20, no. 7, pp. 1483–1510, 2006. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0888327005001512

[239] S. Selcuk, "Predictive maintenance, its implementation and latest trends," *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture*, vol. 231, no. 9, pp. 1670–1679, 2017. [Online]. Available: https://doi.org/10.1177/0954405415601640

[240] L. H. Chiang, R. D. Braatz, and E. L. Russel, "Canonical variate analysis," *Fault Detection and Diagnosis in Industrial Systems.*, 2001.

[241] V. Atamuradov, K. Medjaher, F. Camci, N. Zerhouni, P. Dersin, and B. Lamoureux, "Machine health indicator construction framework for failure diagnostics and prognostics," *Journal of signal processing systems*, vol. 92, pp. 591–609, 2020.

[242] P. Baraldi, R. Canesi, E. Zio, R. Seraoui, and R. Chevalier, "Signal grouping for condition monitoring of nuclear power plant components," in *Reliability and Risk Management ESREL 2011*, 2011, pp. 1–13.

[243] R. Chevalier, D. Provost, and R. Seraoui, "Assessment of statistical and classification models for monitoring edf's assets, sixth american nuclear society int," in *Topical Meeting on Nuclear Plant Instrumentation, Control and Human-Machine Interface Technologies, Knoxville, USA*, 2009.

[244] M. E. Janusz and V. Venkatasubramanian, "Automatic generation of qualitative descriptions of process trends for fault detection and diagnosis," *Engineering Applications of Artificial Intelligence*, vol. 4, no. 5, pp. 329–339, 1991. [Online]. Available: https://www.sciencedirect.com/science/article/pii/095219769190001M

[245] S. Windmann, J. Eickmeyer, F. Jungbluth, J. Badinger, and O. Niggemann, "Evaluation of model-based condition monitoring systems in industrial application cases," in *Machine Learning for Cyber Physical Systems*, O. Niggemann and J. Beyerer, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2016, pp. 45–50.

[246] A. A. Jaber, *Design of an Intelligent Embedded System for Condition Monitoring of an Industrial Robot*, 1st ed. Springer International Publishing, 2017, imprint: Springer.

[247] H. Wang, C. Ma, and L. Zhou, "A brief review of machine learning and its application," in *2009 International Conference on Information Engineering and Computer Science*, 2009, pp. 1–4.

[248] Z. M. Çınar, A. Abdussalam Nuhu, Q. Zeeshan, O. Korhan, M. Asmael, and B. Safaei, "Machine learning in predictive maintenance towards sustainable

smart manufacturing in industry 4.0," *Sustainability*, vol. 12, no. 19, 2020. [Online]. Available: https://www.mdpi.com/2071-1050/12/19/8211

[249] S.-I. Ao, L. Gelman, H. R. Karimi, and M. Tiboni, "Advances in machine learning for sensing and condition monitoring," *Applied Sciences*, vol. 12, no. 23, 2022. [Online]. Available: https://www.mdpi.com/2076-3417/12/23/12392

[250] P. Lanka, D. Rangaprakash, M. Dretsch, J. Katz, T. J. Denney, and G. Deshpande, "Supervised machine learning for diagnostic classification from large-scale neuroimaging datasets," *Brain Imaging and Behavior*, vol. 14, no. 6, pp. 2378–2416, Dec 2020.

[251] C.-L. Yang, Z.-X. Chen, and C.-Y. Yang, "Sensor classification using convolutional neural network by encoding multivariate time series as two-dimensional colored images," *Sensors*, vol. 20, no. 1, 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/1/168

[252] e. a. Chenang Liu, Zhenyu (James) Kong, "An integrated manifold learning approach for high-dimensional data feature extractions and its applications to online process monitoring of additive manufacturing," *IISE Transactions*, vol. 53, no. 11, pp. 1215–1230, 2021. [Online]. Available: https://doi.org/10.1080/24725854.2020.1849876

[253] I. Thoidis, M. Giouvanakis, and G. Papanikolaou, "Semi-supervised machine condition monitoring by learning deep discriminative audio features," *Electronics*, vol. 10, no. 20, 2021. [Online]. Available: https://www.mdpi.com/2079-9292/10/20/2471

[254] X. Zhao, M. Li, J. Xu, and G. Song, "Multi-class semi-supervised learning in machine condition monitoring," in *2009 International Conference on Information Engineering and Computer Science*, 2009, pp. 1–4.

[255] L. Maaløe, O. Winther, S. Spataru, and D. Sera, "Condition monitoring in photovoltaic systems by semi-supervised machine learning," *Energies*, vol. 13, no. 3, 2020. [Online]. Available: https://www.mdpi.com/1996-1073/13/3/584

[256] D. K. Kotary and S. J. Nanda, "Distributed robust data clustering in wireless sensor networks using diffusion moth flame optimization," *Engineering Applications of Artificial Intelligence*, vol. 87, p. 103342, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0952197619302830

[257] M. Gao and G.-Y. Shi, "Ship-handling behavior pattern recognition using ais sub-trajectory clustering analysis based on the t-sne and spectral clustering algorithms," *Ocean Engineering*, vol. 205, p. 106919, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0029801820300032

[258] G. Niu, E. Liu, B. Zhang, M. Golda, and S. Mastro, "A deep residual convolutional neural network based bearing fault diagnosis with multi-sensor data," in *2021 4th IEEE International Conference on Industrial Cyber-Physical Systems (ICPS)*, 2021, pp. 655–660.

[259] P. Tavner, L. Ran, J. Penman, and H. Sedding, "Condition monitoring of rotating electrical machines," *Bibliovault OAI Repository, the University of Chicago Press*, 01 2008.

[260] L. C. Brito, G. A. Susto, J. N. Brito, and M. A. V. Duarte, "Fault diagnosis using explainable ai: A transfer learning-based approach for rotating machinery exploiting augmented synthetic data," *Expert Systems with Applications*, vol. 232, p. 120860, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417423013623

[261] B. Shukla, I.-S. Fan, and I. Jennions, "Opportunities for explainable artificial intelligence in aerospace predictive maintenance," *PHM Society European Conference*, vol. 5, no. 1, p. 11, 2020.

[262] E. Christoforou, K. Blom, Q. Gao, M. Börü, and T. Cataltepe, "Mri condition monitoring with explainable ai and feature selection," in *2022 30th Signal Processing and Communications Applications Conference (SIU)*, 2022, pp. 1–4.

[263] Y. Zhang, S. Hu, Y. Zhang, and L. Chen, "Optimization and analysis of centrifugal pump considering fluid-structure interaction," *The Scientific*

World Journal, vol. 2014, p. 131802, 2014. [Online]. Available: https://doi.org/10.1155/2014/131802

[264] L. Chen, L. Wei, Y. Wang, J. Wang, and W. Li, "Monitoring and predictive maintenance of centrifugal pumps based on smart sensors," *Sensors*, vol. 22, no. 6, 2022. [Online]. Available: https://www.mdpi.com/1424-8220/22/6/2106

[265] M. Tiboni, C. Remino, R. Bussola, and C. Amici, "A review on vibration-based condition monitoring of rotating machinery," *Applied Sciences*, vol. 12, no. 3, 2022. [Online]. Available: https://www.mdpi.com/2076-3417/12/3/972

[266] D. Cornel, F. G. Guzmán, G. Jacobs, and S. Neumann, "Condition monitoring of roller bearings using acoustic emission," *Wind Energy Science*, vol. 6, pp. 367–376, 2021.

[267] Xylem Inc., "Part 1 – how to read a pump performance curve," Xylem Inc., Tech. Rep., 2013, accessed: 2025-09-02. [Online]. Available: https://www.xylem.com/siteassets/support/white-papers/white-paper-resources/pump-curve-white-paper-series_part-1_final-1.pdf

[268] P. S. Matter, "Reliability basics: Understanding npsh, por, and aor," Article / Web resource, 2022. [Online]. Available: https://www.pumps.org/2022/03/21/reliability-basics-understanding-npsh-por-and-aor/

[269] Hydraulic Institute Standards, "ANSI/HI 9.6.3 – Rotodynamic (Centrifugal and Vertical) Pumps – Guideline for Operating Regions," 2017, hydraulic Institute, Parsippany, NJ, USA. [Online]. Available: https://www.pumps.org/HI-Standards

[270] A. R. Budris, "Centrifugal pump flow operating regions and impact on reliability," Article, *WaterWorld*, 2016, argues that pumps should operate near BEP to minimize life-cycle costs and improve reliability. [Online]. Available: https://www.waterworld.com/wastewater-treatment/article/16191262/centrifugal-pump-flow-operating-regions-and-impact-on-reliability

[271] Pumps & Systems, "Pump vibration monitoring basics," Pumps & Systems Magazine, 2020, accessed: 15 September 2025. [Online]. Available: https://www.pumpsandsystems.com/pump-vibration-monitoring-basics

[272] PRUFTECHNIK, "What is pump condition monitoring?" PRUFTECHNIK Condition Monitoring Guide, 2023, accessed: 15 September 2025. [Online]. Available: https://www.pruftechnik.com/condition-monitoring-pilot-pumps/

[273] V. S. Lobanoff and R. R. Ross, "Index," in *Centrifugal Pumps*, 2nd ed. Boston: Gulf Professional Publishing, 1992, pp. 569–577. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780080500850500254

[274] M. H. Garousi, M. Karimi, P. Casoli, M. Rundo, and R. Fallahzadeh, "Vibration analysis of a centrifugal pump with healthy and defective impellers and fault detection using multi-layer perceptron," *Eng*, vol. 5, no. 4, pp. 2511–2530, 2024.

[275] A. Simons, H. Otoo, S. Nunoo, E. Ababio, and C. Addy, "Effects of rotating shaft misalignment on bearing life and lubricant failure," *American Journal of Mechanical and Materials Engineering*, vol. 4, no. 3, pp. 53–59, 2020. [Online]. Available: https://www.sciencepublishinggroup.com/article/10.11648/j.ajmme.20200403.12

[276] Y. Takami, T. Nakazawa, K. Makinouchi, J. Glueck, Y. Ohara, R. J. Benkowski, and Y. Nosé, "Pump power loss and heat generation in a pivot bearing-supported gyro centrifugal pump (c1e3)," *Artif Organs*, vol. 20, no. 7, pp. 794–797, 1996.

[277] D. Baglee and M. Knowles, "Maintenance strategy development within smes: The development of an integrated approach," *Journal of Quality in Maintenance Engineering*, vol. 16, no. 2, pp. 195–209, 2010.

[278] A. Al-Ghamd and D. Mba, "A comparative experimental study on the use of acoustic emission and vibration analysis for bearing defect detection," *NDT & E International*, vol. 39, no. 6, pp. 471–480, 2006.

[279] D. Mba and R. Rao, "Development of acoustic emission technology for condition monitoring and diagnosis of rotating machinery," *Applied Acoustics*, vol. 67, no. 9, pp. 940–965, 2006.

[280] D. B. Jani, Y. Mistry, A. Suthar, J. Shah, and P. Patel, "An overview on cavitation in centrifugal pumps," *International Journal of Innovative Research in Technology*, vol. 6, no. 5, 2019, discusses effects of cavitation on flow/pressure, performance degradation, bearing wear. [Online]. Available: https://www.researchgate.net/publication/336253916_An_overview_on_cavitation_in_centrifugal_pump

[281] J. Sinha and K. Elbhbah, "A future condition monitoring system for industrial plant and equipment," *Tribology International*, vol. 35, no. 10, pp. 631–636, 2002.

[282] F. Hussain, R. Hussain, and E. Hossain, "Explainable artificial intelligence (xai): An engineering perspective," 2021. [Online]. Available: https://arxiv.org/abs/2101.03613

[283] O. Amin, B. Brown, B. Stephen, S. McArthur, and V. Livina, "Ml explanations by design: A case study explaining the predicted degradation of a roto-dynamic pump," in *Proceedings of NDT 2023 – 60th Annual British Conference on NDT*, Northampton, United Kingdom, 2023.

[284] B. Iooss, R. Kenett, and P. Secchi, *Different Views of Interpretability*. Springer International Publishing, 2022, p. 1–20. [Online]. Available: http://dx.doi.org/10.1007/978-3-031-12402-0_1

[285] J. Crabbé and M. van der Schaar, "Evaluating the robustness of interpretability methods through explanation invariance and equivariance," in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, ser. NIPS '23. Red Hook, NY, USA: Curran Associates Inc., 2024.

[286] I. Nunes and D. Jannach, "A systematic review and taxonomy of explanations in decision support and recommender systems," *User Modeling and User-Adapted*

Interaction, vol. 27, no. 3–5, pp. 393–444, Oct 2017. [Online]. Available: https://doi.org/10.1007/s11257-017-9195-0

[287] G. Ras, N. Xie, M. Van Gerven, and D. Doran, "Explainable deep learning: A field guide for the uninitiated," *Journal of Artificial Intelligence Research*, vol. 73, pp. 329–397, Jan 2022. [Online]. Available: https://doi.org/10.1613/jair.1.13200

[288] P. Tfekci and H. Kaya, "Combined Cycle Power Plant," UCI Machine Learning Repository, 2014, DOI: https://doi.org/10.24432/C5002N.

[289] N. S. Santarisi and S. S. Faouri, "Prediction of combined cycle power plant electrical output power using machine learning regression algorithms," *Eastern-European Journal of Enterprise Technologies*, vol. 6, no. 8 (114), p. 16–26, Dec. 2021. [Online]. Available: https://journals.uran.ua/eejet/article/view/245663

[290] P. Tüfekci, "Prediction of full load electrical power output of a base load operated combined cycle power plant using machine learning methods," *International Journal of Electrical Power and Energy Systems*, vol. 60, pp. 126–140, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0142061514000908

[291] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[292] L. E. Peterson, "K-nearest neighbor," *Scholarpedia*, vol. 4, p. 1883, 2009. [Online]. Available: https://api.semanticscholar.org/CorpusID:29611121

[293] D. Garreau and U. von Luxburg, "Explaining the explainers: A first theoretical analysis of lime," in *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 1287–1296. [Online]. Available: https://proceedings.mlr.press/v108/garreau20a.html

[294] SHAP Developers, "An introduction to explainable ai with shapley values," https://shap.readthedocs.io/en/latest/example_notebooks/overviews/An%

20introduction%20to%20explainable%20AI%20with%20Shapley%20values.html, 2018, accessed: 2025-09-16.

[295] P. Hermosilla, S. Berríos, and H. Allende-Cid, "Explainable ai for forensic analysis: A comparative study of shap and lime in intrusion detection models," *Applied Sciences*, vol. 15, no. 13, 2025. [Online]. Available: https://www.mdpi.com/2076-3417/15/13/7329

[296] J. Goldwasser and G. Hooker, "Statistical significance of feature importance rankings," 2025. [Online]. Available: https://arxiv.org/abs/2401.15800

[297] A. Fisher, C. Rudin, and F. Dominici, "All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously," *Journal of Machine Learning Research*, vol. 20, no. 177, pp. 1–81, 2019. [Online]. Available: http://jmlr.org/papers/v20/18-760.html

[298] O. Amin, B. Brown, B. Stephen, and S. McArthur, "A case-study led investigation of explainable ai (xai) to support deployment of prognostics in the industry," *PHME-CONF*, vol. 7, no. 1, p. 9–20, Jun 2022.

[299] K. Sokol and P. Flach, "Explainability fact sheets: A framework for systematic assessment of explainable approaches," in *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, ser. FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. United States: Association for Computing Machinery (ACM), Jan. 2020, pp. 56–67, 3rd ACM Conference on Fairness, Accountability, and Transparency, FAT* 2020 ; Conference date: 27-01-2020 Through 30-01-2020.

[300] F. Poursabzi-Sangdeh, D. G. Goldstein, J. M. Hofman, J. W. Wortman Vaughan, and H. Wallach, "Manipulating and measuring model interpretability," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, ser. CHI '21. New York, NY, USA: Association for Computing Machinery, 2021. [Online]. Available: https://doi.org/10.1145/3411764.3445315

[301] I. Lage, E. Chen, J. He, M. Narayanan, B. Kim, S. J. Gershman, and F. Doshi-Velez, “Human evaluation of models built for interpretability,” *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 7, no. 1, pp. 59–67, Oct. 2019. [Online]. Available: https://ojs.aaai.org/index.php/HCOMP/article/view/5280

[302] B. Stephen, B. Brown, A. Young, A. Duncan, H. Helfer-Hoeltgebaum, G. West, C. Michie, and S. D. J. McArthur, “A quantile dependency model for predicting optimal centrifugal pump operating strategies,” *Machines*, vol. 10, no. 7, p. 557, 2022.

[303] J. Brownlee, “How to convert a time series to a supervised learning problem in python,” 2019, accessed: November 10, 2022. [Online]. Available: https://machinelearningmastery.com/convert-time-series-supervised-learning-problem-python/

[304] K. Sokol and P. Flach, “Explainability fact sheets: a framework for systematic assessment of explainable approaches,” in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, ser. FAT* ’20. New York, NY, USA: Association for Computing Machinery, 2020, p. 56–67. [Online]. Available: https://doi.org/10.1145/3351095.3372870

[305] A. V. Ponce-Bobadilla, V. Schmitt, C. S. Maier, S. Mensing, and S. Stodtmann, “Practical guide to shap analysis: Explaining supervised machine learning model predictions in drug development,” *Clinical and Translational Science*, vol. 17, no. 11, p. e70056, 2024. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC11513550/

[306] K. Aas, M. Jullum, and A. Løland, “Explaining individual predictions when features are dependent: More accurate approximations to shapley values,” *arXiv preprint arXiv:1903.10464*, 2019. [Online]. Available: https://arxiv.org/abs/1903.10464

[307] L. X., "Neurosymbolic approaches for complex systems," Presented at the Alan Turing Institute Neurosymbolic AI Talk Series, 2024, accessed from transcript provided in discussion.

[308] A. K. Coker, "5 - pumping of liquids," in *Ludwig's Applied Process Design for Chemical and Petrochemical Plants*, 4th ed., A. K. Coker, Ed. Burlington: Gulf Professional Publishing, 2007, pp. 303–369. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780750677660500129

[309] Pumps and Systems Magazine, "Best efficiency point and performance curves," https://www.pumpsandsystems.com/best-efficiency-point-performance-curves, Aug. 2021, accessed: 2025-09-10.

[310] C. Frye, C. Rowat, and I. Feige, "Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS'20. Red Hook, NY, USA: Curran Associates Inc., 2020.

[311] E. Anderson, "The species problem in iris," *Annals of the Missouri Botanical Garden*, vol. 23, no. 3, pp. 457–509, 1936. [Online]. Available: http://www.jstor.org/stable/2394164

[312] Y. Kwon and J. Zou, "Weightedshap: analyzing and improving shapley based feature attributions," in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, ser. NIPS '22. Red Hook, NY, USA: Curran Associates Inc., 2022.

[313] A. Chattopadhyay and e. a. Manupriya, "Neural network attributions: A causal perspective," *arXiv preprint arXiv:1902.02302*, 2019.

[314] B. van der Velden, "Explainable ai: Current status and future potential," *Eur Radiol*, 2023.

[315] P. W. Koh and P. Liang, "Understanding black-box predictions via influence functions," in *Proceedings of the 34th International Conference on Machine*

Learning, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. PMLR, 06–11 Aug 2017, pp. 1885–1894. [Online]. Available: https://proceedings.mlr.press/v70/koh17a.html