
Explainability of Rare Events

Bruno Brandão Teixeira

Dissertation

Master in Modelling, Data Analysis and Decision Support Systems

Supervised by:

PhD João Manuel Portela da Gama

PhD Rita Paula Ribeiro

2024

Biographical note

Bruno Brandão Teixeira is the co-founder and manager of SpinnerDynamics, a company specialising in the design and development of complex engineering solutions, with a strong focus on additive manufacturing, product customisation, and process improvement. With extensive experience spanning multiple industries and countries — ranging from finance and education to business development — Bruno has honed his expertise in optimising business operations, strategic management and innovation, particularly within growing enterprises.

Bruno’s educational background includes a Bachelor’s degree in Business Management from the Faculty of Economics of Porto and ongoing studies towards a Master’s degree in Modelling, Data Analysis and Decision Support Systems at the same institution. His academic interests, combined with his professional experience, have led him to explore Explainable AI and its application in detecting and explaining rare events in machine learning models, contributing to the field through his research.

In addition to his entrepreneurial and academic pursuits, Bruno has co-founded initiatives like Centro Infantil Casa Colorida and has actively participated in various community and cultural activities. He has been recognised with multiple awards, including the European 1st Place in the European Enterprise Challenge and the Intel Innovation Award.

Acknowledgments

I would like to express my deepest gratitude to my supervisors, Professors João Manuel Portela da Gama and Rita Paula Ribeiro, for their invaluable guidance, support, and encouragement throughout this journey. Their expertise, insights and their patience have been instrumental in shaping this dissertation, and their mentorship has been an enriching and inspiring experience.

I would also like to extend my heartfelt thanks to my family for their unwavering support and understanding during this demanding period, and to Daniel Braga and Sérgio Tavares for their encouragement and support. A special thank you to Vânia Ramos for her careful revision and continuous encouragement throughout the writing process — your support made a world of difference.

Additionally, I am immensely grateful to António Campos and Paulo Lima for the numerous discussions, study sessions, and intellectual companionship we shared throughout the master's program. Your collaboration and dedication have been a constant source of motivation and inspiration.

Thank you all for being part of this important chapter in my life.

Abstract

This dissertation explores the intersection of Explainable Artificial Intelligence (XAI) with the detection and interpretation of rare events, often referred to as outliers or anomalies. While outliers are typically removed from datasets to ensure model accuracy, they often hold valuable insights when properly understood and explained, which, in some cases, may even be the relevant data. This study introduces a novel approach to outlier explanations by applying clustering techniques to group them, allowing the extraction of meaningful rules and patterns from abnormal events. Using a combination of outlier detection methods, such as Isolation Forest and Random Forest, paired with explainability techniques including SHAP and LIME, the research uncovers significant insights into the detection and explanation of rare events across diverse datasets, including breast cancer, seismic, and cardiotocography data.

The results demonstrate that clustering explanations leads to a deeper understanding of outliers, uncovering relationships and patterns that are otherwise difficult to detect. In the seismic dataset, for instance, clustering the explanations revealed critical feature values related to energy and hazard indicators, while in the breast cancer dataset, key feature values and conditions driving outlier classification, were consistently identified across multiple approaches. The study also successfully validated the derived rules by generating synthetic outliers, with the models achieving high accuracy on the classification of these new instances, thereby reinforcing the utility of explanation clustering for outlier interpretation.

The conclusions highlight the importance of combining robust detection models with explainability techniques to derive actionable insights from rare events. This research contributes to the field by demonstrating the potential of explanation clustering to enhance model transparency and interpretability, especially in complex datasets. Future work is suggested to explore multiclass challenges, advanced feature clustering, and automated synthetic outlier generation to further improve explainability and expand the application of these techniques.

Keywords: Rare events, Outliers, Anomalies, Outlier Detection, Outlier Explanation, XAI, Explanations clustering, Explainability

Resumo

Esta dissertação explora a interseção de *Explainable Artificial Intelligence* (XAI) com a detecção e interpretação de eventos raros, muitas vezes referidos como outliers ou anomalias. Embora os outliers sejam várias vezes removidos dos modelos de decisão para garantir a precisão dos mesmos, muitas vezes contém informação fundamental quando devidamente compreendidos e explicados, podendo, em alguns casos, até ser a informação mais relevante.

Este estudo apresenta uma abordagem inovadora para a explicação de *outliers*, aplicando técnicas de *clustering* para as agrupar, permitindo a extração de regras e padrões significativos de eventos anómalos. Utilizando uma combinação de métodos de detecção de *outliers*, como *Isolation Forest* e *Random Forest*, juntamente com técnicas de explicabilidade, incluindo *SHAP* e *LIME*, o estudo revela descobertas significativas sobre a detecção e explicação de eventos raros em diversos *datasets*, nomeadamente dados sobre o cancro de mama, atividades sísmica e cardiocardiografia.

Os resultados demonstram que as explicações de *clustering* levam a uma melhor compreensão de *outliers*, ao encontrar associações e padrões que são difíceis de detetar. Por exemplo, no *dataset* sísmico, o *clustering* das explicações revela valores críticos das variáveis relacionadas com energia e perigo, enquanto no *dataset* do cancro de mama, os principais valores e condições das variáveis que conduzem à classificação de outliers foram consistentemente identificadas em várias abordagens. O estudo também validou com sucesso as regras extraídas gerando *outliers* sintéticos, com modelos que obtêm alta precisão na classificação dessas novas observações, reforçando a utilidade do *clustering* das explicações para interpretação de eventos raros.

As conclusões destacam a importância de combinar modelos de detecção robustos com técnicas de explicabilidade para obter conhecimento válido de eventos raros. Esta pesquisa contribui para esta temática, demonstrando o potencial do *clustering* de explicações para aumentar a transparência e a interpretabilidade dos modelos, especialmente em desafios complexos. Sugere-se que em trabalhos futuros se explorem classificações multiclasse, técnicas avançadas de *clustering* e automatização da geração de outliers sintéticos para melhorar ainda mais a explicabilidade e expandir a aplicação dessas técnicas.

Palavras-chave: Eventos raros, *Outliers*, Anormalidade, Detecção de *outliers*, Explicação de *outliers*, XAI, *Clustering* de explicações, Explicabilidade.

Contents

Biographical note	iii
Acknowledgements	v
Abstract	vii
Resumo	ix
1 Introduction	1
1.1 The Impact of GDPR on Explainability and Decision Support Systems . . .	2
1.2 Motivation	3
1.3 Structure	4
2 Related Work	5
2.1 Outlier Detection	5
2.2 Explainable AI	7
2.3 Evaluation Metrics	11
3 Methodology	15
3.1 Data	15
3.2 Data Preprocessing	16
3.3 Outlier Detection	16
3.4 Explainable AI	17
3.5 Clustering the explanations	22
3.6 Evaluation Metrics	23
3.6.1 ROC-AUC	23
3.6.2 Other Evaluation Metrics	26
4 Experimental Results	27
4.1 Breast Cancer Wisconsin dataset	27
4.1.1 Outlier Detection	27
4.1.2 Explainable AI	29
4.1.3 Explanations Clustering	35
4.2 The Seismic Sataset	42
4.2.1 Outlier Detection	42
4.2.2 Explainable AI	43
4.2.3 Explanations Clustering	45
4.3 Cardiotocography Dataset	48
4.3.1 Outlier Detection	49
4.3.2 Explainable AI	49
4.3.3 Explanations Clustering	50

4.4	Results Discussion	51
4.5	Perturbing the datasets - Synthetic Outliers	53
5	Conclusion	55
5.1	Contributions	56
5.2	Limitations and Future Work	56
	Bibliography	59
A	Appendix	i
A.1	Outliers & DataSets (ODDS)	i
A.2	Outlier Detection Algorithms & Techniques:	iii
A.3	Outlier Explanations	xiv
A.4	Techniques to generate outliers	xvii
A.5	Experimental Results Aux	xix
A.6	Outlier Explanations	xxiii

Glossary

ABOD - Angle-based Outlier Detection
AI - Artificial Intelligence
AGN - AGNostic black box
AnoGAN - Anomaly Detection with Generative Adversarial Networks
AUC - Area Under the Curve
CAM - Class Activation Mapping
CAM-LIME or LIME-CAM - LIME-based variant of CAM
COF - Connectivity-based Outlier Factor
COIN - Contextual Outlier INterpretation
CBLOF - Cluster-Based Local Outlier Factor
CESI - Comparative Expert Stability Index
CRISP-DM - CRoss Industry Standard Process for Data Mining
CMM - Combined Multiple Models
DBSCAN - Density-Based Spatial Clustering Of Applications With Noise
DILOF - Effective and Memory Efficient Local Outlier Detection in Data Streams
DNN - Deep Nerual Networks
DoS - Denial-of-service
DDoS - Distributed Denial of Service
DR - Decision Rules
DT -Decision Tree
DUE - Dynamic Updated Ensemble
EXAD - Explainable Anomaly Detection
FI - Features Importance
FPR - False Positive Rate
GLADs - Generalised Linear Anomaly Detectors
GDPR - General Data Protection Regulation
ICE - Individual Conditional Expectation
IF - Isolation Forest Algorithm
IMG - IMaGe
KKAS - Key Knowledge Attribute Subspace
LIME - Local Interpretable Model-Agnostic Explanations
LOCI - Fast Outlier Detection Using the Local Correlation Integral
LODI - Local Outlier Detection with Interpretation
LOF - Local Outlier Factor
LSH - locality-sensitive hashing
LSOD - Locality Sensitive Outlier Detection
ML - Machine Learning

NA - Neurons Activation
OD - Outlier Detection
OE - Outlier Explanation
ODD - Outlier Detection DataSets
PALM - Proximal Alternating Linearised Minimization
PCA - Principal Component Analysis
PDP - Partial Dependence Plot
PINN - Projection-Indexed Nearest-Neighbours
PS - Proto-type Selection
RBRP - Recursive Binning and Re-Projection
RDA - Red Deer Algorithm
RF - Random Forest Algorithm
ROC - Receiver Operating Characteristic
ROC-AUC - Area Under the ROC Curve
SA -Sensitivity Analysis
SFE - Simple, Fast, and Efficient (Feature Selection Algorithm)
SHAP - SHapley Additive exPlanations
SM - Saliency Masks
SMOTE - Synthetic Minority Over-sampling TEchnique
SRF - Sapling Random Forest
STORM - Stream Outlier Miner Algorithm
SVM – Support Vector Machines
SVM+P - SVM-Polynomial
TAB - TABular
TE - Tree Ensemble
TPR - True Positive Rate
TXT - TeXT
UCI - University of California Irvine Machine Learning Repository
URCA - Unsupervised Root Cause Analysis
XAI - Explainable Artificial Intelligence
XGBOOST - eXtreme Gradient Boosting
XPACS - eXplaining Patterns of Anomalies with Characterizing Subspaces
XPM - Explainable Predictive Maintenance

List of Figures

1.1	Interest over time in Explainability	3
2.1	Boukerche, Zheng, and Alfandi (2020) Knowledge Map	6
2.2	Guidotti et al. (2018) Knowledge Map	8
2.3	Panji, Gruenwald, Leal, Nguyen, and Silvia (2022) Knowledge Map	9
4.1	ROC Curve: Evaluation metrics	28
4.2	ROC Curve: Random Forest Evaluation metrics	29
4.3	SHAP Summary Plot	29
4.4	Distribution of SHAP Values for each feature Across Outliers	30
4.5	Table and Figure: Difference in SHAP Values Between 2 Outliers	31
4.6	Force Plots of the first five Outliers	32
4.7	All SHAP Force Plots	33
4.8	SHAP Summary Plot	33
4.9	Distribution of SHAP Values for each feature Across Outliers	34
4.10	SHAP (Kernel) Summary Plot	34
4.11	Distribution of SHAP Values (Kernel) for each feature Across Outliers	35
4.12	Elbow Method for Optimal k	36
4.13	PCA of SHAP values with clustering	36
4.14	Distribution of SHAP Values by feature and cluster	37
4.15	Mean absolute SHAP Values by feature and cluster	37
4.16	PCA of SHAP values with clustering	40
4.17	Distribution of SHAP Values by feature and cluster (Kernel)	40
4.18	Mean absolute SHAP Values by feature and cluster Kernel)	41
4.19	PCA of SHAP values with clustering IF SHAP Tree	45
4.20	PCA of SHAP values with clustering RF SHAP Tree	47
4.21	PCA of SHAP values with clustering	51
A.1	Outlier Detection Techniques	ii
A.2	Hierarchical Diagram of Outliers Categorization	iv
A.3	Seismic ROC Curve: IF Evaluation metrics with SMOTE	xix
A.4	ROC Curve: Evaluation metrics (LOF) - seismic	xx
A.5	Breast Cancer ROC Curve: IF Evaluation metrics with SMOTE	xxi
A.6	ROC Curve: Evaluation metrics (LOF) - breast cancer	xxi
A.7	Cardiotocography ROC Curve: IF Evaluation metrics with SMOTE	xxii
A.8	Force Plots of the first five Outliers - Breast Cancer Dataset - IF and SHAP(tree)	xxiii
A.9	Force Plots of the first five Outliers (Kernel) - Breast Cancer	xxiv
A.10	LIME Explanations for the first 5 outliers - Breast cancer dataset	xxv

A.11	LIME Values by Feature condition and Cluster - Breast cancer dataset . . .	xxvi
A.12	Force Plots of the first five Outliers - Breast Cancer Dataset - RF and Tree Explainer	xxvi
A.13	Elbow Method for Optimal k - Kernel explainer	xxvii
A.14	Elbow Method for Optimal k - LIME	xxvii
A.15	Table and Figure: Difference in SHAP Values Between 2 Outliers - IF and Tree Explainer - Seismic dataset	xxvii
A.16	Force Plots of the first five Outliers - Seismic (IF and Tree Explainer) . . .	xxviii
A.17	ROC Curve: Evaluation metrics (Isolation Forest)	xxix
A.18	ROC Curve: Evaluation metrics (Random Forest with SMOTE)	xxix
A.19	SHAP Summary Plot - Tree Explainer based on Isolation Forest	xxx
A.20	Distribution of SHAP Values for each feature across Outliers - Tree Ex- plainer based on Isolation Forest	xxx
A.21	SHAP (Tree) Summary Plot (Random Forest with Feature Selection)	xxxi
A.22	Distribution of SHAP Values (Tree and RF with Feature selection) for each feature Across Outliers	xxxi
A.23	Force Plots of the first five Outliers - Seismic (RF with Feature Selection and Tree Explainer)	xxxii
A.24	Distribution of SHAP Values by feature and cluster (IF Tree SHAP)	xxxiii
A.25	Mean absolute SHAP Values by feature and cluster (IF Tree SHAP)	xxxiv
A.26	ROC Curve: Evaluation metrics (IF) Cardio dataset	xxxiv
A.27	ROC Curve: Evaluation metrics (RF) Cardio dataset	xxxv
A.28	SHAP Summary Plot (Isolation Forest - cardio)	xxxv
A.29	Distribution of SHAP Values for each feature Across Outliers (IF Tree - cardio)	xxxvi
A.30	Difference in SHAP (Tree) Values Between 2 Outliers (IF) - cardio	xxxvi
A.31	Force Plots of the first five Outliers - Cardio (IF and SHAP Tree)	xxxvii
A.32	ROC Curve: Evaluation metrics (LOF) - cardio	xxxviii
A.33	Distribution of SHAP Values by feature and cluster - IF Cardio dataset . . .	xxxviii
A.34	Mean absolute SHAP Values by feature and cluster - IF Cardio dataset . . .	xxxix

List of Tables

4.1	Breast Cancer Dataset Classification Evaluation Report	28
4.2	Dataset Classification Evaluation Report (Random Forest)	28
4.3	Cluster SHAP Values Metrics Summary	37
4.4	Cluster SHAP Values Metrics Summary (Kernel)	39
4.5	Cluster Metrics Summary - Isolation Forest and SHAP Tree Explainer . . .	46
4.6	Cluster Metrics Summary - Random Forest and SHAP Tree Explainer . . .	48
4.7	Cluster Metrics Summary	51
A.1	Examples of possible multi-dimensional point datasets from ODD	iii
A.2	Nearest-neighbor-based Outlier Detection	v
A.3	Clustering-based Outlier Detection	vi
A.4	Projection-based Outlier Detection	vii
A.5	Outlier Detection for High-dimensional Data	viii
A.6	Distance-based Outlier Detection in Data Streams	ix
A.7	Density-based Outlier Detection in Data Streams	x
A.8	Clustering-based Outlier Detection in Data Streams	xi
A.9	Distributed Outlier Detection	xii
A.10	Deep Learning-based Outlier Detection	xii
A.11	Outlier Detection with Feedback	xiii
A.12	Decision Systems Explanations Methods	xv
A.13	Outlier Explanation Techniques Summary	xvi
A.14	Summary of Techniques Generating Numerical Ranking of Outliers	xvii
A.15	Summary of Techniques Finding Outlying Attributes of Individuals and Groups	xviii
A.16	Isolation Forest (with SMOTE) Seismic Test Sample Classification Evalua- tion Report	xix
A.17	Isolation Forest (with SMOTE) Seismic Dataset Classification Evaluation Report	xix
A.18	Isolation Forest (with SMOTE) Breast Cancer Test Sample Classification Evaluation Report	xx
A.19	Isolation Forest (with SMOTE) Breast Cancer Wisconsin Dataset Classifi- cation Evaluation Report	xx
A.20	Test Sample Classification Evaluation Report (Random Forest) - breast cancer	xx
A.21	Isolation Forest (with SMOTE) Cardiocography Dataset Classification Eval- uation Report	xxi
A.22	Isolation Forest (with SMOTE) Cardiocography Dataset Classification Eval- uation Report	xxii
A.23	Cardio Dataset Classification Evaluation Report (Random Forest)	xxix

Chapter 1

Introduction

In data analytics and machine learning, the identification and interpretation of rare events, often referred to as outliers or anomalies, have gained significant attention due to their crucial roles in various applications such as medical research, environmental monitoring, and fraud detection (Panjei et al., 2022). Outlier detection can be accomplished through a wide range of techniques designed to unveil unusual observations that differ from the standard regarding their features. In the past, the norm was to remove outliers to ensure that models fit the training data. However, outliers often contain valuable information, and their utility is greatly enhanced when accompanied by comprehensive explanations that cover both the outlier detection process and the explanation of the identified anomalies.

Distinguishing rare events from standard occurrences involves addressing nuanced distinctions. Anomalies, for example, may arise from a different underlying generative mechanism, while outliers indicate statistical rarity and deviation without necessarily indicating a distinct generative mechanism. The conventional approach of discarding outliers during model training has evolved, recognising their potential to learn valuable information in various domains such as intrusion detection systems, fraud detection, medical anomaly diagnosis, fault detection, and environmental monitoring.

The outlier detection task comprehends multiple phases, including identifying instances deviating significantly from the norm, subsampling and ensembling, density peak clustering, and applying deep learning techniques. However, the significance of outliers is context-dependent, necessitating careful consideration of domain-specific knowledge to discern between genuinely informative outliers and those that may not be of interest.

Supervised, semi-supervised, and unsupervised methods can be applied to detect outliers. Unsupervised techniques gained prominence due to the challenges of obtaining labelled data for model training. The lack of labelled data presents challenges and drives innovation in unsupervised methods to serve diverse applications and scenarios, including high-dimensional data, unbounded data streams, and distributed large-scale datasets.

In the broader context of explainability, the work of Guidotti et al. (2018) dives into the challenges associated with the transparency and interpretability of outlier detection algorithms, seeking to bridge the gap between machine learning and data mining communities, which often focus on describing the workings of black-box models and explaining decisions, respectively.

Understanding the interpretability of models and the concept of explanation itself is a multifaceted challenge, marked by ongoing debates on definitions, properties, and the classi-

fication of methods for opening and explaining black boxes. This dissertation aims to clarify and define what makes a model interpretable, the characteristics of a valid explanation, and the trade-offs between prediction accuracy and interpretability.

1.1 The Impact of GDPR on Explainability and Decision Support Systems

With the enforcement of the General Data Protection Regulation (GDPR) in May 2018, a significant shift occurred in how organisations must handle data privacy, transparency, and accountability, particularly in decision-making processes involving automated systems. The GDPR, which is considered one of the most stringent data protection regulations globally, mandates that individuals have the right to be informed about decisions made by automated systems, especially those that significantly impact them, such as credit scoring, employment screening, or medical diagnoses.

GDPR places conditions on their use, especially when they are involved in making decisions that could have significant impacts on individuals. Key requirements include:

- **Transparency:** Organisations must inform individuals when automated decision-making is being used and provide meaningful information about the logic involved, the significance, and the consequences of such processing;
- **Right to Explanation:** Individuals have the right to request an explanation of the decision-making process and to challenge decisions made by automated systems;
- **Human Intervention:** In many cases, GDPR mandates that decisions with significant effects should not be made solely by automated systems; human oversight or intervention should be available to ensure fairness and accuracy;
- **Consent:** For certain types of automated decision-making, organisations must obtain explicit consent from the individuals concerned.

One of the key provisions of GDPR is the right to explanation, which implies that individuals can request and receive an explanation of how a decision was reached by an automated system. This requirement has propelled the need for explainable AI (XAI) and interpretability in machine learning models, as organisations must ensure that their models not only perform accurately but also provide transparent and understandable justifications for their decisions.

The historical context of GDPR can be traced back to the European Data Protection Directive 95/46/EC, which laid the groundwork for data protection laws across Europe. However, the Directive was considered outdated in the face of rapid technological advancements and the increasing complexity of data processing activities. The GDPR was introduced to address these gaps, emphasising enhanced data subject rights, stricter consent requirements, and accountability for data controllers and processors.

The introduction of GDPR has intensified the focus on the explainability of machine learning models, particularly in sectors where decisions can have far-reaching consequences. Organisations are now required to balance the performance of their models with the ability to provide clear and comprehensible explanations, a challenge that has driven significant research and development in the field of explainable AI.

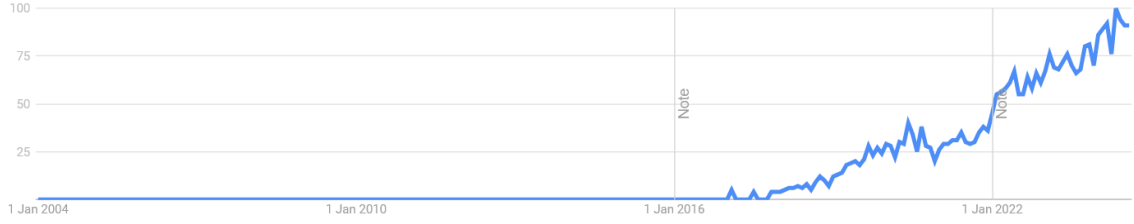


Figure 1.1: Interest over time in Explainability

As observed in the trend depicted in the attached graph, the interest in explainable AI has seen a substantial increase post-GDPR enforcement, reflecting the growing recognition of its importance in complying with regulatory requirements and building trust in AI-driven decisions. The Figure 1.1 shows a marked rise in searches and academic interest in topics related to explainability, interpretability, and transparency in AI, highlighting the critical role that these concepts play in modern data-driven decision-making processes.

In this context, the development of models that can not only detect outliers but also provide meaningful explanations for their identification is not just a technical challenge but a regulatory necessity. The ability to explain why a certain data point is classified as an outlier is essential for ensuring compliance with GDPR, particularly in high-stakes domains such as finance, healthcare, and security.

1.2 Motivation

The awareness of the potential information that can be extracted from a rare event has been steadily growing once the ability to provide meaningful explanations for outliers contributes to their value and utility in an endless range of domains. Therefore, it is not just about identifying outliers but also about comprehensively understand and being able to explain them. The definition of explanation is still an ongoing issue in the relevant literature, revealing a gap in understanding, interpretation and addressing the nuances associated with it. This gap presents a compelling opportunity to explore the research questions:

- Why are the current techniques not able to automate outlier prediction and explanations?
- Why are the current techniques failing to identify and explain outliers?
- Given that each outlier or rare event has a unique explanation, can we still identify the explanations into distinctive groups of explanations?

Delving into these challenges will set up the stage for this dissertation goal: How to improve explainability of rare events and improve the outlier understanding?

This dissertation will explore the intersection of outlier detection, explainability, and regulatory compliance, aiming to contribute to the development of methods that enhance the transparency and interpretability of models used in critical decision-making processes. By addressing the challenges posed by the GDPR and the growing demand for explainable AI, this research aims to answer the ultimate challenge of splitting the outliers' explanations into

groups of distinct explanations, and therefore to advance the state of the art in outlier detection and explanation, ultimately improving the reliability and trustworthiness of automated decision support systems.

In data analytics and machine learning, the identification and interpretation of rare events, often referred to as outliers or anomalies, have gained significant attention due to their crucial roles in various applications such as medical research, environmental monitoring, and fraud detection (Panjei et al., 2022). Outlier detection can be accomplished through a wide range of techniques designed to unveil unusual observations that differ from the standard regarding their features. In the past, the norm was to remove outliers to ensure that models fit the training data. However, outliers often contain valuable information, and their utility is greatly enhanced when accompanied by comprehensive explanations that cover both the outlier detection process and the explanation of the identified anomalies.

1.3 Structure

The dissertation is structured as follow: Chapter 2 reviews outlier detection and explanations, evaluation metrics and identifies research gaps in the literature. Chapter 3 outlines the methodology for outlier detection, extracting their explanations and clustering for deeper explainability. Chapter 4 discusses the findings on clustering XAI. Chapter 5 summarises the contributions and suggests topics for future XAI research.

Chapter 2

Related Work

The detection of outliers is not just about flagging rare events but about understanding the underlying reasons for their occurrence, moving from mere identification to explainability. This chapter starts by diving into the definition of outliers, the algorithms to detect rare events and the techniques to extract knowledge and explain the models applied. In the end, shows and dive into the evaluation metrics of this subject.

2.1 Outlier Detection

According to Boukerche et al. (2020), outliers are data points that exhibit significant deviation from the standard pattern or distribution and are also referred to as anomalies or abnormalities. These rare events exhibit distinctive characteristics that make them stand out from most instances, playing a crucial role in various applications, including medical research, environmental monitoring, and fraud detection.

Given these authors knowledge map in Fig. 2.1 it is possible to conclude that their work, as noted in Fig. 2.1a, has been associated with other authors, and influencing even other relevant authors to extend the knowledge on this field and its corresponding applications, as also noted on Fig. 2.1b.

The use of outlier detection algorithms allows the identification of exceptional observations, thereby providing valuable insights into abnormal behaviour within a system. To fully understand the findings from outlier detection, it is crucial to obtain explainable outcomes, which involve two fundamental tasks: outlier detection and outlier explanation.

Putting into perspective, in a dataset, rare events possess certain dissimilarities compared to standard occurrences, significantly differing from standard occurrences. Anomalies are characterised by a distinct underlying generative mechanism of the instances, whereas outliers are distinguished by their statistical rarity and deviation without necessarily having a different generative mechanism.

In earlier times, the conventional practice was to remove outliers in order to ensure that models would fit the training data. However, outliers tend to represent interesting information, stressing the need to develop outlier detection methods that can be applied to the various fields starting to adopt these methodologies and already cited.

Outliers can be categorised into two big groups based on the number of data instances and the type of input data. Based on the number of data instances, they can be point outliers, which refer to a single data instance that deviates significantly from the standard, or collec-

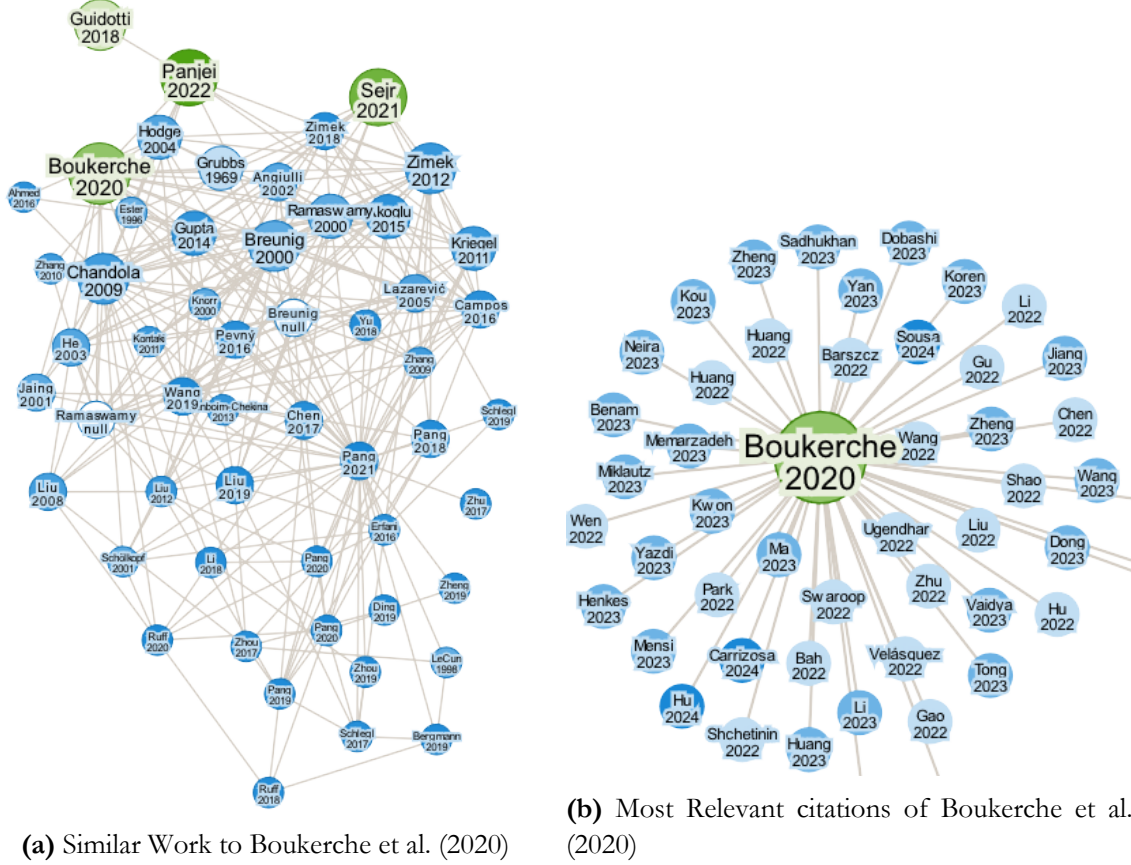


Figure 2.1: Boukerche et al. (2020) Knowledge Map

tive outliers, when there is a collection/group of instances that deviates from the rest of the dataset. However, each of these instances might not be an outlier individually. Point outliers can be further categorised as local outliers, where the deviation is noticeable compared to the nearest neighbouring instances, and global outliers, when considering the differences with the entire dataset. On the other hand, outliers can also be categorised according to the type of input data instances as vector outliers, which involve multi-dimensional data, or graph outliers, where the data is represented as graphs with nodes and edges that points the inter-dependencies among data instances, as shown in the diagram in the Fig. A.2.

Detecting an outlier poses a true challenge and a typical outlier detection task involves some stages. The first one is to identify the instances that deviate significantly from the rest, which imposes the question of what is the meaning of significantly. Let us take, for example, a pregnant woman, which is not considered an outlier by itself, and also a 13 year old, which is also not considered an outlier. However, a 13 years old pregnant woman is clearly identified as an outlier. On the other hand, another example is that there are solar explosions that create very strong magnetic fields affecting most of the electronic systems. Nonetheless, this is not an interesting outlier, although it fits the definition of outlier/anomaly. The following stages are, usually, subsampling and ensembling, density peak clustering, and deep learning.

We can also note that, typically, an outlier detection task involves a binary classification problem, which has inherently imbalanced data - with more instances that are standard than rare events. Additional questions have been researched, for example, should the data be

trained only with standard labels, should unlabelled data be used to build models for outlier score calculation or direct calculation of outlier scores.

Dealing with rare events can be summarised into 2 pillars: outlier detection, which includes phases such as identifying instances that deviate significantly from the rest, subsampling and ensembling, density peak clustering, and deep learning, and outlier explanation to provide insights into the detected anomalies.

After setting the definition of identifying what is significant, we can categorise outlier detection methods as supervised, semi-supervised or unsupervised, where it is hard to obtain labelled data to build models for outlier score calculation. Usually, supervised learning methods are preferred because they provide reliable guidance for modeling abnormal generative mechanisms, while unsupervised learning methods rely on the intrinsic distribution or structure of the data to detect outliers. The lack of labelled data poses some challenges and is also the reason why there are so many unsupervised methods. Given this characteristic, incorporating human feedback into unsupervised models tends to show improvements in the task.

Outlier Detection Algorithms

Per the relevant literature, outlier detection techniques can be arranged into fundamental and advanced approaches. Fundamental approaches are further split into proximity-based and projection based methods, as it can be observed in Figure A.1. Proximity-based methodologies identify outliers by considering their relations with nearby data, and employ techniques based on Nearest-neighbour (see Table A.2, e.g., LOF, COF, LOCI, etc.), or clustering algorithms to assess the extent to which an outlier is close to neighbouring data instances (see Table A.3, e.g., CBLOF, LDcof, etc.). Projection-based techniques transform the initial data into a novel space or structure that exhibits reduced complexity as locality-sensitive hashing (LSH) and Space-filling curves (see Table A.4, e.g., PINN, LSOD, Loda, Isolation Forest, Extended iForest, etc.).

Meanwhile, the emergence of some new challenges, such as data with extremely high dimensionality, unbounded large volumes of data streams, distributed data of large scales, promoted the development of advanced approaches techniques as High-dimensional Outlier Detection (see Table A.5, e.g., RBRP, PINN, ABOD, etc.), Data Streams Outlier Detection (see Table A.6, e.g., STORM, DUE, Incremental LOF, DILOF, etc., and Table A.8, e.g., D-Stream, AnyOut, etc.), Distributed Outlier Detection (see Table A.9), Deep learning-based Outlier Detection (see Table A.10, e.g., RDA, AnoGAN, etc.) and Outlier Detection with feedback (see Table A.11).

2.2 Explainable AI

The relevant literature includes the analysis of Guidotti et al. (2018), who named a black box as a system that hides its internal logic from the user, creating situations of lack of explanations alongside its consequences, both from a practical and ethical perspective. In order to overcome this challenge, many approaches have been researched, sometimes at the expense of compromising accuracy in favour of enhancing interpretability. Their knowledge map can be observed in Fig. 2.2 whereas their work has been related to others as seen in Fig. 2.2a and also its range of influence, Fig. 2.2b.

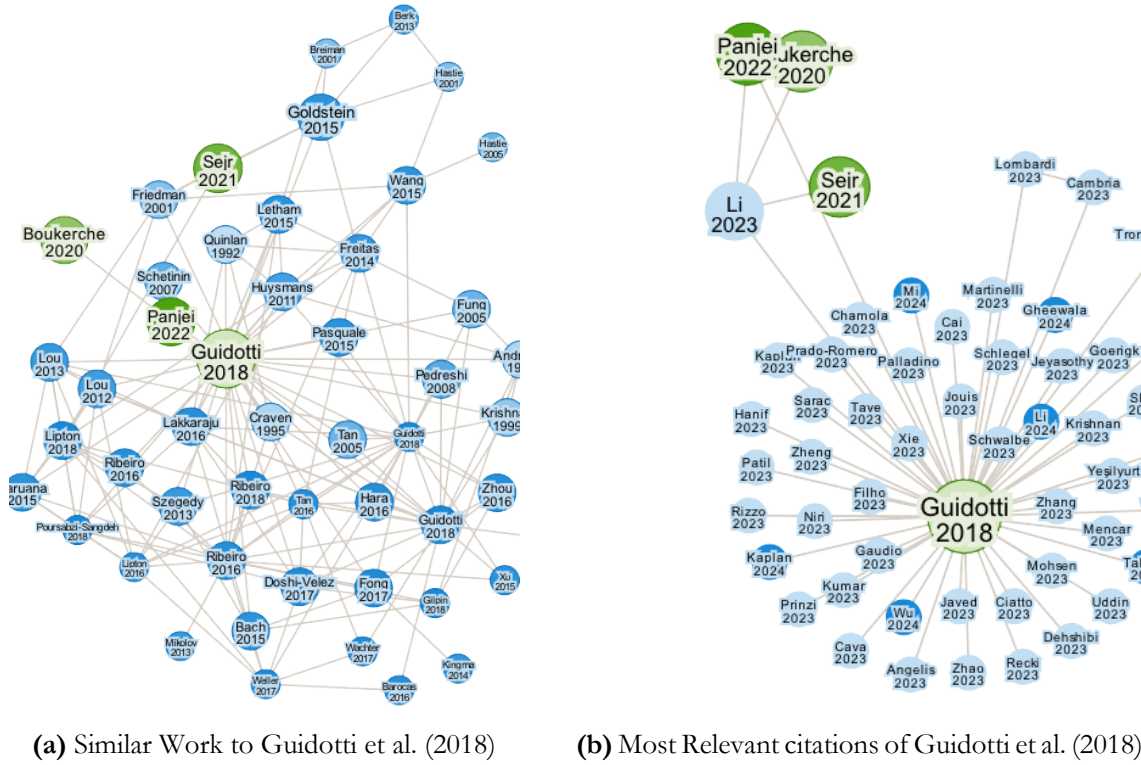


Figure 2.2: Guidotti et al. (2018) Knowledge Map

Over the last few years, with the proliferation of machine learning systems to support and predict sensitive information as credit score, health status, self-driving cars, among others, and also with the introducing by the General Data Protection Regulation (GDPR) of the right of explanation for individuals regarding automated decision-making systems, stressed the awareness and importance of understanding their logic in order to avoid ethical, safety, discrimination, and liability issues. Therefore, explanation technologies became fundamental to assure safer and more trustworthy solutions, and also to validate and accept the machine learning models applied.

Key questions in the literature include: What makes a model interpretable or transparent? What might be an explanation? When is a model or an explanation comprehensible? Which might be the best way to provide an explanation? Which problems might require interpretable models or predictions? What kind of decision data might be affected? Which type of data might be more comprehensible? What might be the best trade-off among prediction accuracy loss versus a gain in any form of interpretability? Considering aspects as interpretability, explanation comprehensibility, and prediction accuracy, a clear classification is stated by Guidotti et al. (2018) as needed to compile the knowledge about research investigating methodologies for opening and understanding the so-called black boxes by the authors.

One aspect to acknowledge is that training classifiers on historical datasets can uncover biases and prejudices that may be considered general rules, leading to unfair decisions. One example brought by the authors is in the image or text recognition, where accurate decision systems classifiers can result from accidental artefacts in the training data, such as differences in lighting conditions or backgrounds. When using Deep Neural Networks (DNNs), the performance is good on pattern-recognition tasks. Still, it can easily be fooled by subtle

perturbations that are invisible to humans, misleading the classifiers and raising doubts about trusting these decision systems.

Another relevant contribution to the starting point of the definition of explanation is presented in the research of Panjei et al. (2022). Their knowledge map, as their related studies, and most relevant citations, can be observed in Figure 2.3, Figure 2.3a, and Figure 2.3b, respectively.

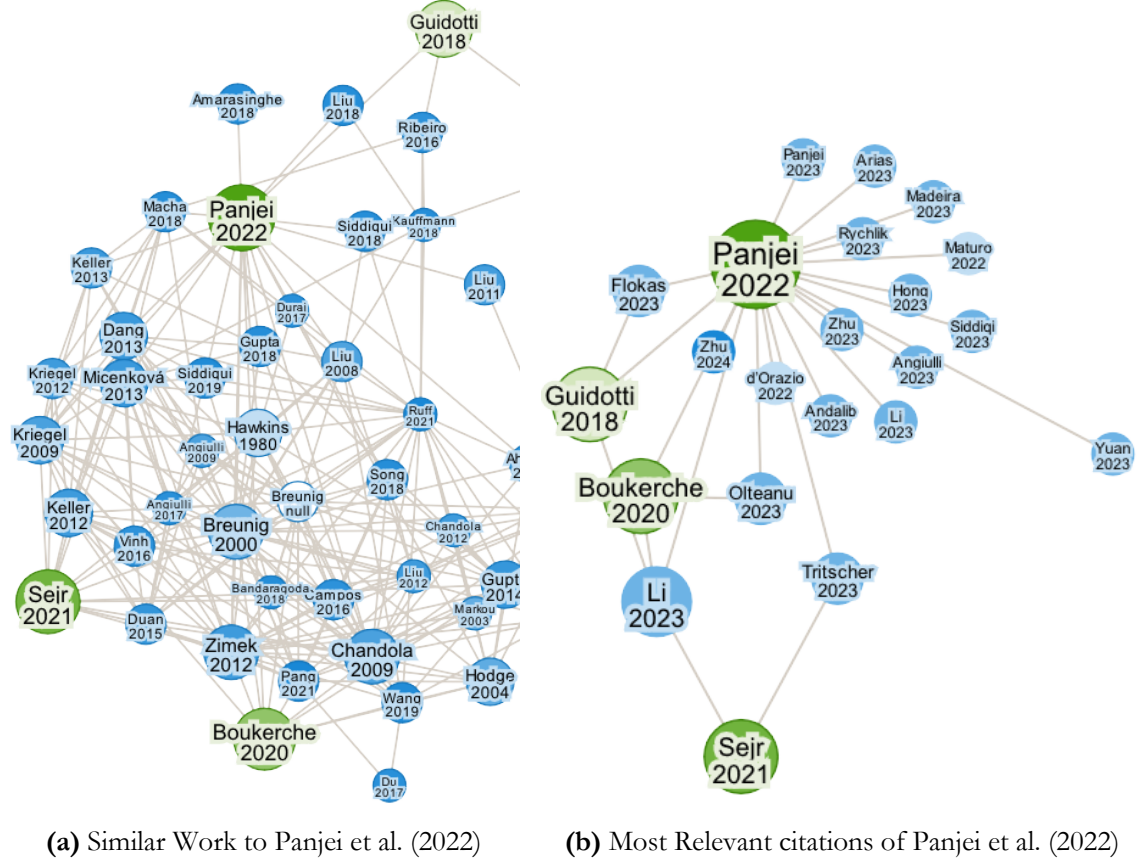


Figure 2.3: Panjei et al. (2022) Knowledge Map

Molnar (2018) emphasizes that interpretability is crucial for building fair, private, reliable, and robust machine learning models. Although there is still no consensus on a universal definition of "explanation" in machine learning, properties such as soundness, completeness, compactness, and comprehensibility are widely considered essential. Human-friendly explanations tend to be contrastive (highlighting significant differences between instances), selected (focusing on one or two causes, even when many exist), focused on the abnormal (emphasizing rare, low-probability events), social (tailored to the intended audience), truthful (aligned with real-world validity), consistent with prior beliefs of the explained since humans are inclined to disregard information that is irrelevant to their prior beliefs, and general and probable (explanations are good when they can explain many events).

Guidotti et al. (2018) identifies four categories of approaches for improving explainable AI: Model explanation - providing an interpretable and transparent model that mimics the behaviour of the black box and is understandable by humans; Outcome explanation - explaining the decisions of any classification method through an explanation vector, highlighting the influential features for a particular decision; Black box inspection - inspecting the black box

model to understand its internal logic and decision-making process; Transparent box design problems - designing a transparent model that is interpretable and provides explanations for its decisions.

The study also identifies that the biggest communities in the literature might be from machine learning, which focuses on describing how black boxes work, and from data mining, which, on the other hand, focuses on explaining the decisions even without understanding the details of how the opaque decision systems work in general.

From both perspectives, the communities try to get to an explanation with the following properties: Soundness, Completeness, Compactness, and Comprehensibility (how much an explanation is understandable for human beings?).

A final observation from (Guidotti et al., 2018) is that that explanations may serve as learning tools, potentially enabling the development of predictors directly from explanations.

Outlier Explanation in the literature

Although there are many survey papers on outlier detection techniques, there is a lack of literature specifically focused on outlier explanations. Panjei et al. (2022) presented a comprehensive survey on outlier explanations, including different types of explanations, challenges in generating them, existing techniques, applications, evaluation methods, and future research directions. They define three categories of non-evaluative aspects of types of outlier explanations: Importance level of outliers, helping to prioritise the effort on investigating the outliers and serving as a contrastive explanation (ranking, either numerical or categorical ranking; position; highlighting the most significant differences between the object of interest and other objects); Causal interactions among outliers (corresponding timestamp order of the outliers - removing one also removes the other one?; which outlier implies other outliers - even if a real-world event has many causes, people usually seek only one or two of these as explanations, it asks if preventing those would lessen other outliers, i.e., would y have been a normal instance if x had been eliminated?); Outlying attributes of outliers (referring to the feature subspace that contributes the most to the deviation from the standard for either an individual or a group of outliers, which has a contrastive characteristic; explanation focuses also on the abnormal where features have abnormal values that make an instance a rare event).

Current techniques do not address all the challenges necessary for enabling effective, application-specific actions. This underscores the need for further research, particularly in developing methods that incorporate additional data references to enhance the automatic generation of outlier explanations, as already stressed in Mozolewski, Bobek, Ribeiro, Nalepa, and Gama (2024).

Considering these aspects, the key challenges in generating outlier explanations include:

- Challenges in generating outlier rankings include unifying different outlier scores, incorporating user feedback to reduce false positives, defining appropriate importance criteria, and selecting efficient mapping functions for outlier categories (see Table A.14, e.g., GLADs, COIN);
- In generating causal interactions among outliers, challenges include ensuring the availability of timestamps and defining data structures that capture these causal relationships;

- For identifying outlying attributes of both individual and group outliers, the challenges involve limiting the subspace search, generating interpretable output with fewer attributes, integrating user knowledge into algorithms, and identifying discriminative attributes (see Table A.15, e.g., LODI, Refour, Explainer, LOGP, Exstream, COIN, EXAD, Lookout, SFE, URCA, KKAS, XPACS).

Panji et al. (2022) presented a summary on the techniques for outlier explanations, as per Table A.13.

Several techniques have been already mentioned by many authors, including Guidotti et al. (2018), as Partial Dependency Plot (PDP) - which assumes weak correlation between the independent features, Individual Conditional Expectation (ICE) - which explains the whole model, Combined Multiple Models (CMM) - which applies surrogate model, Local Interpretable Model Explanation (LIME) - that describes a single prediction with a linear model trained locally, SHapley Additive exPlanations (SHAP), that has a high computational cost to compute the contribution of each feature, among others.

Sejr and Schneider-Kamp (2021) contribute further by highlighting the need to integrate outlier explanations into everyday work practices for data scientists and end users, bridging the gap between theoretical frameworks and real-life applications and addressing the challenges posed by unexplainable, unsupervised outlier algorithms.

Explanation Algorithms

The number of decision systems explanation methods is high, but it is very challenging to find one that could fit any problem. Guidotti et al. (2018) summarised it regarding some of its characteristics, as the problem (Model Explanation, Outcome Explanation, Black Box Inspection, Transparent Design), the explainer (DT - Decision Tree, DR - Decision Rules, FI - Features Importance, SM - Saliency Masks, SA - Sensitivity Analysis, PDP - Partial Dependence Plot, NA - Neurons Activation, PS - Prototype Selection), which black box (NN - Neural Network, TE - Tree Ensemble, SVM - Support Vector Machines, DNN - Deep Neural Network, AGN - AGNostic black box), data type (TAB - TABular, IMG - IMaGe, TXT - TeXT, ANY - ANY type of data), the range of application (indicating if an explanatory approach can be generalized for every black box), and random (indicating if any kind of random perturbation or permutation of the original dataset is required for the explanation), as seen in Table A.12. Some of the techniques are Tree Metrics, SVM+P, GoldenEye, PALM, CAM. LIME, ICE, among others.

2.3 Evaluation Metrics

Evaluating explanations for outliers involves assessing how well they meet user needs, their accuracy, coherence with prior beliefs, and their ability to be generalised. Per the relevant literature, various techniques employ different metrics for this evaluation. For example, ROC AUC is often used to evaluate outlying attributes in both datasets and real-world challenges with ground truth, while the accuracy rate assesses the quality these attributes and their impact on outlier detection. Metrics such as AUC measure the accuracy of the explanation task outputs (the number of true anomalies found in the top-ranking explanations), and F-measure evaluates consistency between ground truth and generated outlying attributes, as well as the predictive power of explanations. While some techniques use real-world data to

analyse causal interactions among outliers, they do not provide precise figures for the number of matched causal events. The evaluation of the distinctive attributes of individual outliers is accomplished through feature visualisation, the identification of anomalous attributes in image datasets, and the measurement of accuracy using the Jaccard Index and Precision (which measures the accuracy rate of explanation task outputs), the minimum number of attributes required to accept a detected outlier as a true positive -which evaluates outlying attributes, incrimination score (that also evaluates outlying attributes), and execution time (which focuses on assessing outlying attributes of a group of outliers). Panjei et al. (2022) also summarised some methods for evaluating outlier explanations.

Evaluation metrics of the explanations, or as we start to see Explainable Artificial Intelligence (XAI) has also been pushed by the field of Predictive Maintenance, namely by the evaluation of explanations generated by Explainable Predictive Maintenance (XPM), as summarised in Cummins et al. (2024). The evaluation of these methods is crucial for ensuring the quality, reliability, and usefulness of these explanations, and the authors ended up labelling evaluation metrics into objective and subjective. In the context of streaming data and modular systems, XAI has been further enriched by recent studies, especially Mozolewski et al. (2024) research, which delves into the challenges of ensuring that AI models remain interpretable and trustworthy in dynamic environments, introducing new evaluation metrics that indicates how well these explanations adhere to expert-derived rules or recognised ground truths. This study also highlight the need for XAI methods that can balance the trade-offs between fidelity and stability, providing explanations that are both trustworthy and understandable in dynamic settings.

The objective evaluation metrics quantify the properties of explanations in a quantitative manner, independent of the user perception. These metrics focus on the internal characteristics of explanations, such as their accuracy, consistency, and sensitivity to changes in input data.

- Metrics that refer to the dependence from the Input Data:
 - Explanation Stability (S) measures how resilient explanations are to input variations. A high S value indicates that the explanations remain consistent across different data points;
 - Sensitivity, measuring the responsiveness of explanations to changes in input data. A high sensitivity suggests that the explanations are reliable and can be applied to make accurate predictions;
 - Robustness, evaluating the consistency of explanations for similar input data. A high robustness indicates that the explanations capture underlying patterns in the data, even with variations in input features;
 - Lipschitz Stability, which evaluates the consistency of explanations over time or across similar data points, which is crucial in streaming data where the model's behaviour may change due to concept drift. A stable XAI method ensures that explanations do not fluctuate wildly with each new data batch, maintaining user trust over time.
- Model-based Metrics:

- Model-Performance-Explanation-Performance Difference (D), measuring the discrepancy between the model’s prediction and the explanation’s prediction. A high D value indicates that the explanation is inaccurate or misleading;
 - Number of Rules (R), which quantifies the complexity of the explanations. A higher R-value suggests more detailed and granular explanations;
 - Number of Features (F), measuring the number of features used to generate the explanations. A higher F value implies more comprehensive explanations.
- Explanation-based Metrics (complexity and interpretability):
 - Monotonicity, which assesses the consistency of explanations regarding feature importance. A high monotonicity value implies that the explanations accurately identify the most influential features of the decision-making process;
 - Explanation Correctness, evaluating the overall accuracy and truthfulness of the explanations and is assessed using techniques such as sensitivity, fidelity, and generalisability;
 - Fidelity, which assesses the accuracy of the explanations in representing the underlying decision-making model process, using techniques such as sensitivity, monotonicity, and robustness. High fidelity can be challenging in some environments, as streaming data, but continuous high-fidelity explanations are important even as the model adapts to new data;
 - Generalisability, measuring the ability of explanations to generalise to other data points. A high generalisability indicates that the explanations can be applied to explain a wider range of scenarios. ;
 - Comparative Expert Stability Index (CESI), a novel metric that compares different XAI methods against expert knowledge, particularly useful in domains where human expertise is critical. CESI contributes in selecting the most appropriate XAI method for a given application by measuring how closely machine-generated explanations match those of human experts.

Subjective metrics, on the other hand, assess the perceived quality of explanations from the user’s perspective, focusing on factors like

- User Trust, measuring how much users rely on the explanations, often assessed through surveys or interviews;
- Explanation Effectiveness, which determines how well the explanations assist users in decision-making or understanding the model’s behavior, using task-based performance assessments and user feedback;
- Explanation Satisfaction, measuring user comprehension and ease of use, often evaluated through questionnaires or user studies.

Another interesting contribution to XAI evaluation is presented in the paper of Sokol and Vogt (2024) which argues that for a more nuanced and context-sensitive approach to XAI evaluation, recognising the modular and sociotechnical nature of these systems, as also, its utility, proposing a Sociotechnical Utility-Based Evaluation Framework that emphasises:

- Functionally Independent Components, i.e., recognise and evaluate the individual components of XAI systems (e.g., algorithmic explainers, explanatory artefacts, communication protocols) separately and in combination;
- Multi-Level Evaluation, or evaluating XAI systems at multiple levels, such as individual component evaluation, module-level evaluation, and end-to-end system evaluation;
- Context-Specific Evaluation, or the need to tailor the evaluation process to the specific context in which the XAI system will be used, considering factors like the user's expertise, task complexity, and operational environment.

By integrating insights from these studies, the evaluation of XAI systems can be significantly enhanced, leading to more reliable, trustworthy, and context-appropriate explainability tools that better meet users' needs across various domains. These approaches offer a comprehensive framework for assessing the effectiveness of XAI methods, particularly in dynamic environments like streaming data, where the ability to maintain accurate and stable explanations over time is crucial.

Despite the advancements in explainable AI, current techniques for outlier explanation still face challenges, as the trade-off between accuracy and interpretability, as well as the lack of an unified framework. The literature suggests that there is an inherent trade-off between model accuracy and interpretability. Methods that prioritise explainability, such as decision trees or rule-based systems, often do so at the cost of predictive accuracy. On the other hand, more accurate models like deep neural networks and ensemble methods tend to be less interpretable, making it difficult to provide reliable explanations without sacrificing performance. Also, there is currently no unified framework for outlier explanations. Different methods handle explanations in distinct ways, with no standard for comparing or evaluating the quality of these explanations across methods. This lack of standardisation creates challenges to ensure consistency or validate the explanations across different models or datasets. This dissertation aims to address these gaps by focusing on model performance to achieve high levels of interpretability. It will propose a foundational framework for detecting outliers, generating explanations, and deriving rules that can be used to interpret and predict observations.

Chapter 3

Methodology

The methodology of this dissertation begins by describing the data used (Section 3.1), followed by the preprocessing techniques applied to prepare the data (Section 3.2). Next, the outlier detection algorithms are presented (Section 3.3), along with the explanation methods used to interpret the outliers (Section 3.4). The explanations are then clustered (Section 3.5), and finally, the evaluation metrics are discussed, including ROC-AUC (Section 3.6.1) and other relevant metrics (Section 3.6.2) used to assess the effectiveness of the models (Section 3.6).

3.1 Data

The data methodology to be employed in this dissertation is designed to make use of the most extensive collection of outlier detection datasets, which might have labelled data, and is available at Outlier Detection DataSets Repository (ODD). This repository is notable as it is the largest collection of outlier detection datasets with ground truth, specifically designed for outlier detection tasks.

Making use of labelled data from the ODD repository is strategic once labelled datasets pose a challenge for the implementation of supervised outlier detection methods, where models are trained on known instances of outliers and non-outliers. This facilitates the development of accurate models and a more comprehensive understanding of the performance of outlier detection algorithms.

ODD repository offers 5 categories of datasets, multi-dimensional point datasets (one record per data point, and each record contains several attributes); time series graph datasets for event detection (temporal graph data where the graph changes dynamically over time in which new nodes and edges arrive or existing nodes and edges disappear); time series point datasets (Multivariate/Univariate), temporal point data where each point has one or more attributes and the attributes change over time; adversarial/attack scenario and security datasets (fraud detection data from online review system - Cyber security data, e.g. intrusion detection with DoS, DDoS etc. attack scenario); crowded scene video data for anomaly detection (video clips acquired with camera).

For the purpose of this dissertation, multi-dimensional point datasets will be used. Inside this category, there are some possibilities as seen in Table A.1, and the datasets to be used are the Breast Cancer, Seismic and Cardiotocography, chosen for their different proportions of outliers in the dataset.

Furthermore, to enhance the comprehensiveness of the research, the methodology will include an innovative approach - the development of an outlier data generator. This generator will aim to ensure that the dissertation has explanations for outliers from the beginning, and by generating synthetic outlier data, the research can systematically control the characteristics, features, and distribution of outliers within the datasets. This approach gives a distinctive advantage as it allows a more thorough examination of the outlier detection models, their explanations, and the robustness of the techniques used.

Developing an outlier data generator represents an innovative aspect of the methodology. This approach addresses the need for a controlled environment in outlier detection research. By creating synthetic outliers, the dissertation can explore various scenarios, including those with known explanations *a priori*, thereby contributing to a deeper understanding of the strengths and limitations of different outlier explanation techniques.

3.2 Data Preprocessing

The preprocessing phase begins by importing the corresponding dataset, including the outlier events which will work as ground truth for the tests.

After loading the data, categorical features (non-numeric) are identified and LabelEncoder is used to convert categorical features into unique integers, assigning a different integer to each category within a feature. To ensure that all numeric features are on the same scale, the MinMaxScaler is applied, assuring that distance-based algorithms won't perform poorly because of the existence of different scales. In addition to normalisation, Standard Scaling is applied, which standardises features by removing the mean and scaling to unit variance, specifically before clustering in order to avoid issues on the algorithms sensitive to feature variance. The issue of class imbalance is addressed through SMOTE, which creates synthetic samples for the minority class, helping to balance the dataset, reducing bias and enhancing overall performance. In some cases, feature selection was performed, in order to reduce dimensionality and remove potential noise within the models.

In this study, SMOTE always presented an inferior precision, recall, F1-score and, most importantly, inferior ROC-AUC performances, as can be seen in Tables A.16, A.18, A.21, A.17, A.19, A.22, and in Figures A.3, A.5 and A.7. That way, it was not applied while preparing the data for the final results in this dissertation.

3.3 Outlier Detection

For outlier detection, this study incorporates multiple models, each offering unique advantages in identifying anomalies within the data. The models used include Isolation Forest, Random Forest, and Local Outlier Factor (LOF).

The Isolation Forest algorithm (Liu, Ting, & Zhou, 2008) isolates anomalies by focusing on how easily points can be isolated from the rest of the data. Instead of modelling normal instances, the algorithm randomly selects a feature and splits the data between its minimum and maximum values. Once anomalous points, are rare and different, this require fewer splits to isolate, making them easier to identify. The key parameters in this algorithm tuned were *n_estimators*, which specifies the number of trees to build in the forest (trade-off between high number of trees to improve performance and computational cost) and contamination, which defined the expected proportion of outliers in the dataset, adjusting its sensitivity. The

Isolation Forest is particularly advantageous when dealing with large datasets containing many outliers because it uses small subsamples of the dataset to construct trees, which enhances the diversity of the ensemble and improves the accuracy of outlier detection. This subsampling approach also mitigates two key challenges: swamping (when normal instances are mistakenly identified as outliers) and masking (when closely clustered outliers are difficult to detect). By using subsamples, Isolation Forest becomes highly efficient, operating with linear time complexity and without needing to calculate pairwise distances, allowing a faster processing, even with large datasets, while maintaining high accuracy in detecting outliers. In this analysis, outliers were identified as observations with abnormal attribute values that differ significantly from typical (non-outlier) instances in the dataset. In one sentence, Isolation Forest asks questions until it can guess the outlier, being an advantage when there are a lot of outliers. In the datasets analysed, the best performance of Isolation Forest corresponded exactly to the one with the highest proportion of outliers.

The Random Forest algorithm is typically used for classification and regression tasks, but it can also be adapted for outlier detection by analysing anomalies in the predicted values. In this approach, RF constructs multiple decision trees during training and averages their predictions, therefore, outliers are identified based on how often instances deviate from the majority of trees' predictions. For outlier detection, RF analyses deviations in the predictive values across its trees and instances that are consistently misclassified or assigned significantly different values compared to the rest of the dataset are flagged as potential outliers. In this analysis, the RandomForestClassifier was used with *n_estimators=100* (100 trees used to enhance the robustness of the predictions) and *class_weight='balanced'*, which automatically adjusts weights to handle class imbalance and give more importance during training.

The Local Outlier Factor (LOF) algorithm (Breunig, Kriegel, Ng, & Sander, 2000) measures the local density of data points and compares each point's density with that of its neighbours. If a point has a significantly lower density than its neighbours, it is considered an outlier. Basically it checks how popular (similar) an instance is when compared to its neighbours: if it's not popular, then it is an outlier. In this dissertation, LOF was tuned to reflect the expected contamination of the dataset and given the number of outliers in each one, the number of neighbours was defined as 20.

After training the models, their effectiveness in detecting outliers is evaluated using the AUC-ROC curve, a crucial metric for anomaly detection tasks (Liu et al., 2008), evaluating the trade-off between true positive and false positive rates and providing a consistent indication of how well the models distinguish between outliers and normal instances. In this dissertation, the ROC-AUC serves as the primary metric to evaluate model performance and also forms the basis for selecting the best-performing model to generate explanations in the subsequent stages of the analysis.

3.4 Explainable AI

LIME: Local Interpretable Model-Agnostic Explanations

LIME consists of a methodology for interpreting predictions made by black box machine learning models. The main objective of LIME is to develop localised surrogate models, which are interpretable models trained to approximate the predictions of the underlying black box model for individual instances. Unlike global surrogate models, LIME focuses on explaining

individual predictions instead of providing a global interpretation.

The fundamental concept of LIME revolves around understanding the logic behind a specific prediction made by a machine learning model through the analysis of variations in the input data. This is accomplished by generating a new dataset of perturbed samples and their corresponding predictions from the black box model. Subsequently, local surrogate models such as linear regression or decision trees are trained using this new dataset. The interpretability of the model is optimised locally, meaning that it accurately approximates the predictions of the black box model for a particular instance. Mathematically, LIME formulates this problem as an optimisation task that considers a loss function, measuring the proximity of the explanation model to the original prediction while simultaneously minimising model complexity.

To train the local surrogate model, LIME perturbs the dataset around the instance of interest and generates predictions using the black-box model. These new predictions assign weights based on their proximity to the instance of interest. Consequently, a weighted and interpretable model is trained on this dataset with perturbations, thereby providing a localised approximation of the black box model's predictions.

LIME offers users flexibility in choosing the type of interpretable surrogate models, such as linear regression or decision trees. Also, users are requested to manage the complexity of the model, i.e., by defining the maximum number of features that can be used. Simplicity is often chosen to enhance interpretability. The process involves selecting the number of features (K) and employing techniques like Lasso with regularisation parameters or forward/backward feature selection to train models with the desired number of features.

The generation of data perturbations depends on the type of analysed data. In the case of text and images, LIME selectively activates or deactivates individual words or super-pixels. For tabular data, new samples are created by perturbing each feature individually, being the perturbations drawn from a normal distribution using the mean and standard deviation derived from the respective feature.

LIME presents a systematic approach to interpreting predictions made by black box models through the localised approximation of these predictions using interpretable surrogate models. As described by Boukerche et al. (2020), LIME can improve interpretability and promote a deeper understanding of the decision-making process employed by complex machine learning models by focusing on individual instances.

Pros LIME can be adapted for tabular, text, and image data, making it a widely applicable tool for different domains. LIME identifies key contributing words in text-based models or highlights important regions in image data, making it adaptable across domains. In cases where the underlying machine learning model is replaced, the same local, interpretable model can still be used for explanation purposes. Assuming that the individuals examining the explanations possess a strong understanding of decision trees, they can be employed as explanations without the need to use them for predictions, e.g., an SVM can be used. Moreover, it is a model-agnostic, meaning it can be applied to any machine learning model—regardless of whether it's a decision tree, a neural network, or a SVM. This versatility is important for practitioners who need to explain a variety of models across different tasks and if it is discovered that an xgboost model performs better, the SVM can be substituted while still using a decision tree to explain the predictions. When Lasso or short trees are used, the resulting explanations are succinct (or selective) and potentially contrastive, meaning they are human-friendly, as

described before, making LIME predominantly observed in applications where the recipient of the explanation is either a layperson or an individual with limited time. However, it is not adequate for complete attributions, thus it is not commonly used in compliance scenarios where a full explanation of a prediction may be legally required. Additionally, for debugging purposes of machine learning models, it is better to have all the reasons instead of just a few. The fidelity measure, which evaluates how well the interpretable model approximates the predictions, provides valuable insight into the reliability of the interpretable model in explaining the predictions in the neighbourhood of the specific data instance of interest. Explanations created using local surrogate models can incorporate other (interpretable) features that were not used in the original model's training. However, these interpretable features must be derived from the data instances themselves, for example, a regression model may employ a non-interpretable transformation of certain attributes, but the explanations can be generated using the original attributes. Yet, a regression model might be trained on components of a principal component analysis (PCA) of an object of study, while LIME can be trained on the original object of study.

Cons On the other hand, as a disadvantage, it can be noted that the precise characterisation of the neighbourhood poses a formidable challenge when employing LIME with tabular data. According to Molnar (2018), this downside represents the most significant obstacle associated with LIME, thereby, the caution needed when using it. For each specific application, it becomes imperative to experiment with diverse kernel configurations and personally assess the intelligibility of the explanations. While local explanations can be very useful, they may fail to capture the overall behaviour of the model, especially in cases where the decision boundaries are complex or highly non-linear. The complexity of the explanation model must be predetermined, although this concern is relatively minor, it remains incumbent upon the user to ascertain the optimal balance between fidelity and sparsity. Yet another substantial issue is the instability of the explanations, as the explanations for two closely situated points exhibited considerable disparities, as Sejr and Schneider-Kamp (2021) point out, LIME struggles to define a meaningful neighbourhood for rare or outlier events, which are inherently far from the majority of the data distribution, making LIME less reliable when used in outlier detection or in cases where the data distribution is highly imbalanced. Meanwhile, LIME is not prone to the possibility of being manipulated in order to overcome biases, increasing the difficulty of trusting the explanations generated. As this methodology remains in the development stage, the user should always consider overcoming the numerous challenges described. The concept of neighbourhood and the sampling method represents a challenge to the application of this methodology when it comes to rare events, which are already deviations from the standard and possess abnormal neighbourhood characteristics.

LIME is a versatile, human-friendly model-agnostic explainability tool that excels in providing localised insights into black-box predictions. However, practitioners need to be mindful of its limitations, particularly around sampling, parameter tuning, and its computational cost.

SHAP: SHapley Additive ExPlanations

SHAP, introduced by Lundberg and Lee (2017), is a method designed to explain individual predictions in machine learning models. It is based on Shapley values from game theory, ensuring a solid mathematical foundation for attribution. Each feature's contribution to a

model's prediction is calculated by considering all possible combinations of features, providing a fair and consistent measure of feature importance.

Pros SHAP can be used for both global explanations (explaining overall model behavior) and local explanations (explaining individual predictions). This flexibility is crucial when providing insights at different levels, whether it's for model developers or end-users. Moreover, SHAP works across various machine learning models, including tree-based models like Random Forests, Gradient Boosting, and XGBoost, as well as neural networks. This adaptability is particularly beneficial when switching between models while maintaining consistent explanations. Furthermore, it enables transparency in black-box models, such as deep learning models, by allowing each feature's contribution to be visualised and explained, as this makes it possible to understand complex, non-linear relationships that models learn from the data. Unlike other techniques as LIME, SHAP offers attribution consistency. The sum of all the SHAP values (across features) for a specific prediction equals the difference between the model's output for the given instance and the average output, ensuring the entire explanation remains balanced. Equally important, SHAP is widely adopted in industries like finance, healthcare, and law, where explainability is a legal requirement due to its rigorous framework and broad applicability.

Cons On the other hand, one of the main limitations of SHAP is its computational cost, as calculating exact Shapley values requires evaluating every possible subset of features, which grows exponentially with the number of features. Although approximation techniques like TreeExplainer exist for tree-based models, it remains expensive for complex, large datasets or models. Also, SHAP needs access to the internal workings of the model for precise attributions, as it can be challenging in black-box models or proprietary environments where full access to the model is not available, and, when working with high-dimensional data, SHAP's effectiveness might reduce due to the complexity of explaining large numbers of features, so it leads to less intuitive visualisations and explanations. In scenarios with numerous features, the contribution of individual features might become diluted or difficult to interpret. While SHAP values provide rigorous explanations, they can be hard to interpret for non-expert users and complex visualisations and mathematical concepts behind Shapley values make them less suited for applications where the end-user is a layperson. In certain models or situations, the SHAP approximation methods (like KernelExplainer for models other than tree-based ones) may lead to suboptimal or noisy explanations. This is particularly relevant when the model has a large number of features, where simplifications are made to expedite the explanation process. Finally, as Molnar (2018) explains, SHAP can run into sampling problems when the dataset is highly imbalanced. In these cases, the Shapley values might not adequately represent minority class behaviour, as the sampling process may not fully capture rare events, leading to poor representation of outliers.

Computing Shapley values to elucidate the contributions of individual feature values in a machine learning model focuses on determining the impact of including a specific feature on the model's predictions when combined with other features. The computation involves simulating various combinations of features, calculating their marginal contributions, and then averaging these contributions to obtain the Shapley values.

There are several key steps to computing it. First, the estimation of Individual Contributions that include that the assessment of a feature's contribution is conducted by simulating its

addition to different combinations of other features, which is achieved by randomly sampling from the dataset and replacing the feature values in the combination. Second, the Shapley values are computed by taking the average of the marginal contributions across all possible combinations, meaning that the computational complexity increases exponentially with the number of features, being important to optimise this process by computing contributions for only a limited number of samples of possible combinations. Third, the determination of the Shapley value for a specific feature should represent all the possible combinations, which can vary in size, ranging from those with no feature values to those that combine multiple features. Fourth, Shapley values offer insights into the contribution of each feature value to the predictions, framing this contribution as comparing a feature value and the average prediction for the entire dataset.

Shapley values are employed to interpret predictions in diverse datasets, and these interpretations involve understanding how each feature value contributes to the predictions relative to the average prediction for the dataset. SHAP values reflect the contribution of each feature to the prediction, where a negative SHAP value means the feature is contributing to a lower prediction score (pushing toward classifying the instance as an outlier in this context). Therefore, Shapley value represents an average contribution but not the difference in the prediction that would arise if a feature were removed from the model.

SHAP introduces an additive feature attribution method, representing the Shapley value explanation as a linear model. The properties of Shapley values, including Efficiency, Symmetry, Dummy, and Additivity, are also satisfied by SHAP. Additionally, SHAP introduces three desirable properties: Local accuracy, Missingness, and Consistency. KernelSHAP, one of the new estimation approaches, estimates contributions by sampling coalitions, computing predictions for each coalition, and fitting a weighted linear model. The SHAP kernel is introduced to determine the weight of each coalition. TreeSHAP, designed for tree-based models, reduces computational complexity when compared to KernelSHAP. The SHAP feature importance is based on the mean absolute Shapley values per feature, while global interpretations are based on SHAP summary plots and SHAP dependence plots.

SHAP offers several visualisation techniques to help interpret feature contributions to model predictions: SHAP Summary Plot (e.g. in Figure 4.3), which provides an overview of the SHAP values for all features across the dataset and each point represents a SHAP value for a feature in a specific instance with colours indicating feature values, with a wider spread indicating a greater impact on the model's output; and Force Plot, which displays the contribution of each feature to a single prediction. A Force Plot is composed by a *Base Value* (on the right side of the plot), which is the average model output across the entire dataset and the starting point for the explanation; a prediction $f(x)$ (in the center) which is the final prediction for the specific instance being examined; *Red Arrows* (Pushing the Prediction Higher) represent features that are pushing the prediction higher than the base value, which have a positive contribution to the final prediction as an outlier; *Blue Arrows* (Pushing the Prediction Lower) represent features that are pulling the prediction downwards from the base value, having a negative contribution to the final prediction as an outlier; and *Feature Contributions* (labels along the arrows) where each arrow is labelled with a feature and its corresponding value, indicating how much that feature is contributing to the deviation of the prediction from the base value.

Finally, advanced interpretation techniques are computed such as: Comparative Analysis, while comparing SHAP values between different outliers to understand how different features influence the model's prediction across various instances (e. g. Figs. 4.4 and 4.5);

and Aggregate Analysis, by aggregating SHAP values by clusters and plotting them, we can get an holistic view of what drives different clusters of outliers (e. g. Figs. 4.14 and 4.15).

These data visualisation techniques are great to communicate complex information to non-technical audiences.

In the context of outlier detection, SHAP assigns each feature an importance value for a specific prediction. This is especially useful for understanding the decision-making process of models like Isolation Forests or other anomaly detection methods. TreeExplainer is used for tree-based models and provides insights into the contribution of each feature to the decision made by the model and it's computationally efficient and accurate for such models. Kernel-Explainer is a commoner approach that can provide explanations regardless of the algorithm type, making it versatile for use with any model but at a computational cost.

For tasks such as outlier detection, SHAP has proven useful, as highlighted by Boukerche et al. (2020). It provides explanations for why certain instances are classified as outliers, thus supporting anomaly detection models like Isolation Forests. However, as noted by Panjei et al. (2022) while SHAP is powerful, it might not always provide explanations that generalise well for outlier detection, especially in datasets with rare event characteristics.

One work that starts to analyse anomalies with clustering is from (Kopp, Pevný, & Holena, 2014). After applying The Sapling Random Forest (SRF) is used to interpret rare events as collections of features in which anomalies can be easily perceived from the rest of the observations, or as combinations of distinct conditions that can be perceived as antecedents of logical rules that are readily comprehensible to humans.

3.5 Clustering the explanations

This section of the dissertation intends to deepen the understanding of different types of rare events by grouping similar anomalies based on the explanations obtained previously. The clustering of these explanations will allow the identification of patterns and rules that govern certain groups of outliers, offering a systematic approach to categorising anomalies. Once the outliers have been identified and explained, clustering techniques are applied to group these outliers based on their feature explanations, being a crucial step in interpreting the dataset, as it highlights similarities and differences in outlier behaviour. In this dissertation, one of the ultimate goals is to categorise each of the groups with possible rules and features values or conditions in order to gain an advanced understanding of the dataset. Before applying any clustering technique, the SHAP values are standardised using the StandardScaler, which is essential because clustering algorithms such as K-Means and DBSCAN rely on distance metrics, and scaling ensures that all features contribute equally to the computation of distances. Then, based on the SHAP values or the features conditions (in the case of LIME), clustering is computed recurring either to K-Means or DBSCAN algorithms (Ester, Kriegel, Sander, & Xu, 1996). DBSCAN groups similar data points and the instances that do not belong to any group are outliers. Advantages: to find outliers that are very different from the rest. These authors introduced a new clustering algorithm called DBSCAN in order to overcome the limitations of using clustering algorithms for class identification in spatial databases. The two types of clustering algorithms are partitioning and hierarchical. The first ones divide a database into clusters using an iterative control strategy and an objective function, while the last ones create a tree-like structure by splitting or merging clusters. The DBSCAN algorithm discovers clusters and noise in a spatial database, defining clusters based on the density of

points within a given radius and a minimum number of points and the shape of the neighborhood is determined by the choice of distance function. It starts with an arbitrary point and retrieves all points that are density-reachable from it based on a specified distance threshold (Eps) and minimum number of points (MinPts). K-Means method groups the outliers into a predetermined number of clusters by minimising the variance within each cluster, which allows to identify groups of outliers that share similar explanatory features.

The outputs of this clustering process are the statistics and descriptions of each clusters (e. g. Table 4.3), its distribution by feature and cluster (e. g. Figs. 4.14, PCA plots (to visually represent the SHAP value clusters in 2D space, showing the separation and similarity between different types of outliers (e. g. Fig. 4.16, and 4.15).

3.6 Evaluation Metrics

In the field of machine learning, particularly in anomaly or outlier detection, selecting the right performance metric is essential to assess the model's predictive power.

3.6.1 ROC-AUC

One objective metric that will be used on all the results will be based on Predictive Performance, such as accuracy or the area under the ROC curve (AUC).

Using ROC-AUC as a metric for evaluating predictive models, particularly for outlier detection and explanations, is widely accepted due to its comprehensive evaluation of both true positive and false positive rates across various decision thresholds.

Understanding ROC-AUC

ROC curve plots the true positive rate (TPR, also known as sensitivity, or how many actual outliers are correctly identified) against the false positive rate (FPR, also known as (1 - specificity), or how many normal points are incorrectly classified as outliers) at different classification thresholds. It provides a visual representation of how well a model distinguishes between classes (e.g., outliers vs. normal observations). Consequently, the Area Under the Curve (AUC) quantifies the overall performance of the model, with 1.0 indicating perfect separation between classes, and 0.5 representing no better than random guessing.

Advantages of Using ROC-AUC for Outlier

The advantages of using ROC-AUC for Outlier Detection are widely documented in the literature, but let's elaborate on the main ones related to this study:

- **Handling Imbalanced Datasets** - Outlier detection is inherently an imbalanced problem because outliers are rare by definition. Traditional metrics like accuracy are inappropriate in these cases as they would be dominated by the majority class (normal data points). For instance, with a dataset that is 95% normal data and 5% outliers, even a model that predicts all data as normal would achieve 95% accuracy but completely fail to identify outliers. ROC-AUC, on the other hand, is threshold-independent and evaluates the trade-off between TPR and FPR, making it robust to class imbalance, as it inherently adjusts for imbalance, as the true positive rate is normalised by the number of actual positive instances and the false positive rate is normalised by the number of

actual negatives. This is crucial when the positive class (outliers) is rare but significant, as AUC focuses on ranking rather than absolute class predictions;

- **Threshold Independence** - ROC-AUC evaluates the performance of the model across all possible classification thresholds, not just one specific threshold. In outlier detection, the optimal threshold for distinguishing outliers may vary depending on the specific application or data distribution, whereas ROC-AUC allows us to evaluate the overall performance without needing to specify a threshold;
- **Comparing Model Quality Across Different Applications** - Outlier detection methods often involve multiple models or algorithms. ROC-AUC offers a consistent and comparable metric for assessing the performance of various models across datasets and scenarios. For instance, Explainable AI (XAI) methods like SHAP or LIME, as discussed in Atzmueller, Fürnkranz, Kliegr, and Schmid (2024), can generate different explanations for the same predictive task, but their performance in terms of classification can be measured uniformly with ROC-AUC;
- **Focus on Detection of Rare Events** - Since outliers are rare, maximizing sensitivity (true positive rate) while controlling false positives (false alarm rate) is crucial. The ROC-AUC curve provides a visual understanding of this trade-off and helps to identify models that are better at detecting rare outliers while maintaining reasonable specificity;
- **Sensitivity to Both Classes** - ROC-AUC evaluates the trade-off between detecting the positive class (e.g., outliers) and minimising false alarms (i.e., misclassifying normal instances as outliers). This is especially important in imbalanced datasets, where correctly identifying the rare class is critical, and avoiding false positives is just as crucial;
- **Comparing Model Quality Across Different Applications** - Outlier detection methods often involve multiple models or algorithms. ROC-AUC offers a consistent and comparable metric for assessing the performance of various models across datasets and scenarios. For instance, Explainable AI (XAI) methods like SHAP or LIME, as discussed in Atzmueller et al. (2024), can generate different explanations for the same predictive task, but their performance in terms of classification can be measured uniformly with ROC-AUC. The work of Bouman, Buehler, and Heskes (2023) also concluded the ability of ROC-AUC to compare across multiple algorithms, which could be promptly interpreted from a probabilistic point of view; even considering other metrics such as the R-precision or average precision and their chance-adjusted variants introduced by Campos et al. (2016), but finding those to be less stable, and harder to interpret.

Why ROC-AUC Is Relevant to Explainability (XAI)?

The relevance of ROC-AUC to Explainability (XAI) goes from supporting stability and fidelity, as in Mozolewski et al. (2024), which highlights that ROC-AUC plays an essential role in evaluating the stability of predictions and, by extension, their explanations. A model with a high AUC consistently distinguishes outliers from normal points, which in turn leads to more stable and faithful explanations. Another relevant topic is Model-Explainer Consistency as for explainability to be trustworthy, the explainer must be consistent with the model's behaviour, and once again, ROC-AUC, by providing an accurate representation of the model's discriminatory ability, indirectly supports the fidelity of XAI methods, helping to assess the alignment

of XAI methods with ground truth and expert knowledge. Finally, in XAI, users must trust that the decisions made by the model are well-founded, and by applying ROC-AUC to validate model performance, you ensure that the model can reliably distinguish outliers, thereby making its explanations more credible and easier for stakeholders to understand and accept, fostering trust in model decisions

Comparing ROC-AUC with other metrics

As mentioned above, there are plenty of options to evaluate a model. Focusing on this study, some of the most common are compared as following:

- Accuracy can be misleading in highly imbalanced datasets. For outlier detection, where the number of normal points far exceeds outliers, a model could achieve high accuracy by predicting everything as normal, completely missing the rare but important outliers. For instance, if 99% of the dataset consists of normal instances and 1% are outliers, a model that classifies all instances as normal could still achieve 99% accuracy, even though it missed all outliers;
- Precision and Recall are widely used in classification problems, but both metrics are threshold-dependent. Precision focuses on the percentage of correctly identified outliers but ignores false negatives, while recall emphasises capturing all actual outliers but may overestimate false positives. ROC-AUC provides a more comprehensive measure by incorporating both precision and recall at all possible thresholds;
- F1 Score, which balances precision and recall, is still threshold-dependent. Although useful, it provides only a single snapshot of performance at a specific threshold, whereas ROC-AUC evaluates the model's performance across all thresholds, offering a more nuanced view.

How ROC-AUC Improves XAI

Incorporating ROC-AUC into XAI frameworks ensures that explanations are based on reliable model predictions. If a model with a high ROC-AUC is used, the resulting explanations (e.g., feature importance rankings from SHAP or LIME) are more likely to reflect meaningful, actionable insights. As Mozolewski et al. (2024) outlines, stability and fidelity are core elements of high-quality XAI systems, which, by ensuring the model itself is performant, ROC-AUC indirectly boosts the interpretability of XAI methods and the **robustness of explanations**. Also, high ROC-AUC scores suggest that the model is making reliable predictions, which increases the **fidelity of feature-importance explanations** provided by SHAP or LIME. High fidelity ensures that explanations align with actual model decisions, improving trust and transparency. Finally, ROC-AUC metrics allow for better communication of model performance to non-technical stakeholders. If the AUC is high, users can more confidently rely on the explanations produced by the model, leading to **enhanced human understanding** and decision-making processes.

Other relevant point of views derive from the work of Sokol and Vogt (2024) stresses the importance of decomposing explainable AI (XAI) systems into independent, modular building blocks and evaluating them separately. This is highly relevant for the argument on ROC-AUC because it introduces the idea that ROC-AUC can serve as an essential technical

evaluation metric at one of these modular stages—specifically the predictive performance of the model being explained. The framework emphasises evaluating each component—algorithmic, social, or mixed—according to its purpose in the system. This modularity highlights the importance of ROC-AUC in assessing the “core building blocks” of prediction accuracy before moving to user-centered evaluations. While traditional literature focuses either on algorithmic metrics (like ROC-AUC) or user-based evaluations (like satisfaction or task performance), this paper advocates for separating the evaluation of technical correctness from social effectiveness. The ROC-AUC metric, as a functionally grounded, algorithmic evaluation, is essential in assessing how well the model distinguishes between outliers and normal observations without conflating it with social validation aspects, such as how users interpret or understand the model’s explanations. Sokol and Vogt (2024) also highlights the relevance of context, arguing that evaluation metrics, including ROC-AUC, should be chosen based on the context of the XAI application. This aligns with the rationale for selecting ROC-AUC for evaluating outlier detection models: outliers typically occur in highly imbalanced contexts, and ROC-AUC effectively handles this by balancing the trade-off between true positive and false positive rates across all classification thresholds, particularly important where the positive class (outliers) is scarce but critically important. Finally, they highlight that trustworthiness is a key requirement of explainability. By ensuring that the underlying model performs well (quantified by a high ROC-AUC score), one can assert that explanations generated from a reliable model are more faithful and stable, which ties into their idea of sociotechnical evaluation, where trust in the model (as validated by ROC-AUC) underpins the trust in the explanations provided to the end-user.

3.6.2 Other Evaluation Metrics

Besides the ROC-AUC, to properly evaluate the clustering technique, objective metrics will be computed, as Mean SHAP values, SHAP values standard deviation (if is low, suggests that outliers in that group share more uniform feature importance), Max SHAP value (implies that those features strongly contribute to model predictions for outliers in that group), Top Features, Intra-cluster SHAP standard deviation (which measures how consistent the SHAP values are within each cluster) and number of outliers.

Perturbated data - Outlier detection evaluation performance

A key aspect of the evaluation involves testing the perturbed datasets based on the generated rules. If the synthetic outliers are correctly classified according to these rules, whether by perturbing the features identified as contributing to outlier classification or those associated with inliers, it indicates that deeper knowledge have been obtained with this approach, which is the ultimate goal of this study.

Chapter 4

Experimental Results

This chapter presents the results from the analysis proposed in the methodology, detailing the performance of the outlier detection and explanation algorithms, the clustering of explanations, and the evaluation metrics used to assess the effectiveness of the models. Through this analysis, the chapter aims to validate the approach and provide insights into the patterns and behaviours identified in the dataset. The ultimate goal is to test the hypothesis of generating synthetic outliers and being able to classify them according to the rules extracted from the clustering.

4.1 Breast Cancer Wisconsin dataset

The Breast Cancer Wisconsin (Original) Dataset is a well-known dataset used for classification tasks, specifically for identifying whether a breast tumour is benign or malignant. The dataset consists of 683 observations, each described by 9 features, which are computed from digitised images of fine needle aspirates (FNAs) of breast masses. Out of these, 239 observations are labelled as outliers (malignant class). The features include attributes such as the thickness of the clump, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, and mitoses, being each of these features represented by an integer value ranging from $\times 1$ to $\times 9$.

In the context of this dataset, outliers are data points that are atypical compared to the majority of observations. These instances may exhibit unusual characteristics compared to benign tumours, making their detection important for improving diagnostic accuracy.

4.1.1 Outlier Detection

Isolation Forest Algorithm

An Isolation Forest model was trained with 100 estimators and a contamination rate of 0.35, which reflects the proportion of outliers expected in the dataset. The model was evaluated using the ROC-AUC metric, as visualised in the ROC curves (Fig. 4.1).

The AUC score of 0.99 on the test dataset (Fig. 4.1a) reflects the model's excellent performance in distinguishing between benign and malignant tumours. The model achieves a near-perfect performance, with only a minimal false-positive rate (FPR), suggesting that it successfully identifies the malignant tumours (outliers) while keeping the false positives low.

Similarly, the ROC curve for the entire dataset also shows an AUC of 0.99 (Fig. 4.1b), reinforcing the model's strong generalisation ability across the whole dataset.

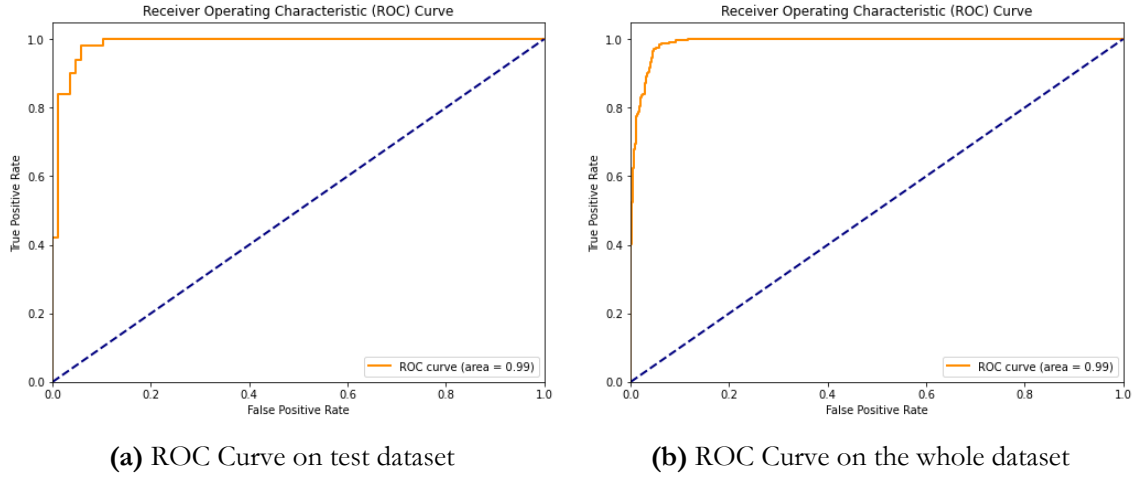


Figure 4.1: ROC Curve: Evaluation metrics

Other evaluation metrics can be observed in Table 4.1 allowing to infer of the good quality of the outlier detection task: from the 683 observations, 239 were classified as outliers, with a very high precision and recall.

Class	Precision	Recall	F1-Score	Support
0	0.96	0.96	0.96	444
1	0.93	0.93	0.93	239
Accuracy			0.95	683
Macro Avg	0.94	0.94	0.94	683
Weighted Avg	0.95	0.95	0.95	683

Table 4.1: Breast Cancer Dataset Classification Evaluation Report

Random Forest Algorithm

Class	Precision	Recall	F1-Score	Support
0	1	0.99	1	444
1	0.99	1	0.99	239
Accuracy			0.99	683
Macro Avg	0.99	0.99	0.99	683
Weighted Avg	0.99	0.99	0.99	683

Table 4.2: Dataset Classification Evaluation Report (Random Forest)

From the 683 observations, 239 were classified as outliers, similar to the results obtained with Isolation Forest algorithm in Fig. 4.1. We can observe the stability and fidelity of the metric ROC-AUC, while precision, recall and F1 score have a very significant improvement just by changing the inherent algorithm (Tables A.20 and 4.2).

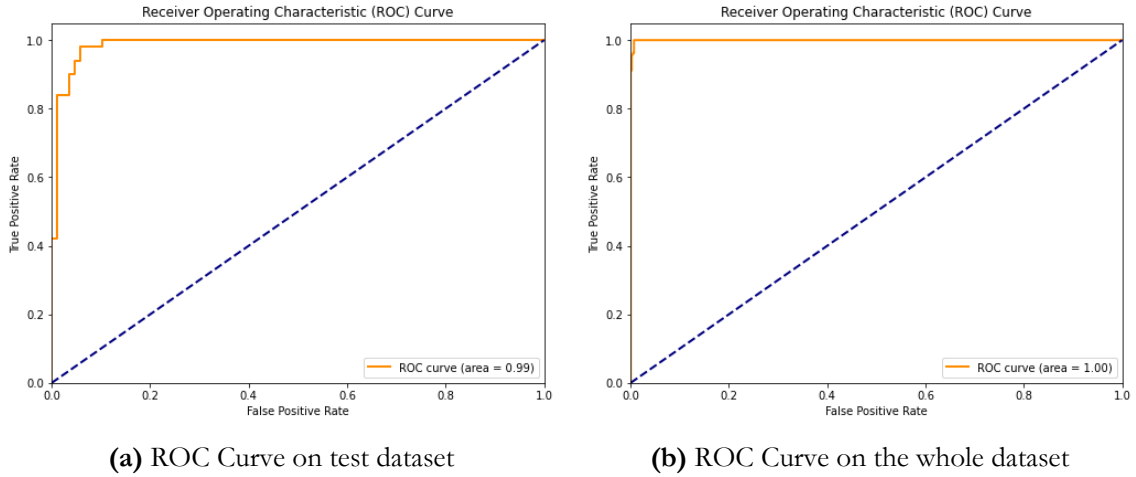


Figure 4.2: ROC Curve: Random Forest Evaluation metrics

LOF

As in the previous dataset, LOF algorithm was applied to this dataset but, once again, with very low results, namely ROC-AUC of 0.41 (as seen in Fig. A.6), precision of 0.55, recall of 0.15 and F1-score of 0.15. Once more, even considering an inversion of the labels, when comparing the predicted outliers with the true ground, the detector performance remains very poor.

4.1.2 Explainable AI

Given the results obtained in the previous task, explanations for the outliers were computed and analysed as following.

SHAP - Tree Explainer (Isolation Forest)

Focusing on explaining outliers within the dataset, SHAP tool with a Tree explainer was used to provide valuable insights into how individual features influence the prediction.

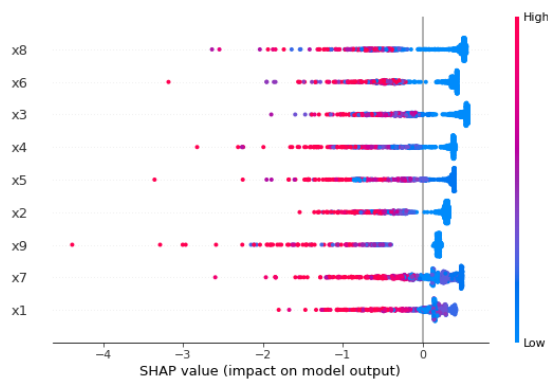


Figure 4.3: SHAP Summary Plot

Observing the SHAP summary plot (see Fig. 4.3) we derive that x_8 , x_6 , and x_5 have the most significant impact, with many instances showing high SHAP values for these features, indicating their strong influence in determining the outliers or inlier status. Regarding the interaction between feature values and SHAP values, for example, high feature values (in red) for x_6 and x_8 push the SHAP values positively, contributing more strongly to classifying an instance as an outlier, while lower values (blue) may contribute less or in the opposite direction. Features like x_8 and x_6 consistently contribute either positively or negatively to the model predictions across the dataset. On the other hand, x_7 ,

x_9 , and x_1 have more distributed impacts, meaning that their contribution to predicting outliers is more varied across different instances. These features influence the prediction both positively and negatively, depending on the specific instance.

The distribution of SHAP values for each feature across all outliers (see Boxplot in Fig. 4.4), shows that x_4 , x_5 , and x_6 exhibit higher variance, indicating that these features contribute significantly to predicting whether an instance is an outlier; features like x_1 , x_2 , and x_3 show narrower SHAP value distributions and closer to zero on the SHAP value scale, suggesting that these features have more neutral and consistent contributions across the outliers; and that the presence of several extreme SHAP values (dots outside the whiskers) indicates that specific instances may have significantly different impacts from the typical SHAP values for a given feature.

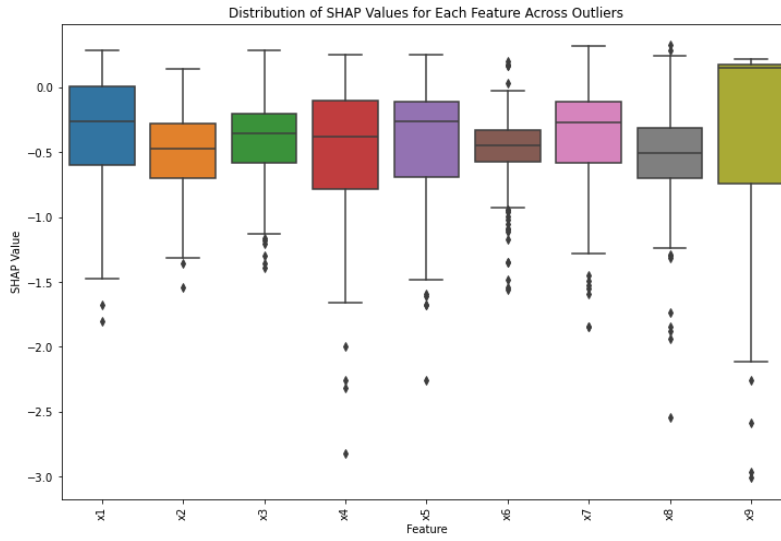


Figure 4.4: Distribution of SHAP Values for each feature Across Outliers

Regarding the Pairwise Comparison and given that each outlier might have an unique explanation, the difference in SHAP values between the first two outliers (Fig. 4.5) provides a comparative insight into how different features contribute to the classification of each outlier. In this specific case:

- x_6 and x_5 show large negative SHAP values for Outlier 1, while for Outlier 2, these values are much smaller (closer to zero or less negative). This indicates that x_6 and x_5 have a much stronger influence on Outlier 1, pushing the prediction towards classifying it as an outlier. For Outlier 2, these features are less influential. The large discrepancy in SHAP values between the two outliers might reveal that these features are pivotal in differentiating the contribution to outlier predictions;
- x_8 and x_3 have consistently negative SHAP values for both outliers, although with slightly different magnitudes. This means that these features consistently push the model toward classifying both instances as outliers, regardless of the outlier. However, the extent of their contribution differs, with Outlier 1 typically showing slightly more negative values than Outlier 2. These features are important across multiple outlier instances, indicating their reliability in contributing to the outlier prediction, albeit to varying extents;

- In contrast, x_7 and x_1 show positive SHAP values for both outliers, suggesting that these features contribute to increasing the prediction score (making the instance less likely to be classified as an outlier). However, the degree of this influence differs slightly between Outlier 1 and Outlier 2;
- The SHAP values for x_4 are opposite in direction for the two outliers. For Outlier 1, x_4 increases the likelihood of being classified as an outlier, while for Outlier 2, it reduces that likelihood. This indicates that x_4 behaves very differently between these two instances, contributing in a contradictory manner. While the SHAP values for x_4 differ in direction, the magnitude of these values is also noteworthy.

Feature	Outlier 1	Outlier 2
x_1	0.282259	-0.135825
x_2	-0.32673	-0.98611
x_3	-0.226681	-1.127075
x_4	-0.491591	0.10008
x_5	-1.029456	-0.010295
x_6	-1.115733	-0.365248
x_7	0.248878	0.119584
x_8	-0.310398	-1.158474
x_9	0.173461	0.18395

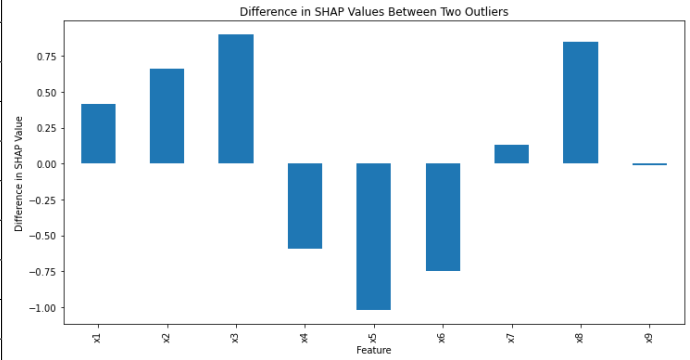
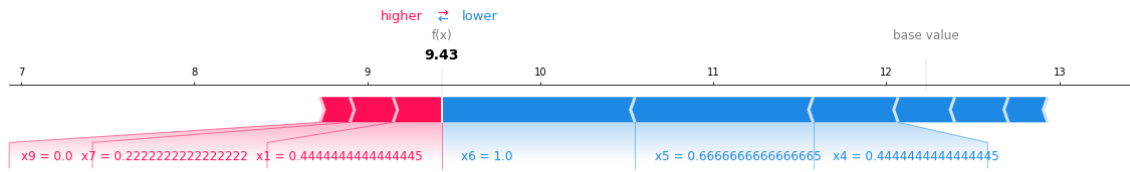


Figure 4.5: Table and Figure: Difference in SHAP Values Between 2 Outliers

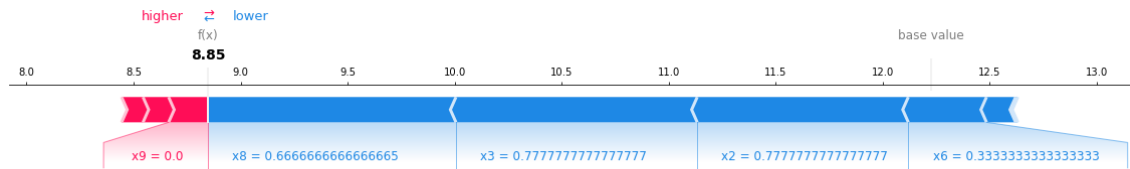
The individual explanation of the first five outliers can be seen in the SHAP Force Plots for Individual Outliers (Fig. 4.6a, 4.6b, 4.6c, 4.6d and 4.6e).

The SHAP force plots visually depict the contribution of each feature to a specific outlier classification, with red representing features pushing the prediction toward outlier status, and blue pushing it toward inlier status.

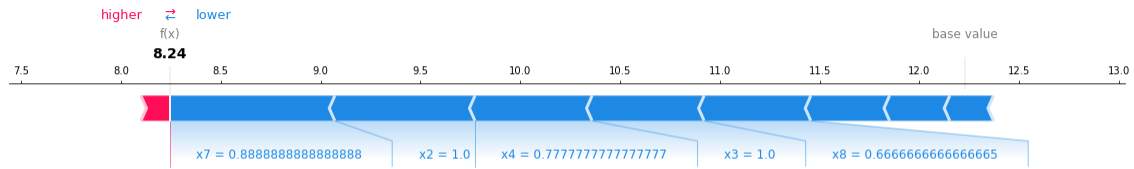
- For Outlier 1 (Force Plot 4.6a), the main driving factors for lowering the predicted value (from the base value of 12.227 to the final prediction of 9.431) are x_6 (-1.12), x_5 (-1.03), and x_4 (-0.49). However, x_1 (+0.28), x_7 (+0.25), and x_9 (+0.17) slightly counteract this and pull the prediction upward.
- For Outlier 3 (Force Plot 4.6b) x_8 lowers the prediction by -1.16, with x_3 and x_2 contributing similarly, reducing the value by -1.13 and -0.99, respectively, while x_1 , x_7 and x_9 positively contributes adding +0.28, +0.25 and +0.17, respectively, pushing the prediction down to 8.85.
- For Outlier 5 (Force Plot 4.6c), the overall prediction of 8.24 is mainly driven by x_7 (-0.83), x_2 (-0.7), x_4 (-0.59), and x_3 (-0.56), with a small upward adjustment from x_9 (+0.14).
- For Outlier 14 (Force Plot 4.6d), The absence of red bars implies that all the features have decreased the predicted value relative to the base value, so in this specific outlier, the features like x_4 , x_9 , x_5 , and x_8 are all pulling the predicted output toward a lower value (7.99), which could mean that the model has recognised an unique pattern that pushes the prediction below the normal range.



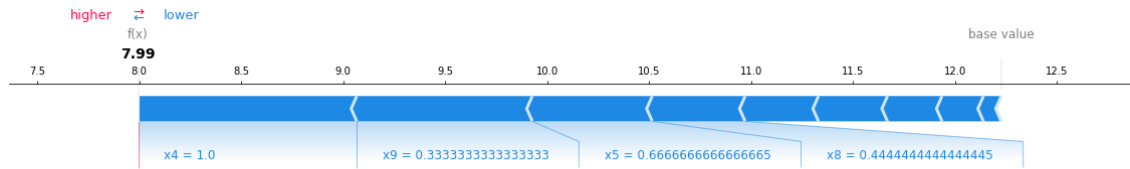
(a) Force Plot of Outlier 1



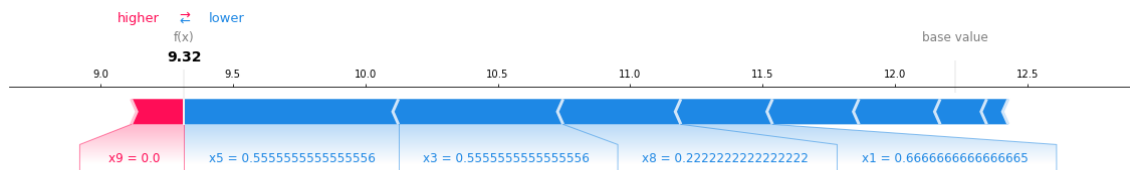
(b) Force Plot of Outlier 3



(c) Force Plot of Outlier 5



(d) Force Plot of Outlier 14



(e) Force Plot of Outlier 15

Figure 4.6: Force Plots of the first five Outliers

- For Outlier 15 (Force Plot 4.6e) x_9 provides a small red contribution (+0.2), pulling the prediction towards outlier classification. On the blue side, x_5 , x_3 , and x_8 strongly contribute, pushing the prediction lower to 9.317, towards an inlier classification.

- x_9 frequently contributes towards outlier classifications (red), particularly when its value is low or zero. This suggests that x_9 may represent a feature tied to abnormal or dangerous conditions when its value is particularly low.
- As already observed in Fig. 4.4, features as x_6 , x_5 , x_4 and x_2 often contribute towards inlier classification (blue), likely representing "healthy" measurements, or conditions in the dataset that correspond to normal fetal heart rate or uterine contractions.
- x_7 shows mixed behaviour: in some outliers (like Outlier 5), it contributes negatively (blue), while in others (like Outlier 1), it contributes positively (red). This suggests that x_7 's contribution is dependent on context, perhaps playing a critical role in identifying abnormal conditions when extreme values are present.

Another view of the same plots can be seen in Fig. 4.6.

If all the force plots explanations are compiled, rotated 90 degrees and stacked horizontally, we can observe explanations for all the dataset, as in Fig. 4.7.

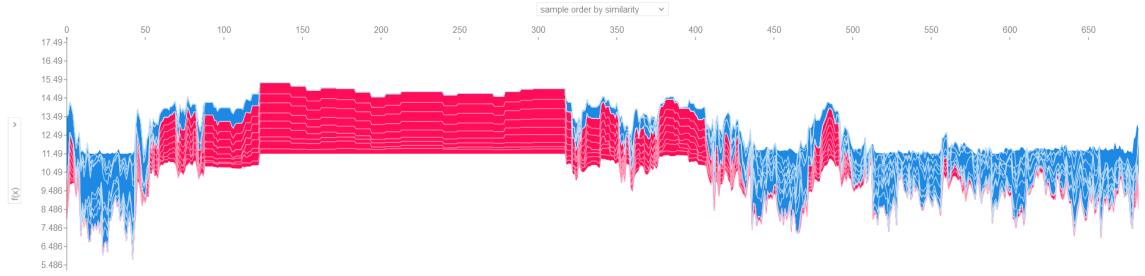


Figure 4.7: All SHAP Force Plots

SHAP - Tree Explainer (Random Forest)

Making use of the same technique but using the basis of Outlier Detection with Random Forest algorithm, we find very interesting results.

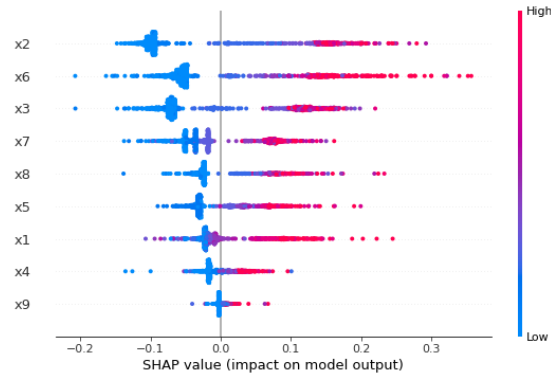


Figure 4.8: SHAP Summary Plot

Observing the SHAP summary plot (see Fig. 4.8), we find very similar results as in the ones based on Isolation Forest (see Fig. 4.3), although the direction of the SHAP values presents a simetric direction.

The distribution of SHAP values for each feature across all outliers (see Boxplot in Fig. 4.9), still shows that features x_8 , x_6 and x_5 are also the most influent.

The individual explanation of the first five outliers can be seen in the SHAP Force Plots for Individual Outliers (Fig. A.12a, A.12b, A.12c, A.12d and A.12e).

The structure of the explanations is very similar to the ones obtained based on Isolation Forest algorithm. For example, see Figures A.12d and 4.6d, where the attributes contributing for the classification are the same and even their magnitudes are similar.

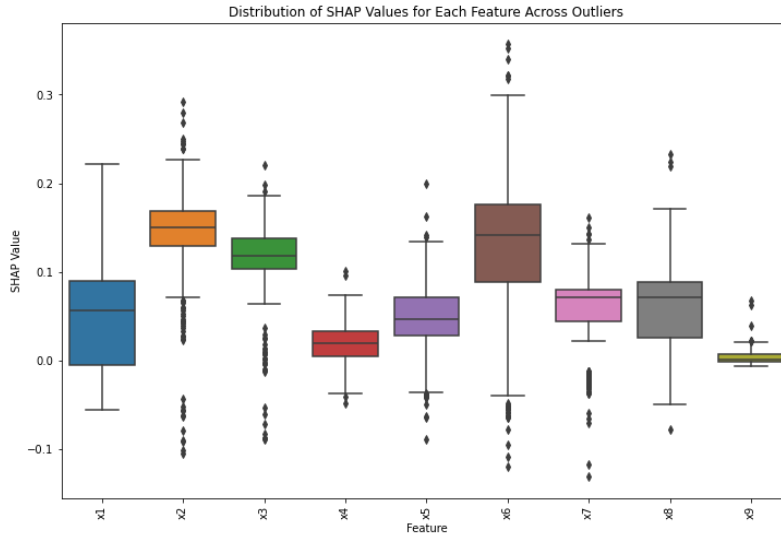


Figure 4.9: Distribution of SHAP Values for each feature Across Outliers

SHAP - Kernel Explainer

In case a Kernel Explainer is used to explain the outliers in this dataset, the following results can be achieved.

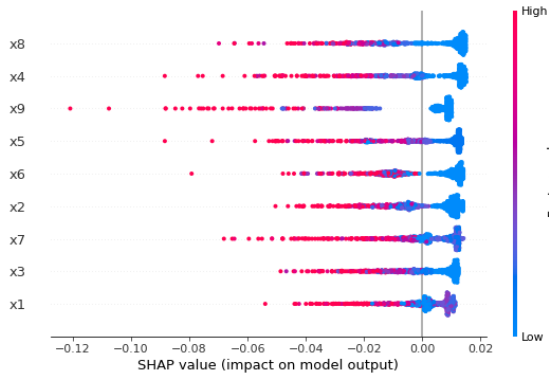


Figure 4.10: SHAP (Kernel) Summary Plot

Observing the SHAP summary plot (see Fig. 4.10), the first difference to the one obtained with the Tree Explainer (Fig. 4.10), which presented a wider range of SHAP values, both positive and negative (ranging from -4 to +4), is that the Kernel Explainer presents a more compressed range of SHAP values, with impacts ranging from -0.12 to +0.02. This suggests a smaller magnitude of feature contribution, which might be typical for this explainer, which is more conservative in estimating feature impacts. On the other hand, as Tree Explainer has a broader variance in feature importance, it might potentially capture more subtle influences of features on the output. Both explainers show that x_8 , x_6 , x_5 are important (although this kernel has smaller SHAP values) and, visually, very similar conclusions can be observed.

The distribution of SHAP values for each feature across all outliers (see Boxplot in Fig. 4.11), shows very similar characteristics to Tree Explainer only with subtle variations, as expected with this kernel.

The individual explanation of the first five outliers can be seen in the SHAP Force Plots for Individual Outliers (Fig. A.9a, A.9b, A.9c, A.9d and A.9e).

As expected, the conclusions are exactly the same as with the Tree Explainer, being the only difference on the Force Plots is the magnitude of the impacts for each attribute.

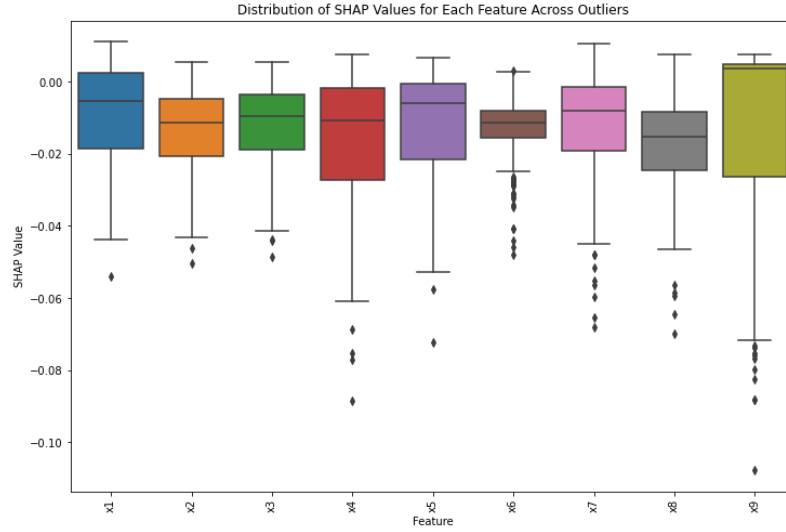


Figure 4.11: Distribution of SHAP Values (Kernel) for each feature Across Outliers

LIME

Instead of using SHAP, LIME method was applied to this dataset, making possible to explain individual predictions and making each observation more interpretable in its local neighbourhood (here, an outlier). These explanations can be observed in Fig. A.10.

For Outlier 1 (Fig. A.10a), the most important feature contributing positively to the prediction being an outlier is $x_9 \leq -0.35$; the higher values of x_4 , x_6 , and x_5 push the instance toward being classified as an inlier (values feature conditions $x_4 > 0.41$, $x_6 > 0.67$, $x_5 > 0.34$), but to a lesser extent; the features x_7 , x_1 , and x_3 have smaller positive contributions (red bars). The feature value can be seen in the right part of the figure as $x_9 = -0.35$, $x_4 = 0.76$, $x_6 = 1.77$, $x_5 = 1.70$ and $x_7 = -0.18$, all falling into the range ranges described.

For Outlier 3 (Fig. A.10b), the most important feature contributing positively to the prediction being an outlier is, once again, $x_9 \leq -0.35$, with smaller contributions from $x_4 \leq -0.64$ and $x_5 \leq 0.34$; on the negative contributions side, the high values of $x_2 > 0.60$, $x_8 > 0.37$, $x_3 > 0.60$ contribute negatively but their impact is not strong enough to classify it as inlier.

For Outlier 5 (Fig. A.10c), the prediction as outlier is mainly driven by $x_9 \leq -0.35$, while $x_4 > 0.41$, $x_7 > 0.64$, $x_2 > 0.60$ and $x_8 > 0.37$ have a negative impact but weaker than the positive contributions.

The same analysis can be done to Outlier 14 and 15 in Fig. A.10d and A.10e.

A very interesting fact is crossing these individual explanations with the explanations obtained through SHAP values in Figs. 4.6 and A.9, with Tree explainer, Kernel explainer, or LIME, the results are very identical.

4.1.3 Explanations Clustering

Isolation Forest and SHAP - Tree Explainer

The clustering process started with calculating the optimal number of clusters applying the Elbow Method. In Fig. 4.12 we can observe that the point where the inertia starts to

decrease at a lower rate is a large number of clusters, which can also be confirmed by the Silhouette Score of 0.18 and the Davies-Bouldin Index of 1.52.

Given this scenario, the next step aimed to reduce the noise in the SHAP values by focusing on more significant contributions from features, specifically, by replacing the SHAP values within the range of $[-1,1]$ by zero, and keeping the most influential SHAP values for further clustering and analysis. After applying this feature contribution pruning, clustering was done through a DBSCAN algorithm where the best parameters were 0.5 for *eps*, 8 *min_samples*, 5 *n_clusters* and a Silhouette Score of 0.45. Although this is not a sublime result, it can still indicate meaningful clusters, even if it means some potential overlapping clusters.

Reducing dimensionality helps visualise the clusters, making it easier to interpret how the outliers are distributed based on their SHAP values. Therefore, the PCA of the clusters is plotted in Fig. 4.13.

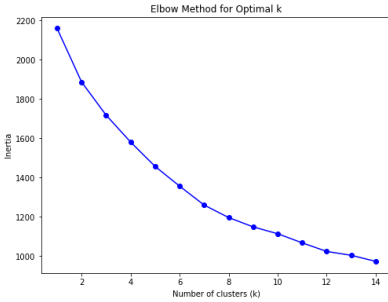


Figure 4.12: Elbow Method for Optimal k

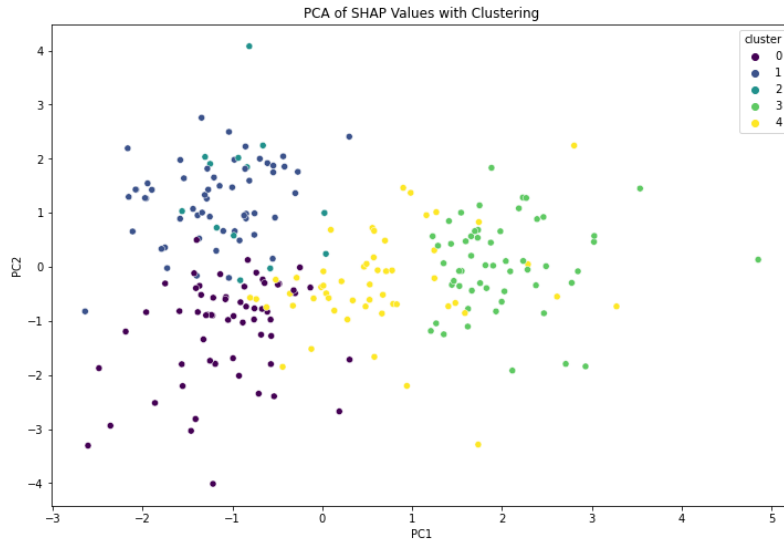


Figure 4.13: PCA of SHAP values with clustering

The characterisation of the clusters can be seen in Table 4.3, the distribution of SHAP values by feature and cluster in Fig. 4.14 and the mean absolute SHAP values by feature and cluster in Fig. 4.15.

Cluster 0 has a Mean SHAP value of 0.2020, indicating that on average, the features contribute moderately to the model predictions for the outliers in these cluster. The SHAP value is not extreme suggesting that the features driving this cluster have a noticeable impact but are not responsible for the most extreme outlier predictions. The relatively low standard deviation (0.0458) indicates that the SHAP values for outliers in Cluster 0 are quite consistent across different instances, suggesting that the feature contributions are relatively uniform within this group and implying that outliers in this cluster exhibit similar behaviours in terms of feature importance. While the maximum SHAP value of 0.2500 is not extremely high, it shows that certain features can still exert a significant influence on the model's predictions for some instances in this cluster. The features x_4 , x_5 , x_8 are the main drivers of outlier

Cluster	0	1	2	3	4
Mean SHAP	0.2020	0.7080	0.5932	0.6265	0.6627
SHAP Std Dev	0.0458	0.4742	0.6004	0.8762	0.4850
Max SHAP Value	0.2500	1.2555	1.2831	1.6377	1.1569
Top Features	x4, x5, x8	x6, x4, x5	x7, x4, x5	x9, x5, x4	x3, x2, x8
Intra-Cluster SHAP Std Dev	0.0964	0.3984	0.4188	0.5346	0.3952
Number of Outliers	159	11	19	39	12

Table 4.3: Cluster SHAP Values Metrics Summary

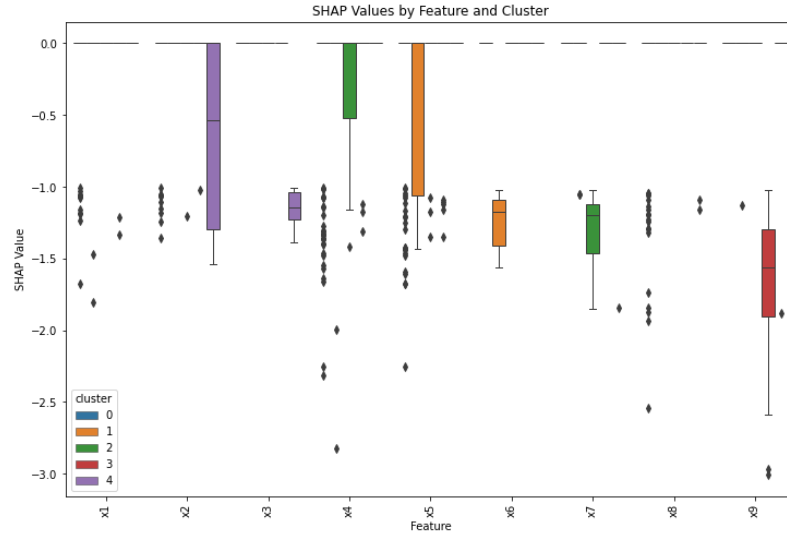


Figure 4.14: Distribution of SHAP Values by feature and cluster

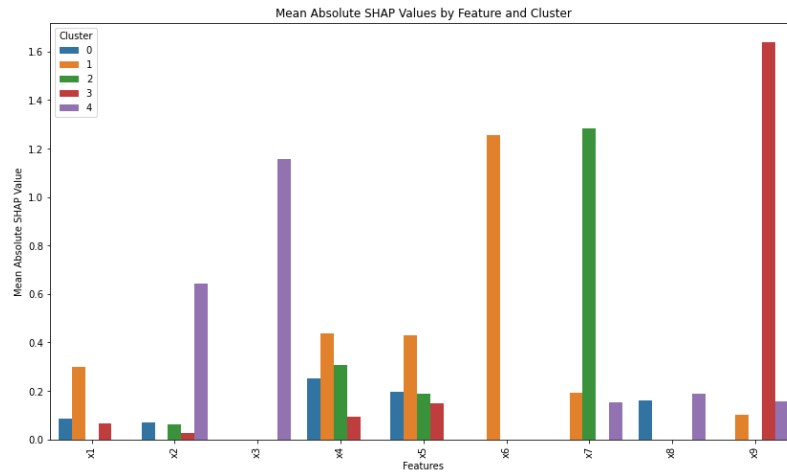


Figure 4.15: Mean absolute SHAP Values by feature and cluster

behaviour in this cluster once they consistently have the highest SHAP contributions for instances within this group. A relatively low intra-cluster standard deviation (0.0964) suggests that feature importance remains stable across outliers in this group, implying that the same features (x_4 , x_5 , x_8) tend to explain similar behaviours in different instances. Regarding the

number of outliers, this is the largest cluster, containing 159 observations, which suggests that a significant portion of the outliers exhibit similar patterns and behaviours and may indicate that a common combination of feature values leads to consistent outlier detection by the model.

Cluster 1 with the high average SHAP value (0.7080) indicates that features in this cluster contribute significantly to the model's predictions, having a major influence on the model's output and making these instances more distinct in their behaviour. The larger SHAP standard deviation (0.4742) suggests that the outliers in this cluster display more variability in how features impact the predictions, while the average contribution is high, different instances may have different combinations of features contributing to their classification as outliers. The maximum SHAP value is notably high (1.2555), indicating that for some instances, specific features (as x_6 , x_4 , x_5) have an exceptionally strong influence on the prediction. The dominant features are x_6 , x_4 , x_5 , which explains why outliers in this cluster have such high SHAP values. The higher intra-cluster SHAP standard deviation (0.3984) shows that there is some diversity in how the features influence different outliers within the group, meaning that not all instances in this cluster behave similarly. Cluster 1 contains a relatively small number of outliers (11), but these are characterised by features that have a significant impact on the model, suggesting that these are strong outliers driven by dominant features.

Cluster 2 Mean SHAP of 0.5932 indicates a high average contribution of features to the predictions for outliers in this cluster. The large SHAP Standard Deviation (0.6004) reflects substantial variability within the cluster, meaning that while the average contribution is high, individual outliers may have quite different feature importance profiles. The maximum SHAP value (1.2831) in this cluster is also high, similar to Cluster 1, indicating that some features have an exceptionally strong influence on the prediction for certain instances. The dominant features are x_7 , x_4 , x_5 . A relatively high intra-cluster deviation (0.4188) suggests that there is significant diversity within Cluster 2 in terms of how feature importance varies between instances. This cluster contains a moderate number of outliers (19), indicating that these feature combinations are less common but still represent significant patterns in the data.

Cluster 3 Mean SHAP value of 0.6265 suggests that features in this cluster have a strong influence on the model's predictions, though not as extreme as Cluster 1. The moderate SHAP values suggest that outliers in this group exhibit notable but not the most extreme feature impacts. The highest standard deviation (0.8762) across the clusters shows that there is a lot of variation in how different instances within this cluster are influenced by the features, suggesting that while the average SHAP value is high, different outliers exhibit widely different feature importance. The maximum SHAP value in Cluster 3 is the highest (1.6377), meaning that for some instances, features like x_9 and x_5 have an exceptionally strong influence, making these outliers stand out more. The features x_9 , x_5 , x_4 have the largest influence on the model's predictions, and the outliers in this group are largely driven by these variables. The high Intra-Cluster SHAP standard deviation within the cluster (0.5346) suggests that outliers vary considerably in how features contribute to their prediction, implying a high degree of heterogeneity in the behaviour of outliers in this group. This cluster contains a moderate number of outliers (39), which indicates that while these feature combinations are less common than in Cluster 0, they still represent a notable portion of the data.

The high average SHAP value (0.6627) in Cluster 4 indicates that features in this group have a significant influence on the model's predictions, but not as extreme as in Cluster 1. A moderately high standard deviation (0.4850) suggests some variability in how features in-

fluence different outliers in this group, having more homogeneity than in Cluster 3, but still some diversity in feature importance. This relatively high maximum SHAP value (1.1569) indicates that certain features (like x_3 , x_2 , and x_8) can exert a strong influence on the prediction for some instances in this cluster. The features x_3 , x_2 , x_8 are the most important for determining outlier behaviour in this cluster, consistently responsible for driving the outlier predictions. A moderate intra-cluster standard deviation (0.3952) suggests that there is some consistency in how features influence predictions, but with a reasonable amount of variation between instances. This is the smallest cluster, containing only 12 outliers, which are less frequent in the data but are driven by specific, impactful feature combinations.

Some general rules we can derive from these analysis are:

- If an observation has moderate values in x_4 , x_5 , or x_8 , it is likely to fall into Cluster 0, where the model is detecting outliers with consistent feature contributions;
- Instances with particularly high values in x_6 or x_5 are likely to be grouped into Cluster 1, where these features significantly influence the outlier behaviour;
- High values in x_7 , x_4 , or x_5 indicate that an instance is likely to be categorised into Cluster 2, once these outliers display strong but somewhat variable feature importance, leading to moderate SHAP contributions.
- Observations characterised by high values in x_9 or x_5 are often found in Cluster 3, where SHAP values are highly variable, and feature contributions can be quite extreme;
- If an instance shows high values in x_3 , x_2 , or x_8 , it may be grouped into Cluster 4.

Isolation Forest and SHAP - Kernel Explainer

The clustering process started with calculating the optimal number of clusters applying the Elbow Method. In Fig. A.13 we can observe that the point where the inertia starts to decrease at a lower rate is a large number of clusters, which can also be confirmed by the Silhouette Score of 0.14 and the Davies-Bouldin Index of 1.79.

Reducing dimensionality helps visualise the clusters, making it easier to interpret how the outliers are distributed based on their SHAP values. Therefore, the PCA of the clusters is plotted in Fig. 4.16

The characterisation of the clusters can be seen in Table 4.4, the distribution of SHAP values by feature and cluster in Fig. 4.17 and the mean absolute SHAP values by feature and cluster in Fig. 4.18.

Cluster	0	1	2
Mean SHAP	0.0317	0.0163	0.0212
SHAP Std Dev	0.0274	0.0008	0.0023
Max SHAP Value	0.0633	0.0173	0.0236
Top Features	x_9 , x_8 , x_5	x_8 , x_4 , x_6	x_2 , x_3 , x_4
Intra-Cluster SHAP Std Dev	0.0174	0.0040	0.0058
Number of Outliers	22	129	89

Table 4.4: Cluster SHAP Values Metrics Summary (Kernel)

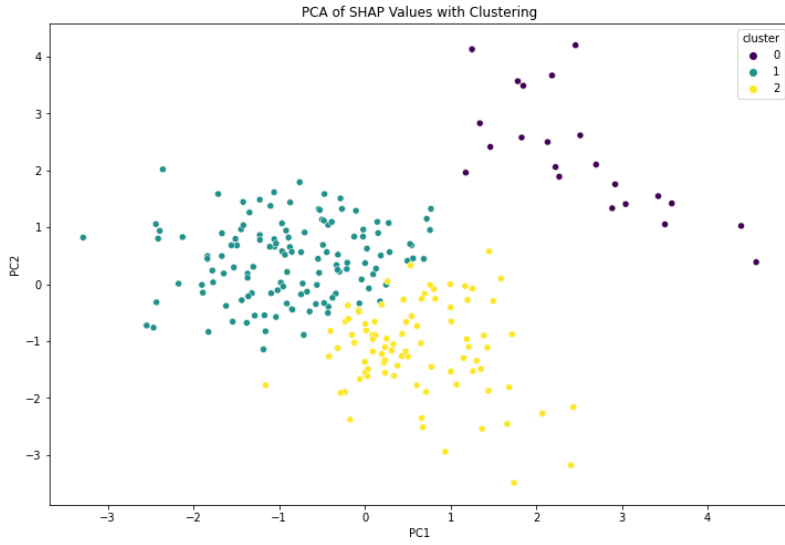


Figure 4.16: PCA of SHAP values with clustering

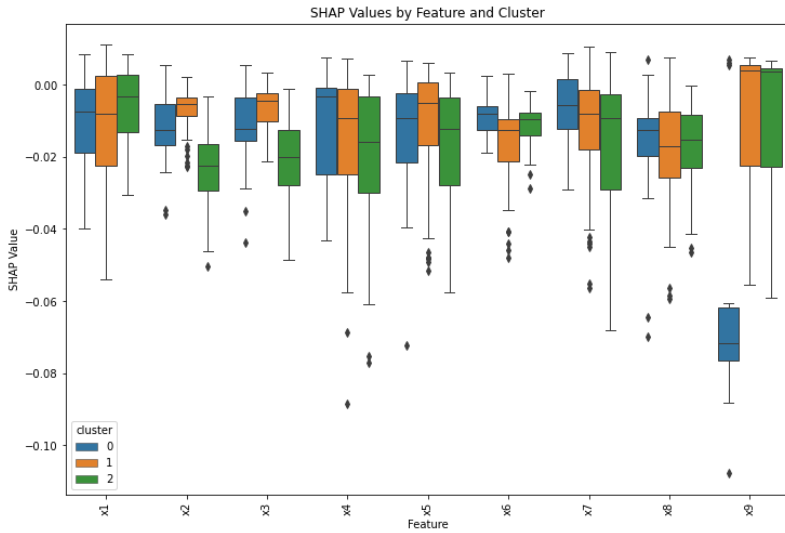


Figure 4.17: Distribution of SHAP Values by feature and cluster (Kernel)

Cluster 0 has the highest mean SHAP value (0.0317), indicating that the outliers in this cluster are significantly impacted by the features, particularly x_9 , x_8 , and x_5 . The higher SHAP values suggest a stronger influence of these features in determining these data points as outliers. Also, the higher SHAP standard deviation (0.0274) and max SHAP value (0.0633) indicate a broader range of feature impacts within this cluster.

Cluster 1 has the highest number of outliers (129), the lowest mean SHAP value (0.0163), and a very low SHAP standard deviation (0.0008), which indicates that the impact of the top features, namely x_8 , x_4 , and x_6 , is highly consistent across all outliers in this cluster. The low max SHAP value suggests that no single outlier in this cluster has an extreme feature contribution, reinforcing the homogeneity of the feature impacts. If an observation shows high values in x_8 , x_4 , or x_6 , it is more likely to be classified in this cluster due to the strong influence of these features.

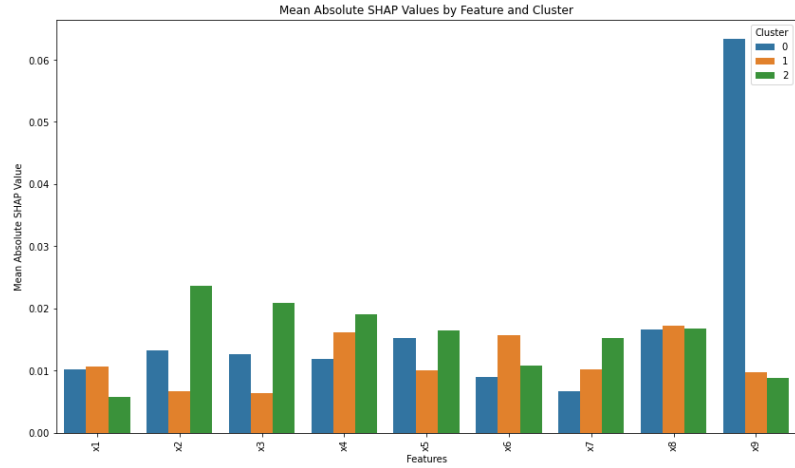


Figure 4.18: Mean absolute SHAP Values by feature and cluster Kernel)

Cluster 2 falls between Cluster 0 and Cluster 1 in terms of SHAP mean and standard deviation. The top features influencing Cluster 2 include x_2 , x_3 , and x_4 with a moderate SHAP impact. It has 89 outliers, which is more than Cluster 0 but less than Cluster 1.

LOF Algorithm and LIME

After detecting outliers with LOF, LIME was applied and the results of the clusters were considerably poor, which goes in line with the requirement of having a good performance on the detection algorithm in order to extract meaningful clusters from the explanations.

Isolation Forest and LIME

In this case, the clustering process started by calculating the optimal number of clusters. By the Elbow Method (see Fig. A.14), the number of clusters is relatively high, while the best Silhouette Score of 0.15, which is relatively low, suggests 2 clusters (with a Davies-Bouldin Index of 1.85).

In order to better visualise the clusters computed, dimensionality was reduced through a Principal Component Analysis of the clusters, plotted in Fig. A.14.

The distribution of LIME values by feature condition and cluster can be seen in Fig. A.11.

Cluster 0 has 127 observations and Cluster 1 has 113 outliers. The feature condition $x_9 \leq -0.35$ has a strong positive LIME value for Cluster 0, meaning that in these situations, it significantly contributes to the outlier classification; for Cluster 1, this feature has a smaller positive contribution. The feature conditions $x_4 > 0.41$, $x_6 > 0.67$ show strong negative contributions for both clusters, suggesting that when x_4 is greater than 0.41 and x_6 is greater than 0.67, they decrease the likelihood of the instance being an outlier; $x_7 \leq -0.55$ has a strong negative impact for both clusters, therefore, instances that satisfy this condition tend to have their likelihood of being an outlier decreased significantly; the feature $x_1 \leq -0.55$ also has strong negative contributions for both clusters. For Cluster 0, when $x_5 > 0.34$, it makes a small but notable positive contribution to the instance being an outlier; however, in Cluster 1, the contribution is slightly negative. The features $x_6 > 0.64$, $x_2 \leq -0.70$ have more subtle contributions, but still play a role in the classification. Clusters 0 and 1 exhibit similar behaviours, but the intensity of certain feature impacts is stronger in Cluster 0 than in

Cluster 1, once that tends to have more extreme positive or negative feature contributions. Some rules we can infer are if $x_9 \leq -0.35$ and $x_5 > 0.34$, the instance is likely to belong to Cluster 0 (higher likelihood of being an outlier); if $x_4 > 0.41$ and $x_6 > 0.67$, the instance is less likely to be classified as an outlier, as these conditions negatively impact the model's decision for Cluster 0; if $x_7 \leq -0.55$ and $x_1 \leq -0.55$, the instance is likely to belong to Cluster 1, as these features have a strong negative contribution to the outlier classification; for both clusters, if $x_4 > 0.41$, $x_6 > 0.67$, or $x_7 \leq -0.55$, the instance is less likely to be an outlier, once these features tend to push the classification away from an outlier; instances where $x_9 \leq -0.35$ and $x_4 > 0.41$ could be likely outliers in Cluster 0, but when x_4 or x_6 is higher, it decreases the outlier likelihood. We can also observe that x_9 , x_4 , and x_6 are important differentiators between the clusters.

4.2 The Seismic Sataset

The Seismic dataset from UCI machine learning repository is a binary classification dataset that contains 19 attributes, one of them, being a binary classification of seismic bumps, which are rare events in underground mining environments that can indicate potential hazards. The features of the dataset contains multiple attributes that describe seismic activities, such as the number of seismic bumps in different time intervals, the energy released by the seismic events, and other related measurements. These features are used to capture the patterns of seismic activity that might precede or indicate a seismic bump. As basic statistics, there is 2,584 observations, whereas 170 observations are labelled as outliers (seismic bumps).

These statistics indicate that the dataset is highly imbalanced, with a small proportion of the observations being classified as outliers, making it a challenging dataset for anomaly detection tasks.

4.2.1 Outlier Detection

Isolation Forest Algorithm

Receiver Operating Characteristic (ROC) curve on the test dataset (see Figure A.17a) shows an area under the curve (AUC) of 0.68. Although an AUC of 0.68 is not outstanding, it does show that the model performs better than random guessing, which would result in an AUC of 0.5. The model was applied to the entire dataset to assess its performance more broadly. In this case, the ROC curve for this evaluation is shown in Fig. A.17b, with an AUC of 0.71, suggesting that the model slightly improves its ability to distinguish outliers when applied to the full dataset, possibly due to greater exposure to the underlying patterns in the data.

The classification report for both the test and full dataset indicated that the precision and recall were consistent with the relatively low AUC scores.

The Isolation Forest model achieved acceptable results, particularly considering the complexity of the problem. Being seismic bumps rare events, an AUC of 0.71, while not perfect, indicates that the model is reasonably capable of identifying outliers within the dataset. However, there remains a significant false positive rate, as demonstrated by the relatively shallow curve in the ROC plots.

LOF Algorithm

LOF algorithm was applied to this dataset but with very low results, namely ROC-AUC of 0.47 (as seen in Fig. A.4), precision of 0.35, recall of 0.06 and F1-score of 0.10. Even considering an inversion of the labels, when comparing the predicted outliers with the true ground, the detector performance remains very poor.

Random Forest Algorithm

Random Forest algorithm performed the best including SMOTE techniques while pre-processing the data. The test dataset achieved a ROC-AUC of 0.68 (see Fig. A.18a) and an outstanding performance of 0.98 in the full dataset when compared with the ground truth (see Fig. A.18b). Given the characteristics of this dataset, feature selection was performed, specifically, eliminating the features *nbumps5*, *nbumps6*, *nbumps7* and *nbumps89*, which achieved similar performances.

4.2.2 Explainable AI

SHAP - Tree Explainer (Isolation Forest)

Applying a SHAP technique with the Tree Explainer on the model based on Isolation Forest, we can obtain the SHAP summary plot (see Fig. A.19), which provides a comprehensive view of how each feature contributes to the prediction of outliers. We can identify features with significant impact like *seismoacoustic*, *nbumps4*, *nbumps3*, and *energy*, having the highest SHAP values and meaning they are the primary contributors to the model's classification of outliers; and feature value impact, as the colours on the plot indicate whether high or low feature values result in higher SHAP values. For example, higher values of *seismoacoustic* seem to be associated with a strong contribution toward outlier prediction, and similarly, lower values of *nbumps3* and *nbumps4* push the prediction towards identifying the data as an outlier.

The distribution of SHAP values for each feature across all outliers shows how each feature's SHAP values are distributed across the detected outliers (see Boxplot in Figure A.20), where each box represents the spread of SHAP values for a given feature, helping to identify the magnitude and variability of that feature's contribution to the outlier classification. The SHAP values for *seismoacoustic* are relatively centred around zero, with a small range of variation. The feature *nbumps4* exhibits a wide spread of SHAP values, with many observations having large negative contributions. The SHAP values range from around -2.5 to near zero, indicating that for many outliers, *nbumps4* significantly pulls the prediction away from the inlier classification. Its large variability suggests that *nbumps4* is critical for the outlier classification, but the degree of its impact varies greatly across different outliers. Both *energy* and *maxenergy* features show broad distributions of SHAP values, indicating their strong influence on outlier detection; for many observations, these features have relatively high SHAP values in both directions, suggesting that these features can either increase or decrease the likelihood of being classified as an outlier, depending on their specific values in each observation. The features *genergy* and *gpuls* display a more consistent negative contribution across outliers, suggesting that lower values of *genergy* and *gpuls* tend to be associated with outlier detection; for some outliers, these features strongly influence the classification decision, contributing negatively to the SHAP value. The *shift* and *seismic* features have relatively small spreads and are clustered around zero SHAP values, indicating that while these features are considered by

the model, they do not play a significant role in the classification of most outliers, remaining neutral for the majority of instances. The bump-related features (*nbumps*, *nbumps2*, *nbumps3*) tend to have mostly negative SHAP values, particularly *nbumps3* and *nbumps2*, which pull the prediction towards outlier classification for many instances; however, there are cases where these features contribute positively, showing a varied impact across outliers.

To demonstrate that each outlier might have a unique explanation, a comparison was made among the first two. In the Figure A.15 we can observe that the most significant discrepancy between the two outliers is seen in *seismoacoustic*, where outlier 2 has a much more significant positive contribution, implying that this feature greatly influenced the classification of Outlier 2 as an outlier, whereas it had a much smaller role in Outlier 1; *nbumps* and related features (*nbumps2*, *nbumps3*, *nbumps4*) show moderate to large negative SHAP values for both outliers but differ in magnitude, suggesting that the bumps count plays a substantial role in defining the outliers, but the extent varies between the two outliers, noting that the discrepancy in *nbumps4* particularly stands out. The pairwise comparison table (Table in Fig. A.15) reveals that while both outliers exhibit some similarities in terms of feature impact (e.g., *genergy*, *gpuls*, *seismic*), the magnitude of these contributions often varies significantly, leading to differing classifications. *Seismoacoustic* and *genergy* are two of the most distinguishing features, with *seismoacoustic* acting as a key positive contributor for Outlier 2 and *genergy* acting as a strong negative contributor for both outliers. The negative SHAP values for *nbumps* and *nbumps2* are much more pronounced for Outlier 1, indicating that these features played a larger role in its classification.

Individual SHAP force plots give us a detailed explanation of how each feature influences the classification of the first five outliers (Figures A.16a, A.16b, A.16c, A.16d and A.16e). The more frequent features in these explanations are *genergy* and *gpuls* and the ones related with *nbumps*, more specifically *nbumps*, *nbumps2* and *nbumps4*.

SHAP - Tree Explainer (Random Forest with Feature selection)

When a SHAP Tree Explainer is applied, after detecting outliers with Random Forest with feature selection, the following results can be achieved.

Observing the SHAP summary plot (see Fig. A.21), the feature *gpuls* has the most significant impact on model output, with both positive and negative SHAP values, and features like *genergy*, *gdpuls*, and *nbumps* also have a considerable influence, but to a lesser extent. For example, *gpuls* shows that high values tend to increase the model's output (towards outliers), while low values (blue) tend to decrease it. *Shift* is a feature that shows a more consistent behaviour, having a strong influence when it has higher values (red) and meaning it significantly pushes the prediction towards identifying outliers. Other features like *nbumps* and *energy* also play important roles but appear to have less variability in their SHAP values, which indicates that their influence is more subtle but still critical.

Regarding the distribution of the SHAP values for each feature across outliers (see Box-plot in Fig. A.22), it can be observed that features such as *gpuls*, *genergy*, and *nbumps* have a wide spread of SHAP values, meaning they have a large range of influence on the model's predictions, with outliers causing higher variance in SHAP values; while features like *shift* and *energy* show narrower distributions, indicating more consistent contributions across different instances. Several features show distinct outliers (e.g., *gpuls* and *genergy*), which suggests that there are specific instances where these features significantly deviate from the norm, contributing more strongly to the outlier predictions.

The distribution of SHAP values for each feature across all outliers, shows very similar characteristics to Tree Explainer only with subtle variations, as expected with this kernel.

The individual explanation of the first five outliers can be seen in the SHAP Force Plots for Individual Outliers (Fig. A.23, individual explanations can be found. For example, for Outlier 162 (Fig. A.23d), the main features pushing to the classification of outlier are *gpuls* and *nbumps*, while *seismoacoustic* and *gdenenergy* have a negative impact, decreasing the prediction. From these 5 outliers, it can also be observed that the features *gpuls*, *nbumps*, *nbumps2*, *energy* are often strongly contributing to pushing the prediction towards outlier classification.

4.2.3 Explanations Clustering

Isolation Forest and SHAP - Tree Explainer

The SHAP values were scaled and clustered using KMeans with three clusters, as suggested by the Elbow method, to determine which features most influenced the outliers within each group. Reducing dimensionality helps visualise the clusters, making it easier to interpret how the outliers are distributed based on their SHAP values. In the PCA scatter plot (see Fig. 4.19, each point represents an outlier, coloured by cluster membership. The clusters are well separated, indicating that the SHAP values are distinctly distributed across these three groups and each cluster has a unique pattern of SHAP values, representing different explanations for the model's predictions.

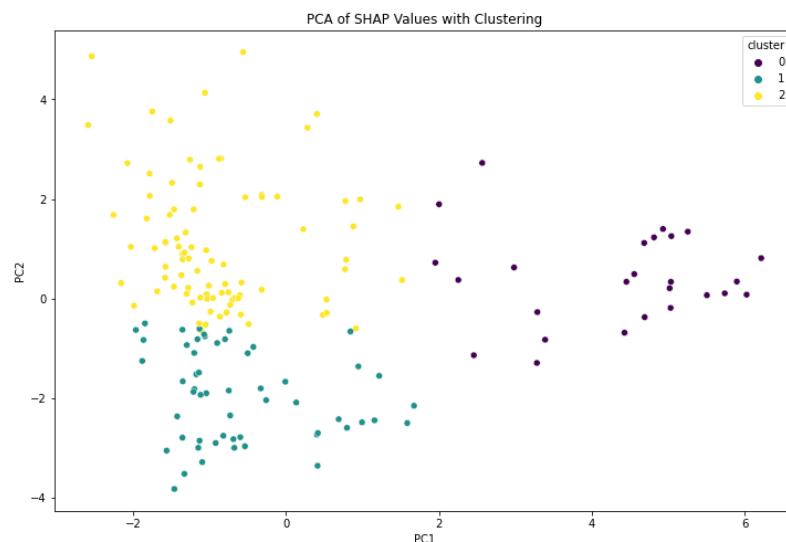


Figure 4.19: PCA of SHAP values with clustering IF SHAP Tree

The characterisation of the clusters can be seen in Table 4.5.

The Cluster 0 is characterised by features related to the number of bumps (*nbumps*, *nbumps3*, *nbumps2*), which play a crucial role in explaining the model's predictions for outliers in this group. The relatively moderate SHAP values indicate that the influence of these features is balanced and consistent across the outliers, with some variability in intra-cluster SHAP deviation.

Cluster 1 has the highest number of outliers, with *nbumps4*, *energy*, and *maxenergy* being the dominant features. These features heavily influence the SHAP values, suggesting that outliers

Cluster	0	1	2
Mean SHAP	0.9114	1.2960	1.2935
SHAP Std Dev	0.2681	0.1720	0.2701
Max SHAP Value	1.2131	1.3978	1.4787
Top Features	nbumps, nbumps3, nbumps2	nbumps4, energy, maxenergy	ghazard, gdpuls, gdenergy
Intra-Cluster SHAP Std Dev	0.3580	0.4681	0.4822
Number of Outliers	55	88	26

Table 4.5: Cluster Metrics Summary - Isolation Forest and SHAP Tree Explainer

in this cluster are significantly impacted by energy-related features and the fourth bump in the data. The standard deviation of SHAP values is lower, suggesting that outliers in this group share more uniform feature importance.

Cluster 2 is the smallest group, and its outliers are mainly influenced by features related to hazards and energy (e.g., *ghazard*, *gdpuls*, and *gdenergy*). This cluster exhibits the highest maximum SHAP value, implying that these features strongly contribute to model predictions for outliers in this group. The intra-cluster variability is higher, indicating more diversity in how these features impact different outliers.

The boxplot of SHAP values by feature and cluster (Fig. A.24) reveals how features affect the predictions differently across the three clusters. For example, features such as *nbumps* and *energy* show distinct differences in SHAP values across the clusters, with each feature having a larger influence in specific clusters; cluster 1 shows consistently higher SHAP values for *nbumps4* and *energy*, while Cluster 0 has a higher focus on *nbumps3*; and cluster 2 shows significant variability, especially for features like *ghazard* and *gdpuls*.

In the bar plot of the mean absolute SHAP values by feature and cluster (Fig. A.25, *nbumps4*, *energy*, and *ghazard* are the top contributors in their respective clusters; Cluster 2 stands out in terms of high SHAP values for *ghazard* and related features, suggesting a distinct explanation pattern compared to the other clusters; and Cluster 0, though smaller in SHAP magnitude, shows a consistent importance of *nbumps* features, differentiating it from the others.

The intra-cluster SHAP standard deviation measures how consistent the SHAP values are within each cluster. Cluster 1 has the lowest variability, indicating more uniform SHAP values across its outliers, suggesting that explanations for predictions in this cluster are more homogenous; meanwhile Cluster 2 shows the highest intra-cluster variability, meaning there is greater diversity in how the SHAP values explain the predictions for outliers within this cluster.

Each cluster represents a group of outliers with similar SHAP explanations, driven by specific features. For instance, outliers in Cluster 2 are more influenced by hazard-related features, while Cluster 1 is dominated by energy and bump-related features. Cluster 1, with the highest number of outliers (88), shows the most homogeneity in SHAP values, indicating that the explanations for the model's predictions are more consistent within this group. Features like *nbumps* and *energy* are critical for most clusters, but their exact impact varies. For example, *nbumps* features dominate in Cluster 0, while *energy* and *ghazard* are more prominent in Clusters 1 and 2, respectively. The clusters reflect diversity in the model's explanations,

with some clusters (like Cluster 2) showing a wide range of SHAP values for key features, which could be due to the varied nature of the outliers within this group.

Isolation Forest and SHAP - Tree Explainer with brute force In this case, a test was made by forcing the SHAP values matrix to replace the values between $[-0.5, 0.5]$ by 0. The optimal number of clusters, by the Silhouette Score (0.357), was 2. Similar conclusions for the clusters were derived for these conditions.

LOF Algorithm and LIME

Once again, after detecting outliers with LOF, LIME was applied and the results of the clusters were considerably poor, which goes in line with the requirement of having a good performance on the detection algorithm in order to extract meaningful clusters from the explanations.

Random Forest and SHAP - Tree Explainer

The PCA plot (see Fig. 4.20) reflects 2 well separated clusters, indicating that the SHAP values are distinctly distributed across the groups and each cluster has a unique pattern of SHAP values, representing different explanations for the model's predictions.

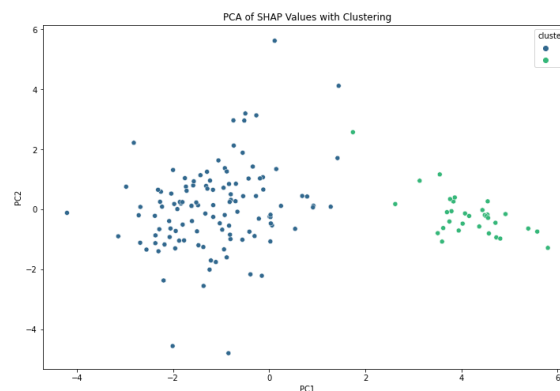


Figure 4.20: PCA of SHAP values with clustering RF SHAP Tree

The characterisation of the clusters can be seen in Table 4.6, achieving a Silhouette Score of 0.357 and a Davies-Bouldin Index of 1.07. In Cluster 0, the features *nbumps4*, *energy*, and *maxenergy* stand out as the most significant in contributing to the model's predictions for outliers, suggesting that the anomalies in this cluster are largely driven by high energy levels and specific *nbumps4* measurements, which likely correspond to anomalous patterns in the data. The maximum SHAP value for this cluster (1.0955) indicates moderate feature contribution to outlier predictions. This cluster contains the largest number of outliers (143 outliers), and lower SHAP standard deviation within the cluster (0.1727) suggests that these outliers have relatively homogeneous explanations, meaning they share similar patterns in terms of how the key features contribute to their classification as outliers. The intra-cluster standard deviation (0.3587) also supports this idea, as it shows that the SHAP values within this cluster don't deviate too much from each other, pointing towards consistent behaviour in the data for these outliers.

The Cluster 1 is dominated by *ghazard*, *gdpuls*, and *gdenergy* as the top contributing features. The presence of *ghazard* suggests that this cluster's outliers are tied to hazard-related data or signals that involve specific *gdpuls* and *gdenergy* patterns, possibly indicating a different type of rare event than in Cluster 0. This cluster has fewer outliers (26 outliers), but with a higher mean SHAP value (1.2634) and a higher max SHAP value (1.4787), suggesting that the model detects stronger feature contributions for these outliers. The larger SHAP standard deviation (0.2647) and intra-cluster SHAP standard deviation (0.4724) indicates more variability in how the features contribute to each outlier within this cluster. This variability could mean that outliers in Cluster 1 are less homogeneous compared to those in Cluster 0, with different combinations of features playing significant roles in each individual case.

Cluster	0	1
Mean SHAP	0.9504	1.2634
SHAP Std Dev	0.1727	0.2647
Max SHAP Value	1.0955	1.4787
Top Features	nbumps4, energy, maxenergy	ghazard, gdpuls, gdenergy
Intra-Cluster SHAP Std Dev	0.3587	0.4724
Number of Outliers	143	26

Table 4.6: Cluster Metrics Summary - Random Forest and SHAP Tree Explainer

Cluster 0 seems to group the energy-driven outliers. If *nbumps4*, *energy*, and *maxenergy* have high SHAP values, the instance is likely an outlier due to energy-related anomalies. Cluster 1 groups hazard-related outliers. If *ghazard*, *gdpuls*, and *gdenergy* are highly significant, the instance is likely an outlier related to potential hazard or dangerous events.

Feature selection Applying the feature selection technique described previously, obtained performances, clustering metrics and dominant features were similarly, originated similar possible rules.

4.3 Cardiotocography Dataset

The Cardiotocography (CTG) Dataset originates from a study that records the fetal heart rate (FHR) and uterine contraction (UC) measurements. These signals are crucial in assessing the well-being of a fetus during pregnancy and the dataset is often used in medical studies for monitoring fetal health and identifying any signs of fetal distress, which could indicate potential complications during pregnancy or childbirth.

The dataset contains 1,831 observations, each described by 21 features (x_1 to x_{21}) and a target class attribute indicating whether an observation is considered an outlier or not based on the fetal state (176 observations with complications during pregnancy). These attributes include various features extracted from the FHR and UC signals, which are numerical measurements describing aspects like the baseline fetal heart rate, variability, accelerations, and decelerations.

In the context of this dataset, outliers are defined as observations that exhibit abnormal or atypical patterns in the fetal heart rate or uterine contractions that deviate significantly

from the normal fetal conditions. These outliers may indicate potential fetal distress or other complications, making their identification crucial in a clinical setting for a timely intervention in clinical settings.

4.3.1 Outlier Detection

Isolation Forest Algorithm

An Isolation Forest model was trained with 100 estimators and a contamination rate of 0.096, reflecting the expected proportion of outliers in the dataset. The model's performance was evaluated using the AUC-ROC metric.

The ROC curves in Fig. A.26 shows the model's performance for detecting outliers both in the test dataset and in the overall dataset. The first ROC curve (Fig. A.26) demonstrates the performance of the model on the test dataset, with an AUC (Area Under the Curve) score of 0.93, which is indicative of a strong model capable of distinguishing between outliers and normal cases. The second ROC curve (Fig. A.26b) demonstrates similar strong performance on the entire dataset, with an AUC score of 0.93 as well. This high performance and consistency indicate the model's ability to distinguish between inliers and outliers.

LOF

LOF algorithm was applied to this dataset but with very low results, namely ROC-AUC of 0.63 (as seen in Fig. A.32), precision of 0.58, recall of 0.57 and F1-score of 0.58.

Random Forest Algorithm

Training a Random Forest on this dataset presented an outstanding ROC-AUC result of 100%, as seen in Fig. A.27.

The classification report presented in Table A.23 reinforce the performance where 176 outliers were correctly identified.

4.3.2 Explainable AI

Isolation Forest and SHAP Tree

The SHAP Summary Plot (in Fig. A.28) shows how features like x_3 , x_7 , and x_{20} (negative SHAP values) contribute significantly to outlier predictions, while features like x_{10} and x_{12} have smaller, more neutral impacts. x_3 (largest negative impact) has a strong negative impact on the model's decision, suggesting it is crucial in pushing observations toward being outliers; x_7 and x_{20} also contribute significantly in pushing instances away from the baseline, albeit to a slightly lesser degree compared to x_3 ; x_{10} , x_{12} exhibit a more neutral impact, which means they don't significantly drive the model's decision in most cases.

The boxplot in Fig. A.29 shows the distribution of SHAP values for each feature across all identified outliers, visualising the variability of contributions and identifying the key features driving the outlier classifications.

The features x_3 , x_7 , and x_{20} show large variations in SHAP values with significant negative impacts, reinforcing the idea that these features play a key role in predicting outliers. x_{10} and x_{12} are distributed closer to zero, suggesting their limited role in defining outliers.

Meanwhile, x_{21} and x_{16} exhibit wider distributions, indicating that in certain cases, they play important roles in either increasing or decreasing the likelihood of outlier classification.

Looking at the bar plot in Fig. A.30 helps to understand which features contribute differently between two specific outliers.

The features x_7 and x_{21} (largest differences) show the most prominent difference between the two outliers, with x_7 having a larger negative SHAP value for Outlier 2 compared to Outlier 1, while x_{21} shows the opposite trend. On the other hand, The features x_6 and x_{19} also exhibit noticeable differences in SHAP values, although the magnitude is smaller when compared to x_7 and x_{21} .

For individual outliers explanations, Force plots for specific outliers (see Fig. A.31) give an intuitive visualisation of how specific features push a prediction toward or away from outlier classification:

- For instance, in the force plot for Outlier 6 (Fig. A.31a), features x_1 , x_{10} , and x_{21} push the prediction towards "outlier" status, as seen by their contributions in red, positively impacting the prediction score. x_1 (0.44) and x_{21} (0.5) both have moderate positive SHAP values, contributing most strongly toward pushing this data point to outlier status. On the other hand, the blue contributions from features x_3 (0.84) and x_{20} (0.28) have high SHAP values, working to pull the prediction away from outlier status, but their impact is insufficient to classify this data point as a non-outlier.
- For Outlier 9 (Fig. A.31b), the red (positive SHAP values) contributions are even more significant for features such as x_1 (0.44) and x_4 (0.38), pushing the prediction significantly towards the outlier classification. Features like x_3 (0.93) and x_7 (0.26) play a negative role, pulling the prediction downward, but insufficient to push the data point to a non-outlier.
- The same analysis can be done to Outlier 11, 12 and 13 (see Figs., A.31c, A.31d and A.31e).

4.3.3 Explanations Clustering

Isolation Forest and SHAP Tree

The clustering analysis conducted on the SHAP values from the outlier explanations using KMeans provides valuable insights into the patterns of feature importance among different groups of outliers in this dataset.

The characterisation of the clusters can be seen in Table 4.7, the distribution of SHAP values by feature and cluster in Fig. A.33 and the mean absolute SHAP values by feature and cluster in Fig. A.34.

Cluster 0 contains the fewest number of outliers (19). The mean SHAP value (0.8779) is the highest among the three clusters, indicating that the features in this cluster have a strong impact on the predictions for these outliers. The highest SHAP standard deviation (0.8923) suggests a larger variability in how the features affect different outliers within the cluster. The top features for this cluster (x_{10} , x_{12} , x_8) show the key contributors to the model's predictions for these specific outliers. **A possible rule** is if a new instance has high values in x_{10} , x_{12} , or x_8 , it may fall into Cluster 0, where the model is likely to exhibit extreme predictions with a high SHAP contribution. Cluster 1 is the largest group, with 82

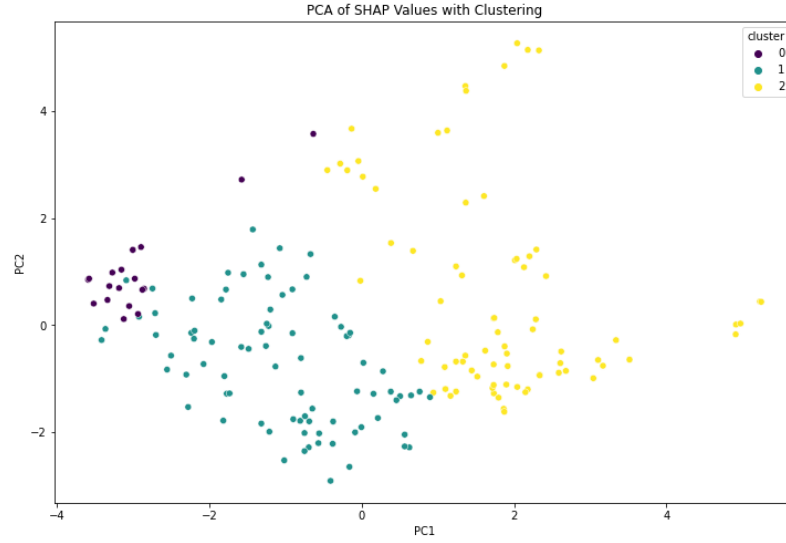


Figure 4.21: PCA of SHAP values with clustering

Cluster	0	1	2
Mean SHAP	0.8779	0.5404	0.5349
SHAP Std Dev	0.8923	0.0894	0.0520
Max SHAP Value	1.9062	0.6269	0.5838
Top Features	x10, x12, x8	x9, x20, x3	x7, x21, x18
Intra-Cluster SHAP Std Dev	0.4091	0.1770	0.1994
Number of Outliers	19	82	77

Table 4.7: Cluster Metrics Summary

outliers. This cluster has the lowest SHAP standard deviation (0.0894), indicating a more homogeneous set of outliers where the influence of features is more consistent across the group. The mean SHAP value (0.5404) and the max SHAP value (0.6269) are lower than in Cluster 0, suggesting that the features in this group have a weaker effect on predictions compared to Cluster 0. The top features for this group are x_9 , x_{20} , and x_3 . **A possible rule** is if instances with moderate shifts in x_9 , x_{20} , and x_3 are likely to be categorised into Cluster 1, where the model's predictions will be consistent but less extreme compared to Cluster 0. Cluster 2 is similar to Cluster 1 in terms of number of outliers (77) and mean SHAP values (0.5349), but it has a slightly lower standard deviation (0.0520). The top features in this cluster are x_7 , x_{21} , and x_{18} , which shows that this group of outliers has distinct driving factors compared to the other clusters. **A possible rule** is if features x_7 , x_{21} , and x_{18} contribute moderately to the prediction for instances in this cluster. Instances with fluctuations in these features are more likely to be outliers categorised here.

4.4 Results Discussion

Given all the knowledge obtained from these datasets, it is important to compile it in order to better prepare and guide the stage of perturbing the data through synthetic outliers. Therefore, by dataset, the main results are described below.

Breast Cancer dataset

In the Outlier Detection task, both Isolation Forest and Random Forest algorithms performed very well, achieving near perfect results. On the other hand, LOF performed on a poor way, which affected the attempt to compute SHAP explanations. For the few outliers that were coincident (either ground truth or any of the other algorithms), the explanations based on LOF were on the opposite direction of a model based on a good ROC-AUC performance.

Regarding global explanation of the outliers, the technique SHAP with Tree Explainer was able to identify x_8 , x_6 and x_5 as features with the most significant impact, which are exactly the same the technique SHAP with Kernel Explainer identified. For the individual explanations, either Tree and Kernel Explainer relate x_9 as frequent contributor towards the outlier classifications, specially when its value is low, and x_6 , x_5 and x_4 as often contributors towards inlier classification. According to LIME technique, one feature condition of x_9 is the dominant one, followed by conditions for x_6 , x_5 and x_4 .

The most consistent features across all techniques seem to be x_4 , x_5 , x_6 , x_8 , and x_9 , which frequently appear in combination across different clusters, regardless of the method used for explainability. The rules derived for this dataset are:

- If an observation has high values for both x_4 and x_5 , it is likely to be classified as an outlier;
- High values of x_6 , x_4 , and x_8 consistently push observations toward outlier classification;
- $x_9 \leq -0.35$ combined with high values of x_4 or x_5 leads to strong outlier predictions (high values of x_4 , and x_8 is present in 3 of the clusters analysed from Tree Explainer and Kernel Explainer);
- Observations with high values in x_7 , x_5 , and x_4 are highly likely to be classified as outliers.

Seismic dataset

The possible rules derived from the combination of Isolation Forest and SHAP (Tree Explainer) with 3 clusters are as follow.

- Cluster 0 (Smaller Seismic Events) where instances are likely influenced by the number of bumps ($nbumps$, $nbumps3$, and $nbumps2$), where high values in these features without extreme energy readings may indicate smaller seismic events or more frequent but less significant anomalies;
- Cluster 1 (Energy-Driven Outliers) is characterised by high *energy*, *maxenergy*, and *nbumps4* values, which suggest that the outliers represent significant seismic events with high energy levels, and the presence of these features in high quantities could serve as a strong indicator of major anomalies;
- Cluster 2 (Hazard-Related Outliers) likely indicate events where *ghazard*, *gdpuls*, and *gdenenergy* significantly contribute to anomaly detection. A combination of high values for these features could serve as an indicator of seismic-related anomalies.

The rules derived from the combination of Random Forest and SHAP (Tree Explainer) with 2 clusters are as follow.

- Cluster 0 seems to group the energy-driven outliers. If *nbumps4*, *energy*, and *maxenergy* have high SHAP values, the instance is likely an outlier due to energy-related anomalies;
- Cluster 1 groups hazard-related outliers. If *ghazard*, *gdpuls*, and *gdenery* are highly significant, the instance is likely an outlier related to potential hazard or dangerous events.

Both approaches derived similar rules, being the only difference that the approach with 3 clusters was able to derive an additional specific rule for smaller seismic events.

Cardiotocography Dataset

The possible rules derived from the combination of Isolation Forest and SHAP (Tree Explainer) with 3 clusters are as follow.

- Instances with extremely high or low values in *x10*, *x12*, and *x8* are likely to be categorised as outliers in Cluster 0;
- Moderate deviations in *x9*, *x20*, and *x3* can lead to outlier detection in Cluster 1;
- Consistent but moderate deviations in *x7*, *x21*, or *x18* are likely to categorise outliers in cluster 2, with a more homogeneous impact from these features.

4.5 Perturbing the datasets - Synthetic Outliers

In this part of the dissertation, outliers were analytical generated based on true inliers, which were then perturbed on the features identified by the possible rules derived in the previous steps.

Perturbing Breast Cancer

The first experience was computed for 10 synthetic outliers. From the original dataset, observations 1 to 11 (excluding the observation 6 which was already an outlier) were the input of the Isolation Forest algorithm, which was able to classify them with an accuracy of 100% accuracy: "df_y_pred: array([1, 1, 1, 1, 1, 1, 1, 1, 1, 1])".

The magnitude of the perturbations took into consideration the distribution of the values of each feature, trying to place each perturbed value outside the Interquartile Range.

After these outstanding results, a wider experiment was conducted, perturbing all the inliers of the original dataset according to the above mentioned rules and obtaining a new dataset composed of 444 potential outliers. The perturbations followed the same methodology. The results were outstanding, where the Isolation Forest identified all the 444 instances as inliers: "sum(df_y_pred) = 444".

In order to fully test the hypothesis of this dissertation, a last experiment was performed, by perturbing features values not pointed by the rules previously derived, mainly *x1* and *x2*. The perturbation followed the same assumptions as before. A small test of 10 instances was performed and the results were really interesting once the algorithm classified all of the perturbed observations as inliers: "df_y_pred = array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0])".

Chapter 5

Conclusion

The explainability of decision-making models is a growing area of focus within the scientific community, and it has begun to gain attention on a global scale. This dissertation has contributed to this rapidly evolving field, specifically, on the explainability of rare events, furthering the research within Explainable AI. XAI plays a crucial role in ensuring that machine learning models provide not only accurate predictions but also clear and transparent reasoning behind those decisions. This field aims to transform the detection of outliers into actionable knowledge, revealing patterns and generating rules that enhance interpretability and foster trust in AI-driven systems.

The methodology consisted in detecting outliers (against their ground truth), explain those classifications and clustering the explanations in order to derive rules from the data. These rules are deeper insights into the challenges covered, which allowed to experiment those findings against synthetic outliers, and even synthetic inliers.

Validating the rules

The culmination of this study indicates the robustness of the rules derived from the clustering of explanations.

For similar ROC-AUC performance levels, the rules obtained from clustering the explanations tend to be consistent across different approaches. Even when adapting the methodology, by changing the outlier detection and/or explanations algorithms, rules with equivalent interpretative and explanatory capacity were generated. As these rules were derived through the clustering process, it is important to highlight the role of a model with good performance, which provided consistency across multiple approaches. Perturbing the features identified by the derived rules of the original inlier observations, consistently led to accurate classification of the generated instances as outliers. On the other hand, perturbing features not implicated by the rules (e.g., x_1 and x_2 in the breast cancer dataset) resulted in inlier classifications, further validating the performance of the derived rules, which indicates that the clustering and rule-generation process not only identifies critical feature combinations but also distinguishes between meaningful and irrelevant perturbations.

The ability of the algorithm to perfectly classify synthetic instances based on the derived rules, whether as outliers or inliers, shows the success of this approach in translating complex patterns into actionable insights. This outcome not only validates the methodology but also highlights its potential for broader applications across various domains, such as fraud detection, medical diagnostics, environmental monitoring, among others, where understanding

and interpreting rare events is critical for decision-making processes.

5.1 Contributions

This dissertation contributes to the field of Explainable AI, particularly in the explanation of rare events.

One of the key contributions is the novel approach of clustering the outlier explanations, which, to the best of my knowledge, has not been extensively explored in the literature. This approach might be useful for extracting valuable insights from abnormal or rare events, where individual outlier explanations may be overwhelming or difficult to interpret individually. By clustering the explanations, stronger patterns and behaviours are revealed, allowing more detailed analysis and interpretation of the data.

Additionally, the research demonstrates the potential of generating synthetic outliers and highlights the opportunities for automating this process in future work. It also raises awareness of the computational challenges associated with more advanced techniques like SHAP Kernel Explainer, due to its high computational cost.

Overall, this dissertation contributes not only to the technical aspects of XAI but also opens pathways for further research into multiclass challenges, feature clustering, and enhanced explainability. It lays important groundwork for making AI systems more transparent, interpretable, and trustworthy, particularly in scenarios where rare events are critical.

5.2 Limitations and Future Work

While the results of this study demonstrate the potential of various outlier detection and explanation methods, certain limitations must be acknowledged. First, the experiments were conducted using specific datasets and algorithms, which may limit the generalisability of the findings to other domains or datasets. Future studies should explore a wider variety of datasets, particularly those from different fields, to validate the methods presented and assess their broader applicability. Similarly, while the methods employed for outlier detection, explanation, and clustering have shown promise, the use of different algorithms and techniques may yield more robust or diverse results, warranting further investigation.

Another significant limitation is the computational cost of SHAP Kernel Explainer. Due to its high computational requirements, this approach was not fully explored within the scope of this research. Future research should aim to investigate this method more thoroughly by optimising computational resources or applying it to smaller-scale datasets to gain deeper insights into its performance for explaining outliers, avoiding limitations to real-time applications.

Several directions for future research can be drawn from this study. One area for future exploration is tackling multiclass challenges. Extending the current methods to multiclass classification scenarios can provide a richer understanding of how outliers behave in more complex environments, where different classes may have varying degrees of abnormality.

Additionally, exploring advanced interpretability techniques such as dependencies among attributes and feature clustering approaches could provide deeper insights and enhance the understanding of the key drivers behind anomalies and further refine the explanation process.

The outliers used in this study were analytically generated, which opens up the opportunity for future work to improve and automate this process. Automating the generation

of synthetic outliers could streamline the analysis and enable a more systematic exploration of outlier detection methods. Furthermore, conducting a sensitivity analysis to refine the outlier generation process based on the rules derived from clustering would provide a more sophisticated approach to extract even more information from the data.

Finally, future research could also focus on enhancing the clustering methods applied to explanations, potentially leading to the discovery of more intricate patterns and rules. By continuing to refine and build upon the current findings, there is great potential to advance the field of explainable artificial intelligence.

Bibliography

- Atzmueller, M., Fürnkranz, J., Kliegr, T., & Schmid, U. (2024, 07). Explainable and interpretable machine learning and data mining. *Data Mining and Knowledge Discovery*, 38. doi: 10.1007/s10618-024-01041-y
- Boukerche, A., Zheng, L., & Alfandi, O. (2020). Outlier detection: Methods, models, and classification. *ACM Computing Surveys (CSUR)*, 53(3), 1–37.
- Bouman, R., Bukhsh, Z., & Heskes, T. (2023). *Unsupervised anomaly detection algorithms on real-world data: how many do we need?* Retrieved from <https://arxiv.org/abs/2305.00735>
- Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). Lof: identifying density-based local outliers. In *Proceedings of the 2000 acm sigmod international conference on management of data* (p. 93–104). New York, NY, USA: Association for Computing Machinery. Retrieved from <https://doi.org/10.1145/342009.335388> doi: 10.1145/342009.335388
- Campos, G., Zimek, A., Sander, J., Campello, R., Micenková, B., Schubert, E., ... Houle, M. (2016, 07). On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study. *Data Mining and Knowledge Discovery*, 30. doi: 10.1007/s10618-015-0444-8
- Cummins, L., Sommers, A., Ramezani, S. B., Mittal, S., Jabour, J., Seale, M., & Rahimi, S. (2024). Explainable predictive maintenance: A survey of current methods, challenges and opportunities. *arXiv preprint arXiv:2401.07871*.
- Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd international conference on knowledge discovery and data mining (kdd-96)* (pp. 226–231).
- Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5), 1–42.
- Kopp, M., Pevný, T., & Holena, M. (2014). Interpreting and clustering outliers with sapling random forests. *Information Technologies—Applications and Theory*.
- Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. In *2008 eighth ieee international conference on data mining* (pp. 413–422).
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon et al. (Eds.), *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf
- Molnar, C. (2018). A guide for making black box models explainable. URL: <https://christophm.github.io/interpretable-ml-book>, 2(3).
- Mozolewski, M., Bobek, S., Ribeiro, R. P., Nalepa, G. J., & Gama, J. (2024). Towards evaluation of explainable artificial intelligence in streaming data. In L. Longo, S. Lapuschkin, & C. Seifert (Eds.), *Explainable artificial intelligence* (pp. 145–168). Cham: Springer Nature

Switzerland.

- Panji, E., Gruenwald, L., Leal, E., Nguyen, C., & Silvia, S. (2022). A survey on outlier explanations. *The VLDB Journal*, 31(5), 977–1008.
- Sejr, J. H., & Schneider-Kamp, A. (2021). Explainable outlier detection: What, for whom and why? *Machine Learning with Applications*, 6, 100172.
- Sokol, K., & Vogt, J. E. (2024). What does evaluation of explainable artificial intelligence actually tell us? a case for compositional and contextual validation of xai building blocks. In *Extended abstracts of the chi conference on human factors in computing systems (chi ea '24)*. ACM. Retrieved from <https://doi.org/10.1145/3613905.3651047> doi: 10.1145/3613905.3651047

Appendix A

Appendix

A.1 Outliers & DataSets (ODDS)

Figure A.1: Outlier Detection Techniques

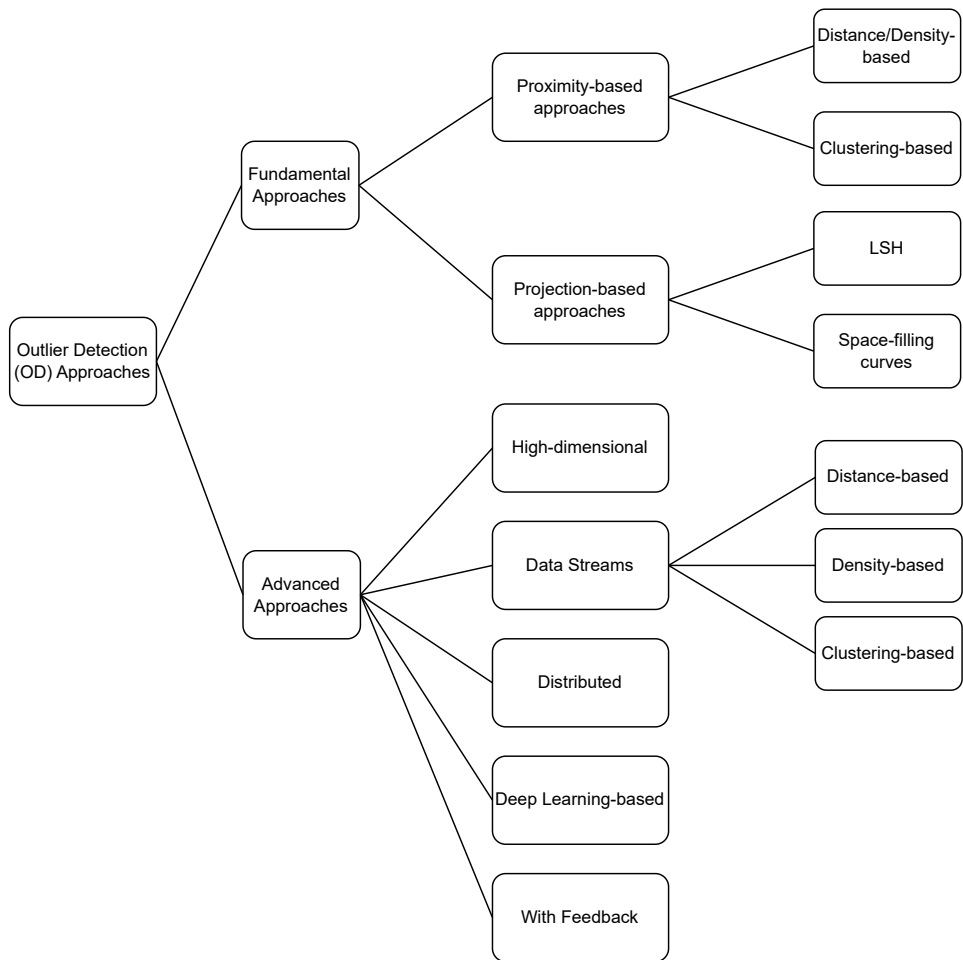


Table A.1: Examples of possible multi-dimensional point datasets from ODD

Dataset	# points	# dim.	# outliers	% outliers
Lympho	148	18	6	4.1%
WBC	278	30	21	5.6%
Glass	214	9	9	4.2%
Vowels	1,456	12	50	3.4%
Cardio	1,831	21	176	9.6%
Thyroid	3,772	6	93	2.5%
Musk	3,062	166	97	3.2%
Satimage-2	5,803	36	71	1.2%
Letter Recognition	1,600	32	100	6.3%
Speech	3,686	400	61	1.7%
Pima	768	8	268	35.0%
Satellite	6,435	36	2,036	32.0%
Shuttle	49,097	9	3,511	7.0%
BreastW	683	9	239	35.0%
Arrhythmia	452	274	66	15.0%
Ionosphere	351	33	126	36.0%
Mnist	7,603	100	700	9.2%
Optdigits	5,216	64	150	3.0%
Http (KDDCUP99)	567,479	3	2,211	0.4%
ForestCover	286,048	10	2,747	0.9%
Mulcross	262,144	4	26,214	10.0%
Smtip (KDDCUP99)	95,156	3	30	0.0%
Mammography	11,183	6	260	2.3%
Annnthyroid	7,200	6	534	7.4%
Pendigits	6,870	16	156	2.3%
Ecoli	336	7	9	2.6%
Wine	129	13	10	7.7%
Vertebral	240	6	30	12.5%
Yeast	1,364	8	64	4.7%
Seismic	2,584	11	170	6.5%
Heart	224	44	10	4.4%
OSAD Benchmark Datasets	Multiple datasets	—	—	—
One-class dataset by David Tax	Multiple datasets	—	—	—

A.2 Outlier Detection Algorithms & Techniques:

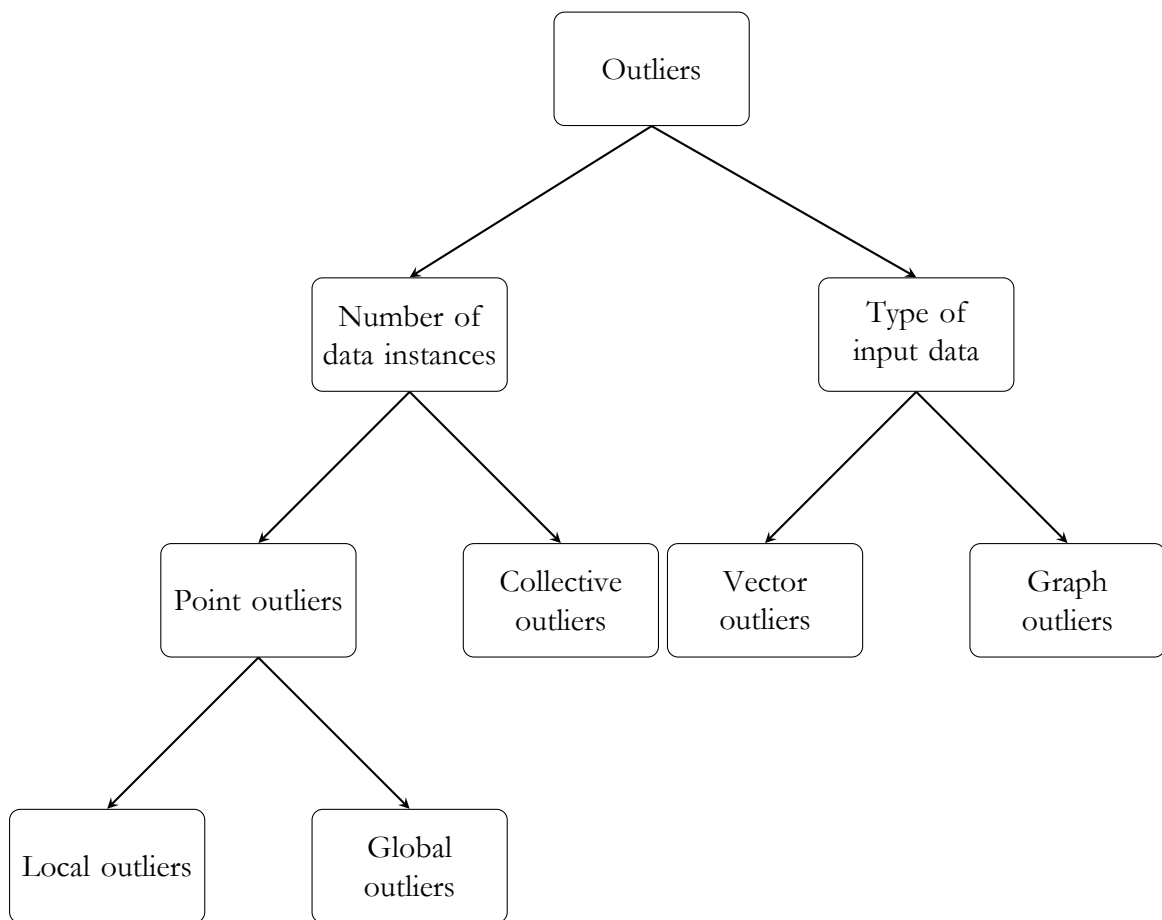


Figure A.2: Hierarchical Diagram of Outliers Categorization

Algorithm	Features	Time Complexity	Local Outlier	Comments
LOF	Density normalized by the density of the neighbors	$O(N^2)$	Yes	First to address local outliers. Fails to appropriately score the instances on the borders of clusters with significantly different densities
COF	Shortest path to connect neighbors	$O(N^2)$	Yes	Address non-spherical distributions
LOCI	Count-based neighborhood density	$O(N^3)$	Yes	Free of parameters but high time complexity
INFLO	Reversed nearest neighbors (k-RNN)	$O(N^2)$	Yes	Improvement on border points between two areas with different densities
LoOP	Assuming Gaussian distribution of distances to k-NN	$O(N^2)$	Yes	Interpretability of output
iNNE	Subsampling and ensembling	$O(N\psi_t)$	Yes	Highly efficient
LeSiNN	Subsampling and ensembling	$O(N\psi_t)$	No	Intuitive & highly efficient

Table A.2: Nearest-neighbor-based Outlier Detection

Algorithm	Clustering Algorithm	Features	Score Based on	Predefined Cluster #	Comments
Jiang et al.	k-means	Minimal spanning tree of cluster centers	Cluster cardinality	Yes	Time efficient but coarse granularity & spherical clusters
CBLOF	Arbitrary	Heuristic small and large clusters	Cluster cardinality & distance to cluster center	N/A	Too many parameters & misuse of cluster cardinality
LDCOF	Arbitrary	Local normalization	Cluster cardinality & distance to cluster center	N/A	Detection of local outliers but too many parameters
Du et al.	Density peak clustering	Chebyshev's theorem for statistical threshold determination	Standard deviations of δ	No	Intuitive & arbitrary-shaped clusters

Table A.3: Clustering-based Outlier Detection

Algorithm	Projection Technique	Base Outlier Detector	Features	Scalability to High Dimensionality	Comments
PINN	Random projection	LOF	Approximate k-NN	Good	Improved time efficiency & high accuracy
LSOD	LSH	kth-NN distance	LSH-based ranking & pruning	Medium	Early detection of top outliers
Schubert et al.	Space-filling curve	LOF	Approximate k-NN & ensemble	Poor	Near-linear complexity & easy distributed deployment
Loda	Sparse random projection clustering	Histogram-based outlier detector	One-dimensional histogram ensemble	Good	Linear complexity & high accuracy & handling missing values
Isolation Forest	Binary tree	N/A	Ensemble & subsampling	Good	Linear complexity & high accuracy
Extended iForest	Binary tree	N/A	Random hyperplane cuts	Good	Improved accuracy

Table A.4: Projection-based Outlier Detection

Algorithm	Features	Improve Efficiency/Effectiveness	Subspace	Comments
RBRP	Recursive partitioning & approximate k-NN	Efficiency	No	Fast search of approximate k-NN
PINN	Random projection & dimension reduction	Efficiency	No	Improved efficiency but approximate results
ABOD	Angle variance	Effectiveness	No	High time complexity
Kriegel et al.	Axis-parallel hyperplane	Effectiveness	Yes	Interpretability of result & High accuracy
HiCS	Statistical test & ensemble	Effectiveness	Yes	Improved accuracy & generalized pre-processing method
RS-Hash	Randomized hashing & ensemble	Efficiency & Effectiveness	Yes	Linear complexity & high accuracy & interpretability of results

Table A.5: Outlier Detection for High-dimensional Data

Algorithm	Features	Memory Complexity	Outlier Storage	Count-based/ Time-based Sliding Window	Comments
STORM	Safe inliers	$O(Wk)$	None	Count-based	High memory usage & high time complexity
Abstact-C	Pre-handling neighbor counts	$O(W^2/S + W)$	None	Both	Improved execution time
DUE	Event queue	$O(Wk)$	Outlier list	Both	Efficient at re-evaluating outlier at runtime
MCOD	Event queue & Micro-cluster	$O(cW + (1 - c)kW)$	Outlier list	Both	Improved execution time & improved memory complexity
Thresh_LEAP	Separate index per slide & Minimal probing	$O(W^2/S)$	Outlier candidate list per slide	Both	Improved execution time but potentially more memory usage

Table A.6: Distance-based Outlier Detection in Data Streams

Algorithm	Features	Window Type	Time Complexity	Memory Complexity	Comments
Incremental LOF	Selectively updating records	Landmark window	$O(N \log N)$	$O(Nk)$	High memory complexity & high time complexity
MiLOF	c-means-based summarization	Sliding window	$O(N \log W)$	$O(Wk)$	Low memory complexity & time complexity but compromised accuracy
DILOF	Density-based summarization	Sliding window	$O(NW)$	$O(Wk)$	Low memory complexity & time complexity & high accuracy

Table A.7: Density-based Outlier Detection in Data Streams

Algorithm	Features	Window Type	Time Complexity	Memory Complexity	Comments
D-Stream	Grid-based clustering	Damped window	Density of a grid	No	Arbitrary-shaped clusters but poor scalability to high dimension
Elahi et al.	k-means & delayed determination of outliers	Non-overlapping sliding window	Distance to cluster centers	Yes	Assuming spherical clusters & too many parameters
AnyOut	Hierarchical clustering & ClusTree	Damped window	Distance to cluster centers & Gaussian probability density	No	Real-time & varying granularity but assuming spherical clusters
Salehi et al.	Hyperellipsoidal clustering	Non-overlapping sliding window	Whether belongs to a cluster	No	Time efficient & addressing switching data streams
HyCARCE Chenaghlou et al.	Hyperellipsoidal clustering & Gaussian clusters	Non-overlapping sliding window	Gaussian probability density	No	Time efficient & addressing emerging distributions

Table A.8: Clustering-based Outlier Detection in Data Streams

Algorithm	Base Outlier Detector	Distributed Infrastructure	Features	Comments
Bhaduri et al.	ORCA	Network of ring topology	Top-N pruning	High communication overhead
Angiulli et al.	Solvingset	Ethernet network with TCP sockets	Top-N pruning	Impaired scalability due to broadcasting
DLOF	LOF	Hadoop MapReduce	Grid-based partitioning & duplication reduction	Reduced communication overhead but not scalable to high dimensionality
DTOLF	LOF	Hadoop MapReduce	Grid-based partitioning & top-N pruning	Reduced communication overhead but not scalable to high dimensionality
OW-OCRF	One-class random forest	Wireless sensor network	Weighted ensemble	Real-time response & high accuracy

Table A.9: Distributed Outlier Detection

Algorithm	Techniques	Data Label	Comments
Chen et al.	Autoencoder & ensemble	Unlabeled	Increased computation cost
RDA	Robust autoencoder	Unlabeled	High sensitivity to hyperparameters
AnoGAN	GAN	Inlier label	High computation cost
ALAD	GAN	Inlier label	Improved efficiency and accuracy
RAMODO	Artificial neural network	Unlabeled	Customized representation but dependent on the quality of candidate sets

Table A.10: Deep Learning-based Outlier Detection

Table A.11: Outlier Detection with Feedback

Algorithm	Features	Base Out-lier Detector	Active Learning Strategy	Comments
Görnitz et al.	Tailored active learning strategy	SVDD	Clusters near decision boundary	Robust against obfuscation & quick accuracy boost over a few labels
Das et al.	AATP & adjusted ensemble weights	Loda	The most anomalous instances	High accuracy & generalizable
Vercruyssen et al.	Label propagation	Constrained k-means based	Decision boundary	High accuracy
Siddiqui et al.	Online convex optimization	Isolation Forest	The most anomalous instances	High accuracy & high efficiency & generalizable

A.3 Outlier Explanations

Table A.12: Decision Systems Explanations Methods

Technique	Authors	Year	Problem	Ex-planator	Black-Box	Data Type	General	Random	Ex-amples	Code	Data set
Trepan	Craven et al.	1996	ModelExpl.	DT	NN	TAB	x			x	
-	Krishnan et al.	1999	ModelExpl.	DT	NN	TAB	x	x	x		
DecText	Boz	2002	ModelExpl.	DT	NN	TAB	x	x		x	
GPD ^T	Johansson et al.	2009	ModelExpl.	DT	NN	TAB	x	x	x		x
TreeMetrics	Chipman et al.	1998	ModelExpl.	DT	TE	TAB				x	
CCM	Domingos et al.	1998	ModelExpl.	DT	TE	TAB	x	x		x	
-	Gibbons et al.	2013	ModelExpl.	DT	TE	TAB	x	x			
STA	Zhou et al.	2016	ModelExpl.	DT	TE	TAB		x			
CDT	Schetinin et al.	2007	ModelExpl.	DT	TE	TAB			x		
-	Hara et al.	2016	ModelExpl.	DT	TE	TAB	x	x	x		
TSP	Tan et al.	2016	ModelExpl.	DT	TE	TAB				x	
ConjRules	Craven et al.	1994	ModelExpl.	DR	NN	TAB	x				
G-REX	Johansson et al.	2003	ModelExpl.	DR	NN	TAB	x	x	x		
REFNE	Zhou et al.	2003	ModelExpl.	DR	NN	TAB	x	x	x		x
RxREN	Augusta et al.	2012	ModelExpl.	DR	NN	TAB	x	x	x		x
SVM+P	Nunez et al.	2002	ModelExpl.	DR	SVM	TAB		x		x	
-	Fung et al.	2005	ModelExpl.	DR	SVM	TAB		x		x	
inTrees	Deng	2014	ModelExpl.	DR	TE	TAB		x		x	
-	Lou et al.	2013	ModelExpl.	FI	AGN	TAB	x	x	x	x	
GoldenEye	Henelius et al.	2014	ModelExpl.	FI	AGN	TAB	x	x	x	x	x
PALM	Krishnan et al.	2017	ModelExpl.	DT	AGN	ANY	x	x	x		
-	Tolomei et al.	2017	ModelExpl.	FI	TE	TAB		x			
-	Xue et al.	2015	OutcomeExpl.	SM	DNN	IMG		x	x	x	
-	Fong et al.	2017	OutcomeExpl.	SM	DNN	IMG		x		x	
CAM	Zhou et al.	2016	OutcomeExpl.	SM	DNN	IMG		x	x	x	
Grad-CAM	Selvaraju et al.	2016	OutcomeExpl.	SM	DNN	IMG		x	x	x	
-	Lei et al.	2016	OutcomeExpl.	SM	DNN	TXT		x		x	
LIME	Ribeiro et al.	2016	OutcomeExpl.	FI	AGN	ANY	x	x	x	x	x
MES	Turner et al.	2016	OutcomeExpl.	DR	AGN	ANY	x		x	x	
NID	Olden et al.	2002	Inspection	SA	NN	TAB		x			
GDP	Bachrens	2010	Inspection	SA	AGN	TAB	x	x	x		
IG	Sundararajan	2017	Inspection	SA	DNN	ANY		x	x		
VEC	Cortez et al.	2011	Inspection	SA	AGN	TAB	x	x	x		
VIN	Hooker	2004	Inspection	PDP	AGN	TAB	x	x	x		
ICE	Goldstein et al.	2015	Inspection	PDP	AGN	TAB	x	x		x	x
Prospector	Krause et al.	2016	Inspection	PDP	AGN	TAB	x	x	x	x	
Auditing	Adler et al.	2016	Inspection	PDP	AGN	TAB	x	x	x	x	
OPIA	Adebayo et al.	2016	Inspection	PDP	AGN	TAB	x	x			
-	Yosinski et al.	2015	Inspection	NA	DNN	IMG		x	x		
TreeView	Thiagarajan et al.	2016	Inspection	DT	DNN	TAB		x	x		
IP	Shwartz et al.	2017	Inspection	NA	DNN	TAB		x			
-	Radford	2017	Inspection	NA	DNN	TXT		x			
CPAR	Yin et al.	2003	Transp. Design	DR	-	TAB				x	
FRL	Wang et al.	2015	Transp. Design	DR	-	TAB		x	x	x	
BRL	Letham et al.	2015	Transp. Design	DR	-	TAB		x			
TLBR	Su et al.	2015	Transp. Design	DR	-	TAB		x		x	
IDS	Lakkaraju et al.	2016	Transp. Design	DR	-	TAB		x			
RuleSet	Wang et al.	2016	Transp. Design	DR	-	TAB		x	x	x	
1Rule	Malioutov et al.	2017	Transp. Design	DR	-	TAB		x		x	
PS	Bien et al.	2011	Transp. Design	PS	-	ANY		x	x		
BCM	Kim et al.	2014	Transp. Design	PS	-	ANY		x	x		
-	Mahendran et al.	2015	Transp. Design	PS	-	IMG		x	x	x	
-	Kononenko et al.	2010	Transp. Design	FI	-	TAB		x		x	
OT-SpAMs	Wang et al.	2015	Transp. Design	DT	-	TAB		x	x	x	

Table A.13: Outlier Explanation Techniques Summary

Technique	Reference	1 MA	2 DSA	3 NR	4 CR	5 CIO	6 OAI	7 OAG	8 PK	9 UF	10 DS	11 Par	12 DSy	13 PC
–	Knorr and Ng				x		x							
–	Chen et al.							x	x					
URCA	Silveira and Diot	x	x					x						
KKAS	Huang and Yang							x						
STOTree	Liu et al.		x			x								
–	Kriegel et al.	x		x										
–	Viswanathan et al.	x	x	x							x		x	
LODI	Dang et al.	x					x							
–	Micenkova et al.	x					x							
Refout	Keller et al.	x					x							
Explainer	Kopp et al.	x					x							
LOGP	Dang et al.						x							
–	Xing et al.		x			x								
Exstream	Zhang and He		x				x				x		x	
COIN	Liu et al.	x		x			x		x					x
EXAD	Song et al.		x				x				x		x	
–	Amarasinghe et al.						x							
XPACS	Macha and Akoglu	x						x				x		x
GLAD	Siddiqui et al.	x		x						x			x	x
Lookout	Gupta et al.	x					x					x		x
SFE	Siddiqui et al.						x							
–	Takeishi						x							

1 - Model Agnostic, 2 - Domain-specific Application, Explanations provided: [3 - Numerical Ranking, 4 - Categorical ranking, 5 - Causal Interaction among Outliers, 6 - Outlying Attributes of an Individual Outlier, 7 - Outlying Attributes of a Group of Outliers], 8 - Prior Knowledge of the Dataset, 9 - Incorporating Users' Feedback, 10 - Data Stream, 11 - Parallelism, 12 - Distributed System, 13 - Public availability of Source Code

A.4 Techniques to generate outliers

Table A.14: Summary of Techniques Generating Numerical Ranking of Outliers

Name	References	Outlier Detection Model	How Outlier Scores Are Generated	Challenges Addressed
–	Chandola et al.	Neural network	Reconstruction error	None
		Rule-based	Inverse of the confidence associated with the best rule	None
		Clustering	Inverse of the confidence associated with the best rule	None
		Gaussian model	The distance of each instance from the estimated mean	None
		Regression	Residual’s magnitude	None
		Density-based	The inverse of the probability density value	None
		Histogram-based	Bin’s height where each instance falls into	None
		SVM	Probabilistic prediction value of each instance	None
–	Kriegel et al.	Model-agnostic	Regularization and normalization-based scoring	Unifying various outlier scores
–	Viswanathan et al.	Model-agnostic	False positive rate-based scoring	Unifying various outlier scores
GLADs	Siddiqui et al.	Model-agnostic	User’s feedback adaptive scoring	Incorporating user’s feedback
COIN	Liu et al.	Model-agnostic	Nearest-normal neighbors-based scoring	Unifying various outlier scores

Table A.15: Summary of Techniques Finding Outlying Attributes of Individuals and Groups

Name	Reference	Outlier Detection Model	Explanation Methodology	Output	Data Type Supported	Future Outlier Prediction	Challenges Addressed
–	Knorr and Ng	Distance-based	Heuristic-based subspace search	Sets of outlying attributes	Numerical	No	Limiting subspace search & Generating readily interpretable output
LODI	Dang et al.	Model-agnostic	Local neighborhood	Contribution score for each attribute	Numerical	No	Generating readily interpretable output
–	Micenková et al.	Model-agnostic	Local neighborhood	Contribution score for each attribute	Numerical	No	Generating readily interpretable output
Refout	Keller et al.	Model-agnostic	Heuristic-based subspace search	Sets of outlying attributes	Any	No	Limiting subspace search & Generating readily interpretable output
Explainer	Kopp et al.	Model-agnostic	Decision Tree	DNF	Any	No	Generating readily interpretable output
LOGP	Dang et al.	Graph	Local neighborhood	Contribution score for each attribute	Numerical	No	Generating readily interpretable output
Exstream	Zhang and He	User-annotated CEP monitoring	Entropy-based reward	CNF	Numerical	Yes	Generating readily interpretable output
COIN	Liu et al.	Model-agnostic	Local neighborhood	Contribution score for each attribute	Numerical	No	Generating readily interpretable output & Incorporating user's prior knowledge
EXAD	Song et al.	Neural network	Decision Tree	DNF	Numerical	No	Generating readily interpretable output
–	Amarasinghe et al.	Neural network	Layer-wise relevance propagation	Contribution score for each attribute	Numerical	Yes	Generating readily interpretable output
Lookout	Gupta et al.	Model-agnostic	Heuristic-based subspace search	Set of focus plots	Numerical	No	Limiting subspace search & Generating readily interpretable output
SFE	Siddiqui et al.	Density-based	Branch-and-bound search	Sets of outlying attributes	Numerical	No	Limiting subspace search & Generating readily interpretable output
–	Takeishi	PCA	Shapley values	Shapley values of all	Numerical	No	Generating readily interpretable output
–	Chen et al.	Distance-based	Apriori-based	Set of outlying attributes and their values	Any	No	Incorporating user's prior knowledge & Finding discriminative outlying attributes
URCA	Silveira and Diot	Model agnostic	Graph-based	Graph of attribute flows and their values	Any	Yes	Finding discriminative outlying attributes
KKAS	Huang and Yang	Distance-based	Attribute Reduction-based	Set of outlying attributes	Numerical	No	Finding discriminative outlying attributes
XPACS	Macha and Akoglu	Model-agnostic	Subspace Clustering-based	Conjunction of attributes' value intervals	Numerical	Yes	Limiting subspace search & Generating readily interpretable output & Finding discriminative outlying attributes

A.5 Experimental Results Aux

Class	Precision	Recall	F1-Score	Support
0	0.50	0.95	0.65	471
1	0.64	0.08	0.14	495
Accuracy			0.51	966
Macro Avg	0.57	0.52	0.40	966
Weighted Avg	0.57	0.51	0.39	966

Table A.16: Isolation Forest (with SMOTE) Seismic Test Sample Classification Evaluation Report

Class	Precision	Recall	F1-Score	Support
0	0.94	0.95	0.94	2414
1	0.13	0.11	0.12	170
Accuracy			0.89	2584
Macro Avg	0.53	0.53	0.53	2584
Weighted Avg	0.88	0.89	0.89	2584

Table A.17: Isolation Forest (with SMOTE) Seismic Dataset Classification Evaluation Report

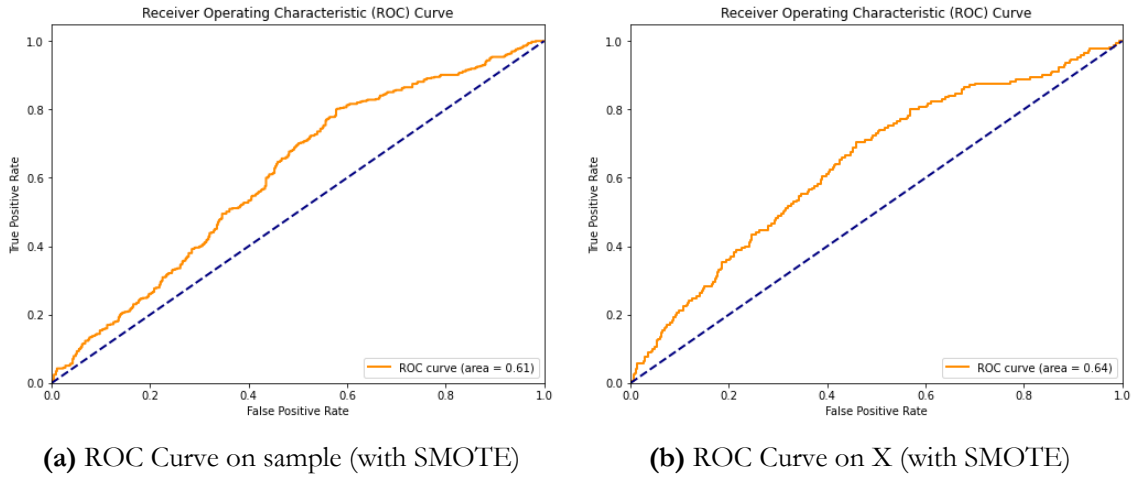
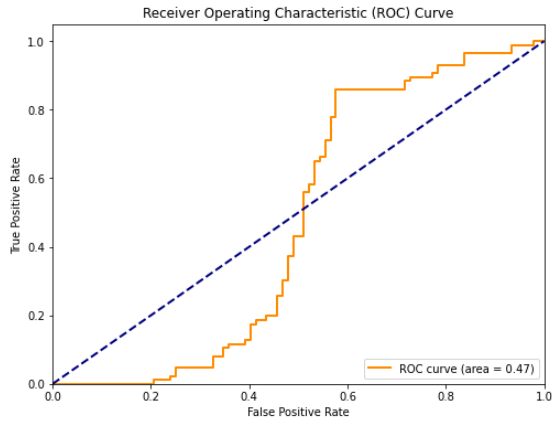
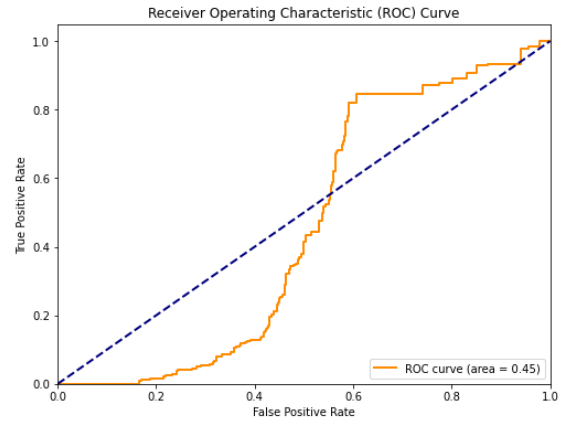


Figure A.3: Seismic ROC Curve: IF Evaluation metrics with SMOTE



(a) ROC Curve on X_{test} (LOF) - seismic



(b) ROC Curve on X (LOF) - seismic

Figure A.4: ROC Curve: Evaluation metrics (LOF) - seismic

Class	Precision	Recall	F1-Score	Support
0	0.80	0.93	0.86	92
1	0.92	0.76	0.83	86
Accuracy			0.85	178
Macro Avg	0.86	0.85	0.85	178
Weighted Avg	0.86	0.85	0.85	178

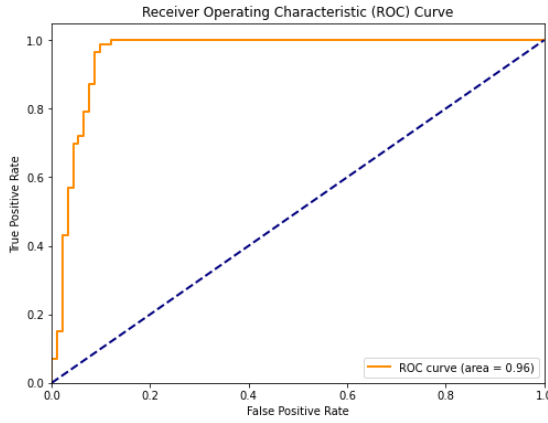
Table A.18: Isolation Forest (with SMOTE) Breast Cancer Test Sample Classification Evaluation Report

Class	Precision	Recall	F1-Score	Support
0	0.87	0.96	0.91	444
1	0.91	0.73	0.81	239
Accuracy			0.88	683
Macro Avg	0.89	0.85	0.86	683
Weighted Avg	0.88	0.88	0.88	683

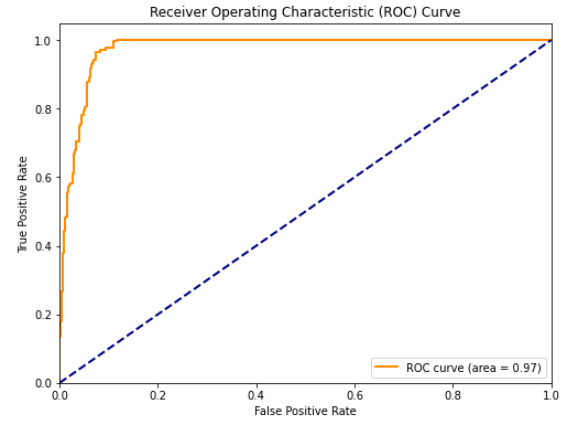
Table A.19: Isolation Forest (with SMOTE) Breast Cancer Wisconsin Dataset Classification Evaluation Report

Class	Precision	Recall	F1-Score	Support
0	0.99	0.97	0.98	87
1	0.94	0.98	0.96	50
Accuracy			0.97	137
Macro Avg	0.97	0.97	0.97	137
Weighted Avg	0.97	0.97	0.97	137

Table A.20: Test Sample Classification Evaluation Report (Random Forest) - breast cancer

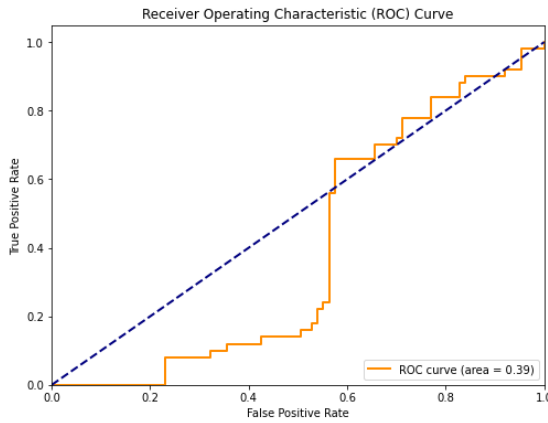


(a) ROC Curve on sample (with SMOTE)

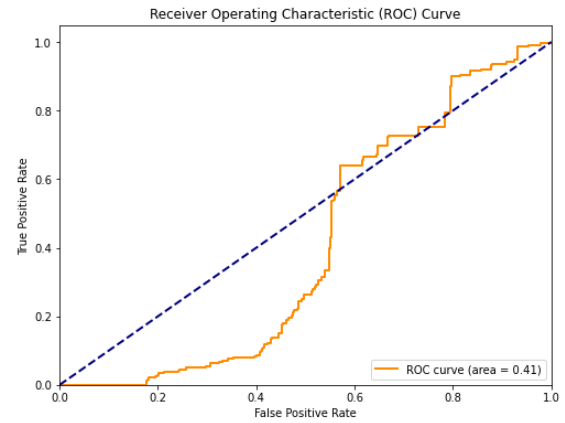


(b) ROC Curve on X (with SMOTE)

Figure A.5: Breast Cancer ROC Curve: IF Evaluation metrics with SMOTE



(a) ROC Curve on X_test (LOF) - breast cancer



(b) ROC Curve on X (LOF) - breast cancer

Figure A.6: ROC Curve: Evaluation metrics (LOF) - breast cancer

Class	Precision	Recall	F1-Score	Support
0	0.54	0.92	0.68	342
1	0.66	0.16	0.26	320
Accuracy			0.55	662
Macro Avg	0.60	0.54	0.47	662
Weighted Avg	0.60	0.55	0.48	662

Table A.21: Isolation Forest (with SMOTE) Cardiotocography Dataset Classification Evaluation Report

Class	Precision	Recall	F1-Score	Support
0	0.92	0.93	0.92	1655
1	0.23	0.19	0.21	176
Accuracy			0.86	1831
Macro Avg	0.57	0.56	0.57	1831
Weighted Avg	0.85	0.86	0.86	1831

Table A.22: Isolation Forest (with SMOTE) Cardiotocography Dataset Classification Evaluation Report

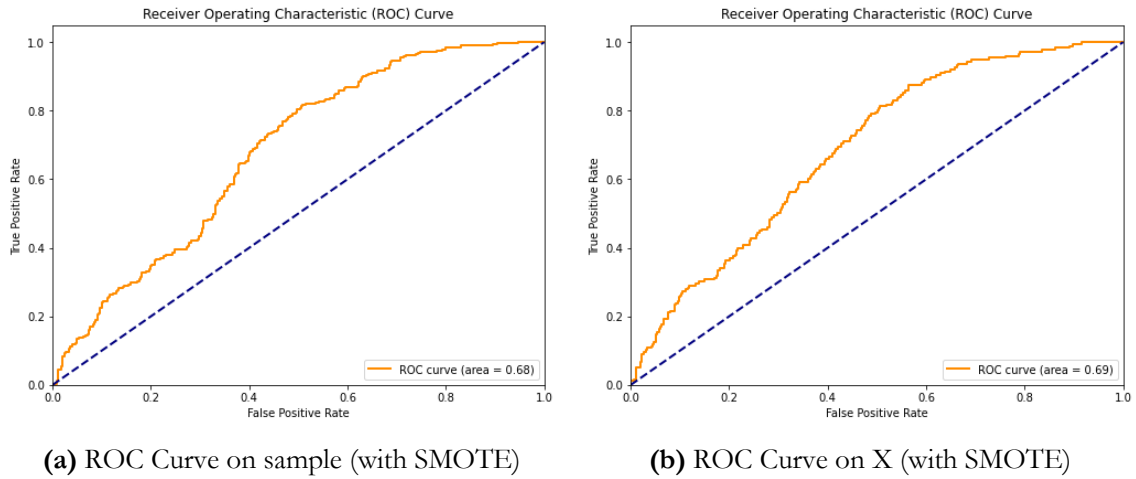


Figure A.7: Cardiotocography ROC Curve: IF Evaluation metrics with SMOTE

A.6 Outlier Explanations

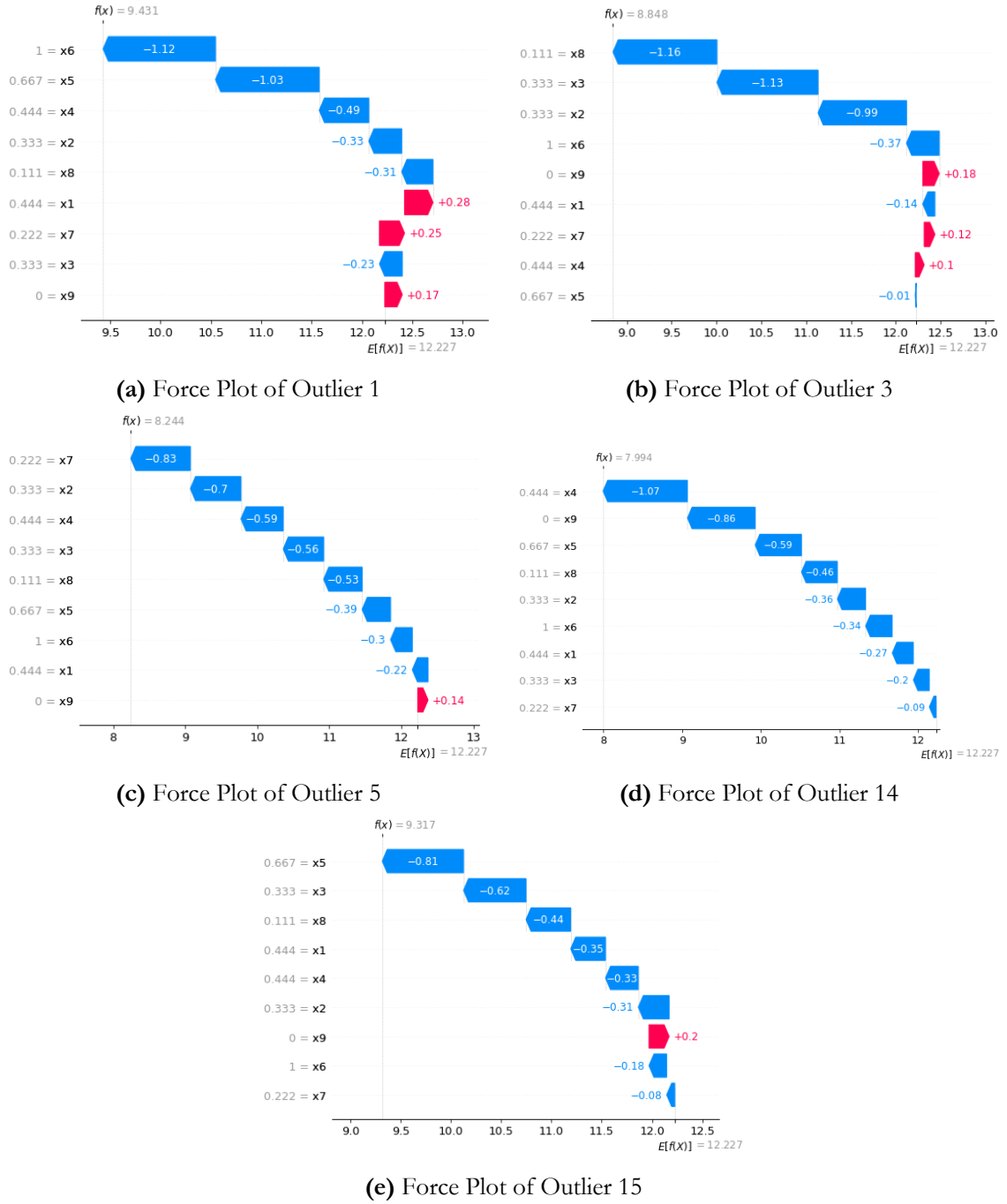
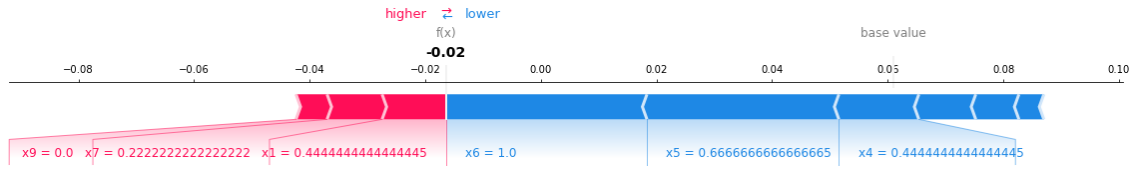
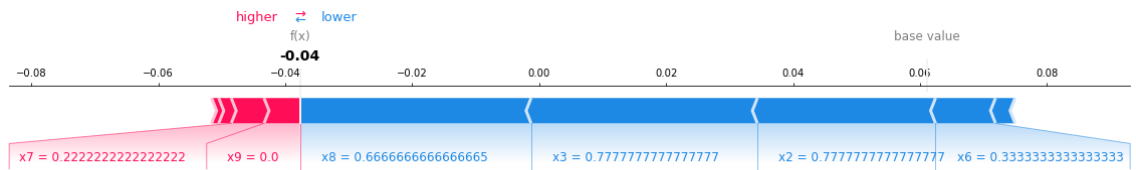


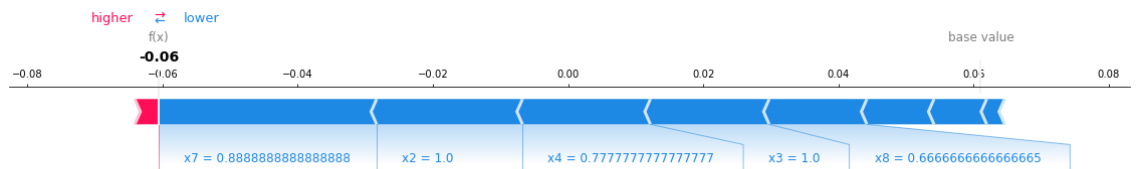
Figure A.8: Force Plots of the first five Outliers - Breast Cancer Dataset - IF and SHAP(tree)



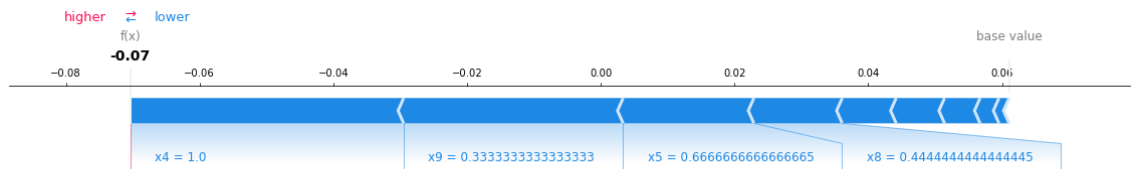
(a) Force Plot of Outlier 1



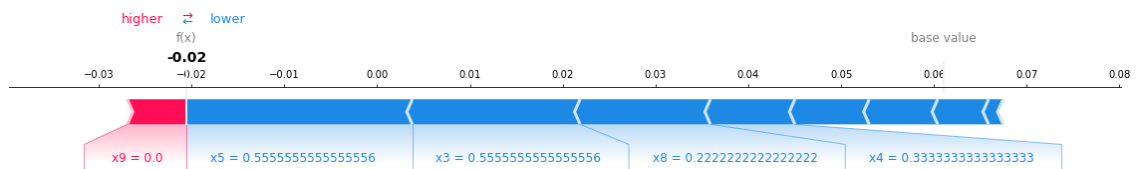
(b) Force Plot of Outlier 3



(c) Force Plot of Outlier 5

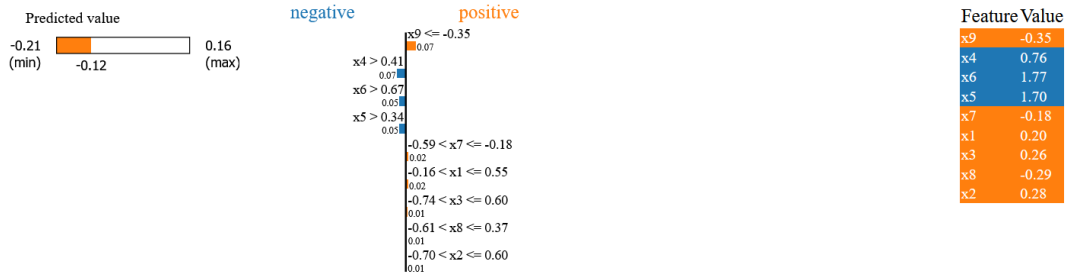


(d) Force Plot of Outlier 14

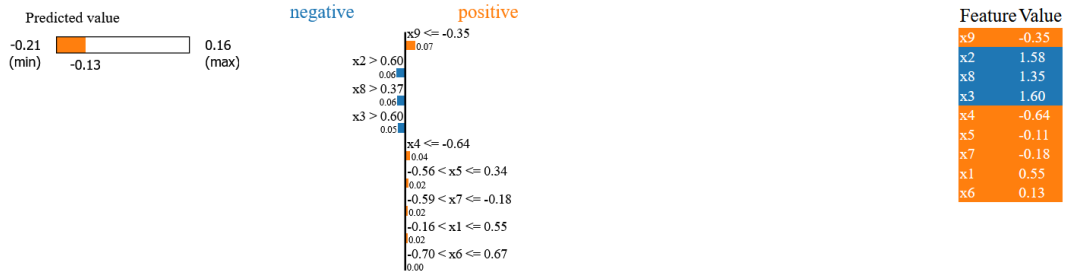


(e) Force Plot of Outlier 15

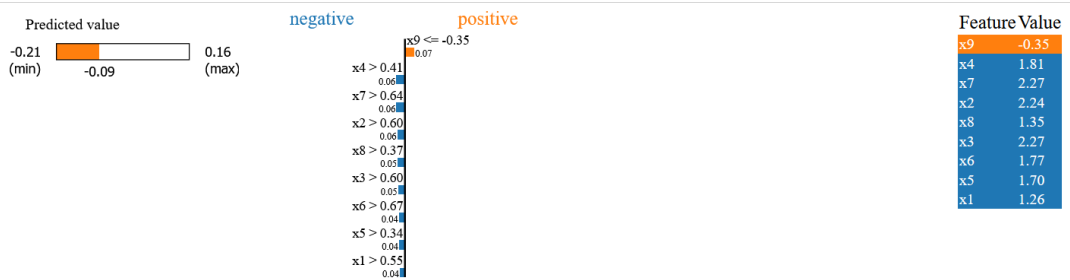
Figure A.9: Force Plots of the first five Outliers (Kernel) - Breast Cancer



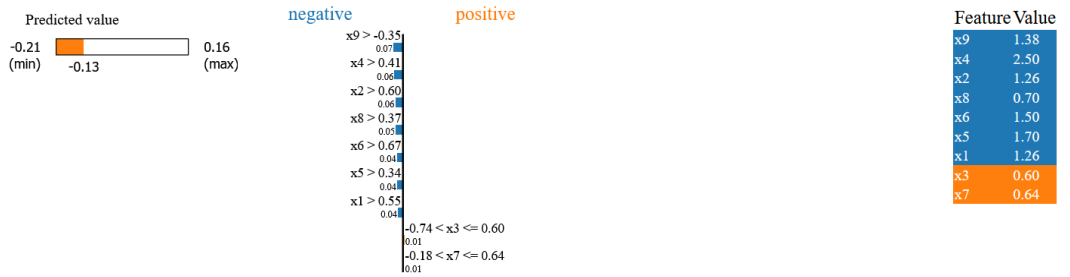
(a) LIME Explanation of Outlier 1



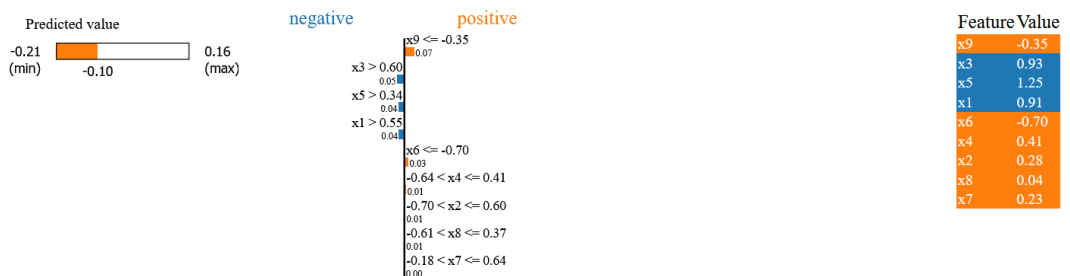
(b) LIME Explanation of Outlier 3



(c) LIME Explanation of Outlier 5



(d) LIME Explanation of Outlier 14



(e) LIME Explanation of Outlier 15

Figure A.10: LIME Explanations for the first 5 outliers - Breast cancer dataset

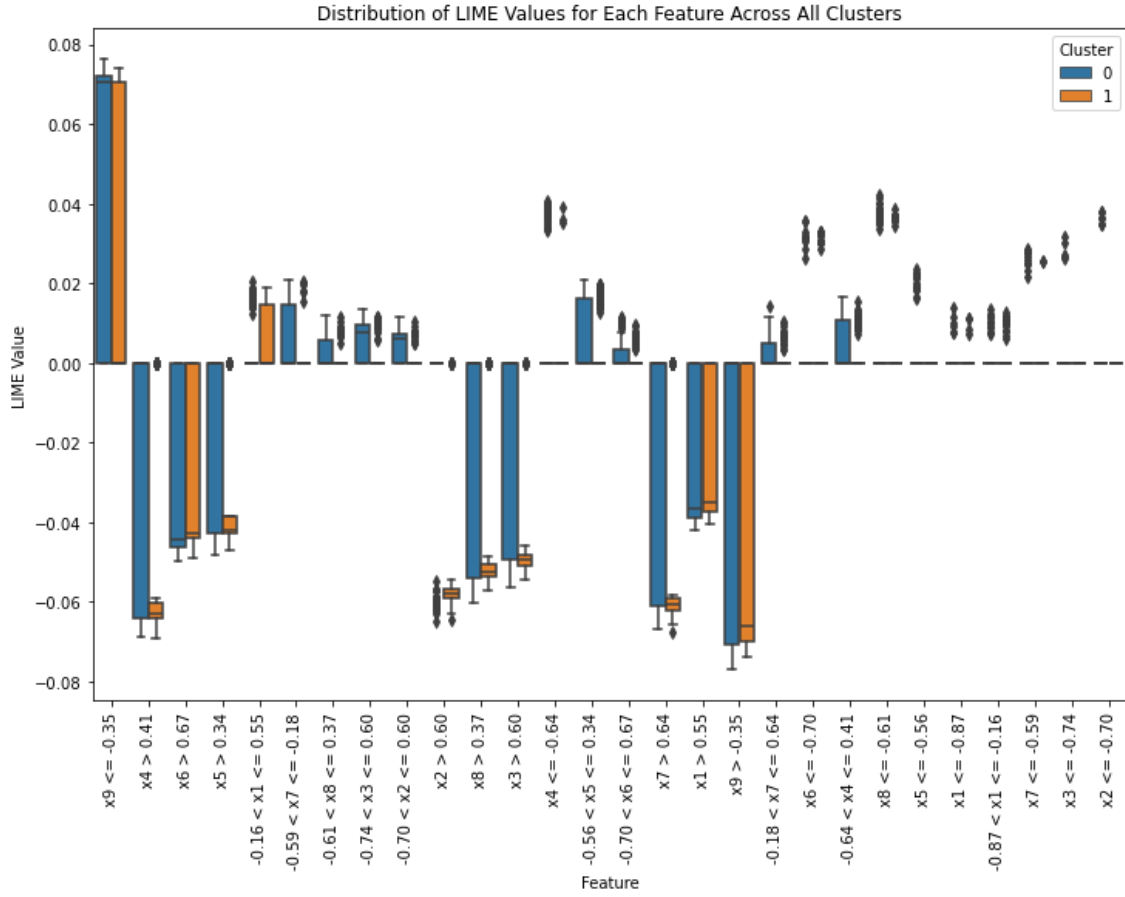


Figure A.11: LIME Values by Feature condition and Cluster - Breast cancer dataset

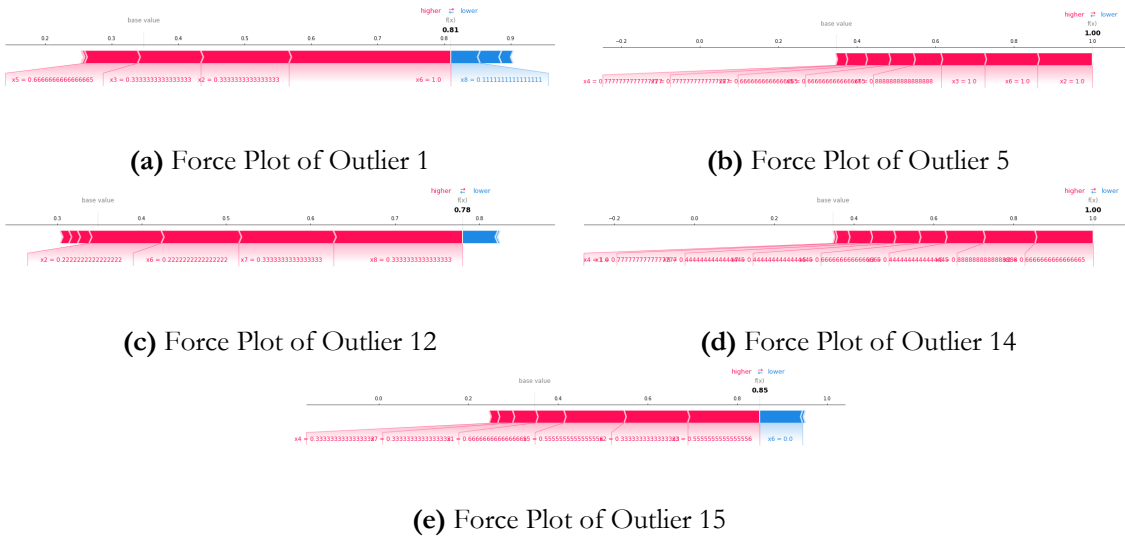


Figure A.12: Force Plots of the first five Outliers - Breast Cancer Dataset - RF and Tree Explainer

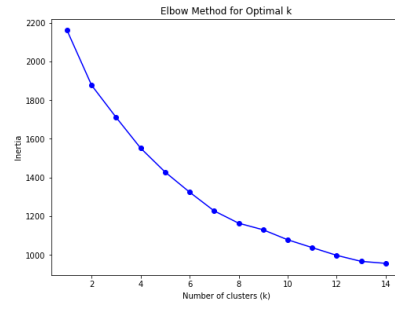


Figure A.13: Elbow Method for Optimal k - Kernel explainer

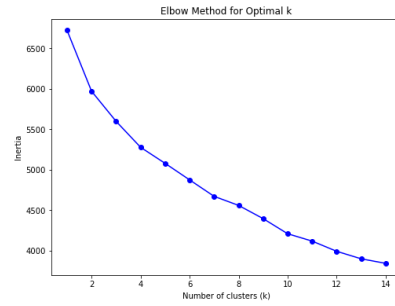


Figure A.14: Elbow Method for Optimal k - LIME

Feature	Outlier 1	Outlier 2
seismic	-0.05278	-0.225935
seismoacoustic	0.090569	-1.569.577
shift	0.155961	0.141411
genergy	-0.515153	-1.109.627
gpuls	-0.4592	-0.799201
gdenergy	-0.195599	0.076887
gdpuls	-0.272156	0.108036
ghazard	0.097652	0.084462
nbumps	-1.394.162	-0.19404
nbumps2	-1.080.016	-0.188335
nbumps3	-0.512093	0.050426
nbumps4	-1.433.250	-1.326.941
nbumps5	0.007213	0.008747
nbumps6	0	0
nbumps7	0	0
nbumps89	0	0
energy	-0.398228	-0.573321
maxenergy	-0.441314	-0.586059

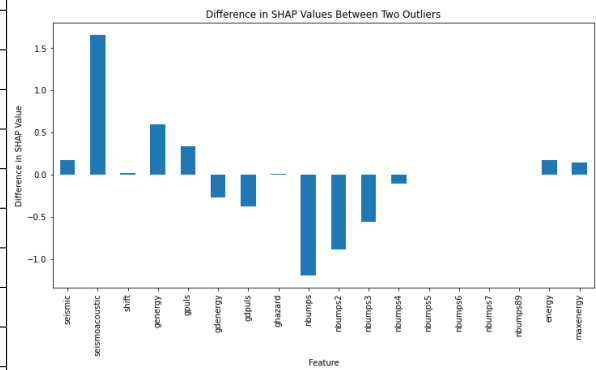
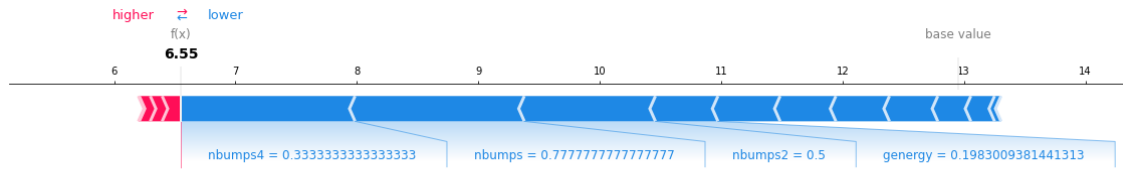
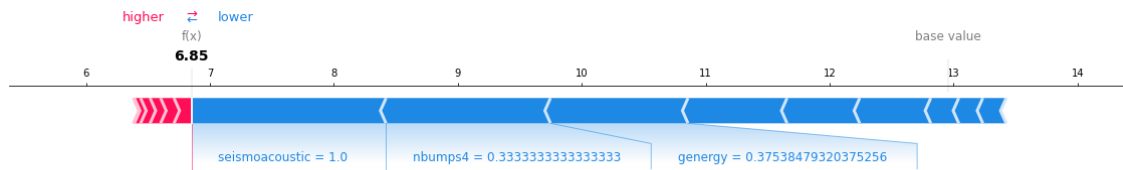


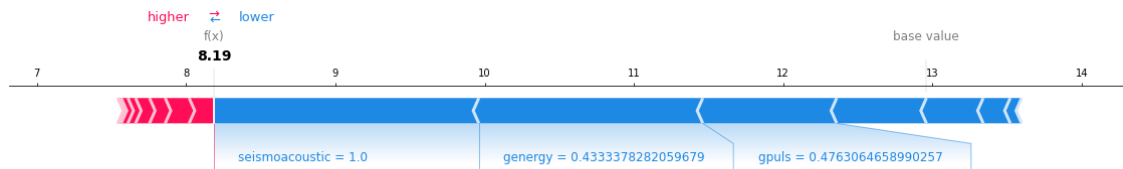
Figure A.15: Table and Figure: Difference in SHAP Values Between 2 Outliers - IF and Tree Explainer - Seismic dataset



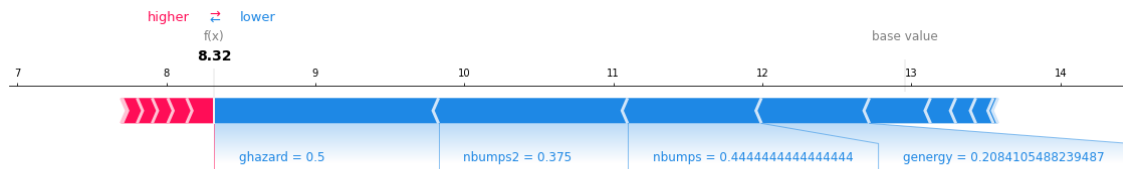
(a) Force Plot of Outlier 36



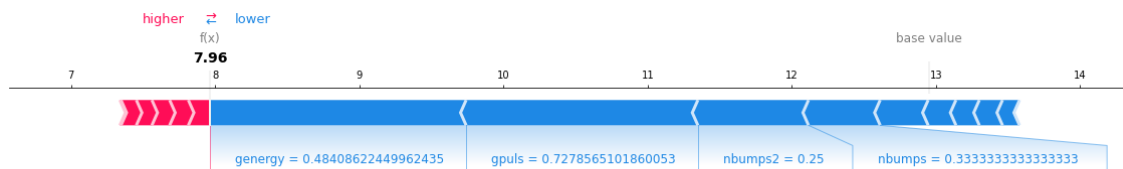
(b) Force Plot of Outlier 161



(c) Force Plot of Outlier 164

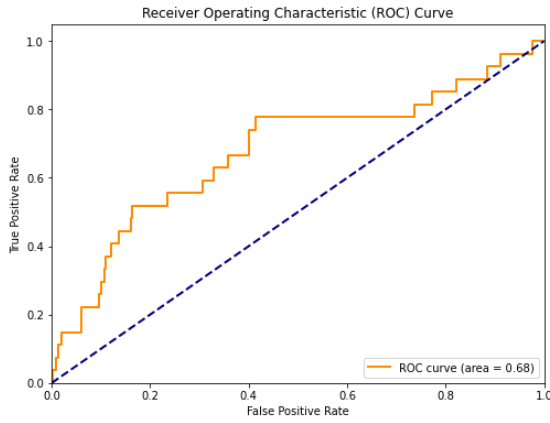


(d) Force Plot of Outlier 190

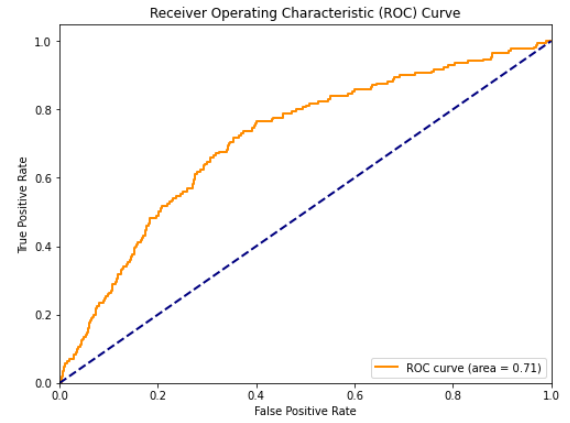


(e) Force Plot of Outlier 197

Figure A.16: Force Plots of the first five Outliers - Seismic (IF and Tree Explainer)

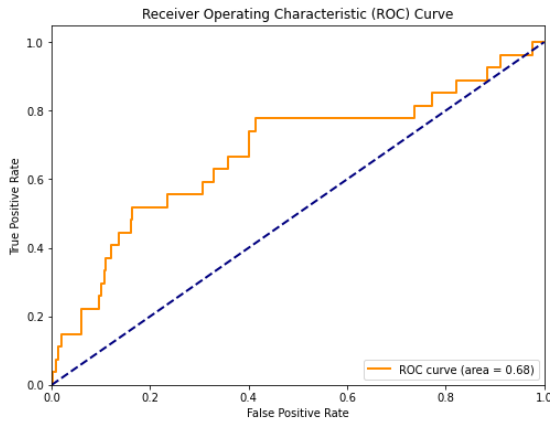


(a) ROC Curve on test dataset

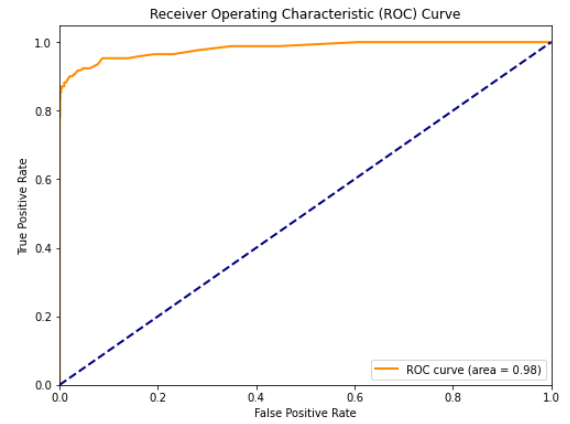


(b) ROC Curve on the whole dataset

Figure A.17: ROC Curve: Evaluation metrics (Isolation Forest)



(a) ROC Curve on test dataset



(b) ROC Curve on the whole dataset

Figure A.18: ROC Curve: Evaluation metrics (Random Forest with SMOTE)

Class	Precision	Recall	F1-Score	Support
0	1.00	1.00	1.00	1655
1	0.99	0.98	0.99	176
Accuracy			1.00	1831
Macro Avg	1.00	0.99	0.99	1831
Weighted Avg	1.00	1.00	1.00	1831

Table A.23: Cardio Dataset Classification Evaluation Report (Random Forest)

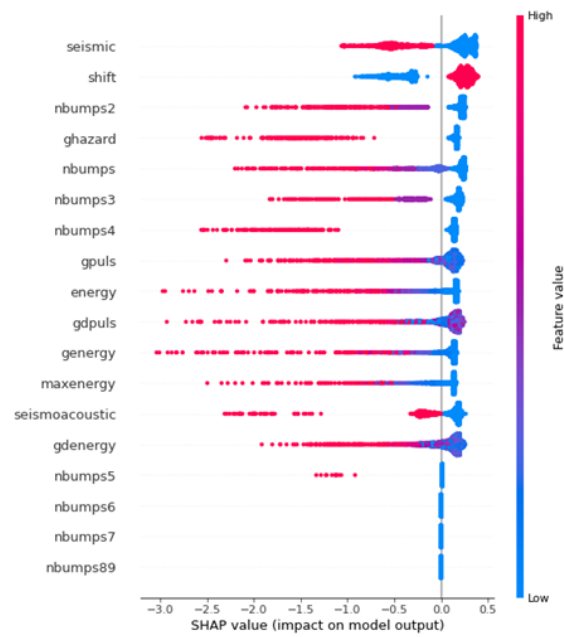


Figure A.19: SHAP Summary Plot - Tree Explainer based on Isolation Forest

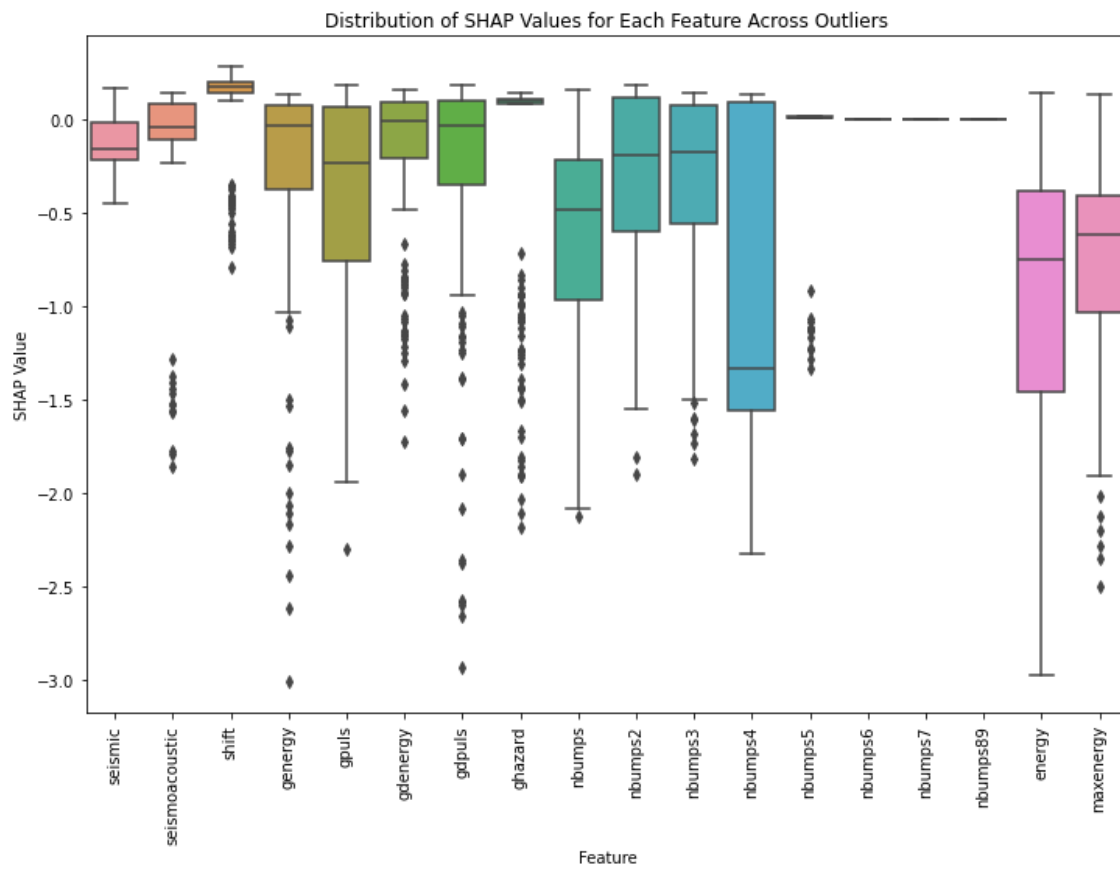


Figure A.20: Distribution of SHAP Values for each feature across Outliers - Tree Explainer based on Isolation Forest

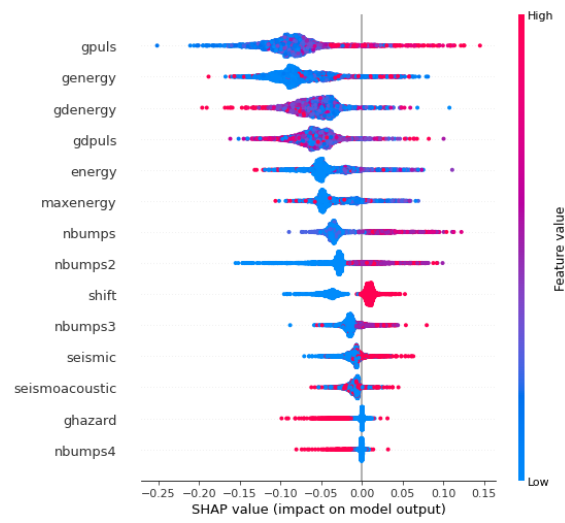


Figure A.21: SHAP (Tree) Summary Plot (Random Forest with Feature Selection)

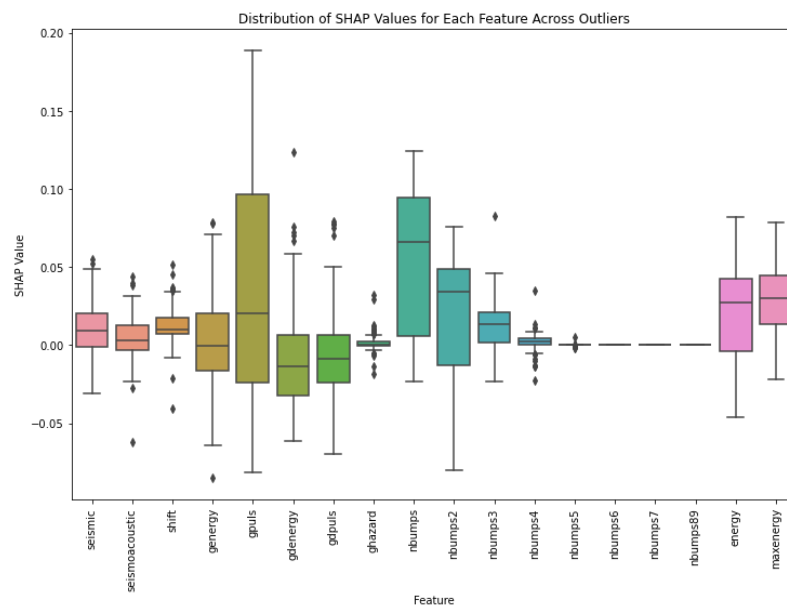
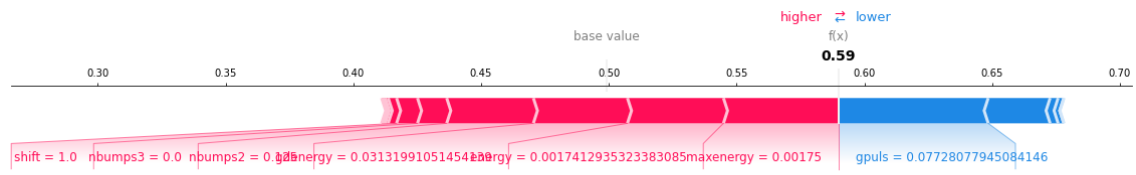
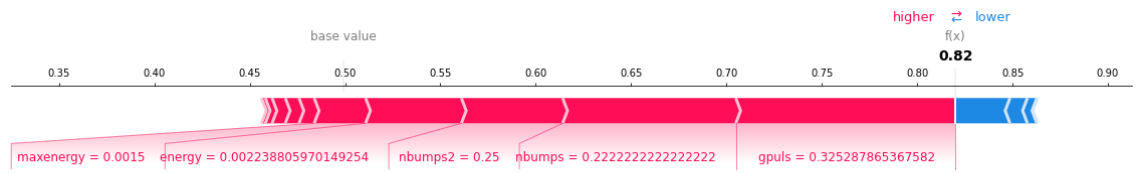


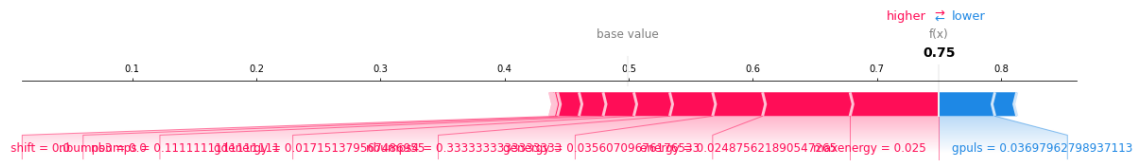
Figure A.22: Distribution of SHAP Values (Tree and RF with Feature selection) for each feature Across Outliers



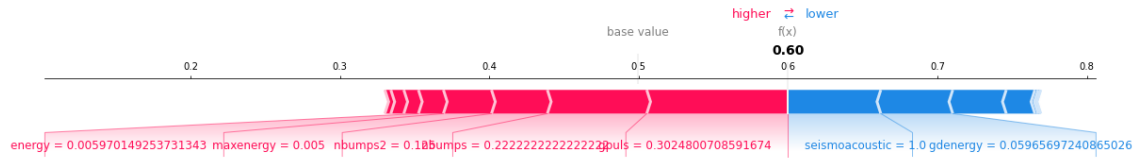
(a) Force Plot of Outlier 35



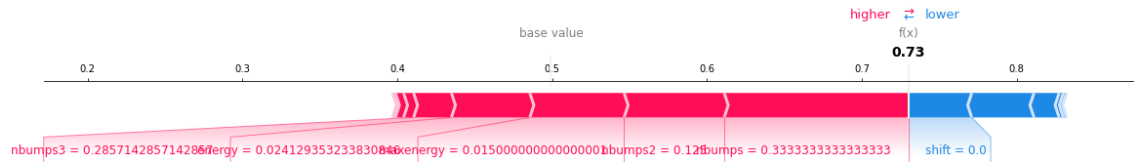
(b) Force Plot of Outlier 159



(c) Force Plot of Outlier 160



(d) Force Plot of Outlier 162



(e) Force Plot of Outlier 175

Figure A.23: Force Plots of the first five Outliers - Seismic (RF with Feature Selection and Tree Explainer)

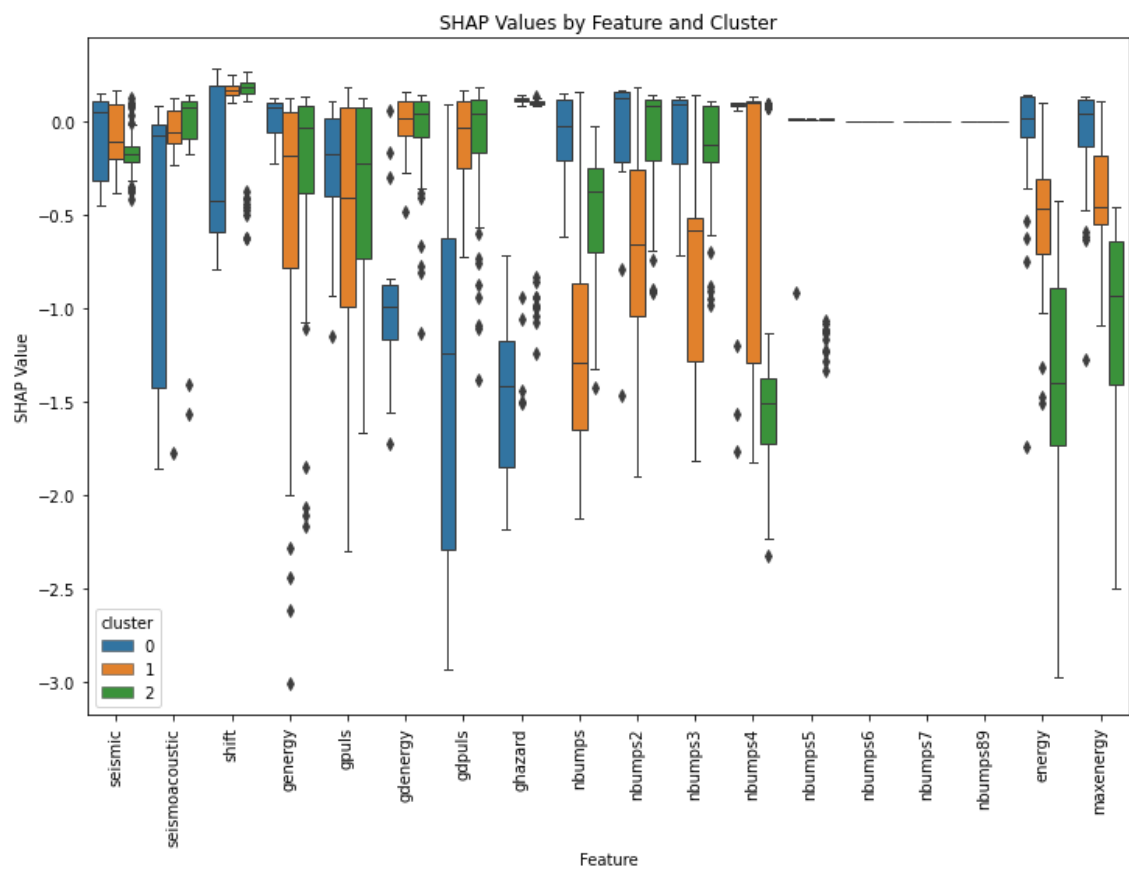


Figure A.24: Distribution of SHAP Values by feature and cluster (IF Tree SHAP)

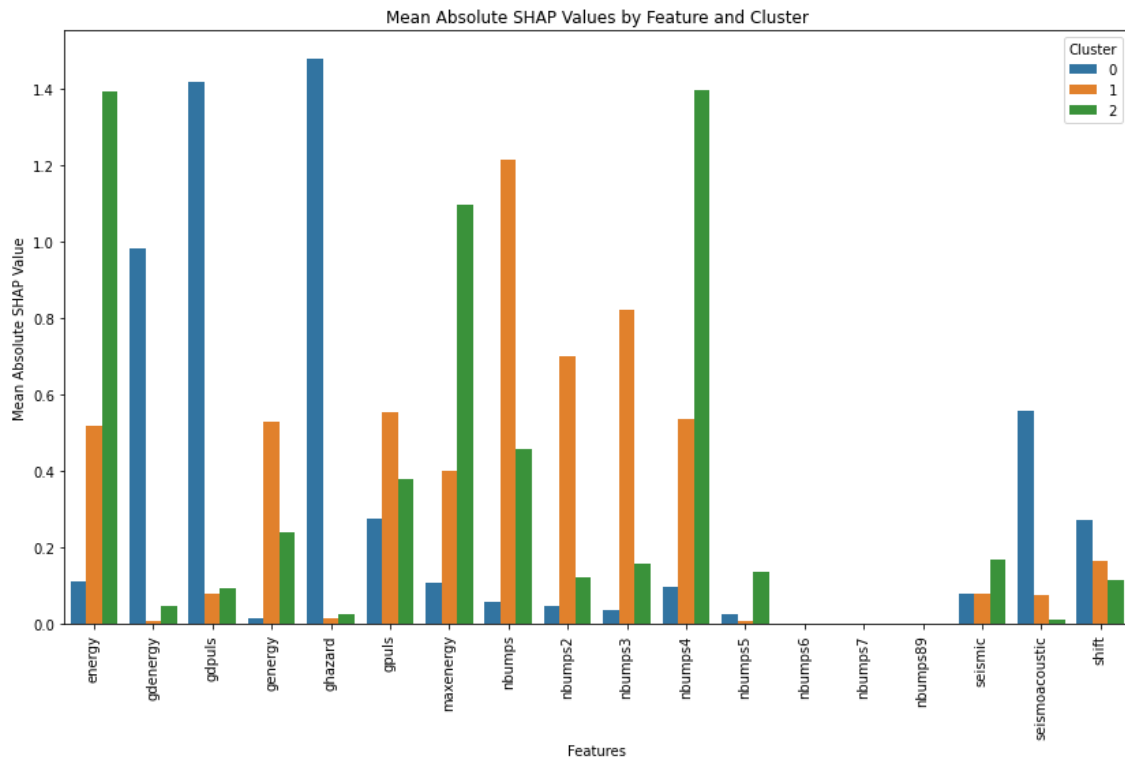


Figure A.25: Mean absolute SHAP Values by feature and cluster (IF Tree SHAP)

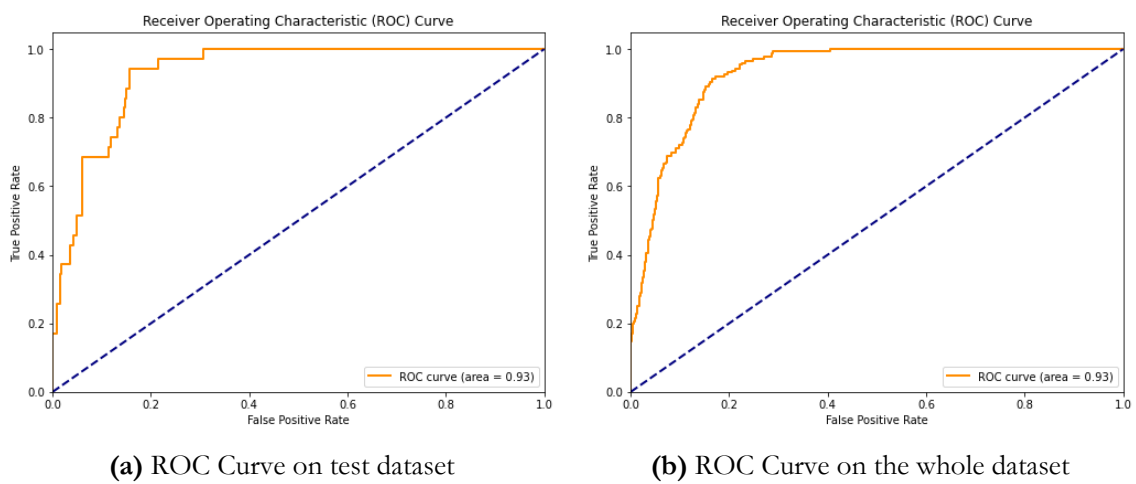


Figure A.26: ROC Curve: Evaluation metrics (IF) Cardio dataset

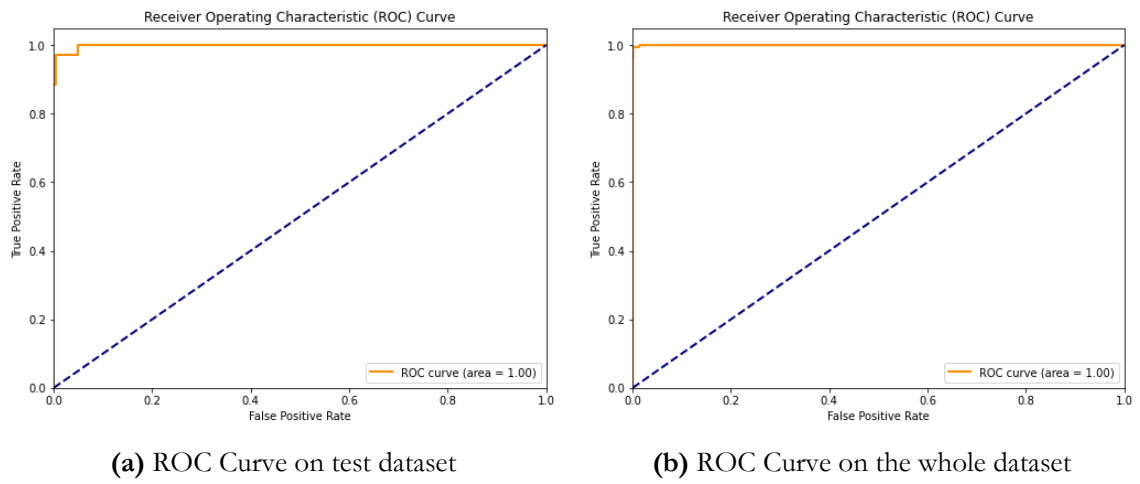


Figure A.27: ROC Curve: Evaluation metrics (RF) Cardio dataset

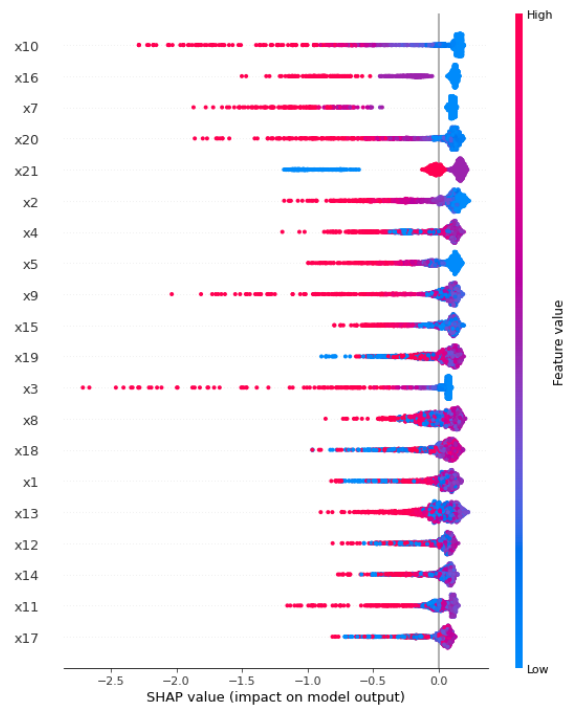


Figure A.28: SHAP Summary Plot (Isolation Forest - cardio)

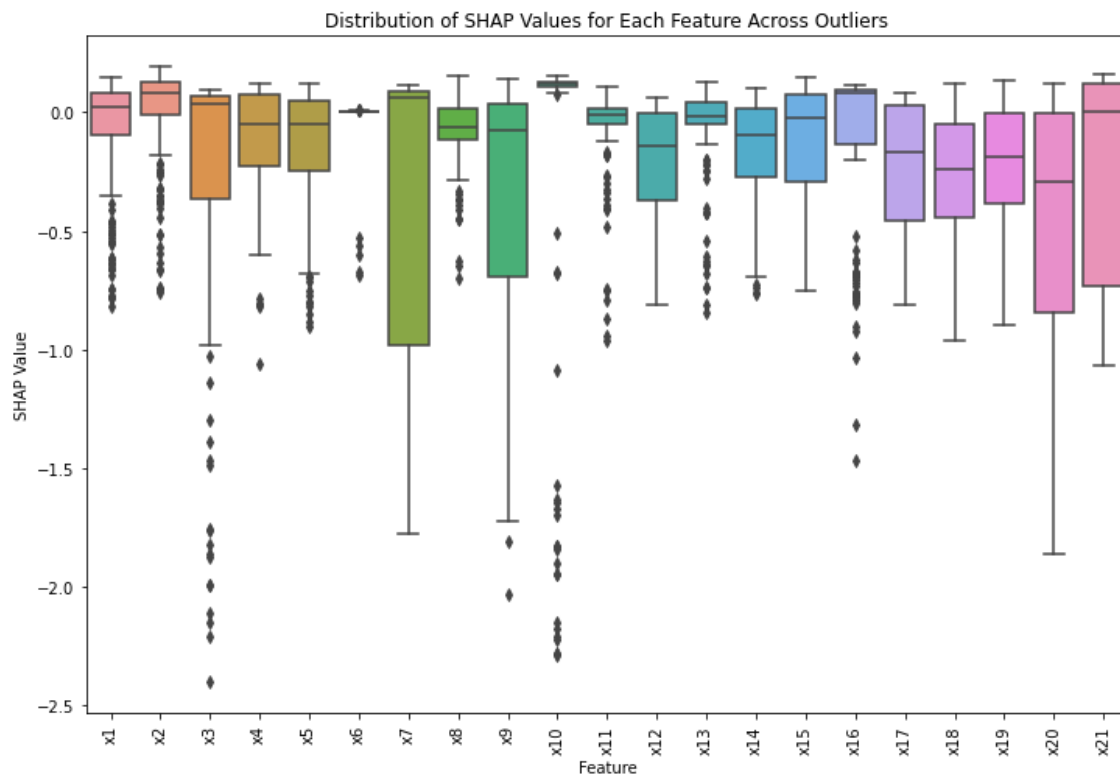


Figure A.29: Distribution of SHAP Values for each feature Across Outliers (IF Tree - cardio)

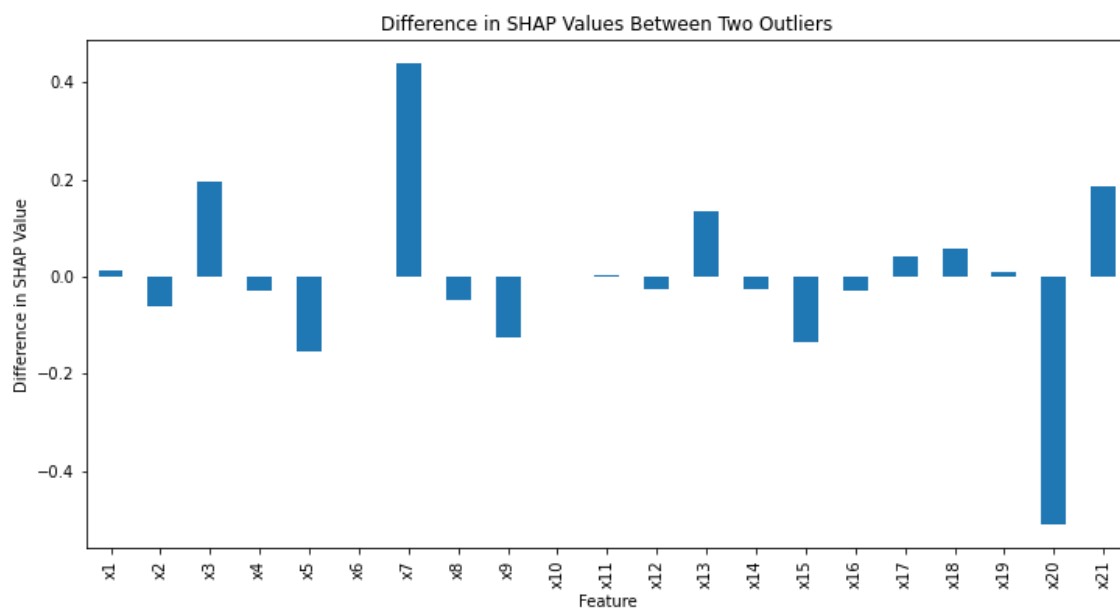
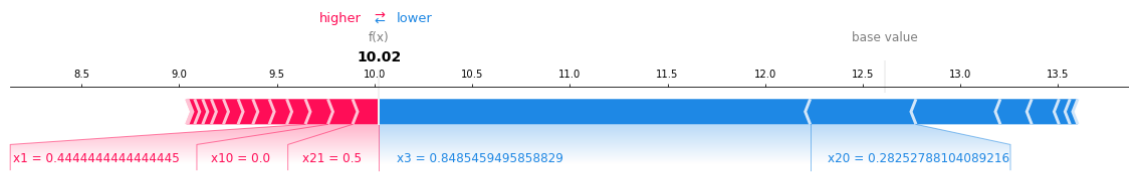
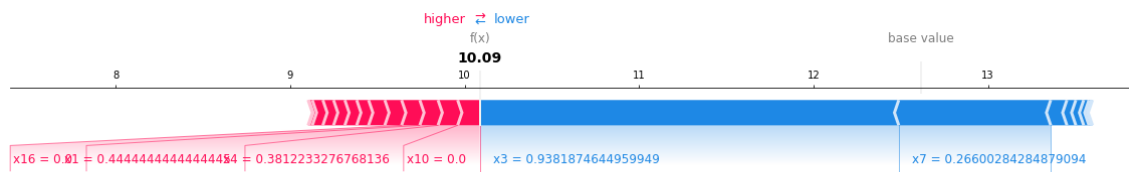


Figure A.30: Difference in SHAP (Tree) Values Between 2 Outliers (IF) - cardio



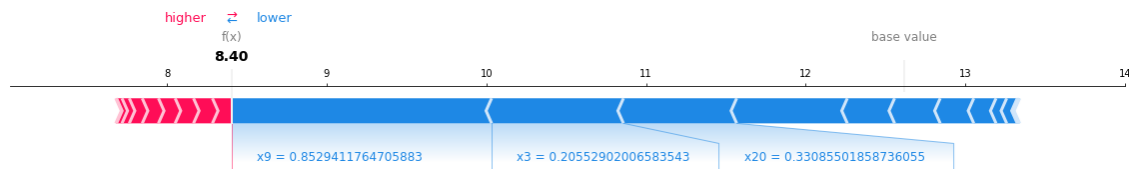
(a) Force Plot of Outlier 6



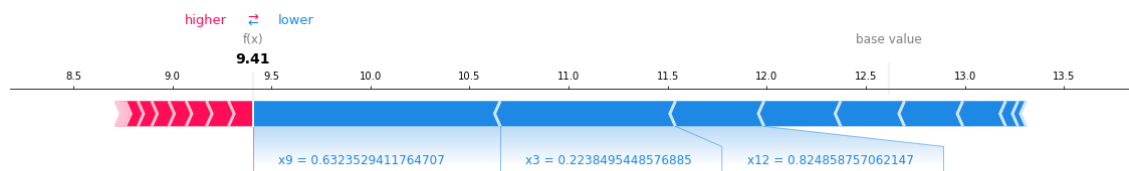
(b) Force Plot of Outlier 9



(c) Force Plot of Outlier 11

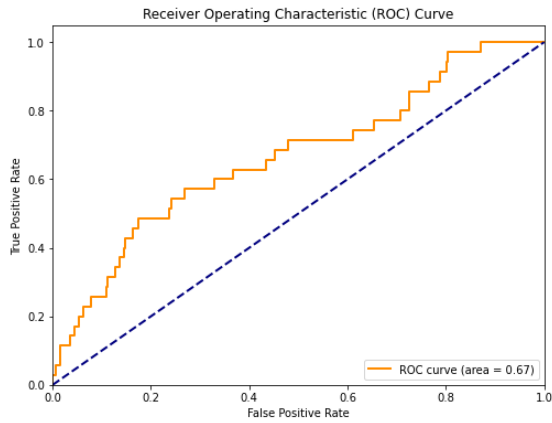


(d) Force Plot of Outlier 12

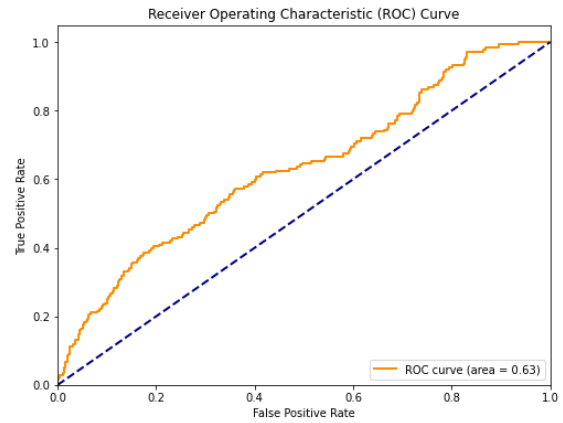


(e) Force Plot of Outlier 13

Figure A.31: Force Plots of the first five Outliers - Cardio (IF and SHAP Tree)



(a) ROC Curve on X_test (LOF) - cardio



(b) ROC Curve on X (LOF) - cardio

Figure A.32: ROC Curve: Evaluation metrics (LOF) - cardio

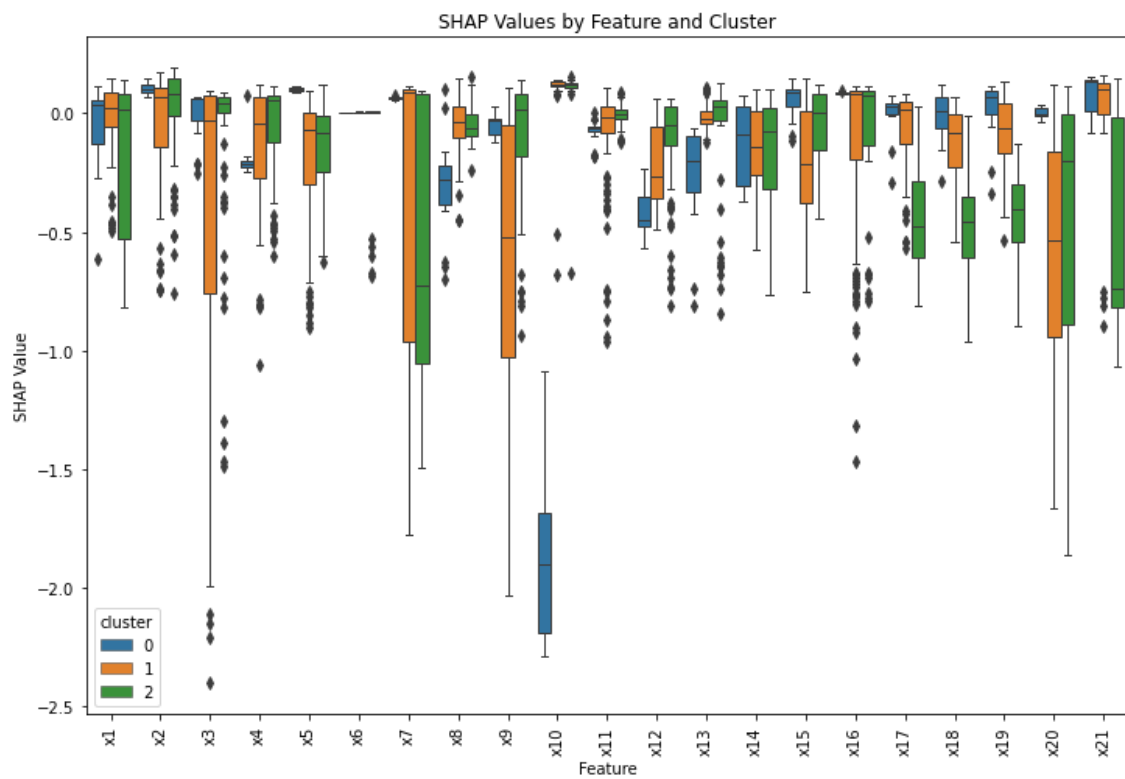


Figure A.33: Distribution of SHAP Values by feature and cluster - IF Cardio dataset

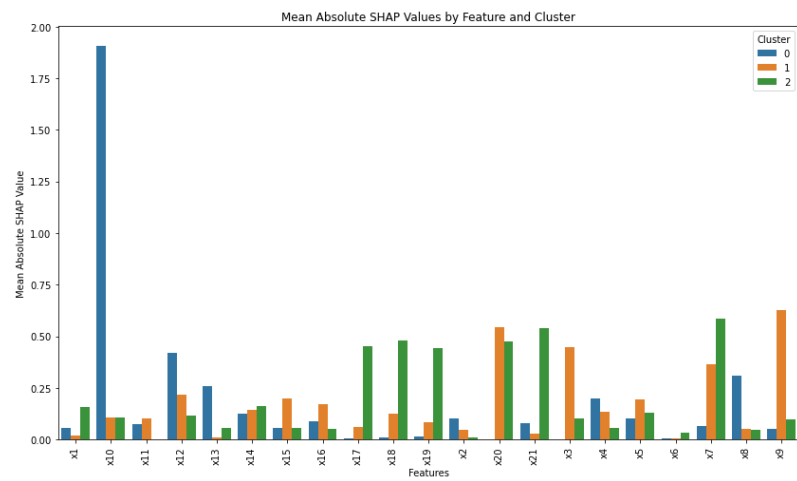


Figure A.34: Mean absolute SHAP Values by feature and cluster - IF Cardio dataset