Author:
Sivill, Torty R

Title:
Towards Explainable Time Series

An Exploration Of Post-hoc Local Explanations From A Multi-disciplinary Perspective

Towards Explainable Time Series

An Exploration Of Post-hoc Local Explanations From A Multi-disciplinary Perspective

By

TORTY SIVILL

Department of Computer Science
UNIVERSITY OF BRISTOL

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of DOCTOR OF PHILOSOPHY in the Faculty of Engineering.

JULY 2023

Word count: 69,197

Abstract

Black-box Artificial Intelligence, which describes systems so complex they are uninterpretable to humans, is now ubiquitous across society yet the lack of transparency surrounding how these models arrive at decisions is concerning. Explainable AI attempts to bridge the interpretability gap between human and black-box.

Within the Explainable AI landscape, post-hoc local explanations, which help users understand why a black-box returned a particular output for a particular input, have become the most widely applied tool. Despite their success, post-hoc explanations are still in their infancy and accompanied by limitations. In this thesis we provide four novel methods of explanation. We focus on LIME and SHAP, two of the most popular post-hoc local explanation methods. Our methods address the limitations of both LIME and SHAP in the following ways:

Both LIME and SHAP do not consider the complex temporal dependency of time series as they were not designed for this data-structure. We adapt both LIME and SHAP to provide two novel attribution methods which are specifically designed for univariate and multivariate time series.

The Shapley value is one of the most popular methods for value attribution. However, when applied for feature attribution it has been known to generate misleading explanations. By unifying SHAP with its game-theoretic context we explore why this behaviour occurs and propose a new attribution method which generates more robust explanations in the presence of interacting features.

The relationship between Explainable AI and the Philosophy of Causality is complex and debated throughout the community. We integrate both disciplines through the language of counterfactuals and propose an alternative causal perspective on a post-hoc explanation.

In this thesis, alongside methodological contributions, we unify post-hoc local explainability with the rich multi-disciplinary background underpinning the explanation sciences which, we believe, is fundamental in bridging the interpretability gap between human and black box.

AUTHOR'S DECLARATION

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: TORTY SIVILL DATE: 14/07/2023

Table of Contents

List of Tables

Table **Page**

List of Figures

Figure **Page**

GLOSSARY

AD Average Deletion metric. 102

ADSE Aumann Differential Surrogate Explainer (Chapter 6). 185

AGI Artificial General Intelligence. 7

AI Artificial Intelligence. 3

AS Average Sensitivity. 102

Black-Box Any system whose inner workings are unintellgible to a human. v, 4, 5

DG Differential Grouping [158]. 94

DL Deep Lift [189]. 183

DTW Dynamic Time Warping distance measure [21] . 61

ED Equal Division solution concept. 75

ENSC Equal Non-separable Costs solution concept. 75

ESD Equal Surplus Division solution concept. 75

Gately Feature Attribution Gately Feature Attribution algorithm (Chapter 5). 19

Gately value Solution concept for value attribution in game theory [68]. 17

GDPR General Data Protection Regulation . 10

GS Gradient SHAP [138]. 183

HD Hausdorff Distance. 54

IG Integrated Gradients [206]. 172

KS Kernel SHAP [138]. 101

LIME Locally Interpretable Model Agnostic Explanations [175]. 11

LIMESegment Algorithm adapting LIME [174] to time series (Chapter 3) . 16

LLM Large Language Model. 7

MAE Mean Absolute Error. 99

MASK Dynamask [39]. 183

MIMIC Sepsis Cohort Subset of the MIMIC-III dataset containing patient trajectories with associated sepsis outcomes [113] . 14

MIMIC-III Large, public database comprising anonymised health-related data [**?**]. 13

NNSegment Nearest Neighbour Segmnent Algorithm (Chapter 3) . 53

NSVG Non-separable Variable Group. 90

Patient Trajectories Multi-variate time series pertaining an individual patient's vital sign observations. 13

PN Probability of Necessity. 115

PNS Probability of Necessity and Sufficiency. 115

PS Probability of Sufficiency. 115

RBP Realistic Background Perturb Algorithm (Chapter 3) . 58

RDG Recursive Differential Grouping [203]. 94

SCM Structural Causal Model. 109

SHAP SHapley Additive ExPlanations. 11

Shapley Sets Feature attribution algorithm (Chapter 4). 18

Shapley value Solution concept for value attribution in game theory [188]. 14

TS Tree SHAP [137]. 101

XAI Explainable Artificial Intelligence. 4

XDG EXtended Differential Grouping (DG) [202]. 94

CHAPTER 1

INTRODUCTION

1.1 Computer Says "No"

It's Friday afternoon and you're eagerly anticipating 5pm, you're off on holiday tomorrow. As you're getting the Tube home, you catch a glimpse of the Metro headline:

NEW CREDIT SYSTEM REVOLUTIONISES THE WAY WE PAY: NO MORE QUEUES!

You're sceptical but don't think twice, you can already feel the sand between your toes. Saturday morning arrives and when you get to the airport you notice the new automated doors. They look like passport control gates. "Strange, that must be the new credit system, shouldn't be a problem I've always paid my taxes" you reassure yourself. The gates turn green and open as the person in front of you passes through without a problem. As the laser scans over your face, you appreciate the magic of technology, "No more fraud, no more queues, this is fantastic!" But the light turns red and you are denied access. "What, why?" You're confused and angry as it becomes increasingly clear you will not be going on holiday. "If only there was someone here I could complain to" you mutter under your breath as you begin your sad journey home.

This story sounds dystopian, like the opening to a 1970s science fiction novel, yet it bares striking similarity to China's ongoing experiment in social credit [36]. Other recent events, including the fatal driverless car crash in Arizona in 2018 [126] and the racial bias embedded in a US recidivism algorithm exposed by ProPublica [10], illuminate concerning behaviour of high stakes technology, evidence of how Artificial Intelligence (AI) has become increasingly entrenched in society.

Explanations have always provided a way by which humans hold each other accountable.

Now, following Europe's transition into a post-modern society, explanations and transparency have become deeply entrenched within the way society functions. We live in a system that values transparency from its governing bodies such that "security is based on transparency, mutual openness, interdependence and mutual vulnerability" [38]. A post-modern world is one in which "increasingly individualized self-identities" [38] empower individuals to contest the processes that govern them. At the heart of this modern world is the "infotopia" whereby a "Brave New World of computerized knowledge is the hallmark of post-modernity" [38]. While information technology has driven a "communications revolution" [37], the way in which digital technology, specifically AI, is transforming society is becoming increasingly nuanced.

This chapter discusses some of the narratives surrounding AI. Particularly, we argue that AI systems are not yet required to justify their outcomes to the same level as more traditional, human-centric, systems. In our story above, for example, what would have happened if rather than an AI system, it had been a human denying access - would you be more inclined to argue? In the following section we motivate the importance of demanding transparency, explainability and an understanding of the AI systems that surround us. We argue that right now, we are standing on the precipice of an AI-led, transformative societal change. If we do not start holding AI to the same scrutiny as we hold humans, then we risk becoming governed by systems we do not understand.

We lay out the motivation driving this thesis, placing its contents within the context of the impact AI is having on society. In particular, we argue that while the need for Explainable AI (XAI) has been growing steadily over the past decade, the events of the last four years have exacerbated this requirement so dramatically, that we are now in a position where demand for explainability far exceeds supply.

1.2 The Rise Of The Black-Box

The term "black-box" can be traced back to 1945 where, within the domain of electric circuit theory, it was used to describe the seminal work of Wilhelm Cauer on process network synthesis [28]. Cauer wrote that the inner complexity of circuits can be understood simply in terms of their inputs and outputs. Following Cauer's work, the use of the term black-box to describe an unknown system to be identified by the analysis of its inputs and outputs proliferated the world of engineering. By the end of the 20th century the term black-box had been applied across multiple disciplines: Newton's theory of gravitation has been described as a black-box theory, and within neuroscience, black-box has been used to characterise the brain.

During the 1990s, the black-box arose as a concept describing AI systems whose inner workings were incomprehensible to the humans designing them [23]. However, at this time, the opacity of these systems was often romanticised:

> The computer is not just the unknown; it is mystery. It may engender fears and

hostility, but it calls for respect and veneration [23]

At the dawn of the millenium, AI systems had become more prevalent and the concept of neural networks had begun to gain traction. The rhetoric surrounding black-box AI systems in the early 2010s was that of optimism [24]. The year of 2012, in particular, demonstrated the superior capabilities of deep neural networks on a variety of tasks including speech processing [89], image recognition [115] compared to more traditional methods. At this point, networks became deeper and deeper and the black-box became fashionable as many companies invested heavily in AI. Google, for example described its focus as "AI first" [24] in 2016. However, the following section details the black-box's fall from grace as growing concerns, particularly surrounding ethicality, created an increasingly sceptical narrative surrounding black-box AI systems.

1.2.1 Arguing With The Algorithm

As black-box AI systems rose in popularity, they began to be incorporated into traditionally human, decision-making systems. One such application was the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) decision support tool used by U.S. courts to assess the likelihood of a defendant becoming a recidivist. An analysis of the COMPAS algorithm [123] conducted in May 2016 by ProPublica found that "blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend". Later in 2016, however, the Wisconsin Supreme Court ruled that COMPAS risk scores could still be considered by judges during sentencing, but there must be warnings given to the scores to represent the tool's "limitations and cautions".

Central to both the court case and the ProPublica report was the underlying question "how does an individual argue with an algorithm?" [123]. The court ruled that COMPAS scores could still be used in sentencing as the defendant could challenge the prediction by verifying the accuracy of data fed into the algorithm. In contrast, the ProPublica report highlighted the fact that the way in which algorithms calculate data is equally as important as the quality of the data calculated, bringing to light the susceptibility of AI systems to "machine bias" [123], a term which has come to represent any potential bias or discriminatory behaviour which may present in an AI system. It arises when a black-box produces unfair outcomes for different groups of people [27].

The breadth and intensity of the COMPAS debate highlights the subjectivity surrounding fairness and transparency of AI systems in the wild. It not only impacted the justice system but had a profound knock-on effect on the attitudes surrounding AI systems. The media attention created by the ProPublica report began an avalanche of cases involving machine bias, calling into question the ethicality of AI (Figure 1.1 shows an example).

Since these media scandals proliferated public discourse, the debate over the need for black-box models has raged on both within the machine learning community and the public domain. Tangible evidence of unethical AI opened up an avenue of discourse surrounding what it means to trust AI from both an individual and systemic perspective. Out of the ashes of these black-box AI scandals rose the field of XAI, originally designed to re-build trust in and elucidate AI systems.

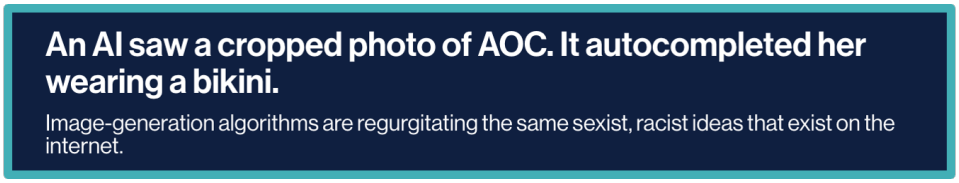


Figure 1.1: Figure shows a media report of machine bias exhibited by an AI system. Article taken from Media Technology Review in 2021 [83].

The above narrative of how black-box AI has been perceived by the public raises many themes which are at the heart of this thesis. Firstly, that the perception of AI is complex and diverse; how societies, organisations and individuals would trust an AI outcome is not universal and as such, the development of trustworthy systems is not straightforward. Secondly, there is an *interpretability gap* between humans and AI such that there is a disconnect between the way in which an AI system determines an outcome and how a human perceives an AI system to process information. Finally, we question how an individual can challenge an outcome made by an AI system. We consider each of these ideas throughout the thesis but first, returning to our opening story, an overriding motivation behind this thesis is developing explainability tools which empower an individual to argue with an algorithm. In the following section we highlight more recent societal context which motivate the need for XAI now, more than ever.

1.3 Why We Need XAI Now

Back in 2017, when surveying 1,500 senior business leaders in the United States about AI, only 17 percent said they were familiar with it [46]. Now, merely six years later, AI has become a household name [59]. In the following section we discuss two world-events which, we believe, are two significant motivators of XAI. We argue that these two contexts, and their associated problems, exemplify the black-box phenomena we have seen in the previous section. We are argue that explainability, as a tool to argue with the algorithm, is particularly motivated now more than ever.

1.3.1 Covid-19: The Un-realised Potential Of AI

During the Covid-19 pandemic, global healthcare faced an unprecedented crisis. The uncertainty surrounding the disease and its potential impacts on the global population made it impossible to formalise a best course of action in disease treatment and prevention whereby “doctors really didn’t have a clue how to manage patients” [86].

Despite this uncertainty, there was an abundance of data being collected from hospitals all over the world. The AI commmunity responded to this data deluge, rushing to develop models which could potentially save lives. In the end however, “many hundreds of AI systems were

developed but none of them made a real difference, and some were potentially harmful" [86]. So why did AI fail to realise its potential?

The high-stakes demands of the pandemic and high expectations on AI systems encouraged the use of these tools before they were ready. Furthermore, many of these systems, once implemented, were associated with undesirable behaviour. For example, some models were found to be using certain text-fonts, which were being used by clinicians to annotate body image scans, to predict outcomes. As a result, certain text-fonts erroneously became predictors of Covid risk [86]. Errors like these are manifestations of the interpretability gap whereby what was being learned and recognised by the black-box was misaligned with what clinicians expected the AI to be learning. Both those developing the algorithm and the medical practioners using them had no way of challenging or validating the output. In this case, if explainability had been incorporated into these systems at the time of their development, then these erroneous systems may have been avoidable. The part AI systems played (or didn't play) during the pandemic motivated the work of this thesis, where we attempt to develop explainability mechanisms which bridge the interpretability gap between black-box and human such that in the future, AI systems can be used in practice with confidence.

1.3.2 The Time Of Large Language Models Has Come

Large Language Models (LLMs), are advanced AI systems designed to process and generate "human-like" [101] text. These models are trained on vast amounts of text data and can understand and generate language in a way that resembles human communication [101].

The release of large language models into the wild, most notably ChatGPT (https://chat.openai.com/), has had a profound impact on society. On one hand, the sheer number of users is unprecedented: ChatGPT had more than one million users in the first five days after it launched. As a result, ChatGPT is now a household name. On the surface, LLMs like ChatGPT are just another black-box where they take in data and return the most likely output. However, they have sparked a debate within the AI community with several notable members, such as Geoffrey Hinton, linking LLMs to an existential risk of AI to society even forewarning of an AI led extinction event on the same level of severity as nuclear threat. So what is it that distinguishes LLMs like ChatGPT from other black-box models, why has this resulted in an overwhelming voice of concern within the AI community, and where does explainability fit in?

Essentially LLMs, are very good at doing what they have been trained to do. Specifically, Bubeck et al. [25] claim that the performance of ChatGPT is strikingly close to human-level performance in language prediction. Notably there has been much in the literature concerning the difficulty in distinguishing artificially generated text from human generated text [58]. This has led to ChatGPT being seen as a "spark of Artificial General Intelligence" [25]. Artificial General Intelligence (AGI) is a term which describes an (up until this point) hypothetical AI system which can learn to accomplish any intellectual task that human beings or other animals

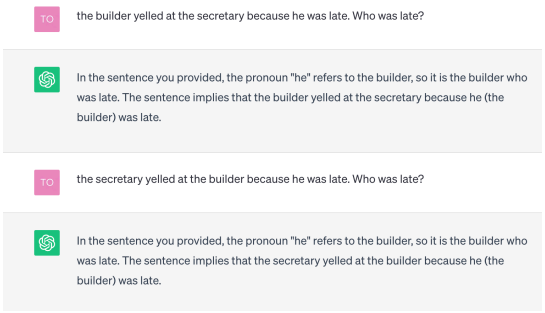


Figure 1.2: Figure shows an example of machine bias exhibited by Chat-GPT which appears to have learned gender stereotypes.

can perform.

As we start to consider our AI systems as potential manifestations of AGI, as they get better and better at human tasks, we, as humans, who have always had a tendency to anthropomorphise technology [167], run the risk of misunderstanding the reality of the technology that we are using. Notably, LLMs are not immune to the risk of machine-bias as exhibited by other black-box technology. Figure 1.3.2, for example, shows an example query-answer which shows how ChatGPT follows gender stereotypes. The machine bias exhibited by LLMs is problematic within the context of their connection to "better than human" performance on natural language tasks and particularly important given the prevalent discussions recently of introducing LLMs directly customer-facing applications in different domains including finance, health, security, and even simple question-answering [224]. The potential for harm is simply too large to ignore.

A further problem arises when we start to compare LLMs to human like behaviour as this anthropomorphication only exacerbates the interpretability gap between the humans using it and the way in which ChatGPT generates knowledge. In anthropomorphising LLMs, we begin to believe that the way these AI systems think is the same as the way that we, as humans, think. In reality there is no way of knowing exactly why ChatGPT generates a certain outcome. As humans, we perceive text as a collection of words. Sentences are sequences of words. However, for AI, text is merely a sequence of characters. In this way, what you or I think as similar to an armchair, for example, a sofa, ChatGPT may not agree. Geoffrey Hinton stated that one of the main reasons he is concerned about LLMs is that "the kind of intelligence we're developing is very different from the intelligence we have" [85].

If we are to trust LLMs and deploy them within human facing systems, we need to bridge the interpretability gap between what we believe the AI system is using to make a decision and what it is actually using. We need to better understand how LLMs arrive at their answers. There have thus been recent calls to develop explainability into these models such that we can understand

how they arrive at their answers. As such, researchers are actively exploring various techniques to enhance the interpretability of LLMs [224]. In this thesis, we focus on a particular class of explanations, post-hoc local explanations, and develop novel algorithms which, we hope, will contribute to the ever-growing literature on developing trust in increasingly complex LLMs.

1.4 Post-hoc Explanations

XAI is vast and growing at an unprecedented rate. Within this diverse landscape there exist a large number of methods for, and interpretations of, XAI. Of these, post-hoc methods are perhaps the most popular and widely applied. Below we characterise these explainability mechanisms and use the context we discussed in Section 1.2 to motivate why they became the front-runner in XAI and remain integral to the future of the discipline.

Post-hoc explanations are techniques that are applied after an AI system has been trained. They provide insights into how a model arrives at its decision without modifying the underlying model. Most often, the investigative goal of post-hoc explanations is to extract the reason a black-box gave a particular output in terms of a set of inputs. In Section 1.2 we have seen how media coverage of machine bias transformed the narrative surrounding AI from one of optimism to one of scepticism. What made this period interesting from an explainability perspective was the fact that black-box AI systems were already being deployed across many public-facing applications. Therefore, as the narrative shifted to demand transparency from black-box systems, the machine learning community *reacted* via the field of XAI. In its initial stages, XAI therefore played a game of catch-up with their black-box counterparts wherein the AI systems were already out there in the wild - the challenge was to make them explainable as quickly as possible [230].

Post-hoc explanations, by definition, are a form of retroactive interpretability where they enable users to analyze and interpret the decisions made by a model. They can investigate past predictions and understand the factors that influenced them, even if the model was initially deployed without explainability. This retrospective analysis can help identify biases, errors, or undesirable behavior. In this way, post-hoc explanations, compared to other forms of explainability, which we discuss further in Chapter 2, appealed to the needs of the time and are still, to this day, the most popular form of explainability, partly because they are designed to be model-agnostic and thus explainability can be added to the model after the model has been developed or deployed.

1.5 From Global To Local Explanations

If we understand post-hoc explanations to be mechanisms which extract the reason why a given model arrived at a particular outcome, then we can further characterise post-hoc explanations as global or local. Global explanations aim to provide a holistic understanding of a model's behavior. Local explanations, in contrast, focus on explaining the prediction of a specific instance. While global explanations provide an overall understanding of the model's behavior, local explanations

offer instance-specific insights, allowing users to understand individual predictions in a more interpretable and actionable manner. Both types of explanations are important for comprehensive interpretability and can be used together to gain a deeper understanding of the model's behavior at different levels of granularity [152]. Below, however, we motivate local explanations which are the focus of this thesis.

The objective of global post-hoc explanations is to determine the general relationship between the inputs and outputs of a black-box over an entire data distribution. As such, global explanations aim to capture the overall behavior and patterns learned by a machine learning model. As we have discussed throughout this chapter, black-box AI systems often have complex architectures with millions of parameters. Both extracting and summarising the behavior of such models in high dimensional space in a way the end user can understand is often impossible [187]. In contrast, if you focus on an individual prediction, as is the case with local explanations, the relationship between features and outcome may be less complex [152]. In this way, the reasons why a black-box model made a particular prediction for a local sample may be simpler to extract and summarise for this less complex part of the decision space.

We have seen in Section 1.2 and Figure 1.1 that the populist debate surrounding black-box systems has been centred around the harms of these systems on normal people interacting with them. In this way, much of the initial motivation for XAI was focused on providing explanations intended to safeguard individuals against the undesirable effects we had seen in the media. Local explanations enable users to take actionable steps based on the model's output. In understanding the reasons behind a particular prediction, users can make informed decisions, provide feedback, or take appropriate actions based on the model's recommendations.

We have seen that the COMPAS debate was instrumental in re-framing the narrative surrounding the impact black-box decision support systems were having on individuals. Since 2016 therefore, XAI has had an intertwined relationship with the law. The ability to give a reason for a decision is essential to legal practice, as such, legal decisions are documented so they can be interpreted. Documentation is an extension of how governing systems are held accountable by the public for what they do and why they do it [221]. The implementation of the European Union's General Data Protection Regulation (GDPR) [173] has sparked a legal interest in the explainability of algorithms. In addition, Article 22 of GDPR, in essence, grants an individual a "right of human intervention" [173].

Under the GDPR, an individual's ability to argue with an algorithm has been actualised as the right of an individual to ask for a human to review the AI's decision to determine whether or not the system made a mistake [173] which places a legal obligation on the AI system owner to understand what happened, and then make a reasoned judgment as to if a mistake was made. Local explanations, which are designed to provide the reasons for an individual's outcome,can thus help organizations meet regulatory requirements that demand explanations for specific predictions or decisions. By providing clear justification at the instance level, models can be

audited and validated for fairness, transparency, and compliance.

1.6 Post-hoc Local Explanations And Data Types

As we have seen above, post-hoc local explanations are well motivated within the XAI landscape and consequently, a great number of techniques have been introduced over the past ten years which are now being used to explain a number of complex models in high stakes domains such as medicine, finance, law, and science [44]. Chapter 2 provides a more detailed introduction to some of the most commonly used post-hoc explanation methods. However, here we mention that of existing methods for post-hoc local explanations, there are two frameworks which have been widely recognized as the state-of-the-art in XAI [74]. These are:

- Locally Interpretable Model Agnostic Explanations [175] (LIME)
- SHapley Additive ExPlanations [138] (SHAP)

XAI, while intended to address the malevolent behaviour of AI black-box systems, is in itself not a magic bullet to ensure the trustworthiness of AI systems. Furthermore, LIME and SHAP have themselves been shown to produce misleading explanations [66, 117]. In these cases, an explanation has the potential to cause more harm than good and arguably, we would be better off without the explanation.

This thesis focuses on the limitations of LIME and SHAP and related approaches for post-hoc local explanations. We attempt to address their shortcomings in order to reduce the likelihood of future occurrence of misleading explanations.

A particular issue affecting state-of-the-art explanations is the asymmetric attention paid to different data-types. Most of the post-hoc local explanation methods which currently exist in the literature (including LIME and SHAP) have been developed with tabular, image or natural language data in mind. Time series data, on the other hand, which refers to any collection of observed variables as a function of time, has received relatively little attention from the XAI community. Despite this, the growing ubuiquity of time series data is undeniable [157]. With the increasing connectivity of our world, sensors and systems are continuously generating vast quantities of time series data which has led to an increasing demand for accurate and interpretable time series models [157]. Time series have traditionally been considered as a complex data-structure [129] yet the kind of deep learning model being introduced now means we are required to understand less the underlying data and still obtain good predictive performance. There are several reasons why there may be fewer XAI techniques specifically designed for time series data compared to other data types.

One of the most prominent barriers to explainable time series is that these data-structures often exhibit complex temporal dependencies, making it challenging to analyze and interpret the patterns and relationships within the data. The challenges invoked by these dependencies are

exacerbated by the fact that time series data can have high dimensionality and long sequences, which pose challenges for XAI techniques which rely on the assumption that the resulting explanations should be interpretable to humans. Many popular explanation techniques, such as LIME or SHAP, may thus not directly translate to the temporal domain.

Developing specialized time series techniques that can effectively capture and convey temporal dynamics is an ongoing research area [77, 156]. As the demand for explainable time series increases, researchers are actively working towards developing more specialized and effective techniques for explainability in this domain. There exists a real need to explain these models both in a capacity that allows us to reason as to why a black-box reached a certain outcome but also as a way of understanding the underlying data-structure being modelled by the black-box.

1.7 Digital Healthcare

One industry where the need for explainability is particularly well motivated is healthcare. As we have seen in Section 1.3.1, AI has the potential to revolutionize healthcare by harnessing the increasing amount of patient data. AI is already being applied in medical imaging, diagnostics and decision support, drug discovery and personal medicine [29]. Even though AI-driven systems have been shown to outperform humans in many of these tasks, the lack of explainability continues to spark criticism. Thanks to the combination of grave consequences of mistakes, propensity for subjectivity encoded in data and ubiquity of confounding variables, healthcare is a prime example of an industry where the explainability of machine learning is paramount to its success. However, explainability implemented by computer scientists is not a magic bullet for trustworthy AI, as creating suitably trustworthy systems which can be confidently used by respective clinicians invokes a host of medical, legal, ethical, and societal questions that require thorough exploration [103].

1.7.1 Need For Post-hoc Local Explanations In Healthcare

Given the vast number of black-box models which exist within healthcare, the number of potential explainability mechanisms is also vast. One approach is to make a global explanation by listing what features are generally more important while making the prediction [29]. In healthcare however, it is desirable to obtain an instance-specific explanation which allows for more individualized decision making, thus providing the patient with more personalized care [29]. As such, within digital healthcare, most of the current explainability techniques are developed as individual and post-hoc [29]. The most likely reason is their ease of use for target users [118].

Kumarakulasinghe et al. [118] evaluate the post-hoc explanations generated by LIME for clinical machine learning classification models and found that LIME provides a patient-specific explanation for a given classification. However, Kumarakulasinghe et al. note that the explanations provided by LIME can sometimes be inconsistent or unstable which reduces trust [118]. While

promising, and well motivated in clinical decision support systems, post-hoc local explanations have a long way to go before they can be widely applied and more importantly before we can trust their implications. In this thesis, we build upon LIME to address some of its limitations particularly when applied to time series data.

1.7.2 Time Series In Healthcare

Time series analysis has been even slower to come into mainstream medicine than other branches of statistics and data analysis, likely because time series analysis is more demanding of record-keeping systems [157]. During the 20th century, time series data within healthcare largely consisted of electrocardiograms (ECGs) and electroencephalogram (EEGs) which measure electrical signals in the brain and in the heart. However, nowadays following the introduction of wearable technology and "smart" digital devices there is an increasing availability of time series data available which is collected on both sick and healthy individuals.

A prominent source of time series data within healthcare is electronic health records which store the history of an individual's medical history including diagnoses, medications, laboratory values, and treatment plans, in this thesis, we refer to these time series data as *Patient Trajectories*. Patient trajectories are a rich and useful data source which can be predictive of the future progression of a disease. As such, modeling patient trajectories is a motivated research discipline which could help develop robust models of diseases that capture disease dynamics [4]. Recently there have an increasing number of promising deep-learning models which use patient trajectories to make predictions [169, 211]. In this thesis, we explore the kinds of explanation needed for this kind of data-type which would increase the confidence in the outcomes determined by the associated models.

1.8 The MIMIC Sepsis Cohort

An underlying goal of this thesis is the development of post-hoc local explanations for time series within healthcare. While all of our methods are applicable to multiple domains, we contextualise our methodological contributions within the healthcare setting by considering a real-world setting and associated dataset, the MIMIC Sepsis Cohort. Throughout this thesis, we use the MIMIC Sepsis Cohort as a running example which unifies many of the ideas and demonstrates how our contributions can be used in the future of XAI in healthcare. In particular the MIMIC Sepsis Cohort, which we describe in full below, is an example of a patient-trajectory dataset, which as we have discussed above, is one of the most rapidly growing types of data source within healthcare.

MIMIC-III is a large, public database comprising anonymised health-related data associated with over forty thousand patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012 [98]. The MIMIC-III database has facilitated a large number of research studies in epidemiology, clinical decision support management and predictive

modelling [98]. One of these applications is the development of patient trajectory models for sepsis mortality detection.

Sepsis can be defined as severe infection leading to life-threatening acute organ dysfunction [113] and is among the leading causes of death in intensive care units (ICUs) worldwide. Its recognition, however, particularly in the early stages of the disease, remains a medical challenge. In an attempt to address this, the model proposed by Komorowski et al. [113] aimed to build a predictive model for sepsis mortality using the MIMIC Sepsis Cohort dataset [113] which we detail below.

The MIMIC Sepsis Cohort was created by Komorowski et al. [113] where a subset of individuals was selected from MIMIC-III. Vital sign patient trajectories of a total of 19598 individual patients were included from up to 24h preceding until 48h following the estimated onset of sepsis. A total of 52 predictor variables were extracted from MIMIC-III including demographics, vital signs, laboratory values, fluids and vasopressors (For a full list see Komorowski et al. [113]). Patient data were coded as multivariate time series with 4-h time steps. As such, the length of individual patients trajectories varied across the cohort. The binary outcome variable was 90-day mortality where a value of 1 indicates that the patient died from sepsis and an outcome of 0 indicates a patient survives.

In this thesis we extract univariate and multivariate time series datasets from the MIMIC Sepsis Cohort by selecting varying subsets of patient trajectories. In each setting we build time series classifiers, which distinguish between patient trajectories that result in death from those that result in survival, upon which we apply the post-hoc local explanation methods developed as part of this thesis.

1.9 Research Aims

Following the above motivation of post-hoc local explanations, with a particular focus on the development of time series explanations within a healthcare setting, we now explicitly lay out the research objectives of this thesis.

- **1. Adapting LIME for Time Series** Our first objective is to understand the challenges arising from the application of LIME to time series data. We have discussed how there is a lack of existing explainability methods for time series. However, missing from the literature is a rigorous evaluation of why the application of existing methods to time series is challenging. Our objective in this thesis is therefore to understand exactly what must be adapted when applying LIME to time series. We specify three open challenges which should be addressed to adapt local surrogate models for time series data.

- **2. Shapley value in the Presence of Interacting Features** Our second objective is motivated by the ongoing criticism of the Shapley value when applied to feature attribution.

Our objective therefore is to introduce the Shapley value within its game-theoretic context, something which is largely overlooked in the literature. We use this context to understand why, in the presence of interacting features, the Shapley value sometimes generates misleading attributions. Our objective, given the above exploration, is to develop an alternative mechanism for post-hoc explanations which are more robust to feature interaction.

- **3. Addressing the Gap between the Shapley value and Counterfactuals** With the ever-growing landscape of XAI solutions, it is increasingly difficult to unify approaches. Furthermore, an understanding of which approach should be used, and when, is lacking in the literature. One of our aims in this thesis is to unify existing approaches to post-hoc local explanations with the kind of investigative question they ask. We unify the research discipline of XAI with that of causality which allows us to unify Shapley values with the idea of counterfactual explanations: two approaches which are often viewed as distinct in the literature. Our aim is to develop a novel method for feature attribution which is more closely aligned with a counterfactual explanation [173].

- **4. Conceptualisation of a Time Series** One of our open questions, introduced as part of Research Aim 1, is the challenge of decomposing a time series into an interpretable representation. A further research objective is to explore alternative ways of decomposing, or conceptualising, a time series. Firstly, we question whether an imposed decomposition captures the behaviour of the black-box model. Our objective is to explore whether time series can be automatically decomposed into interpretable concepts which more accurately reflect the behaviour of the underlying AI system. Secondly, we question whether we actually need a conceptualisation at all. Our objective is to consider whether there is an alternative attribution technique which takes as input the high-dimensional raw time series but returns an interpretable and minimal explanation without the need for conceptualisation.

- **5. Explanations of Multivariate Time Series in Healthcare** Our objective is to design post-hoc local explanation for patient trajectories within healthcare which take into account the temporal dependencies which exist between heterogeneous variables recorded over time.

1.10 Outline Of Thesis

In Chapter 2 we first establish the background underlying the diverse research landscape of XAI. We define the nebulous concept of an "explanation" and present our approach to taxonomising state-of-the-art approaches, in particular motivating the importance of insights from the explanation sciences. We continue by establishing what we mean by a "time series" and give an overview of time series analysis focusing on the rise of deep learning on this datatype, motivating the need for explanations.

Chapter 3 is dedicated to Research Aim 1. We introduce the mathematical background underpinning the LIME algorithm [174] which enables us to communicate why LIME has become so popular as a post-hoc local explanation method. We continue by exploring why applying LIME to time series data is challenging and encapsulate these challenges as the following open questions: How do we effectively decompose a time series into a semantically meaningful representation? How do we remove salient information from a time series? How do we define a local neighbourhood around a time series?

For each of our open questions we propose a solution which is motivated by particular properties of time series data. For the challenge of decomposition, we propose a segmentation algorithm, Nearest Neighbour Segment which decomposes a given time series into an interpretable representation comprised of motifs and anomalies. For the challenge of information removal we propose a perturbation algorithm Realistic Background Perturb which uses the frequency domain representation to remove high frequency bands in a given time series segment. For the challenge of defining a neighbourhood, we propose the use of the Dynamic Time Warping measure. We unify each of the algorithms above into our adaptation of LIME for time series, LIMESegment. We continue by demonstrating experimentally the benefits of LIMESegment over LIME when used to explain time series. We conclude Chapter 3 on a example patient trajectories from the MIMIC Sepsis Cohort to show the applicability of LIMESegment in a healthcare setting.

Chapter 4 is dedicated to Research Aim 2. We begin by introducing the game theoretic research domain of value attribution and show how solution concepts, such as the Shapley value, can be characterised within this research landscape by their associated axiomatisation. We motivate the standard rhetoric within game theory that there does not exist a "one size fits all" solution concept for value attribution despite the Shapley value's widespread adoption. Furthermore, we connect solution concepts and their associated fairness axioms to the challenge of feature attribution where we argue that game theory offers multiple alternatives to the Shapley value for explanations which approach value division from different perspectives of fairness. We then begin our exploration of how the Shapley value was introduced as the backbone of post-hoc local explanations and introduce the large number of value functions which map the Shapley value from game theory to machine learning. We continue by investigating the different vulnerabilities of the Shapley value when used for local explanations. We argue, using the concept of function decomposition, that the Shapley value generates misleading explanations when the underlying black-box function is not fully-additively separable. We then propose our algorithm, Shapley Sets, as a method inspired by the function decomposition literature which automatically finds the non-separable variable groups for any underlying black-box function. We experimentally and theoretically show the robustness of the attributions generated by Shapley Sets compared to those of the Shapley value in the presence of interacting features in both model and data.

Chapter 5 is concerned with Research Aim 3. We begin by connecting the diverse landscape of XAI approaches with the equally diverse set of investigative goals each approach is designed to

answer. We continue by establishing the importance of causal relationships when determining explanation and introduce Pearl's causal hierarchy as a formalisation of the different types of causal relationship we can extract from a system. We explore different philosophical perspectives of the notion of a cause and distinguish between a general and singular notion of causality, introducing the concepts of the probability of necessity and sufficiency in the general case and actual, counterfactual and bifactual causes from a singular perspective. We then show how value attribution from game theory can be connected to causal questions regarding an underlying system and show how under a binary outcome game, the Shapley value is equivalent to the probabilistic necessity and sufficiency of a player generating a winning outcome in the game. We then consider post-hoc local explanations and argue the value of bifactuals in providing minimal, actionable explanations. We introduce and motivate the concept of determining the "bifactual effect" of a variable on an outcome and show how the Gately value, an alternative solution concept to the Shapley value, attributes in a way which is proportional to the bifactual effect of a player on the game. we introduce a novel method for feature attribution, Gately Feature Attribution, which is more in line with a bifactual explanation than the Shapley value. We experimentally and theoretically motivate Gately Feature Attribution across multiple settings of post-hoc local explanations.

Chapter 6 unifies the approaches introduced in Chapter 4 and Chapter 5 with Research Aims 4 and 5. We begin by extending the challenges associated with applying LIME to time series data mentioned in Chapter 3 to the use of feature attributuion methods based on the Shapley value. We motivate the use of Shapley Sets as a way of automatically decomposing a time series into interpretable concepts which address some of the limitations of NNSegment. We continue to motivate the use of the Gately Feature Attribution on univariate time series as a way of producing minmal attribution vectors which avoid the need for conceptualisation altogether.

Chapter 6 continues to explore Research Aim 5 by introducing the concept of *Differential Attribution* as the setting where we want to attribute a change in model outcome in terms of the changes in features between two temporally consecutive samples. We introduce the Aumann-Shapley value [15] as an alternative solution concept to the Shapley value from the game theoretic literature and show how the Aumann-Shapley value produces more realistic Differential Attributions. We then propose our novel Differential Attribution method which first constructs a local surrogate model around a multivariate time series and then determines the Aumann-Shapley value to provide Differential Attributions for patient trajectories and multivariate time series models. We experimentally validate our approach on a variety of datasets, including the MIMIC Sepsis cohort and show how our proposed attribution is more faithful to the underlying model than state-of-the-art explanation approaches.

	Data Type	Faithfulness/Deletion	Robustness	Sensitivity	Accuracy
Chapter 3	Univariate Time Series	X	X		
Chapter 4	Tabular Data	X		X	
Chapter 5	Tabular Data	X			
Chapter 6	Univariate Time Series	X			
Chapter 6	Multivariate Time Series				X

Table 1.1: Table provides the chapter breakdown of data type and metrics covered in this thesis

1.11 Data Types and Metrics

As outlined in the previous section, this thesis is concerned with an exploration of post-hoc local explanations. While each of the approaches proposed in this thesis are motivated by time series, they are not limited to this data structure. Indeed the methods of Chapters 4 and 5 are first introduced on tabular data which illustrates the improvements each attribution method offer over the Shapley value. Chapter 6 later applies these methods to univariate time series. For clarification Table 1.1 displays the data type which takes the main focus of each chapter.

It is well known that Explainable AI mechanisms are difficult to evaluate as there is no standardised metric system within the XAI community. Why this is the case is discussed further in Chapter 2. However, here we illustrate with Table 1.1 the metrics used to evaluate each of our contributions in each of the chapters. Faithfulness a measure of how well the explainability mechanism identifies the most important features and also known as Deletion, is used to evaluate each method and is the most commonly applied evaluation metric across the XAI literature [119]. Note that we adapt Faithfulness metric in each chapter dependent on the data type. Robustness is used to evaluate how sensitive to random noise our adaptation of LIME to time series is in Chapter 3. It is well documented that LIME is sensitive to change in input [194] and as such this metric was well motivated within this chapter. Sensitivity, a measure of how successful the Shapley value is approximated, is used to evaluate the method of Chapter 4, Shapley Sets where a comparison of approximation between our method and the Shapley value is meaningful. Finally, Model Accuracy, an aggregated measure of Faithfulness over an entire dataset, is used to evaluate our method for multivariate time series explanations in Chapter 6 as it provided a more reliable metric in the presence of high dimensional data structures.

1.12 Research Contributions

Below we summarise the novel contributions of this thesis as our four methods for post-hoc local explanations.

LIMESegment: An adaptation of LIME for time series which replaces the original LIME components with three separate algorithms which are specifically designed for time series.

Shapley Sets: A feature attribution algorithm which automatically decomposes any black-

box function into non-separable variable groups which together are attributed value, resulting in more robust explanations in the presence of interaction either within the model or in the data.

Gately Feature Attribution: A method for feature attribution which is motivated as an alternative to the Shapley value to be used with an off-manifold value function. Gately Feature Attribution is inspired by the game theoretic Gately value and the individual contribution of a feature to a local explanation given only its impact in the example to be explained and a baseline sample.

Aumann Differential Surrogate Explanations: A novel explanation method for Differentia Attribution which we apply in two settings inspired by patient trajectories in healthcare. The method constructs a differentiable surrogate model around an individual multivariate time series sample via a Multi-adaptive regression spline. The Aumann-Shapley value [15] is then determined for each feature between selected temporally indexed observations. Aumann Differential Surrogate Explanations explicitly takes into account the temporal dependence between heterogeneous feature values in attributing a particular outcome to individual features.

1.13 Research Outputs

Below we detail the research articles developed during this thesis, including the articles which form parts of the thesis and those, conducted additionally, which are out-of-scope of this thesis.

1.13.1 Articles Related To The Thesis

- [192] Sivill, Torty, and Peter Flach. "Limesegment: Meaningful, realistic time series explanations." International Conference on Artificial Intelligence and Statistics. PMLR, 2022.
- [193] Sivill, Torty, and Peter Flach. "Shapley Sets: Feature Attribution via Recursive Function Decomposition." arXiv preprint arXiv:2307.01777 (2023). To be submitted to AAAI Conference on Artificial Intelligence (AAAI-24).

[192] underlies the content of Chapter 3 of this thesis and [193] underlies the content of Chapter 4. Both publications are Torty's original ideas investigated, implemented and written-up under Peter's supervision.

1.13.2 Additional Research Articles

- [191] Sivill, Torty. "Ethical and statistical considerations in models of moral judgments." Frontiers in Robotics and AI 6 (2019): 39.
- [53] Divya Balasubramanian, Kai Hou Yip, Indira Sen, Matthew Forshaw, Nikita Vala, Prakhar Rathi, Ridda Ali, Sami Alabed, Sara Masarone, Stephen Kinns, Torty Sivill, Tatiana Alvares-Sanches. "Communicating high-street bakery sales predictions using

counterfactual explanations". Data Study Group, Alan Turing Institute (Zenodo). https://doi.org/10.5281/zenodo.5562660 (2021)

- Sivill Torty , Ljevar Vanja, Goulding James, and Skatova Anya. "What Can Transactional Data Reveal About the Prevalence of Menstrual Pain in England?". 2023. To be submitted to the International Journal of Epidemiology.

CHAPTER 2

BACKGROUND

2.1 What Is Meant By An Explanation?

In this section we provide background on the discipline of Explainable AI. Having discussed its origins and growing motivation in Chapter 1, this section connects its standing within the machine learning community with multi-disciplinary perspectives of the Explanation Sciences.

Due to the rapidly increasing number of research outputs falling under the Explainable AI (XAI) umbrella, navigating the discipline is often challenging, prompting numerous attempts to survey the research landscape [55, 56, 234]. Taxonomising XAI however, is difficult and there is no unified consensus on how different branches of the discipline connect [44]. This is due, in part, to its accelerated expansion, the heterogeneity of associated approaches, and diversity of explainability applications. In this section, we intend to highlight, by connecting XAI to its counterparts in the Explanation Sciences, what we believe to be one of the most significant challenges facing XAI. XAI, as a research discipline, has been founded within computer science, and has oftem overlooked the rich multi-disciplinary literature which brings together philosophy, sociology, and psychology of how humans generate explanations [147].

A significant challenge arising from the detachment of XAI from the Explanation Sciences and one which particularly complicates taxonimisation, is that of definitions. To begin explaining our AI systems we must first ask "What is meant by an explanation?". This question is loaded with philosophical interpretation and has inspired countless theories of explanation emerging from the social sciences [149]. The ambiguity of defining explainability has undoubtedly led to the diversity of approaches within XAI which encode different and sometimes conflicting definitions and measures of explanations [149].

Lipton [133] argues that explainability is ill defined within the XAI literature and is composed of several ideas including transparency and post-hoc explainability. Doran et al. [55] examine a

corpus of literature from various well established communities in XAI where they distinguish between "opaque systems" that offer no insight to the algorithmic workings, "comprehensible systems" which offer user friendly explanations of algorithmic workings and "interpretable systems" that can be mathematically analysed to understand the underlying algorithmic workings [55]. Other attempts at taxonomisation include that of Molnar [152] who uses a taxonomy of four categories, post-hoc or intrinsic; result of the method; model specific model agnostic and local or global methods.

A common occurrence within the XAI literature, which is noted by many of the above surveys, is that approaches often fail to explicitly define an explanation. We argue, that perhaps the first step in addressing the difficulty of navigating the XAI landscape is being explicit with definitions. If the XAI community become more transparent about the assumptions encoded in the associated approach to explainability, this in turn, will inform policy makers and government and will address some of confusion plaguing the discipline. We now present our understanding of an explanation.

Our definition of an explanation (Definition 2.1) is comprised of a four stage process which we detail in Section 2.2. Our definition is inspired by Miller [147] to be human-centric. Using our definition of an explanation, we attempt to categorise and summarise the XAI landscape detailing current trends and approaches. Particularly, we emphasize that most of the current literature exists within Stage 1 of the explanation process. We argue that future work in XAI should adopt a more holistic approach to explainability by considering the later stages included in our definition. We also situate the work of this thesis within the context of the stages of an explanation.

2.2 The Stages Of An Explanation

One of the most highly cited definitions of an explanation is that of Josephson and Josephson [99], who state that "An explanation is an assignment of causal responsibility". The dependency of an explanation on causal attribution, which is the process of extracting a causal chain and displaying it to a person, is widely accepted within the Explanation Sciences [190]. However, our definition of an explanation is inspired by the argument of Miller, who states "while a person could use such a causal chain to obtain their own explanation, this does not constitute giving an explanation" [147]. Miller argues that causal attribution comprises just part of an explanation, which in total reflects a process by which a human receives and understands the reasons for why an event occurred. As such, inspired by the argument of Miller [147], below we define an explanation as a four part process which involves an identification of causes, a selection of causes, a communication of the explanation, and finally, an evaluation of the explanation.

Definition 2.1 (An Explanation)**.** Within this thesis, our definition of an explanation refers to some output from a system that is a combination of the following stages. We argue that an ideal

explanation for a specific event X must go through the following generation process

- The causes of X are identified (Stage 1)
- A subset of the (potentially extensive) causes of X are selected as an explanation (Stage 2)
- The explanation is communicated to a human (Stage 3)
- The explanation is appropriately evaluated (Stage 4)

Given Definition 2.1, we now attempt to categorise the different types of approaches within the XAI landscape. We do this by aligning existing work with the stages of an explanation defined above. First however, we formalise what we mean by a post-hoc local explanation of a black-box function which, as we have seen in Chapter 1, is the focus of this thesis. We use the following notation to represent our black-box function $f : \mathscr{X} \rightarrow \mathbb{R}$. The function f can represent any machine learning model which takes input and generates output as defined above. In this thesis, we discuss functions f which are trained on tabular, image, univariate and multivariate time series data. As the focus of this thesis is on model-agnostic methods for explainability, we use a selection of arbitrarily selected models throughout this thesis to exemplify the kinds of explanation we can generate. A detailed background description of existing machine learning models therefore, is out of the scope of this thesis. We assume that the reader is somewhat familiar with standard classification and regression models such as Logistic and Linear Regression, Support Vector Machines, Random Forest Classifiers and the concept of neural networks and deep learning. Given our black-box function f, to generate a post-hoc local explanation, we take our prediction event as $f(\mathbf{x})$ where $\mathbf{x} = \{x_1, ..., x_n\}$ represents an individual vector of n distinct values of the predictor values $\mathbf{X} = \{X_1, ..., X_n\}$. An explanation will therefore try to communicate the causes of $f(\mathbf{x})$ in terms of its input values $\{x_1, ..., x_n\}$.

2.3 Stage 1: Causal Attribution

Stage 1 of explanation generation is concerned with causal attribution, or methods in the XAI literature that attempt to collect the possible causes for an outcome. This phase of the explanation generation process has attracted the most attention in the literature and can be characterised as either approaches that look at extracting causes from inherently transparent models or those that extract causes of outcomes already determined by the model (post-hoc explanations). While reading the following, we draw attention to the intertwined yet complex relationship between the disciplines of the philosophy of causality and XAI. Many of the following techniques approach explainability from a machine learning perspective, neglecting the philosophical literature underpinning causality [87, 163], which we explore in Chapter 5 of this thesis. In light of this, the extent to which the following techniques of causal attribution can be considered truly causal is a point of debate within the literature [87, 95].

2.3.1 Transparent Models

By taking our event to represent an individual prediction $f(\mathbf{x})$, one of the ways by which we can identify the possible causes for this prediction is by using something which is referred to in the literature as a "transparent model" as our model f. Rudin [179] expresses concern over the contemporary approach of designing solutions to high stakes problems using black-box methods. She argues that instead of trying to make these models explainable we should instead be focusing on applying transparent models to these problems [179]. Arriteta et al. [13] define a model to be transparent if it can, by itself, be understood completely. These models include linear and logistic regression, decision trees, bayesian models, rule based learning, generative additive models, and K-nearest-neighbours [13]. Under a transparent model the event $f(x)$ can be casually attributed to each feature value using the inherently interpretable mathematical properties of the model. For example, the linear coefficients given a linear or logistic regression model. Arrieta et al. claim that linear and logistic regression models meet all three characteristics of transparent models. However, note the difficulty in maintaining this transparency when applied to large datasets as well as the need to translate this transparency into some sort of human understandable explanation [13].

Deep learning is famous throughout the literature for bringing high predictive accuracy to a multitude of data-based problems. Rudin [179] claims however, that there is often no significant difference in accuracy between these and simpler, more interpretable algorithms [179]. Similarly, Ba et al. [18] question the value added by increasing network depth. Hagras et al. [78] however, argue that "transparency rarely comes for free". As data complexity increases so does the trade-off between accuracy and interpretability [78]. Kim et al. [107] argue that an effective explanation is not necessarily entirely transparent but it should aid the user's understanding of a particular result. These arguments characterise the "Accuracy Interpretability Trade-off" which we elaborate on in the following section.

2.3.2 The Accuracy Interpretability Trade-off

In Chapter 1, we have discussed how the proliferation of black-box AI systems has been accompanied by a narrative pitting more simple, or transparent AI systems against more complex black-boxes. We have seen how the performance of black-boxes on certain tasks ultimately won out in this debate, which has seen the proliferation of highly complex models such as Large Language Models. The accuracy interpretability debate, however, is still ongoing, as it is widely believed that in optimising for performance there is a trade-off to be made in terms of the resulting interpretability. It is this debate which, in part, has driven the success of post-hoc local explanations whereby if we accept that while models are becoming increasingly complex and powerful, they are simultaneously becoming less interpretable, then if we want explainability, we must develop post-hoc explanations. However, there are two assumptions, which are the main drivers behind the continuation of this argument, which we challenge below.

2.3.3 Are More Interpretable Models Less Accurate?

Interpretable models are often simpler and have fewer parameters compared to more complex models. Simpler models may not capture all the intricate patterns in the data, which can lead to lower accuracy in certain scenarios. However, simpler models can still perform well, especially when the data is relatively straightforward or when the problem does not require high complexity. Furthermore, deep learning can sometimes overfit, especially when the problem being modelled can be solved with a less complex model [127]. The importance of both interpretability and performance varies across different domains and applications. In some domains, such as healthcare or finance, interpretability is crucial for understanding the decision-making process, ensuring transparency, and addressing legal or ethical concerns [179]. In these cases, the focus may be on developing more interpretable models even if they perform less well. The accuracy interpretability debate definitely perpetuates the assumption that deep learning will provide a better performing AI system. In Chapter 7, we debate this and argue that moving forward, we should shift from the accuracy interpretability discussion to a discussion surrounding the appropriate selection of models given a particular context.

2.3.4 Are More Complex Models Less Interpretable?

In general, more complex models tend to be less interpretable compared to simpler models. As models become more complex, such as deep neural networks with numerous layers and parameters, understanding their inner workings and decision-making processes becomes increasingly challenging. However, it has been recently argued [20] that contrary to a common belief held by AI practitioners, black-box models may often be both the most accurate and the most explainable models to end users. With respect to black-box models, users are able to understand what the model does without necessarily having to understand how it works [20]. In this sense, the use of complex black-box models with explainability incorporated into these systems, i.e. post-hoc explanations which we explore in the following section, is prioritised over "inherently transparent" models. From the above, we question the extent to which the accuracy interpretability trade-off manifests in reality. Despite this, we must accept that there will be the ongoing application of complex models, with strong performance which are considered black-boxes. This thesis is concerned with the development of methods which attempt to generate explanations for these kind of model.

2.3.5 Post-hoc Explanations

Post-hoc explanations, which have been introduced and motivated in Chapter 1, are techniques which identify the causes of prediction events given functions f which are themselves not transparent, and in doing so treat them as an impenetrable black-box that can be probed or reasoned about. Arrieta et al. [13] argue that this form of explanation reflects the way

in which humans explain things themselves. Lipton describes post-hoc analysis as methods including "natural language explanations, visualizations of learned representations or models and explanations by example" [133]. Ribeiro et al. [175] argue that model agnostic approaches to post-hoc explanations address the limitations of transparent approaches, offering model flexibility that treats the underlying model as a black-box, allowing for the choice of any machine learning model [175]. Arrieta et al. [13] categorise model agnostic approaches as either model simplification, feature attribution explanations or visual explanation techniques.

Feature attribution methods are the most common form of post-hoc explanation and characterise the methods which attempt to rank the relevance of each feature in a prediction [13]. These techniques include LIME [174] and SHAP [138], which are the main focus of this thesis.

One of the key selling points of both LIME and SHAP is that both techniques are designed to be both model and data-agnostic and customisable to individual investigative goals. The initial generality of the algorithms perhaps seemed like the utopian dream for the domain of XAI where a one size fits all explanation is more desirable than individual alternatives. In hindsight, the limitations of LIME and SHAP have been acknowledged [117] and it is well known now that within XAI, one explanation does not fit all. A detailed background on both LIME and SHAP are given in Chapter 3 and Chapter 4 respectively where we explore these limitations in more detail. Counterfactual explanations, like those proposed by Wachter et al. [216], are an alternative mechanism to feature attribution for post-hoc explanation. We distinguish between the two in Section 2.4.

Examples of model simplification techniques include, for example, GREX, Genetic Rule Extraction [114], extended by Johansson et al. [97] to show how the explanations generated by the method for rule extraction impacts accuracy. It is worth noting that LIME [174], despite being a feature attribution technique, can also be considered a model simplification method for post-hoc explainability as it constructs a simple surrogate to approximate a complex model. Explanations by example have been developed in the form of activation clusters [13] or influential instances [110]. Model specific approaches include post-hoc methods that attempt to explain shallow ML models that are not inherently transparent. This can include tree ensembles or random forests, made explainable by Auret et al. [17] who extract variable importance from a random forest and support vector machines. The most popular methods for explanations of deep neural networks are local feature extractions [13] such as LIME and SHAP. However, alternative methods include the adaption of convolutional neural networks to assign each filter layer with an object which generates explanations [232], and the approach of Wisdom et al. [225] who design explainability for recurrent neural networks by modelling a a group of correlated observations with a set of sparse linear vectors.

2.3.6 Difference Between Attribution And Explanation

Most of the current approaches to XAI, including LIME and SHAP, are focused on the attribution part of the explanation generation pipeline, such that they are concerned only with the identification of the causes of a particular prediction event [226]. These methods often output an attribution vector as the resulting explanation such that an end-user can understand why the black-box model gave a particular output in terms of a ranked vector of the individual's set of inputs.

The attribution vector as determined by these methods, according to our definition of an explanation, only represents part of the explanation generation process. Miller [147] argues that returning an extensive list of possible causes for a prediction event may be misleading or non-useful to the human. Similarly, Halpern et al. [81] distinguish explanation from causality as the following: "causality is the problem of determining which events cause another, whereas explanation is the problem of providing the necessary information in order to establish causation". The movement from Stage 1 to Stage 2 of the explanation generation process can therefore be considered as the consideration of how best to present the causes of a particular event to a human. These sorts of approaches are considered in Stage 2. Furthermore, the original methods of LIME [175] and SHAP [138] have since been extended in a variety of ways. Some of these methods have included visualising the attribution vector, or restricting the kind of attribution vector presented to an end user. In the following sections we shall see how this extension of an explanation beyond an attribution vector is more in line with our definition of an explanation as it considers the optimal transferal of knowledge from human to end-user.

In this thesis, each of our proposed methods for post-hoc local explanations, are ultimately feature attributions, each produce an attribution vector in terms of an ordering over inputs. In this way, whether our proposed solutions can truly be considered as explanations under our definition is debatable. However, we argue that for each of our proposed methods, which improve original LIME [174] and SHAP [138], we have considered how best to optimise the attribution vector with the end-user in mind.

2.4 Stage 2: From Cause To Explanation

A key difference between XAI research in Stage 1 and research in Stage 2 is the consideration of a human within the cause extraction process. Anjomshoae et al. [11] argue that humans anthropomorphise AI agents to allow them to explain their behaviour and as such, explanations should establish a "theory of mind" which allows humans to incorporate the behaviour of the AI system into their own world view. Anjomshoae et al. argue that we must therefore turn to the social sciences to start explaining this behaviour [11]. Madumal et al. [141] suggest that causal models represent the method of explanation most closely reflecting that of human reasoning [141] whereas Mittelstadt et al. motivate explanations that are contrastive and communicative

[150]. Contrastive explanations have been explored by comparing necessarily present and absent features for classification [50]. Walton et al. [217] combine contrastive explanations with argumentation suggesting that not only should explanations transfer knowledge of the causal chain but should also give argumentative support of this knowledge.

Counterfactual explanations, have received attention within the XAI literature, distinguishing them from feature attribution methods, as an optimal form of explanation as vehicles that "provide information to users on what might be done to change the outcome of an automated decision" [102]. A counterfactual explanation is a causal statement which attributes an event to a set of causes in the form *If X had not had happened then Y would not have happened*, *If my alarm had gone off then I wouldn't have been late for work*. Counterfactual reasoning requires imagining counterfactual worlds which might have happened but didn't. Counterfactual explanations can be used to explain predictions of individual instances where the event is the model's prediction on the instance and the causes are the particular feature values of the instance [152]. Counterfactual explanations are the focus of Chapter 5 where we delve into the mathematical and philosophical foundations of these forms of explanation.

2.5 Stage 3: Communication Of Explanations

Stage 3 of the explanation generation process is concerned with how best to communicate explanations to a human. We have discussed how many of the existing attribution methods often produce an attribution vector as the resulting explanation. Methods considering Stage 3 of the explanation process, may therefore be concerned with how best to communicate that vector. Explanations can be delivered numerically, graphically/visually, in text and as a part of a formal argumentation framework. Numerical explanations are those which could, for example, output relative feature importance and weightings such as the attribution vector generated by LIME and SHAP, or the vector of coefficients pertaining to a Linear Regression Model. Visual Graphical and visual explanations can include techniques such as saliency graphs for images [84] or partial dependence plots [63]. Other visual explanation techniques include a visual analytic interface that shows connections between instance level explanations [208], or a visual tool that uses decomposition to generate explanations of predictions that are then averaged and visualised to compare [176].

Since their theoretical introduction, both LIME and SHAP have been transformed into software packages which provide visualisations of the attribution, examples of which are shown in Figure 2.1. These visualisations, particularly for the SHAP package, are ubuiquitous across the literature and are therefore commonly assumed to be the optimal way of presenting attributions. However, this assumption is subject to the fact that most people viewing these visualisations were originally within the XAI community. Actually, there is a big difference in understanding of these graphs between those who are within the XAI community, those who are within the ML

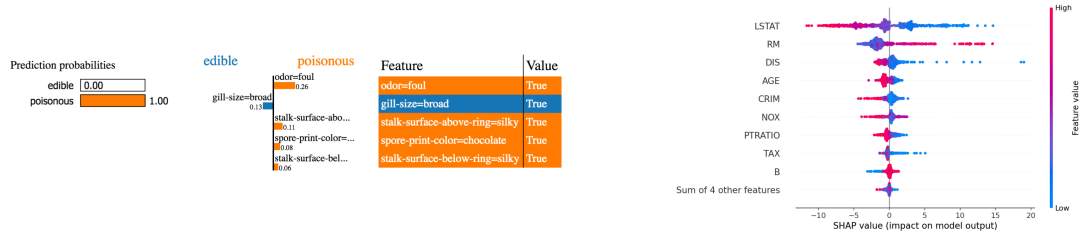


Figure 2.1: Figure shows an example visualisation of the attribution vector generated by the LIME library [175] (left) and the SHAP library [138] (right)

community and those who have no ML experience. A study assessing the capability of humans to meaningfully interpret LIME explanations [52] found that interviewees with both XAI and no XAI prior knowledge expressed uncertainty about what the LIME illustrations show. As such, research in Stage 3 of the explanation process should consider the best method of communication with respect to the end-user.

Communication of explanations may extend beyond a one-off output. Many explanations of systems need to be interactive and adapt over a relationship with a user [106]. These can include interactive explanations of personal AI devices such as a smart home device or chatbot. Interactive explanations may incorporate user feedback into future explanations [106], adapt explanations to suit the user [197]. We explore the significance of interactive explanations in the future of explainability in Chapter 7.

2.6 Stage 4: Evaluating Explanations

We have discussed at length the many open ended questions facing XAI. One of the main challenges for the discipline however, is the question "what constitutes a successful explanation?" Evaluation of XAI is therefore a challenging area of research, again plagued by a lack of standard definitions, attracting much speculation but not yet reaching a unified consensus. Alongside the ever-growing number of XAI methods which are populating the research landscape, there is an increasing need to design and apply suitable evaluation measures [44]. Unlike the evaluation of machine-learning models, for which there exist a large number of quantiative performance measures, the performance of an explainability tool cannot so easily be quantified. This difficulty, akin to the lack of standardisation surrounding explainaility terms, is part of the reason for the huge variation in explanation techniques [155]. The optimal evaluation methods could depend on the application domain, the type of explanation, the type of data, the background knowledge of

the user and the investigative question to be answered [155].

The XAI community, has not yet agreed on a standardised way of evaluating approaches. As such, each contribution to the literature is often evaluated in its own way, using a bespoke metric, which is often selected as "anecdotal evidence showing individual, convincing examples that pass the first test of having face-validity" [155]. The lack of quantitative evaluation disrupts the application for XAI methods, since "anecdotal inspection is not sufficient for robust verification" [155]. If the XAI community were to collectively agree upon a set of standardised, rigorous and holistic evaluation metrics then in the future, we could apply explanation approaches in practice with confidence.

One of the most commonly applied quantitative measures by which to evaluate and compare existing post-hoc explanation methods is Deletion (also referred to as Faithfulness), first introduced by Alvarez et al. [5] which relies on the assumption that if a feature is essential for a model if its removal will lead to a significant change in prediction when it is removed (or perturbed) from the instance. While Deletion has been widely applied across the XAI literature, it has been criticised [134]. In this thesis, we use a variety of quantitative metrics to evaluate our proposed methods, including an adapted measure of Deletion.

When creating XAI **for humans**, we argue that human evaluation should always be used for testing success. When evaluating explanations designed for humans further metrics should be considered including how explanations increase trust and satisfaction, Sokol et al. [195] identify further evaluation metrics of this kind including coherence, soundness and contextuality as ideal properties for human explanations [195].

In this section, we have provided background on the discipline of XAI. We began by discussing the lack of standardised definitions plaguing the domain and proposed our own, multi-stage definition of an explanation. Using this definition, we categorised some of the approaches within XAI and connected the work of this thesis to the stages of an explanation. This thesis is largely dedicated to the evaluation of existing, and the development of novel, methods for feature attribution. As we have argued above, attribution is an important yet not singular component of an explanation. The above section therefore, has not only set out the current lay of the land within XAI but has also shown the vast potential for future research in this domain.

2.7 What Is A Time Series?

In Chapter 1 we introduced one of the main aims of this thesis as the development of post-hoc explanation methods for time series. In this section therefore we introduce the background on time series analysis and discuss why, given the increasing application of deep learning methods on this kind of data, explainability is motivated. First, however we define a time series. A time series is a collection of observations made sequentially through time [32]. A time series can either be univariate, in which case the observations refer to only a singular variable through time,

or multivariate where the collection contains observations of multiple variables recorded over time. In this thesis, we adopt the following notation for a univariate time series $\mathbf{x} = \{x_1, ..., x_t\}$, comprised of t observations such that $\mathbf{x} \in \mathscr{X} \subset R^t$. Here, $j < k$ for $j, k \in \{1, ..., t\}$ indicates that observation x_j precedes observation x_k. For a multivariate time series comprised of a collection of n variables, each with t observations, we adopt the following notation such that each $\mathbf{x} \in \mathscr{X} \subset \mathbb{R}^{t \times n}$

$$\mathbf{x} = \{\{x_1^1, ..., x_n^1\}, \{x_1^2, ..., x_n^2\}, ..., \{x_1^t, ..., x_n^t\}\}$$

.

2.8 Why Are Time Series Different To Other Data Types?

Given the temporal dependence between observations in a time series, analysis of this data-type has traditionally involved harnessing this correlation structure to make some sort of prediction about the future or understanding temporal dynamics [157]. If there exists no temporal dependence between observations in a time series then usually, there is no point in applying traditional time series models to these data as traditional linear models which assume individual observations are independent and identically distributed (i.i.d) (i.e. linear regression) can be applied instead [157]. However, if the observations of a time series can not be treated as i.i.d, then we must assume that there is temporal dependency encoded within the time series and as such, we must find ways of modelling these data which account for this structure.

2.9 Traditional Time Series Analysis

One common application of time series analysis is that of forecasting which accounts for models f taking as input a time series $\mathbf{x}$ of length t and output the most likely value of the variable at the following index $t+1$. Some of the most popular methods of traditional time series forecasting models include, Autoregressive Models (AR), Moving Average (MA) mdoels, and Autoregressive Moving Average Models (ARMA), each of these models encodes a certain set of assumptions regarding how a future value of the underlying process depends on the past [157].

Traditional time series forecasting models are most commonly parametric models informed by domain expertise [130] and as such, because of the simple mathematical expressions that define these models, they can be understood clearly in terms of their parameters. Applying these models to relatively small data sets often obtains good performance on forecasting tasks. However, due to their simplicity these models often do not scale as well to larger data sets. By definition, these models are not built to handle nonlinear dynamics and will perform poorly in the presence of nonlinear patterns. [157].

In contrast, modern machine learning forecasting methods provide a means to learn temporal dynamics in a purely data-driven manner. Deep learning, however, has not yet delivered the

superior results for forecasting that it has for other areas, such as image processing and natural language processing [157]. Despite this, motivated by the fact that applying deep learning to time series forecasting removes the need to enforce assumptions and technical requirements common for traditional forecasting models, recent deep learning architectures have had success on forecasting. These methods include that of Salinas et al. [182] which proposes an autoregressive recurrent neural network model, or Lim et al. [130] who propose a Bayesian recurrent neural network architecture.

2.10 Classification And Regression For Time Series

A different application of time series analysis is that of time series classification and regression whereby, unlike forecasting where the value to be predicted is generally dependent on the most recent preceding values, the prediction is based on the whole time series. Many of the traditional models for time series forecasting can be adapted for time series classification such as that proposed by Kini et al. [108] which uses an AR model for classification. However, the use of these traditional time series models still requires certain assumptions. In contrast, machine learning models which were originally developed for other data-types are being applied to time series classification and regression where, similarly to machine learning forecasting models, there is no need to posit rules about the underlying process. Machine learning on time series classification and regression instead focuses on identifying patterns that describe the process's behavior in ways relevant to predicting the outcome of interest, such as the appropriate classification label for a time series [157].

The earliest types of time series classification involve the application of classification or regression models which were originally designed for tabular data methods such as tree-based methodologies for example, whereby formulating features of the time series is a necessary step along the way of using the methodology. Underpinning feature-based time series models is dimension reduction by using a set of features representing the time series [49]. Many machine learning algorithms however, tend to be fairly brittle in terms of the dimensionality and kinds of input data required [157]. For example, many time series feature-based models assume stationarity [157]. Deep learning offers a more flexible way to develop architectures specific to temporal data. Deep learning, when applied to time series classification and regression problems addresses many of the difficulties of pre-processing data to fit a model's assumptions. There is no requirement of stationarity. There is no need to carefully adjust the model based on properties of the underlying phenomenon, such as assessing seasonality and order of a seasonal ARIMA model. Deep learning is highly flexible as to the model and the nature of the inputs. Example recent deep architectures for time series classification include that of Wang et al. [220] who proposes a Fully Convolutional Network, and that of Dempster et al. [48] who propose a convolution kernel based architecture. For time series regression, recent approaches include that of Bloemheuvel

et al. [22] who apply graph neural networks to multivariate time series and that of Zhang etal. [233] who propose an attention-based network.

However, deep learning is not a magic bullet for time series analysis. Although there is no requirement of stationarity for deep learning applied to time series, in practice, deep learning does not do a good job of fitting data with a trend unless standard architectures are modified to fit the trend [157]. Furthermore, deep learning optimization for time-oriented neural networks like Recurrent Neural Networks, (RNNs) are not as well developed as those for image processing networks like Convolutional Neural Networks (CNNs) [157]. Time series analysis and modelling has yet to reach its golden period [157], and, to date, time series analysis remains dominated by traditional statistical methods as well as simpler machine learning techniques, such as ensembles of trees [157].

This section has introduced the background behind time series analysis, showing the progression from traditional time series models to the more recent application of deep learning. Deep learning for time series is a relatively new endeavor, but is growing in popularity [93]. Because deep learning is a highly flexible technique, it can be advantageous for time series analysis. Most promisingly, it offers the possibility of modeling highly complex and nonlinear temporal behavior without having to approximate an appropriate functional form [157]. As a result, as the application of deep learning to time series data increases, as we have seen from the proliferation of the black-box applied to other data types, there will be an accompanying motivation to explain these models.

CHAPTER 3

LIMESegment: Meaningful, Realistic Post-hoc Local Explanations for Univariate Time Series

This chapter focuses on one of the first and most popular existing mechanisms for post-hoc local explanations, Locally Interpretable Model Explanations (LIME) [175]. We begin by introducing local surrogate models as the class of explainers which categorises LIME and distinguish the components of LIME from other explanation methods in this category. We outline the challenges which may manifest when it is applied to time series.

We introduce three principal challenges of adapting local explainability to univariate time series data: 1) The challenge of conceptualisation: How do we group time series observations into semantic concepts? 2) The challenge of perturbation: How do we remove concepts from a time series while maintaining a realistic sample? 3) How do we define a local neighbourhood around a time series? We then introduce our adaptation of LIME for time series explanations, LIMESegment, which combines three solutions to the above challenges. We show experimentally how LIMESegment can be used to generate meaningful and realistic explanations for time series classification models. We apply LIMESegment to the MIMIC Sepsis cohort, demonstrating the utility of explanations in a healthcare setting. As outlined in Chapter 3, a significant proportion of the work contained in this chapter has been published by Sivill and Flach [192]. This chapter is focused on the adaptation of LIME to univariate time series, the extension of post-hoc explanations to multivariate time series is tackled in Chapter 6.

3.1 Post-hoc Local Explanations: The Challenge

As we have discussed in Chapter 1 and Chapter 2, post-hoc local explanations are built on the assumption that the local relationship between features and output as determined by the model

is both easier for a human to understand *and* more meaningful for the human to interpret than an explanation involving a global representation of this relationship. Post-hoc local explanations are unified by their treatment of the underlying function as a black-box which can be queried or approximated but not *opened*. Each of these methods results in an explanation relating the individual feature value of a given instance to its relative importance for the resulting prediction. The ways in which these explanations are constructed can vary from perturbation based attribution to counterfactual reasoning. In this chapter, we focus on LIME which, as discussed in Chapter 2, is a method for feature attribution built using a surrogate explainer.

3.1.1 Surrogate Explainers

Surrogate models can be categorised as either global or local. They broadly refer to algorithms that can be used in lieu of original models for various purposes. Surrogate models were first introduced within the context of "explaining" a more complicated underlying model by Craven et al. [40]. Here, decision trees were used to approximate the behaviour of an underlying neural network. First, the neural network's prediction was obtained on a subset of the original dataset. The transformed dataset was then used to train a decision tree whose "inherently interpretable" representation was then used to explain the original model. Generally, global surrogate models take a data distribution, which is assumed to be representative of the original decision space used to train the original complex model and train a new simpler, "inherently interpretable model" on this data using the prediction obtained under the complex model as labels.

Global surrogate explainers have the advantage of being post hoc, model-agnostic and data-universal. However, they are limited by the same interpretability questions that plague "transparent" models as discussed in Chapter 2. Furthermore, the accuracy interpretability trade-off is particularly relevant in the global decision space, where due to the high-dimensionality of the problem, there is no guarantee that the behaviour learned by the surrogate is a true representation of that of the model.

Local surrogate models approximate the underlying model around a given instance. Now, rather than using a dataset which is representative of the global decision space, the local surrogate model is designed to approximate the local behaviour of the complex model for a given instance. As such, the surrogate model takes as input the instance to be explained and generates a new data distribution *in the neighbourhood* of this instance. The original model's predictions are obtained on this local data distribution, upon which the surrogate model is then trained. There have been multiple approaches to applying local surrogate models for post-hoc local explanations [76, 96, 174]. Of these, Locally Interpretable Model Agnostic Explanations (LIME) [174] have become the most ubiquitous and are thus the focus of this chapter. Below we outline the main assumptions of LIME [174] which distinguish it from other work in the explainable surrogate landscape.

In this section, we have introduced the class of Explainable Surrogates as mechanisms for explainability and differentiated between local and global surrogate models.

3.1.2 Locally Interpretable Model Agnostic Explanations

The overarching objective of post-hoc local explanations within the context of surrogate models is encapsulated as the following: given a pre-trained model, or the black box function $f : \leftarrow \mathbb{R}$, and the variable set $\mathbf{X} = \{X_1, ..., X_n\}$, where each of the samples $\mathbf{x} = \{x_1, ..., x_n\}$ in a given input dataset $\mathbf{X}_{input}$ is characterised by n features, a local surrogate model will take as input a particular sample to be explained $\mathbf{x}$ and return a vector of attributions $\epsilon(\mathbf{x})$ which is a solution to the following optimisation problem.

$$\epsilon(\mathbf{x}) = \arg\min_{g \in G} L(f, g, \pi_{\mathbf{x}}) + \Omega(g) \tag{3.1}$$

Here ϵ is the specific surrogate model which satisfies Equation 3.1. The surrogate explainer is considered "transparent" such that we can extract the feature importances from $\epsilon(\mathbf{x})$ in some way. For example, under LIME [175], $\epsilon(\mathbf{x})$ is a sparse linear model from which $\mathbf{w}$ represents the vector of linear coefficients which we interpret as our feature importances for the given instance $\mathbf{x}$

L is the measure which determines how faithful the surrogate model g, selected from the class of interpretable models G, is to the behaviour of the underlying black-box function f in the local neighbourhood around $\mathbf{x}$, given as $\pi_{\mathbf{x}}$. In order to ensure both interpretability and local faithfulness we must minimize $L(f, g, \pi_{\mathbf{x}})$ while also minimising the complexity of the surrogate model $\Omega(g)$ such that the resulting explanation is interpretable to humans. The selection of the class of surrogate models G, faithfulness measure $L(f, g, \pi_{\mathbf{x}})$ and complexity measure Ω therefore characterises the class of local surrogate models.

The way in which LIME optimises faithfulness, $L(f, g, \pi_{\mathbf{x}})$ is characterised by three central building blocks which are detailed in the following section. First, however we discuss the importance of selecting the class of surrogate models G and associated complexity measures $\Omega(g)$ which lies at the very heart of local surrogate explainers and is a practical realisation of the "accuracy-interpretability" trade-off.

3.1.3 Faithfulness, Complexity Trade-off

The two terms on the RHS of Equation 3.1 together characterise the struggle of designing effective local surrogate explainers. Under the assumed complexity of f, we want to find a surrogate model g which matches as closely as possible the behaviour of f, while simultaneously being itself simple enough to be interpreted. The way of balancing the optimisation of these two terms is known as the "faithfulness, complexity trade-off" [196]. There are a wide number of "interpretable

models" which can be used as local surrogates, these include linear models or decision trees. In the original formulation of LIME, G is defined as the set of sparse linear models, and $\Omega(g)$ restricts the number of features which can be incorporated into the explanation to be less than a user specified parameter K, (an enforcement of sparsity). While this selection of the complexity measure Ω renders the exact computation of Equation 3.1 intractable, it is approximated via a K-Lasso procedure which is further detailed in Section 3.2.7. While the local application of surrogate explainers reduces the complexity of the decision subspace being approximated, the specification of G enforces certain assumptions on how we expect the model to behave locally.

In this section, we have specified the overarching objective of local surrogate explainers when used as post-hoc explanations. We have shown that local surrogates optimise the faithfulness of the model while minimising the complexity.

3.2 Building Blocks Of LIME

In Section 3.1.2 we have shown how local surrogate models are characterised by the way in which they balance the faithfulness and complexity of a surrogate. We have discussed the way in which LIME minimises complexity $\Omega(g)$. In the following section we introduce the three central components of LIME which together, form the term $L(f, g, \pi_{\mathbf{x}})$ from Equation 3.1. These building blocks therefore signify how LIME ensures the faithfulness of the resulting surrogate model. Sections 3.2.1 to 3.2.8 explicate previous work applying LIME to tabular, text and image data and may be skipped. Section 3.2.9 introduces the challenge of adapting LIME for time series data and begins the novel contribution of this chapter.

3.2.1 Conceptualisation Of The Input Space

In LIME, the original feature space is transformed into an interpretable representation which is dependent on data type. Interpretable representations provide the backbone for many XAI approaches. They translate the "language" (actual features) of the underlying model, which are often low-level data representations of high-dimensional objects, into high-level representations which are semantically meaningful to humans.

Consider an image classification model, where the original feature space contains a large number of pixels, which individually, are meaningless to a user, i.e pixel 145 was most important for the classification of this image as a dog. An interpretable representation of this data structure would group individual pixels into spatially related visual concepts (ears or eyes of a dog) which would result in explanations of the form "the ears in this image were the most influential feature in its classification of a dog". The transformation into an interpretable representation, which we term the *conceptualisation* of an input, allows the explainability mechanism to return the

statistical and mechanical reasons for a particular outcome in terms an end-user can understand and interpret. Many of the explainability mechanisms which depend on conceptualisation assume that the resulting interpretable representations are independent of each other [223]. As such, how a datatype is conceptualised encodes assumptions about how we believe the underlying function behaves and thus controls the kind of investigative question the explanation mechanism targets. By varying the kind of interpretable representation, we control the kind of information captured by the explanation. While this customisation is beneficial from an interpretability perspective, and also facilitates the model agnosticism of the associated explainability method, the ambiguity underlying conceptualisation have been shown to sometimes produce misleading or non-robust explanations [120].

The conceptualisation of both images and language make more intuitive sense due to the structure of these data structures in comparison to tabular or time series data. This is due to the fact that both these data structures can be viewed as a collection of meaningful objects. For example, a sentence is composed of semantically meaningful words and an image is composed of semantically meaningful visual objects yet for tabular data or time series data, concepts do not naturally arise from the raw data structure.

3.2.2 Conceptualisation Of Text

Text data, at its lowest granularity can be viewed as a collection of letters, which like raw pixels in an image, are assumed to be non meaningful in isolation. With increasing granularity we can view a piece of text as a collection of words, phrases and sentences with varying semantic meaning. For example, given the sentence, “Not all patients survive sepsis, yet for those that do, long term illness is common” how we conceptualise this sentence encodes certain assumptions about the model. Conceptualising the sentence into individual words may lose information regarding word ordering and the information carried by their co-occurrence where distinguishing between “not all patients survive” and “yet for those that do” is intrinsic to the meaning of the given statement.

Most applications of surrogate explainers to text data have used word level concepts as the interpretable representation [143]. Conceptualisation of text data, however, can be achieved in a variety of ways thanks to the rich landscape of natural language representation learning. These conceptualisation methods can include ontology-based groupings where each concept set is a set of correlated terms, words, and concepts which have been semantically encoded [120], or context dependent groupings, where relations mediated by nouns, adjectives, and verbs are conceptualised as a triplet [186]. It is important to note that the conceptualisation method is, in most cases, separated from the grouping behaviour of the underlying model, and as such, while the resulting concepts are understandable, we may be mis-assuming how the model operates. It was shown by Lai et al. [120] how conceptualising at the word level may result in misleading explanations due to the disregard of semantic correlations among words, which is particularly problematic for language models capable of extracting context dependent relationships within

the data. In contrast, higher-dimensional text representations have been linked to redundant and wordy explanations [120]. Under the original LIME specification, for text classification, the interpretable representation is a bag of words where the number of words K is fixed at a constant for all samples to be explained.

3.2.3 Conceptualisation Of Images

In the same way that letters organise themselves naturally into semantically meaningful words and phrases, pixels can be grouped into semantically meaningful visual concepts which are referred to as super-pixels. However, much like the discussion surrounding the granularity of text conceptualisation, the scale of visual concepts encoded by each super-pixel is customisable and again subject to the same dependence assumptions associated with text data. For example, consider a classifier which detects the presence of cancerous lesions from images of lungs. Conceptualising the raw pixel format into regions of the lung is more meaningful to the clinician who can interpret the regional importances relative to their own domain expertise. However, if the conceptualisation is too granular then this could mis-attribute regions which are on their own not identifiable as cancerous but are part of a bigger cancerous region. If the conceptualisation on the other hand is too high-level, then the attribution may not reveal the actual segment of the image which was truly influential.

While a conceptualisation determined by the domain-expert is appealing from an interpretability perspective, the attributions to these may be misguiding due to bias from domain expertise. For example, if the clinician is presented with an attribution which confirms what they already believe to be true about the manifestation of cancer in the lungs then they would not question this attribution despite this high level representation not actually being influential for the machine learning model's prediction in the way they expect.

Like with natural language, the conceptualisation of image data can be achieved in many different ways. For example, conceptualisation can be achieved using edge based methods which prioritise not the semantic meaning of the group of pixels but other statistical properties such as a change in shape or texture. Semantic segmentation strategies instead are those concerned with grouping pixels into meaningful concepts although this often required hard coded knowledge about the concept the images represent. As the segmentation of an image is constructed for an individual instance, the segmentation may vary between attribution images and as such, two similar images with the same prediction may obtain very different attributions in different terms [6].

Surrogate models have been applied most widely to image data when compared to other data types and as such there exists a substantial body of work on "super-pixels" with many different approaches which include graph-based algorithms which treat the image as an undirected graph and partition this graph based on edge-weights which are often computed as color differences or similarities [92], or more recent deep learning inspired segmentation such as that of Yang et al.

[227] who use a fully convolutional network to predict super-pixels on a regular image grid.

Automated image segmentation sometimes exhibits undesirable variance across generated super-pixels of similar images which motivates a locality assumption which is often enforced on image conceptualisation. This assumes that the most granular object in the data structure, i.e pixels for images, are locally dependent such that nearby letters or nearby pixels are assumed to be connected. However, this spatial coherence property may fail to capture or occlude meaningful information about the underlying image. For example, there may be global properties of the image which are occluded when conceptualised. This is particularly problematic for sensitive attributes which are encoded by visual properties of the image. How to conceptualise therefore has applications within the fairness literature [218].

3.2.4 Conceptualisation Of Tabular Data

While the dimensionality of the raw input space of both image and text requires an interpretable representation to meaningfully interpret the explanation, this may not be the case for tabular data. For example, if there are a small number of discrete features used to train the model then these are already interpretable. However, there are many cases where an interpretable representation of tabular data is useful, particularly for continuous features. As is the case with image and text conceptualisation, for tabular data we are interested in grouping feature values into concepts which can be either "on" or "off". One approach is to treat each feature value in the given instance to be explained as an "on" concept. While this makes sense if the feature is categorical, treating each potential value of a feature as an individual concept does not make sense over a continuous domain. In this case, it is common to use the input data distribution to discretise the original feature domain into discrete bands. For example, if a temperature was used as a feature, a discretisation of this variable would conceptualise it into the concepts, "between 24 and 25" or "between 25 and 26". In the original implementation of LIME, the dataset $\mathbf{X}_{input}$ is required as input for the tabular setting. The boundaries of the bins are obtained as the quantiles of $\mathbf{X}_{input}$ across each dimension such that for p bins, $1/p$ of the data is associated with each bin. Each bin which contains the feature value in the example to be explained is considered an "on" concept.

For tabular data, unlike the conceptualisation for image and text data which is inherent to the particular instance under consideration, after being conceptualised, the instance to be explained in its binary form may also describe other samples which also fall into the discretised bin. The impossibility to differentiate samples belonging to the same discretised region of the decision space in the binary interpretable representation reduces the surrogate explainer's ability to approximate the local behaviour of the black-box classifier [196]. Since the explanations generated by the surrogate model rely on its ability to approximate the local behaviour of the underlying model, care should be taken when discretising the continuous features such that the characteristics of the local neighbourhood are well modelled [196].

In this section we have introduced the first building bloack of LIME as the specification of a *conceptualisation* method whereby the raw input is transformed into an interpretable representation.

3.2.5 Sampling Hypothetically Via Concept Removal

The conceptualisation strategies discussed above result in the following process for the local surrogate explainer. First, the original sample $\mathbf{x} = \{x_1, ..., x_n\}$ is conceptualised into the binary vector $\sigma(\mathbf{x}) = \{\sigma(\mathbf{x})_1), ..., \sigma(\mathbf{x})_{n'}\} = \{1, ..., 1\}$ where each $\sigma(\mathbf{x})_i = 1$ represents that the i'th interpretable concept of the original instance is turned "on". In this section we introduce the second building block of LIME as the selection of the method by which concepts are removed from the instance to be explained. Following conceptualisation, local surrogate explainers generate hypothetical samples within the local neighbourhood of the instance to consider counterfactually, "What would have happened to the prediction if this concept had been removed from the input instance?". From this counterfactual perspective we assume that there will be a large enough variation over the prediction distribution of counterfactual samples to be meaningfully interpreted in terms of influential concepts.

LIME creates a local neighbourhood around the sample to be explained by sampling d samples over the interpretable representation domain such that each $\sigma(\mathbf{z}_i) = \{\sigma(\mathbf{z}_i)_1, ..., \sigma(\mathbf{z}_i)_{n'}\}$ for all $i \in \{1, ..., d\}$ such that $\sigma(\mathbf{z}_i)_j) \in \{1, 1\}$ for all $j \in \{1, ..., n'\}$ Once d samples have been generated these are then converted back into the original input domain to generate the neighbourhood dataset upon which, the original model's predictions are obtained.

Given that the original instance, in its conceptualised form, is a vector of ones indicating that each concept is turned on, to convert each $\sigma(\mathbf{z}_i)$ into the original domain, each individual concept where $\sigma(\mathbf{z}_i)_j = 1$ is taken as the corresponding feature value in the original instance such that $\mathbf{z}_j = \mathbf{x}_j$. Each individual concept within an individual sample vector where $\sigma(\mathbf{z}_i)_j = 0$ requires that the concept and associated group of raw feature values be "turned off". For text data, individual concepts (words, phrases) can be removed from the input text naturally, resulting in samples $\mathbf{z}_i$ which can be fed back into the original model (assuming these language models can handle arbitrarily sized input).

For image data, concepts (parts of an image) cannot be removed from an input as the pre-trained model requires an input of fixed size n. A such, the removal of super-pixels must be approximated via an occlusion strategy. This poses several challenges for the trustworthiness, robustness and computational soundness of the resulting explanations as well as their consistency with the explainee's intuition. Similarly, for tabular data, feature values cannot simply be dropped from a sample as the underlying function is trained on a fixed input size n. The occlusion strategies for tabular data are even more complex than that of an image due to the ambiguity in selecting a

non-informative value for a continuous variable.

3.2.5.1 Occlusion For Images

When occluding concepts of an image, the overriding intuition is to remove all the salient information represented by that concept in the original instance. The most commonly applied approaches for occlusion within the context of local explanations include:

- **Whiting Out:** The replacement of all the pixel values contained within the super-pixel with a constant value. This approach was adopted by Zeiler et al. [231] who identified the most salient parts of an input image for a neural network's classification by systematically occluding different portions of the input image with a grey square, and monitoring the output of the classifier.
- **Averaging:** The replacement of all pixel values contained within the super-pixel with the expected pixel value taken over the distribution of pixel values belonging to that super-pixel. This approach was adopted by the original LIME specification [174].
- **Random Noise:** Randomly generated Gaussian noise is added to the region characterised by the super-pixel. This approach was adopted by Fong et al. [60].
- **Blurring**: The Gaussian blur kernel is applied to the region characterised by the super-pixel. This approach was adopted by Fong et al. [60]

The Average occlusion strategy used by both the original LIME formulation [174] and SHAP [138] has been associated with undesired effects which make the occlusion redundant [196]. Under Average occlusion, super-pixels that have a relatively uniform colour across the individual pixels they contain may, effectively, be impossible to remove; this is especially common for fragments that are in the background or out of focus [196]. The size of super-pixel also influences the ability of Average occlusion to remove salient information from that region. The smaller the super-pixel, the more likely it is that nearby pixels share the same color. Thus, after applying Average occlusion, we may find that the expected value retains salient information due to the "mosaic effect" exhibited by this occlusion strategy [196].

It was noted by Dabkowski et al. [42] that both the Averaging and Constant occlusion strategies result in super-pixel masks which are non-smooth and as such introduce "adversarial artifcats" into the sample, which may be imperceptible to humans but obstruct the true influence of super-pixels. Following this, the Blurring and Noise Occlusion strategies were introduced by Fong et al. [60] to counteract the introduction of artifacts. Noise occlusion makes the resulting sample more unpredictable for hypothetical samples with a small number of super-pixels removed and therefore reduces the probability of artifacts. Similarly, Blurring occlusion generates hypothetical samples with less high-frequency artifacts. Even with Blurring and Noise occlusion,

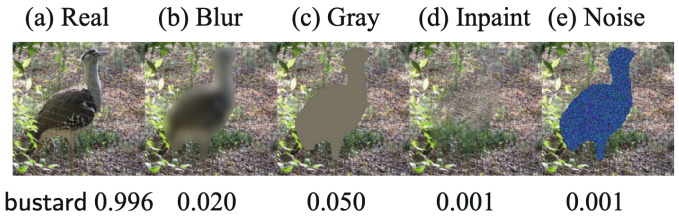


Figure 3.1: Figure taken from Agarwal et al. [3] which shows the effect of four different occlusion strategies, Blur (b), Gray (c), Inpaint (d) and Noise (e), applied to the original image (a). Each occlusion strategy is accompanied by its associated predicted class probability.

adversarial artifacts may still occur and both strategies have demonstrated an inability to remove salient class information [3]. Dabowski et al. [42] motivate the importance of generating smooth occlusion masks, and propose an occlusion strategy which enforces smoothness by minimising the mask variance such that the contrast between "on" and "off" pixels is restricted enough to reduce the "mosaic effect".

Even with smoothness, each occlusion strategy above can result in a generated sample which is visually unrealisitic with regards to the natural distribution of images. Figure 3.1, taken from Agarwal et al. [3] exemplifies how each occlusion strategy (Blur, Noise and Gray (Constant)) results in an image which is visually unrealistic. When evaluating the function on out of distribution samples, it has been shown that the resulting predictions can be misleading due to the evaluation of the function on off-manifold samples upon which the model was not trained [3]. Agarwal et al. [3] propose an occlusion strategy which utilises an Inpainter [229] to generate realistic, in-distribution hypothetical samples. Inpainters are generative models remove pixels from an input image and fill in with content that is plausible under the true data distribution. The result of the inpainting occlusion strategy is shown in Figure 3.1 which visually creates a more realistic sample while also successfully removing the salient information from the sample.

3.2.5.2 Occlusion For Tabular Data

Within the local surrogate framework, which relies on the binarised transformed feature set, we have discussed how for tabular data, the conceptualisation of the given input is not necessary but can be useful. As such there are two approaches such that either the raw feature values in the given instance are considered as the "on" concepts, or the raw feature values have been "conceptualised" into binary or discrete concepts which are all present in the outcome to be explained. How to remove these concepts from a feature vector is is an open challenge within the XAI community, not just limited to local surrogate explainers and is discussed at length in Chapter 4 in this thesis. However, here we frame the feature removal challenge from a surrogate explainer perspective and relate the challenges with those of image occlusion discussed. In the original LIME formulation, we have discussed in Section 3.2.4, how continuous features are

discretised into p quantiles reflecting the input data distribution. Each bin index of a given sample gives the quantile that a particular feature value falls into. To generate a single sample $\sigma(\mathbf{z}_i)$, LIME samples bin indices randomly, where the quantiles are distributed uniformly on $\{1,...,p\}$ for each dimension. As LIME requires the binarisation of concepts, if a generated bin index $j \in \{1,...,n\}$ is the same as that in the original sample to be explained then $\sigma(\mathbf{z}_i)_j = 1$ otherwise $\sigma(\mathbf{z}_i)_j = 0$

To then convert the hypothetical sample into the real domain, LIME uses each sampled bin index to generate the hypothetical instance $\mathbf{z}_i$ which is sampled dimension by dimension. Each $z_j \in \mathbf{z}_i$ is distributed according to a truncated Gaussian random variable which is characterised by the mean and variance of the associated quantile. In this way, the sampling of the hypothetical samples does not depend directly on the example to be explained but on the discretised example, therefore we can see that two potential examples to be explained may generate exactly the same samples as long as their feature values fall into the same quantiles.

Problems associated with image occlusion strategies are even more troublesome for tabular data. Unlike image data, for tabular data, we are unable to visually compare the resulting sample with the original sample in regards to the realism of the hypothetical sample. Furthermore, the adversarial artifacts created by replacing individual features with their "missing value" are even harder to detect on tabular data.

From the above discussion it has become apparent that despite the assumed modularity of the building blocks which underpin the local surrogate explainer framework, it is actually imperative to consider the assumptions made by the conceptualisation and the occlusion strategies in tandem. Below we discuss the assumptions made by the LIME algorithm when defining a local neighbourhood around an instance and argue that this strategy is also intrinsically connected to conceptualisation and concept removal.

In this section, we have introduced the second building block of LIME as the way by which concepts are occluded from the input to be explained. We have discussed the differences between occlusion for images and for tabular data.

3.2.6 Local Neighbourhood Weighting

In this section we introduce the third building block of LIME as the way in which a local neighbourhood is contructed around the sample to be explained. Once the hypothetical samples $\mathbf{Z} = \{\mathbf{z}_1,...,\mathbf{z}_d\}$ and their corresponding labels given the underlying model $f(\mathbf{Z})$ have been generated, a surrogate model is trained using these samples and labels which approximates the behaviour of the underlying black-box in the local neighbourhood around the example to be explained. Prior to this, however, a key component of the LIME algorithm is the weighting of the hypothetical samples to enforce the notion of locality such that samples which are more similar

to the original instance are more heavily weighted in the resulting surrogate model. Integral to the selection of a similarity kernel is the definition of a "meaningful" neighbourhood around the instance to be explained.

Under the original LIME specification, the distances between the hypothetical samples and the original instance in the interpretable representation are calculated under a specified distance measure e.g., the Manhattan, Euclidean or Cosine distance depending on data-type [196]. Then, these distances are transformed into a weight vector via the use of a similarity kernel. According to the original paper [175], the specification of a distance measure over the interpretable domain has to be a distance function that is intuitive to the user, i.e. the user should be able to grasp if two examples are far or close given the intuitive representation. The use of a particular distance measure, however, encodes certain assumptions about how two samples may be related in high-dimensional space. For example, for tabular data, the Manhattan distance is known to encourage sparsity [216], and is considered a better metric than the Euclidean for high-dimensional data. Furthermore, for image data, it was shown by Garreau et al. [67] that the Cosine distance is dependent only on the **number** of inactivated super-pixels which may not capture the fact that removing certain super-pixels may have varying impact on the similarity between original and generated samples. Concerning the specification of the kernel, the original implementation of LIME uses a Gaussian smoothing kernel to define the neighborhood. When LIME is applied to high-dimensional objects like images, the notion of capturing a meaningful neighbourhood surrounding an instance is made challenging due to the type of high-dimensional spaces within which these data-structures inhibit. Furthermore, it has been shown that while the Gaussian kernel performs well on low dimensional data, when applied to high-dimensional data the notion of similarity itself loses its significance [61]. This means that close distances and large distances are hardly distinguishable from their kernel values. There are variants of the Gaussian kernel which have been adapted for high-dimensional data [61].

The specification of the distance measure and the associated similarity kernel has a significant impact on the resulting explanation generated by LIME. To this end, Sokol et al. [198] argue that data should be weighted according to the particular investigative goal when training the surrogate, for example, using similarity scores computed with kernelised distance between the explained instance and the sampled data could be determined in *either* the intepretable or the original data representation. Saito et al. [181] show how determining the distance in the original domain, which takes into account the original data distribution, produces explanations which are more robust to adversarial attack.

Alternative adaptations of LIME [124] question the assumption that generated samples which are "closest" to the example to be explained should be the most influential in the resulting attribution. They argue that by defining locality this way, LIME does not properly capture the relevant local classification behaviour of the black-box. As a result, the effect of locally important features can be hidden by globally important ones. Laugel et al. [124] show that this issue is not

just related to kernel parametrisation, but actually a problem with centering the sampling on the instance $\mathbf{x}$. Laugel et al. [124] work under the intuition that in order to best approximate a local decision boundary, we should sample around the decision boundary itself. Laugel et al. propose GSLS [124] which directly exploits the structure of the local decision boundary by finding the nearest point of the opposite class and using this instance to centre the hypothetical instances by sampling uniformly within a hyper-ball within the vicinity of this contrastive sample. Under the GSLS approach, a meaningful neighbourhood now characterises the samples which lie in the locality around the closest threshold to the opposite class.

In this section, we have introduced the third building block of LIME as the selection of a measure of locality around the sample to be explained.

3.2.7 LIME

Once the similarity kernel has been selected, LIME trains the weighted surrogate model on the hypothetical samples $\mathbf{Z}$ and their corresponding labels $f(\mathbf{Z})$. Formally this corresponds to the maximisation of the faithfulness of the surrogate model, or the minimisation of $L(f,g,\pi_{\mathbf{x}})$ from Equation 3.1 as specified by LIME.

$$L(f,g,\pi_{\mathbf{x}}) = \sum_{\mathbf{z},\sigma(\mathbf{z})\in\mathbf{Z}} \pi_{\mathbf{x}}(\mathbf{z})(f(\mathbf{z}) - g(\sigma(\mathbf{z})))^2 \tag{3.2}$$

The original specification of the LIME algorithm [175], which we refer to as "out-of-the-box LIME" is displayed in Algorithm 1.

Algorithm 1 LIME

Require: Classifier f, Number of samples d
Require: Instance $\mathbf{x}$, Conceptualised version $\sigma(\mathbf{x})$
Require: Similarity Kernel $\pi_{\mathbf{x}}$, Length of explanation K
  $\mathbf{Z} \leftarrow \{\}$
  for $i \in \{1,...,d\}$ **do**
    $\sigma(\mathbf{z}_i) \leftarrow Sample(\sigma(\mathbf{x}))$
    $\mathbf{Z} \leftarrow \mathbf{Z} \cup \{\sigma(\mathbf{z}_i), f(\mathbf{z}_i, \pi_{\mathbf{x}}(\mathbf{z}_i))$
  end for
  $\mathbf{w} \leftarrow K - Lasso(\mathbf{Z},K)$ where $\sigma(\mathbf{z}_i)$ are features and $f(\mathbf{z}_i)$ as target
  return $\mathbf{w}$

LIME, as formalised by Algorithm 1 returns the attribution vector $\mathbf{w}$. This vector which corresponds to the linear co-efficients of the local surrogate model as specified by LIME which translates into an explanation which informs the end-user of the relative importance of each input feature (in the interpretable domain).

We have thus far decomposed LIME into its three essential building blocks: conceptualisation, concept removal and the definition of a local neighbourhood. Although LIME is intended to be customised, accommodating the explanation of any possible model or data-type, we have discussed the range of assumptions invoked via the selection of each of LIME's components. We therefore motivate the consideration of the entire LIME pipeline, particularly focusing on how the selection of each individual building block has repercussions on the rest of the algorithm. Below we draw attention to the fact that LIME's data-agnosticism, while being one of the main selling points of LIME, can also present challenges when applied to specific data types.

3.2.8 Data-agnosticism: Blessing or a Curse?

The data-agnosticism of LIME has motivated its adoption in explaining AI systems in healthcare [94, 160, 207, 213]. The documented prior success of LIME in generating useful explanations for human experts [175] motivated its use in explaining patient-specific outcomes generated by clinical decision support systems [118]. The explanations obtained with LIME on healthcare tabular systems have been extensively evaluated [71, 118, 207]. Tajgardoon et al. [207] applied LIME to explain a a predictive model determining the risk of pneumonia onset. The model was trained with over 150 variables including demographic information history and physical examination information, laboratory results, and chest X-ray findings. Tajgardoon et al. [207] found that the human reviewers agreed with 78% of LIME-generated explanatory features for actual explanations and agreed with only 52% of explanatory features for fabricated explanations. This result provides evidence that the human reviewers were able to distinguish between valid and invalid explanations.

However, predictive systems in healthcare are renowned for the multi-modality of input data. In this respect, the data-agnosticism of LIME, despite being one of the key motivators of the original algorithm, has been associated with undesirable behaviour when applied to data-types it has not been specifically designed for. Ghassemi et al. [71] have shown that for structured data, electronic health-care records or electroencephalogram waveform data, for example, the overlap between explanations generated by LIME and those generated by domain experts suffers from an interpretability gap. This problem is exacerbated by the conceptualisation stage of LIME such that domain knowledge informs the kind of concepts raw inputs are grouped into. While these concepts are semantically meaningful to the user, they may not accurately reflect the behaviour of the model. For example, although explanations for language tend to revolve around highlighting the words in the text that contributed to the decision, this does not reveal the associative meaning the model has learned for those words. As with heat maps, the human tendency is to assume that a model has used words in the same way we would. However, deeper investigation often reveals that these models rely on unacceptable shortcuts, such as strongly associating the word doctor with maleness and using this reductionist interpretation to inform decision making.

As noted by Tonekaboni et al. [213], post-hoc local explanations such as LIME may not

be directly applicable to clinical settings. Some methods may work well for specific kinds of data, like images, but may have to be non-trivially extended to be applicable to other data. For example, LIME has been used by Palatnik et al. [160] to generate explanations on how a Convolutional Neural Network (CNN) detects tumor tissue in patches extracted from histology whole slide images. Palatnik et al. found that generally, the explanations generated by LIME in this setting align with the expected important image regions as identified by clinicicans. However, Palatnik et al. showed how the granularity of image segmentation employed by LIME impacts the resulting explanation. This result evidences the "interpretability gap" whereby it becomes hard to differentiate the ground truth model behaviour from the use of LIME explanations to justify a pre-existing hypothesis about what the model is learning. Furthermore, it was argued that this problem is especially under-studied for clinical time-series models that heavily employ deep learning methods [51, 213] . Ito et al. [94] applied LIME to health risk prediction from electronic health records (EHRs), a time series prediction application. It was shown by Ito et al. [94] that for EHRs, if a local neighbourhood contains anomalous observations the model will be distorted. From the above discussion, it becomes apparent that as the growing demand for explanations of AI systems extends to specialist data-types it is important to carefully consider the correct components of LIME which are best suited to explain the respective models. In the following section we consider the potential problems which may arise when applying LIME to time series data and propose three building blocks which are designed specifically to permit the application of LIME to time series which result in meaningful and realistic attributions.

In this section, we have discussed some of the problems associated with applying LIME to data structures it has not been designed for.

3.2.9 Adapting LIME to Time Series

In this section we explore the challenges of extending LIME to explain univariate time series classification. As we have outlined in Chapter 2, a univariate time series is a temporally ordered set of t observations $\mathbf{x} = \{x_1, x_2, ..., x_t\}$. Given a time series to be explained $\mathbf{x}$, a black box classifier $f : \mathscr{X} \rightarrow \mathbb{R}$, and the predicted label $y = f(\mathbf{x})$, we are interested in generating a surrogate model g around $\mathbf{x}$ in order to identify the most salient observations in $\mathbf{x}$ with regards to y. For example, if we consider time series pertaining to an individual's activity levels over the course of a week, measured by some wearable device, alongside a machine learning classifier which has been trained to use activity level data to detect depression in individual time series. A time series explanation in this case may resemble a set of indexes where the associated activity level was most influential in the classifier's outcome of depressed.

As discussed in Chapter 1, there is limited literature on time series explainability compared to image and natural language for which a plethora of established frameworks exist. Feature

importance based explanations have limited success when applied to time series as any adaptation requires the consideration of the temporal nature of the input space. LIME has been previously applied to explain time series prediction by Rovzanec et al. and Schlegel et al. [178, 185] where "out-of-the-box" LIME is used and each time step is taken as an individual concept. In [185], the attributions generated by LIME were compared to those generated by alternative local post-hoc methods, including SHAP, where it was found to perform the worst, "most likely because of the large dimensionality by converting time to features". Despite this, out-the-box LIME has still been used to attribute individual observations of time series in applications including forecast models [185] and stock price prediction [170].

There has also been a (limited) recognition in the literature that to apply LIME to time series attributions, changes to the original algorithm must be made to account for the temporal structure of this kind of data [77, 156]. In the following section we discuss how the considerations previously addressed concerning the different building blocks of LIME: conceptualisation, sampling and locality extend to a time series setting. We propose desiderata for these building blocks when constructing LIME for time series and present our three proposed solutions to this challenge, ultimately culminating in our algorithm, LIMESegment, which extends the original LIME algorithm to the time series setting to produce meaningful, realistic explanations.

3.3 Conceptualising A Time Series

The need to transform a time series into an interpretable representation follows the same intuition as images where an explanation involving a single observation would be non-useful to an end user and fail to capture salient properties of the time series. It follows that this transformation requires segmenting the time series into a lower dimensional representation. However, a time series does not lend itself naturally to conceptualisation in the same way that super-pixels intuitively conceptualise images as there is no visually obvious grouping of observations into semantically meaningful concepts. In their respective adaptions of LIME to time series, Guilleme et al. [77] and Neves et al. [156] use arbitrarily determined, fixed length time slices as the "concepts" to be input to the LIME algorithm. Mujkanovic et al. [153] argue that when conceptualising a time series, we can either adopt a segmentation in the time or frequency domain whereby each approach is accompanied by its own set of assumptions.

Time slice conceptualisation makes the following two assumptions implicitly [153].

- **Temporal space assumption**: The underlying model f learns patterns in the data solely in the time domain [153].

- **Temporal coherence assumption**: Temporally local observations have a similar impact on the predictions of the model [153].

For our earlier activity level example, a time slice conceptualisation would makes the assumption that the depression classifier uses patterns in the time domain to make decisions. For example, it may recognise that decreased activity levels at the weekend are indicative of depression. Furthermore, a time slice conceptualisation would assume that temporally local activity level observations, for example, observations in the same hour would impact the model in the similarly, allowing us to conceptualise observations into hourly segments. Many time series models seem to fulfill these assumptions, however, some time series estimators may not only work on the temporal structure of a time series but also examine its frequency spectrum. This might even be done by an explicit conversion of the series from time domain to the frequency domain via the Fourier transform.

Many time series models may adopt the time domain assumptions however, time series analysis and modelling may also be performed in the frequency domain.

While time slice mapping allows us to compute interpretable impacts of temporal slices of a time series, it fails to capture any frequency bands that may be considered impactful by some models [153]. Frequency domain conceptualisation makes two analoguous assumptions to those for the time domain [153].

- **Feature space assumption**: The underlying model learns patterns in the frequency domain of the data [153].
- **Frequency coherence assumption:** Local frequencies have a similar impact on the predictions of the model [153].

For our running activity level example, the depression classifier may have learned patterns in the frequency domain which it uses to make decisions. For example, it may have learned that high frequency components (all signals can be decomposed into sinusoids of varying frequency) are indicative of depression. In this case, if we were to conceptualise a time series into hourly segments, our explainer would not be able to correctly identify the influential signal component as these high frequency bands persist over the global signal. Under frequency coherence therefore we would conceptualise a time series into local frequency bands which would capture the behaviour of the classifier in the frequency domain.

3.3.1 Why Is Conceptualisation A Challenge?

Whether we choose to conceptualise in the time or frequency domain enforces assumptions on how we expect the underlying model to be learning patterns in the data.

When occluding a segment in the time domain, for example, we could occlude a spike in temperature recorded as a part of a patient trajectory from the MIMIC Sepsis Cohort, the removal will have a knock-on effect in the frequency domain as well. The removal of a significant spike in temperature will lower the magnitude of the associated frequency band [153]. If the

model had learned that the global behaviour of this frequency, rather than the individual spike, was important, then the resulting explanation, which would indicate that the spike was important for the resulting classification, would be misleading.

Arguably the most important consideration when considering frequency vs. temporal conceptualisation is the interpretability of the resulting concepts. It is unclear as to whether frequency band concepts would provide meaningful explanations to an end-user. In contrast, conceptualistaion in the time domain aligns more naturally with how humans perceive time series data as the progression of a concept chronologically through time.

Assuming coherence alleges that neighbouring values have a similar impact on the model's prediction, both in the time and frequency domain [153]. Just as we have argued for super-pixels of images, the granularity of the resulting time series series influences the ability of the explainer to make an accurate approximation of the reasons a model makes a decision. If the segments are too small, they run the risk of being non-interpretable, or, not being realistic representations of how the model learns patterns. However, if the segments are too large, they may not identify the salient information in the time series [153].

From the above discussion it is clear that the approach for conceptualising a time series into interpretable concepts is non-trivial. We therefore propose the following open question, which we argue any approach to developing local post-hoc explanation for time series should consider and explicitly justify their approach. **Open Question 1:** *"How do we meaningfully conceptualise a time series into an interpretable representation where each concept corresponds to an impactful fragment?"*

3.3.2 Nearest Neighbour Segmentation

To address Open Question 1, we return to the overarching goal of LIME, in that it generates attributions in terms of meaningful concepts that the end user can understand. To this end we adopt a temporal conceptualisation approach, decomposing a time series into semantically meaningful *super-segments*. When considering the extraction of impactful fragments, intuitively, we would want the super-segments of a time series to capture homogeneous regions of behaviour. For example, given a time series recording an individual's activity over the period of a day, super-segments could correspond to the various activities (sleeping, walking). An arbitrary segmentation may result in non-homogeneous segments with conflicting properties or homogeneous regions spanning multiple segments.

There are many alternative approaches to time series segmentation which attempt to minimise the artifacts induced by the temporal coherence assumption. Change point analysis decomposes a time series using changes in statistical properties of the time series [2, 57]. In contrast, semantic segmentation uses the shape of a time series to segment [70]. We propose a segmentation algorithm which uses both the shape *and* the statistical properties of a time series to identify change points.

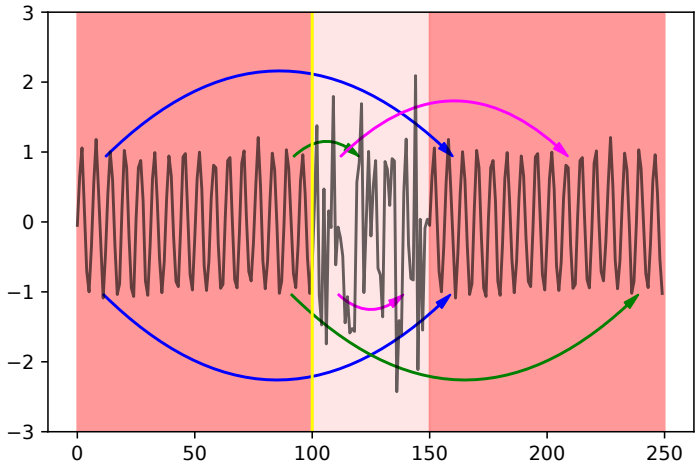


Figure 3.2: Figure shows the intuition behind NNSegment. time series composed of motifs (shaded deep red) and anomalies (shaded pale red). Arrows connect current window with its nearest neighbour. Blue arrows at indexes 10 and 11 represent windows where adjacency holds. Magenta arrows at indexes 110 and 111 indicate adjacent windows where adjacency is broken. In this case $ws = 10$ and $\rho(\mathbf{w}_{100}, \mathbf{w}_{110}) > \rho(\mathbf{w}_{110}, \mathbf{w}_{120})$, indicating a cp at the beginning of the window, at index $i = 110$ which is added to the set of potential change points **cp**. Green arrows at indexes 90 and 91 also indicate windows where adjacency is broken. In this case, $\rho(\mathbf{w}_{90}, \mathbf{w}_{100}) > \rho(\mathbf{w}_{80}, \mathbf{w}_{90})$ thus $i = 100$ is added to **cp**. In this example, $\rho(\mathbf{w}_{90}, \mathbf{w}_{100}) > \rho(\mathbf{w}_{100}, \mathbf{w}_{110})$ implying that the cp at $i = 100$ is more likely than that at $i = 110$.

Given a time series $\mathbf{x}$ of length t, our segmentation algorithm, NNSegment returns a set of change points $\mathbf{cp} = \{cp_1, cp_2, ..., cp_{m'}\}$, where m' is the number of change points in $\mathbf{x}$. Each $cp_j = i$ indicates the temporal index i of $\mathbf{x}$ where there is a change in behaviour. NNSegment is built on the assumption that neighbouring observations are likely to represent the same behaviour. We can thus group these neighbouring observations together into super-segments which may be either repeating motifs or randomly occurring anomalies. Our segmentation approach is based on the assumption that regularly occurring motifs will be similar in shape. To identify "similar" fragments of the time series we use normalised cross correlation (Definition 3.1) which has had much success on pattern matching tasks [235].

Definition 3.1 (Cross Correlation)**.** Given a time series $\mathbf{x} = \{x_1, ..., x_t\}$, the similarity between two sub-sequences of length ws is denoted as $\psi(\mathbf{x}_{s1}, \mathbf{x}_{s2})$ where $\mathbf{x}_{s1} = \{x_{s1}; ...; x_{s1+ws}\}$ and $\mathbf{x}_{s2} = \{x_{s2}, ..., x_{s2+ws}\}$, can be defined as the normalised cross correlation function such that

$$\psi(\mathbf{x}_{s1}, \mathbf{x}_{s2}) = \frac{\mathbb{E}[\mathbf{x}_{s1} - \mu_{\mathbf{x}_{s1}}][\mathbf{x}_{s2} - \mu_{\mathbf{x}_{s2}}]}{\sigma_{\mathbf{x}_{s1}} \sigma_{\mathbf{x}_{s2}}}. \tag{3.3}$$

Here, μ and σ represent sub-sequence mean and variance respectively.

NNSegment first decomposes $\mathbf{x}$ into overlapping windows of size ws which are then used to identify super-segments in $\mathbf{x}$. $\mathbf{x}$ decomposed into its windowed representation is the set $\mathbf{w} = \{\mathbf{w}_1;...;\mathbf{w}_{t-ws}\}$. For each window $\mathbf{w}_i = \{x_i,...,x_{i+ws}\}$, the index of its nearest neighbour $w_z(i)$ is determined by finding the window with which it shares minimal cross correlation: $w_z(i) = argmin_j(\psi(\mathbf{w}_i\mathbf{w}_j))$. After finding the nearest neighbours of all windows, our intuition, demonstrated in Figure 3.2, is that adjacent windows which belong to a homogeneous region of behaviour will follow an adjacency pattern (Definition 3.2). When adjacency is broken, we assume the behaviour of the time series has changed and the corresponding index represents the end of a super segment.

Definition 3.2 (Adjacency)**.** Given a time series $\mathbf{x}$ and two adjacent windows at index i and $i+1$, adjacency holds at window index i if $w_z(i)+1 = w_z(i+1)$.

For anomalies, adjacent windows are likely to break the adjacency property. However, this does not necessarily signify the end of the super-segment. To account for these potentially erroneous change points we include a normalisation term which uses the statistical properties of the preceding and following window to determine if a true change point has occurred. We define the difference in statistical properties between two windows as

$$\rho(\mathbf{w}_i, \mathbf{w}_j) = |(\frac{\mu(\mathbf{w}_i)}{\sigma(\mathbf{w}_i)} - \frac{\mu(\mathbf{w}_j)}{\sigma(\mathbf{w}_j)})|. \tag{3.4}$$

Here μ and σ represent window mean and variance respectively. For each window $\mathbf{w}_i$ which breaks the adjacency property, we calculate the values of $\rho(\mathbf{w}_i, \mathbf{w}_{i-ws})$ $\rho(\mathbf{w}_i, \mathbf{w}_{i+ws})$. If $\rho(\mathbf{w}_i, \mathbf{w}_{i-ws}) > \rho(\mathbf{w}_i, \mathbf{w}_{i+ws})$ the change point is more likely to have occurred at the beginning of $\mathbf{w}_i$ whereas if $\rho(\mathbf{w}_i, \mathbf{w}_{i-ws}) < \rho(\mathbf{w}_i, \mathbf{w}_{i+ws})$ the change point is more likely to have occurred at the end of $\mathbf{w}_i$. We sort the vector $\rho(.)$ in increasing order and return the set of m' indexes corresponding to the most likely change points $\mathbf{cp}$. Under the assumption that every observation $x_i \in \mathbf{x}$ either forms part of a motif or anomaly, we formally introduce $NNSegment$ in Algorithm 2 which requires parametrising with window size, ws and the number of user-specified change points m'.

3.3.3 Experimental Validation of NNSegment

To evaluate NNSegment in addressing Open Question 1, we use Hausdorff Distance (Equation 3.5) and F-Score (Equation 3.6). The Hausdorff Distance (HD) is the greatest temporal distance between a true change point $\mathbf{cp}_i^r$ and the predicted change point $\mathbf{cp}_i^p$ as returned by a segmentation algorithm for a given time series of length t [8].

$$HD(\mathbf{cp}^r, \mathbf{cp}^p) = \frac{max(\mathbf{cp}_i^r - \mathbf{cp}_i^p)}{t} \forall i \in \{1,...,|\mathbf{cp}^r|\} \tag{3.5}$$

The F-score is the harmonic mean of the Precision and Recall of the predicted change points.

Algorithm 2 Nearest Neighbour Segment (NNSegment)

Require: TS $\mathbf{x}$, window length ws, number of Change Points m'
  $\mathbf{w} \leftarrow \{\mathbf{w}_i\}$ for $i \in \{1,..,t-ws\}$
  $\mathbf{w_z} \leftarrow argmin_j(\psi(\mathbf{w}_i\mathbf{w}_j))$ for $i,j \in \{1,...,t-ws\}$
  $\mathbf{cp} \leftarrow \{\}$
  for $i \in t-ws$ **do**
    if $w_z(i+1) \neq w_z(i)+1$ **then**
      if $|(\rho(\mathbf{w}_i, \mathbf{w}_{i-ws}))| > |(\rho(\mathbf{w}_i, \mathbf{w}_{i+ws}))|$ **then**
        $\mathbf{cp} \leftarrow \mathbf{cp} \cup \{i\}$
      end if
    else
      $\mathbf{cp} \leftarrow \mathbf{cp} \cup \{i+ws\}$
    end if
  end for
  $\mathbf{cp} \leftarrow sort(\mathbf{cp})$ by $\rho(.)$
  return $\mathbf{cp} \leftarrow \{cp_1, cp_2, ..., cp_{m'}\}$

$$FScore(\mathbf{cp}^p, \mathbf{cp}^r) = \frac{\mathbf{cp}^r \cap \mathbf{cp}^p}{\mathbf{cp}^r \cap \mathbf{cp}^p + \frac{(\mathbf{cp}^r \setminus \mathbf{cp}^p + \mathbf{cp}^p \setminus \mathbf{cp}^r)}{2}}. \tag{3.6}$$

To evaluate both F-score and Hausdorff distance we assume access to univariate time series datasets which are annotated with true change point indexes. For our first dataset we construct a synthetic colllection of univariate time series which we construct by concatenating six super segments as per the method in Algorithm 3. Each segment and its respective frequency composition is taken to represent a homogeneous region of activity, as such the true change points are located at indexes $\{100, 200, 300, 400\}$ for every time series t we generate. For our second annotated univariate time series dataset we use examples from the Apnea-ECG dataset [166] for evaluation. The Apnea dataset contains ECG recordings of 70 participants with labelled apnea events. Here, our ground truth change points are the temporal indexes which are labelled as either the start or end of an apnea event.

Algorithm 3 Synthetic Time Series Generation for Segmentation Evaluation

  $\mu, \sigma \leftarrow 10, 0.2$
  $s1 \leftarrow N(\mu, \sigma, 100)$
  $\mu, \sigma \leftarrow 10, 0.03$
  $s2 \leftarrow N(\mu, \sigma, 100)$
  $\mu, \sigma \leftarrow 10, 0.1$
  $s3 \leftarrow N(\mu, \sigma, 100)$
  return $\mathbf{x} \leftarrow \{s1, s2, s3, s2, s1\}$

We compare NNSegment with FLUSS, the semantic segmentation algorithm closest to our work [70]. Table 3.1 shows how NNSegment outperforms FLUSS across both metrics and both

	NNSegment	FLUSS
F-Score		
Synthetic	**0.90**	0.00
Apnea	**0.42**	0.16
HD		
Synthetic	**0.05**	0.76
Apnea	**0.30**	0.40

Table 3.1: Table presents the F-score and HD obtained by NNSegment and FLUSS when applied to the Synthetic and Apnea Datasets. Higher F-score, lower HD reflects better segmentation.

datasets evaluated. We attribute the superiority of NNSegment over FLUSS to the difference in intuition driving both algorithms. Unlike NNSegment, FLUSS assumes all similar behaviour occurs in the same segment and fails to take into account for repeating patterns.

In this section we have explained why conceptualising a time series is challenging. We introduced our novel algorithm NNSegment for time series conceptualisation. NNSegment decomposes a univariate time series into a vector of motif and anomaly segments. We compared the ability of NNSegment with that of FLUSS, a state of the art time series segmentation algorithm, in generating a meaningful segmentation of a time series.

3.4 Time Series Occlusion

After conceptualising a time series into its interpretable super-segment representation it is then a challenge to generate new samples in its local neighbourhood. LIME does this by "turning off" concepts. To apply this intuition to super-segments, we must specify what it means to delete information, or occlude, from a time series. Occlusion of a time series suffers from the same challenges as both image and tabular data, as unlike natural language, segments (in the time domain) cannot just be removed from the time series as the underlying model is trained on fixed input size. As discussed in Section 3.2.5, occlusion approaches for image data include replacing the super-pixel with some constant value, injecting noise, or blurring the image.

These techniques can also be applied to time series where Neves et al. [156] replace segments with mean valued segments and Guilleme et al. [77] replace segments with randomly selected authentic segments from the original dataset. However, as is the case with image super-pixels we are generally interested in simulating in-distribution time series samples, leading to more realistic perturbations. Figure 3.3 shows the application of blur perturbation to a time series. Unlike perturbed images, where we can visually inspect the artifacts induced on the image via blur perturbation (Figure 3.1), there is no way of visually validating the realism of the resulting sample in Figure 3.3. Furthermore, there is a lack of literature surrounding the extension of the

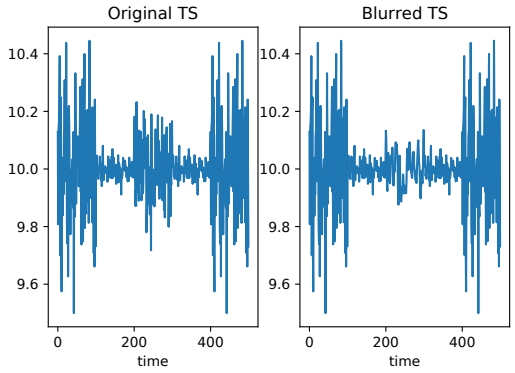


Figure 3.3: Figure shows the effect of applying blur (via a Gaussian Filter) to an example super-segment located at index $\{200:300\}$ (Original TS). From the figure, we argue that unlike image occlusion, we are unable to confirm visually whether the resulting perturbed time series, (Blurred TS), is realistic.

generative inpainting process to time series. While time series inpainting is addressed by Pires et al. [168], the resulting inpainting procedure is intended to fill in the gaps, or impute, associated time series. In this sense, the in-painting incorporates, rather than removes, salient information to the time series. We therefore raise the following open question **Open Question 2:** *"How do we remove salient super-segment information while retaining the realism of the resulting time series?"*

3.4.1 Background Frequency Perturbation

To address Open Question 2 we employ findings from harmonic analysis: Any time series $\mathbf{x}$ can be represented as a composition of harmonic oscillations in the frequency domain. Maxima in the frequency domain reflect a high proportion of the signal oscillating at that frequency. It has been shown that realistic background content represents a global property of an image and is not necessarily the local low frequency content but the most commonly occurring global frequency information [3]. A time series' frequency distribution varies considerably over time. Applying a low pass filter to, or blurring a segment, is therefore not necessarily a true reflection of removing the salient signal content. Instead we propose replacing a super-segment with a background content segment, artificially generated by identifying the frequency band which has the highest representation, with lowest variance in the original signal.

To understand how the spectral density of a time series changes over time we use the Discrete Short Time Frequency Transform (STFT). The STFT is a Fourier-related transform used to determine the sinusoidal frequency and phase content of local sections of a signal as it changes over time. Given a time series $\mathbf{x}$, the STFT converts $\mathbf{x}$ into its time-frequency representation by taking the Fourier transform of $\mathbf{x}$ multiplied by a sliding window of length ws.

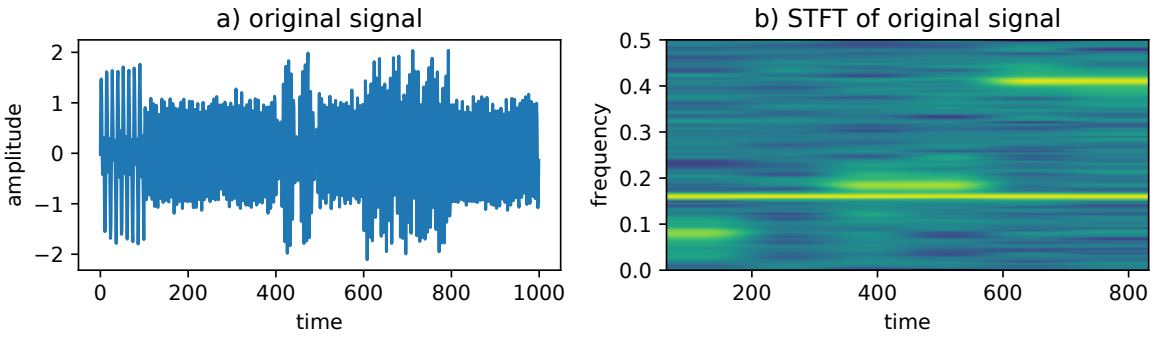


Figure 3.4: Figure shows the intuition behind *RBP*. The original signal (a) is composed of background signal and varying frequency sine waves at indexes: [0 : 100], [400 : 500] and [600 : 800]. b) shows the spectrogram obtained by applying STFT to the original signal. The spectrogram captures the background signal which remains constant through time as well as the shorter length "content" sine waves at their respective frequencies.

$$STFT(\mathbf{x}, ws, \omega) = \sum_{t=-\infty}^{\infty} \mathbf{x}\rho(t - ws)DFT_t \tag{3.7}$$

Here, $DFT_t = e^{-i\omega t}$ represents the Fourier transform. ω is the frequency parameter and $\rho(.)$ is a window function parametrised by window size ws.

$STFT(\mathbf{x}, ws, \omega)$ is a complex function representing the phase and magnitude of the signal over time and frequency. To obtain only the magnitude component of $STFT(\mathbf{x}, ws, \omega)$ we compute $|STFT(\mathbf{x}, ws, \omega)|$. This results in a matrix whereby f_t indicates the magnitude of frequency band f at temporal index t. For our perturbation algorithm we are interested in filtering the signal by selecting only the background content. Given $|STFT(\mathbf{x}, ws, \omega)|$ we find the most persistent frequency by identifying the frequency band which has the highest value over time with minimal variance

$$F_{persist} = \arg\max_{f} \frac{\mu(f_t)}{\sigma(f_t)} \forall f, t \in |STFT(\mathbf{x}, ws, \omega)| \tag{3.8}$$

Here $\frac{\mu(f_t)}{\sigma(f_t)}$ is the mean magnitude response normalised by its standard deviation of a selected frequency band over time. To use this background content to meaningfully perturb our original time series we convert $F_{persist}$ into the original time domain via the inverse STFT, from which, the relevant segments of background content can be chosen to replace parts of the original signal. Our realistic background perturbation algorithm, *RBP* is shown in Algorithm 4 and Figure 3.4 shows the STFT applied to an example time series, demonstrating our background frequency intuition.

Algorithm 4 Realistic Background Perturbation (RBP)

Require: TS $\mathbf{x}$ of length t, window size ws, frequency parameter ω, change point indexes **cp**, perturbation segments $\sigma(z)$

$\mathbf{x}_{stft} \leftarrow STFT(\mathbf{x}, \omega, ws)$

$F_{persist} \leftarrow \mathbf{x}_{stft}[argmax_f \frac{\mu(|f_t|)}{\sigma(|f_t|)}]$ for all $f, t \in \mathbf{x}_{stft}$

$R = STFT^{-1}(F_{persist}, \omega, ws)$

$\mathbf{z} \leftarrow \mathbf{x}$

for i in T' **do**

  if $\sigma(\mathbf{x}_i) == 0$ **then**

    $\mathbf{z}[cp_i : cp_{i+1}] \leftarrow R[cp_i : cp_{i+1}]$

  end if

end for

return perturbed time series $\mathbf{z}$

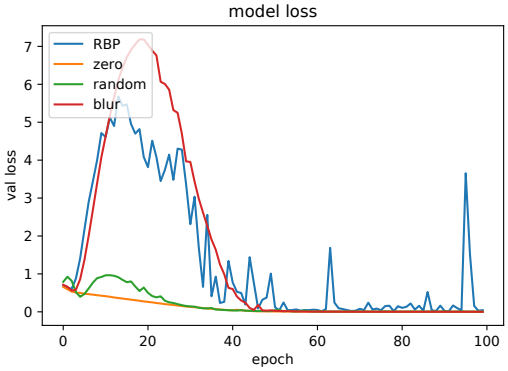


Figure 3.5: Figure shows the validation loss of a classifier on the synthetic datasets generated by each perturbation method. A validation curve which falls quickly to near zero indicates that the model has successfully learned to separate each class and has generalised well to the validation set. RBP does not have a smoothly decreasing loss curve and has not reached stable low loss which indicates that the black box is unable to differentiate between perturbed and non-perturbed time series.

3.4.2 Experimental Validation Of Realistic Background Perturbation

To evaluate RBP on Open Question 2 we ask the following: How capable is RBP at generating background content? How capable is RBP at generating realistic new samples? To evaluate the former we use the intuition of Agarwal et al. [3] adapted to time series: the more successful the perturbation, the worse the classification performance when differentiating between real and generated samples. We generate a synthetic binary time series dataset where each time series has five super segments. The class-wise difference is contained solely in the final super-segment.

We select a 1D Convolutional Neural Network as our black-box classifier. We perturb the final segment of each evaluation time series with either RBP, blurring, zeroed, or random values and obtain classification accuracy on each perturbed dataset. Results are shown in Table 3.2 where

	Original	RBP	Zero	Random	Blur
Acc	1.0	**0.36**	0.49	0.52	0.47

Table 3.2: Table shows the classification accuracy of the CNN applied to synthetic time series datasets under each perturbation strategy.

we can see a significant accuracy decrease following all perturbations. However, the accuracy decrease is most significant for RBP.

To evaluate whether the time series perturbed RBP produces more realistic time series than blurring, noise and zero perturbations we build on the theory of Chen et al. [33], who show how unrealistic samples allow an adversary to differentiate between data points coming from the input distribution and instances generated via perturbation. To test how realistically RBP generates new samples we train a new classifier, a 1D CNN, on a binary class synthetic time series dataset. Class A contains time series with five super segments. Class B contains the time series of Class A after undergoing a perturbation. We train the classifier on the dataset and compare the validation loss curve for varying perturbation strategies including RBP, blurring, zeroed, random. The more realistic the perturbation, the more difficult it will be for the classifier to learn and generalise. Results are displayed in Figure 3.5 which confirms our earlier claim that blur, zero and random perturbation result in more unrealistic time series than RBP.

In this section we have explained why occlusion for time series is challenging and motivated our algorithm RBP as an occlusion strategy which filters out salient information in the frequency domain of a time series. We have compared RBP with other occlusion strategies to experimentally shown that RBP is capable of removing salient information from a time series.

3.5 Defining A Time Series Neighbourhood

An important concept in LIME as discussed in Section 3.2.6 is the weighting over generated samples, used as input to the interpretable model, to encapsulate the intuition that samples closer to the instance to be explained should have more influence on the generated explanations. It has been shown by [67] that for image data, weights depend only on the number of inactivated super-pixels in each generated sample. For a time series, distances measured in the interpretable domain fails to take into account the global distance between the generated sample and the original instance. Figure 3.6 visualises an example time series $\mathbf{x}^0$ with six super segments at indexes: $\{100, 200, 300, 350, 400\}$ where $\sigma(\mathbf{x}^0) = \{1,1,1,1,1,1\}$. We generate three new samples as $\sigma(\mathbf{x}^1) = \{0,1,1,1,1,1\}$, $\sigma(\mathbf{x}^2) = \{1,1,0,1,1,1\}$, $\sigma(\mathbf{x}^3) = \{1,1,1,1,0,1\}$. Under the Euclidean distance each of the generated samples would be equidistant from $\mathbf{x}^0$. However, we can see that perturbation

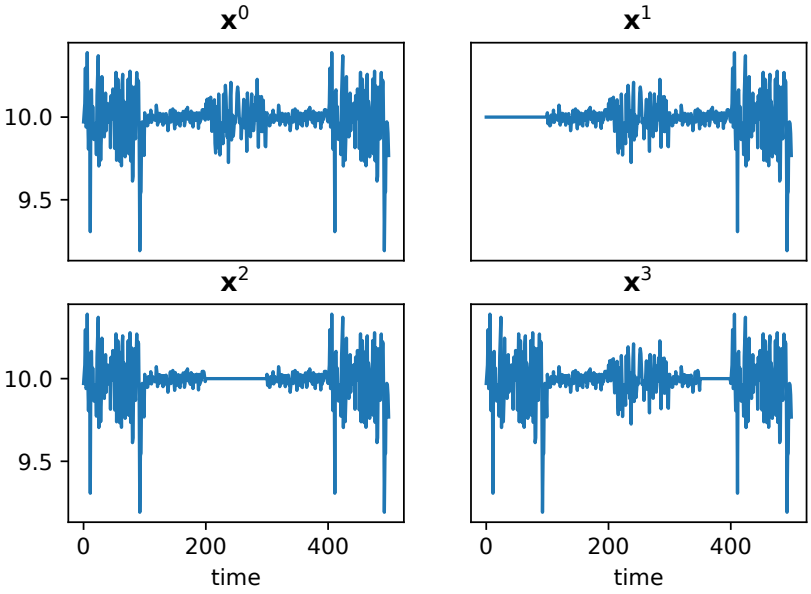


Figure 3.6: Figure shows an example time series, $\mathbf{x}^0$, and three generated samples with zero perturbations at different locations: $\mathbf{x}^1$ at index $\{0:100\}$; $\mathbf{x}^2$ at index $\{200:300\}$; $\mathbf{x}^3$ at index $\{350:400\}$

.

to shorter length super segments ($\mathbf{x}^3$), are more similar to the original instance than longer perturbations. Moreover, even for super-segments of equal length we can use Figure 3.6 to argue that $\mathbf{x}^2$ is closer to $\mathbf{x}^0$ than $\mathbf{x}^1$ as $\mathbf{x}^2$ maintains more motif structure than T1 which raises **Open Question 3:** *"How do we measure distance between two time series that accurately reflects local neighbourhood round* $\mathbf{x}$*?"*

To address Open Question 3, we work under the assumption that a more similar sample maintains more of the original global structure of the time series to be explained. In Figure 3.6 we would assume that $\mathbf{x}^3$ is more similar to $\mathbf{x}^0$ than $\mathbf{x}^1$. We thus employ Dynamic Time Warping (DTW), introduced by Bellmanbuhr et al. [21] to address the limitations of the Euclidean distance in measuring the similarity between two time series in the original time domain. It is not necessary, for our explainaility method, to explicate the workings of the DTW algorithm which is complex and heavy in notation. However, we direct the interested reader to the original paper of Bellman [21]. The superiority of DTW over Euclidean distance for these tasks has been demonstrated in the literature [172]. Particularly, Euclidean distance metric is widely known to be very sensitive to distortion in time axis. DTW accounts for this distortion by allowing non-linear alignments between two time series to accommodate sequences that are similar, but locally out of phase [172] . For generating explanations, and more specifically, generating the weighting between the instance to be explained and the generated samples, this property of DTW is very useful. For the example samples shown in Figure 3.6, we obtain $d_{DTW}(\mathbf{x}^0,\mathbf{x}^1) = 16$,

	Euclidean	DTW
Simple Synthetic	0.11	**0.49**
Complex Synthetic	0.17	**0.22**
ECG200	0.55	**0.63**

Table 3.3: Table shows the Mean RSSI of the explanations generated by LIMESegment with DTW and Euclidean distance to the classification of the Simple, Complex Synthetic and the ECG200 datasets.

$d_{DTW}(\mathbf{x}^0, \mathbf{x}^2) = 10$ and $d_{DTW}(\mathbf{x}^0, \mathbf{x}^3) = 1$ which captures the intuition that perturbations applied to the more significant super-segments of the original time series should represent a more dissimilar resulting sample. While this is a strong assumption to make, we adopt the argument of Gorecki et al. [73], that while there exist many methods which measure the similarity of time series, the DTW often emerges as the optimum [73]. In the future we look to explore cases where this assumption may not hold and explore more holistically what it means to be in the local neighbourhood of a time series.

3.5.1 Experimental Validation Of DTW

To evaluate DTW on Open Question 3, we follow the same approach of Visani et al. [215] and assume that a failure to correctly sample in the locality of $\mathbf{x}$ results in unstable explanations. To measure the stability of explanations we adapt the explanation stability metric of Visani et al. [215] to introduce Ranked Segment Stability Index (RSSI) as the proportion of concordant (equal) pairs out of all $\frac{n(n-1)}{2}$ pairs of segment importance vectors for a given time series after running the explanation algorithm for n iterations. To show how DTW results in more stable explanations than the Euclidean distance measure we generate explanations via our adaptation of LIME for time series, LIMESegment, which we introduce in formally Section 3.6, under both the Euclidean and DTW distance measures.

We generate two synthetic binary class time series datasets. Across both datasets each time series has five super segments. In “simple synthetic”, the class difference occurs in the final super segment of each time series. In “complex synthetic” the class difference is spread evenly across the initial and final super segments. We also evaluate the stability of LIMESegment on the ECG200 dataset from the UCR time series repository [35]. We use a 1D CNN as our black box classifier. For each time series we obtain two sets of segment importances by running LIMESegment using DTW as well as running LIMESegment with Euclidean distance. To establish the stability of each set of segment importances we repeat this process 50 times for each time series. Table 3.3 shows RSSI of explanations under DTW is greater than those of Euclidean distance implying that comparing the similarity of generated time series to the original time series in their raw form according to DTW results in more stable explanations.

In this section we have discussed why defining a meausre of locality around an individual univariate time series is challenging. We have motivated both theoretically and experimentally the use of Dynamic Time Warping over Euclidean distance when defining a neighbourhood around an individual time series to be explained by LIME.

3.6 LIMESegment: An Adaptation Of LIME For Time Series

In the following section we bring together the ideas from the previous three sections, that is, NNSegment, RBP and DTW, to formalise our extension of LIME to time series. Algorithm 5 details our framework, LIMESegment, for generating local time series explanations. LIMESegment is designed for the following problem statement. Given an example time series to be explained $\mathbf{x} = \{x_1, ..., x_t\}$ and an underlying black box classifier $f : \mathscr{X} \rightarrow \mathbb{R}$, we build a surrogate model g in the locality of $\mathbf{x}$ to generate explanations in the interpretable domain $\sigma(x)$.

We first transform $\mathbf{x}$ into its interpretable representation $\sigma(\mathbf{x})$ which corresponds to a vector of ones for each super-segment found by NNSegment. We generate random samples in the locality of $\sigma(\mathbf{x})$ according to a Bernoulli sampler. Each coordinate of $\sigma(z_i)$ is i.i.d Bernoulli distributed with parameter $\frac{1}{2}$. We use RBP to convert the samples $\sigma(\mathbf{Z})$ into time series $\mathbf{Z}$. Given our transformed time series dataset $\mathbf{Z}$ we can obtain predicted sample labels $Y_{\mathbf{Z}} = f(\mathbf{Z})$. To determine the distances to be used as weightings π to the surrogate model g we compute for each $\mathbf{Z}_i \in \mathbf{Z}$, $\pi_i = exp(\frac{DTW(\mathbf{x},\mathbf{Z}_i)}{1}$. We use a Linear Ridge Regression as our surrogate model g and interpret the feature weight vector $\mathbf{w}$ as our super-segment importances and resulting explanations. We argue that the combination of a meaningful segmentation algorithm NNSegment and realistic perturbation RBP alongside the use of DTW results in a more appropriate adaptation of LIME to time series than existing methods [77, 156], which we evaluate extensively Section 3.6.1.

3.6.1 Experimental Validation Of LIMESegment

To evaluate LIMESegment as a tool for generating meaningful, realistic explanations we evaluate:

1) How faithful is LIMESegment to the original black box classifier? We define faithfulness as the decrease in classification confidence of the black-box when removing the most important super-segment from the time series as returned by its explanation. If LIMESegment has correctly identified the most important super-segment, its removal will result in large decrease in prediction confidence. We adopt the same approach as Neves et al. [156] whereby removing a selected segment from the time series corresponds to replacing it with reversed segment values. To measure the faithfulness of LIMESegment we measure the mean drop in prediction probability after segment removal for each time series in the test set.

Algorithm 5 LIMESegment

Require: TS $\mathbf{x}$, Model f, no. of samples d
Require: Bernoulli Sampler B, $RidgeRegression$
Require: NNSegment, RBP,ws,ϵ
  $\mathbf{cp} \leftarrow NNSegment(\mathbf{x}, ws, \epsilon)$
  $\sigma(\mathbf{x}) \leftarrow \{1\}^{|\mathbf{cp}|}$
  $\mathbf{Z} \leftarrow \{\}$
  for $i \in \{1, ..., d\}$ **do**
    $\sigma(\mathbf{z}_i) \leftarrow \{B_0, ..., B_{|\sigma(\mathbf{x})|}\}$
    $\mathbf{z}_i \leftarrow RBP(\mathbf{x}, ws, cp, \sigma(\mathbf{z}_i))$
    $\mathbf{Z} \leftarrow \mathbf{Z} \cup \mathbf{z}_i$
  end for
  $Y_{\mathbf{Z}} \leftarrow f(\mathbf{Z})$
  $\pi \leftarrow exp(\frac{-DTW(\mathbf{x},\mathbf{Z})^2}{2l^2})$
  $\mathbf{w} \leftarrow RidgeRegression(\sigma(\mathbf{Z}), Y_{\mathbf{Z}}, \pi)$
  Return Segment Importances $\mathbf{w}$

2) How robust is LIMESegment to small variations in the input space? Ideally, an end-user would want their explanations to be robust to small changes in the input space such that anomalous observations don't influence the resulting explanation. In this work we measure robustness by observing the difference in explanation generated for a time series before and after it has been perturbed with randomly generated noise. We report the proportion of time series whose explanations are unchanged following perturbation as our measure of robustness.

To evaluate robustness and faithfulness of LIMESegment we use three black box classifiers: K Nearest Neighbour, a 1D CNN and the state-of-the-art time series classification LSTM proposed by [100]. We train each classifier on twelve randomly selected binary time series datasets from the UCR repository [35]. We evaluate the performance of LIMESegment against the performance of the LIME time series adaptations of Guilleme et al. [77], and Neves et al. [156]. For each dataset we measure the mean faithfulness and robustness across classifiers alongside the standard deviation. Table 3.4 reports five individual datasets alongside the mean faithfulness and robustness across all twelve datasets evaluated (all). Explanations generated under LIMESegment are significantly more Robust than the framework of Neves et al. [156] and Guilleme et al. [77] for all datasets evaluated. While LIMESegment is also more faithful than the approach of Neves et al. [156] and that of Guilleme et al. [77], individual dataset results are nuanced: LIMESegment applied to instances of the Strawberry dataset achieves significantly superior faithfulness across all three classifiers than the frameworks of Guilleme et al. [77] and Neves et al. [156].

We observe that generally, the most important segment as returned by LIMESegment occurs in the middle of these time series and is roughly of length 20. De et al. [47] introduce a wavelength importance classifier and show how the Strawberry dataset can be accurately classified by retaining just 4% of the original signal. Our result confirms this finding and shows how LIMESegment successfully identifies the most significant segment of the time series whose removal results in

	strawberry	handout	yoga	ecg200	chinatown	all
			faithfulness			
L	**0.35** ± 0.10	0.08 ± 0.05	**0.10** ± 0.06	**0.20** ± 0.18	**0.05** ± 0.09	**0.10** ± 0.06
G	0.05 ± 0.04	0.05 ± 0.06	0.06 ± 0.03	0.13 ± 0.18	0.03 ± 0.06	0.04 ± 0.05
N	0.07 ± 0.04	**0.16** ± 0.14	0.05 ± 0.04	0.16 ± 0.14	**0.05** ± 0.09	0.06 ± 0.05
			robustness			
L	**1.00** ± 0.20	**0.40**± 0.20	**0.70** ± 0.36	**0.88** ± 0.20	**0.98** ± 0.03	**0.74** ± 0.14
G	0.20 ± 0.35	0.00 ± 0.00	0.60 ± 0.53	0.33 ± 0.58	0.67 ± 0.58	0.42 ± 0.22
N	0.00 ± 0.00	0.10 ± 0.17	0.52 ± 0.40	0.45 ± 0.05	0.67 ± 0.58	0.34 ± 0.23

Table 3.4: Table shows the mean and standard deviation of the faithfulness (F) and robustness (R) of LIMESegment (L), The explanation approach of Guilleme et al.[77] (G) and that of Neves et al. [156] (N) after training KNN, CNN and LSTM on 12 datasets from UCR repository [35] (all) alongside individual results of five datasets. As L and G require user defined segmentation we report the best results obtained with segment length of 5%, 10%, or 20% of time series length.

a significant decrease in performance, exemplifying the benefits of meaningful over arbitrary segmentation. For the HandOutlines dataset we find that the framework of Neves et al. [156] is more faithful than that of LIMESegment. Each time series of HandOutlines is of length 2709 and the framework of Neves et al. [156] performs best when using five super segments. LIMESegment identifies 12 super segments where we observe that generally, segment importance is spread evenly across five super segments. As faithfulness calculates performance decrease after removal of the singularly most important segment we can understand why Neves et al. [156] results in higher faithfulness as more of the salient information is contained in its longer length segments.

3.6.2 Sepsis Cohort Case Study

LIMESegment in Practice: We now present LIMESegment in a healthcare setting to exemplify the insights time series classification explanations offer. We apply LIMESegment to the MIMIC Sepsis Cohort as described in Section 1.8. We generate a univariate time-series dataset described by selecting only the temperature vital sign observations of all those patients in the original cohort.

We train a 1D CNN on 1130 univariate time series and randomly select time series from a test set to generate explanations via LIMESegment. Four example time series and associated segment importances are shown in Figure 3.7. For the True Negative (individual dies) and True Positive (individual survives), the segment following sepsis onset is most influential for the resulting classification which aligns with sepsis research reporting elevated peak temperature in the first 24 hours following sepsis onset is associated with decreased in-hospital mortality [228], This kind of explanation guided observation demonstrates how LIMESegment could be used to offer medical insight to an end-user.

Figure 3.7 shows how, for the False Negative example, the most influential segment in its

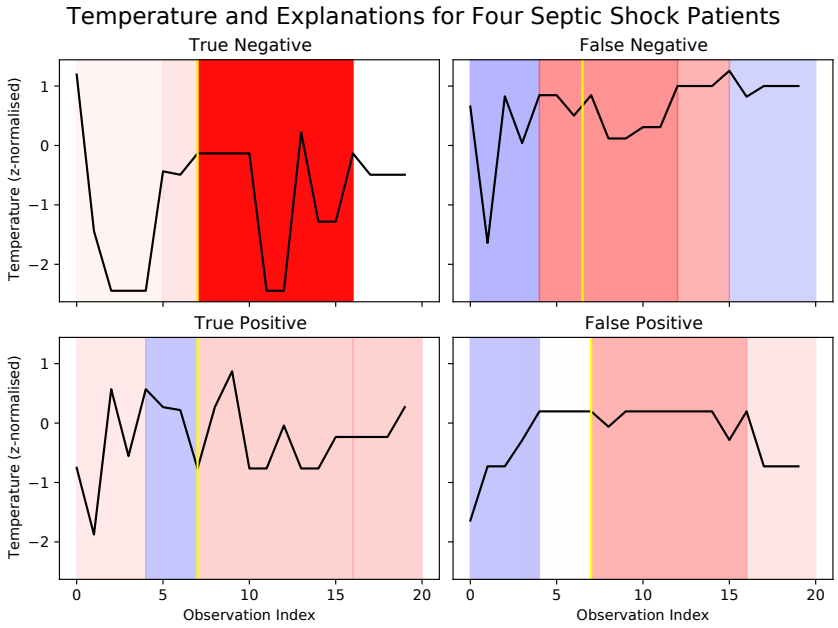


Figure 3.7: Figure shows the application of LIMEsegment to patient trajectories from the MIMIC Sepsis Cohort. Each time series is labelled as either a True Negative or Positive, where the black box has correctly classified the instance or, as a False Positive or Negative where the black box has misclassified the sample. Each super-segment as returned by LIMESegment is shaded either blue or red. Red shading indicates the segment importance supports the black box prediction and blue indicates the segment importance contradicts the black box prediction. Opacity indicates greater segment importance. The yellow vertical line indicates sepsis onset for each individual. For both correctly classified instances LIMESegment has detected the time of sepsis onset in its segmentation.

incorrect classification is of similar shape to the True Negative's most influential segment: temperature significantly dips and sharply rises, giving insight into why this sample was incorrectly classified by the black box. To further evaluate the LIMESegment algorithm on the Sepsis Cohort we compare faithfulness and robustness of LIMESegment with Guilleme et al. [77] and Neves et al. [156] on each test sample. LIMESegment obtains faithfulness of 0.29 and robustness of 1.0, Guilleme et al. [77] obtains faithfulness of 0.02 and robustness of 0.90, and Neves et al. [156] obtains faithfulness of 0.12 and robustness of 0.72, demonstrating how LIMESegment outperforms state-of-the-art time series LIME adaptations on real-world data.

3.7 LIMESegment: Concluding Remarks

In this chapter we have overviewed LIME [174] and outlined how its customisability requires a plethora of decisions to be made when applying LIME for a particular investigative task. We focused on the kind of assumptions which are encoded by customisation when LIME is used to explain time-series data. We have introduced three open questions which should be addressed

in future work when adapting post-hoc local explanations for this kind of data structure. We have presented our application of LIME to univariate time series data, LIMESegment and experimentally validated the utility of the resulting explanations. We discuss the limitations of LIMESegment in Chapter 7. Furthermore, here we note that LIMESegment is limited to univariate time series, conceptualisation for multivariate time series and how these concepts may be occluded from a signal is an interesting research challenge, particularly considering the dependence between different variables over time ubiquitous in this data type, which we motivate for future study. One particular limitation of LIMESegment - a general critique of LIME itself, which motivates the shift in focus in the following chapters to consider an alternative method for feature attribution, the Shapley value.

Despite the popularity of existing post-hoc local explainers, their formal properties are under-studied [154]. This makes their comparison difficult. We have discussed in Chapter 1 how evaluating post-hoc explanations is challenging due to the absence of a "ground-truth" explanation. As such, a plethora of evaluation metrics have been devised to assess the relative advantages of different approaches. However, these evaluation metrics are also sensitive to weaknesses and have been observed, in the literature, to lack both rigour and consistency when evaluating post-hoc explanations [154]. Concerning LIME, it has been shown that the quality of LIME explanations can vary, indicating that depending on dataset, LIME can be both accurate, or inaccurate, at explaining the model [154].

LIMESegment is also sensitive to this instability of evaluation metric. We draw attention to the sensitivity of robustness to the length of a given time series as exemplified by Table 3.4 for the UCR time series examples. Particularly we compare robustness on the HandOutlines dataset where $m = 2709$ and the Chinatown dataset where $m = 24$. robustness for Chinatown is over double than that obtained for HandOutlines with twelve times the variance between time series, not only confirming the observation made by Narody et al. [154], that evaluation measures can significantly vary between datasets but also between instances of the same dataset. As an explanation as to why this variance occurs we note that the heuristic nature of the original LIME algorithm (and therefore LIMESegment) is not guaranteed to provide the same explanation for the **same** sample. This is due to the fact that LIME generates a user-specified number of hypothetical samples. Each time we compute LIMESegment, a different combination of hypothetical samples may be used to determine the resulting attribution. A such, we cannot guarantee the stability of evaluation measures even on the same instance! The fact that we cannot reliably interpret evaluation metrics is problematic for the future of post-hoc local explanations and as a consequence, attention has been paid to mathematical formalisation [7, 154]. In the same vein, the following chapters see us turn our attention to feature attribution methods which are accompanied by mathematical guarantees. In particular we explore the application of the game theoretic Shapley value to XAI.

CHAPTER 4

SHAPLEY SETS: INTERACTION-ROBUST ATTRIBUTIONS

Feature attribution based on the Shapley value has become one of the most ubiquitous techniques for post-hoc local explanations, inspiring a rapidly expanding sub-genre of approaches for applying the Shapley value to the explainability landscape. These feature attribution methods are considered the state-of-the-art in the XAI (XAI) community yet the Shapley value has itself existed since the 1950s. The history of the Shapley value set within the game-theoretic context from which it was designed is often ignored by the machine learning community, where its application to feature attribution is usually justified from a statistical perspective. Following the seminal paper of Llundberg and Lee in 2014 [138], the Shapley value immediately became the front runner for feature attribution. Despite its theoretical appeal, the application of the Shapley value for feature attribution isn't without its problems which we explore in this chapter.

In the following three chapters, we challenge the Shapley value from three distinct perspectives and expose a variety of assumptions, limitations and potential use cases associated with its application in feature attribution. We propose three alternative feature attribution techniques which are inspired by our analysis of the Shapley value and address some of the limitations we outline. While not restricted to time series, the particular dependency structure of these high-dimensional objects is particularly sensitive to simplifying assumptions employed by the Shapley value for feature attribution and thus, we motivate each of our attribution methods for the use on time series attribution challenges (Chapter 6).

In this chapter, we begin by introducing the Shapley value within its original game-theoretic context and discuss the process by which it was adopted by the feature attribution community. We present the challenges which arise from the transition between game theory and machine learning and argue that the misleading attributions under the Shapley value, in the presence of feature interaction, occur due to the approximations required to make it work for functions with features rather than games with players.

We propose an alternative attribution approach, Shapley Sets, which awards value to sets of features. Shapley Sets decomposes the underlying model into Non Separable Variable Groups (NSVGs) using a recursive function decomposition algorithm with log-linear complexity in the number of variables. Shapley Sets attributes to each NSVG their combined value for a particular prediction. We show that our approach, Shapley Sets, is equivalent to the Shapley value over the transformed feature set and thus benefits from the same axioms of fairness. Shapley Sets is value function agnostic and we show theoretically and experimentally how Shapley Sets avoids pitfalls associated with Shapley value-based alternatives and are particularly advantageous for data-types with complex dependency structure. As outlined in Chapter 1, much of this work in this chapter is published by Sivill and Flach as [193].

4.1 Value Attribution In Game Theory

Game theory formalises situations of conflict and ventures of co-operation, of decision problems with multiple decision makers, whose decisions impact one another and the outcome. Coalitional game theory models the collaboration of a group of decision makers who co-operate and take joint action in a coalition to increase their value in some venture. Section 4.1 introduces the Shapley value within its game theoretic context, comparing it to alternative economic game values, such as the Gately value which underpins Chapter 5, it can therefore be skipped if the reader is interested in the novel contribution of this chapter, Shapley Sets which is motivated in Section 4.3 and formalised in Section 4.4.

Example 4.1. *Imagine a group of countries all part of an agreement to reduce the number of global deaths from sepsis. All the countries must work together to maximise the total reduction and as each country in the group has their own resources, time, influence (and political agenda), the overall reduction in sepsis mortality will vary depending on which countries participate. At the end of the year, all countries attend a global summit where it is revealed that the agreement managed to lower global sepsis deaths by a factor of three. As part of the review, the summit wish to understand the difference in contributions between countries in the agreement.*

The challenge of fairly dividing an overall payout between n parties has persisted since the earliest human societies. However, it has been the focus of coalitional game theory since the 20th century. Coalitional game theory is the formal study of interacting decision makers, applying a mathematical framework to the problem of value allocation. Coalitional games have since found application in economics, finance, politics, computing and problem resolution of various kinds.

Definition 4.1 (Coalitional Game)**.** A coalitional game is a tuple (N, v), where N is a set of players and $v : 2^N \to \mathbb{R}, v(\varnothing) = 0$ is the characteristic function for coalitions which takes as input a coalition of players, $S \subseteq N$ and returns the value of that coalition in the game v. The class of coalitional n-person games is denoted by G_n.

Formalising Example 4.1 as a coalitional game would set the player set N as the group of countries taking part in the sepsis deaths agreement: Britain, France, Japan and the game v as the reduction in global deaths from sepsis. v quantifies for sub-coalitions, e.g. France and England, their isolated reduction in deaths.

The central idea behind coalitional game theory is that the players will work together as their coalitional payoff will be at lest as good as their individual payoff. Under this assumption, the formation of the grand coalition, N, will lead to the highest benefit of the coalition since $v(N)$ is always better, or equal to the payoff earned by any of the possible coalitions which could form. $v(\varnothing)$ represents the value earned by the coalition containing no players and is assumed to be zero within coalitional games: if no countries took part in the agreement then the reduction in sepsis deaths would be zero.

In this section we have introduced to concept of value attribution within coalitional games whereby the joint value earned by a group is split between each individual fairly.

4.1.1 Solution Concepts: No Free Lunch

The goal in coalitional game theory is to divide up the value of the grand coalition, $v(N)$ such that each player receives a proportion which is representative of their contribution to the overall game. Solution concepts are used to calculate the worth at the individual marginal level using the characteristic function $v(S)$ of the game (N,v). Central to the definition of a solution concept is the notion of a marginal contribution, which can be understood as the amount by which the value of a coalition increases, upon introducing a given player to that coalition [65].

Definition 4.2 (Marginal Contribution [65])**.** The marginal contribution of player i to coalition S is the difference in value when player i joins the coalition: $v_i(S) = v(S \cup \{i\}) - v(S)$

For a coalitional game v, the payoff vector as determined by the solution concept, $\boldsymbol{\omega}(v) = (\omega_1, \omega_2, ...\omega_n)$ for a coalition $S \subseteq N$, is the proposed amount distributed among the players, such that player i is to receive ω_i. In coalitional games, the players are assumed to form multi-level coalitions.

The cornerstone of a solution concept is therefore the fair distribution of coalitional value among all coalition members. First, let $(N,v) \in G_n$. The solution concept $\boldsymbol{\omega}(v)$ is an **imputation** of v if $\sum_i^n \omega_i = v(N)$ and $\omega_i \geq v(\{i\})$. The first condition, the axiom of efficiency, enforces that the sum of all individual payouts must equal the overall value of the grand coalition. The second condition, individual rationality, enforces that each player receives at least their individual contribution to the game.

Definition 4.3 (Additively Separable Function)**.** The function $f : \mathscr{X} \to \mathbb{R}$, with variable set $\mathbf{X} = \{X_1, ..., X_n\}$, is separable if it has the following form, $f(\mathbf{X}) = \sum_{i=1}^k f_i(\mathbf{X}_i)$ for all $1 < k \leq n$.

Here, $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_k$ are k non-overlapped sub-vectors of $\mathbf{X}$ and f_i refers to a function of only the variables $\mathbf{X}_i \subset \mathbf{X}$. Specifically, the function f is also called fully additively separable if $k = n$, while it is regarded as fully non-separable if $k = 1$.

Under efficiency, imputations decompose the value of the grand coalition $v(N) - v(\varnothing)$ to attribute worth to each individual player. An imputation is a fully separable function such that $v(N) - v(\varnothing) = \sum_{i=1}^{n} \omega_i(v)$. In contrast, the set function v is not guaranteed to be fully separable. Within coalitional games this is due to the interaction between players. Consider Example 4.2 for the game (N, v) with player set $N = (France, England, Japan)$

Example 4.2. $v(\{Japan\}) = 1$
$v(\{England\}) = 0$
$v(\{France\}) = 0$
$v(\{Japan, England\}) = 2$
$v(\{France, Japan\}) = 3$
$v(\{France, England\}) = 1$
$v(\{France, England, Japan\}) = 3.$

From Example 4.2, as $v(\{England\}) + v(\{France\}) + v(\{Japan\}) \neq v(\{France, England, Japan\})$ the game is not fully separable. The non-separable interaction effects within coalitional games are dealt with by imputations which map partially separable into fully separable functions, allowing an individual attribution of worth to each player. There is no optimal way of dividing up the interaction effects of a coalition and as a result there is no universal solution concept within the vast landscape of candidates. To divide up the total payoff (in this case the reduction in deaths from sepsis) among each country so that each country is assigned a fair allocation of how much it contributed, the definition of fair is critical.

Fairness, akin to explainability, is notoriously difficult to define. Arguments surrounding the distribution of joint value constitute the topic of distributive justice. Principles of distributive justice are therefore best thought of as providing guidance for the structures that affect the distribution of benefits and burdens in societies. There exist many solution concepts which adopt different principles of distributive justice, which can include the following two principles.

The Utilitarian principle asserts that the best social policy for the division of joint goods is the one which gives the greatest total welfare to the individual members of society, where "total welfare" is measured by summing the value attributed to each individual [121]. Maximising the classical utilitarian sum of individual welfare across all members of a society may mean sacrificing some individuals.

The Strict Egalitarian principle is one of the simplest principles of distributive justice. In contrast to the Utilitarian perspective, Strict Egalitarianism stipulates that every member should be allocated the same amount of value regardless of their contribution.

Equality of Opportunity is an alternative form of egalitarianism which asserts that the best social policy is the one which gives the greatest total welfare to individual members of society subject to the constraint that all individual members should have equal opportunities within the

society. This egalitarian principle maximizes the utility of the most unfortunate individuals in society [121].

While the solution concept landscape is rich and diverse, in this section we introduce some of the most commonly used, connecting them to the above principles of fairness.

In this section we have introduced the class of solution concepts within coalitional value attribution. The attribution afforded by a solution concept is characterised by its axiomtisation of fairness.

4.1.2 Utilitarian Division: The Shapley Value

Historically the first and undoubtedly most prominent value for coalitional games, the Shapley value [188], is defined as the payoff vector which assigns to each player $i \in N$, a weighted average of their marginal contributions over all possible coalitions they could join $S \subseteq N \setminus \{i\}$.

Definition 4.4 (Shapley value)**.** For the game v the Shapley value of player $i \in N$ is given as

$$\phi_i(v) = \Sigma_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [v(S \cup \{i\}) - v(S)] \tag{4.1}$$

To understand Equation 4.1 it is helpful to consider players joining the game sequentially. There exist $|N|!$ possible orders over which players could enter the game and this is the denominator of Equation 4.1. Similarly, there exist $|S|!$ possible orders over which players, belonging to coalition S, can enter the coalitions. As player i enters coalition S, the new value of the characteristic function will become $v(S \cup i)$. The numerator is therefore the number of potential orders in which the $(|S| - 1)$ members of $(S \setminus \{i\})$ precede player i and the number of potential orders in which the $(|N| - |S|)$ members of $(N \setminus S)$ follow player i. The Shapley value is therefore just the average amount player i contributes to each coalition, if the players sequentially form this coalition in a random order.

An alternative characterisation of the Shapley value explicitly accounts for each of the possible orderings [188]. Let Π be the set of all orderings or permutations of the set N. For any player $i \in N$ and any permutation of the player set $\pi \in \Pi$, the following are denoted:

$$H_\pi(i) = \{j \in N | \pi_j < \pi_i\} \tag{4.2}$$

Here $H_\pi(i)$ is the set of all predecessors of i and all in the permutation π (excluding player i). Given the equation above, an alternative definition of the Shapley value is given below.

Definition 4.5 (Shapley Value)**.** For the game v the Shapley value of player $i \in N$ is given as

$$\phi_i(v) = \sum_{\pi \in \Pi} \frac{1}{n} [v(H_\pi(i) \cup \{i\}) - v(H_\pi(i))] \tag{4.3}$$

If we consider Example 4.2 the Shapley values for England (E) Japan (J) and France (F) would be,

$$\phi_{Japan} = (\frac{2}{6} \times 1) + (\frac{1}{6} \times 2) + (\frac{1}{6} \times 3) + (\frac{2}{6} \times 2) = \frac{11}{6},$$

$$\phi_{France} = (\frac{2}{6} \times 0) + (\frac{1}{6} \times 1) + (\frac{1}{6} \times 2) + (\frac{2}{6} \times 1) = \frac{5}{6},$$

$$\phi_{England} = (\frac{2}{6} \times 0) + (\frac{1}{6} \times 1) + (\frac{1}{6} \times 1) + (\frac{2}{6} \times 0) = \frac{2}{6},$$

$v(\{J\}) - v(\varnothing) = 1$	$v(\{F\}) - v(\varnothing) = 0$	$v(\{E\}) - v(\varnothing) = 0$
$v(\{J,E\}) - v(\{E\}) = 2$	$v(\{F,E\}) - v(\{E\}) = 1$	$v(\{F,E\}) - v(\{F\}) = 1$
$v(\{J,F\}) - v(\{F\}) = 3$	$v(\{F,J\}) - v(\{J\}) = 2$	$v(\{E,J\}) - v(\{J\}) = 1$
$v(\{J,F,E\}) - v(\{F,E\}) = 2$	$v(\{J,F,E\}) - v(\{J,E\}) = 1$	$v(\{J,F,E\}) - v(\{J,F\}) = 0$

The Shapley value can be treated as a utilitarian allocation in the sense that every player claims the entirety of their average marginal contribution without sharing with others. We can see straight away that within Example 4.2, Japan is more valuable than both France and England from the perspective of its marginal contribution. The Shapley value is built on this intuition that those who contribute more to the coalitions that include them should be allocated more. This persepctive of fairness is further characterised by the following axioms which the Shapley value uniquely satisfies [188]:

- Efficiency: the total worth distributed among the players is equal to the value of the grand coalition $\sum_{i \in N} \phi_i(v) = v(N)$
- Symmetry: If i and j are such that $v(S \cup \{i\}) = v(S \cup \{j\})$ for every coalition $S \subseteq N$, where $i, j \notin S$ then $\phi_i(v) = \phi_j(v)$. Two players i, j are said to be symmetric with respect to coalition game (N, v) if they make the same marginal contribution to all coalitions.
- Dummy: If it holds for player i that $v(S) = v(S \cup \{i\})$ for every coalition S where $i \notin S$, then $\phi_i(v) = 0$. The contribution by a dummy participant is null and his payoff vector is null.
- Additivity: For coalitional games u and v, then $\phi(u + v) = \phi(u) + \phi(v)$. When two games are played at the same time, the sum of the Shapley values of the games are as if they are played separately.

In this section we have introduced the Shapley value as a utilitarian solution concept and given its two most prominent definitions.

4.1.3 Strict Egalitarian Solution Concepts

The most extreme of allocations under the principle of Strict Egalitarianism, the Equal Division allocation (ED), divides the grand payout equally between all players regardless of their marginal contribution.

$$ED_i(v) = \frac{1}{n} v(N) \tag{4.4}$$

ED satisfies additivity, efficiency and symmetry but unlike the Shapley value, whose dummy axiom reflects the utilitarian perspective that players who do not contribute to any coalition should receive no share of the payout, the Equal Division solution concept considers these null players as equally valuable as marginally contributing players in the game.

Two alternative solution concepts which are built on the principle of Strict Egalitarianism is the Equal Surplus Division allocation (ESD) and the Equal Non-separable Cost allocation (ENSC), both of which – unlike the Shapley value which considers only the structure imposed by the value function – also consider the incentives of individual players in joining certain coalitions and are thus founded on the following two concepts.

Definition 4.6 (Minimum Rights Vector)**.** The minimum rights vector is the vector of the individual value of each singleton player i, $v_i = v(\{i\})$ for $i \in \{1,...,n\}$

Definition 4.7 (Utopia vector)**.** The utopia vector contains the marginal, or utopia, value of individual players defined by their "separable" contribution to the grand coalition $M_i = v(N) - v(N \backslash \{i\})$.

The ESD first allocates each player their individual value, $v(\{i\})$ in the game v and then splits the leftover equally between all players. The ENSC first allocates each player their marginal contribution to the grand coalition and then splits the (negative) leftover between each player in the game.

$$ESD_i(v) = v(\{i\}) + \frac{1}{n}(v(N) - \sum_{j \in N} v(\{j\})) \tag{4.5}$$

$$ENSC_i(v) = M_i(v) + \frac{1}{n}(v(N) - \sum_{j \in N} M_j(v)) \tag{4.6}$$

The ESD minimises the Euclidean distance between the minimal rights vector and the allocation vector $\boldsymbol{\omega}$. The ENSC minimises Euclidean distance between the utopia vector and the allocation vector $\boldsymbol{\omega}$. Both ESD and ENSC satisfy additivity, efficiency and symmetry. For Example 4.2, the allocations under the above Strict Egalitarian allocations are as follows:

$$ED_{France} = ED_{Japan} = ED_{England} = 1$$

$$ESD_{France} = ESD_{England} = \frac{2}{3}, ESD_{Japan} = \frac{5}{3}$$

$$ENSC_{France} = 1, ENSC_{Japan} = 2, ENSC_{England} = 0$$

The above attributions show that despite being built on the premise of Strict Egalitarianism, each of the above solution concepts result in different yet valid attributions to each player. It is particularly worth noting that both ENSC and ESD, despite being characterised by the same principle of an equal division of the surplus, in situations where the surplus is zero, these solution concepts align more closely with the utilitarian principle of division: players are allowed to first claim their individual value for the ESD, or their utopia value for the ENSC, before the remainder is split among all players equally.

This is exemplified by the ENSC attribution for Example 4.2 as the only of the Strict Egalitarian solution concepts to award zero attribution to England which is arguably even more utilitarian than the Shapley value. The above discussion highlights the nuances between solution concepts and how they navigate the fairness landscape. Below we introduce a further two solution concepts which despite being built on the same egalitarian principle ultimately result in allocations which are diometrically opposed in their perspective of fair attribution.

In this section we have introduced the ED, ESD and the ENSC as egalitarian solution concepts.

4.1.4 From Egalitarianism To Utilitarianism: The Nucleolus and The Gately value

The Nucleolus applies the concept of egalitarianism to coalitional games whereby the Nucleolus minimizes the "unhappiness" of the most-unhappy coalition during attribution. Schmeidler [125] defines "unhappiness" or excess $e(S,\boldsymbol{\omega})$ of a coalition S as the difference between what the members of the coalition could get by themselves and what they are actually getting if they accept the allocations suggested by the solution concept $\boldsymbol{\omega}$.

The Nucleolus thus minimises the excess or unhappiness of the most unhappy coalition and can therefore be seen, in principle, as an equal opportunity solution concept where penalties incurred to the most unfortunate of players are minimised. The Nucleolus is the imputation which also satisfies symmetry and dummy axiom but does not satisfy the additivity axiom. Furthermore there does not exist any established formula for calculating the Nucleolus, it must be computed via linear programming techniques [12] and thus, while being mathematically attractive is difficult to apply practice.

Gately [68] proposed a variant of the Nucleolus, a weighted Nucleolus, which is easier to compute. Gately's Nucleolus minimises an alternative form of unhappiness, the propensity to disrupt. For a given payoff vector, the propensity to disrupt of any player is defined as the ratio of the total amount the other players would lose if the grand coalition broke up, to the amount which that player himself would lose.

Definition 4.8 (Propensity to disrupt)**.** For any payoff vector $\boldsymbol{\omega}$, the propensity to disrupt of player $i \in N$ denoted $d(\boldsymbol{\omega}, i)$ is the ratio of loss incurred by the complementary coalition $N \backslash \{i\}$ to the loss incurred by i if that payoff vector $\boldsymbol{\omega}$ is abandoned.

$$d(\boldsymbol{\omega}, i) = \frac{\boldsymbol{\omega}_{N \backslash \{i\}} - v(N \backslash \{i\})}{\omega_i - v(\{i\})} \tag{4.7}$$

Thus, a payoff vector $\boldsymbol{\omega}$ could be considered less disruptive than another payoff vector $\boldsymbol{\omega}'$ if the maximum propensity to disrupt over all the players is less with $\boldsymbol{\omega}$ than with $\boldsymbol{\omega}'$. The payoff vector which is least disruptive in this sense is Gately's Nucleolus. The way to minimize the maximum propensity to disrupt over all players is to find the payoff vector which makes each individual propensity to disrupt equal, which is done by setting $d(\boldsymbol{\omega}, i) = d*$.

$$d* = \sum_{i \in N} \frac{M_i - v_i}{\sum_{j \in N} M_j - v_j} \tag{4.8}$$

Minimising the propensity to disrupt over all players leads to the solution concept, the Gately value [68]. Where the Nucleolus treats all coalitions as equally important, the Gately value minimises the unhappiness of the most valuable players as best as possible and in this way can be seen as a more utilitarian variant of the Nucleolus.

Definition 4.9 (The Gately Value [68])**.** For a regular game $v \in G_n$ the Gately value (GA) of player $i \in N$ is given as

$$GA_i(v) = v_i + (v(N) - \sum_{j=1}^{n} v_j) \frac{M_i - v_i}{\sum_{j=1}^{n} M_j - v_j} \tag{4.9}$$

The Gately value satisfies the following fairness axioms, for a full account of these axioms and associated proofs please see [72].

- Efficiency: $\sum_{i \in N} GA_i(v) = v(N)$
- v-Compromise: For game $(N, v), GA(v) = v' + GA(v - v')$, where $v - v'$ is the zero-normalisation of v defined by $(v - v')(S) = v(S) - \sum_{i \in S} v(\{i\})$ for every coalition $S \in 2^N$. The v-compromise property is a reduced form of additivity and as such decomposes the allocation rule in a translation of the allocation assigned to the zero-normalisation of the game [72].
- Restricted Proportionality: For every zero normalised game $w = v - v'$, $GA(w) = \gamma_w M(w)$ for some $\gamma_w \in \mathbb{R}$. The restricted proportionality property imposes zero-normalised games are assigned an allocation that is proportional to the utopia vector [72].

Applying the Gately value to Example 4.2 for each country, which for this example, coincide with the ENSC value, are as follows: $GA_{France} = 1$, $GA_{Japan} = 2$, $GA_{England} = 0$. The Shapley value satisfies efficiency and v-compromise but does not satisfy Restricted Proportionality which characterises Gately's unique perspective on fairness whereby each player should be attributed an amount lower bounded by their minimal rights value and upper bounded by their utopia value.

Similarly to ESD (Equation 4.5) the Gately value allows each player to first claim their minimal rights value. However, if there is non-zero remainder, in contrast to ESD's equal division, this is divided between players in proportion to the difference between their utopia vector and their minimal rights vector. In this sense, the Gately value can be considered as a specialised solution concept under the utilitarian principle of fairness. Where the Shapley value attributes according to each players marginal contribution to every possible coalition, the Gately value attributes according to each player's marginal contribution to the empty coalition and their marginal contribution to the grand coalition. We explore the implications of this in Chapter 5

In this section we have introduced the Nucleolus and the Gately value. The Gately value is the focus of Chapter 5 of this thesis.

4.1.5 Feature Attribution And Fairness

The previous section has introduced just a handful of solution concepts from the coalitional game theory literature to illustrate the range of perspectives of fairness which govern value division. Example 4.2 has demonstrated that these solution concepts generate valid attribution vectors which are equally justifiable from the relevant perspectives of fairness. This motivates the question, why has the Shapley value become the most ubiquitous of them all? Within coalitional game theory, the extensive generalisations of the Shapley value have shown its robustness and versatility over a wide range of applications, emerging as the "the crown jewel of coalitional game theory" [210]. It is therefore unsurprising that the last decade saw the Shapley value materialise within the machine learning community across parallel derivations of fair feature attribution and it is now the most popular method for feature attribution within the machine learning community. Section 4.2 details the Shapley value's journey within the feature attribution literature. First, however, we provide a word of caution against the blanket use of the Shapley value for feature attribution inspired by the principles of fairness we have discussed in this section.

Feature attribution is used to realise a set of typically exclusive investigative goals which could include succinctly describing a data generating process; improving the predictive performance of the model; maximising the power of the model; and producing a more cost-effective or computationally efficient predictor [65]. A method employed for feature attribution should correspond to a particular investigative question [65]. From Example 4.2 we observe that from the

perspective of the grand coalition, England is a redundant player, including it in the agreement, once Japan and France have already joined brings no marginal benefit. However, its Shapley value is non-zero, reflecting its non-zero marginal contribution to sub-coalitions. If only France and not Japan was part of the agreement, for example, England would not be a redundant player.

Within feature attribution, the decision as to whether features which display the same behaviour as England should be considered as redundant features within a model should be encapsulated by the design of the investigative goal, which in turn selects the appropriate feature attribution methodology. Furthermore, Shapley values are a model averaging procedure, being the weighted average of marginal contributions [65]. Fryer et al. [65], argue that using model averaging for feature selection, which assesses the model's performance under a large number of settings, may not be representative of the performance of a feature in the "set of optimal sub-models". If the goal of the feature attribution was to generate a more computationally efficient model, the allocation afforded by the ENSC and the Gately value would be more beneficial, both implying that the redundant feature could be dropped and there would be no impact on the outcome.

Alternatively, if the goal of feature attribution was to describe the data generating process, the Shapley value attributions, which imply that the redundant feature has a non-zero impact on certain coalitions would be useful. We have discussed in Chapter 2 how XAI has suffered from ambiguity and lack of transparency. The above discussion emphasises the importance of considering the overall objective when developing technqiues for feature atttribution. In Chapter 5 we dig deeper into the relationship between investigative goals and appropriate feature attribution methods. Specifically we formulate the range of feature attribution goals as contrastive investigative questions of the form, "Why P rather than Q" and explore the relationship between different investigative questions and the solution concepts described above.

Despite its axiomatisation, it has been acknowledged in the literature that the Shapley value, when applied to feature attribution, generates misleading attributions [117]. This is largely down to the mapping of the Shapley value from a context regarding players in games to the feature attribution context regarding features and model outcomes. In the following section we illustrate this mapping, exploring the multitude of ways in which features are modelled as players. We explore why spurious explanations arise, aligning feature interaction with Additively Separable Functions (Definition 4.3).

In this section we have unified our discussion of the game theoretic solution concept landscape to the kind of investigative question posed by feature attribution within XAI.

4.2 The Shapley Value For Feature Attribution

In this section we describe the process by which the Shapley value became the cornerstone for many of the most commonly used approaches for feature attribution within the machine learning community. Since 2001, there have been a series of parallel approaches [131, 200], which each, in their own way, devised a method for explaining a non-linear model output in the presence of feature interaction. It wasn't until the seminal paper of Llundberg and Lee [138] that these approaches were unified as applications of the Shapley value to feature attribution. Section 4.2 introduces the Shapley value within its machine learning context, showing how it can be approximated under certain statistical assumptions, it can therefore be skipped if the reader is interested in the novel contribution of this Chapter, Shapley Sets, which is motivated in Section 4.3 and formalised in Section 4.4.

As we have discussed in Chapter 1 there are many different approaches to explaining the output of a machine learning model. Arguably the most intuitive way of explaining a model $f : \mathscr{X} \to \mathbb{R}$ and variable set $\mathbf{X} = \{X_1, X_2, ...X_n\}$ is via feature attribution: quantifying, for each variable, its individual contribution to the model. One of the reasons why linear models, such as linear and logistic regression, are argued to be *transparent* is that their structure lends itself naturally to this form of explanation. Consider the linear function below:

$$f(\mathbf{X}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_i X_i ... + \beta_n X_n \tag{4.10}$$

Assuming that features $X_i \in \mathbf{X}$ are all independent we simply take each coefficient $(\beta_j, \forall j \in n)$ as the corresponding global importance for the relevant feature. If, however there is multi-collinearity between features, then the co-efficients cannot be interpreted in this way. One of the first applications of the Shapley value for feature attribution, Shapley Regression Values [131], provide a solution to the above problem, which assigns an importance value to each feature that represents the effect on the linear model's R^2 score of including that feature. Shapley Regression Values retrain the model f with only the features present in coalition S for every possible coalition $\S \subseteq N$. The impact of these sub-functions on the overall R^2 score is then averaged together according to Definition 4.4 resulting in the Shapley value for each feature.

The use of the Shapley value in this approach corresponds to global feature attribution as the resulting feature attributions correspond to the relative importance for the predictive performance, its R^2 score, of the overall model. It wasn't until 2014 where the Shapley value was used for local feature attributions [200]. We focus our attention in this chapter on local feature attribution and below we reformulate the argument of Strumbelj et al. [200] to show how the Shapley value emerged as the optimum local feature attribution in the presence interacting features.

For local feature attribution, the overarching objective is to understand the influence of each feature value, for a given sample, on its prediction in relation to the expected behaviour of the

model. Or in other words, to quantify the extent to which setting feature X_i's value to x_i impacts the resulting prediction $f(\mathbf{x})$ when compared to setting X_i to its expected value $\mathbb{E}[X_i]$:

$$\phi_i = f(x_1, x_2, ..., x_i, ..., x_n) - E[f(x_1, x_2, ..., X_i, ..., x_n)] \tag{4.11}$$

For a linear model, like that in Equation 4.10, the above equation can be simplified to the following, also known as the "situational importance" [200] of X_i.

$$\phi_i = \beta_i x_i - \beta_i E[X_i] \tag{4.12}$$

Computing the situational importance for the model above is simple because none of the features interact and the model is known to be additive. Therefore, the contribution of a particular setting of $X_i = x_i$ is the same across all instances, regardless of the other feature values. Now consider a setting where the function is not additive $f(\mathbf{X}) = X_1 \vee X_2$. Assuming that the features $\mathbb{E}[X_1] = \mathbb{E}[X_2] = 0$, determining the influence of X_1 and X_2 given the example $f(\mathbf{X}) = X_1 \vee X_2 = f(1,1) = 1$ gives both features a contribution of zero as perturbing x_1 or x_2 to their expected value does not change the function output. This is undesirable as both have an equal yet positive influence on the prediction. In reality, we know that features often interact. To account for this, rather than considering X_i's influence in isolation from all other features it was put forward by Strumbelj et al. [200] that we must consider all possible combinations of other features which X_i could interact with. To do this Strumbelj et al. [200] construct a mapping from the function f to return a single value when evaluated on a subset of features, $S \subseteq N$ where $N = \{1, 2, ..n\}$ such that S indexes variables in $\mathbf{X}$ over which f was defined.

$$f(\mathbf{x}, S) = \mathbb{E}[f(\mathbf{X}) | \mathbf{X}_S = \mathbf{x}_S]. \tag{4.13}$$

For the empty set, the above equation reduces to $f(\mathbf{x}, \varnothing) = \mathbb{E}[f(\mathbf{X})]$. Strumbelj et al. thus define the value of a subset of features, which is a generalisation of the situational importance of a feature, as the following,

$$v(\mathbf{x}, S) = f(\mathbf{x}, S) - f(\mathbf{x}, \varnothing) \tag{4.14}$$

The above equation is the change in model output caused by setting the features $\mathbf{X}_S = \mathbf{x}_S$. Now, given the set containing the values of all subsets $S \subseteq N$, to attribute the overall prediction $f(\mathbf{x})$ to each feature Strumbelj et al. map the 2^n subset values into n contributions, one allocation for each feature which is achieved by equating the value of a subset of features to the sum of all interactions across all subsets of those features, which leads to the following definition (See [200] for a full proof and further details)

$$\phi(v)_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} v(\mathbf{x}, S \cup \{i\}) - v(\mathbf{x}, S) \tag{4.15}$$

The equation above is equivalent to the Shapley value as defined in Equation 4.4, where the player set $N = \{1, ..., n\}$ indexes the variable set $\mathbf{X} = \{X_1, ..., X_n\}$ and the value function $v : 2^n \to \mathbb{R}$ is replaced by the set function which measures the contribution of a subset of feature values, $v(\mathbf{x}, S)$.

When applying the Shapley value to feature attribution, there are three functions to consider: The model to be explained $f : \mathscr{X} \to \mathbb{R}$ which operates on a variable set $\mathbf{X} = \{X_1, ..., X_n\}$. The set function v, which takes as input a set of features $S \subseteq N$ and obtains f's prediction on this coalition of features. The Shapley value $\phi(v)$ which maps the set function v into a fully separable function. Given a particular prediction to attribute, $f(\mathbf{x})$ where $\mathbf{x} = \{x_1, ..., x_n\}$, the value function $v(\mathbf{x}, S)$ specifies how the subset of features, $\bar{S} = N \setminus S$, should be removed from $\mathbf{x}$.

In this section we have shown how the Shapley value originally came to be used for feature attribution within the XAI community.

4.2.1 Value Functions

Within coalitional game theory, it is assumed that the set function v is well defined over coalitions of players such that players can be easily removed from coalitions and the value of the resulting coalition is known. In feature attribution however, the underlying model $f : \mathscr{X} \to \mathbb{R}$ is defined over a variable set N of fixed size n. Furthermore, the true probability density of machine learning models is rarely known so the computation of Equation 4.14 must be approximated as it relies on the removal of features $\bar{S}$ where $\bar{S} = N \setminus S$.

There exists a plethora of ways in which coalitional feature values are approximated, encapsulated by various value functions each with their own set of assumptions and limitations. In the absence of ground truth conditional probabilities, the conditional expectations from Equation 4.24 can be taken over the empirical input distribution.

$$f(\mathbf{x}, S) = \mathbb{E}_{Input}[f(\mathbf{X}) | \mathbf{X}_S = \mathbf{x}_S] \tag{4.16}$$

The conditional value function assumes that the empirical dataset is large and representative enough to give a meaningful approximation of the expectation. Additionally, to obtain the Shapley value of a given feature, it is necessary to calculate this expectation for every subset $\mathbf{X}_S \subseteq \mathbf{X}$ which is an exponential number of computations, all for a single feature attribution.

Tree SHAP [137] exploits the structure of tree-based models such as Random Forest models or Gradient Boosting Machines to efficiently approximate the conditional value function by

observing the proportion of samples from the empirical input distribution, which satisfy the condition $\mathbf{X}_S = \mathbf{x}_S$, and fall into each leaf node. Alternatively, Lundberg et al. [138] propose two simplifying assumptions which can reduce the complexity of estimating the conditional value function. The first is the assumption that features $X_i \in \mathbf{X}$ are each independently distributed, which allows Equation 4.14 to be approximated as the following:

$$f(\mathbf{x}, S) = \mathbb{E}[f(\mathbf{X}|\mathbf{X}_S = \mathbf{x}_S] = \mathbb{E}_{\mathbf{X}_{\bar{S}}|\mathbf{X_S}=\mathbf{x_S}}[f(\mathbf{X}_S = \mathbf{x}_S, \mathbf{X}_{\bar{S}})] \approx \mathbb{E}_{\mathbf{X}_{\bar{S}}}[f(\mathbf{X}_S = \mathbf{x}_S, \mathbf{X}_{\bar{S}})] \tag{4.17}$$

Independence allows each feature to be treated separately so we can avoid the conditional aspect of trying to find a data point that "matches" the feature values in $\mathbf{X}_S$. Instead Equation 4.17 implies that we can sample values of the features in $\mathbf{X}_{\bar{S}}$ independently to approximate their removal and average over them to estimate the expectation. The independence assumption is adopted by many of the Shapley-based feature attribution techniques in the literature which adopt a sampling technique [45, 200].

The second assumption of Lundberg et al. [138] assumes linearity of the model f which allows the following approximation of Equation 4.14.

$$f(\mathbf{x}, S) = \mathbb{E}[f(\mathbf{X})|\mathbf{X}_S = \mathbf{x}_S] = \mathbb{E}_{\mathbf{X}_{\bar{S}}|\mathbf{X_S}=\mathbf{x_S}}[f(\mathbf{X}_S = \mathbf{x}_S, \mathbf{X}_{\bar{S}})] \approx f(\mathbf{X}_S = \mathbf{x}_S, E[\mathbf{X}_{\bar{S}}]) \tag{4.18}$$

Linearity additionally allows us to simply compute the expectation over each dimension of features in $\mathbf{X}_{\bar{S}}$ separately and plug it into the model, removing the need to do sampling to estimate the expectation.

4.2.2 From LIME To SHAP

A full account of LIME is given in Chapter 3 but it is re-framed here in connection to the Shapley value. As we recall from Algorithm 1, LIME relies on a mapping from the original domain into an "interpretable domain" $\sigma(\mathbf{x})$ where $\sigma(\mathbf{x})$ is a binary vector such that a feature value of 1 indicates a feature is "turned on". We therefore view the function evaluated on the sample $\mathbf{z}$, $f(\mathbf{z})$ as being equivalent to $f(\mathbf{x}, S)$ where the indices in $\sigma(\mathbf{x})$ with a value of 1 are equivalent to the features in S. As we have discussed in Chapter 3, LIME occludes feature values, with an arbitrary background value, an average over the empirical input distribution under the Average Occlusion Strategy (Section 3.2.5). In this way, under the assumptions of linearity of the model and independence of the features, LIME can be unified with Shapley-based methods such that $f(\mathbf{x}, S) = f(\mathbf{X}_S = \mathbf{x}_S, E[\mathbf{X}_{\bar{S}}]) = f(\mathbf{z})$. To find the attributions, LIME minimises the following objective function (Equation 3.1):

$$\epsilon(\mathbf{x}) = argmin_{g \in G} L(f, g, \pi_{\mathbf{x}}) + \Omega(g) \tag{4.19}$$

Lundberg et al. [138], showed how the components of LIME could be adapted to approximate Shapley values. Specifically, by setting Ω, $\pi_{\mathbf{x}}(\mathbf{z})$ and $L(f,g,\pi_{\mathbf{x}})$ as defined in Equation 3.2 to the following equations, LIME approximates the Shapley values under $\Omega(g)=0$.

$$\pi_{\mathbf{x}}(\mathbf{z}) = \frac{|\mathbf{X}|-1}{(|\mathbf{X}|choose|\sigma(\mathbf{z})|)|\sigma(\mathbf{z})|(|\mathbf{X}|-|\sigma(\mathbf{z})|)} \tag{4.20}$$

$$L(f,g,\pi_{\mathbf{x}'}) = \sum_{\mathbf{z}\in Z} \pi_{\mathbf{x}}(\mathbf{z})[f(\mathbf{z})-g(\sigma(\mathbf{z}))]^2 \tag{4.21}$$

The parametrisation of LIME components as above equates to the Kernel SHAP approach of approximating the Shapley value for feature attribution as proposed in [138] which, is to this day, the most widely applied Shapley-based approach. The kernel $\pi_{\mathbf{x}}(\mathbf{z})=\infty$ when $|\sigma(\mathbf{z})|=0$ or $|\sigma(\mathbf{z})|=|\mathbf{X}|$ which enforces that the attribution of the empty set is $f(\mathbf{x},\varnothing)$ and that the sum of each individual feature attribution equals $f(\mathbf{x},N)$. In practice, these infinite weights can be avoided during computation. LIME uses the simplified input mapping which is equivalent to the approximation of the conditional value function from Equation 4.18. The intuition behind the derivation of the Shapley values via linear regression is that the summation in Equation 4.21 is equivalent to the Shapley value under the linearity and independence assumptions in Equations 4.17 and 4.18 given that the set of samples, $\mathbf{Z}=2^{|\mathbf{X}|}$. The resulting linear function g is therefore the closest linear model to the Shapley vales. As the size of $\mathbf{Z}$ increases, the coefficients of g have been shown to converge on the Shapley values. As a consequence, the Shapley values can be computed using weighted linear regression.

The big difference between Kernel SHAP and vanilla LIME is the weighting of samples in the approximating regression model. LIME weights samples according to how similar they are to the instance to be explained. Kernel SHAP weights samples dependent on the size of the coalition in consideration (the number of features "turned on" in sample $\sigma(\mathbf{z})$ such that coalitions with small number of features or large number of features get a higher weight.

In this section we have described how the limitations of regression coefficients in the presence of feature interaction and non-linear models originally motivated the use of the Shapley value for feature attribution. We have discussed the original intuition behind the approximation of feature removal by the conditional set function $f(\mathbf{x},S)=f(\mathbf{X}|\mathbf{X}_S=\mathbf{x}_S)$ and shown how the independence and linearity assumptions as proposed in [138] allow for computationally practical approximations including Kernel SHAP. We have discussed how the approximation of feature removal is non-trivial with each approach characterised by its own set of assumptions. In the next section we adopt a more critical perspective, introducing more recent value functions and categorise each according to its limitations.

In this section we have introduced the different value functions within the XAI literature which each approximate feature removal in a different way.

4.3 Value Functions And Feature Interaction

In Section 4.1.1 we discussed how, for a game in characteristic form (N,v) the Shapley value is a fully separable function which decomposes the value of the grand coalition $v(N)$ into a vector of length $|N|$ despite the function v not being fully separable as a result of coalitional interaction effects. The Shapley value for feature attribution decomposes the value of the grand coalition $v(\mathbf{x},N) = f(\mathbf{x},N) - f(\mathbf{x},\varnothing)$ among each feature in N. Similarly to coalitional games, the set function $v(\mathbf{x},S)$ may not be fully separable due to coalitional interaction effects.

Feature interaction may occur in the data, for example, $f(X_1,X_2,X_3) = X_1 + X_2 + X_3$ where $X_2 = \alpha X_3$ and/or in the model $f(X_1,X_2,X_3) = X_1 + 2X_2X_3$. The choice of value function determines the kind of interactive effects the Shapley value must allocate between features. In this section we discuss how the choice of value function determines the extent of feature interaction which is attributed by the Shapley value. While the following ideas have previously been discussed [1, 87, 117], we re-frame them here within the context of separability which allows us to motivate our proposed attribution method, Shapley Sets.

4.3.1 When Interaction Occurs In The Data

Example 4.3. *Consider the binary variable set* $\mathbf{X} = \{X_1,X_2,X_3\}$ *and function* $f(\mathbf{X}) = X_1 + X_3$ *where* X_2 *is the causal ancestor of* X_3 *such that* $X_3 = X_2$*. It is given that* $\mathbb{E}[X_1] = \mathbb{E}[X_2] = \mathbb{E}[X_3] = 0$

From Example 4.3, it is clear that X_2 has no impact on $f(\mathbf{X})$ from the perspective of the model. However, from the perspective of the data distribution, X_3 is dependent on X_2. Changing X_2 will result in a change in X_3 therefore, changing X_3 to a value non-consistent with X_2 does not make sense. Whether to consider X_2 as a separate player in the game and attribute value despite it having no direct influence on the model output is an open debate in the literature [117].

Off-manifold Value Functions There are those who argue that features with no impact on the model should receive no attribution [95, 146]. These methods re-frame the feature attribution challenge by considering feature values as causes of the prediction of a given sample $\mathbf{x}$. Rather than considering the causal system being modelled by the function f, these methods, which were originally formalised by [95], consider only the causal structure of the input/output system for a given prediction, separating the true values of the variables from their value as inputs to the model. These methods break all statistical relationships between variables by using a value function which calculates the impact of each feature on the model independently of its impact on the distribution of other features. This approach was formalised as v_{marg} by Janzing et al. [95]

and samples each feature from their marginal joint distribution without conditioning.

$$v_{marg}(\mathbf{x}, S) = f(\mathbf{X}_S = \mathbf{x}_S, \mathbb{E}[\mathbf{X}_{\bar{S}}]) - f(\mathbb{E}[\mathbf{X}]) \tag{4.22}$$

We note that v_{marg} is equivalent to the approximation of the conditional value function under the independence and additivity assumptions as given in Equation 4.18 by Lundberg et al. [138] but Janzing et al. [95] were the original motivators of v_{marg} from a causal perspective. The expectation is usually taken over the input distribution $\mathbf{X}_{input}$ which attributes to each feature the difference between the prediction of the sample and the expected prediction of model over the empirical input distribution. If this is replaced by an arbitrary distribution, v_{marg} can be generalised to the baseline value function, v_{bs} [205]. v_{bs} uses an arbitrary baseline sample $\mathbf{z}$ (which could be the vector $\mathbb{E}[\mathbf{X}]$), to approximate the missing features $\mathbf{X}_{\bar{S}}$

$$v_{bs}(\mathbf{x}, \mathbf{z}, S) = f(\mathbf{X}_S = \mathbf{x}_S, \mathbf{X}_{\bar{S}} = \mathbf{z}_{\bar{S}}) - f(\mathbf{X} = \mathbf{z}) \tag{4.23}$$

The baseline value function can thus be used to attribute the difference between the prediction of the sample to be explained and an arbitrary reference sample, which is often taken as an uninformative baseline: a vector of zeros, for example.

There are those who argue that attributions independent of the statistical interactions in the data are inherently misleading [87, 91]. Firstly from a causal perspective, if we consider Example 4.3, the Shapley value via v_{marg} would assign zero importance to X_2. An attribution ignoring that X_2 is directly responsible for X_3 could be misleading. If we consider Example 4.3 as a model where $f(\mathbf{X})$ is the probability of a cancer diagnosis, X_2 is smoking and X_3 is tar buildup on the lungs, an attribution which ignores that smoking is directly responsible for tar build up on the lungs is misleading, especially if the person receiving the attribution is looking for recommended changes to reduce their probability of a cancer diagnosis.

v_{marg} evaluates the model on out of distribution samples. To see this, if we break the causal relationship between X_2 and X_3, and using their expected values $\mathbb{E}[X_2] = \mathbb{E}[X_3] = 0$ in v_{marg}. The model is evaluated on samples $(x_1, 1, 0)$ which is a complete misrepresentation of the dependency between variables which exist in reality, where in our example, we would never expect to see an individual who smoked who did not also have tar buildup on the lungs. Furthermore, these samples lie well outside the sample distribution on which the model was trained and the resulting predictions may be unreliable [117]. The predictions on off-manifold samples are thus not necessarily relevant to the task of explaining an in-distribution sample yet the attributions will be affected by them.

On-manifold Value Functions To combat this problem, on-manifold samples can be calculated by the use of the conditional value function (Equation 4.24). It is re-defined here as v_{cond}, and is the original value function which, unlike v_{bs} and v_{marg} does consider statistically related features as separate players in the game, allowing the distribution of out of coalition features to be impacted by the feature in question.

$$v_{cond}(\mathbf{x}, S) = \mathbb{E}[f(\mathbf{X}_S = \mathbf{x}_S, \mathbf{X}_{\bar{S}})|\mathbf{X}_S = \mathbf{x}_S] - \mathbb{E}[f(\mathbf{X})] \tag{4.24}$$

v_{cond} is often taken as the observational conditional probability whereby the expected conditional is calculated over $\mathbf{X}_{input}$. This generates on-manifold data samples which remove the problems of off-manifold value functions discussed above. Furthermore, features which have no direct impact on the model but an indirect impact through other features are assigned a non zero importance, more accurately reflecting reality.

However, there are two significant issues with v_{cond}. The first is its computational complexity, requiring the evaluation of the model on 2^N multivariate conditional distributions. The second being the undesirable impact of considering all features as players combined with the efficiency axiom. In assigning non-marginal features a non-zero importance, v_{cond} can give misleading explanations which indicate features to change despite having zero impact on the outcome. This weakness of v_{cond} has been formalised as a violation of sensitivity: *When the relevance of ϕ_i is defined by $v_{cond}, \phi_i \neq 0$ does not imply that f depends on X_i* [95].

The failure of sensitivity exhibited by v_{cond} leads to further issues with the generated attributions. To see this consider Example 4.4 where X_3, X_2, X_1 are binary variables as in Example 4.3 but now we have that $X_2 = X_3$ (X_3 is the causal ancestor of X_2).

Example 4.4. *Consider the binary variable set $\mathbf{X} = \{X_1, X_2, X_3\}$ and function $f(\mathbf{X}) = X_1 + X_3$ where X_3 is the causal ancestor of X_2 such that $X_2 = X_3$.*

The simplified reference distribution from which we calculate the conditional expectations for Example 4.4 is given in the following table.

X_1	X_2	X_3	$f(\mathbf{x})$
1	1	1	2
1	0	0	1
0	1	1	1
0	0	0	0

Given the input $\mathbf{x} = (x_1, x_2, x_3) = (1, 1, 1)$ and $f(x_1, x_2, x_3) = 2$ Under v_{cond}, the Shapley attributions are as follows: $\phi_{X_2} = \phi_{X_3} = \frac{1}{4}$, $\phi_{X_1} = \frac{1}{2}$ where the full calculations can be found in Appendix A.1.

The Shapley values for X_2 and X_3 are equal and less than than the attribution for X_1 despite X_3 having an equal influence on the prediction as X_1 from the perspective of the model. Clearly, the attribution of X_2 violates sensitivity. Now, consider an alternative function which is just trained on two features X_1, X_3. As $X_3 = X_2$, $f_2(X_1, X_3) = f(X_1, X_2, X_3)$. However, now the Shapley values for X_1 and X_3 are equal. The relative apparent importances of X_1 and X_3 depend on whether X_2 is considered to be a third feature, even though the two functions are effectively the same. This problem is exacerbated when v_{cond} is used in the calculation of high-dimensional

Shapley values where there may be a large number of statistically interacting features without a direct impact on the model. which opens up the following challenge: How do we capture the interactive effects of the data while remaining robust to how we define the model?

Frye et al. [64] propose a solution to the failure of sensitivity exhibited by v_{cond} following the intuition: If X_i is known to be the deterministic causal ancestor of X_j ,one might want to attribute all effect to X_i and none to X_j. They propose asymmetric v_{cond} which, rather than equally dividing the interactive gain between interacting features, all the interactive gain is given entirely to the causal ancestor. For Example 4.4 all the interactive gain would be given to X_3 rather than X_2 and therefore the Shapley values of X_1 and X_3 would be the same under both f and f_2.

In contrast, Heskes et al. [87] argue that the only way to remove the problems arising from the failure of sensitivity is to replace the observational v_{cond} with the interventional conditional distribution based on the intuition that to work out the true impact of intervening on a feature, interventions need to be placed on the the observational distribution, not just conditional probabilities which gives rise the following interventional value function.

$$v_{interventional}(\mathbf{x}, S) = \mathbb{E}[f(\mathbf{X}|do(\mathbf{X}_S = \mathbf{x}_S)] - E[f(\mathbf{X}] \tag{4.25}$$

$do(\mathbf{X}_S = \mathbf{x}_S)$ refers to Pearl's do-calculus [164], which facilitates the identification of causal effects from observational data and is introduced formally in Chapter 5. It has been argued that the interventional value function is the only value function which results in truly causal Shapley values [87].

Both the asymmetric and interventional attributions above require the specification of the causal structure of the phenomenon being modelled. It has been argued [117] that this requirement is a significant limiting factor in the adoption of either approach (although both methods provide a framework for incorporating only partial causal knowledge). In this chapter we propose a novel attribution method, Shapley Sets (Section 4.4) which can be used with on and off-manifold value functions. Under v_{cond}, Shapley Sets finds a partitioning of the variable set which is robust to the inclusion of statistically interacting yet dummy features, generating on-manifold attributions which avoid the failure of sensitivity without requiring any knowledge of the causal structure of the underlying data distribution.

In this section we have shown how and why, in the presence of feature interaction in the data, the Shapley value sometimes generates misleading explanations.

4.3.2 When Interaction Occurs In The Model

While off-manifold value functions ignore interaction in the data, both on and off-manifold value functions capture interaction in the model. It has been recognised, however, that the Shapley value, in the presence of feature interaction in the model, sometimes generates misleading attributions [116].

Example 4.5. *Consider the function* $f(X_1, X_2, X_3) = X_1 + 2X_2X_3$ *and assume that each of the three features are statistically independent, i.e. all interaction between features is defined entirely by the model.*

Given the baseline value function v_{bs} with $z = \{0,0,0\}$, the Shapley value, applied to Example 4.5 gives equal attribution to each feature $\phi_{X_1} = \phi_{X_2} = \phi_{X_3} = 1$. Full calculations can be found in Appendix A.2.

While this attribution makes sense from the perspective of how much each feature contributes to the change in prediction between sample and baseline, it does not reflect the true behaviour of the model where changing the value of X_2 or X_3 would have double the impact on the prediction as changing X_1. Furthermore, the Shapley value under both on and off-manifold value functions are sensitive to the approximation of missing features and has been shown to violate the symmetry axiom [95, 205].

Example 4.6. *Consider the simple function* $f = X_1 + X_2$ *where* X_1 *and* X_2 *are both independent variables distributed uniformly over the set* $\{0,1,2\}$.

Given Example 4.6 and the sample $\mathbf{x} = (2,2)$ and two different reference values, $\mathbf{z}_1 = (0,0)$ and $\mathbf{z}_2 = (0,1)$, the Shapley values under v_{bs} given $\mathbf{z}_1$ are $\phi_{X_1} = \phi_{X_2} = 2$ but under $\mathbf{z}_2$, the Shapley values are $\phi_{X_1} = 2$, $\phi_{X_2} = 1$. From the perspective of the model, X_2 and $X1$ are interchangeable yet under $\mathbf{z}_2$, the attribution of X_2 is double that of X_1 and this, as argued by Sundararajan et al. [205], constitutes a violation of the symmetry property. As the features X_2 and X_1 are independent, this violation of symmetry is exhibited by both on and off-manifold value functions. It was argued by Janzing et al. [95] however, that if we consider the effect of *changing* each feature from its "off" value to its "on" value then the larger attribution awarded to X_2 is intuitive as the difference between its "on" and "off" value is greater than that of X_1 and thus has a greater impact on the overall change in prediction. Janzing et al. [95] argue therefore that this behaviour is not undesirable given careful comparison with the relative values of missing features (which we call a removal value) in the baseline.

We now show that the sensitivity of the Shapley value to the selection of reference values **does** exhibit undesirable behaviour in the presence of feature interaction in the model such that the relative feature attributions of interacting features not only depends on their removal value but also on the removal values of the features they interact with. If we consider again Example 4.5, but now under the baseline $\mathbf{z}_2 = (0,0,\frac{1}{2})$, we obtain the Shapley attribution vector of $\phi_{X_1} = 1$,

$\phi_{X_2} = \frac{3}{2}$, $\phi_{X_3} = \frac{1}{2}$. Full calculations can be found in Appendix A.2. Comparing these to the Shapley values obtained under the baseline $\mathbf{z} = (0,0,0)$ We can see that although the removal value for X_2 is constant over both $\mathbf{z}_1, \mathbf{z}_2$, its Shapley value is different as it is dependent on the removal value of X_3. Unlike in Example 4.6, where the violation of symmetry can be interpreted meaningfully given each feature's removal value, as we are unaware of the true feature interactions in the model, we are therefore unable to meaningfully interpret the violation of symmetry between interacting features as we are unable to cross-reference the relevant removal values (we would be uncertain as to why the Shapley value of X_2 is non-constant over the two baselines $\mathbf{z_1}, \mathbf{z_2}$).

Similarly to the problems invoked by the violation of sensitivity under v_{cond}, the dependence of feature attributions on the removal values **of other features** are particularly problematic in high-dimensional feature sets where the ground truth interaction is unknown. Our proposed attribution method, Shapley Sets, unlike the Shapley value, finds a partition of the variable set which attributes to groups of interacting variables in the model. Under both on and off-manifold value functions, Shapley Sets generates feature attributions which are robust to the removal values of other interacting features.

In this section we have shown how and why, in the presence of feature interaction in the model, the Shapley value sometimes generates misleading explanations.

4.4 Shapley Sets Of Non-Separable Variable Groups

The problems with Shapley value attributions discussed above occur as it assigns individual value to variables belonging to Non-Separable Variable Groups (NSVGs) in regards to the underlying partially separable function f (Definition 4.3). NSVGs are used to describe the formed variable groups $\{\mathbf{X}_1, ..., \mathbf{X}_k\}$ after a complete (or ideal) decomposition of f. An NSVG can also be defined as the minimal set of all interacted variables given the function f which we explicate in Definition 4.10.

Definition 4.10 (Non-Separable Variable Group (NSVG))**.** Let f be a partially separable function $f : \mathscr{X} \to \mathbb{R}$ satisfying Definition 4.3 with variable set $\mathbf{X} = \{X_1, X_2, ..., X_n\}$. Given the notation $f(\mathbf{x})_S = f(\mathbf{X}_S = \mathbf{x}_S, \mathbf{X}_{\bar{S}} = \mathbf{x'}_{\bar{S}})$ such that $S \cup \bar{S} = N$, if there exists any two candidate decision vectors $\mathbf{x} = \{x_1, ..., x_n\}$ and $\mathbf{x'} = \{x'_1, ..., x'_n\}$, sampled from $\mathscr{X}$, such that the following property holds for any two mutually exclusive subsets $I, J \subset N$, $I \cap J = \varnothing$,

$$f(\mathbf{x})_{I \cup J} - f(\mathbf{x})_J \neq f(\mathbf{x})_I - f(\mathbf{x})_\varnothing, \tag{4.26}$$

then the sets $\mathbf{X}_I, \mathbf{X}_J$ are said to interact. As an NSVG refers to the minimal set of interacted variables, if $|I|$ and $|J|$ is minimized such that Equation 4.26 still holds, then $\mathbf{X}_I \cup \mathbf{X}_J$ is a NSVG.

(For proof, see [203]).

Mapping Definition 4.10 into a feature attribution context, given that $f(\mathbf{x})_S$ is a function over the domain of all the possible subsets of $S \subseteq N$, we can rewrite Equation 4.26 in terms of $v(\mathbf{x}, S)$, where v could represent any of the value functions we have previously discussed but here we focus on $v \in \{v_{cond}, v_{bs}\}$ as two representative on and off-manifold value functions. If we set $I = \{i\}$ and $J = S$, we can then say that for the baseline value function v_{bs}, given that $|S|$ is minimised, if there exist any candidate vectors $\mathbf{x}, \mathbf{x}' \in \mathscr{X}$ such that

$$v_{bs}(\mathbf{x}, \mathbf{x}', \{i\} \cup S) - v_{bs}(\mathbf{x}, \mathbf{x}', S) \neq v_{bs}(\mathbf{x}, \mathbf{x}', \{i\}) \tag{4.27}$$

then $\{i\} \cup S$ is a non-separable variable set. For example, given the partially separable function from Example 2, settings $\mathbf{x} = (1,1,1)$ and $\mathbf{x}' = (0,0,0)$ and the baseline value function v_{bs}, $\{X_2, X_3\}$ is a NSVG as

$$v_{bs}(\mathbf{x}, \mathbf{x}', \{3,2\}) - v_{bs}(\mathbf{x}, \mathbf{x}', \{2\}) \neq v_{bs}(\mathbf{x}, \mathbf{x}', \{3\}) \tag{4.28}$$

In contrast, there exists no candidate vector $\mathbf{x}$ or $\mathbf{x}'$ from the domain of $\mathbf{X}$ for which Equation 4.27 holds for X_1, X_1 is thus a singleton NSVG, resulting in the partition of the variable set $\{\{X_1\}, \{X_2, X_3\}\}$ for the function as defined in Example 2 under baseline value function.

For the conditional value function v_{cond}, given that $|S|$ is minimised, if there exists any candidate vector $\mathbf{x} \in \mathscr{X}$ such that

$$v_{cond}(\mathbf{x}, \{i\} \cup S) - v_{cond}(\mathbf{x}, S) \neq v_{cond}(\mathbf{x}, \{i\}) \tag{4.29}$$

then $\{i\} \cup S$ is a NSVG. Given the partially separable function from Example 1, setting $\mathbf{x} = (1,1,1)$ and the conditional value function, v_{cond}, the set $\{X_2, X_3\}$ is a NSVG

$$v_{cond}(\mathbf{x}, \{3,2\}) - v_{cond}(\mathbf{x}, \{2\}) \neq v_{cond}(\mathbf{x}, \{3\}) \tag{4.30}$$

There exists no candidate vector $\mathbf{x}$ for which Equation 4.29 holds for X_1. X_1 is thus a singleton NSVG, resulting in the partition of the variable set $\{\{X_1\}, \{X_2, X_3\}\}$ for the function as defined in Example 1 under the conditional value function.

In this section we have introduced the concept of a Non-Separable Variable Group which describe groups of interacting features with regards to an underlying model.

4.4.1 Shapley Sets

In this section, we introduce our attribution method which, unlike the Shapley value does not separate NSVGs to assign attribution. We work under the intuition that any interacting feature

whether that be in the model or in the data should not be considered as separate players in the coalitional game but should be awarded value together. In both the examples above, X_2 and X_3 would receive joint attribution under our proposed method.

Given the partially separable function f, variable set $\mathbf{X} = \{X_1, X_2..., X_n\}$, and a specified value function $v_{cond}(\mathbf{x}, S)$ or $v_{bs}(\mathbf{x}, S)$, our proposed solution concept φ, which we term Shapley Sets, first decomposes the variable set $\mathbf{X}$ into $m > 1$ NSVGs $\{\mathbf{X}_1, ..., \mathbf{X}_m\}$. The resulting variable grouping $\{\mathbf{X}_1, ..., \mathbf{X}_m\}$ satisfies Definition 4.10 such that each $\mathbf{X}_i$ is an NSVG, each a minimum set of interacting features. From Definition 4.3, we know that the sum of the function evaluated on each NSVG is equivalent to the value of the function evaluated on the entire variable set for every possible candidate vector $\mathbf{x} \in \mathscr{X} : f(\mathbf{x}) = \sum_{i=1}^{m} v(\mathbf{x}, \{i\})$. Given a particular prediction to be attributed, $f(\mathbf{x})$, our proposed attribution Shapley Sets φ, therefore returns the attribution for each variable group $\mathbf{X}_i, \forall i \in m$ as:

$$\varphi_{\mathbf{X}_i} = v(\mathbf{x}, \mathbf{X}_i) \tag{4.31}$$

Proposition 4.1. *If we model each NSVG, $\mathbf{X}_i \in \{\mathbf{X}_1, ..., \mathbf{X}_m\}$ as a super-feature Z_i such that $\mathbf{Z} = \{Z_i, ..., Z_m\}$, $z_i = \mathbf{x}_i$ and $\mathbf{z} = \{z_1, ..., z_m\}$, then the Shapley value of each super feature $\phi_{Z_i}(v, \mathbf{z})$ is equivalent to $v(\mathbf{z}, \{i\})$*

Proof. Consider the game (M, v) described above, where M is the variable set of super features $\{Z_1, ..., Z_m\}$ and v is either v_{cond} or v_{bs}. The Shapley Value for super feature Z_i can be written as

$$\phi_{Z_i}(v, \mathbf{z}) = \sum_{S \subseteq M \setminus \{i\}} \alpha [v(\mathbf{z}, \{i\} \cup S) - v(\mathbf{z}, S)]$$

where $\alpha = \frac{|S|!(|M|-|S|-1)!}{|M|!}$

Given that each $Z_i \in \{Z_1, ..., Z_m\}$ is a NSVG, from Definition 4.10 we know that for any two $I, J \subseteq M$

$$v(\mathbf{z}, I \cup J) - v(\mathbf{z}, J) = v(\mathbf{z}, I)$$

Therefore, setting $I = \{i\}$ and $J = S$ gives the following

$$v(\mathbf{z}, \{i\} \cup S) - v(\mathbf{z}, S) = v(\mathbf{z}, \{i\})$$

.

It follows that given

$$\alpha = \sum_{S \subseteq M \setminus \{i\}} \frac{|S|!(|M| - |S| - 1)!}{|M|!} = 1$$

the Shapley value for super feature Z_i can be rewritten as

$$\phi_{Z_i}(v,\mathbf{z}) = \sum_{S \subseteq M \setminus \{i\}} \alpha v(\mathbf{z},\{i\}) = v(\mathbf{z},\{i\})$$

■

Proposition 4.1 states that $\varphi_{\mathbf{X}_i}(v,\mathbf{x})$ is equivalent to the Shapley value when played over the feature set $\mathbf{Z}$ containing the set of NSVGs $\{Z_1,...,Z_m\} = \{\mathbf{X}_1,...,\mathbf{X}_m\}$ for a given $v \in \{v_{cond}, v_{bs}\}$. Shapley Sets therefore satisfies the same axioms of fairness as the Shapley value: efficiency, dummy, additivity and symmetry when played over this feature set. However, we have seen in Section 1.2 how these axioms can be violated when approximating the Shapley value for feature attribution. In Section 4.6 we show how Shapley Sets, by attributing interacting features together, avoids these violations and displays practical advantages over the Shapley value. First however, we provide a method for computing the decomposition of f into its NSVGs.

In this section we have proposed our attribution method which, given the set of NSVGs with regards to an underlying function, attributes value to each NSVG rather than individual features. We have shown that our attribution is equivalent to the Shapley value when played over the transformed feature set and thus satisfies the same set of axioms as the Shapley value. The following section shows our method for automatically decomposing a function into its NSVG set.

4.5 Computing Shapley Sets

Determining the NSVGs of a function f could be achieved manually by partitioning the variable set and determining interaction over every possible candidate vector. However, this would be computationally intractable. Instead, there exists a large body of literature surrounding function decomposition in global optimization problems. Of this work, automatic decomposition methods identify NSVGs.

4.5.1 Automatic Function Decomposition Methods

Function decomposition refers generally to the process by which a functional relationship is broken down into its constituent parts in such a way that the original function can be reconstructed. Function decomposition has many applications including providing insight into the identity of constituent function components or to gain a compressed representation of the function. Within the world of global optimisation, function decomposition splits the optimisation problem into sub-components which can be optimised separately, allowing for the application of various optimisation methods at scale via a divide-and-conquer approach. Ideally, the sub-components should

be formed according to the interaction pattern of the decision variables so that the interactions between the sub-components are kept to a minimum [140].

The processes by which the optimisation problem is decomposed can be categorised as either manual or automatic. Manual methods rely on a decomposition structure engineered by hand. Automatic methods, in contrast determine the structure of the sub-components by identifying interacting decision variables, or NVSGs. One of the first of these methods, Differential Grouping (DG)[158], iterates over every decision variable to determine the interacting groups. In the Extended DG algorithm (XDG) [202], XDG was extended to identify both interacting and conditionally interacting variables. Conditionally interacting variables interact only when they are placed in their sub-component. For example, for the function $f(\mathbf{X}) = X_1 \vee (X_2 \wedge X_3)$, X_1 conditionally interacts with X_2 and X_3 but not X_2, X_3 individually.

While a theoretical improvement on DG, the uptake of the XDG was limited by its computational complexity of n^2 where $|\mathbf{X}| = n$. More recently the Recursive Differential Grouping method, RDG, was proposed [203] which adapts XDG to recursivley identify the interaction between sets of variables with log-linear computational complexity, overcoming the computational barrier of XDG. The following section gives our method for calculating Shapley Sets which is based on the Recursive Decomposition Grouping algorithm (RDG) as introduced in [203].

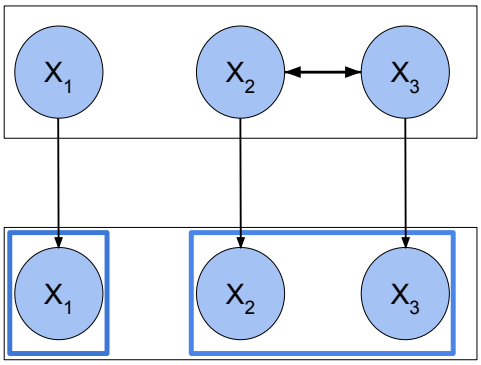


Figure 4.1: Figure shows the interaction structure of the variables (box above) and resulting grouping (box below) for the Shapley Sets algorithm under v_{bs} when applied to the function and variable set in Example 4.5

.

4.5.2 The Shapley Sets Algorithm

Given the variable set $\mathbf{X}$ and partially separable function f, to identify whether two sets of variables $\mathbf{X}_I$ and $\mathbf{X}_J$ interact, DG, XDG, and RDG and all use a fitness measure, Equation 4.32, with candidate vectors, $\mathbf{x}, \mathbf{x}'$ set as the lower and upper bounds of $\mathscr{X}$

Algorithm 6 ValueInteract(I, J)

Require: $v \in \{v_{bs}, v_{cond}\}, \mathbf{X}_{input}, \epsilon$
  if v_{bs} **then**
    Sample $\mathbf{x}, \mathbf{x}'$ from input distribution $\mathbf{X}_{input}$
    $\sigma_1 = v(\mathbf{x}, \mathbf{x}', I \cup J) - v(\mathbf{x}, \mathbf{x}', J)$
    $\sigma_2 = v(\mathbf{x}, \mathbf{x}', I)$
  end if
  if v_{cond} **then**
    Sample $\mathbf{x}$ from input distribution $\mathbf{X}_{input}$
    $\sigma_1 = v(\mathbf{x}, I\cup) - v(\mathbf{x}, J)$
    $\sigma_2 = v(\mathbf{x}, I)$
  end if
  if $|\sigma_1 - \sigma_2| > \epsilon$ **then**
    if $|J| == 1$ **then**
      $I = I \cup J$
    else
      Split J into two equal groups G_1, G_2
      I^1 = ValueInteract(I, G_1)
      I^2 = ValueInteract(I, G_2)
      $I = I^1 \cup I^2$
    end if
  end if
  Return I

$$|f(\mathbf{x})_{I\cup J} - f(\mathbf{x})_J - [f(\mathbf{x})_I - f(\mathbf{x})_\varnothing]| \geq \epsilon \tag{4.32}$$

If the difference between the left and right hand side of Equation 4.26 meets some threshold $\epsilon = \alpha \min\{|f(\mathbf{x}_1)|, ..., |f(\mathbf{x}_k)|\}$ where $\{\mathbf{x}_1, ..., \mathbf{x}_k\}$ are randomly selected candidate vectors, then $\mathbf{X}_I$ and $\mathbf{X}_J$ are deemed by RDG to interact. We note that the RDG algorithm is an approximation of the true NSVGs due to the use of singleton candidate vectors $\mathbf{x}, \mathbf{x}'$ to determine interaction between variable sets where the definition of NSVGs (Definition 4.10) requires a search over the space of all possible candidate decision vectors. As noted by Omidvar et al. [158], the selection of $\mathbf{x}, \mathbf{x}'$ as the upper and lower bounds of $\mathbf{X}$ is arbitrary, these candidate vectors can be selected randomly as long as they do not coincide. To adapt RDG for v_{cond} and v_{bs}, we therefore propose an alternative fitness measure, Definition 4.11, with candidate vectors $\mathbf{x}, \mathbf{x}'$ randomly sampled from the empirical input distribution $\mathbf{X}_{input}$ which can identify NSVGs in the function and/or in the model.

Definition 4.11 (Shapley Sets Fitness Measure)**.** Given two sets of variables $I, J \subseteq N$ and a specified value function, $v \in \{v_{bs}, v_{cond}\}$ If $|v_{cond}(\mathbf{x}, I \cup J) - v_{cond}(\mathbf{x}, J) - v_{cond}(\mathbf{x}, I)| > \epsilon$ then there is interaction between $\mathbf{X}_I$ and $\mathbf{X}_J$. Or, if $|v_{bs}(\mathbf{x}, \mathbf{x}', I \cup J) - v_{bs}(\mathbf{x}, \mathbf{x}', J) - v_{bs}(\mathbf{x}, \mathbf{x}', I)| > \epsilon$ then there is interaction between $\mathbf{X}_I$ and $\mathbf{X}_J$.

Algorithm 7 Shapley Sets (Adapted from RDG [203])

Require: $v \in \{v_{cond}, v_{bs}\}, \epsilon$, $\mathbf{x}_{inp}$, $\mathbf{x}_{ref}$ (if $v = v_{bs}$)
  Intialise $seps$ and $nonseps$ as empty groups
  Assign the first variable in $\mathbf{X}$ to $\mathbf{X}_I$
  Assign the rest of the variables in $\mathbf{X}$ to $\mathbf{X}_J$
  while $\mathbf{X}_J$ is not empty **do**
    $\mathbf{X'}_I \leftarrow ValueInteract(\mathbf{X}_I, \mathbf{X}_J)$
    if $\mathbf{X'}_I$ is the same as $\mathbf{X}_I$ **then**
      if $\mathbf{X}_I$ contains one variable **then**
        $seps \leftarrow \mathbf{X}_I$
      else
        $nonseps \leftarrow \mathbf{X}_I$
      end if
      Empty $\mathbf{X}_I$ and $\mathbf{X}'_I$
      Assign the first variable of $\mathbf{X}_J$ to $\mathbf{X}_I$
      Delete the first variable of $\mathbf{X}_J$
    else
      $\mathbf{X}_I = \mathbf{X'}_I$
      Delete the variables of $\mathbf{X}_I$ from $\mathbf{X}_J$
    end if
  end while
  For each NSVG $\mathbf{X}_I \in seps \cup nonseps$ return $v_{cond}(\mathbf{x}_{inp}, I)$ or $v_{bs}(\mathbf{x}_{inp}, \mathbf{x}_{ref}, I)$

We substitute the Shapley Sets fitness measure into the RDG algorithm which identifies NSVGs by recursively identifying sets of interacting variables. If $\mathbf{X}_J$ and $\mathbf{X}_I$ are said to interact they are placed into the same NSVG $\mathbf{X}_I$. At which point conditional interaction between $\mathbf{X}_I$ and the remaining variables is identified. The algorithm iterates over every variable $X_i \in \mathbf{X}$ and returns the set of NSVGs. To compute the Shapley Sets attributions for a given prediction $f(\mathbf{x})$ we compute $v(\mathbf{x}, I)$ for each NSVG. Our full algorithm is shown in Algorithm 4.5.1. The runtime of Shapley Sets is $O(nlogn)$ as proven in [203]. Figure 4.1 shows the interaction structure and resulting variable grouping when applied to the prediction Example 2 under v_{bs}. The resulting Shapley Set values attributions for the prediction $f(1,1,1)$ with $z = (0,0,0)$ would be $\varphi_{X_1} = 1$, $\varphi_{\{X_2,X_3\}} = 2$

In this section we have proposed our algorithm Shapley Sets, inspired by Recursive Function Decomposition, as a method which first decomposes a value function into its NSVGs and then attributes value to each group.

4.6 Motivating Shapley Sets

While Shapley Sets is designed to be used with any value function and thus model agnostic, the selection of the value function v determines the partition of the variable set. Used with an off-manifold value function, v_{bs}, as interacting features in the model are placed in the same NSVG, the attributions resulting from Shapley Sets will be more faithful to the underlying model.

The Shapley Sets attribution for v_{bs} in Example 4.5 with sample $\mathbf{x} = (1,1,1)$ and reference sample $\mathbf{z} = (0,0,0)$ would be $\varphi_{X_1} = 1$ and $\varphi_{X_2,X_3} = 2$. To see why the attribution afforded by Shapley Sets is more useful than that of the Shapley value consider Example 4.5 to be a model used by a bank to approve or deny a loan. If we take X_1, X_2 and X_3 to represent the applicant's education level, credit score and current overdraft respectively and $f(\mathbf{x})$ to represent the undesirable outcome and $E[f(X)]$ as the desirable outcome. The attribution provided by the Shapley value under v_{bs} would indicate that changing any of the individuals current feature values would all equally move the individual's current score towards the desirable prediction. However, the attribution provided by Shapley Sets under v_{bs} would imply that changing X_2 and X_3 would result in double the impact on the outcome than changing X_1.

This attribution is particularly useful in situations where there is varying difficulty in changing feature values. Changing an individual's level of education is, for example, much more challenging than changing a credit score or overdraft limit. Furthermore, under both on and off-manifold value functions, as the attributions under Shapley Sets automatically groups interacting features together they are robust to the removal values of other interacting features.

We have seen how the Shapley value applied to Example 4.5 under the two reference samples $\mathbf{z} = (0,0,0)$ and $\mathbf{z_2} = (0,0,\frac{1}{2})$ generates misleading attributions due to the dependence on removal values of interacting features and an unknown interaction structure (Section 4.3.2). In contrast, under Shapley Sets, we gain knowledge of the interaction structure, thus we are able to cross-reference the relevant removal values of interacting features to meaningfully interpret the non-symmetric attributions. For Example 4.5 with reference $\mathbf{z_2} = (0,0,\frac{1}{2})$, the Shapley Sets attributions remain the same as under $\mathbf{z} = (0,0,0)$. As we know that $X_2$and X_3 interact, from this attribution we can infer that the relationship between $X_2$and X_3 is multiplicative due to the constant attribution over the two baselines despite the non-constant removal value X_3. While the attribution generated by Shapley Sets is not explicit in the type of interactive effect that exists between features, the fact that interacting features are identified allows their resulting attributions to be interpreted and compared meaningfully, offering insight into the type of interactive effect.

Used with an on-manifold value function, v_{cond}, as interacting features are placed in the same NSVG, the attributions resulting from Shapley Sets do not suffer from the problems of sensitivity as described in Section 4.3.1. Consider again Example 4.4, as X_2, X_3 now belong to a NSVG, the Shapley Sets attributions for X_1, X_3 are now equal across both f and f_2 and therefore robust to whether non-directly impacting features are included in the model. Shapley Sets offer a further

advantage when used to compare the attributions under on and off-manifold examples. Consider again Example 4.5 yet now with $X_1 = \alpha X_2$. The Shapley Sets attribution via v_{marg} would be $\varphi_{X_1} = 1$ and $\varphi_{X_2,X_3} = 2$. However, if Shapley Sets was calculated via v_{cond} $\varphi_{\{X_1,X_2,X_3\}} = 3 - \mathbb{E}[f(\mathbf{X})]$ indicating that f is non-separable and all the features interact. The comparison between on and off-manifold Shapley Sets therefore indicate *where* the feature interaction takes place.

We have thus far provided an alternative attribution method to the Shapley Value, Shapley Sets which can be computed in $O(nlogn)$ time with n being the number of features. Shapley Sets can be adapted for arbitrary value functions and offers several advantages over Shapley value-based attributions when used with on and off-manifold value functions.

In this section we have argued how Shapley Sets addresses the limitations of the Shapley value when used with on and off-manifold value functions.

4.7 Experimental Motivation Of Shapley Sets

We have previously discussed the challenges surrounding the quantitative evaluation of local explanation methods in Chapter 1. These challenges are further exacerbated by the fact that Shapley Sets attributes to sets of features rather than singletons. This section therefore takes care to outline the assumptions and justification underpinning our evaluation of Shapley Sets, focusing on tabular data to experimentally validate the claims of Section 4.6.

We begin here with two synthetic experiments. The first of these motivates the use of Shapley Sets in the presence of interaction in the model. The second motivates the use of the Shapley Sets in the presence of interaction in the data. We then compare Shapley Sets to existing Shapley value (SV) based attribution methods on three benchmark datasets. We first outline how the value functions v_{bs}, v_{cond} are computed for our experiments and detail the experimental setups.

Calculation Of Our Off-Manifold Value Function For our experiments we select v_{marg} (Equation 4.22). The expectation is taken over the empirical input distribution $\mathbf{X}_{input}$.

Calculation Of Our On-Manifold Value Function: For the calculation of v_{cond} (Equation 4.24, as the true conditional probabilities for the underlying data distribution are unknown we approximate $p(\mathbf{X}_{\bar{S}}|\mathbf{X}_S = \mathbf{x}_s)$ using the underlying data distribution. Approximating conditional distributions can be achieved by directly sampling from the empirical data distribution. However, as noted in [1], this method of approximating $p(\mathbf{X}_{\bar{S}}|\mathbf{X}_S = \mathbf{x}_s)$ suffers when $|\mathbf{X}_S| > 2$, due to sparsity in the underlying empirical distribution.

We adopt the approach of Aas et al. [1], where under the assumption that each $X \in \mathbf{X}$ is sampled from a multivariate Gaussian with mean vector μ and covariance matrix Σ, the conditional distribution $p(\mathbf{X}_{\bar{S}}|\mathbf{X}_S)$ is also multivariate Gaussian such that $p(\mathbf{X}_{\bar{S}}|\mathbf{X}_S = \mathbf{x}_s) = \mathcal{N}_{\bar{S}}(\boldsymbol{\mu}_{\bar{S}|S}, \Sigma_{\bar{S}|S})$ where $\boldsymbol{\mu}_{\bar{S}|S} = \boldsymbol{\mu}_{\bar{S}} + \Sigma_{\bar{S}S}\Sigma_{SS}^{-1}(\mathbf{x}_S - \boldsymbol{\mu}_S)$ and $\Sigma_{\bar{S}|S} = \Sigma_{\bar{S}\bar{S}} + \Sigma_{\bar{S}S}\Sigma_{SS}^{-1}\Sigma_{S\bar{S}}$. We can therefore

sample from the conditional Gaussian distribution with expectation vector and covariance matrix given by $\boldsymbol{\mu}_{\bar{S}|S}$ and $\Sigma_{\bar{S}|S}$ where $\boldsymbol{\mu}$ and Σ are estimated by the sample mean and covariance matrix of $\mathbf{X}_{input}$.

4.7.1 Synthetic Experiment: Interaction In The Model

We first construct three functions with linear and non-linear feature interactions which are shown in Table 4.1. We next construct a synthetic dataset of seven features drawn independently from $\mathcal{N}(-1,1)$. For each of 100 randomly drawn samples we compute Shapley Sets under v_{marg}. As $|\mathbf{X}| = 7$ we are able to compute the true SVs under v_{marg} for each feature, without relying on a sampling algorithm. As we know the ground truth we calculate the Mean Absolute Error (MAE) across all n features and k samples to be used as our evaluation metric,

$$MAE = \frac{1}{k}\sum_{j=1}^{k}\frac{1}{n}\sum_{i=1}^{n}|m(X_{ij}) - gt(X_{ij})|. \tag{4.33}$$

Here $m(X_{ij})$ is the attribution given by Shapley Sets m_{SS} or the Shapley value, $m_S V$ to feature i in sample j. As Shapley Sets calculates an attribution for a set of features, $m_{SS}(X_{ij}) = \varphi_{\mathbf{X_{ij}}}$ where $X_{ij} \in \mathbf{X}_{ij}$. The ground truth attribution $gt(X_{ij})$ is the ground truth value of all the non-separable components of the function. For example, given $f = 2(X_1X_2)$ and $\mathbf{x}_j = (1,1)$, $gt(X_{1,j}) = 2$ and $gt(X_{2,j}) = 2$.

Results are shown in Table 4.1. Shapley Sets is successful in decomposing each function into its NSVGs and the attributions awarded to each set matches the ground truth of the function giving MAE of zero for all samples and functions. Shapley Value attributions deviate from ground truth by dividing the value of non-separable variable set between each individual feature which results in misleading attributions, particularly in the presence of inverse relationships between features. For example consider the following function sub-component $(X_1)/(1-X_2)$, and a particular sample $\mathbf{x} = (1, 0.2)$. Shapley Value gives X_1 a positive attribution but X_2's attribution is negative. Under Shapley Sets, X_1 and X_2 are considered as non-separable and awarded a positive attribution together. From its Shapley Value attribution, a user may opt to change X_2 rather than X_1, however, as these features jointly move the outcome from the reference to the target, the impact of changing X_2 in isolation could be cancelled out by the impact of X_1.

	Shapley Sets	Shapley Value
$f_1(\mathbf{X}) = X_0 + (X_1/(2+X_4)) + 2(X_2 * X_3) + sin(2(X_5) + X_6)$	$\mathbf{0.000 \pm 0.000}$	0.335 ± 0.400
$f_2(\mathbf{X}) = 2(sgn(X_0)) + sgn(X_1X_2X_3) + sgn(X_4X_5X_6)$	$\mathbf{0.000 \pm 0.000}$	1.143 ± 0.990
$f_3(\mathbf{X}) = 2(X_0x_2X_3) + 4(X_4x_5) - 3(X_1)^2 - (X_6))$	$\mathbf{0.000 \pm 0.000}$	0.540 ± 0.580

Table 4.1: Table shows the MAE ± std, for Shapley Sets and Shapley Value attributions under v_{marg} for three functions. Shapley Sets perfectly identifies NSVGs for all three functions.

4.7.2 Synthetic Experiment: Interaction In The Data

We adopt the approach of Hooket et al. [91] and propose an underlying linear regression model $f(\mathbf{X}) = X_0 + 0.5X_1 + 0.8X_3 + 0.2X_2 + 0.5X_4$. We construct a synthetic dataset comprising five features $n = 5$. (X_2, X_3, X_4) are all modeled as i.i.d and drawn independently from $\mathcal{N}(-1, 1)$. X_0, X_1, however are modeled as dependent features where $X_1 = \rho X_0$. We generate a synthetic dataset X_{train}, X_{test} consisting $k = (2000, 100)$ samples of each feature and obtain the ground truth labels $\mathbf{y}_{train}, \mathbf{y}_{test} = f(\mathbf{X}_{train}), f(\mathbf{X}_{test})$. We next select a model g which is trained on $\mathbf{X}_{train}, \mathbf{y}_{train}$ to approximate f. We calculate the attributions for each sample in $\mathbf{X}_{test}$ generated by the Shapley Value under both v_{marg} and v_{cond} and the attributions from Shapley Sets under v_{cond}. To evaluate attributions we use the coefficients of the linear regression model as our ground truth attributions $c = \{1, 0.5, 0.8, 0.2, 0.5\}$. We use MAE (Equation 4.33) where the ground truth for feature i in sample j is given as $gt_{X_{ij}} = c_i x_{i,j}$. Off-manifold attributions in the presence of interaction in the data recover the ground truth attributions reliably when g is a linear model, however, that breaks down when non-linear models are used as the approximating function g [91]. We therefore compare attributions under g_1, a linear regression model, and g_2, an XGBoost model.

Results are shown in Table 4.2 where Shapley Sets outperforms Shapley Value on both g_1 and g_2. We now show experimentally the claim that Shapley Sets under on-manifold value function avoid the issues related to sensitivity. To do this we add a dummy variable $X_5 = X_0$ to the dataset $\mathbf{X}$ such that X_5 is not used by f. We train another XGBoost model, g_3 using the new dataset and generate the three sets of attributions as before. Results are shown in Table 2. Under the influence of the dummy, MAE of Shapley Value under v_{cond} increases, as the attribution of each of the non-dummy variables moves further away from its true value to accommodate the attribution of the new feature despite it having no effect on the true output. In contrast, as Shapley Sets includes this dummy feature in the non-separable set $\{X_0, X_1\}$. The resulting attribution to the existing features is unchanged and thus the MAE remains constant under the inclusion dummy variables, demonstrating SS's robustness to how the underlying phenomenon is modelled.

	Shapley Sets	Shapley Marg	Shapley Cond
g_1	$\mathbf{0.204 \pm 0.114}$	0.226 ± 0.121	0.211 ± 0.127
g_2	$\mathbf{0.071 \pm 0.031}$	0.082 ± 0.032	0.073 ± 0.031
g_3	$\mathbf{0.074 \pm 0.044}$	0.110 ± 0.068	0.150 ± 0.059

Table 4.2: Table shows the MAE ± std for Shapley Sets under v_{cond} and the Shapley Value under v_{cond} and v_{marg} for the three experiments outlined in Section 5.2. Shapley Sets has lower MAE than the Shapley Value for all models

4.7.3 Shapley Sets Of Real World Benchmarks

We now evaluate Shapley Sets on real data: the Diabetes, Boston and Correlation datasets from the SHAP library [139] as detailed below.

Models: The models used in the experiments include: Linear Regression (LR), Random Forest Regressor(RF), XGBoost Regressor (XGB). All the models are imported via the Scikit learn python library. Random Forest is initialised with a maximum depth of ten. The XGboost model is initialised with 100 estimators and a maximum depth of three. The R^2 metric was used to assess performance.

Boston Dataset: The Boston dataset, downloaded from the Shap python package [136] contains 506 samples of 14 features. The regression task predicts the median price of a home based on the regional attributes. We split the dataset into a train and test set of ratio $\frac{2}{3}, \frac{1}{3}$ and obtain a R^2 score of 0.90 using RF model.

Diabetes Dataset: The Diabetes dataset, downloaded from the Shap python package [136] contains 442 samples of 10 features. The regression task predicts the quantitative progression of the disease one year after a specific baselin. We split the dataset into a train and test set of ratio $\frac{2}{3}, \frac{1}{3}$ and obtain a R^2 score of 0.89 using the RF model.

Correlation Dataset: The synthetic Correlation dataset, downloaded from the Shap python package [136] contains 1000 samples of 60 tightly correlated features. We split the dataset into a train and test set of ratio $\frac{2}{3}, \frac{1}{3}$ and obtain a R^2 score of 0.47 using the XGBoost model. We standardise all real world datasets to lie in the range $(0, 1)$.

We compute Shapley Sets attributions for 100 randomly selected samples from the test set under both v_{marg} and v_{cond}. As the dimensionality of the datasets now exceed that capable of being computed by the true Shapley values we compare the Shapley Sets attributions with the most commonly used approximation techniques: Tree Shap (TS) [136] and Kernel SHAP (KS) [139]. KS is an approximation of off-manifold SHAP and breaks the relationship between input features and the data distribution. TS does not make this assumption and is presented as an on-manifold Shapley value approximation. However, in practice TS performs poorly when there is high dependence between features in the dataset [1].

To evaluate the attributions generated by SS, KS and TS in the absence of a ground truth attribution we use modified versions of the deletion and sensitivity measures which have been used widely across the literature [69]. Deletion is built on the intuition that the magnitude of a feature's importance score should reflect its impact on the output. If the features that were marked as important are truly important, we would expect the output of the model to drop rapidly as more of them are removed from the original instance.

There are many different formalisations of Deletion across the literature [69], where it has also been described as a measure of "Faithfulness" of the attribution method. While the intuition behind these implementations is the same, they vary in the number of important features that are removed, how the features are removed, and how the change in model output is determined. One of the simplest implementations of Deletion is that of Samek et al. [184] which iteratively removes the top p most important features from the input. It has been noted in the literature that with increasing p, or as we remove a larger number of features from the instance, the resulting

sample becomes vulnerable to out-of-distribution effects [69]. As such, we restrict $p = 1$ for our implementation of Deletion.

A further problem with Deletion is that it was originally designed for the evaluation of explanations on classification models. In this way, once the p most influential features have been removed from the instance, we would expect the resulting predicted class probability to also decrease from the original probability. Using the change in model output therefore makes intuitive sense as we assume a directionality in the way in which the prediction changes following feature removal.

When we apply the Deletion intuition to explanations of Regression models however, we do not make the same assumption of a decrease in model output following a feature removal. Under the Regression setting, an optimal attribution method would recognise the most influential features in the prediction $f(\mathbf{x})$, which as we have seen in this chapter compose additively under efficiency such that, $f(\mathbf{x}) = \sum_{i \in n} \phi_i(\mathbf{x}) + f(\mathbf{x}, X_\varnothing)$. In this way, a good attribution method would identify the most salient feature such that when removed from the input, the distance between the function evaluated on the perturbed sample and the function evaluated on the baseline $f(\mathbf{x}, X_\varnothing)$ is minimised. This intuition is behind our adaptation of the Deletion metric. Our metric, Average Deletion (AD), Equation 4.34, therefore measures the absolute distance between the baseline prediction, $v(\mathbf{x}, \mathbf{X}_\varnothing)$ and the prediction of a given sample $v(\mathbf{x}, \mathbf{X})$ after the most important feature $X_i' = x_i'$, determined by the attribution method under consideration m, has been masked in the original sample.

$$AD = \frac{1}{k} \sum_{i \in k} |v(\mathbf{x}_i, \varnothing) - v(\mathbf{x}_i, N \backslash \{i\})| \tag{4.34}$$

where $\mathbf{X}_i'$ is the most important feature, or NSVG for the sample $\mathbf{x}_i$ as identified by the attribution method.

Low AD indicates that the attribution technique has correctly identified an important feature to remove. As Shapley Sets attributes to sets of features we allow $\mathbf{X}'$ to be a NSVG as generated by Shapley Sets. This may influence the reliability of AD due to a varying number of features being removed from an instance. We therefore also assess the Average Sensitivity (AS) of the attribution technique, Equation 4.35, which calculates the difference between the sum of all the attributions given by the attribution technique and the total change in prediction between the sample and reference value. Ideal attributions have low AS.

$$AS = \frac{1}{k} \sum_{j=1}^{k} v(\mathbf{x}, N) - \sum_{i=1}^{n} m(X_{ij}) \tag{4.35}$$

Tables 4.3 and 4.4 show how Shapley Sets has lower (better) AD than TS and KS across all three datasets. However, KS has the lowest AS score on the Diabetes dataset, we note that for this dataset, there is high variance of the sensitivity score for both Shapley Sets attributions. This can be largely explained by the sensitivity of Shapley Sets to the setting of ϵ, i.e. the amount of

	Shapley Sets Marg	Shapley Sets Cond	KS	TS
B	0.020 ± 0.022	$\mathbf{0.007 \pm 0.006}$	0.046 ± 0.047	0.047 ± 0.048
D	0.081 ± 0.075	$\mathbf{0.050 \pm 0.039}$	0.103 ± 0.085	0.010 ± 0.082
C	$\mathbf{0.005} \pm 0.007$	0.033 ± 0.029	0.075 ± 0.057	0.072 ± 0.055

Table 4.3: Table shows AD $\pm$ std for the attributions generated by Shapley Sets under v_{marg} and v_{cond}, KS and TS for the Boston (B), Diabates (D) and Correlation (C) datasets. Shapley Sets attributions have lowest deletion score across all datasets.

	Shapley Sets Marg	Shapley Sets Cond	KS	TS
B	0.015 ± 0.049	$\mathbf{0.006 \pm 0.031}$	0.029 ± 0.000	0.030 ± 0.000
D	0.021 ± 0.099	0.017 ± 0.067	$\mathbf{0.004 \pm 0.000}$	0.076 ± 0.000
C	$\mathbf{0.000 \pm 0.010}$	0.008 ± 0.020	0.001 ± 0.000	0.035 ± 0.000

Table 4.4: Table shows AS $\pm$ std for Shapley Sets under v_{marg} and v_{cond}, KS and TS for the Boston (B), Diabates (D) and Correlation (C) datasets. Shapley Sets results in the lowest sensitivity for B and C yet KS achieves lowest sensitivity for D.

statistical interaction permitted between two variables before they are considered to be dependent. Figure 4.2 shows the advantage of sets rather than individual attributions. The red and green curves (KS and Shapley Sets respectively) show the change in prediction as each feature in the sorted attributions is masked consecutively from the input. The blue and black horizontal bars show the original and target prediction respectively. An optimum feature attribution technique would result in a curve which sharply approaches the target. By considering the effect of sets of interacting features rather than individual features we can see that Shapley Sets avoids the sub-optimal behaviour of KS which arises due to the interaction effects between features in the model masking each other's importance. Figure 4.2 also validates the use of the AD to compare individual and set attributions as it is clear that masking more features does not guarantee a lower AD score.

In this section wedemonstrated on a variety of synthetic and real-world datasets, how in the presence of interacting features be that in the m odel or in the data, Shapley Sets generates more robust and meaningful explanations than that of Shapley value based alternatives.

4.8 Shapley Sets: Concluding Remarks

This chapter introduced the Shapley value within its original coalitional game theory context, comparing its perspective on fair division of value with alternative solution concepts. The purpose of this discussion was to demonstrate the ambiguity surrounding the concept of fairness and how this translates into division of value. We provided an account of how the Shapley value

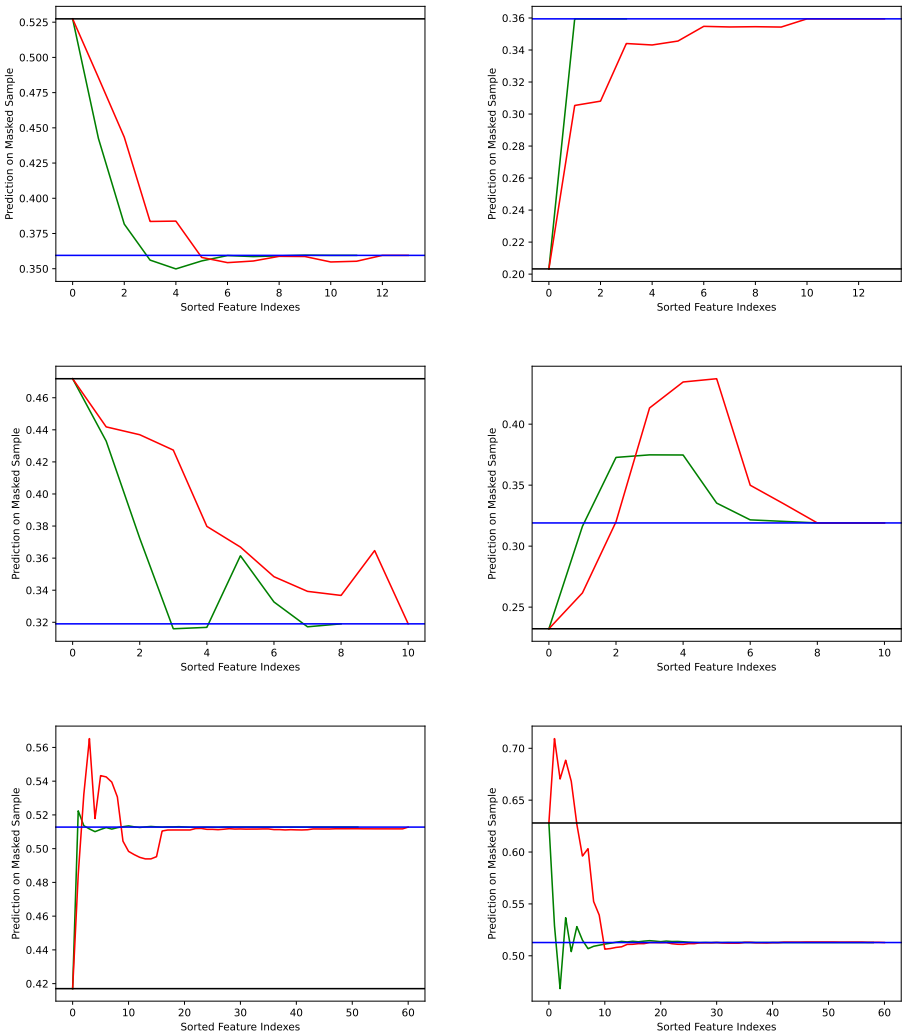


Figure 4.2: Figure shows the change in prediction of two individual samples from the Boston dataset (top row), Diabetes dataset (middle row) and Correlation Dataset (bottom row) as increasing features, as sorted in order of importance by the attributions returned by Shapley Sets Marginal (green) and KS (red), are perturbed from the instance. Original and target predictions are shown by the black and blue horizontal line. An ideal attribution would result in a sharp increase or decrease towards the target. In both samples, Shapley Sets results in a quicker and smoother transition from original to target prediction across all datasets and example samples.

has been applied to feature attribution and highlighted some of the most common assumptions required to map the game-theoretic value into a feature attribution tool. We showed how these assumptions can result in violations of the original Shapley fairness axioms which connect the approximated Shapley value for feature attribution to games under partition structure such that the interaction structure invoked on features by the data distribution or by the prediction model invoke a partition structure on the variable set. Shapley Sets operates under the perspective

that it is unfair to consider interacting features independently and thus we offer an attribution method which decomposes the variable set into its constituent NSVGs under a specified value function. We demonstrated empirically the advantages of Shapley Sets on tabular data compared to state-of-the-art Shapley value-based alternatives.

In Chapter 6 we illustrate how the difficulties in attributing time series data first introduced in Chapter 3 extend to the application of the Shapley value and demonstrate the advantages of Shapley Sets within this context. We believe that Shapley Sets is a stepping stone connecting feature attribution with the concept of function separability, a research discipline with rich mathematical background and recent algorithmic development. We note however, the vulnerability of Shapley Sets to weaknesses of the RDG algorithm which include a sensitivity to correlation in the data as well as an assumption that the predictive function is at least partially additively separable. In Chapter 7 we elaborate on the weaknesses of Shapley Sets and outline directions for future work. The introduction of the Shapley value alongside alternative solution concepts we have provided in this chapter also grounds the argument presented in the following chapter where we revisit the Shapley value from a causal perspective, using the mathematical framework underpinning contrastive questions to connect the Shapley value and the Gately value to differing investigative goals within the XAI landscape.

CHAPTER 5

POST-HOC LOCAL EXPLANATIONS AS CONTRASTIVE QUESTIONS: FROM THE SHAPLEY VALUE TO THE GATELY VALUE

5.1 Explanations As Investigative Goals

In this chapter, we continue our exploration of the Shapley value but from a causal perspective, unifying feature attribution with counterfactuals, two methods from the post-hoc local explanation literature which are often considered as distinct. The overarching investigative goal of Explainable AI is widely accepted as the endeavour to help humans understand the decisions made by an AI system [122]. Despite this, Explainable AI captures an ever-expanding set of use-cases, methods and ideas which are themselves characterised by a wide variety of investigative goals. Each time we delve into a particular branch of the Explainable AI landscape we become increasingly precise in the associated investigative goal.

The feature attribution methods discussed in the previous two chapters, LIME and Shapley value-based attribution are examples of post-hoc local explanations where the investigative goal is to understand the behaviour of an *existing* AI system by extracting relationships between the *feature values* of a given instance and its *prediction*. These types of explanation have captured the attention of the machine learning community and remain as the focus of this chapter. However, even after specifying the investigative goal of post-hoc local explanations, we uncover a diverse set of investigative questions which are characterised by the semantic interpretation to the questions, "what sort of relationship do we wish to understand between features and prediction?" and "what constitutes an understanding?"

When specifying an investigative goal for post-hoc local explanations we must consider the kind of relationship between input and outputs we want the explanation to capture. The types of relationship that exist between system components underpins the discipline of the philosophy of

explanation. Accounts of explanation both philosophical and psychological stress the importance of causality — that is, an explanation refers to causes [41].

If we view a post-hoc explanation as an identification of the causal relationship between the feature values and the output where the feature values of the input instance "caused" a certain prediction, then we can align the disciplines of causal inference and XAI. However, the extent to which various post-hoc explanations can be considered causal is an intensely disputed topic within the machine learning community [87, 95]. In the following section we introduce Pearl's causal hierarchy [164] which illustrates the central arguments underpinning causal explanations, and unify the feature attribution methodologies of the previous chapters with causal questions using the language of counterfactuals. Furthermore, we recall our earlier discussion of the role feature attribution plays in answering investigative questions (Section 4.2), motivating the importance of selecting a feature attribution method which makes most sense given a specified goal.

In this section we revisit this discussion and compare the different types of investigative questions answered by post-hoc local explanations. We discuss the investigative goals which are best answered by different levels of the causal hierarchy. Pearl's causal hierarchy details the different levels of causal reasoning which, as we descend further down the "ladder of causation", reveal more about the underlying causal system [164]. The hierarchy consists of three layers encoding different concepts: the associational, the interventional, and the counterfactual [19], corresponding roughly to the ordinary human activities of seeing, doing, and imagining, respectively. We discuss below how the hierarchy corresponds to varying levels of understanding in terms of causes and outcomes.

This chapter focuses on counterfactual questions and causes and how feature attribution methods relate to causal attribution. Even if we were to perfectly attribute a model output to its causes, whether this constitutes an explanation is tied to the definition of "understanding". Understanding is an innately human concept: how best to encode statistical information about the causal relationship between the model, feature values and individual output which is optimum for the explainee. An important concept is the relationship between causal attribution and explanation. Extracting a causal chain and displaying it to a person is causal attribution, not (necessarily) an explanation. While a person could use such a causal chain to obtain their own explanation there are many conflicting arguments in the literature which argue for the selection of different subsets of the causal chain to be presented as an optimal explanation. In this chapter we will explore the different types of counterfactual question which each encode different perspectives on causation and, in turn, extract different elements of the causal chain which can be used to explain AI systems.

In this section, we have argued that XAI encompasses a diverse set of investigative goals. We have identified the interdependence between the sciences of explanation and causality and outlined the objective of this chapter as the unification of feature attribution methods and the causal definition of counterfactuals.

5.2 The Causal Hierarchy

The causal mechanisms underpinning a system of interest are normally assumed to be unobservable, however they do produce observable data [19]. In this way, reality and the data generated by it are distinct. Causal inference, as a research discipline, attempts to identify the true causal components of a system from the non-perfect data it generates [19]. Causal methods often rely on a Structural Causal Model (Definition 5.1) as the mathematical object which formalises causal knowledge about a system. Section 5.2 formalises a Structural Causal Model as introduced by Pearl [19], it can be therefore skipped if the reader is familiar with this concept. Section 5.3 introduces the notion of a "general perspective of causality".

Definition 5.1 (Structural Causal Model [19])**.** A Structural Causal Model (SCM) M is a tuple $(\mathbf{U},\mathbf{V},F)$, where

- $\mathbf{U}$ is a set of exogeneous variables
- $\mathbf{V}$ is a set of endogeneous variables $\{V_1,...,V_n\}$
- F is a set of functions $\{f_1,...,f_n\}$ which map the variables in $\mathbf{U}$ to those in $\mathbf{V}$ such that $v_i \rightarrow f_i(\mathbf{Pa}_i = \mathbf{pa}_i, \mathbf{U}_i = \mathbf{u}_i)$ where $\mathbf{Pa}_i \subseteq \mathbf{V}$ and $\mathbf{U}_i \subseteq \mathbf{U}$

A probabilistic Structural Causal Model is defined as above but with the addition of a probability distribution $P(\mathbf{U})$ over the set of exogeneous variables.

5.2.1 Exogeneous And Endogeneous Variables

The definition of an SCM above relies on the partitioning of the variables into endoegenous and exogeneous sets. A variable is said to be endogeneous within the causal model if its value is determined by one or more of the other variables included in the model [82]. A variable is said to be exogeneous if it is independent of all other variables included in the model but may be dependent on variables not included as part of the model [82].

The distinction between exogeneous and endogeneous variables when defining a SCM relies on assumptions made about the system of interest which in reality, is likely to be far more complex than we are capable of modelling. For example, how do we know for sure which variables truly influence the probability of death from sepsis? The perfect specification of a causal model

which entirely represents reality is practically impossible so certain assumptions must be made which are specified by the model. $P(\mathbf{U})$ represents the probability distribution over the state of the world. In most circumstances, these exogenous variables correspond to the units, or elements of the population under investigation i.e. individual patients. A context, $\mathbf{u}$ is a vector that gives a unique value to each exogenous variable $U \in \mathbf{U}$. A model/context pair $(M, \mathbf{u})$ is called a situation [148].

5.2.2 An SCM Example

We now introduce a simple SCM (Example 5.1) which we will use to compare the different types of causal relationships as specified by Pearl's causal hierarchy.

Example 5.1. *Consider the (simplified) causal system which underlies whether an individual will die or survive sepsis which is modelled by the Output variable* $O = 1$ *for survival and* $O = 0$ *for death. The model includes two binary endogeneous variables: diabetes and lung cancer where* $V_1 = 1$ *indicates the presence of diabetes and* $V_2 = 1$ *indicates the presence of lung cancer. There are also two exogeneous variables* U_1 *and* U_2*, which represents a source of variation outside the model affecting* V_1, V_2*. Table 5.1 outlines the distribution* $P(\mathbf{U})$ *for each possible combination of endogeneous variables. In this model, the exogeneous variables* U_1 *and* U_2 *determine the values of* V_1 *and* V_2 *respectively and the Outcome, whether a patient survives sepsis is determined only by the value of* V_2 *as follows:*

$$F = \begin{cases} V_1 \leftarrow U_1, \\ V_2 \leftarrow U_2, \\ O \leftarrow 1 - V_2. \end{cases} \tag{5.1}$$

The definition of a probabilistic SCM induces a mapping between $P(\mathbf{U})$ and $P(\mathbf{V})$. For Example 5.1, each entry of Table 5.1 thus corresponds to an individual in the space of $\mathbf{U}$ and the corresponding realization of $\mathbf{V}$ according to the functions in F.

U_1	U_2	V_1	V_2	O	$P(\mathbf{u})$
1	1	1	1	0	0.25
0	1	0	1	0	0.25
1	0	1	0	1	0.25
0	0	0	0	1	0.25

Table 5.1: Table showing the probability distribution $P(\mathbf{U})$ for Example 1 as the mapping of events in the space of $\mathbf{U}$ to $\mathbf{V}$ in the context of Example 1

It is important to note that the simplicity of Example 5.1 encapsulates the challenges of interpreting correlative machine learning algorithms in a causal manner. Of course, it is not

only the presence of diabetes or lung cancer which determines whether a patient will survive sepsis, this is in reality determined by a large number of observable and non-observable factors. However, here we model that a correlation is found between the variables lung cancer, diabetes and the outcome which could be used as features within a predictive machine learning model. These variables cause the outcome even though they may represent a simplified version of reality.

In many practical settings, it may be impossible to specify the exact form of the underlying causal relationships. Nevertheless, we often assume that they exist [164]. It was shown by Pearl [164] that each SCM induces a causal hierarchy (or "ladder of causation"), which highlights qualitatively different aspects of the underlying reality and thus equates to causal questions encapsulating varying investigative goals regarding the system under investigation.

5.2.3 Level 1: Seeing

The first level of Pearl's hierarchy characterises the notion of "seeing" a particular relationship unfold and can be connected to the investigative goal of observation, e.g., "If I see that a particular patient has lung cancer what will that tell me about their likelihood of surviving sepsis"?

Definition 5.2 (Seeing [19])**.** The probabilistic SCM $M =< \mathbf{U}, \mathbf{V}, F, P(\mathbf{U}) >$ from Definition 5.1, defines a joint probability distribution $P^M(\mathbf{V})$ such that for each $\mathbf{W} \subseteq \mathbf{V}$

$$P^M(\mathbf{w}) = \sum_{\mathbf{u}|\mathbf{W}(\mathbf{u})=\mathbf{w}} P(\mathbf{u}) \tag{5.2}$$

Here, $\mathbf{W}(\mathbf{u})$ is the solution for $\mathbf{W}$ after evaluating F with $\mathbf{U} = \mathbf{u}$

The above definition characterises the probability mass $P(\mathbf{U} = \mathbf{u})$ obtained by collecting all units $\mathbf{U} = \mathbf{u}$ for which $\mathbf{W} = \mathbf{w}$ after their endogeneous variables $\mathbf{V}$ have been determined by the specification of the model M. As noted by Bareinboim et al. [19], most machine learning models fall into the seeing category of the causal ladder, for instance expressions such as $P(\mathbf{Y}|\mathbf{X})$ with $\mathbf{X}, \mathbf{Y} \subseteq \mathbf{V}$ are all "seeing" questions in that they passively observe the system of interest. For the sepsis SCM from Example 5.1, a "seeing" question may be *What is the probability of death given lung cancer?* which would be formalised as the following:

$$P(O = 0|V_2 = 1) = \frac{\sum_{\mathbf{u}|O(\mathbf{u})=0, V_2(\mathbf{u})=1} P(\mathbf{u})}{\sum_{\mathbf{u}|V_2(\mathbf{u})=1} P(\mathbf{u})} = \frac{0.5}{0.5} \tag{5.3}$$

5.2.4 Level 2: Doing

The next level of the ladder allows the representation of the notion of "doing" or intervening in the world to bring about some action. For example, if a patient didn't have lung cancer would they still die from sepsis? Performing an external intervention or action is modelled through the replacement of the natural mechanisms associated with the variables in $\mathbf{X} \subseteq \mathbf{V}$ with a constant

$\mathbf{x}$ [19]. This intervention is characterised by the causal sub-model $M_{\mathbf{X}} = (\mathbf{U}, \mathbf{V}, F_{\mathbf{X}}, P(\mathbf{U}))$ such that the original natural function F as specified by the SCM M is replaced by the function $F_{\mathbf{X}} = \{f_i : V_i \notin \mathbf{X}\} \cup \{\mathbf{X} \leftarrow \mathbf{x}\}$ [162]. This act of intervention is represented by the "do-operator" $do(\mathbf{X} = \mathbf{x})$ [164]. The impact of the intervention on an outcome variable $\mathbf{W} \subset \mathbf{V}$ is called the Potential Response (Definition 5.3). An SCM permits the valuation for interventional quantities by collecting the units for which the Potential Response for certain outcome variable $\mathbf{W} \subset \mathbf{V}$ matches certain condition $\mathbf{w}$ after a particular intervention. A further measurable quantity, the Causal Effect of a variable (Definition 5.4) determines the difference in Potential Response of an outcome variable relative to two settings of the variable and is the later focus of this chapter.

Definition 5.3 (Potential Response [162])**.** Let $\mathbf{X}, \mathbf{W}$ be two sets of variables in $\mathbf{V}$ and let $\mathbf{u}$ be a unit. The Potential response $\mathbf{W}_{\mathbf{X}=\mathbf{x}}(\mathbf{u})$ is the solution to $\mathbf{W}$ of the set of equations $F_{\mathbf{X}}$ with respect to the model M [163].

Definition 5.4 (Causal Effect [162])**.** The difference between two different interventions of an endogeneous variable, $\mathbf{X} = \mathbf{x}$ and $\mathbf{X} = \mathbf{x}'$, in unit $\mathbf{u}$,

$$\mathbf{W}_{\mathbf{X}=\mathbf{x}}(\mathbf{u}) - \mathbf{W}_{\mathbf{X}=\mathbf{x}'}(\mathbf{u}), \tag{5.4}$$

is known as the Causal Effect of the variable setting $\mathbf{X} = \mathbf{x}$ relative to $\mathbf{X} = \mathbf{x}'$ on the outcome variable $\mathbf{W}$ [163].

Definition 5.5 (Doing [19])**.** The SCM $M =< \mathbf{U}, \mathbf{V}, F, P(\mathbf{U}) >$ from Definition 5.1, defines a joint probability distribution over the endogenous variables $\mathbf{V}$, one for each intervention $\mathbf{x}$, such that for each $\mathbf{W} \subseteq \mathbf{V}$:

$$P^M(\mathbf{w}_{\mathbf{X}=\mathbf{x}}) = \sum_{\mathbf{u}|\mathbf{W}_{\mathbf{X}=\mathbf{x}}(\mathbf{u})=\mathbf{w}} P(\mathbf{u}) \tag{5.5}$$

Here, $\mathbf{W}_{\mathbf{X}=\mathbf{x}}(\mathbf{u})$ is the solution for $\mathbf{W}$ after evaluating $F_{\mathbf{x}}$ with $\mathbf{U} = \mathbf{u}$

The above definition characterises the probability mass $P(\mathbf{U})$ obtained by collecting all the units $\mathbf{U} = \mathbf{u}$ for which $\mathbf{W}_{\mathbf{X}=\mathbf{x}} = \mathbf{w}$ after the functions F have been replaced by the functions $F_{\mathbf{X}}$. For Example 5.1, a possible "doing" question would be, *What is the probability that if a patient didn't have lung cancer would they survive sepsis?* This question is equivalent to the following:

$$P(O = 1|do(V_2 = 0)) = \sum_{\mathbf{u}|O_{V_2=0}(\mathbf{u})=1} P(\mathbf{u}) = 1 \tag{5.6}$$

To obtain the above probability, refer to the probability distribution as shown in 5.1. If we set all individual values of $V_2 = 0$ and evaluate F_O with these settings then for each individual $O_{V_2=0} = 1$ as F_0 only depends on the value of V_2.

5.2.5 Do Calculus: The Difference Between Seeing And Doing

The main difference between seeing and doing within the causal hierarchy is the notion of intervening on a variable, setting it to a particular value rather than observing a variable with that value. Intervening requires replacing that variable's function within the SCM with the given value, breaking its dependence on its causal ancestors and the downstream variables dependence on that variable's ancestors.

Observing a variable at a particular value does not replace its function and therefore may potentially influence (the belief in) every variable in the network. As such, "seeing" the Causal Effect of a particular variable at a certain value may be affected by its causal ancestors, intervening on the variable instead "guarantees" that the resulting effect will be causal [19]. Interventional causal quantities, unlike observational statistical quantities, are defined relative to a causal model M and not only relative to the distribution $P^M(\mathbf{V})$. Similarly, observed data provides information about $P^M(\mathbf{V})$ alone, and since several SCMs can generate the same distribution, the danger exists that the Causal Effect of $do(\mathbf{X} = \mathbf{x})$ will not be discernible unambiguously from the data, even when we have enough of it [164].

Identifiability ensures the assumptions necessary for the condition that the causal model M will supply the missing information without having to explicate M in full. The Causal Effect of $\mathbf{X}$ on $\mathbf{Y}$ is identifiable from a graph G if the quantity $P(\mathbf{W}_{\mathbf{X}=\mathbf{x}})$ can be computed uniquely from any positive probability of the observed variables [164].

The identifiability of $P(\mathbf{W}_{\mathbf{X}=\mathbf{x}})$ ensures that we can infer the interventional effect $do(\mathbf{X} = \mathbf{x})$ on $\mathbf{W}$ from the following sources: 1) 'seen" observations, from the probability function $P^M(\mathbf{V})$ 2) the causal graph G, which specifies which inter-dependencies between variables. There are many different methods which estimate Causal Effects from experimental data using various identification strategies including front and back-door criteria [164] which are out of the scope of this thesis. We explore, in Section 5.5.7, the extent to which value functions, employed by various feature attribution methods, calculate the true Causal Effects of feature values on a prediction.

5.2.6 Layer 3: Imagination

The final layer of Pearl's causal hierarchy allows the mathematical formalisation of the notion of "imagining" alternative worlds. This would include questions like *"If the patient didn't have diabetes and they had survived would they still have survived had they had diabetes?*. In reality, the individual in question had diabetes *and* they did not survive so this kind of imaginative question requires a mechanism for conceiving and grounding all the potential possible worlds to evaluate an answer.

Definition 5.6 (Counterfactual)**.** The SCM M from Definition 5.1, induces a family of joint

distributions over counterfactual events $\mathbf{W_X}$,...,$\mathbf{Z_Y}$ for any $\mathbf{Y},\mathbf{X},\mathbf{W},\mathbf{Z} \subseteq \mathbf{V}$:

$$P^M(\mathbf{w_x}, \mathbf{z_y}) = \sum_{\mathbf{u}|W_\mathbf{X}(\mathbf{u})=\mathbf{w},...,\mathbf{Z_Y}(\mathbf{u})=\mathbf{z}} P(\mathbf{u}) \tag{5.7}$$

Where $\mathbf{W_X}(\mathbf{u})$ is the solution for $\mathbf{W}$ after evaluating $F_\mathbf{X}$ with $\mathbf{U} = \mathbf{u}$ and $\mathbf{Z_Y}(\mathbf{u})$ is the solution for $\mathbf{Z}$ after evaluating $F_\mathbf{Y}$ with $\mathbf{U} = \mathbf{u}$

Definition 5.6 contains the intervention of variables under varying settings of other variables which characterise different counterfactual "worlds". The probability mass $P(\mathbf{U})$ is therefore the collection of each unit $\mathbf{U} = \mathbf{u}$ which is consistent with the events over the counterfactual variables $\mathbf{W_X},\mathbf{Z_Y}$ after each of the relevant natural mechanisms $F_\mathbf{X},...,F_\mathbf{Y}$ have been replaced with the appropriate constants.

Continuing with Example 5.1 a possible counterfactual question would be *"What is the probability that given a patient had lung cancer and died, would they have survived had they not had lung cancer?* Given that the patients we consider in this counterfactual did actually have lung cancer and died from sepsis this question is non-trivial. This question is equivalent to the following equation

$$P(O_{V_2=0} = 1|O = 0, V_2 = 1) = \frac{\sum_{\mathbf{u}|O_{V_2=0}(\mathbf{u})=1,V_2(\mathbf{u})=1,O(\mathbf{u})=0} P(\mathbf{u})}{\sum_{\mathbf{u}|V_2(\mathbf{u})=1,O(\mathbf{u})=0} P(\mathbf{u})} = \frac{0.5}{0.5} = 1 \tag{5.8}$$

The difference between doing and imagining within the hierarchy is captured by the numerator of the above equation which involves the simultaneous evaluation of two possible worlds, the one in which patients had lung cancer and died and also the world in which patients did not have lung cancer and survived. "Doing" does not allow for these hypothetical worlds (expressions with more than one subscript) and thus the third layer is required to answer these counterfactual questions. Within the landscape of counterfactual reasoning, formalised by the imagining layer of the hierarchy, there are many possible counterfactual questions which can be extracted from the structural causal model. The following section introduces some example counterfactual questions with different semantic interpretation.

In this section, we have introduced Pearl's Causal Hierarchy as a way of formalising the kinds of causal questions we can ask about a system of interest.

5.3 Different Types Of Counterfactual Questions

When talking about counterfactual attribution, it is common to differentiate between necessary and sufficient causes [163]. A necessary condition is a condition that must be present for an event

to occur. A sufficient condition is a condition that will produce the event. A necessary condition must be there, but it alone does not provide sufficient cause for the occurrence of the event.

In the following we use the notation $P(y)$ to denote the probability that an event y occurred and the notation $P(y')$ to denote the probability that the event y did not occur. The Probability of Necessity (PN) $P(y'_{x'}|x,y)$ encodes how the outcome $Y = y$ is attributable to a particular exposure, the singleton, $X = x$, interpreted counterfactually as the probability that an event would have not occurred in the absence of a particular factor given that the event and factor did, in reality, occur [163]: *What is the probability of survival had those who had lung cancer and died not had lung cancer?* The Probability of Sufficiency (PS) $P(y_x|x',y')$ instead captures the probability that the event $Y = y$ would have counterfactually occurred under the counterfactual treatment $X = x$ given that in reality, the event and factor did not occur [163]: *What is the probability of death if those who did not have lung cancer and survived had had lung cancer?*

While both the Probability of Necessity and Sufficiency ascertain what causes death from sepsis, the nuances in the framing of the counterfactual question directly impact the set of hypothetical situations considered. Under necessity, the reasoning is tailored to a specific event under consideration – dying from sepsis and the counterfactual estimates the effect of lung cancer in the individuals who did actually die from sepsis. The Probability of Necessity therefore assesses the presence of an active causal process capable of producing the effect [163]. In contrast, under sufficiency, we are interested in studying a general tendency of a given effect whereby we consider the impact of lung cancer across all individuals who did not die.

Choosing between necessary and sufficient causes when designing a counterfactual question has particular significance when explaining systems. If we focus an explanation on general tendencies of a causal factor, i.e., determining sufficient causes for an event then that explanation may misrepresent the actual scenario at hand. For example, if there was a group of healthy individuals who, if they had had diabetes would have been taken to a different hospital at which they would have died from sepsis, then diabetes in this situation is sufficient for them to die from sepsis. However, it does not capture the fact that diabetes itself does not cause death from sepsis. In contrast, if we focus only on a single situation, i.e., identify the necessary causes of an event, then irrelevant background factors may constitute part of the explanation simply because if they happened to not have occurred then the event would not have happened.

As argued by Pearl [163], some balance must be struck between the necessary and the sufficient components of causal explanation. The Probability of Necessity and Sufficiency (PNS) $P(y_x, y'_{x'})$, stands for the probability that Y would respond to $X = x$ both ways, and therefore measures both the sufficiency and necessity of $X = x$ to produce $Y = y$

We have discussed the difficulty in estimating Causal Effects from observational data in Section 5.2.5 and have shown how the estimation of Causal Effects relies on an identifiability assumption. Since both the probability of necessity and sufficiency require conditioning on $Y = y$ and since $Y = y$ is presumed affected by $X = x$, the antecedent of the counterfactual $Y_{X=x}$, we

know that none of these quantities is identifiable from knowledge of the structure of the model and the data $P(\mathbf{U})$ alone, even under condition of no confounding. Pearl [163] showed how to identify both probabilities of necessity and sufficiency we must be able to identify the Causal Effect of $X = x$ on outcome $Y = y$. To do this, we must assume Exogeneity" (Definition 5.7) which specifies that the way in which Y would respond to intervening on $X = x$ or $X = x'$ is independent of the actual value of X.

Definition 5.7 (Exogeneity Assumption [163])**.** A variable X is said to be exogeneous relative to Y in model M iff

$$P(y_x, y_{x'}|x) = P(y_x, y_{x'}) \tag{5.9}$$

Alongside the assumption of Exogeneity, Pearl outlines a further requirement for the identification of the Probabilities of Necessity and Sufficiency, Monotonicity (Definition 5.8). Monotonicity expresses the assumption that a change from $X = 0$ to $X = 1$ cannot, under any circumstance make Y change from 1 to 0.

Definition 5.8 (Montonicity [163])**.** A variable Y is said to be monotonic relative to variable X in a causal model M iff the junction $Y_{X=x}(u)$ is monotonic in x for all u. Equivalently, Y is monotonic relative to X iff

$$y'_x \cap y_{x'} = false \tag{5.10}$$

If Y is monotonic relative to X, then the Probabilities of Necessity and Sufficency are all identifiable whenever the Causal Effect $Y_{X=x} - Y_{X=x'}$ is identifiable and are given by

$$PNS = P(y_x) - P(y_{x'}) \tag{5.11}$$

$$PN = P(y'_{x'}|x, y) \tag{5.12}$$

$$PS = P(y_x|x', y') \tag{5.13}$$

When determined over a population, probabilistic and necessary and sufficient causes measure the global relationships $Y_{X=x}$ and $Y_{X=x'}$ which is "too crude" [164] to fully characterize the "many nuances" of causation. We often need more detail of the causal mechanisms connecting X to Y to explicate more refined notions, such as an "Actual Cause" which we explore further below.

5.3.1 Singular And General Causes

Causal relationships can be categorised into those which are singular and general [90]. For example, we can argue that generally, lung cancer causes death from sepsis and regarding a single case, Amy who died from sepsis yesterday, it was having lung cancer which caused her to die. It has been argued that general causal relations encapsulate causal relations in singular cases, i.e lung cancer causes death from sepsis only because in general, patients like Amy who have lung cancer usually die from sepsis [90]. However, it has been argued that singular causal relations can exist even if they are not instances of the general causal laws [90]. A singular notion of causation which seeks to understand why a specific event occurred involve an "amiable jumble" of regularities, counterfactual dependence, statistical correlation and so on [90]. In this way, we can see why probabilistic accounts of general causation are considered "too crude" to account for all the multitude of factors for a given event; they are but one (yet important) part of the puzzle.

The extent to which probabilistic causation can be used as a generalisation of singular causes, or that singular causes can be extended to define general causal relations is widely debated in the literature [90] encompassing many different perspectives. In this chapter, we question the type of causal relation which are identified by feature attribution methods. We first however outline the ways in which singular causation differs from general causation.

So far, we have explicated the necessary and sufficient conceptions of causation in terms of their probabilities and thus as the general causal relationship i.e. the Probabilistic Necessity of lung cancer in causing death from sepsis, but not as properties of a singular scenario, dictated by a specific state of $\mathbf{u}$. singular causes are formalised, like probabilistic causes, with a SCM yet now we do not require a probability distribution over individual units $P(\mathbf{U})$ as we consider individual units $\mathbf{u}$ in isolation. In this context, rather than the probabilistic interpretation of necessity and sufficiency whereby the necessity of an effect is measured over a population. We talk about whether a particular cause was necessary or sufficient for an individual event in a given situation $\mathbf{u}$. In this context, the difference between necessity and sufficiency becomes even starker as, given a particular situation $\mathbf{u}$, we are given a particular setting of variables and outcome which actually happened. This has direct implications on the definitions of singular Necessity (Definition 5.9) and Sufficiency (Definition 5.10).

Definition 5.9 (Singular Necessity [163])**.** Event $X = x$ is said to be a necessary cause of event $Y = y$ in a world $\mathbf{u}$ just in case the following hold

- $Y(\mathbf{u}) = y$
- $X(\mathbf{u}) = x$
- $Y_{X=x'}(\mathbf{u}) \neq y$

Definition 5.10 (Singular Sufficiency [163])**.** Event $X = x$ is said to be a sufficient cause of event $Y = y$ in a world $\mathbf{u}$ just in case the following hold

- $Y(\mathbf{u}) \neq y$
- $X(\mathbf{u}) \neq x$
- $Y_{X=x}(\mathbf{u}) = y$

Example 5.2 adapts Example 5.1 from a general to singular perspective of causation. The following details how we may identify the causes for the singular situation: Amy dying from sepsis.

Example 5.2. *Adapting our SCM from Example 5.1, we now focus on a specific individual, Amy, who has both diabetes and lung cancer and who died from sepsis as shown in Table 5.3.1.*

u	V_1	V_2	O
Amy	1	1	0

If we are to find the necessary causes for Amy's outcome. We not that this is solely V_2 as under the intervention $O_{V_2=0} = 1$, Amy would have survived. To find the sufficient causes of Amy's death from sepsis, we imagine that Amy had not died from sepsis and ascertain the impact of lung cancer and diabetes on these hypothetical Amys. We find that lung cancer is sufficient for her death. The definition of singular Sufficiency requires both the factor $X = x$ and the event $Y = y$ to be false in *any* possible world u - this implies an imaginative action where we must step outside of reality for a moment and imagine a hypothetical world in which $X = x$ and $Y = y$ are absent, apply $X = x$ and see if $Y = y$ happens.

In this section, we have distinguished between singular and general causality and have introduced sufficient and necessary causes from both perspectives.

5.3.2 Which Counterfactual Worlds Should We Imagine?

We have seen that the definition of singular Necessity requires the hypothesization of the singular counterfactual world in which $X = x$ is not true and $Y = y$ is not true. In contrast, the definition of singular Sufficiency requires the hypothesization of all possible counterfactual worlds in which the the event (which we are looking to explain) and the factor (which we may believe caused the event) are both false. singular Sufficiency and Necessity therefore invoke different imaginary scenarios in which we compare what actually happened with what could have happened in order to attribute a singular event to a cause. Sections 5.3.2, 5.3.3 and 5.3.4 formalise the notions of an Actual and Contrastive Cause as per the definitions of Halpern [82] and Miller [148]. They exemplify a singular perspective of causality which underpins our proposed feature attribution method in this chapter. This section therefore may be skipped if the reader is interested in our novel "singular" feature attribution method which is motivated and introduced in Section 5.5.4.

An alternative perspective of a singular Cause is that of an Actual Cause (Definition 5.11). The notion of an Actual Cause was first formalised using structural equations by Halpern and Pearl [82], this definition was later modified [81] and then modified again in [80] with each definition reflecting a different perspective on singular Causality. To this day there is no "accepted" definition of an Actual Cause reflecting the ambiguity surrounding causality we have alluded to throughout this chapter. Below however, we give the most recent of the definitions, from Halpern [80]. Fundamentally, an Actual Cause attributes an event y to a set of causes. Unlike singular Sufficiency and Necessity which consider a singular conjunct $X = x$, an Actual Cause finds the set of conjuncts $\mathbf{X} = \mathbf{x}$ which together caused an outcome y.

Definition 5.11 (Actual Cause)**.** $\mathbf{X} = \mathbf{x}$ is an Actual Cause of y if the following conditions hold:

- Property 1: $Y(\mathbf{u}) = y, \mathbf{X}(\mathbf{u}) = \mathbf{x}$
- Property 2: There is a set $\mathbf{W} \subseteq \mathbf{V}$ and a setting $\mathbf{x}'$ of variables $\mathbf{X}$ such that if $W(\mathbf{u}) = \mathbf{w}$ then $Y_{\mathbf{X}=\mathbf{x}',\mathbf{W}=\mathbf{w}} = y'$ If $\mathbf{X}$ did not have the value $\mathbf{x}$ and all variables in $\mathbf{W}$ are fixed at $\mathbf{w}$ then event y would not have occurred
- Property 3: The set $\mathbf{X} = \mathbf{x}$ is minimal there are no unnecessary primitive events in the conjunction $\mathbf{X} = \mathbf{x}$

The first condition of Definition 5.11 requires that the conjunct $\mathbf{X} = \mathbf{x}$ and the event y must both be true in situation $\mathbf{u}$. The second condition stipulates that for the hypothetical setting in which $Y_{X=x,W=w}(\mathbf{u}) = y$, if fixing $\mathbf{X} = \mathbf{x}'$ while fixing $\mathbf{W} = \mathbf{w}$ causes the output to change from y to y', then $\mathbf{X} = \mathbf{x}$ is an Actual Cause of y. The third property of Definition 5.11 constrains the conjunct $\mathbf{X} = \mathbf{x}$ to be minimal.

A Partial Cause is a subset of an Actual Cause [80]. Actual Causes can be considered as answers to "Why" questions, *Why did Amy die from sepsis?*. To answer the above we would find that Amy having lung cancer was an Actual Cause of her dying from sepsis, as given the setting of diabetes at its actual value $V_2 = 1$, had Amy not had lung cancer, she would have survived sepsis. Amy having diabetes was not an Actual Cause of her dying from sepsis as, given that the presence of lung cancer was fixed at its actual value $V_1 = 1$ had she not had diabetes, she still would have died from sepsis.

5.3.3 Contrastive Questions

According to Miller [148], most consider "Why" questions to be inherently contrastive: *Why P rather than Q?*, *Why did Amy die from sepsis rather than survive?* In this format, the event that occurred, the patient dying from sepsis, is the fact while the alternative, the patient surviving is the contrast case or foil. To explain, or provide an answer to a contrastive question, Lipton [132] argues that a contrastive explanation should consist of the Difference Condition:

Definition 5.12 (Difference Condition [132])**.** To explain *Why* $Y = y$ *rather than* $Y = y'$, we must cite a causal difference between y and y', consisting of a cause of y and the absence of a corresponding event in the history of not y'.

Lipton [132] argues that the explainer does not need to reason about or even know about all causes of the fact, only those relative to the contrast case. The explainer does not need to know about all the reasons a particular patient died from sepsis, only those that separated them from the alternative outcome of surviving. There are two types of contrast which can occur in a contrastive explanation: Compatible and Incompatible Contrasts. Below we distinguish between these two types of contrast and how they relate to two different types of contrastive question, a Counterfatual and a Bifactual.

Incompatible Contrasts occur when the occurrence of the fact precludes the occurrence of the foil: *Why did this patient die from sepsis rather than survive?* [148] The death of the patient precludes their survival, and this factor is something which the Difference Condition cited in the explanation must take into account. The explanatory Difference Condition must reveal the causal factor that led to the patient's death and prevented its survival.

Compatible Contrasts occur when the occurrence of the fact does not preclude the occurrence of the foil. *Why did patient A die from sepsis but patient B survived?* [148] Clearly, patient A dying from sepsis does not affect patient B dying and certainly does not preclude it. When people ask questions involving compatible contrasts, it is usually because they expect both the fact-event and the foil-event to occur. In this case, the explanation must cite a difference in the causal histories of the events which made the difference between the fact occurring rather than both the fact and foil occurring.

A Counterfactual Question (Definition 5.13) is a realisation of the contrastive question, *Why* $Y = y$ *rather than* $Y = y'$ under the Incompatible Contrast, y'. For a counterfactual question, this means that, in some situation, the fact y occurred and the explainee is asking why the foil y' did not occur in that situation instead.

Definition 5.13 (Counterfactual Question [148])**.** Why $Y(\mathbf{u}) = y \wedge Y(\mathbf{u}) \neq y'$?

A Bifactual Question (Definition 5.14) is a realisation of a contrastive question under a compatible contrast where both the $\mathbf{u}$ and $\mathbf{u}'$ are two different situations including two potentially different models, y is the fact, and y' is the surrogate. Linguistically, the change of *rather than* to *but* can be interpreted as the transition of the counterfactual question to an environment where both the fact and the foil are true, there is no hypothetical outcome.

Definition 5.14 (Bifactual Question [148])**.** Why $Y(\mathbf{u}) = y \wedge Y(\mathbf{u}) \neq y'$ but $Y(\mathbf{u}') = y' \wedge Y(\mathbf{u}') \neq y$

5.3.4 Contrastive Causes

To answer contrastive questions, we need contrastive causes [148]. A contrastive cause between y and y' is a pair, in which the first element is Partial Cause of y and the second element

is a Partial Cause of y'. Miller [148] distinguishes between two different types of contrastive causes, Counterfactual and Bifactual, which can be used to answer Counterfactual and Bifactual questions accordingly.

Definition 5.15 (Counterfactual Cause)**.** The pair of events $\mathbf{X} = \mathbf{x}, \mathbf{X} = \mathbf{x}'$ is a counterfactual Cause of event y in a situation $(M, \mathbf{u})$ if the following properties hold:

- Property 1: $\mathbf{X} = \mathbf{x}$ is a Partial Cause of event y in situation $\mathbf{u}$
- Property 2: $Y(\mathbf{u}) = \neg y'$ the foil is not true in the situation $\mathbf{u}$
- Property 3: There is a non-empty set $\mathbf{W} \subseteq \mathbf{V}$ and a setting $\mathbf{w}$ of variables in $\mathbf{W}$ such that $\mathbf{X} = \mathbf{x}'$ is a Partial Cause of y' under this hypothetical situation where $Y_{\mathbf{W}=\mathbf{w}, \mathbf{X}=\mathbf{x}'}(\mathbf{u}) = y'$.
- Property 4: $[\mathbf{X} = \mathbf{x}] \cap [\mathbf{X} = \mathbf{x}'] = \varnothing$
- Property 5: The set $\mathbf{X} = \mathbf{x}$ is maximal

A Counterfactual Cause identifies the factors necessary to change y to y' in situation $\mathbf{u}$. We note here how a Counterfactual Cause is constructed by the maximal intersection of those causes which were in part necessary for the event y as well as those which were in part necessary for the counterfactual event y'. Therefore if $\mathbf{X} = \mathbf{x}$ is a Partial Cause for y but $\mathbf{X} = \mathbf{x}'$ is not a Partial Cause of event y' then $\mathbf{X} = \mathbf{x}$ will not be considered as a Counterfactual Cause of the event y with regards to y' despite it being an Actual Cause of y. Returning to Example 5.2, to answer the Counterfactual Question *Why did Amy die from sepsis rather than survive*, we know that that having lung cancer was an Actual Cause of Amy dying. We can now find the Counterfactual Cause by identifying hypothetical counterfactual situations for Amy, where, under some adjustment of her variables she would have survived sepsis. We identify $O_{V_2=0}(Amy) = y'$ as the counterfactual situation where if Amy had not had lung cancer then she would have survived sepsis. We are then able to return the pair $(V_2 = 1, V_2 = 0)$ as the Counterfactual Cause of Amy dying from sepsis as there exists a hypothetical situation in which Amy not having lung cancer was a Partial Cause of her surviving sepsis.

Definition 5.16 (Bifactual Cause)**.** The pair of conjuncts $\mathbf{X} = \mathbf{x}, \mathbf{X} = \mathbf{x}'$ is a Bifactual Cause of event y in a situation $\mathbf{u}$ given situation $\mathbf{u}'$ with outcome y' if the following properties hold:

- $\mathbf{X} = \mathbf{x}$ is a Partial Cause of y in the situation $\mathbf{u}$
- $\mathbf{X} = \mathbf{x}'$ is a Partial Cause of y' in the situation $\mathbf{u}'$
- $[\mathbf{X} = \mathbf{x}] \wedge [\mathbf{X} = \mathbf{x}'] = \varnothing$
- The set $\mathbf{X} = \mathbf{x}$ is maximal

From the above definition we can see that a Bifactual Cause pair consists of the causes where $\mathbf{X} = \mathbf{x}$ is a Partial Cause for event y and $\mathbf{X} = \mathbf{x}'$ is a Partial Cause of y'. The difference therefore between the notion of a Counterfactual and Bifactual Cause is in the determination of the counterfactual Causes of y'. For a Bifactual Cause we consider only the natural settings under $\mathbf{u}'$ for which $\mathbf{X} = \mathbf{x}'$ is a Partial Cause for y'.

Now consider the Bifactual Question, *Why did Amy die from sepsis but John survived?* Where John is known to have diabetes but not lung cancer. To find a Bifactual Cause of this question satisfying definition 5.16 we again note that Amy having lung cancer was an Actual Cause of her dying from sepsis and John not having lung cancer was an Actual Cause of him surviving sepsis. Under the Bifactual setting, as both John and Amy have diabetes, this would never be included in the Bifactual Cause due to the Difference Condition, (Property 3 of Definition 5.16). The Bifactual Cause of Amy dying from sepsis but John surviving is the pair $(V_2 = 1, V_2 = 0)$.

In this section we have outlined some of the different approaches for identifying the causes of a singular event y in a particular situation $\mathbf{u}$. We have shown how the probabilities of sufficiency and necessity have been extended to singular situations, discussed the ambiguity in defining an Actual Cause and have introduced Miller's distinction between Counterfactual and Bifactual Causes. In the following section we start to question what sort of causal relationships between input and output are most suited for post-hoc local explanations and explore how existing counterfactual explanations adopted by the Explainable AI community fit into the philosophical discussion.

In this section, we have introduced the notions of an Actual, Counterfactual and Bifactual Cause which encapsulate a singular perspective on causality.

5.4 Contrastive Questions In AI Systems

The definitions of Counterfactual and Bifactual Causes as specified by Miller exemplify the way in which Lipton's Difference Condition can be used to generate explanations which are relevant to a given situation. In providing a foil, we are encoding prior knowledge into our investigative question, or, as Miller puts it, being specific about our "window of uncertainty" [147].

Consider the inclusion of a further endogeneous variable V_3 to Example 5.2, whereby $V_3 = 1, V_3 = 0$ indicates the presence or absence of a cold respectively. In this scenario, the presence of a cold is an Actual Cause of death from sepsis, however the absence of a cold is *not* an Actual Cause of survival. Asking why Amy died from sepsis could include a number of Actual Causes which each would offer a valid attribution, because she had lung cancer, or because she had a cold. However, let us consider a situation where our investigative goal is a direct comparison of the truth with the counterfactual outcome of survival, asking *why did Amy die rather than*

survive? For this investigative question, the attribution that refers to the presence of a cough makes a poor explanation, as even if Amy hadn't had a cough, she is not guaranteed to survive. Instead, a good explanation would point to what is different between the two outcomes, identifying the Counterfactual Cause of lung cancer. Importantly, the explanation fits directly within the questioner's "window" of uncertainty, and is smaller and simpler [148].

AI models are typically high-dimensional, complex and structured. Section 5.4 introduces recent work within the AI community which attempts to derive counterfactual explanations for AI outcomes, it may be skipped if the reader is interested in the novel contribution of this chapter, the unification of the Shapley value with a general perspective of causality, which is detailed in Section 5.5.3.

Causal attributions for a particular outcome which are non-contrastive thus run the risk of extracting irrelevant or superfluous detail which is not specific to a given scenario. How best to communicate a causal attribution has received much attention in the philosophical literature [148].

In particular, the linguist Grice [75] laid out the principles which intuitively guide humans in making effective conversation and can be applied to how best to communicate a contrastive explanation [148]. Grice's conversational maxims [75] lay out the desiderata for effective communication as: make your contribution only as informative as is required; and only provide information that is related to the conversation. Following our discussion regarding how the Difference Condition in contrastive causes reduces the set of applicable causes we consider, it is clear that counterfactual questions and associated causes have potential to provide meaningful explanations which adhere to Grice's maxims for AI systems.

Counterfactual explanations have received much attention in the literature for post-hoc local explanations [43, 216] whereby the overarching objective is, given an instance to be explained $\mathbf{x}$ and an associated outcome y, identify the reasons as to why $f(\mathbf{x}) = y$ and not y' where the function f signifies the trained AI system. The way in which existing methods approach counterfactual generation varies, they are however, unified by the following desiderata [152], which we term Counterfactual Criteria.

- 1. Minimise the number of changes to the instance $\mathbf{x}$ needed to transform the prediction y to that of y'
- 2. The generated counterfactual $\mathbf{x}'$ corresponds to a valid data point

Given a machine learning model f, an instance to be explained $\mathbf{x}$ and a counterfactual outcome y, Wachter et al. [216] propose generating counterfactual instances as the set of solutions to the following optimisation problem.

$$\arg\min_{\mathbf{x}'} \max_{\lambda} \lambda(f(\mathbf{x}') - y)^2 - d(\mathbf{x}', \mathbf{x}) \tag{5.14}$$

Here, $d(.,.)$ is a distance function, recommended to be the sparsity inducing L1 norm by Wachter et al. [216], that measures how far the counterfactual $\mathbf{x}'$ and the original data point $\mathbf{x}$ are from one another. In practice, maximisation over λ is done by iteratively solving for $\mathbf{x}'$ and increasing λ until a sufficiently close solution is found. This solution fulfills counterfactual Criterium 1 in the sense that the minimal number of changes to $\mathbf{x}$ are fulfilled, which is enforced by the penalisation term minimising the distance, or maximising the similarity, between the counterfactual instance $\mathbf{x}'$ and original instance $\mathbf{x}$. Of particular importance therefore is the selection of the distance measure which implicitly encodes what we mean by "minimal amount of changes".

The method of Wachter et al. [216] however, does not satisfy Counterfactual Criterium 2 as it does not consider the "realism" of the resulting feature values recommended by $\mathbf{x}'$. A more recent approach for counterfactual explanations suggested by Dandl et al. [43] minimises the following objective

$$L(\mathbf{x},\mathbf{x}',y',X_{obs}) = (o_1(f(\mathbf{x}')-y'),o_2(\mathbf{x},\mathbf{x}'),o_3(\mathbf{x},\mathbf{x}'),o_4(\mathbf{x}',X_{obs})) \tag{5.15}$$

Equation 5.15 is composed of four minimisation objectives. o_1 minimises the Manhattan distance between the prediction of the counterfactual and the target counterfactual outcome, o_2 and o_3 minimise the (Gower) distance between $\mathbf{x}$ and $\mathbf{x}'$ and encourage sparsity respectively. The Gower distance is selected for its capability of handling both continuous and categorical distances. o_4 minimises the (Gower) distance between the generated counterfactual and nearest observation taken from a specified input dataset X_{inp}. In this way, the counterfactual explanations generated by the method of [43] fulfill both Counterfactual Criteria 1 and 2.

Both the methods above are well motivated for post-hoc explanations, particularly in the light of GDPR regulation [173] as discussed in Chapter 1. Furthermore, we can see that both counterfactual generation methods proposed by Wachter et al. [216] and Dandl et al. [43] align with Miller's definition of a counterfactual in the sense that they apply the Difference Condition to the fact, $f(\mathbf{x})$ and the set of hypothetical worlds $\mathbf{x}'$ which evaluate to the desired counterfactual outcome y'. However, we have seen how the determination of counterfactuality according to Miller relies on the identification of Actual Causes which, in turn, relies on the identifiability of Causal Effects required to perform interventions, opening up the question as to whether these methods can be considered as true *answers* to counterfactual questions. We discuss this further in Section 5.5.7.

First, however, we discuss the limitations of asking counterfactual questions to explain AI systems. Once central issue with counterfactuals used for post-hoc local explanations is the Rashomon effect [9] Rashomon is a Japanese film where the murder of a Samurai is recounted by different people. Each of the stories is a valid depiction of reality yet but the stories contradict each other. The same can happen with counterfactuals, since we generate counterfactuals by considering any possible hypothetical world which evaluates to the foil outcome, there are usually multiple different counterfactual explanations which, like Rashomon, may confuse or

contradict. This opens up the question, how do we select the optimum counterfactual? While counterfactual explanations have been developed within the XAI community, the distinction between Compatible and Incompatible Contrasts and the exploration of Bifactuals for use in post-hoc local explanations has received relatively little attention [147].

A Bifactual Cause, by Definition 5.16, satisfies both Counterfactual Criteria. Similarly to a Counterfactual Cause, the identification of a Bifactual Cause applies the Difference Condition to the fact $f(\mathbf{x})$ and the counterfactual outcome in situation $f(\mathbf{x}') = y'$. However, in contrast to a counterfactual, a bifactual explanation will always use a real instance in generating the cause. A bifactual, by requiring the specification of a desirable foil prior to explanation generation, is thus more aligned with the notion of an actionable explanation than that of a counterfactual. When investigating whether Amy died from sepsis, we can compare her to the actual counterfactual case John, who survived, or Adam, who also survived. We can therefore interpret the associated contrasting bifactual explanations, because Amy had a cold whereas John didn't, or because Amy had lung cancer whereas Adam didn't, meaningfully within the context we, ourselves, have selected.

Central to the above discussion is the identification of a relevant contrasting outcome for the explanation. We therefore motivate the importance of being explicit about our investigative goal. Do we want to know what the impact of lung cancer is on death from sepsis, or why Amy died, or why Amy died rather than survive, or why Amy died but why John survived? We have seen in Section 5.3 how each of the questions above, while appearing as superficially similar, are loaded with contrasting mathematical and philosophical meaning.

The Shapley value has been previously described as a counterfactual question of the form Why $f(\mathbf{x})$ rather than $\mathbb{E}[f(\mathbf{X})]$? or Why $f(\mathbf{x})$ rather than $f(\mathbf{x}')$? where $\mathbb{E}[f(\mathbf{X})]$ and $f(\mathbf{x}')$ are the the two kinds of contrast invoked by on and off-manifold value functions respectively [117]. Conflictingly, [112, 222] argue that feature attribution techniques such as LIME and SHAP are distinguished from counterfactual methods via the axis of necessity and sufficiency where counterfactual approaches extract the Necessary Causes of an output but attribution methods instead extract the Sufficient Causes of an outcome. In the following section we try and reason about these conflicting views and propose our own way of unifying feature attribution techniques with counterfactual explanations. First, we explore the kind of contrastive question imposed by solution concepts in coalitional games. Second, we explore where value functions are situated on the causal hierarchy and and argue as to whether these methods can be unified with counterfactuals.

In this section, we have related counterfactual post-hoc local explanations to the philosophical definitions of a counterfactual we introduced in Section 5.3. We have motivated a bifactual explanation as one which satisfies both Counterfactual Criteria.

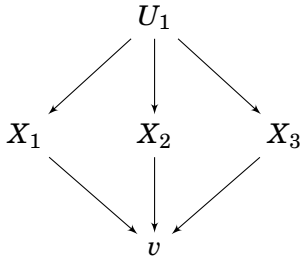
5.5 Contrastive Questions In Coalitional Games

In this section we unify the ideas of singular and general causality with value attribution in coalitional games. Sections 5.5.1 and 5.5.2 translate solution concepts from game theory into the language of counterfactuals, they can be skipped if the reader is interested in one of the novel contributions of this thesis, the unification of the Shapley value with a general perspective of causality, the PNS, which can be found in Section 5.5.3. We first introduce Example 5.3.

Example 5.3. *We consider the following coalitional game with three players: Player 1, Player 2 and Player 3, represented as the set* $N = \{X_1, X_2, X_3)$ *and value function* $v : \mathbb{R}^3 \to \mathbb{R}$*. The coalition* $S = (X_i)$ *represents the setting where Player* i *is present such that variable* $X_i = 1$ *and variables* $X_j = 0 \forall X_j \neq X_i \in N$*. The following set function corresponds to a binary outcome game played between three players where each setting of players can either win or lose.*

	$v(X_2) = 0$	$v(X_1, X_3) = 1$
$v(\varnothing) = 0$	$v(X_3) = 0$	$v(X_2, X_3) = 1$
$v(X_1) = 0$	$v(X_1, X_2) = 0$	$v(X_1, X_2, X_3) = 1$

To be able to ask causal questions, we need to encode the coalitional game in Example 5.3 as a Structural Causal Model. We fix the set of endogenous variables $\mathbf{V}$ as the union of player set N with the value function of the game v. We assume for simplicity one single exogenous variable U_1 which represents the background information which causes a particular coalition $\mathbf{u}$ to form. We take the set of all possible coalitions over the player set $S \subseteq N$ as individual units $\mathbf{u}$ with an associated probability distribution as determined by the Shapley coefficient (Definition 4.4). The causal diagram below shows how the endogenous variables – the players – are independent and the set functions listed above show how the variable v depends on each of the three endogenous variables.



5.5.1 Singular Causal Players

We know from Section 4.1 that within coalitional games we assume that the grand coalition has formed and that the empty coalition has zero value. Value attribution considers the importance

of each player to the grand coalition and determines the payoff he or she can reasonably expect. From a causal perspective, we are therefore looking to understand the extent to which the inclusion of each player caused the outcome $v(N)$.

Continuing our argument from Section 5.3.1, the kind of causes we extract from the SCM depends on whether or not we consider the causal relationships between the players and the game to be a generalisation of the causal relationships between the players and the outcome in individual settings (or coalitions) of the game. As we have discussed in Section 5.3.1, this places certain assumptions about how the effects of players in singular settings of the game $\mathbf{u}$ and the associated outcome $v(\mathbf{u})$ relate to the general causal relationships between players and outcome.

We know from Chapter 4 that games in coalitional form permits interactive effects between players: for example, Player 1 from Example 5.3 may have a greater impact on the game when working with Player 2 than Player 3. As such, how we extract causal relationships from the game will depend on the kind of settings we consider. We show below how focusing on singular and general causal questions changes the kind of causal explanation we extract.

We begin by showing how we may extract the Actual Causes of the grand coalition winning in Example 5.3. Under a singular causal perspective, if we were to ask, *Why did the grand coalition win?* then we would find the set of conjuncts which satisfy Definition 5.11 for the singular world $\mathbf{u} = N$ which represents the grand coalition where $X_1(\mathbf{u}) = 1$, $X_2(\mathbf{u}) = 1$, $X_3(\mathbf{u}) = 1$ and $v(\mathbf{u}) = 1$. For the game exemplified in Example 5.3, $X_3 = 1$ is an Actual Cause of $v(\mathbf{u}) = 1$ as $v_{X_3=0}(\mathbf{u}) = 0$. Player 1 or Player 2 are not individual Actual Causes of $v(\mathbf{u}) = 1$ as $v_{X_1=0}(\mathbf{u}) = 1$ and $v_{X_2=0}(\mathbf{u}) = 1$. However, the conjunct $(X_1 = 1, X_2 = 1)$ is an Actual Cause of $v(\mathbf{u}) = 1$ as $v_{X_1=0,X_2=0}(\mathbf{u}) = 0$.

Now, if we want to start asking contrastive questions, following Miller, we could ask the Counterfactual Question *Why did the grand coalition win rather than lose?* To generate the Counterfactual Cause pair (Definition 5.15) we would now find the Actual Causes of the counterfactual outcome of losing $v(\mathbf{u}') = 0$. For Example 5.3, these are: $X_3 = 0$, $X_2 = 0$, $X_1 = 0$ as there exist possible settings of the game where the inclusion of Player 3, Player 2 or Player 1 individually would have resulted in a win. Returning the maximum intersection of these partial causes we obtain the Counterfactual Cause: $(X_1 = 1, X_2 = 1, X_3 = 1)$ and $(X_1 = 0, X_2 = 0, X_3 = 0)$.

We note that for this example, the Counterfactual Cause determines that the absence of all three players would cause the counterfactual outcome of losing the game. This is representative of the situation where Player 3 on its own does not cause a win, $v(\{X_3\}) = 0$. If, under a different game formulation $v(\{X_3\}) = 1$, then the Counterfactual Cause would be restricted to the following: *Because Player 3 is present, had Player 3 been absent the game would have been lost*.

If we instead ask the Bifactual Question why $v(\mathbf{u}) = 1$ but $v(\mathbf{u}') = 0$ where $\mathbf{u}' = (X_3)$ describes the coalition containing only Player 3. We would find the Actual Causes of $v(\mathbf{u}) = 1$ which are the same as above. Now however, when we find the Actual Causes of the outcome $v(\mathbf{u}') = 0$, we get the Actual Cause $X_1 = 0, X_2 = 0$ as this is the minimal conjunct of Actual Cause for the loss in setting $\mathbf{u}'$. Returning the maximum intersection of these Partial Causes we obtain the counterfactual

Cause of $(X_1 = 1, X_2 = 1)$, $(X_1 = 0, X_2 = 0)$. The game was won (and not lost) in situation u because Player 1 and Player 2 were present. The game was lost (and not won) in situation $\mathbf{u}'$ because Player 1 and Player 2 were absent.

Counterfactual and Bifactual Causes capture the minimum set of causes necessary for the outcome y and the outcome y'. Under a binary outcome and given binary variables, the negation of the Actual Causes of outcome y can be viewed as sufficient to cause y' under that hypothetical setting. For Example 5.3, Player 1, Player 2 and Player 3 are all individually sufficient to cause y given the set of all possible hypothetical worlds which evaluate to y'.

Restricting the set of counterfactual worlds in which we determine the Actual Causes of the event y', which is what we do when determining a Bifactual Cause, therefore enforces a more singular notion of causality on the counterfactual outcome. Moving from a Counterfactual to a Bifactual can thus be seen as becoming increasingly more specific in our window of uncertainty.

In this section, we have shown how we can extract the Actual, Bifactual and Counterfactual Causes for a coalitional game given the singular event $\mathbf{u}$ corresponding to the outcome of the grand coalition $v(N)$.

5.5.2 Are Solution Concepts Causal?

Solution concepts are not only designed to understand which groups of players caused a certain outcome but also the **extent** to which an **individual** player caused a certain outcome. In this sense, we can view solution concepts as mechanisms which identify individual Causal Effects of each player on the game. We must therefore extend our language concerning contrastive causes to those which consider the Causal Effects of a player on an outcome. Solution concepts are built upon the notion of marginal contributions - they identify the effects of singletons on the game. From our definition of a Causal Effect (Definition 5.4), if we set $\mathbf{u} = S$ such that $\mathbf{u}$ corresponds to a particular setting of the game where the variables $X_i = 1 \forall i \in S$ and $X_i = 0 \forall i \notin S$, the marginal contribution (Definition 4.2) of Player i to coalition S, $v(S \cup i) - v(S)$ can be treated as its Causal Effect on the outcome of the game v in world $\mathbf{u}$, $v_{X_i=1}(u) - v_{X_i=0}(u)$.

Before we consider unifying solution concepts with causal and contrastive questions we must first ask whether the Causal Effects of players on the game are identifiable. From Definition 5.7, we know that for the Causal Effects to be identifiable, each player must be exogeneous with regards to the outcome v. From our causal diagram we know that there is an absence of a common ancestor of each player and the outcome. Thus, exogeneity holds for the SCM of a coalitional game. While monotonicity (Definition 5.8) of the co-operative game is not guaranteed it was shown by Tijs [212] that this assumption holds for a large subclass of coalitional games and the corresponding set function. Furthermore, Pearl [163] showed how the Causal Effects of variables could be identified in the case where monotonicity does not hold. However, for simplicity in the

following section we assume montonicity. This means that the Causal Effects $v_{X_i=1}(\mathbf{u}) - v_{X_i=0}(\mathbf{u})$ of each player are identifiable.

In this section, we have shown how the marginal contribution of player i in coalitional game v and coalition S can be treated as the Causal Effect of that player in context $\mathbf{u}$ such that $\mathbf{u} = S$ assuming Exogeneity and Montonicity.

5.5.3 The Shapley Value: Probability of Necessity And Sufficiency

Whether solution concepts consider the singular or general Causal Effects of players on the game to attribute the value of the grand coalition is dependent on the particular concept and associated approach to fairness as discussed in Chapter 4. For some solution concepts, such as Equal Division (Definition 4.4), for example, the Causal Effects of each individual player are not taken into account by the solution concept when attributing value, where the total Causal Effect of all players on the game is divided equally.

Kommiya et al. [112] claim that, under a binary outcome, the Shapley value calculates the Probability of Sufficiency (PS) of each player generating the outcome 1. For Example 5.3, the Shapley value attribution vector is $(\frac{1}{6}, \frac{1}{6}, \frac{4}{6})$. Below we show that the Shapley value under a binary outcome does not correspond to the PS. Intuitively we can understand why these two measures are not equivalent as under PS, Equation 5.13, we condition the probability on situations where $X_i = 0$ **and** $v(\mathbf{u}) = 0$ both hold. This is not the probability calculated by the Shapley value which does not condition on the outcome. To see this, for Example 5.3, the probability of each player being sufficient for the outcome of winning is given as the following

$$\frac{P(v=1|X_1=1) - P(v=1|X_1=0)}{1 - [P(v=1|X_1=0)}] = \frac{\frac{1}{6}}{\frac{4}{6}} = \frac{1}{4} \tag{5.16}$$

$$\frac{P(v=1|X_2=1) - P(v=1|X_2=0)}{1 - [P(v=1|X_2=0)}] = \frac{\frac{1}{6}}{\frac{4}{6}} = \frac{1}{4} \tag{5.17}$$

$$\frac{P(v=1|X_3=1) - P(v=1|X_3=0)}{1 - [P(v=1|X_3=0)} = \frac{\frac{4}{6}}{1} = \frac{4}{6} \tag{5.18}$$

Similarly, the Probability of Necessity is conditioned on instances where both the player i is present and the outcome of winning is true.

$$\frac{P(v=1|X_1=1) - P(v=1|X_1=0)}{P(v=1|X_1=1)} = \frac{\frac{1}{6}}{\frac{3}{6}} = \frac{1}{3} \tag{5.19}$$

$$\frac{P(v=1|X_2=1)-P(v=1|X_2=0)}{P(v=1|X_2=1)}=\frac{\frac{1}{6}}{\frac{3}{6}}=\frac{1}{3} \tag{5.20}$$

$$\frac{P(v=1|X_3=1)-P(v=1|X_3=0)}{P(v=1|X_3=1)}=\frac{\frac{4}{6}}{\frac{4}{6}}=1 \tag{5.21}$$

The equations above demonstrate how the Shapley value is not equivalent to either the PS or the PN. Below we show that actually, the Shapley value determines the Probability of Necessity *and* Sufficiency (PNS) of each player in generating the outcome 1.

If we use Definition 4.5 of the Shapley value, then we can see that the value is determined by the average effects of players on the game: the value is generated as the sum of the marginal contribution of each individual player to all possible permutations $\pi \in \Pi$ divided by the total number of possible permutations $n!$ given n players. From here, we note that for each player, there are $n!$ possible orderings $\pi \in \Pi$, or permutations of the game $H_\pi(i)$ from which, the value of the coalition containing the players preceding player i are determined both with player i $v(H_\pi(i) \cup \{i\})$ and without player i, $v(H_\pi(i))$. Each permutation $H_\pi(i)$ therefore corresponds to a possible setting of the game $\mathbf{u}$. From our mapping of the coalitional game to the causal formalisation, we know that the marginal contribution of player i, $v(H_\pi(i) \cup \{i\}) - v(H_\pi(i))$, is equivalent to the Causal Effect of player i in setting $\mathbf{u} = H_\pi(i)$, $v_{X_i=1}(\mathbf{u}) - v_{X_i=0}(\mathbf{u})$. Taking each permutation $H_\pi(i)$ for all $\pi \in \Pi$ as a possible world $\mathbf{u}$, we present one of the main contributions of this chapter in Proposition 5.1.

Proposition 5.1. *Given the coalitional game (N, v) with binary outcome variable v, if we let each permutation of the game $H_\pi(i)$ represent an individual setting $\mathbf{u}$ then the Shapley value $\phi_i(v)$ (Definition 4.5) is equivalent to the PNS (Equation 5.11) of an individual player i causing a winning outcome in the game v under the assumption of Exogeneity (Definition 5.7) and Monontonicity (Definition 5.8).*

Proof. Given Definition 4.5, the Shapley value for each player i is calculated as the following

$$\phi_i(v) = \frac{1}{n!}\sum_{\pi\in\Pi} v(H_\pi(i)\cup\{i\}) - v(H_\pi(i)) = \mathbb{E}[v_{X_i=1}(\mathbf{u}) - v_{X_i=0}(\mathbf{u})] = \mathbb{E}[v_{X_i=1}] - \mathbb{E}[v_{X_i=0}]. \tag{5.22}$$

Under the binary outcome set function v,

$$\mathbb{E}[v_{X_i=1}] - \mathbb{E}[v_{X_i=0}] = P[v_{X_i=1}=1] - P[v_{X_i=0}=1] \tag{5.23}$$

Taking outcome $P(y_x) = P[v_{X_i=1}=1]$ as the probability that the presence of player i causing a winning outcome and event $P(y_{x'}) = P[v_{X_i=0}=1]$ as the probability that the absence of player i causes a winning outcome

$$P[v_{X_i=1} = 1] - P[v_{X_i=0} = 1] = P(y_x) - P(y'_x), \tag{5.24}$$

which matches the definition of PNS from Equation 5.11 and thus completes the proof.

■

From Proposition 5.1, we can see that as the expectation is taken over every possible setting (or permutation of players) of the game, the Shapley value can be seen less as a singular account of causation but general, where we consider the Causal Effect of a players in each possible setting of the game. The PNS of each player for the outcome of winning in Example 5.3 is given as the folllowing set of equations

$$P(v = 1|X_1 = 1) - P(v = 1|X_1 = 0) = \frac{3}{6} - \frac{2}{6} = \frac{1}{6}, \tag{5.25}$$

$$P(v = 1|X_2 = 1) - P(v = 1|X_2 = 0) = \frac{3}{6} - \frac{2}{6} = \frac{1}{6}, \tag{5.26}$$

$$P(v = 1|X_3 = 1) - P(v = 1|X_3 = 0) = \frac{4}{6} - 0 = \frac{4}{6}, \tag{5.27}$$

which matches the Shapley value of each player in Example 5.3. From what we know about the Shapley value, the fact that it determines the PNS of each player makes intuitive sense as the Shapley value considers all possible settings of the game and determines the marginal impact of each player in each of these settings regardless of whether those settings generated a particular outcome. In this sense, the Shapley value identifies the general (or probabilistic) causal relationships between individual players and the outcome v.

In this section, we have shown that the Shapley value determines the general Causal Effects of individual players on the game. Specifically we have shown how it determines the PNS of each player on the game under the assumption of exogenity and montonicity. In the following section we motivate the Gately value as an alternative to the Shapley value which identifies the Singular Causel Effects of individual players on the game.

5.5.4 The Gately Value: A Bifactual Solution Concept

We have discussed in Section 5.3.1 the diversity of opinion among philosophers of causation as to the extent to which singular Causal Effects can be unified probabilistically into general Causal

Effects. Furthermore, we have shown in Section 5.5.1 how for coalitional games, the notions of individual Necessity and Sufficiency, Counterfactual and Bifactual Causes all change the kind of causal explanation we can extract from the game.

Within coalitional games, we know that players may have a varying effect on the game depending on the presence or absence of other players. For instance, a player may exhibit a large Probability of Necessity yet not actually be individually Necessary for the value of the grand coalition. Averaging the Necessity and Sufficiency of each player therefore, as with the Shapley value, may lose important information about the outcome that actually happened. Following the argument we have established surrounding singular vs. general accounts of causation, we motivate the question: **When attributing the value of the grand coalition, do we care about the effect of each player in every hypothetical world or only in the world which actually happened?**

If we are to restrict the possible worlds over which we measure the PNS of a player, we could select the settings of the game we consider to be important and thus, we can become more singular in the type of causal relation we consider. In Section 5.4, we have discussed how a Bifactual aligns itself with the "Counterfactual Criteria" and Grice's maxims [75]. We know that under Miller's specification [148], a Bifactual is concerned only with the causes of outcome y in setting $\mathbf{u}$ and the causes of y' in setting $\mathbf{u}'$ under the Difference Condition.

In this way, if we are to restrict the calculation of the PNS of a player to certain worlds, we could restrict the Probability of Necessity to consider only the probability that each player was individually necessary in a specific world $\mathbf{u}$. Given that under a binary outcome, the negation of the necessary causes of outcome y' can be seen as the sufficient causes of outcome y, by restricting the Probability of Sufficiency to consider only a singular hypothetical world $\mathbf{u}'$ we obtain a PNS measure which is grounded in singular causality and closely aligned with that of a Bifactual. We term this restriction of the PNS of a player to the world $\mathbf{u}$ and $\mathbf{u}'$ as the Bifactual Effect of a player (Definition 5.17)

Definition 5.17 (Bifactual Effect)**.** The Bifactual Effect of $\mathbf{X} = \mathbf{x}$ relative to $\mathbf{X} = \mathbf{x}'$ in situation $\mathbf{u}$ on outcome Y, given situation $\mathbf{u}'$ is defined as

$$\frac{(Y_{\mathbf{X}=\mathbf{x}}(\mathbf{u}) - Y_{\mathbf{X}=\mathbf{x}'}(\mathbf{u})) + (Y_{\mathbf{X}=\mathbf{x}}(\mathbf{u}') - Y_{\mathbf{X}=\mathbf{x}'}(\mathbf{u}'))}{2} \tag{5.28}$$

Under a binary value function $Y = v$, we can see from our earlier definition of the PNS (Equation 5.11), the Bifactual Effect of an individual player in game v is equivalent to the PNS of that player taken as an expectation over only the two settings of the game $\mathbf{u}$ and $\mathbf{u}'$

In Example 5.3, if we consider only the two settings of the game such that $\mathbf{u}$ is the grand coalition and $\mathbf{u}'$ is the empty coalition then the Bifactual Effect of Player 1 and Player 2 given the specific worlds $\mathbf{u}$ and $\mathbf{u}'$ are both zero but the Bifactual Effect of Player 3 is $\frac{1}{2}$. By measuring

only the Bifactual Effect, the resulting PNS is more aligned with a notion of a Bifactual Cause as defined in Definition 5.16 such that if a player has no casual effect in the world $\mathbf{u}$ or $\mathbf{u}'$ then they are not considered a cause to either situation. We note, however, that the Bifactual Effect (Definition 5.17) is not equivalent to the notion of a Bifactual Cause (5.16) as it is restricted to the effect of singletons on the game. By restricting the set of possible worlds over which we calculate the PNS, we are no longer guaranteed the set of fairness axioms as guaranteed by the Shapley value. When determining the Bifactual Effect of a player in Example 5.3, it is clear that the PNS over the limited worlds $\mathbf{u}$ and $\mathbf{u}'$ no longer satisfy Efficiency (Definition 4.1.2). We have discussed throughout this thesis the importance and utility of axioms. The above discussion therefore motivates the following question,

Is there a solution concept which determines the Bifactual Effect of a player which is also guaranteed by a set of fairness axioms?

In this section, we motivate the use of the Gately value as a solution concept which addresses the above question and our proposed feature attribution method, Gately Feature Attribution (Definition 5.18), is one of the main contributions of this chapter.

The Gately value, introduced in Chapter 4, Definition 4.9, attributes to each player an amount bounded from above by their Utopia Value M_i (Definition 4.7), which corresponds to a player's Causal Effect in the world $\mathbf{u}$, and from below by their Minimum Rights vector v_i (Definition 4.6) which corresponds to a player's Causal Effect in world $\mathbf{u}'$. In this way, we can view the Gately value as a solution concept whose attributions are proportional to the Bifactual Effect of an individual player given the two settings of the game: the grand and empty coalition, in such a way that is axiomatically justified. For Example 5.1, the Minimum Rights vector is $(0,0,0)$ and the Utopia Vector is $(0,0,1)$, resulting in the following attribution under the Gately value: $gv_{X_1} = gv_{X_2} = 0$ and $gv_{X_3} = 1$ which is proportional to the Bifactual Effect of each player. Furthermore, we can see that the Gately value satisfies Efficiency (Definition 4.1.2), which is part of its unique axiomatisation as discussed in Chapter 4 Section 4.1.1.

The Gately value also satisfies v-compromise and restricted proportionality. V-compromise ensures that the attribution afforded by the Gately value to each player is proportional to its Minimum Rights vector and the Restricted Proportionality axiom ensures that the attribution afforded to each player is proportional to its utopia value. As we have seen above, it is the combination of these two axioms which make the Gately value an appropriate solution concept to generate bifactual explanations such that if a player has no marginal contribution in the world $\mathbf{u}$ or $\mathbf{u}'$, they are awarded zero attribution.

Thus far, we have connected the Shapley value and the Gately value to general and singular perspectives of causality. We have presented the Gately value as an attribution mechanism which is more applicable to generate bifactual explanations, using only the Bifactual Effect of players in two singular settings of game, the grand coalition and the empty coalition. The Gately value is axiomitised by efficiency, v-compromise and restricted proportionality making it a desirable

alternative to the Shapley value in value attribution. In the following section we motivate the Gately value for use in feature attribution. First, we extend our discussion to consider contexts where the outcome variable v is continuous which is the reality represented by both coalitional games and feature attribution tasks.

In this section, we have shown that as the Gately value considers only the Marginal Contribution of players in the grand and empty coalition, it can be viewed as a solution concept which determines an attribution which is proportional to the Bifactual Effect of a player in the two settings of the game: the grand and empty coalition. The Gately value thus operates under a more singular notion of causality than the Shapley value while satisfying a desirable set of fairness axioms.

5.5.5 From Binary To Continuous Outcomes

As discussed in Chapter 4, the Shapley value, applied to feature attribution, depends on the specification of a value function $v(\mathbf{x},\mathbf{X}_S)$ whereby $\mathbf{x}$ corresponds to the instance to be explained and $\mathbf{X}_S$ is the set of feature values to be "turned on" in the given instance. In this setting, each X_i is treated as a binary variable yet both coalitional games and feature attribution model continuous outcomes represented by the set function $v(\mathbf{x},\mathbf{X}_S)$. If we maintain our assumption of identifiability, under a continuous outcome, the Shapley value no longer determines the probability that a particular player generates a particular outcome (i.e. winning the game) as defined Equation 5.23 but instead measures the average Causal Effect of $X_i = 1$ relative to $X_i = 0$ as defined in Equation 5.22. For the Shapley value, this average is taken over every possible setting of the game $\mathbf{u} = H_\pi(i)$ for all $\pi \in \Pi$, but for the Gately value, the Causal Effect is determined only over the bifactual settings of the game.

Below we define our feature attribution method built on the Gately value. Given a feature set $X = \{X_1, X_2, ...X_n\}$, consisting of n features, a pre-trained function $f : \mathscr{X} \rightarrow \mathbb{R}$ and a corresponding set function $v \in \{v_{int}, v_{bs}, v_{cond}\}$. The feature attribution given by the Gately value for an individual instance to be explained $\mathbf{x}$ and reference value ϵ is given as the following equation

Definition 5.18 (Gately Feature Attribution)**.** Given the Minimum Rights vector, $\mathbf{v}$ such that $v_i = v(\mathbf{x},\mathbf{X}_i)$ for all $i \in N$ and the utopia vector $\mathbf{M}$ such that each $M_i = v(\mathbf{x},\mathbf{X}) - v(\mathbf{x},\mathbf{X}_{N\setminus\{i\}})$ The attribution vector afforded by the Gately value ρ for instance to be explained $\mathbf{x}$ is defined as

$$\rho_i = v_i + (f(\mathbf{x}) - \Sigma_{j=1}^{n} v_j)\frac{M_i - v_i}{\Sigma_{j=1}^{n} M_j - v_j} \forall i \in N \tag{5.29}$$

From Definition 5.18, given that the Minimum Rights vector represents the Causal Effect of $X_i = 1$ relative to $X_i = 0$ in the setting $\mathbf{u}'$ and that the utopia value measures the Causal Effect of

$X_i = 1$ relative to setting $X_i = 0$ in the setting $\mathbf{u}$, we can see that the attribution to each feature afforded by the Gately value is thus proportional to its Bifactual Effect in only two singular settings of the game invoked by the value function where the attribution is determined in such a way which guarantees its unique axiomatisation.

In this section, we have shown that under a continuous outcome v, the Shapley value determines the average Causal Effect of a player over all settings of the game $\mathbf{U}$. By restricting the number of worlds over which we determine the marginal contributions of players, we become increasingly singular in the Causal Effect we extract. The Gately value, applied to a continuous outcome game, considers the Bifactual Effect of each player on the game. We have defined our *Gately Feature Attribution method*.

5.5.6 Identifying Causal Effects With Value Functions

From the previous section, we know that the Shapley value determines the average Causal Effect of X_i on outcome v under setting $X_i = x_i$ relative to $X_i = x_i'$. The average is taken as an expected value over all possible settings of the game induced by the Shapley value.

Within coalitional game theory, the "game" induced by the Shapley value, i.e., the set of all possible permutations, entirely covers the decision space of the problem due to the binarity of players. When we are using the Shapley value for feature attribution, where features are not necessarily binary, we must therefore understand what "the game" is specifying. When used for local feature attribution, the Shapley value determines the average Causal Effect of a feature relative to it being "on", or set to the value in $\mathbf{x}$, and "off", where it is set to an approximated value (x') as determined by a value function.

For off-manifold value functions (v_{bs} in Equation 4.23 and v_{marg} in Equation 4.22), the Shapley value calculates the Causal Effect of features over all the possible settings where a feature can take only the corresponding value in $\mathbf{x}$ and the corresponding value in $\mathbf{x}'$. In this sense, the Shapley value breaks all the causal relationships between features that may exist in the real-world. Rather than modelling the true underlying system, the Shapley value under off-manifold value functions models feature values as system inputs. This way, it considers hypothetical settings of the game where the inputs to the model are changed rather than the the values of the true features. The causal diagram constructed by the assumptions of off-manifold value functions is exemplified in Figure 5.1.

For on-manifold value functions such as v_{cond} in Equation 4.24 and v_{int} in Equation 4.25, the Shapley value does not make the distinction between the features in the real world and the inputs of the prediction model. The "off" value of each feature corresponds to an expected value of each feature over the dataset which is allowed to change as a result of conditioning on

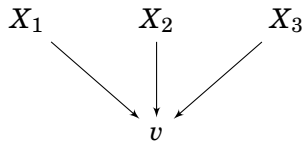


Figure 5.1: Figure shows an example causal structure enforced by off-manifold value functions which separate the true variables from the independent inputs which are fed into the value function as features.

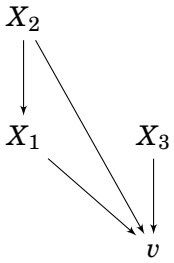


Figure 5.2: Figure shows an example of the causal structure enforced by on-manifold value functions which allow the dependencies between variables which exist in reality to influence the resulting value function.

other feature values. The causal diagram constructed by the assumptions of on-manifold value functions is exemplified in Figure 5.2.

We discussed in Section 5.5.2 how, when mapping coalitional games to counterfactual questions, we assume independence of players which allows us to assume exogeneity and thus, the Causal Effects of each feature are identifiable. For local feature attribution, given $f(\mathbf{x})$ we have no explicit knowledge of how the prediction of $f(\mathbf{x})$ would have changed has X_i not been set to x_i. We must therefore estimate these counterfactual situations from data which is done, for feature attribution, via a value function. The extent to which the Shapley value for feature attribution can be considered causal is thus dependent on the kind of value function we use and whether or not they allow for the identification of Causal Effects.

In this section, we have shown how different value functions, from the feature attribution literature, construct different "local explanation games" which determines the kind of feature effects captured by the Shapley value.

5.5.7 Value Functions And Causal Effects

We have discussed in Section 5.2.4 that identifiability of Causal Effects relies on the assumption of exogeneity. We have also seen how the SCM constructed to model a coalitional game meets this requirement as it is assumed that each player enters the game independently. When, however, we are using the Shapley value to determine Causal Effects, we know that features, in reality, are often inter-dependent and thus may share a common causal ancestor, or alternatively, the Causal Effect of each feature will be vulnerable to confounding.

If we consider the causal structure represented by Figure 5.2, in this situation, exogeneity cannot be assumed and as such, the Causal Effects each feature are not identifiable as $v_{x_i} \neq \mathbb{E}[v|X_i = x_i]$. If we instead consider the local explanation game induced by the Shapley value under off-manifold value functions as exemplified by Figure 5.1, as these value functions break all the statistical ties between variables when determining the Shapley value, similarly to the coalitional game setting, the Causal Effects of features when considered only as independent inputs to the model are identifiable and can be determined via Equation 5.22.

In contrast, on-manifold value functions take into account the dependencies between features when determining the Causal Effect of each feature. The conditional value function v_{cond} is computed via an observational expectation, which given the fact we cannot assume exogeneity, is not a true representation of the Causal Effect of each feature and is thus vulnerable to confounding effects. As an adaptation of the conditional value function, the interventional value function v_{int}, introduced by Heskes et al. [87] replaces the observational conditional value function with an interventional value function. This value function identifies the Causal Effects of each feature under the absence of exogeneity by first intervening on each feature and then observing its value, essentially transporting the conditional value function from the "seeing" rung of the causal hierarchy to that of "imagination".

We have seen above how the Shapley value under different value functions correspond to different assumptions about whether dependencies between features should be considered when determining the Causal Effect of features. However, moving beyond the discussion surrounding identifiability, it is important to remember that for feature attribution, we are usually explaining the output of a particular AI model. This is in stark contrast to Section 5.2 where we introduced SCMs as representing a real system. For feature attribution, when using the Shapley value within a causal framework we are using an SCM to represent a machine learning system which, as we have discussed in Section 5.2.3, usually only operates at the "seeing" level of the causal hierarchy. Therefore, while constructing counterfactual explanations for a machine learning model will highlight the correlations made by the underlying model, making correlations transparent does not make them necessarily causal, even if we appropriately identify the Causal Effects of features on the model output.

In this section, we have shown how the Shapley value, under different value functions, corresponds to different approximations of the Causal Effect of variables on a model outcome.

5.6 Motivating Gately Feature Attribution

Implicit in the use of the Shapley value for feature attribution is the comparison with the reference prediction $f(\mathbf{x}')$ with which the Shapley value determines the difference in Causal Effect of each feature between settings $X_i = x_i$ and $X_i = x_i'$. The Shapley value essentially asks the question, how does each $X_i = x_i$ move us towards the local prediction $f(\mathbf{x})$, and, how does each $X_i = x_i'$ move us further from the prediction $f(\mathbf{x})$? We have seen in Section 5.5 how the Shapley value is a general account of the causal relationship between features and output and as such cannot be interpreted as a counterfactual in the way formalised by Miller. Despite this, we can still view $f(\mathbf{x}')$ as a form of contrastive foil.

For on-manifold value functions, the foil is $\mathbb{E}[f(\mathbf{X})]$. Treating this expectation as our foil, we are asking the investigative question "Why $f(\mathbf{x})$ rather than $E[f(\mathbf{X})]$?" over which the Shapley value induces all the hypothetical settings of the game to determine Causal Effects of features. As the foil is an expectation determined over a given data distribution, $E[f(\mathbf{X})]$ can be assumed to preclude $f(\mathbf{x})$ [148], thus rendering it an incompatible contrast and more in line with a Counterfactual Question, *Why the fact rather than the expected?*

Selecting a foil allows us to focus the explanation on causes that are relevant to the question. Lipton argues that determined foils are more pragmatic than counterfactual questions, if a person provides a foil, they are implicitly pointing towards the part of the model they do not understand [132]. The selection of a desirable foil can be introduced via the use of the baseline value function. However, even if a desirable foil is specified, it has been noted by Kumar et al. [117] that the Shapley value produces undesirable results as it still induces $n!$ possible worlds, incorporating the effects of each feature on all the possible hypothetical worlds which lie between the fact and the foil. In this way, the effects of a feature may be calculated for conflicting hypothetical worlds [117].

The baseline value function specifies the fact as $f(\mathbf{x})$ and foil as $f(\mathbf{x}')$ which correspond to two real predictions. In this sense, the foil is the prediction of $f(\mathbf{x}')$, an incompatible contrast as its existence does not preclude $f(\mathbf{x})$. In contrast to the Counterfactual Question implied by the use of v_{cond}, under v_{bs} we are asking, *Why* $f(\mathbf{x})$ *but* $f(\mathbf{x}')$*?* Despite this implied bifactuality, the Shapley value will still induce the same number of "hypothetical worlds" between $f(x\mathbf{x})$ and $f(\mathbf{x}')$ in the way it does for on-manifold value functions.

Under the baseline value function, we therefore question whether the Shapley value is the right solution concept for attribution as the effect of a feature on all the hypothetical worlds will

be just as influential its effect in the grand coalition. Under v_{bs}, we argue that the contrastive question implied by the value function is that more closely aligned with singular causality and specifically, a Bifactual Question. In using Gately Feature Attribution (Definition 5.18) under v_{bs} would thus provide more of a singular account of feature attribution. In only considering the effects of features in restricted settings of the game, we are more likely to obtain a minimal attribution vector which were originally motivated in Section 4.1.5. To see this in practice we introduce Example 5.4

Example 5.4. *Let us consider the following function* $f(\mathbf{X}) = 2(X_1 - 1) + X_2X_3$. *If we set our example to be explained as* $f(1,1,1) = 1$ *and our reference foil as* $f(-1,1,3) = -1$*, the Shapley value under the baseline value function can be calculated using the following set of equations,*

$$\begin{array}{llll} & v(1) = 3 & v(3) = -3 & v(1,3) = 1 & v(1,2,3) = 1 \\ v(\varnothing) = -1 & v(2) = -1 & v(1,2) = 3 & v(2,3) = -3 & \end{array}$$

Determining the Shapley value for Example 5.4 returns the attribution vector $\phi = (3\frac{1}{3}, \frac{2}{3}, \frac{2}{3})$. The Gately attribution vector for Example 5.4 is $\rho = (4, 0, -2)$. We can see how the attribution awarded by the Shapley value incorporates the casual effect of features in hypothetical worlds which are different from $\mathbf{x}$ and $\mathbf{x}'$. As such, the attribution awarded to X_2 is non-zero despite this feature having the same value in both fact and foil. Applying the Difference Condition between fact and foil within the context of a Bifactual Cause as discussed in Section 5.3 gives an attribution which awards zero attribution to X_2, as does the Gately Feature Attribution value. In practice, an attribution technique which only awards value to features which satisfy the Difference Condition is useful, particularly for high-dimensional feature vectors as it results in a minimal attribution.

In this section, we have argued that when using the baseline value function v_{bs}, we are asking a Bifactual question and thus recommend Gately Feature Attribution in this case. We have shown how the Gately Feature Attribution can offer minimal attribution vectors compared to the Shapley value.

5.6.1 Robustness To Off-manifold Artifacts And Reduced Computation

We have discussed in Chapter 4 how off-manifold value functions, including v_{bs}, are sensitive to vulnerabilities incurred by the generation of out-of-distribution samples on which the function is evaluated. Further to the weaknesses of these out-of-distribution permutations as discussed in Section 4.3.2, it was first discussed by Slack et al. [194] how the explanations generated by

off-manifold value functions could be arbitrarily controlled by an adversary due to the ease of distinguishing between the hypothetical samples generated during the attribution process and the samples from the real data-distribution.

The potential for an "adversarial attack" [194] on explanations could potentially allow certain important features to be masked from an explanation. For example, Slack et al. [194] show how race can be completely hidden from the explanation after the attack, even though it was the sole important feature for the original classifier. This has serious ramifications when post-hoc local explanations such as SHAP are deployed in practice where the objective of the party delivering the explanation is in opposition to the party receiving it.

Slack et al. [194] motivate the development of adversarially robust explanations that can withstand such attacks. Gately Feature Attribution generates only two off-manifold samples per feature attribution. This is in contrast to the Shapley value which computes 2^n of these off-manifold samples. Furthermore, the hypothetical samples generated by the Gately value only differ from the original and foil real sample by one feature value. This limits the potential exploitation of the resulting explanations as it reduces the probability a classifier will be able to differentiate between generated and real data samples. We experimentally validate this claim in Section 5.7.2.

A further advantage of Gately Feature Attribution requiring only two off manifold samples per feature attribution is in the reduced computational complexity compared to the Shapley value. Gately Feature Attribution is linear in the number of features where the Shapley value is exponential. We have seen in Chapter 4 how the Shapley value is approximated by methods such as KS. However, KS is still computationally limited by the number of samples it must generate to adequately fulfill the Shapley axioms. In contrast, the Gately value has no need to be approximated and is thus guaranteed to fulfill its axiomatisation.

In this section, we have argued that as the Gately value only considers the effects of features in certain settings of the local explanation game, compared to the Shapley value, it is less vulnerable to adversarial attacks and it is easier to compute.

5.7 Experimental Motivation Of Gately Feature Attribution

This section experimentally motivates the use of Gately Feature Attribution over the Shapley value. We begin with two synthetic experiments where the first demonstrates the Gately value's ability to generate minimal attributions compared to the Shapley value under v_{bs}. The second demonstrates the robustness of the attributions generated by the Gately value compared to the Shapley value under v_{bs}. Finally, we compare the Gately Feature Attribution and state-of-the-art Shapley-based attributions on real world datasets.

5.7.1 Minimal Explanations Synthetic Experiment

In this section we compare the ability of the Gately value and the Shapley value to generate minimal attributions. We begin by constructing a function

$$f(\mathbf{X}) = sgn \sum_{X_i \in \mathbf{X}} X_i,$$

over the variable set $\mathbf{X} = \{X_1, ..., X_7\}$ where X_1 is sampled from the distribution $U(-\alpha, \alpha)$ and α is sampled from the distribution $U(-10, 10)$. All other variables, $X_i \in \mathbf{X} \backslash \{X_1\}$ are sampled from the distribution $U(-0.1, 0.1)$.

$f(\mathbf{X})$ returns -1 if the sum of features is negative and 1 if the sum of features is positive. We operate under the assumption that an optimum attribution returns the minimal set of features needed to change the output from $f(\mathbf{x})$ to $f(\mathbf{x}')$. We set our optimum attribution as the difference between the negative and positive outputs $f(\mathbf{x}) - f(\mathbf{x}')$ equally divided between the set of all features which, when individually replaced by their value in the reference sample $\mathbf{x}'$, change the sign of the summation and therefore change the output. Due to the non-linearity of sgn, these features are all equally necessary for the change in prediction. All features which do not cause an individual change in prediction would ideally be awarded zero attribution.

We generate 200 samples and associated references of the opposite output and compute Gately Feature Attributions as well as the Shapley values and the attributions obtained by KS with 1000 samples and a background set containing only the reference sample which replicates the behaviour of v_{bs} [205].

We determine the proportion of the 200 attributions which exactly match that of our optimum attributions for each attribution technique and report this result in Table 5.2. We repeat the experiment for varying settings of α which controls the number of features which are necessary for a prediction to change. A high value of α makes it more likely that only X_1 is singularly necessary whereas a low value of α increases the randomness of the necessary feature set.

An example attribution under Gately Feature Attribution and KS is shown in Figure 5.7.1 under $\alpha = 0.5$. The only necessary feature in this example is X_1. Our results show that the Gately Feature Attribution and the Shapley value differ in the type of attribution generated. The Shapley value assigns non-zero attribution to non-necessary features. This matches our earlier discussion of the Shapley value as a manifestation of the average effect of a feature over all possible coalitions. Gately Feature Attribution, due to the reduction in number of coalitions considered, attributes zero weight to non-impactful features in the grand coalition or their individual coalition and thus attributes the entire prediction change to X_1.

From Table 5.2 we see that the Gately Feature Attribution matches the optimum attribution more frequently than the Shapley value over all weightings of α. Unsurprisingly, the higher the value of α, the more each attribution method coincides, as the impact of X_1, when averaged over all coalitions, becomes more apparent. These results show that Gately Feature Attribution

is better, compared to the Shapley value-based comparisons, at finding the minimum set of features for a prediction change, particularly in the presence of multiple necessary features, which motivates the use of Gately Feature Attribution over the Shapley value in situations where a minimal attribution is preferred, for example, if an end-user is interested in an explanation which is sparse, or if they are looking to make as few changes as possible which change their outcome.

	Gately Feature Attribution	Shapley Value	KS
$\alpha = 0.1$	**0.78**	0.24	0.24
$\alpha = 0.5$	**0.90**	0.54	0.54
$\alpha = 1$	**0.94**	0.79	0.79
$\alpha = 10$	**0.99**	0.97	0.97

Table 5.2: Table shows the proportion of the 200 attributions, generated under Gately Feature Attribution, the Shapley value and KS, which match the optimum attributions described in Section 5.7.1 under varying settings of α. The results show that Gately Feature Attribution outperforms the Shapley value and KS when determining necessary effects of features over all values of α.

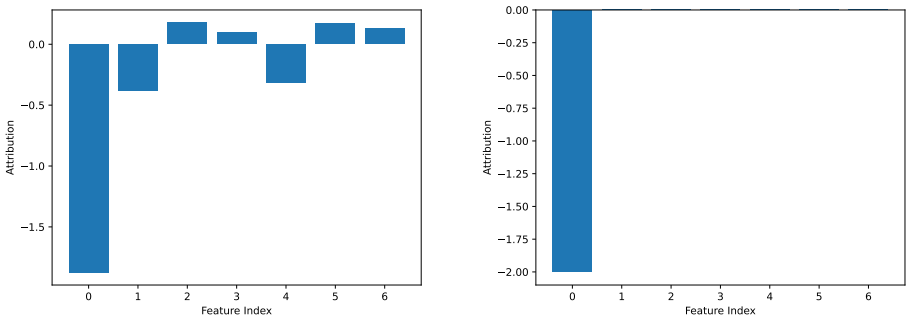


Figure 5.3: Figure shows an example attribution generated by the Shapley value (left) and Gately Feature Attribution (right). We can see that Gately Feature Attribution identifies only the necessary feature index for the given sample and baseline and attributes all the prediction change to this feature. The Shapley value, despite assigning the highest importance to this feature also gives non-zero attribution to the other features despite them not being necessary for the example.

In this section, we have experimentally shown how the Gately Feature Attribution results in minimal feature attribution explanations when compared to Shapley value alternatives.

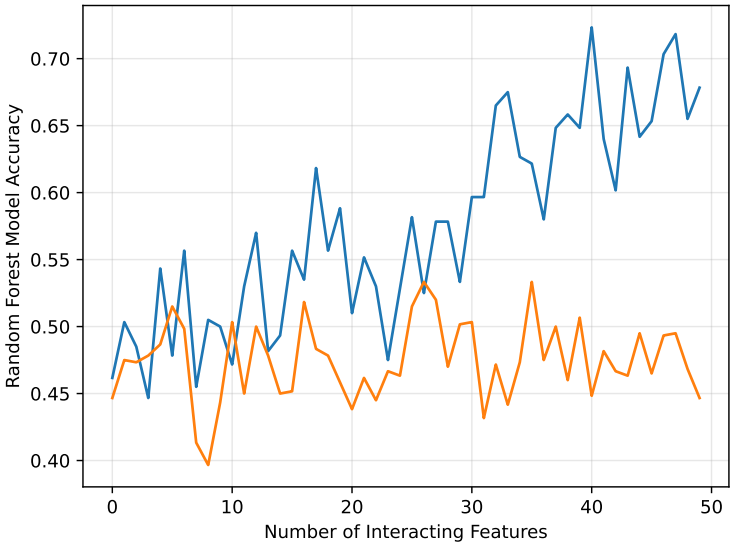


Figure 5.4: Figure shows the Random Forest Classifier's ability to distinguish between real and hypothetical samples as the number of correlated features increases. The blue line shows the classifier's performance trained with samples generated by the Shapley value and the orange line shows the performance on samples generated by Gately Feature Attribution. The figure validates our claim that the attributions generated by Gately Feature Attribution are more adversarially robust to those of the Shapley value, particular in the presence of dependent features.

5.7.2 Off-manifold Synthetic Experiment

In this section, we experimentally show that that the attributions generated by Gately Feature Attribution are less susceptible to the adversarial attacks which are known to affect the Shapley value. The argument of Slack et al. [194] assumes that there is an adversary with an incentive to deploy a biased classifier f for making a critical decision. The adversary must provide black-box access to end-users, if the end-users determine that f is biased, they will refuse to deploy it in practice.

Slack et al. [194] argue that if we can differentiate between data points coming from the input distribution and instances generated via perturbation, an adversary can create a scaffolding classifier g that behaves like f with potentially biased behaviour on the input data points, but appears to be unproblematic on the perturbed instances which are then used to generate an explanation.

For our experiment we adopt the approach of Slack et al. [194] and first construct 50 synthetic datasets, each containing 1000 samples of 50 features, with each dataset containing a varying number of interacting features. Dataset i contains i interacting features and $50-i$ independent features. For interacting features we first sample $\mathbf{X}_i \sim U(-1,1)$ and then we construct each X_j by first sampling the correlation parameter $\rho \sim U(0,10)$ and then constructing $X_j = \rho X_i$.

For each sample in each dataset we generate the perturbation datasets using both the Shapley value and Gately Feature Attribution under v_{bs}. For our scaffolding classifier we use a Random Forest with 100 estimators. We construct our classification dataset using 100 randomly selected samples of the perturbation dataset and 100 randomly selected samples of the original dataset for each attribution technique and each interaction dataset. We use a train-test split of $(0.7, 0.3)$ and evaluate the mean model accuracy on the held-out test over ten instances of the model for each dataset.

Our results are shown in Figure 5.4 which shows that as the number of dependent features increases, the scaffolding classifier is able to distinguish between real and hypothetical samples generated under the Shapley value attribution with increasing accuracy. In contrast, for Gately Feature Attribution, the scaffolding classifier's ability to distinguish between samples remains constant. In Figure 5.4, therefore, we can see that the attributions generated by Gately Feature Attribution are less vulnerable to the kind of adversarial attack as specified by [194] than those generated by the Shapley value, particularly in the presence of a large number of interacting features.

In this section, we have experimentally shown how the Gately Feature Attribution is more robust to explanation adversarial attacks than Shapley value alternatives

5.7.3 Attributions On Real World Data

We now evaluate Gately Feature Attribution on real-world datasets: the Crime, Boston and Correlation datasets from the Shap library [139]. We first split each dataset into a test and train set split of 0.7 vs. 0.3. We train an XGBoost regression model on the provided train set obtaining R^2 score of 0.57 0.91 and 0.72 on the held-out test respectively. We compute the Gately value attributions for 100 randomly selected samples from the test set. We compare Gately Feature Aattributions with the most commonly used Shapley value approximation techniques: Tree SHAP (TS) [136] and Kernel SHAP (KS) [139] as introduced in Section 4.7. To evaluate the attributions generated by Gately Feature Attribution, KS and TS in the absence of a ground truth attribution we use Average Deletion which we introduced in Chapter 4 (Equation 4.34). A lower AD is an indication of a better attribution technique.

Table 5.3 shows that the attributions under Gately Feature Attribution, when masking the single most important feature returned by the attribution method to measure the deletion score, outperform those generated by KS and TS across on the Boston and Correlated datasets yet are outperformed by KS on the Crime dataset. When instead using the top 50% of features to mask (Table 5.4) we note that the Gately value outperforms both KS and TS across all datasets. We note that the most important feature as determined by Gately Feature Attribution coincides with that of KS on 95% of test samples for the Boston dataset, 68% on the Correlation dataset and 91%

	Gately Feature Attribution	KS	TS
Boston	$\mathbf{0.487 \pm 0.610}$	0.492 ± 0.623	0.492 ± 0.627
Correlated	$\mathbf{0.804 \pm 0.645}$	0.858 ± 0.645	0.848 ± 0.631
Crime	0.659 ± 0.665	$\mathbf{0.647 \pm 0.671}$	0.659 ± 0.684

Table 5.3: Table shows AD and standard deviation for each of the attribution techniques evaluated on the Boston, Correlated and Crime real-world data benchmarks. In this experiment only the most important feature as indicated by the attribution technique was masked. Gately Feature Attribution outperforms both KS and TS for the Boston and Correlated dataset yet is outperformed by KS on the Crime dataset.

	Gately Feature Attribution	KS	TS
Boston	$\mathbf{0.057 \pm 0.073}$	0.078 ± 0.118	0.072 ± 0.113
Correlated	$\mathbf{0.036 \pm 0.058}$	0.068 ± 0.080	0.038 ± 0.046
Crime	$\mathbf{0.060 \pm 0.078}$	0.103 ± 0.100	0.069 ± 0.075

Table 5.4: Table shows AD and standard deviation for each of the attribution techniques evaluated on the Boston, Correlated and Crime real-world data benchmarks. In this experiment the top 50% features as indicated by the attribution technique were masked. Gately Featuree Attribution outperforms both KS and TS for all three datasets.

on the Crime dataset. However, when comparing the top 50% of features ordered by attribution, the orderings generated by Gately Feature Attribution coincide only with those generated by KS in 15% of the Boston test samples and 0% of the Correlation and Crime test samples.

We note that the most important feature as determined by Gately Feature Attribution coincides with that of TS on 96% of test samples for the Boston dataset, 69% on the Correlation dataset and 85% on the Crime dataset. However, when comparing the top 50% of features ordered by attribution, the orderings generated by the Gately value coincide only with those generated by KS in 7% of the Boston test samples and 0% of the Correlation and Crime test samples.

All three attribution methods identify the same most important feature in the majority of test samples across all three datasets. On those samples where the Gately Feature Attribution disagrees with the Shapley value-based alternatives we find that the Manhattan distance between samples and their reference is greater than the average over all samples and references for both the Boston and Correlated dataset, 18% more than average for the Boston dataset and 5% more than average for the Correlated dataset. For the Crime dataset, in contrast, we find that the samples and associated references for which all three attribution methods disagree on the most important feature have a mean Manhattan distance 20% below average. From these results, we argue that the Gately value is a preferred solution concept to the Shapley-based alternatives on samples and associated references from distant regions of the input space. Intuitively, this makes sense as Gately Feature Attribution, unlike the Shapley value, does not consider a great number of hypothetical samples where, particularly for samples with very different feature co-ordinates, these hypothetical samples may detract from the importance of that feature for the given output.

This behaviour is reinforced by the results in Table 5.4 which mask the top 50% of features from the input. As we increase the number of features value we mask from the original sample, we expect the resulting prediction to be a lot closer to that of the reference value compared to only masking the most important feature, particularly for the Correlation and Crime datasets which are of greater dimensionality than the Boston dataset. Furthermore, for this experiment, Gately Feature Attribution outperforms both KS and TS for all three datasets.

As we increase the number of feature values we mask from the original sample we expect the effect of misleading hypothetical samples on the resulting masked prediction to also increase and this is what we observe, particularly for the Crime dataset where Gately Feature Attribution now achieves a lower deletion score than both TS and KS.

We also draw attention to the variability of KS which arises due to the sampling procedure upon which this approximation of the Shapley value relies on. We construct a sample to be at the upper limit of the input space (i.e feature values are set as the maximum over the domain) with the corresponding reference sample to lie in the middle of the input space for all three datasets (i.e. feature values are set as the mean over the domain). We generate 10 separate attribution vectors under 50 distinct parametrisations of KS, increasing the number of samples used to approximate the Shapley value, ranging from 100 to 5000. We plot the mean (over the ten attribution vectors) AD score after the top 50% of features have been masked. For comparison we also compute 10 attribution vectors of the Gately value and TS.

These results are shown in Figure 5.5. We first note that as both TS and the Gately value attribution methods do not rely on sampling, the variance of these measures over the all settings of KS remains constant. In contrast, we note the high variability of KS under the sample size parametrisation. This experimentally confirms the problems associated with KS whereby the relative feature attributions generated for the same sample are likely to vary across multiple calls to the KS algorithm. This occurs as the sampling procedure means that different hypothetical samples are used to weight the attributions with each call to the KS algorithm. In contrast, the attributions generated by Gately Feature Attribution and TS are robust across multiple calls to each algorithm. The attributions generated by Gately Feature Attribution value outperform both KS and TS for the sample from the Correlation and Crime dataset and coincides with those generated by TS for the sample from the Boston dataset. We also re-emphasise that TS can only be used for tree-based underlying models.

In this section, we have experimentally motivated the use of the Gately value for feature attribution compared to state-of-the-art Shapley-based alternatives in real world settings.

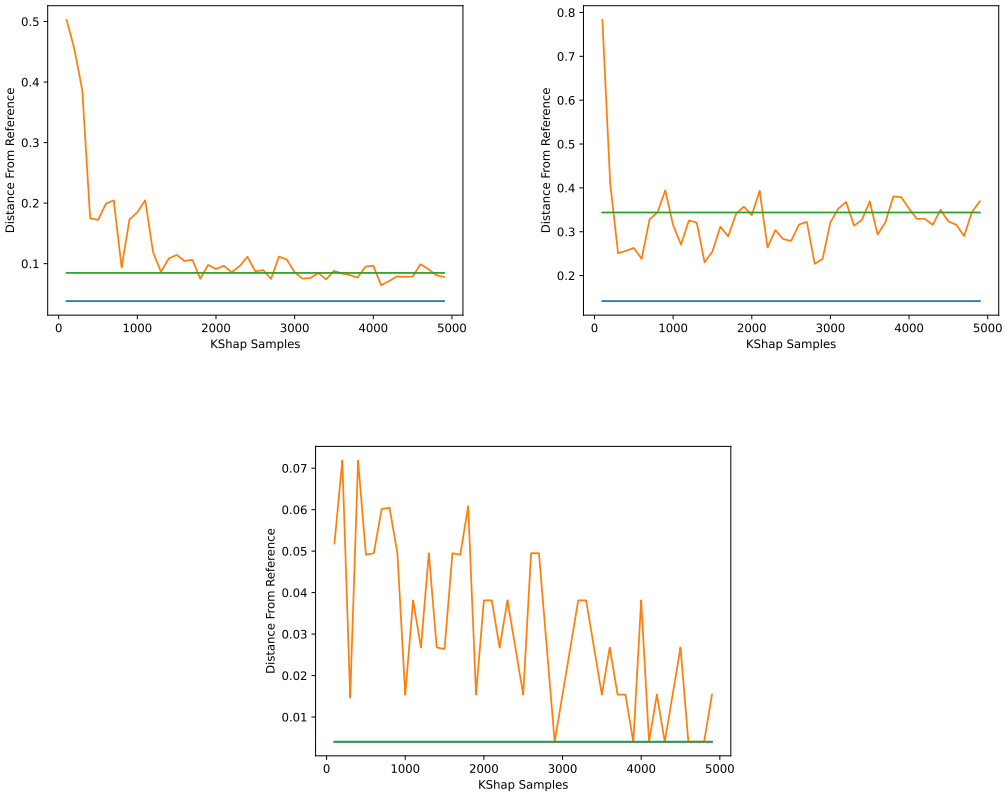


Figure 5.5: Figure shows how AD varies for different parametrisations (number of generated samples) of KS, indicated by the orange line for the Crime (top left), Correlated (top right) and Boston dataset (bottom). In contrast, as TS and Gately Feature Attribution do not rely on sample size parameter AD is constant over multiple iterations.

5.8 Gately Feature Attribution: Concluding Remarks

In this chapter we have explored the extent to which the disciplines of the philosophy of causality, causal inference and XAI can be aligned. We have compared the different ways in which causes can be identified, showing how different identification practices relate to different investigative questions, above all, we hope we have demonstrated the ambiguity and subtlety surrounding the definition of a cause. We have discussed the differences between a general and singular account of causality and shown how the Shapley value measures the probability of necessity *and* sufficiency of a player in a game with a binary outcome. We have motivated the use of bifactuals for post-hoc local explanations and shown how the Gately value can provide an axiomatically justified feature attribution which is in line with a bifactual question. We have experimentally and theoretically demonstrated the advantages of the the Gately value of the Shapley value for feature attribution.

The use of the Gately value to generate post-hoc local explanations, akin to the Shapley value is accompanied by its own set of limitations which we discuss further in Chapter 7. We

have dedicated a significant proportion of this thesis to the analysis of the Shapley value and its application to feature attribution. The Shapley value is, and will continue to be one of the main approaches for feature attribution. As such there is a large amount of Shapley-based feature attribution literature. However, the argument that we have constructed throughout the previous two chapters has contextualised the Shapley value within a game theoretic context and has connected the Shapley value to causality and counterfactuals.

In addition to presenting two alternative feature attribution methods, the overarching goal of this thesis is to draw attention to the benefits of considering the Shapley value, and more generally, post-hoc local explanations from a multidisciplinary perspective. Furthermore, by relating the Shapley value to contrastive questions we have demonstrated the importance of considering the investigative goal of a post-hoc local explanation when selecting an appropriate method. In the following chapter we connect the ideas in the previous two chapters to the problem of time series attribution. We generate time series explanations using both the Gately value and Shapley sets and demonstrate how both solutions address some of the challenges we discussed in Chapter 3.6. We re-visit the game theoretic literature to propose an alternative attribution method, based on the Aumann-Shapley value which is better suited to attribute time series predictions.

CHAPTER 6

DIFFERENTIAL ATTRIBUTION: TOWARDS TEMPORALLY AWARE ATTRIBUTIONS FOR MULTIVARIATE TIME SERIES

We now return to time series attribution. We first connect the work of Chapter 4 and Chapter 5 with that of Chapter 3 by discussing some of the challenges which must be addressed when using Shapley value-based attribution to explain univariate time series models. We continue by motivating both Shapley Sets and Gately Feature Attribution for this data-type.

The second half of this chapter shifts focus to multivariate time series attribution where we first explore why attribution on this data structure is particularly challenging. We introduce the concept of Differential Attribution as an explanation setting which requires the consideration of a multivariate feature set with underlying temporal dependence.

We connect Differential Attribution with game theoretic literature and motivate the Aumann-Shapley value over the Shapley value for this kind of attribution. We develop a post-hoc local attribution method based on the Aumann-Shapley value which constructs a local surrogate model surrounding individual multivariate time series. We experimentally demonstrate the benefits of our approach over existing multivariate attribution methods.

6.1 Applying The Shapley Value To Univariate Time Series Prediction

A univariate time series prediction problem is defined over the variable set $\mathbf{X}$ each individual sample $\mathbf{x}$ in contains t temporally ordered observations such that $\mathbf{x} = \{x_1, ..., x_t\}$. Each individual time series is associated with a label, $y_i \in \{c_1, ..., c_m\}$ for classification problems and $y_i \in \mathbb{R}$ for regression. The associated time series model either outputs a vector of class probabilities for classification problems $f : \mathbb{R}^t \rightarrow \mathbb{R}^m$ or a singleton value for regression problems $f : \mathbb{R}^t \rightarrow \mathbb{R}$. When

we consider post-hoc local explanations for univariate time series we are concerned with finding the set of observations $\mathbf{X}_j \subset \mathbf{X}$ which were most influential for the local prediction $f(\mathbf{x})$.

Historically, Shapley-based attribution has been applied to univariate time series data by treating each observation in a time series as an individual feature. This results in an individual attribution for each observation [151, 214]. Due to the high-dimensionality of time series data, Shapley values must be approximated. Kernel SHAP [138] is the most commonly Shapley-based approximation used to compute attributions for time series data.

While attributions generated by Kernel SHAP for time series data have been shown to be more robust than those generated by other feature importance mechanisms (e.g. LIME [175] and DeepLift [189]), applying Shapley value-based feature attributions to time series data suffers from similar limitations as the application of LIME which were discussed in Chapter 3. Below we outline some of the problems surrounding the application of Shapley value-based attribution to univariate time series data.

Comprehending High Dimensional Attributions. An explanation which comprises individual attributions for each observation in high-dimensional objects is difficult for humans to comprehend as there are too many attributions to make sense of [14]. For this reason, interpretable concepts, or super-pixels have been incorporated into feature attribution methodology for images. However, we have seen in Chapter 3 how the concept of super pixels does not naturally transfer to time series as this datatype does not lend itself naturally to semantically meaningful segmentation. Chapter 3 has shown that even if we were to decompose a time series into interpretable concepts prior to attribution, these groupings, while being semantically meaningful, may be susceptible to misleading attributions. As the Shapley value, is by nature, an average of a feature's contribution over all possible combination of features, for high-dimensional objects, we have seen in Chapter 5 that the likelihood of the majority of Shapley values being non-zero is high, despite many of these features having a negligible impact on the prediction of the sample to be explained.

Fully Separable Approximation For Non-Linear Models. The Shapley value, as we have seen in Chapter 4, imposes a fully additively-separable function on the underlying machine learning model. We have seen in Chapter 4 how this results in misleading attributions in the presence of a high level of interaction in the model. This behaviour is particularly problematic for time series models which are becoming increasingly non-linear to capture the complexity of the underlying time series data they are modelling [237].

Off-Manifold Value Functions. When computed with off-manifold value functions, the Shapley value for time series is likely to suffer from issues relating to off-manifold samples. Compared to other data types, the features, or observations which make up a time series are likely to be dependent. Thus, the use of off-manifold value functions for time series is, in itself, questionable as the off-manifold samples are likely to be very different from the real data distribution.

On-Manifold Value Functions. When computed with on-manifold value functions, the Shapley value for time series will have a large number of statistically interacting observations, which, when considered as individual features, will be vulnerable to the failure of sensitivity as outlined in Chapter 4.

Given the above discussion regarding the issues relating to the application of the Shapley value to attribute individual observations belonging to a univariate time series, the following section motivates both Shapley Sets and Gately Feature Attribution as explanation mechanisms which overcome some of the problems outlined above for time series attribution.

This section has outlined the limitations affecting the application of Shapley-value-based explanations to univariate time series when each observation is treated as an individual feature.

6.2 Applying Shapley Sets To Time Series Attribution

Shapley Sets decomposes a variable set into non-overlapping interacting groups to provide its attributions. For high-dimensional objects like time series this kind of attribution is particularly useful as it reduces the dimensionality of the resulting attribution vector. An end-user will be able to compare attributions over a set of Non Separable Variable Groups (Definition 4.10) (NSVGs) of dimensionality $m \leq t$ reducing the complexity burden associated with Shapley value-based attributions of high-dimensional objects. Furthermore, the groups of NSVGs, whether they are computed via an off or on-manifold value function, can be viewed as an alternative for a super-pixel or a super-segment which are generated automatically rather than specified manually prior to the attribution.

In Chapter 3 we constructed a segmentation approach built on the temporal coherence assumption where observations which are close together in time are attributed together. The segmentation of the variable set according to Shapley sets operates via a different assumption whereby interacting observations either in the model or/and in the data are attributed together. Under the temporal coherence assumption, a conceptualisation of a time series into super-segments prioritises what would, we assume, to be the most intuitive dimensionality reduction for the end-user. The conceptualisation under Shapley Sets prioritises a dimensionality reduction from the perspective of the model.

We have seen in Chapter 4 how, when used with the conditional value function, Shapley Sets is capable of capturing interaction between variables reflected in the underlying data distribution. When applying Shapley Sets to time series via the conditional value function we therefore capture the statistical interactions between observations. We have seen in Chapter 4 that correlated features, above the interaction threshold ϵ are placed within the same NSVG. Just as correlation

measures the extent of a linear relationship between two variables, autocorrelation measures the linear relationship between two observations of the same variable. Shapley Sets, under the conditional value function, therefore decomposes a time series according to the autocorrelation between observations [180].

Is this decomposition semantically meaningful? It is highly likely that observations which are statistically related reflect the same homogeneous behaviour of the underlying time dependent variable. However, the aurocorrelation function decays substantially over time, as such long-ranging temporal dependence may not be captured by the measure [145]. This motivates the importance of the interaction threshold ϵ when identifying NSVGs for time series as there is likely to be a higher level of interaction in the data due to autocorrelation. Indeed, the only type of time series with zero interaction between observations is a white noise process and these processes are the exception rather than the norm in time series modelling. We discuss the selection of ϵ in the Shapley Sets algorithm further in Chapter 7.

To demonstrate the utility of Shapley Sets for post-hoc local explanations for univariate time series we apply Shapley Sets to explain time series classification. We use the Italy Power Demand dataset [104] which contains 1086 samples of 24 hourly observations of household power demand. Each time series represents a day in winter or summer. We train a Convolutional Neural Network (CNN) classifier with 20 epochs and batch size of four from the SKTime library [135] to obtain a predictive accuracy of 0.96 on the held out test set. We generate the attributions for a random sample from the test set via Shapley Sets under v_{cond}, v_{marg}. As the CNN is not a tree model, we obtain Kernel SHAP attributions as a comparison technique. Figure 6.1 shows the power demand time series for a day in the winter (left) and a day in the summer (right). The blue curve represents the expected hourly demand.

Kernel SHAP identifies 6pm as the most important hour. However, Shapley Sets under v_{marg} offers a more meaningful attribution: the model considers the hours between 6 and 11 pm as most influential in classification. It is commonly known that the evening hours experience greater power demand during the winter months in the northern hemisphere [109] thus, it is not surprising the model has learned this behaviour. Furthermore, Shapley Sets under v_{cond} identifies 12 to 4 am as interacting with 6 to 11 pm, this again is not surprising as Shapley Sets Conditional has learned the temporal dependency between the night time hours. From Figure 6.1, we can see how “the most important part of the input’ as determined by Shapley Set attribution under the conditional value function provides the most detailed information out of the three attribution methods evaluated, where the “most important part of the input” as determined by Kernel SHAP is limited to a single observation.

This section has shown that Shapley Sets provides an automatic conceptualisation of a univariate time series which result in interpretable, low dimensional attribution vectors.

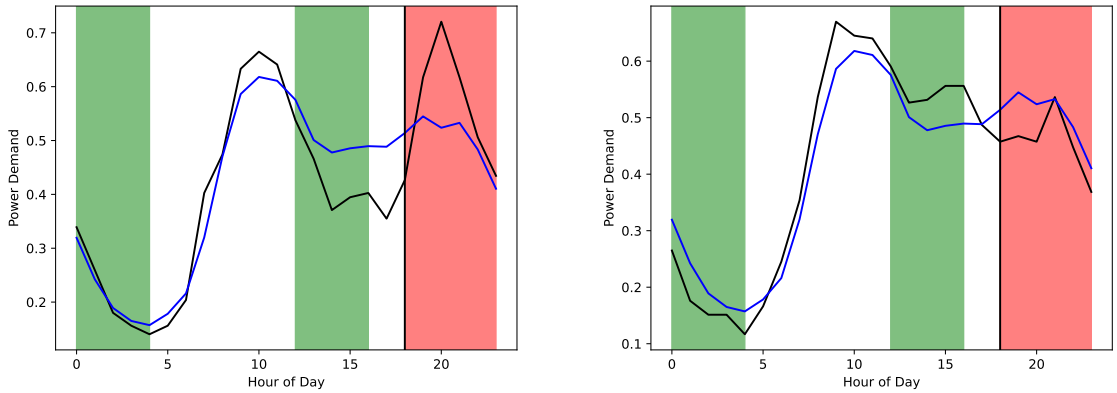


Figure 6.1: Figure shows two individual time series (black) and expected time series (blue) representing power demand for a day in winter (left) and summer (right). Black vertical line corresponds to the most important hour as returned by Kernel SHAP, the red segment indicates the most important hours as returned by Shapley Sets Marginal and the green segments show the additional most important hours as returned by Shapley Sets Conditional. Shapley Sets attributions offer more insight into the underlying phenomenon than the individual attributions of Kernel SHAP

6.3 Applying Gately Feature Attribution to Time Series Attribution

We now motivate the use of Gately Feature Attribution (Definition 5.18) for univariate time series attribution. As we have discussed in Chapter 5, the Shapley value under an off-manifold value function generates a large number of hypothetical samples over which it determines a feature's marginal contribution.

For high-dimensional objects such as time series there will be a large number of hypothetical samples which represent time series which neither resemble closely the instance to be explained nor the reference value under consideration but some hybrid instance which could lie well outside the data distribution. The marginal impact of an individual observation on these hypothetical samples will be weighted equally regardless of how closely they resemble the sample and reference.

Under Gately Feature Attribution, in contrast, the only hypothetical samples which are considered for each feature's attribution are in fact the original sample and reference sample with just the feature in question removed. In Chapter 5 we showed how Gately Feature Attribution results in minimal explanations compared to Shapley value-based alternatives.

To experimentally show the benefits of these minimal explanations for time series attribution we apply both Baseline SHAP and Gately Feature Attribution to the Italy Power Demand classification task and the GunPoint classification task. For our implementation of Baseline

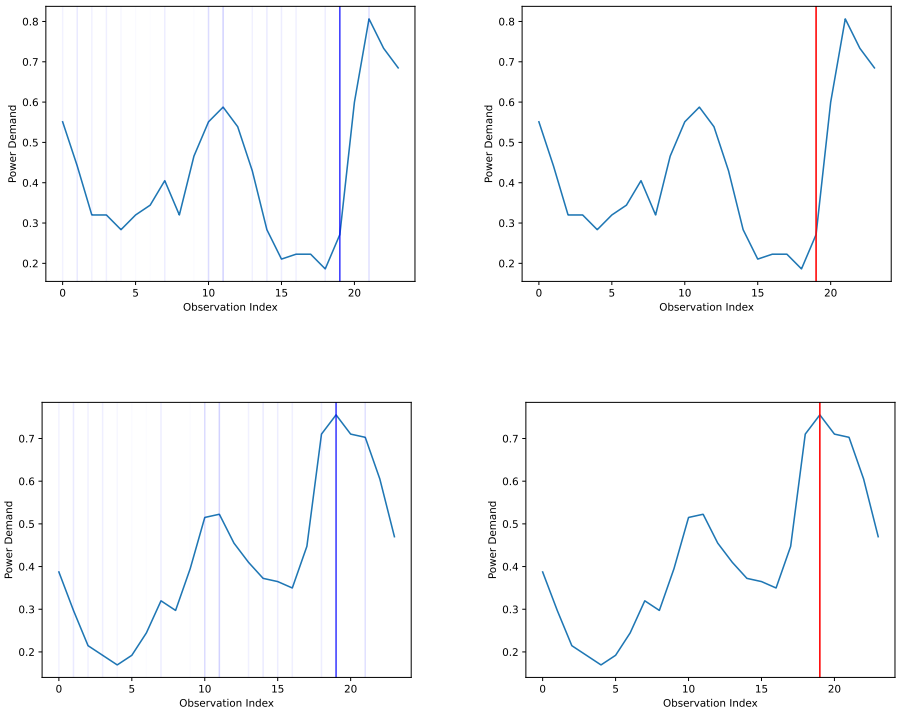


Figure 6.2: Figure shows Two individual time series from the Italy Power Demand dataset, set as the reference sample (bottom row) and target sample (top row) showing the attributions generated by Baseline SHAP (left column) and Gately value (right column). The vertical lines represent that the attribution technique has identified the associated time step as being influential in the local classification. The darker the colour, the more influential that feature is. The figure demonstrates the minimality of Gately Feature Attribution compared to Baseline SHAP

SHAP we use the same implementation of Kernel SHAP as outlined in Section 4.7 but with a singular reference value. The Italy Power Demand classification is implemented as described in Section 6.2 for two randomly selected samples of the test set we obtain Gately Feature Attribution and the Baseline SHAP attribution by treating one sample as the reference and the other as the target. Figure 6.2 shows the Baseline SHAP (left) and Gately Feature Attribution (right) for the reference (bottom) and corresponding sample (top). The GunPoint dataset is a univariate time series dataset containing 50 train and 150 test samples of length 150 recording an individual's hand movement. the classification is to distinguish male and female hand movement from the time series. We train a Convolutional Neural Network (CNN) classifier with 20 epochs and batch size of four from the SKTime library [135] to obtain a predictive accuracy of 0.88 on the held out test set.

We select two random samples from the test set to represent our reference and sample time series for which we obtain Kernel SHAP and Gately Feature Attributions. The resulting attributions are shown in Figure 6.3 which shows the Kernel SHAP attribution (left) and Gately

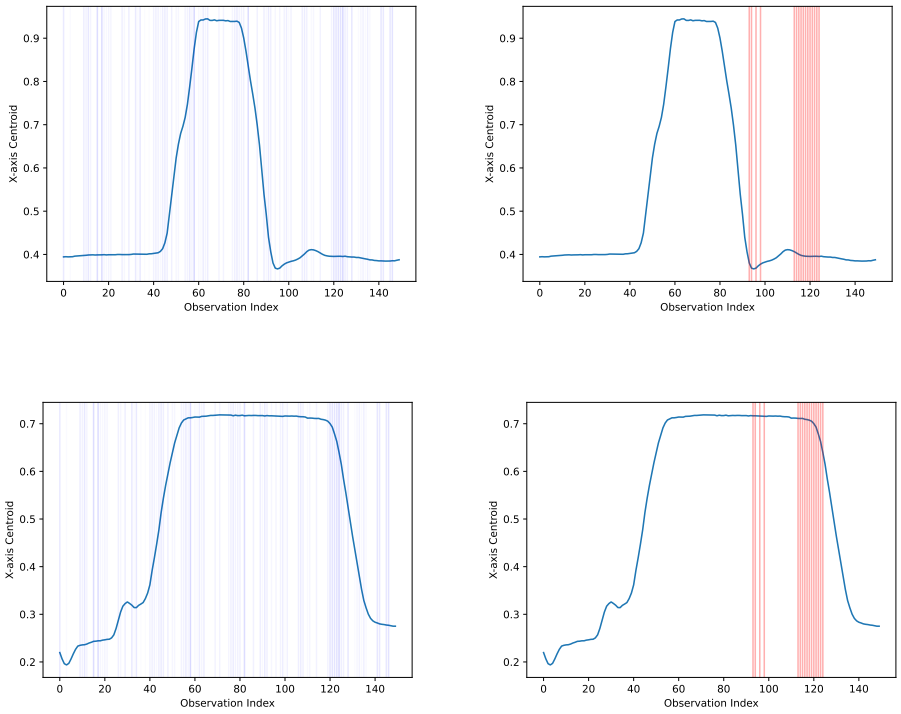


Figure 6.3: Figure shows two individual time series from the GunPoint dataset, set as the reference sample (bottom row) and target sample (top row) showing the attributions generated by Baseline SHAP (left column) and Gately Feature Attribution (right column). The vertical lines represent that the attribution technique has identified the associated time step as being influential in the local classification. The darker the colour, the more influential that feature is. The figure demonstrates the minimality of Gately Feature Attribution compared to Baseline SHAP

Feature Attribution (right) for the reference (bottom) and sample (top). For both classification datasets we can see that Gately Feature Attributions are more minimal than those afforded by Kernel SHAP.

In practice, an attribution which gives an end-user only the necessary features to change would be useful in many applications, particularly those tasKernel SHAP where there is an overhead costs associated with changing certain features. Similarly, for the GunPoint classification, Figure 6.3 shows that under Gately Feature Attribution, a considerably smaller proportion of observations are indicated as influential compared to the attribution of Kernel SHAP.

We argue that Gately Feature Attribution provides a simpler, clearer visual explanation of the influential regions of the time series. Furthermore, perturbing the values of each observation of the sample indicated as influential by Gately Feature Attribution are enough to change its outcome whereas for the attribution offered by Kernel SHAP there is no clear identification of the minimal set of perturbations necessary for this change in outcome.

Where Shapley Sets provides an automatic conceptualisation of the input time series, Gately Feature Attribution offers an alternative solution to the conceptualisation challenge. Rather than grouping individual observations into meaningful concepts, when using Gately Feature Attribution, we have experimentally shown how the bifactuality of this attribution technique result in minimal attribution vectors which are more interpretable than the Shapley value.

This section has shown that Gately Feature Attribution provides a conceptualisation via minimisation of a univariate time series which result in interpretable, low dimensional attribution vectors.

6.4 From Univariate To Multivariate Explanations

So far in this thesis we have studied univariate time series. Now, we turn our attention to generating explanations for multivariate time series. As discussed in Chapter 2, due to the increasing ubiquity of edge computing and the Internet of Things, there exists a growing volume of multivariate time series data. Harnessing this rich data source via the development of multivariate time series models is an important research topic in the machine learning community.

A multivariate time series prediction problem takes as input a time series variable set $\mathbf{X}_n^t$ consisting of n features each recorded at t timesteps. Each individual time series $\mathbf{x}$ is associated with a label, $y_i \in \{c_1, ..., c_m\}$ for classification problems and $y_i \in \mathbb{R}$ for regression. The associated time series model either outputs a vector of class probabilities for classification problems $f : \mathbb{R}^{n \times t} \rightarrow \mathbb{R}$ or a singleton value for regression problems $f : \mathbb{R}^{n \times t} \rightarrow \mathbb{R}$.

In Chapter 2 we have motivated the importance of developing post-hoc explanations for multivariate time series. When considering the application of existing post-hoc local explanations to multivariate time series, the challenges discussed in Section 6.1 persist, however, these problems are further exacerbated by the inter-dependencies which now exist between variables. For example, consider a model which predicts the probability that a given patient will survive sepsis given their patient trajectory. The model has learned that a spike in temperature, if followed by a delayed spike in blood pressure, dramatically increases the probability that a patient will die. In this example, both the relationship between the two variables, temperature and blood pressure, over time is crucial for the model. Considering either each observation or each time series as individual features would fail to capture the temporal dependency.

The limitations of existing post-hoc explanations on multivariate time series have been noted in the literature [183, 219]. Saluja et al. [183] conducted a user study on both LIME and SHAP explanations for multivariate time series, where each individual time series within a sample were treated as independent features [183]. The study found that nearly half of those surveyed said that they would not be able to trust the model output even after receiving the

explanation. Furthermore, it was argued by Wang et al. [219] that existing methods for post-hoc local explanations, when used to attribute the outcome of Recurrent Neural Networks, which encode temporal structure, fail to identify salient features.

Existing post-hoc explanations, such as LIME and SHAP, are static and as such, the temporal dependency and context are forgotten when treating all the time steps as separate features [39]. In the following section we develop an attribution mechanism which is specifically designed to consider the temporal dependency across multivariate features such that we attribute a change in model output which occurs over time with respect to the change in feature values. First, however we introduce the general context underpinning our attribution technique which we term *Differential Attribution*. We highlight two explainability contexts which are suited to Differential Attribution as they both handle multivariate datasets with a temporal dependency structure. First we give two motivating examples representing both Differential Attribution settings from a healthcare perspective.

In this section we have motivated and dicusssed the challenges of designing post-hoc local explanations which can be appropriately applied to multi-variate time series.

6.5 Differential Attribution

In this section we introduce our concept of Differential Attribution which characterises a particular kind of post-hoc local explanation intended to attribute a **change** in model outcome between two temporally ordered instances. We distinguish between Differential Attribution and the other types of post-hoc local explanations we have studied in this thesis, henceforth termed *Static Attribution*. Differential Attribution takes as input two samples and associated model outcomes which are separated by some temporal interval. The Differential Attribution challenge therefore, is to explain the change in model outcome, which occurs between the two samples, in terms of the change of their feature values. In contrast, Static Attribution, for which the Shapley value is a method, take as input a single sample and associated model outcome and attribute in terms of feature values with reference to a baseline sample. Throughout this thesis we have discussed the implications of selecting a baseline sample which is generally selected to represent an “uninformative” reference value.

From this discussion we can see how Differential and Static Attribution differ in that Differential Attribution places equal importance on both samples when attributing change in model outcome. Below we outline two healthcare settings which require a post-hoc local explanation and argue why Differential, rather than Static, Attribution is more aligned with the associated investigative question.

In Chapter 1, we motivated explanations of individual patient trajectories within healthcare.

Implicitly encoded in this kind of data structure is the way in which a particular outcome, e.g. death from sepsis, depends on the progression of a set of factors. As we have discussed in Chapter 2, deep learning models have the potential to take as input a large volume of these patient trajectories and learn the underlying patterns and dependencies which can be used to predict a particular outcome. In this chapter, we argue that when we are explaining these kinds of models, an attribution which considers the temporal dependence between inputs and outputs is a better reflection of reality than one which attributes in terms of an individual sample and an uninformative baseline. Inspired by patient trajectory data, below we exemplify two different settings, Example 6.1 and Example 6.2, for which we motivate Differential over Static Attribution. The examples describe two different kinds of model which can be trained on patient trajectory data to predict death from sepsis.

Example 6.1. *First, consider a scenario where medical researchers are attempting to understand the reasons why certain patients admitted to ICU with sepsis do not survive. They have access to historical data containing the vital signs of patients taken each hour during their stay in ICU. These multivariate time series are used to train a black-box model which achieves high predictive accuracy on classifying surviving and non-surviving individuals. Given a set of new sepsis patient trajectories, the researchers seek to understand which vital sign observations and at which time during their stay in ICU were most influential for the corresponding classification.*

Example 6.2. *Now imagine a patient is admitted to the ICU with sepsis, the attending physician wants to monitor the patient's vital signs periodically to determine their immediate risk of dying from sepsis. The physician has access to an AI system which has been trained on historical static vital sign observations to predict the chance of that patient's survival in that instance. Each hour, the patient's vital sign observations are input to an AI system which predicts the patient's risk of death. Between two consecutive hours, the patient's prediction of survival has dropped from 0.8 to 0.3. The attending clinician needs to understand which of the patient's observations caused this drop in survival chance so they can treat the patient accordingly.*

Both Example 6.1 and Example 6.2 describe situations where an explanation in terms of the input features is required of a model. Each describes a multi-variate feature set with n input vital signs taken at t time-steps. Both examples describe a model outcome which is, in some way, dependent on time. Both exemplify the investigative question, *Which vital sign caused a change in outcome between two time steps?* It is the focus on a change in model outcome between two selected time-steps which characterises Differential Attribution.

While both examples require Differential Attribution, the kind of model and thus the kind of explanation required by each example is distinct. Both examples above take as input a multivariate time series. The difference between the settings in Example 6.2 and Example 6.1 lies in the fact that for Example 6.2 we have access to intermediate model outcome at each timestep whereas for Example 6.1, we have access to only one outcome as the model is trained

on a fixed input size of $n \times t$. Below, therefore, we formalise the kind of Differential Attribution question characterised by Example 6.1 and Example 6.2 respectively.

- **Multivariate Time Series Explanations:** How do we adapt local explanation methods to consider the temporal dependency in multivariate time series models?
- **Temporal Explanation:** How do we explain the change in model prediction between two temporally ordered instances?

To illustrate why Static Attribution methods such as LIME and SHAP may generate misleading Temporal or Multivariate Explanations let us adapt 6.2 to a simplified real-world setting with Example 6.3

Example 6.3. *Let us assume that we have three continuous features which are used to train the sepsis survival model: temperature, blood pressure and oxygen. A patient's initial values are 27, 42 and 18 which give an outcome of 0.2 probability of death from sepsis. However, after an hour, the vital signs are measured again, obtaining a reading of 27.5, 45 and 12 which now gives an outcome of 0.8. Here, we want to understand which feature was most influential in the change in outcome.*

Under the Static Attribution techniques of the previous three chapters (LIME and SHAP), we could obtain a post-hoc local explanation for both of these samples $\mathbf{x} = \{27, 42, 18\}$ or $\mathbf{x}' = \{27.5, 45, 12\}$. However, by explaining both samples independently, this attribution would not reflect the behaviour of the model between $\mathbf{x}$ and $\mathbf{x}'$, instead encoding the idea that the change in model is from the reference sample to $\mathbf{x}$ and then from the reference sample to $\mathbf{x}'$.

For example, under the Shapley value with v_{cond} (Equation 4.24), we may receive an attribution where temperature is the most influential feature in both samples with respect to the expected value. However, for the change $f(\mathbf{x}') - f(\mathbf{x})$, the change in blood pressure was actually the most influential. Section 6.5.1 elaborates further on why applying static methods to differential settings may generate misleading explanations.

One of the main contributions of this chapter is the motivation of the Aumann-Shapley value [15] for Differential Attribution. In Section 6.6 we introduce the game-theoretic context behind the Aumann-Shapley value and distinguish it from the Static Attribution provided by the Shapley value. In Section 6.7 we motivate the Aumann-Shapley value both theoretically, with examples, and axiomatically for Differential Attribution.

While the Aumann-Shapley value is theoretically and axiomatically desirable for Differential Attribution, just like we have seen for the Shapley value, mapping the Aumann-Shapley value from game theory to XAI is challenging. Primarily, the Aumann-Shapley value requires the underlying black-box function to be continuously differentiable. In Section 6.8 we discuss prior XAI research [206] which has applied the Aumann-Shapley value to feature attribution. However,

this existing method for post-hoc explainability can only be applied to continuously differentiable black-box functions and is a further example of Static Attribution.

A main contribution of this chapter therefore is our proposed Differential Attribution method, termed Aumann Differential Surrogate Explanations (ADSE) which we introduce in Section 6.9. ADSE constructs a continuously differentiable surrogate model around a multivariate time series and its associated function outcome which allows for the calculation of the Aumann-Shapley value. Unlike prior work mapping the Aumann-Shapley value to feature attribution [206], ADSE is model-agnostic and a method designed explicitly for Differential Attribution.

Having distinguished between the kind of Differential Attribution required by Temporal and Multivariate Explanations, Section 6.10 and Section 6.11 show how we adapt ADSE to generate explanations in both of these settings. We conclude the chapter experimentally comparing ADSE for Temporal and Multivariate Explanations with other state-of-the-art methods for multivariate time series explanations.

In this section we have introduced our concept of Differential Attribution as a specialised stype of post-hoc local explanation. We have presented an outline of the remainder of this chapter and stated our two main contributions.

6.5.1 Formalising Differential Attribution

Having introduced the intuition which separates Differential from Static Attribution, this section formalises Differential Attribution. Consider a function of several variables $f(X_1,...,X_n)$. Given a change in the values of these variables, we ask what portion of the overall change is due to the change in each particular variable, X_i. We have seen in Chapter 4, how this problem can be modelled as a coalitional game when each X_i is binary. We have explored in Chapter 4 and Chapter 5 how, and why, the Shapley value is considered the optimal attribution method and presented our own alternatives under the binary variable setting.

In this chapter, we explore the more general situation where variables are not binary but continuous, which applies to both Temporal and Multivariate Explanations as motivated in Example 6.2 and Example 6.1. This attribution problem has many names in the game theoretic literature including values of non-atomic games or attribution between variable demands of heterogeneous goods. In this chapter, we refer to the continuous value attribution problem as Differential Attribution.

Formally, given a real valued characteristic function $f : \mathscr{X} \to \mathbb{R}$ of n variables and initial and final values $\mathbf{x}, \mathbf{x}' \in \mathscr{X}$, our objective is to find the attribution vector $\mathbf{z} = \{z_1(\mathbf{x},\mathbf{x}',f),...,z_n(\mathbf{x},\mathbf{x}',f)\}$ where we can interpret $z_i(\mathbf{x},\mathbf{x}',f)$ as the portion of the change in f due to the change in the i'th variable such that

$$z_1(\mathbf{x},\mathbf{x}',f)+\ldots+z_n(\mathbf{x},\mathbf{x}',f)=f(\mathbf{x}')-f(\mathbf{x}). \tag{6.1}$$

A naive approach to Differential Attribution might be to determine the attributions $z_i(\mathbf{x},\mathbf{x}',f)$ as the difference of two individual post-hoc local static explanations which could be computed by LIME, for example. We would first calculate $\rho(\mathbf{x})$ and then $\rho(\mathbf{x}')$ and return the Differential Attribution as $\rho(\mathbf{x}')-\rho(\mathbf{x})$, where ρ corresponds to the static post-hoc local explanation method. However, this approach would not reflect the behavior of the function f between $\mathbf{x}$ and $\mathbf{x}'$, instead expressing the idea that the change in f is from $\mathbf{x}$ to some reference value, as determined by the perturbation method (Chapter 3), or the value function (Chapter 4), and then from $\mathbf{x}'$ to the reference value.

As we have discussed in Chapter 3, when we binarise continuous features, we encode several assumptions regarding the underlying data distribution. For Example 6.3, under the LIME formulation, the two temperature readings of 27.5 and 27 may both lie inside the same discretised bin and be considered as the same interpretable concept. Alternatively, the oxygen readings of 18 and 12 may lie in different bins giving the interpretable representation of 1 and 0. However, an alternative second oxygen reading of, say, 8 will also give an interpretable representation of 1 and 0 despite both situations representing contrasting changes over time with potentially different impact on the model.

Alternatively, adopting v_{bs} (Equation 4.23) we could use the Shapley value to determine the attribution. In this setting, however, we must first discretise each of the continuous features, indicating that in the initial reading, which we treat as the reference, the feature values are all "off" and in the sample to be explained, the final reading, the feature values are all "on". In this setting, each feature attribution reflects its contribution in changing the prediction of $\mathbf{x}$ to $\mathbf{x}'$. In this way, unlike the formulation under LIME, we encode a direction to the attribution such that we move from $f(\mathbf{x})$ to $f(\mathbf{x}')$. However, in the following section we shall show that under the Shapley value, we assume that the changes in each feature value happen independently and at once. i.e. for Example 6.3 the change in temperature from 27 to 27.5 happens instantaneously, before or after the change in blood pressure. We know that, in reality, changes in multiple continuous variables happen simultaneously as a function of time, not consecutively. Below we explicate some of the weaknesses of the Shapley value when applied to Differential Attribution.

In this section we have formalised Differential Attribution and have argued why Static Attribution may fail to accurately reflect the way in which continuous features evolve over time.

6.5.2 The Shapley Value And Continuous Features

When determining the Shapley value of the game characterised by v_{bs}, we consider the initial and final feature readings $\mathbf{x}$ and $\mathbf{x}'$ as two discrete values. In reality, for continuous features we know that these feature values move from the initial to the final values as a function of time. By ignoring the continuous nature of these feature values, the Shapley value may not correctly account for the way in which the underlying function f is attributable to each variable over time.

From our definition of the Shapley value (Equation 4.3), we know that the Shapley value assumes that each player, or variable, joins the game independently and in a random order, with a uniform distribution over the set of all possible orderings. Implicit in this definition is the assumption that players join the game *sequentially*. Within the context of coalitional value attribution, as each of the variables represents a player who can only be present or absent, this assumption makes sense. However, when variables represent features at initial and final values, an individual possible permutation, or ordering, places certain assumptions over the way in which variables move from initial to final values.

For Example 6.3, given the permutation $\{X_1, X_2, X_3\}$, where X_1 corresponds to temperature, X_2 to blood pressure and X_3 to oxygen, the Shapley value assumes that the temperature change happens first, and then the change in blood pressure and finally the change in oxygen levels. As the Shapley value is an expectation of each variable's marginal impact over all possible orderings it may take into account impact of features in orderings which are a complete mis-representation of reality. For example, the function may rely on the values of temperature and blood pressure together such that considering the ordering $\{X_1, X_3, X_2\}$ does not capture the true behaviour of the function over the attribution period. This behaviour of the Shapley value is particularly problematic in the presence of dependencies between variables as we show with Example 6.4 below.

Example 6.4. *Consider the function $f(\mathbf{X}) = 2^{X_1+X_2}X_3$ with variable set (X_1, X_2, X_3) where each variable is continuous and $X_2 = X_1$ such that X_1 is an exact replica of X_2. We first set initial feature readings $\mathbf{x} = (0,0,0)$ such that $f(\mathbf{x}) = 0$ and the final feature readings $\mathbf{x}' = (1,1,1)$ such that $f(\mathbf{x}') = 4$. The Differential Attribution task is such that we want to know which features, X_1, X_2, X_3 were most influential for the change in prediction $\Delta f = f(\mathbf{x}') - f(\mathbf{x}) = 4 - 0$.*

	$v(X_1) = 0$	$v(X_3) = 1$	$v(X_2, X_3) = 2$	$v(X_1, X_2, X_3) = 4$
$v(\varnothing) = 0$	$v(X_2) = 0$	$v(X_1, X_2) = 0$	$v(X_1, X_3) = 2$	

Applying the Shapley value to Example 6.4, under v_{bs}, the resulting attribution would be as follows: $\phi_{X_1} = 0.83, \phi_{X_2} = 0.83, \phi_{X_3} = 2.33$. Now, given the fact that $X_2 = X_1$ we can reformulate f as a new function f' on a modified variable set $\mathbf{X}' = (X_1, X_3)$ in such a way that $f'(\mathbf{X}') = 2^{2X_1}X_3 = f(\mathbf{X})$ such that f' behaves identically to f on the restricted variable set. As a result, the outcome

of f' evaluated on the initial $x = (0,0)$ and final $x' = (1,1)$ feature readings is identical to that of f' such that $f'(\mathbf{x}) = 0$ $f'(\mathbf{x}') = 4$. The Shapley values of the modified function f' are as follows

$$\begin{array}{lll} & v(X_1) = 0 & v(X_1, X_3) = 4 \\ v(\varnothing) = 0 & v(X_3) = 1 & \end{array}$$

where $\phi_{X_1} = 1.5, \phi_{X_3} = 2.5$. Under the modified function, the Shapley value attribution for X_3 has changed, despite it having identical impact across both f and f'. In using the Shapley value, we obtain a non-unique attribution [236]. This is particularly problematic due to the ubiquity of dependency between multivariate time series. For example, a machine learning algorithm may recognise the fact that a spike in temperature causes a spike in oxygen levels which together, have a positive impact on the outcome of death from sepsis. From the function's perspective, the spike in both temperature and oxygen should be considered together. However, under the Shapley value, the impact of the spike in both features are considered separately, impacting the attribution of the other features.

Given the above example we provide the following Desideratum for Differential Attribution which addresses the undesirable behaviour of the Shapley value above. **Differential Attribution Desideratum 1**: We want our attribution to reflect the way in which continuous variables change as a function of time.

In this section we have shown with Example 6.4, how the Shapley value, through its discretisation of features, generates attributions which do not reflect how the function depends on the change in variables over time.

6.5.3 The Shapley Value And The Attribution Region

A further problem which arises from the discretisation of initial and final values $\mathbf{x}$ and $\mathbf{x}'$ is that the function f is evaluated only at the endpoints of a feature change. The function may behave differently at the endpoints than it does over the entire attribution region as we illustrate with Example 6.5

Example 6.5. *Consider the function $f(\mathbf{X}) = (X_1 + X_2)X_3$ with variable set (X_1, X_2, X_3) which are all continuous features. Given the initial feature readings $\mathbf{x} = (0,0,0)$ such that $f(\mathbf{x}) = 0$ and the final feature readings $x' = (1,2,-1)$ such that $f(\mathbf{x}') = -3$. The Differential Attribution task is such that we want to know which features, X_1, X_2, X_3 were most influential for the change in prediction $\Delta f = f(\mathbf{x}') - f(\mathbf{x}) = -3 - 0$.*

$$\begin{array}{llllll} & v(X_1) = 0 & v(X_3) = 0 & v(X_2, X_3) = -2 & v(X_1, X_2, X_3) = -3 \\ v(\varnothing) = 0 & v(X_2) = 0 & v(X_1, X_2) = 0 & v(X_1, X_3) = -1 & \end{array}$$

Applying the Shapley value to Example 6.5, under v_{bs}, the resulting attribution would be as follows: $\phi_{X_1} = -0.5, \phi_{X_2} = -1, \phi_{X_3} = -1.5$. Now, suppose we introduce a new function as $f'(X_1^2, X_2, X_3) = (X_1 + X_2)X_3$ and want to attribute the change in f' when evaluated on $\mathbf{x}, \mathbf{x}'$ as defined above. Given that the evaluation of f' at the feature value endpoints $\mathbf{x} = (0,0,0)$ and $\mathbf{x}' = (1,2,-1)$ is the same as that of f, the resulting Shapley value attribution of f' is the same as that of f: $\phi_{X_1} = -0.5, \phi_{X_2} = -1, \phi_{X_3} = -1.5$, despite the fact that the modified function's treatment of X_1 is different to that of f over the entire attribution region.

This behaviour of the Shapley value is particularly problematic in situations where the initial and final values, i.e. the endpoints, mask the true dependency of the function on a given variable over the attribution region resulting in attributions which may indicate that a function is not attributable at all to a particular feature but in reality there is a non-zero dependency between this variable and the outcome. This discussion prompts the following **Differential Attribution Desideratum 2**: We want our attribution to account for the function's behaviour over the entire attribution region.

In this section we have set out the ways in which treating Differential Attribution as Static Attribution and using the Shapley value causes undesirable behaviour which is a misrepresentation of reality. We have introduced two desiderata for Differential Attribution. In the following section we introduce the game theoretic background on the Differential Attribution challenge. We motivate the Aumann-Shapley value [15] for Differential Attribution. Adapting the work of [159], we show the situations where the Shapley value and the Aumann-Shapley value coincide and argue that for Differential Attribution, their coincedation is rare which prompts us to motivate the use of the Aumann-Shapley value for Differential Attribution.

In this section we have shown with Example 6.5 how the Shapley value, through its discretisation of features, only attributes according to the behaviour of the function f evaluated at the endpoint (initial and final) values. In this way, it may not capture the true dependency of the function on a variable over the entire attribution region.

6.6 Path Values And The Aumann-Shapley Value

Originally, the study of value attribution in coalitional games centred around games with a finite and small number of players, i.e., 2 or 3 player games [16], which saw the introduction of "static" solution concepts like the Shapley value. Since the 1960s, however, there has been increased attention served to games with a larger number of players, such as voters in an election whereby no individual player can affect the overall outcome [16]. In these situations, it is conventional to consider these games as being played by a continuum of players. These games are termed "non-atomic" in the value attribution literature [16] and describe the type of game modelled by

the Differential Attribution challenge we introduced in Section 6.5. Owen [159] shows, via a multi-linear extension, how an n person game can be generalised to model a non-atomic game whereby each of the n players in the game v is replaced by a continuum of players.

The following section details this generalisation and introduces path values, which provide an attribution, for non-atomic games. Of these path values, we introduce and motivate the Aumann-Shapley value [15] for Differential Attribution. Furthermore, we show how the Aumann-Shapley value and the Shapley value coincide when the extension of the game v is multi-linear [159]. Relating this back to Example 6.4 and Example 6.5, we argue that for the Differential Attribution task, the assumption of multi-linearity does not often hold in practice and in this setting, we motivate the Aumann-Shapley value as an attribution method more in line with reality which fulfills both Differential Attribution Desiderata. Section 6.6 outlines the game theoretic background underpinning our novel contribution, which we motivate and formalise in Sections 6.7 and 6.9 respectively. Before we discuss the extension of games to a continuum of players we define a multi-linear function (Definition 6.1), upon which the argument of Owen [159] is constructed.

Definition 6.1 (Multi-linear Function [204])**.** A function $f : \mathscr{X} \to \mathbb{R}$ is multi-linear if we can write f as the following

$$f(r_1,...,r_n) = \sum_{I \subset N} c_I \prod_{i \in I} r_i. \tag{6.2}$$

As we recall from Chapter 4, a game in characteristic form is the tuple (N, v) with player set N and the set function v with the domain 2^N. If we consider 2 as the set $\{0,1\}$ then we can see that the domain of v is a subset of the unit n cube I^N where $I = [0,1]$, as shown in Figure 6.4. Discrete solution concepts, such as the Shapley value, can thus be seen as Static Attribution methods which operate over a subset of the unit n cube. Owen [159] extends the game characterised by v to the entire unit cube by introducing the function below.

$$f(x_1,...x_n) = \sum_{S \subset N} \{\prod_{j \in S} x_j \prod_{j \notin S} (1 - x_j)\} v(S), \tag{6.3}$$

for $0 \leq x_i \leq 1$, $i = 1,...,n$. We let α^S represent the value of each variable in the S corner of the cube such that $\alpha_i(S) = 1$ if $i \in S$ and $\alpha_i(S) = 0$ if $i \notin S$ [159]. Substituting, $\alpha_i(S)$ into Equation 6.3, we obtain the following value of f at the S corner of the cube.

$$f(\alpha^S) = \sum_{T \subset N} \{\prod_{j \in T} (\alpha_j^S) \prod_{j \notin T} (1 - \alpha_j^S)\} v(T). \tag{6.4}$$

From Equation 6.4, it is easy to see that $f(\alpha^S) = v(S)$ and that $\{\prod_{j \in T}(\alpha_j^S) \prod_{j \notin T}(1 - \alpha_j^S)\}$ will evaluate to zero for all $T \subset N$ bar $T = S$. Thus f is an extension of v, whose domain contains

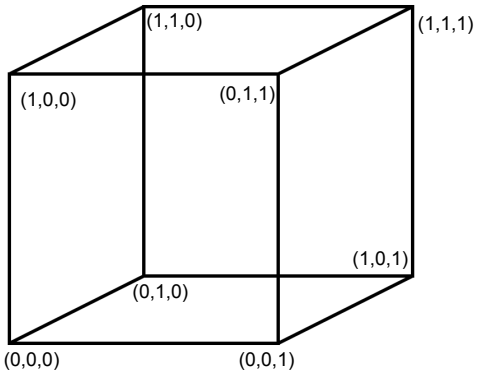


Figure 6.4: Figure shows the unit 3 cube which corresponds to the 3 player game.

the set of all corners of the unit cube, to the unit cube I^N. Since it is a multi-linear function (Definition 6.1) and the only multi-linear function which coincides with v at the corners of the cube α^S, it is termed the multi-linear extension of v [159]. If we think of x_i as the probability that player i will join a certain coalition, then we can consider the extension f as the expected value of the (potentially yet unformed) coalition [159].

In this section we have documented the non-atomic, multi-linear extension f of game v as proposed by Owen [159]. We have shown how games with binary variables can be extended to the continuous setting.

6.6.1 Path Attribution Methods

The path attribution methods are a natural class of attribution methods which assign to each variable its marginal effect along some path from the initial point to the final point of the unit n cube [204]. Given the multi-linear extension f [159], values for non-atomic games, or path values, can be obtained by integrating along the paths which start at the corner of the hypercube $\alpha^{\varnothing}$ and finish at the corner α^N [159]. If we let $\phi_i(t)$ for $i = 1,...,n$ be continuous, monotone functions with $\phi_i(0) = 0, \phi_i(1) = 1$ for all i, then the equations,

$$x_i = \phi_i(t), 0 \leq t \leq 1, \tag{6.5}$$

will represent a monotone path from the origin $(0,0,...,0,0)$ to the unit cube corner $(1,1,...,1,1)$ [159]. The path attribution methods assign to each player i the following value for given characteristic function f.

$$z_i = \int_0^1 f_i(\phi(t))d\phi_i(t). \tag{6.6}$$

Here f_i is the partial derivative of the function f in dimension i. In the special case that f corresponds to the multi-linear function f from Definition 6.1, Owen [159] shows that the partial derivative f_i can be expressed as the following.

$$f_i(x) = \sum_{S' \subset N; i \in S'} \{ \prod_{j \in S'; j \neq i} x_j \prod_{j \in S'} (1 - x_j) \} v(S') - \sum_{S \subset N; i \notin S} \{ \prod_{j \in S} x_j \sum_{j \notin S; j \neq i} (1 - x_j) \} v(S) \tag{6.7}$$

Letting $S' = \{S \cup \{i\}\}$, the above simplifies to

$$f_i(x) = \sum_{S \subset N; i \notin S} \{ \prod_{j \in S} x_j \prod_{j \notin S} (1 - x_j) v(S \cup \{i\}) - v(S) \}. \tag{6.8}$$

Here, the sum of the braces is 1. The partial derivative is thus a weighted average of the term $v(S \cup \{i\}) - v(S)$. It was shown by Owen [159] that the vector $(z_1, ..., z_n)$ from Equation 6.6 will always be an imputation, characterised by the path $\phi(t)$. An example path from the unit corner $(0,0,0)$ to $(1,1,1)$ is shown in Figure 6.5. Intuitively, if we consider players as continuous features, the path can represent the heterogeneous way in which features change from the initial to final values. The associated path value allocates to each type its average marginal contribution to the process along the production path. The selection of the path function $\phi(t)$ therefore characterises different values for non-atomic games. Below we introduce the Auman-Shapley value as an example of one such value.

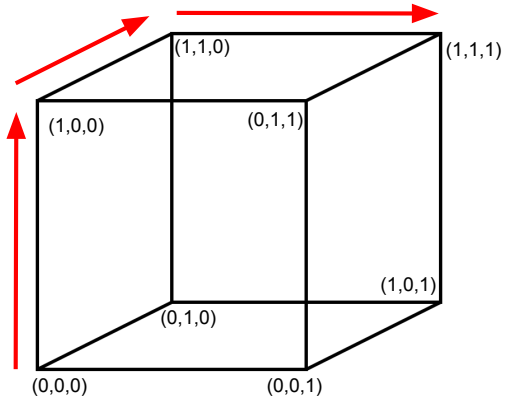


Figure 6.5: Figure shows an example path $\phi(t)$ which would correspond to the multi-linear extension of the 3 player game v whereby the path of production would see a sequential change in values of player 1 followed by player 2 followed by player 3.

In this section we have introduced the family of path values which assign value to players in non-atomic games. We have shown how for multi-linear functions, the partial derivative is a weighted average of the marginal contribution of each player to each coalition $S \subseteq N$. The selection of the path value encodes how we believe the variables move from initial to final values.

6.6.2 Integrating Along The Main Diagonal: The Aumann-Shapley Value

Given the path value defined in Equation 6.6, the Aumann-Shapley value is the unique path attribution method corresponding to the path function

$$\phi(t) = t \tag{6.9}$$

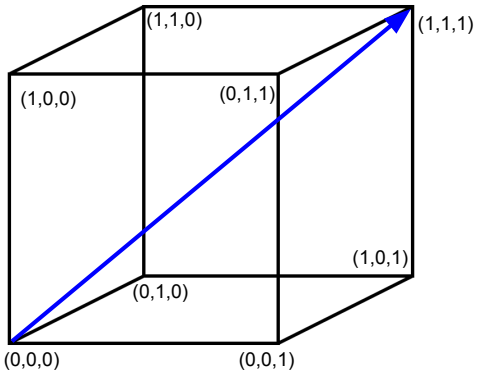


Figure 6.6: Figure shows the path of integration $\phi(t) = t$ which corresponds to the path of production used by the Aumann-Shapley value [15]

In words, the Aumann-Shapley value integrates the partial derivative along the straight line segment connecting the initial feature values to the final feature values i.e. the path of integration will be along the main diagonal as shown in Figure 6.6. Substituting $\phi(t) = t$ into the equation of the path values (Equation 6.6) renders Definition 6.2 [204].

Definition 6.2 (Aumann-Shapley Value [204])**.** For two samples $\mathbf{x}, \mathbf{x}'$ corresponding to initial and final feature values, the Aumann-Shapley value z_i is determined as the following attribution for variable $X_i \in \mathbf{X}$

$$z_i = (x_i' - x_i) \int_0^1 \frac{\delta f}{\delta X_i}(1-z)\mathbf{x} + z\mathbf{x}')dz \tag{6.10}$$

Here, z corresponds to the straight line segment between $\mathbf{x}$ and $\mathbf{x}'$.

Alternatively the Aumann-Shapley value can be seen as a three step process:

- 1) Calculate the partial derivative of the function f with respect to the variable under consideration X_i, $\frac{\delta f}{\delta X_i}$
- 2) Integrate that partial derivative along the line segment between $\mathbf{x}$ and $\mathbf{x}'$
- 3) Multiply the result by the change in that variable's value at the initial and final points $x_i' - x_i$

The integral given in Definition 6.2 is not guaranteed to have a closed-form solution. However, it can be computed numerically under Simpson's rule, for example [236]. We explore approximation methods further in Section 6.9.

In this section we have defined the Aumann-Shapley value as the unique path value for which the path of production corresponds to the main diagonal through the unit n cube.

6.6.3 Equivalence Of The Shapley Value And The Aumann-Shapley Value

Owen [159] showed that given the multi-linear extension f (Equation 6.3) of game v, the Shapley value and the Aumann-Shapley value coincide. To see this, we substitute the function $\phi(t) = t$ into Equation 6.8 which gives the following

$$f_i(t,t,...,t,t) = \sum_{S \subset N; i \notin S} t^{|S|}(1-t)^{n-|S|-1} v(S \cup \{i\}) - v(S), \tag{6.11}$$

and so,

$$z_i = \sum_{S \subset N; i \notin S} \frac{|S|!(n-|S|-1)!}{n!} v(S \cup \{i\}) - v(S) \tag{6.12}$$

This leads to the main result of Owen [159]: Given the multi-linear extension f of game v, if the path of integration is the main diagonal then z_i is the Shapley value. The equivalence of the Aumann-Shapley and the Shapley value requires that the function f is multi-linear. It was later shown by Sun et al. [204] that if the Aumann-Shapley and the Shapley value coincide for any function f then f must be multi-linear. In the section below we argue that within the context of Differential Attribution, we cannot assume the multi-linearity of the underlying function f, in which case we motivate the use of the Aumann-Shapley value to provide an attribution.

In this section we have presented the argument of Owen [159] whereby in the case of a multi-linear function f, the Shapley value and the Aumann-Shapley value coincide.

6.7 Motivating The Aumann-Shapley Value For Differential Attribution

The Shapley value under v_{bs} is the expected attribution of a monotone random walk along the edges of the hypercube with opposite vertices at $\mathbf{x}$ and $\mathbf{x}'$ [205]. If we subdivide the hypercube with opposite vertices at $\mathbf{x}$ and $\mathbf{x}'$ into a grid of smaller hypercubes – which is the intuition behind Owen's multi-linear extension [159] – and consider monotonic random walKernel SHAP in this structure, the density of the resulting walKernel SHAP will be focused on the diagonal. Hence, when the underlying function is multi-linear, the average of these path values will tend to the Aumann-Shapley method in the limit [205]. When, however, the function f is not multi-linear, the two values do not coincide. To see this, we note that the partial derivative $f_i(x)$ in Equation 6.7 relies on the linearity of each x_j. When the underlying function is not multi-linear, the partial derivative, calculated by the Aumann-Shapley value is not guaranteed to take the form of 6.7 and is thus not equivalent to the Shapley value.

Having seen the relationship between the Shapley value and the Aumann-Shapley value, we can now motivate the Aumann-Shapley value for Differential Attribution. Firstly we argue that it is unlikely that the function to be explained can be assumed to be multi-linear. We have seen throughout this thesis, how many deep learning models, particularly those employed for time series modelling, rely on non-linear operations. Indeed, both functions f, f' within Example 6.4 and f' in Example 6.5 are not multi-linear despite being relatively simple. In the case where we cannot assume multi-linearity of our black-box function we motivate the use of the Aumann-Shapley value over the Shapley value as the Aumann-Shapley value fulfills both Differential Attribution Desiderata outlined in Section 6.5.

By definition, the Aumann-Shapley value integrates along the diagonal corresponding to the progression of each feature through time such that $\phi(t) = t$. Furthermore, as the integration is taken over the entire attribution region such that if the partial derivative of the function is non-zero with regards to a particular variable, the attribution will reflect this dependence even if the values of the feature at the end-points do not. We now show, using Example 6.4 and Example 6.5, how the Aumann-Shapley value fulfills Differential Attribution Desiderata 1 and 2.

To calculate the Aumann-Shapley value for Example 6.4 we follow the approach detailed in Section 6.6.2 where the full calculations can be found in Appendix B.1. We obtain the Aumann-Shapley attribution for X_1, X_2, X_3 given $f(\mathbf{x})$ and $f(\mathbf{x}')$ in Example 6.4 as $\mathbf{z} = (0.92, 0.92, 2.16)$. As the Aumann-Shapley value is calculated via the partial derivative of the function with respect to each individual variable we can see why the Aumann-Shapley value is robust to the reformulation of the function f, f' from Example 6.4 as the partial derivative of X_3 is constant in both formulations. This essentially encapsulates the fact that the Aumann-Shapley value considers change in variables as they happen over time rather than consecutively. For the reformulated function $f'(\mathbf{X}') = 2^{2X_1}X_3$, we obtain the following Aumann-Shapley values for

X_1 and X_3, $\mathbf{z} = (1.84, 2.16)$. From this attribution, we can see that the Aumann-Shapley value, unlike the Shapley value, generates an attribution which is robust to the reformulated function f', due to the equal attribution of X_3 for both f and f', demonstrating empirically how the Aumann-Shapley value fulfills Desideratum 1.

To show how the Aumann-Shapley value satisfies Desideratum 2 we apply it to Example 6.5. Again for brevity, we omit the details of the calculations, leaving these to Appendix B.2. For the function f we obtain the Aumann-Shapley attribution vector $\mathbf{z} = (-0.5, -1, -1.5)$. We note that for this function which is multi-linear, the Aumann-Shapley value coincides with the Shapley value as shown in Section 6.6.3. However, when applying the Aumann-Shapley value to the modified function f', we obtain the following attribution vector $\mathbf{z} = (-\frac{2}{3}, -1, -\frac{4}{3})$ which, when compared to the attribution of f, has increased the attribution awarded to X_1 in accordance with its greater impact on the function as determined by its partial derivative. This attribution captures the behaviour of the function over the entire attribution region despite the same outcome at the endpoints. This empirically shows how the Aumann-Shapley value fulfills Desideratum 2 in the absence of a multi-linear underlying function.

In the section above we have have motivated the use of the Aumann-Shapley value for Differential Attribution by demonstrating how it fulfills Desideratum 1 and 2 for underlying functions which are not multi-linear. In the following section we discuss the axiomatisation of the Aumann-Shapley value.

In this section we have argued that within the context of Differential Attribution, we cannot usually make the assumption of multi-linearity and that in this case, the Aumann-Shapley value is better justified in comparison with the Shapley value.

6.7.1 Axiomatisation Of The Aumann-Shapley Value

The Aumann-Shapley value satisfies Efficiency, Dummy and Additivity which are three out of the four axioms which characterise the Shapley value (Definition 4.1.2). The Aumann-Shapley value also satisfies Scale Invariance [204] (Definition 6.3). Scale Invariance conveys the idea that attributions should be independent of the (possibly incomparable) units in which individual variables are measured. It is especially compelling in the context of Differential Attribution because the different variables may refer to quantities of entirely different things.

Definition 6.3 (Scale Invariance)**.** [204] Given two functions f, g,variable set $\mathbf{X} = \{X_1, ..., X_n\}$, initial and final samples $\mathbf{x}, \mathbf{x}'$, the attribution vector $\mathbf{z}$ satisfies Scale Invariance if the attributions are independent of linear re-scaling of individual variables. That is, for any $c > 0$, if $g(x_1, ..., x_n) = f(x_1, ..., \frac{x_j}{c}, ..., x_n)$, then for all i we have

$$z_i(\mathbf{x}, \mathbf{x}', f) = z_i((x_1, ..., cx_j, ..., x_n), (x'_1, ..., cx'_j, ..., x'_n), g). \tag{6.13}$$

Unlike the Shapley value, the Aumann-Shapley value does not satisfy Symmetry but instead a weaker notion of Symmetry, Type-Symmetry [79]. We now explore what distinguishes these two notions of symmetry and why the weaker axiom is preferential for Differential Attribution.

The Symmetry axiom requires an impartial treatment of players such that a relabelling leaves each player (under their new name) the same payoff as in the original game [79]. In reality, however, there are numerous situations in which the set of players is naturally partitioned into natural subsets (types) such that players can switch their identities only if they are of a certain type [79]. This is the case for continuous variables of a particular type which are not necessarily comparable, for example, temperature and blood-pressure. A weaker form of symmetry, type-restricted symmetry, requires that the payoffs of players *of the same type* do not depend on a labelling. If each type is infinite, as is the case with continuous variables, type-symmetric values are generated by path values as defined in Equation 6.6.

The Shapley value, under Symmetry, enforces that the arrival time of each player are independently and identically distributed [79]. In this sense, the Shapley value is a symmetric-path value, the same distribution is attached to each player and the outcome does not depend on a particular distribution. The Aumann-Shapley value, in contrast, is a type-symmetric path value such that only players of the same type do not depend on a labelling [79]. The weakening of the Symmetry axiom is what motivates the Aumann-Shapley value for Differential Attribution such that variables are not assumed to change from their initial to final values in the same way.

So far, we have introduced the Aumann-Shapley value as a value for non-atomic games and shown how it compares to the Shapley value. We have motivated the use of Aumann-Shapley value for Differential Attribution in the setting of a non multi-linear underlying function. One of the main requirements of the Aumann-Shapley value is that the underlying function f is differentiable. Within the context of post-hoc local explanations this is often not guaranteed. Below we discuss Integrated Gradients (IG), which is an existing Static Attribution method [171], also based on the Aumann-Shapley value. We show how IG overcomes the differentiable requirement and distinguish it from our use of the Aumann-Shapley value for Differential Attribution.

In this section we have given the axiomatisation of the Aumann-Shapley value, showing that it satisfies Efficiency, Additivity, Dummy and Scale Invariance. We motivated the axiom of Scale Invariance for Differential Attribution. We have also discussed how the Aumann-Shapley value satisfies a weaker form of symmetry: Type symmetry which we have motivated for Differential Attribution.

6.8 Integrated Gradients

In this section, we introduce Integrated Gradients [206] which is an existing Static Attribution method also based on the Aumann-Shapley value. Integrated Gradients (IG) is a widely used post-hoc local attribution tool for generating explanations on computer vision and deep learning task [206]. IG assumes an underlying function f which is continuous and almost everywhere differentiable. Then, given an instance to be explained $\mathbf{x}'$ and a baseline $\mathbf{x}$, IG is defined in Equation 6.14

$$IG(\mathbf{x},\mathbf{x}') = (x_i' - x_i)\int_0^1 \frac{\delta f}{\delta X_i}(\mathbf{x} + \alpha(\mathbf{x}' - \mathbf{x}))d\alpha. \tag{6.14}$$

From Equation 6.14, we can see that IG is exactly the Aumann-Shapley value as we have defined in Definition 6.2. However, the use of IG has mostly been restricted to image data and deep learning systems where the partial derivatives of the the machine learning model can be extracted. To our knowledge, IG has not been applied to Differential Attribution and more generally, not been applied to univariate time-series or multivariate time series models. Furthermore, IG enforces two requirements on the attribution:

Integrated Gradients, just like v_{bs} and in contrast to Differential Attribution is a method of Static Attribution such that it places emphasis on a single sample to be explained, $\mathbf{x}'$, requiring an "uninformative" baseline sample $\mathbf{x}$ to be used as a reference. IG then applies the Aumann-Shapley value to attribute the change in function output between the reference and the sample to be explained. We have previously explored in Chapter 4 and Chapter 5 how the choice of baseline impacts the resulting attribution. This dependence of an attribution on the selection of the baseline sample extends to IG where it was noted by Sturmfel et al. [201] that despite the common practice of selecting the baseline to be the vector of all zeros, these arbitrary feature values may have unintended meaning [201].

For example, if the model has learned that low values of blood pressure correspond to a high probability of death following sepsis onset then is a reference value of 0 a suitable "uninformative" feature value? A large part of this problem stems from the fact that under both the Shapley value and IG, the reference value is assumed to be an uninformative baseline. In contrast, when applying the Aumann-Shapley value to Differential Attribution we assign importance not only to the individual sample $\mathbf{x}'$ which represents the final feature values but equally to the reference sample $\mathbf{x}$ as we construct an attribution of the change in function output between the two samples $\Delta f = f(\mathbf{x}') - f(\mathbf{x})$. In other words, when applied to Differential Attribution, the Aumann-Shapley value requires no pre-specification of an uninformative baseline as this sample is already implicitly included in the investigative question of Differential Attribution.

In providing a post-hoc local explanation of the input sample $\mathbf{x}'$ in regards to it's model outcome $f(\mathbf{x}')$ IG require that the function f is continuously differentiable. Formally, this means the function f is continuous everywhere and the partial derivative of f along each input dimension

satisfies Lebesgue's integrability condition, i.e., the set of discontinuous points has measure zero. Deep networKernel SHAP built out of Sigmoids, ReLUs, and pooling operators satisfy this condition. This is a strict requirement of the attribution technique and IG is most widely applied to deep learning and computer vision applications [201]. The use of IG with tabular data, for example, requires smooth (differentiable) response surfaces, and hence they are not applicable to tree-based algorithms.

It has been noted by Dombrowski et al. [54] that even if the underlying function is continuously differentiable, the attributions provided by IG are sensitive to noise due to the large curvature of the output manifold of the underlying function [54]. Specifically, Dombrowski et al. [54] show that larger ReLu non-linearities in the network decrease the robustness of the resulting attributions. Our application of the Aumann-Shapley value to Differential Attribution (detailed in Section 6.9), in contrast to IG, does not assume the differentiability of the underlying function. Instead, we use a continuously differentiable local surrogate model. In this way, our attribution method is not only model-agnostic, it can be applied to tree-based algorithms, but it also avoids the undesirable behaviour of IG in the presence of functions which are differentiable yet contain discontinuities.

In the next section we present one of the main contributions of this thesis, our method for Differential Attribution termed Aumann Differential Surrogate Explanations (ADSE). Our method builds a continuously differentiable surrogate model around a multivariate time series such that given an initial and final sample $\mathbf{x}$ and $\mathbf{x}'$, and associated function output $f(\mathbf{x})$ and $f(\mathbf{x})$, ADSE generates a Differential Attribution, attributing the change in model outcome to each of the changes in feature values. We then apply ADSE to the two different Differential Attribution settings outlined in Section 6.5. The first is Temporal Explanations which describe the type of attribution we would require in Example 6.2 and the second is Multivariate Time Series Explanations, described by Example 6.1. In both settings our method ADSE, unlike Integrated Gradients, does not require a differentiable original model and does not require the selection of an "uninformative" baseline.

In this section we have introduced Integrated Gradients [206], a popular technique for post-hoc local explanations, also built using the Aumann-Shapley value. We have distinguished between our use of the Aumann-Shapley value for Differential Attribution, which we explicate in the following section, and the use of the Aumann-Shapley value in Integrated Gradients.

6.9 Aumann Differential Surrogate Explanations

In this section we formalise our method for Differential Attribution, Aumann Differential Surrogate Explanations (ADSE). For our Differential Attribution setting we assume that we have

access to a multivariate variate $\mathbf{X}$. Each individual sample in that set, $\mathbf{x}_i$, consists of n features recorded at t time-steps as given in Equation 6.15.

$$\mathbf{x}_i = \{\{x_1^1,...,x_1^t\}\{x_2^1,....,x_2^t\},...,\{x_n^1,...,x_n^t\}\} \tag{6.15}$$

For example, $\mathbf{X}$ may correspond to a set of patient trajectories where each $\mathbf{x}_i \in \mathbf{X}$ represents an individual patient's n vital sign measurements taken over t intervals.

As ADSE generates post-hoc explanations, we assume access to a black-box function $f : \mathscr{X} \to \mathbb{R}$ such that we have access to the model outcome evaluated on n features at each time-step $j \in$\{1,...,t\} such that $\mathbf{x}^j = \{x_1^j,...,x_n^j\}$ which gives following the outcome vector,

$$\mathbf{y}_i = \{f(\mathbf{x}^1),...,f(\mathbf{x}^t)\}. \tag{6.16}$$

As ADSE is a local explanation method, we take each individual sample (or patient trajectory) $\mathbf{x}_i$ as our local instance. Our Differential Attribution investigative question is thus of the form, given an initial sample $\mathbf{x} \in \mathbf{x}_i$ and final sample $\mathbf{x}' \in \mathbf{x}_i$, which features were most influential in the change in model outcome $f(\mathbf{x}') - f(\mathbf{x})$?.

To answer the above Differential Attribution question, we intend to apply the Aumann-Shapley value (Definition 6.2), using $\mathbf{x}$ and $\mathbf{x}'$ as our initial and final values respectively such that the Aumann-Shapley attribution for feature $X_i \in \{X_1,...,X_n\}$ is given as the following equation (which is taken from Definition 6.2).

$$\mathbf{z}(\mathbf{x},\mathbf{x}',i,f) = (x_i' - x_i)\int_0^1 \frac{\delta f}{\delta X_i}(1-z)\mathbf{x} + z\mathbf{x}')dz \tag{6.17}$$

However, as we have discussed in Section 6.8, the calculation of Equation 6.17 relies on the continuous differentiability of the function f, which, as we have discussed in Section 6.8 limits the applicability of the value. To avoid this limitation we therefore propose to approximate the underlying function f with a continuously differentiable surrogate model $\hat{f}$, such that Differential Attribution for feature $X_i \in \{X_1,...,X_n\}$ under ADSE is given as Equation

$$\mathbf{z}(\mathbf{x},\mathbf{x}',i,\hat{f}) = (x_i' - x_i)\int_0^1 \frac{\delta \hat{f}}{\delta X_i}(1-z)\mathbf{x} + z\mathbf{x}')dz \tag{6.18}$$

In the following section we discuss how we construct the continuously differentiable surrogate model $\hat{f}$ around our local sample $\mathbf{x}_i$.

6.9.1 Continuously Differentiable Surrogate Model

Having introduced and explored the concept of surrogate models in Chapter 3, we propose to overcome the differentiable requirement of the Aumann-Shapley value by building a local

surrogate model around the vector of multivariate observations $\mathbf{x}_i$ pertaining to an individual sample (or individual patient). Unlike the surrogate models we saw in Chapter 3, which are optimised for simplicity and faithfulness (Equation 3.1), our surrogate model is required to be continuously differentiable. This allows the computation of the partial derivatives in Equation 6.18. Given our local sample $\mathbf{x}_i$ as defined in Equation 6.15, and the associated vector of model outcomes $\mathbf{y}_i$ as defined in Equation 6.16 we construct the continuosly differentiable surrogate model $\hat{f}$ as the following optimisation,

$$\arg\min_{\hat{f}} \hat{f}(\mathbf{x}_i) - \mathbf{y}_i. \tag{6.19}$$

We relate the above minimisation to regression analysis. Regression analysis describes the problem, common to many disciplines, of adequately approximating a function of multiple variables, given only the value of the function evaluated at various points in the independent variable space. The system that generated the data is presumed to be described by

$$y = f(x_1, ..., x_n) + \epsilon \tag{6.20}$$

The single-valued deterministic function f captures the joint predictive relationship of y with $x_1, ..., x_n$. The additive stochastic component ϵ reflects the dependence of y on factors outside the system of interest. The goal of regression analysis is to use the data to construct a surrogate function $\hat{f}(x_1, ..., x_n)$ that can reasonably serve as an approximation of $f(x_1, ..., x_n)$ over the domain of interest n. Below we investigate what a reasonable function approximation, within the context of Differential Attribution, constitutes.

As we recall from Chapter 3, an optimal local surrogate model is one that is faithful to the original model and one that is simple. Unlike the local surrogate models discussed in Chapter 3, which are directly used as explanations themselves (i.e the coefficients of the linear model underpinning LIME), our local surrogate model is not used directly in explanation generation. Instead, we compromise on the interpretability constraint $\Omega(g)$ (Equation 3.1) to instead prioritise finding a local surrogate which is continuously differentiable. Furthermore, we have explicated the prevalence of non-multi-linearity under Differential Attribution. As such, we want our surrogate model to be able to capture non-linearities between variables. For our surrogate model $\hat{f}$ we use Multivariate Adaptive Regression Splines (MARS) [62] due to the following properties

Capable of modelling non-linearity of features MARS is a flexible regression technique that merges traditional linear regression with non-linear methods which results in a model capable of capturing complex relationships between variables. MARS is particularly useful when dealing with variables that have non-linear and interactive effects [62]. This property of MARS is particularly useful when we cannot assume multi-linearity of an underlying black-box function, which, as argued in Section 6.7, is common in the context of functions requiring Differential Attribution.

Continuous Differentiability: In the MARS algorithm, the approximating model is built from a collection of basis functions. These basis functions are created by splitting the input space at various cut-off points and estimating a linear relationship within each segment. The model then adapts the complexity of these basis functions by adding or removing cut-off points, allowing it to capture both linear and non-linear patterns in the data. Unlike prior basis-based approaches to regression analysis the approximating function constructed by MARS is continuously differentiable [62] which fulfills the requirement for our local surrogate model.

To construct our local surrogate model $\hat{f}$ we build a MARS model around our local sample, using $\mathbf{x}_i$ as our independent variables and $\mathbf{y}_i$ as our single dependent variable such that $\hat{f} = MARS(\mathbf{x}_i, \mathbf{y}_i)$. For our implementation of MARS we use the PyEarth Python library [30]. Using the PyEarth implementation, we first construct the model $\hat{f}$ as above and are then able to access the first partial derivatives of the function $\hat{f}$ with regards to each individual variable X_i.

Having outlined how we construct our local surrogate model $\hat{f}$ we combine our MARS model with Equation 6.18 in Algorithm 8. To numerically approximate the integral along the straight line segment as specified by Equation 6.18, we adopt the same approach as Sun et al. [204] whereby we sum the gradients at points occurring at sufficiently small intervals along the straight-line path from the initial observation $\mathbf{x}$ to the final observation $\mathbf{x}'$, replacing Equation 6.18 with Equation 6.21. We set $m = 100$ for each of our experiments which we document in the following section. Algorithm 8 details in full our Differential Attribution method, Aumann Differential Surrogate Explanations (ADSE).

$$\mathbf{z}'(\mathbf{x}, \mathbf{x}', i, \hat{f}) = (x_i' - x_i) \times \sum_{k=1}^{m} \frac{\delta \hat{f}}{\delta X_i}((1-z)\mathbf{x} + z\mathbf{x}') \times \frac{1}{m} \tag{6.21}$$

Algorithm 8 Aumann Differential Surrogate Explanations (ADSE)

Require: Individual sample $\mathbf{x}_i$, associated outcome vector $\mathbf{y}_i$
Require: Initial $\mathbf{x} \in \mathbf{x}_i$ and final $\mathbf{x}' \in \mathbf{x}_i$ values
  $\hat{f} \leftarrow MARS(\mathbf{x}_i, \mathbf{y}_i)$
  $\mathbf{w} \leftarrow \{\}$
  for $i \in \{1, ..., n\}$ **do**
    $w_i \leftarrow \mathbf{z}'(\mathbf{x}, \mathbf{x}', i, \hat{f})$
    $\mathbf{w} \leftarrow \mathbf{w} \cup w_i$
  end for
  Return $\mathbf{w}$

In this section we have introduced our method, Aumann Differential Surrogate Explanations for Differential Attribution

6.10 ADSE For Temporal Explanations

A Temporal Explanation is the kind of post-hoc explanation characterised by Example 6.2. We take as input a black-box function $f : \mathscr{X} \rightarrow \mathbb{R}$ which has been trained on a multivariate feature set $\mathbf{X} = \{X_1, ..., X_n\}$ consisting of n continuous features. For Temporal Explanations, as detailed in Section 6.5, we assume access to a vector of observations $\mathbf{x}_i$ which represent a number of observations pertaining to an individual i. For example, for the MIMIC Sepsis Cohort (Section 1.8) we would treat $\mathbf{x}_i$ as the collection of vital sign observations belonging to an individual patient recorded for t time-steps.

We also require an initial instance $\mathbf{x} = \{x_1, ..., x_n\}$ and a final instance $\mathbf{x}' = \{x'_1, ..., x'_n\}$ such that both $\mathbf{x}, \mathbf{x}' \in \mathbf{x}_i$. We assume that the sample $\mathbf{x}$ has pre-dated that of $\mathbf{x}'$.

The Differential Attribution problem within this setting is to determine the extent to which each feature is attributable in the change of prediction $\Delta f = f(\mathbf{x}') - f(\mathbf{x})$. From the above characterisation of the Temporal Explanation setting we can see that by treating an individual's patient trajectory $\mathbf{x}_i$ as our individual sample, we can obtain the prediction vector as $\mathbf{y}_i = \{f(\mathbf{x}^1, ...f(\mathbf{x}^t)\}$ and directly apply Algorithm 8 to obtain our Temporal Explanation for individual i given initial and final samples $\mathbf{x}$ and $\mathbf{x}'$.

Having shown how we apply ADSE to the Temporal Explanation setting from Example 6.2, below we apply ADSE to the MIMIC Sepsis Cohort as a real-world manifestaion of the Temporal Explanation setting. We experimentally compare ADSE with a Static Attribution method, Baseline SHAP [205].

6.10.1 Experimental Validation Of ADSE For Temporal Explanations

The motivation for Aumann Differential Surrogate Explanations applied to Temporal Explanations was originally inspired by the healthcare setting and specifically, the MIMIC Sepsis Cohort where the presence of individual patient time series inspired our first example in Section 6.5. Alternative settings which may require Temporal Explanations could include, for example, weather forecasting e.g why the weather forecast for a particular region changed overnight? or for individual recommender systems, why has the recommended movie for a particular user changed? However, in the following section we evaluate Aumann Differential Surrogate Explanations on the MIMIC sepsis cohort in Section 6.10.2.

6.10.2 MIMIC Sepsis Cohort

To evaluate ADSE on a real world setting suited to Temporal Explanations we compare the attributions of Baseline SHAP and ADSE on the MIMIC cohort case study. As we know from Section 1.8, the MIMIC Sepsis Cohort contains a collection of 19598 individual patients each associated with 51 observed variables over a varying number of time steps and a single outcome indicating whether that individual survived sepsis. For this experiment, we select individuals

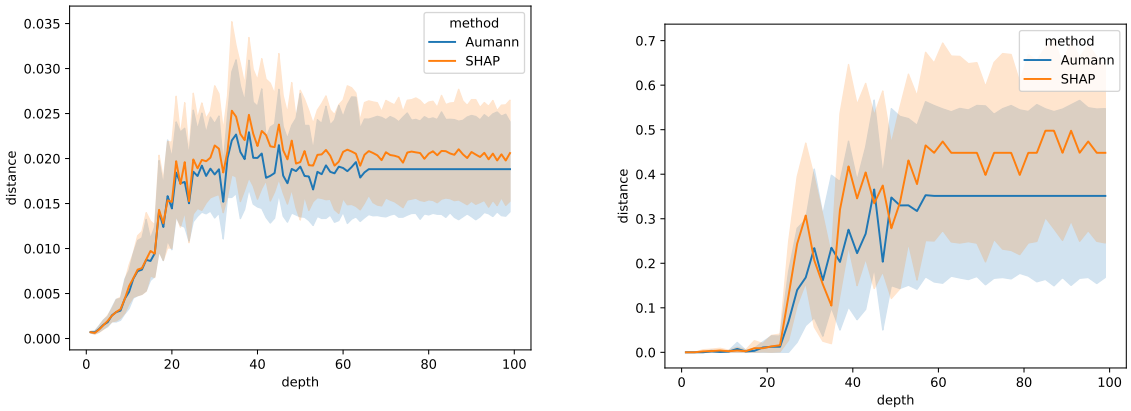


Figure 6.7: Figure shows the results of the Temporal Explanation experiment applied to the MIMIC Sepsis Cohort described in Section 6.10.2 Figures show AD distance for the Random Forest (left) and XGBoost (right) models as we increase maximum depth for the ADSE (blue) and Baseline SHAP (orange) attributions. Our results show that as we increase the number of interactions in the underlying classifier, the attributions afforded by ADSE method generate increasingly improved attribution than Baseline SHAP.

with a substantial number (20) of observed feature recordings, which results in a set of three thousand individuals each with 20 samples of 51 observed variables and one binary outcome variable. We select 33 of the features to train our sepsis model which discounts discontinuous features such as age.

For our Temporal Explanation setting we train the model using each feature recording as a single sample such that $f : \mathbb{R}^{33} \to \mathbb{R}$, we use the final outcome variable for the individual to whom each sample belongs as a label. We split the data into a train set of 2700 and a test size of 300. For this experiment, we wish to compare the attributions generated by Baseline SHAP and ADSE under varying feature interaction in the model. We therefore select a Random Forest Classifier (RFC) and an XGBoost classifier due to their ability to be parametrised with an increasing number of feature interactions using the depth parameter.

We use the Scikit-learn implementation of the RFC [165] with 100 estimators and Gini a splitting criteria. For our XGBoost model, we use the Scikit-learn implementation [165] with 100 estimators, log loss function and a learning rate of 0.1. For both models we train 100 instances of the classifier letting the max depth to range from 1 to 100. The greater the max depth parameter, the greater level of the feature interaction in the model [26]. Under each trained model we obtain ADSE and Baseline SHAP attributions for two randomly selected samples as the initial and final samples from each of the 300 test individuals. Our evaluation metric is Average Deletion as motivated in Section 4.7 and defined in Equation 4.34. Applied to our Temporal Explanation setting we measure the Absolute distance between the initial sample and the perturbed final sample once the most important feature, as returned by the attribution

method has been perturbed.

Results are shown in Figure 6.7 which show how, for both classifiers, the difference in Average Deletion scores between Baseline SHAP and ADSE increase as the level of interaction in the model increases. For both models, when the maximum depth parameter is set below 20, the two attribution techniques appear to coincide. As this parameter is increased above 20 however, the Average Deletion scores of both attribution methods diverge and Differential Attribution outperforms Baseline SHAP for a maximum depth of above 50 in both models. This result confirms our theoretical discussion in Section 6.7 where, under an underlying multi-linear function, ADSE and Baseline SHAP converge.

Despite the fact that ADSE approximates the underlying model for attribution where Baseline SHAP is computed directly on the classifier model, we can still expect to observe the convergence of both attributions for increasingly simpler underlying functions. As machine learning models become increasingly complex, being able to supply a surrogate which is capable of capturing the locally non-linear behaviour of the underlying model is motivated. Through our empirical analysis on the MIMIC sepsis cohort we have shown the utility of our attributions in the presence of an increasingly non-linear underlying model.

In this section we have shown how to apply ADSE to generate Temporal Explanations. We have compared the attributions generated by ADSE with those of Baseline SHAP in generating Temporal Explanations for the MIMIC Sepsis Cohort. ADSE results in attributions with a lower Average Deletion score than those of Baseline SHAP, particularly for models with a high level of feature interaction.

6.11 ADSE For Multivariate Time Series Explanations

In this section we apply ADSE to the second Differential Attribution setting which was motivated in Example 6.1, Multivariate Time Series Explanations. In this setting, we now have a predictor function f which takes as input a multivariate variable set $\mathbf{X}$ where each sample $\mathbf{x}_i$ (Equation 6.15) is composed of n predictor variables, each recorded for t observations. However, under the Multivariate Time Series Explanation setting, we now have a black-box function $f : R^{n \times t} \rightarrow \mathbb{R}$ which is trained on multivariate time-series.

Given an individual's multivariate time series $\mathbf{x}_i$ and its associated prediction $f(\mathbf{x}_i)$, the clinician wishes to understand which features and at which temporal locations were most important for the resulting prediction.

One way of generating the above explanations for multivariate time series classifiers is to use the existing SHAP mechanism, treating each individual observation $\{x_i^j\}$ for $i \in \{1, ..., n\}$ and $j \in \{1, ..., t\}$ as an individual player in the game. However, we have discussed in Section 6.5 how

this causes spurious attributions. We therefore motivate multivariate time series explanations as requiring Differential Attribution which takes into account the temporal dependence between observations in a multivariate time series. What distinguishes the Multivariate Time Series setting from Temporal Explanations is the fact that now, not only do we approximate the black-box function f with our continuously differentiable MARS model $\hat{f}$ but now, we also do not have access to the outcome vector $\mathbf{y}_i = \{\mathbf{x}^1, ..., \mathbf{x}^t\}$ as the underlying function f is trained on multivariate time series of dimension n and length t. The section below shows how we overcome this restriction such that given an individual multivariate time series we want to be explained $\mathbf{x}_i$, we approximate the outcome vector $\mathbf{y}_i$ in order to apply Algorithm 8.

Given an individual multivariate time series to be explained $\mathbf{x}_i$, to approximate the function f's output at each time-step $j \in \{1, .., t\}$ we adopt a perturbation strategy such that the approximated function output at time $j \in \{1, ..., t\}$ is evaluated on a sample which includes only observations up until that time-step. We remove observations from each of the multivariate samples by replacing them with the mean value for that feature such that

$$\mathbf{x}^j_{app} = \{x^0_0, ..., x^j_0, {x'_0}^{j+1}, ..., {x'_0}^t\}\{x^0_1, ..., x^j_1, {x'_1}^{j+1}, ..., {x'_1}^t\}\{x^0_n, ..., x^j_n, {x'_n}^{j+1}, ..., {x'_n}^t\}$$

Here, each ${x'_i}^j = \mathbb{E}[\{x^0_i, ..., x^t_i\}]$. This approximation allows us to obtain the underlying function's prediction on intermediate time steps, i.e. for all time steps $j \in \{1, ..., t\}$ such that $\mathbf{y}'_i = \{f(\mathbf{x}^0_{app}), ..., f(\mathbf{x}^t_{app})\}$. Having obtained the function output for each time-step and collected these as our approximated outcome vector $\mathbf{y}'_i$ we apply ADSE (Algorithm 8) using $\mathbf{x}_i$ as our individual sample and our approximated outcome vector, $\mathbf{y}'_i$, as input. We can then select an initial $\mathbf{x}$ and $\mathbf{x}'$ and final sample from $\mathbf{x}_i$ to answer the Differential Attribution question of which feature was most influential for the change in function output between the two selected time-steps.

For multivariate time series explanations, rather than selecting the feature values at two random time-steps as the initial and final samples $\mathbf{x}, \mathbf{x}'$, it may be useful to construct a Differential Attribution vector which ranges the entire time-step interval $\{1, ..., t\}$, or in words, we may wish to understand how the function depends on each feature across the entire temporal interval spanned by the dataset. The following section shows how we use ADSE to construct this kind of explanation.

In this section we have shown how we can apply ADSE to generate Multivariate Time Series Explanations by approximating the outcome vector $\mathbf{y}_i$.

6.11.1 Multivariate Time Series Attribution: How Attribution Varies Over Time

For an attribution which considers how the underlying function depends on each feature over time we can determine the attribution vector for each variable X_i for all $i \in \{1, ..., n\}$ by calling

Algorithm 8) for each feature $i \in \{1,...,n\}$ *and* for each time-step $j \in \{1,...,t\}$. As a result, each w_i^j (as returned by Algorithm 8) determines the attribution to feature X_i between the time-step j and $j' = j+1$ for all $j \in \{1,...,t\}$ such that $\mathbf{x} = \mathbf{x}^j$ and $\mathbf{x}' = \mathbf{x}^{j'}$. We collect each ADSE attribution for each feature at each time step as our attribution vector $\mathbf{w}$.

In this section we have shown how we can obtain an attribution vector for a multivariate time series $\mathbf{x}_i$ by collecting the vector comprised of ADSE applied to $\mathbf{x}^j$ and $\mathbf{x}^{j+1}$ for each $j \in \{1,...,t\}$

6.12 Experimental Validation For Multivariate Time Series Explanations

In this section we conduct experiments comparing ADSE with state-of-the-art methods for Multivariate Time Series Explanations. We select a variety of multivariate time series datasets (detailed in Section 6.12.1). For each of our datasets, we train a State RNN classifier (detailed in Section 6.12.2) on the train set, achieving a classification accuracy of 0.97 on NATOPS, 0.96 on RacketSports, 0.93 on Epilepsy, 1.0 on Basic Motions, 0.89 on Sepsis. For each of the individual multivariate time series in the test set we generate the vector of ADSE as per the method in Section 6.11 and the related multivariate attributions for each of the benchmark attribution methods (Section 6.12.2). We evaluate each of the attribution methods using Accuracy metric described in Section 6.12.3. We run the experiment five times which generates a mean Accuracy score and error bands for each attribution method as reported in Figure 6.8.

6.12.1 The Datasets

NATOPS: This dataset, taken from the UCR archive [35] is a gesture recognition task. The dataset is comprised of a train and test set each with 180 individual multivariate time series comprised of 24 features and length 51. The data is generated by sensors on the hands, elbows, wrists and thumbs. The data are the x,y,z coordinates for each of the eight body locations. There are six classes corresponding to distinct separate actions.

RacketSports: This dataset, taken from the UCR archive [35] is a classification task to identify which sport and action a player is making from accellerometer data. The dataset is comprised of a train set of size 151 and test set of size 152 individual multivariate time series comprised of 6 features and length 30. The data order is accelerometer x, y, z then gyroscope x, y, z. There are 4 classes and the challenge is to predict whether a player played either a forehand/backhand in squash or a clear/smash in badminton.

Epilepsy: This dataset, taken from the UCR archive [35] is a classification task to distinguish epileptic activity from other activities (walking, running and sawing) from accellerometer

data.The dataset is comprised of a train set of size 137 and test set of size 138 individual multivariate time series comprised of three features and length 207. The data order is accelerometer x, y, z.

BasicMotions: This dataset, taken from the UCR archive [35] is a classification task to distinguish walking, resting, running and badminton activity from accelerometer data. The dataset is comprised of a train set of size 40 and test set of size 40 individual multivariate time series comprised of six features and length 100. The data order is accelerometer x, y, z and gyroscope x,y,z.

Sepsis: This dataset, taken from the MIMIC Sepsis Cohort Section 1.8, is a classification task to identify whether a patient survives or dies from sepsis ggiven 48 hours of vital sign observations from sepsis onset. The dataset is comprised of a train set of size 1034 and test set of size 300 individual multivariate time series comprised of 33 features and length 15.

6.12.2 The Benchmarks

We choose to compare our multivariate attribution on state-of-the-art gradient SHAP-based methods (which require a differentiable underlying model) and a state-of-the-art non-Shapley-based method Dynamask.

Integrated Gradients: Integrated Gradients [171] (IG) is described in detail in Section 6.8. We use the default implementation of Integrated Gradients from the Captum Pytorch library [111], with a baseline of $E[\mathbf{x_i}]$ where $\mathbf{x}_i$ indicates feature i of the multivariate time series sample to be explained.

DeepLift: DeepLift [189] (DL) explains the difference in output from some 'reference' output in terms of the difference of the input from some 'reference' input. In contrast to to Shapley-based approaches DeepLIFT propagates an importance signal from an output neuron backwards through the layers of a neural network to the input layer which is the signal transformed into an attribution vector. We use the default implementation of DeepLift from the Captum Pytorch library [111], with a baseline of $E[\mathbf{x_i}]$ where $\mathbf{x}_i$ indicates feature i of the multivariate time series sample to be explained.

Gradient SHAP: Gradient SHAP [138] (GS) approximates SHAP values by computing the expectations of gradients by randomly sampling from the distribution of baselines/references. It selects a random baseline from baselines' distribution and a random point along the path between the baseline and the input, and computes the gradient of outputs with respect to those selected random points. We use the default implementation of Gradient SHAP from the Captum Pytorch library [111], with a baseline of $E[\mathbf{x_i}]$ where $\mathbf{x}_i$ indicates feature i of the multivariate time series sample to be explained.

Dynamask: Unlike existing perturbation approaches, Dynamask [39] (MASK) generates multivariate attribution for each feature at each time step by fitting a perturbation mask to the original multivariate time series. To account for time dependency in the data structure,

Dynamsk incorporates a dynamic mask to the attribution method. We use the original authors' implementation of Dynamask [39] with a baseline of $E[\mathbf{x_i}]$ where $\mathbf{x}_i$ indicates feature i of the multivariate time series sample to be explained.

STATE Model We use the State Recurrent Neural Network of [39] with 200 hidden GRU cells as our multivariate time series classifier. We adopt a learning rate of 0.1, use the Adam Optimiser and train from 200 epochs for each of the classification tasKernel SHAP described above.

6.12.3 The Metric

To evaluate each of the benchmark datasets we propose an alternative metric to Average Deletion which has been used throughout this thesis. We choose this measure, Accuracy (Equation 6.22) rather than Average Deletion, which is designed to measure the effect of a single most influential feature, as in the presence of a large number of influential features, Average Deletion may not accurately reflect the ability of the attribution technique at identifying the most influential observation.

For each test sample, we remove each individual x_i^j (by replacing them with the relevant $x_i^{j\prime}$ value as defined by $\mathbf{X}_{app}^j$) in the order as determined by the attribution technique in question. For each feature that we perturb, we measure the average predictive accuracy over the test set as a measure of how faithful the attribution technique is, a greater reduction in predictive accuracy over the entire test set indicates a more faithful attribution technique.

$$Acc(i) = \frac{f(\mathbf{X}_p^i) - \mathbf{y}}{|\mathbf{y}|} \tag{6.22}$$

Here, $i \in \{1, ..., txn\}$ indicates the index of the vector of sorted features by importance as indicated by the attribution technique under-question. $\mathbf{X}_p^i$ therefore refers to the entire test set of multivariate time series after they have had their i-th most important feature perturbed. $\mathbf{y}$ indicates the ground truth labels of the test set. $Acc(i)$ therefore assesses the accuracy of the model when evaluated on the perturbed test set, capturing the intuition that the more influential a feature is, the greater the drop in accuracy of the classifier when it is removed from the instance.

Figure 6.8 shows the results of the experiment described above. An optimal attribution method would correctly recognise the most salient features and time-steps which, when removed would result in a sharp degradation of model accuracy. For the Epilepsy dataset we can see that the Aumann Differential Surrogate Explainer outperforms all other attribution methods. However, the results on the other datasets evaluated are more nuanced. On the NATOPS dataset, for example, we can see that although our Aumann Differential Surrogate Explainer presents the sharpest decrease in model accuracy for the removal of the first 800 observations, it is overtaken by other methods including Gradient SHAP and Deep Lift for proceeding observations. This behaviour is also observed for the RacketSports and the Sepsis datasets.

Interpreting these results we could argue that while our Aumann Differential Surrogate Explainer consistently outperforms other metrics in identifying the most salient features and associated timesteps, as the saliency of observations decreases, our method struggles to distinguish more and less influential observations. This could be due to limitations of the approximating surrogate model, or the approximated intermediate function outputs. However, we leave a more rigorous evaluation of this behaviour to future work.

Comparing the settings of the BasicMotions classification to the other datasets evaluated, we note that the range in $Acc(i)$ for all the attribution methods on the BasicMotions dataset represented by the shaded regions is significantly greater than those on all other datasets indicating that there is high variability over the multiple instantations of the experiment. This occurs due to the high variation in the predictive accuracy of the state model. From this, we can observe that in a setting where the underlying model is not accurate, the resulting attribution is also likely to display high variance. In this case, the differential multivariate attribution which undergoes a further layer of abstraction via the local surrogate mars model results in an average Acc score which is outperformed by the MASK method.

6.13 Differential Attribution: Concluding Remarks

In this chapter we have applied both attribution methods from Chapter 4 and Chapter 5 to univariate time series. We have motivated both Shapley Sets and Gately Feature Attribution for these high-dimensional data structures. However, we have also argued, as we did in Chapter 3, about the dangers of applying existing post-hoc feature attribution methods which were not designed specifically for time series to this data structure. Furthermore, we have considered the problem of Differential Attribution and proposed a method, Aumann Differential Surrogate Explanations (ADSE), based on the Aumann-Shapley value, which specifically attributes in the context of time varying features and function output. We have experimentally compared ADSE with existing approaches to multivariate attribution and demonstrated the ability of ADSE at recognising salient regions of multivariate time-series.

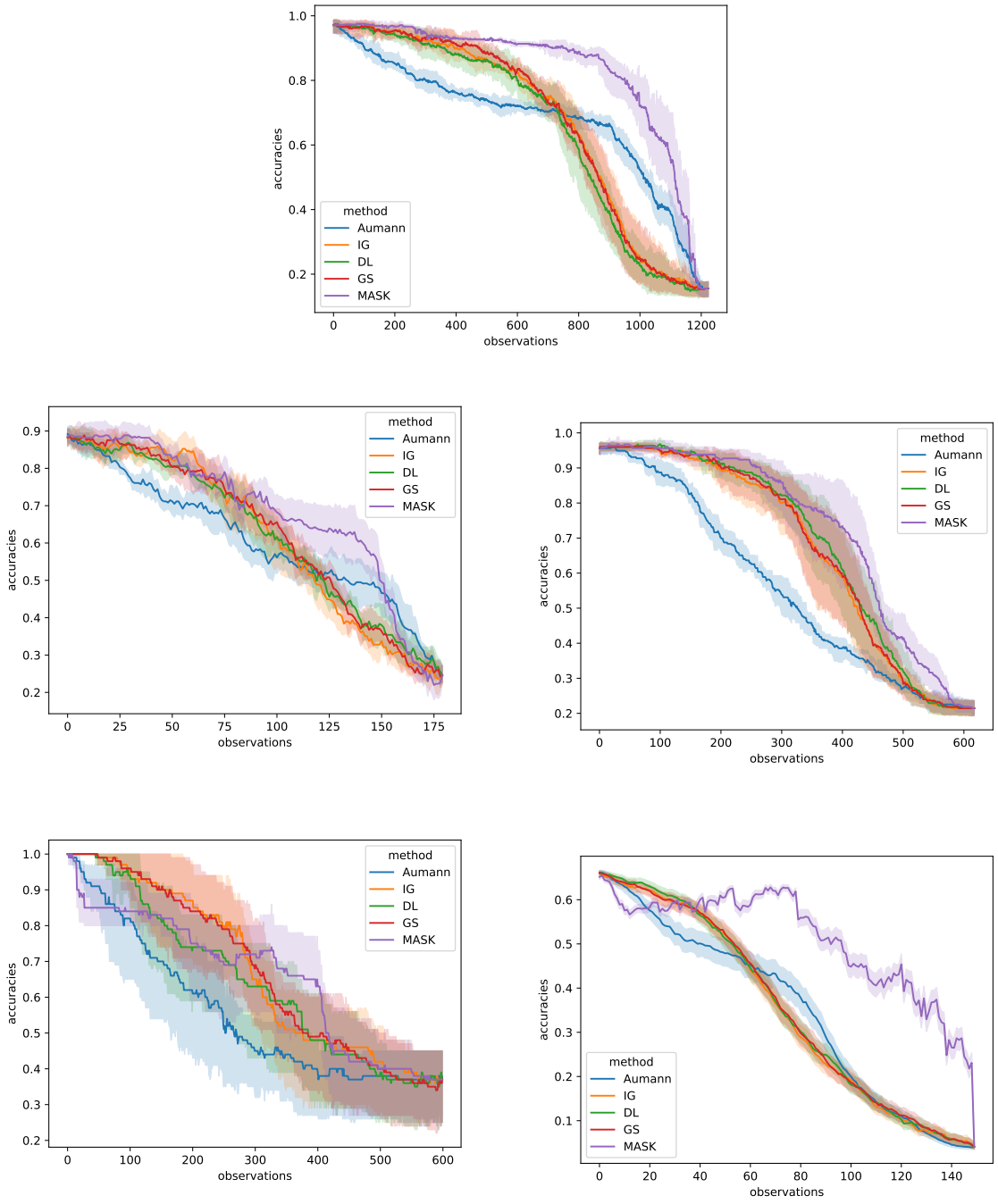


Figure 6.8: Figure shows the results of the multivariate attribution experiments as discussed in Section 6.12. The Figure records the mean model accuracy (Equation 6.22) and shaded variance (across experiment iterations) for all test samples evaluated as each observation $i \in \{1, ..., n \times t\}$ is removed from the sample as indicated by the attribution method. A curve where accuracy degrades sharply as observations are removed reflects an optimal attribution method. Figure shows the results for the NATOPS dataset (top), RacketSports (middle left), Epilepsy (middle right), BasicMotions (bottom left), Sepsis (bottom right). The figure records the average accuracy measured for the attributions of Aumann Differential Surrogate Explanations against the benchmarks.

CHAPTER 7

SUMMARY, CONCLUSIONS AND FUTURE WORK

In this thesis we have explored the diverse research landscape which underpins post-hoc local explanations in AI systems. We have focused on two of the most popular approaches, LIME and SHAP which, through an analysis of their limitations, presented four novel methods for post-hoc local explanations. However, aside from these methodological contributions to the feature attribution literature, this thesis has attempted to unify post-hoc local explainability with the rich multi-disciplinary background which underpins the explanation sciences. Through our extensive discussion, ranging from the fairness principles encoded by game-theoretic solution concepts to the philosophical perspectives underpinning the concepts of counterfactual and bifactual explanations, we hope that we have motivated the value of grounding XAI within a multi-disciplinary context, not only as a way of providing novel ways by which to explain AI systems, but also as a way understand better what we encode in an explanation method when it is applied from another discipline.

Our adaptation of LIME in Chapter 3 to univariate time series outlined the central challenges in adapting post-hoc local explanations for this data structure, focusing on the three open challenges: how to conceptualise, how to perturb and how to define a neighbourhood. Through our attribution method LIMESegment, we introduced a novel way of addressing these challenges, resulting in a modular algorithm which, we hope, will pave the way for future explainability mechanisms which are designed specifically for time series.

In Chapter 4, we turned our attention to the Shapley value, first contextualising it within its game-theoretic origins and aligning it, among other solution concepts, with the principles of fairness each attribution method encodes. We continued to show the limitations of the Shapley value when applied to feature attribution. Particularly, we unified the Shapley value with the concept of additively-separable functions and showed how, when features interact, either in

the data or in the model, attribution via the Shapley value generates misleading explanations. We introduced our algorithm, Shapley Sets, as a mechanism which automatically identifies the non-additively separable groups of variables given a black-box model, resulting in more robust attributions in the presence of interacting features than Shapley value alternatives.

In Chapter 5 we continued our analysis of the Shapley value, now from a causal perspective. We explored different types of causal question and differentiated between singular and general Causal Effects and mapped these ideas to those of counterfactual, bifactuals and probabilitstic causality. We showed how the Shapley value corresponds to the Probability of Necessity and Sufficiency, a general perspective of causality. We motivated the utility of a bifactual explanation with regards to the individual right to an explanation as outlined by Wachter et al. [216] and defined the Bifactual Effect of a player on a coalitional game. We then motivated the use of the Gately value as a solution concept which derives an attribution which is proportional to the Bifactual Effect of a player and is axiomatically justified. We mapped these ideas to the feature attribution space, showing how different value functions correspond to different levels of causal question and motivated the use of our proposed method, Gately Feature Attribution, under the off-manifold value function v_{bs}.

In Chapter 6, we returned to our discussion in Chapter 3 regarding the difficulties in explaining time series data, instead focusing on the limitations of the Shapley value on this data-structure. We showed how both Gately Feature Attribution and Shapley Sets address some of these problems for this data-structure. We then turned our attention to the open challenge of explaining multivariate time series data and defined the challenge of Differential Attribution. We showed how, and when, we cannot assume multi-linearity of the underlying black-box model, the Aumann-Shapley value rather than the Shapley value offers attributions which are more in line with how continuous features evolve over time. We finally presented our novel Differential Attribution method, Aumann Differential Surrogate Explanations in two settings, for Temporal Explanations and for Multivariate Time Series Explanations.

Despite the theoretical and conceptual focus of this thesis, we have approached the problem of developing post-hoc explanations for time series from a healthcare perspective. In Chapter 2, we acknowledge that XAI approaches should not be considered without a consideration of how they will fare in practice and as such we have drawn inspiration from digital healthcare, using the MIMIC sepsis cohort as the thread unifying our approaches and grounding our methodological approach in reality. The rest of this chapter first outlines future work for each of the aforementioned explanation approaches comprising the last four chapters. We conclude with a discussion of some exciting avenues for the future of XAI as a research discipline.

7.1 LIMESegment: Future Work

In Chapter 3, we presented our adaptation of LIME [174] for univariate time series explanations, LIMESegment. One of the limitations of our approach is the need for parametrisation of each of the comprising algorithms, NNSegment (Algorithm 2), RBP (Algorithm 4). In particular for NNSegment, the user must specify the total number of change points, which corresponds to the total number of time series super-segments they want the algorithm to find. As such, we require the user to have some intuition as to how, roughly, the time series they are explaining can be decomposed into homogeneous regions of activity. As we have discussed in Section 3.7, LIMESegment is sensitive to the length of time series and resulting number of super-segments. An appropriate parametrisation of the NNSegment algorithm is therefore important to ensure reliable attributions. A further limitation of NNSegment is that we assume that homogeneous regions of activity can be well characterised by the adjacency property (Definition 3.2), while this may apply to some types of time series there may exist some sorts of time series where similar segments do not follow adjacency, particularly if the window size parameter (Algorithm 2) is not appropriately specified.

In the future, we will therefore explore alternative segmentation approaches. For example, Kidger et al. [105] propose a time series segmentation algorithm which learns an appropriate metric of similarity between segments. A further assumption made in Chapter 3, is that a segmentation which operates under the temporal coherency assumption is more intuitive to a human than one which operates under the frequency coherence assumption. While this is motivated, in the future we intend to explore what a decomposition based on the frequency domain would look like.

7.2 Shapley Sets: Future Work

In Chapter 4 we introduced our novel method, Shapley Sets to determine automatically grouped attributions. We showed theoretically and empirically how Shapley Sets generates more robust explanations in the presence of interacting features compared to Shapley value based alternatives. However, the Shapley Sets algorithm requires parametrisation of the variable ϵ, Algorithm 4.5.1, which determines the degree to which two sets of variables are considered to interact. The original RDG algorithm recommends the setting ϵ as proportional to the magnitude of the objective space. While this setting works well for Shapley Sets under v_{marg} as its design is similar to the original RDG fitness measure, we noticed a large variation in the variable grouping generated by Shapley Sets under v_{cond} and this setting of ϵ. This is not surprising as it is known that v_{cond} is sensitive to feature correlations in the data and it is difficult to know how much interaction to allow before two features are considered to be causally dependent. Future work should therefore look at alternative methods of function decomposition which are robust to the parametrisation of ϵ [34].

One of the main assumptions of Shapley Sets assumes is that of partially separability of the

underlyinf function. If we consider the function $f(\mathbf{X}) = X_1X_2X_3$, Shapley Sets would result in a single attribution to all three features of $f(\mathbf{x})$. This is not useful from an explanation perspective although does inform us about the nature of the underlying model. However, this assumption is also made by the Shapley value. Under the marginal value function where $E[X_1, X_2, X_3] = 0$, the Shapley value for the above example would be $\frac{1}{3}f(\mathbf{x})$ regardless of the magnitude of each of the variables which is not a true account of the situations at hand. Future work should consider function decomposition under a wider class of separability such as multiplicative separability where associated algorithms decompose a function into its additive and multiplicative separable variable sets [34].

7.3 Gately Feature Attribution: Future Work

In Chapter 5, we defined the Gately value as the asolution concept satisfying Efficiency, v-Compromise, and Restricted Proportionality for the **regular** game $v \in G_n$ where G_n is the class of all n player games (Definition 4.9). Unlike the Shapley value which can be applied to any game $v \in G^n$, the restriction of games for which the Gately value satisfies the above axioms to those which are regular have implications on the kind of attribution problem the Gately value can be applied to. Staudacher et al. [199] showed that when the propensity to disrupt $d* = -1$ (Equation4.7), then the Gately value is undefined. The conditions equating to $d* = -1$ occur in weakly-constant sum games, or in other words in certain games where the value of coalitions $v(S)$ $S \subseteq N$ is greater than the value of the grand coalition $v(N)$. In our attribution of the Gately value for feature attribution Definition 5.18, we account for this edge case to return a "non identifable" explanation in the setting detailed above. However, we leave it to future work to explore the extent to which $d* = -1$ occurs in natural setting for feature attribution.

There has been considerable work in the game-theoretic literature, connecting classes-of games with approximation methods of the Shapley value. Liben et al. [128], for example, show how the Shapley value can be approximated in polynomial time for super-modular games. This body of work, however, has largely been under-studied within the context of feature attribution where most of the approximation methods of the Shapley value are based on sampling strategies. We have motivated throughout this thesis how connecting feature attribution with other disciplines can yield alternative methods for feature attribution which answer different investigative goals. We therefore motivate te future study, connecting feature attribution to classes of games, e.g monotonic or super-modular games, to develop alternative methods of approximating the Shapley value.

7.4 Aumann Differential Surrogate Explanations: Future Work

In Chapter 6 we motivated the use of the Aumann-Shapley value for the task of Differential Attribution, showing that in the presence of a non-multi-linear underlying function, the attributions

afforded by the Aumann-Shapley value are a better representation of reality than those of the Shapley value. As shown in Definition 6.2, the Aumann-Shapley value involves an integration of the partial derivative along the straight line segment between the initial and end points.

When using the Aumann-Shapley value for Differential Attribution, where the initial and final endpoints correspond to initial and final values of features $\mathbf{x}, \mathbf{x}'$, we must ensure that representing the change in feature values as a straight line makes sense. This requirement translates to ensuring that our initial and final feature values $\mathbf{x}, \mathbf{x}'$ are close enough together such that the change in each variable can be approximated linearly. For example if we consider a single variable in the multivariate feature set of the sepsis cohort, temperature, for example. If we know that in reality, over the course of an individual patient's sepsis trajectory, temperature spikes and then falls. If their initial and final values are thus 36.5 and 36.2 but an intermediary reading is of blood pressure 37.3, an attribution taken at the endpoints would not account for this intermediary spike in value. If the underlying black-box had recognised that temperature was very influential in the probability of death from sepsis, a better attribution would take into account that the patient's temperature first increased and then decreased. We would therefore motivate determining two sets of attributions, one from the initial and intermediary value and then from the intermediary to the final values.

The above is a manifestation of the more general challenge of using discrete time series to represent continuous phenomena which is ubiquitous across time series modelling, not just in our setting of Differential Attribution whereby, time series datasets are subject to the assumption that recorded values of the variable are a true representation of how it progresses over time. A direction of future work would therefore be to explore how impactful our linearity assumption is on attribution under various kinds of multivariate time series.

7.5 Explaining The Model vs. Explaining The Data

In Chapter 1 we explored how XAI arose as a research discipline largely in response to the proliferation of machine bias associated with black-box AI systems. As a result, post-hoc explanations have dominated the research landscape thus far which is why they have been our focus. As we have shown throughout this thesis, existing methods of generating post-hoc explanations are subject to limitations. One of the key ideas we explored in Chapter 4 was whether we, in order to adequately understand a model output, need explanations of the model or of the data?

This is a question which lies at the heart of many criticisms of post-hoc local explanations. By definition, post-hoc explanations make no assumptions of the target model they explain and, as such, they have a low barrier to adoption as they can be applied easily to any model without requiring the user to understand either the model or the data. Essentially, the argument between explaining data or model is a manifestation of a more general challenge which plagues machine learning in that by applying black-box models to large datasets, no specific assumptions, or prior

knowledge have to be encoded, allowing many of these parameter-free models to learn their own rules about the underlying data in an unsupervised fashion. When considering the future of XAI, however, we must consider whether this truly is the best approach. Should we, as a research community apply predictive models without first understanding the data? Furthermore, how useful is an explanation of a model outcome if we do not understand the phenomenon that the model is approximating?

This line of questioning has started to gain traction in the XAI community whereby recently, explainable models are trying to include richer objectives through auxiliary tasks. In particular, with the proliferation of generative modelling, we are seeing the development of models which attempt to learn explanations of the underlying data being modelled. Of these methods, Disentangled Representation Learning (DRL) aims to learn a model capable of identifying and disentangling the underlying factors hidden in the observable data in representation form.

The process of separating underlying factors of variation into variables with semantic meaning imitates the human comprehension process when observing an object or relations. DRL has achieved wide success in image processing where, for example, it has had great success on separating class-invariant content and class-variant style factors from images. For example, paintings by both Picasso and Cezanne are of the same content (e.g., a lake) but they are of different styles. The semantically meaningful representations paired with the generative architectures often means that the user can interact with the model to understand the underlying phenomenon being modelled. When applied to medical images, DRL has been shown to afford representations of the data which are semantically meaningful but allow manipulations of the latent space, and as such, facilitate interpretation. Such interpretability and interactivity has been considered integral to the future of clinical AI systems [31].

Again, much like the methods we have explored in this thesis there exists a development gap between DRL developed for image data and time series data. An interesting avenue of future work would be to explore DRL for time series data. Particularly the concept of content-style disentanglement to formalise what a content and style vector would correspond to for a time series.

7.6 Let's Do Better With Better Metrics

We discussed in Chapter 1 how the XAI community is lacking a unified way of evaluating explanations. As such, the way in which XAI methods are evaluated varies considerably across the literature whereby measures Faithfulness and Deletion remain the most commonly applied quantitative measures. Indeed, throughout this thesis we have tried to be explicit about the assumptions encoded by our evaluation metrics. For example, in Chapter 5, we assumed that minimal explanations are optimal while in 4 we assumed that Deletion under grouped attributions (via Shapley Sets) could be compared to Deletion under singular attribution (the Shapley value).

While we justified both assumptions, by selecting a particular metric, we enforce our own opinion of what constitutes an optimal explanation. The subjectivity surrounding explanation evaluation in the literature is prevalent and particularly concerning when evaluation decisions are not appropriately justified.

As we have argued throughout this thesis, explanation methods should be considered within a multi-disciplinary context. Current evaluation measures, we believe prioritise a quantitative measure with which to benchmark various methods against each other. Particularly the use of Faithfulness has been criticised across the literature [177]. The future of XAI we believe, should seek out evaluation measures which are more holistic in the way they quantify the benefit of XAI towards achieving the an investigative goal.

7.7 Inherently Interpretable Models

As we have discussed in Chapter 2, methods within XAI can largely be categorised as either transparent or post-hoc. The debate as to which type of explainability we must demand from our AI systems lies at the heart of the "accuracy interpretability" trade-off discussion. Due to the proliferation of black-box AI, explainability, in the sense of the machine learning community focuses mostly on post-hoc methods, to make black-box models explainable to a human.

When considering the future of XAI, we should consider what AI systems of the future may look like and how we may go about explaining them. Currently, driven by the success of Large Language Models, it would seem that the AI community is on a trajectory dominated by increasingly powerful neural net architectures. Recently, however, particularly given the connections being made between Large Language Models and Artificial General Intelligence (AGI), there are growing questions concerning the limitations of neural networks, or more generally statistical AI, for AGI. Furthermore, there have been several in the community arguing for the need of a new AI paradigm that moves beyond data-driven machine learning. One of these paradigms, primarily motivated by the un-interpretability of black-box deep learning is neuro-symbolic learning, which addresses the limitations of black-boxes by combining statistical learning with knowledge representation, integrating machine learning with logical reasoning. These neuro-symbolic architectures, which are explainable by design as the behaviour of a neuro-symbolic network can be represented in a set of human-readable expressions, appear to be an exciting potential avenue for the future of XAI with notable recent work [142, 161]. However, neuro-symbolic AI is a long way from widespread adoption on the same level as deep learning. In the meantime, we motivate further scrutiny of the "accuracy-interpretability" trade-off. We argue that what is needed is more transparency and justification surrounding model selection. If it is well-known that one model does not fit all, then practitioners should, at the very least, be aware of what their model is and isn't capable of doing. Furthermore, in those contexts where post-hoc explanations are used, in light of the exploration within this thesis, then the selection of

explanation method should be rigorously justified too.

7.8 Arguing With The Algorithm: Interactive Explanations

We have seen in Chapter 2 how, from the perspective of the social sciences, an explanation entails more than solely an identification of the most probable causes for a particular event. In particular, Hilton [88] argues that explanation takes the form of conversation. Explanations are selected by questions and are thus governed by general rules of discourse. Relating this back to our argument in Chapter 1 where we motivated the empowerment of the individual to argue with the algorithm. An explanation should facilitate this explanatory discourse which allows an end user to interact with the AI system until they feel satisfied. Furthermore, throughout this thesis we have seen how post-hoc local explanations can be wrong. The user should therefore not only be empowered to question why they received a particular outcome from the model but also perhaps why they received a particular explanation. Through interactive explanations, the user can even correct the explanation presented to guide the AI-system. This correction step is crucial for more directly affecting the learner's beliefs and is integral to modulating trust [209]. Interactive explanations will become increasingly important with the proliferation of AI based agents in society whereby an ongoing relationship between user and AI is determined. Surprisingly, despite the motivation from the explanation sciences, the link between interacting, explaining and building trust has been largely ignored by the machine learning literature [209]. We therefore hypothesise that an important part of future will focus on constructing interactive explanation where an explanatory discourse is facilitated between end-user and AI system.

Finally, in Chapter 1, we set out that one of the main intentions of this thesis, via an exploration of post-hoc local explanation methods, was to pave a way to a future where individuals are able to argue with an algorithm in such a way that generates trust and mitigates undesirable behaviour of a black-box model. What we have learned from the work conducted in this thesis is that developing methods for XAI is challenging for many reasons: the numerous assumptions, the difficulty in evaluation, the importance of context. As such, our real hope for the future is that XAI methods become more transparent about their capabilities and limitations.

APPENDIX A

APPENDIX: SHAPLEY VALUE CALCULATIONS

This appendix details the full calculations when applying the Shapley value to Example 4.3 and Example 4.5 as detailed in Chapter 4.

A.1 Interaction in the Data Calculations

This section details the full calculations of the application of the Shapley value to Example 4.4.

Given the input $\mathbf{x} = (x_1, x_2, x_3) = (1, 1, 1)$ and $f(x_1, x_2, x_3) = 2$ Under v_{cond}, the Shapley attributions are as follows:

$$v(\mathbf{x}, \{X_1\}) - v(\mathbf{x}, \varnothing) = \mathbb{E}[f(X)|X_1 = 1] - \mathbb{E}[f(X)] = \frac{3}{2} - 1 = \frac{1}{2}$$

$$v(\mathbf{x}, \{X_1, X_2\}) - v(\mathbf{x}, \{X_2\}) = \mathbb{E}[f(X)|X_1 = 1, X_2 = 1] - \mathbb{E}[f(X)|X_2 = 1] = 2 - \frac{3}{2} = \frac{1}{2}$$

$$v(\mathbf{x}, \{X_1, X_3\}) - v(\mathbf{x}, \{X_3\} = \mathbb{E}[f(X)|X_1 = 1, X_3 = 1] - \mathbb{E}[f(X)|X_3 = 1] = 2 - \frac{3}{2} = \frac{1}{2}$$

$$v(\mathbf{x}, \{X_1, X_2, X_3\}) - v(\mathbf{x}, \{X_2, X_3\}) = \mathbb{E}[f(X)|X_1 = 1, X_3 = 1, X_2 = 1] - \mathbb{E}[f(X)|X_3 = 1, X_2 = 1] = 2 - \frac{3}{2} = \frac{1}{2}$$

$$v(\mathbf{x}, \{X_2\}) - v(\mathbf{x}, \varnothing) = \mathbb{E}[f(X)|X_2 = 1] - \mathbb{E}[f(X)] = \frac{3}{2} - 1 = \frac{1}{2}$$

$$v(\mathbf{x}, \{X_1, X_2\}) - v(\mathbf{x}, \{X_1\}) = \mathbb{E}[f(X)|X_1 = 1, X_2 = 1] - \mathbb{E}[f(X)|X_1 = 1] = 2 - \frac{3}{2} = \frac{1}{2}$$

$$v(\mathbf{x}, \{X_2, X_3\}) - v(\mathbf{x}, \{X_3\}) = \mathbb{E}[f(X)|X_2 = 1, X_3 = 1] - \mathbb{E}[f(X)|X_3 = 1] = \frac{3}{2} - \frac{3}{2} = 0$$

$$v(\mathbf{x}, \{X_1, X_2, X_3\}) - v(\mathbf{x}, \{X_1, X_3\}) = \mathbb{E}[f(X)|X_1 = 1, X_2 = 1, X_3 = 1] - \mathbb{E}[f(X)|X_1 = 1, X_3 = 1] = 2 - 2 = 0$$

$$v(\mathbf{x},\{X_3\}) - v(\mathbf{x},\varnothing) = \mathbb{E}[f(X)|X_2 = 1] - \mathbb{E}[(\mathbb{X})] = \frac{3}{2} - 1 = \frac{1}{2}$$

$$v(\mathbf{x},\{X_1,X_3\}) - v(\mathbf{x},\{X_1\}) = \mathbb{E}[f(X)|X_1 = 1, X_3 = 1] - \mathbb{E}[f(X)|X_1 = 1] = 2 - \frac{3}{2} = \frac{1}{2}$$

$$v(\mathbf{x},\{X_2,X_3\}) - v(\mathbf{x},\{X_2\}) = \mathbb{E}[f(X)|X_2 = 1, X_3 = 1] - \mathbb{E}[f(X)|X_2 = 1] = \frac{3}{2} - \frac{3}{2} = 0$$

$$v(\mathbf{x},\{X_1,X_2,X_3\}) - v(\mathbf{x},\{X_1,X_2\}) = \mathbb{E}[f(X)|X_1 = 1, X_2 = 1, X_3 = 1] - \mathbb{E}[f(X)|X_1 = 1, X_2 = 1] = 2 - 2 = 0$$

$\phi_{X_2} = \phi_{X_3} = \frac{1}{4}$, $\phi_{X_1} = \frac{1}{2}$

A.2 Interaction in the Model Calculations

The following details the full calculations of the application of the Shapley value to Example 4.5 under baseline sample $z = \{0,0,0\}$.

$$v(\mathbf{x},\{X_1\}) - v(\mathbf{x},\varnothing) = (1 + 2 \times 0 \times 0) - (0 + 0 \times 0 \times 0) = 1 - 0 = 1$$

$$v(\mathbf{x},\{X_1,X_2\}) - v(\mathbf{x},\{X_2\}) = (1 + 2 \times 1 \times 0) - (0 + 2 \times 0 \times 0) = 1 - 0 = 1$$

$$v(\mathbf{x},\{X_1,X_3\}) - v(\mathbf{x},\{X_3\}) = (1 + 2 \times 0 \times 1) - (0 + 2 \times 0 \times 0) = 1 - 0 = 1$$

$$v(\mathbf{x},\{X_1,X_2,X_3\}) - v(\mathbf{x},\{X_2,X_3\}) - = (1 + 2 \times 1 \times 1) - (0 + 2 \times 1 \times 1) = 3 - 2 = 1$$

$$v(\mathbf{x},\{X_2\}) - v(\mathbf{x},\varnothing) = (0 + 2 \times 1 \times 0) - (0 + 0 \times 0 \times 0) = 0 - 0 = 0$$

$$v(\mathbf{x},\{X_1,X_2\}) - v(\mathbf{x},\{X_1\}) = (1 + 2 \times 1 \times 0) - (1 + 2 \times 0 \times 0) = 1 - 1 = 0$$

$$v(\mathbf{x},\{X_2,X_3\}) - v(\mathbf{x},\{X_3\}) = (0 + 2 \times 1 \times 1) - (0 + 2 \times 0 \times 1) = 2 - 0 = 2$$

$$v(\mathbf{x},\{X_1,X_2,X_3\}) - v(\mathbf{x},\{X_1,X_3\}) = (1 + 2 \times 1 \times 1) - (1 + 2 \times 0 \times 1) = 3 - 1 = 2$$

$$v(\mathbf{x},\{X_3\}) - v(\mathbf{x},\varnothing) = (0 + 2 \times 0 \times 1) - (0 + 0 \times 0 \times 0) = 0 - 0 = 0$$

$$v(\mathbf{x},\{X_1,X_3\}) - v(\mathbf{x},\{X_1\}) = (1 + 2 \times 0 \times 1) - (1 + 2 \times 0 \times 0) = 1 - 1 = 0$$

$$v(\mathbf{x},\{X_2,X_3\}) - v(\mathbf{x},\{X_2\}) = (0 + 2 \times 1 \times 1) - (0 + 2 \times 1 \times 0) = 2 - 0 = 2$$

$$v(\mathbf{x},\{X_1,X_2,X_3\}) - v(\mathbf{x},\{X_1,X_2\}) = (1 + 2 \times 1 \times 1) - (1 + 2 \times 1 \times 0) = 3 - 1 = 2$$

This gives the Shapley values $\phi_{X_1} = \phi_{X_2} = \phi_{X_3} = 1$.

The following details the full calculations of the application of the Shapley value to Example 4.5 under baseline sample $\mathbf{z}_2 = (0,0,\frac{1}{2})$.

$$v(\mathbf{x},\{X_1\}) - v(\mathbf{x},\varnothing) = (1 + 2 \times 0 \times \frac{1}{2}) - (0 + 2 \times 0 \times \frac{1}{2}) = 1 - 0 = 1$$

$$v(\mathbf{x},\{X_1,X_2\}) - v(\mathbf{x},\{X_2\}) = (1+2\times 1\times \frac{1}{2}) - (0+2\times 1\times \frac{1}{2}) = 2-1 = 1$$
$$v(\mathbf{x},\{X_1,X_3\}) - v(\mathbf{x},\{X_3\}) = (1+2\times 0\times 1) - (0+2\times 0\times 1) = 1-0 = 1$$
$$v(\mathbf{x},\{X_1,X_2,X_3\}) - v(\mathbf{x},\{X_2,X_3\}) = (1+2\times 1\times 1) - (0+2\times 1\times 1) = 3-2 = 1$$

$$v(\mathbf{x},\{X_2\}) - v(\mathbf{x},\varnothing) = (0+2\times 1\times \frac{1}{2}) - (0+2\times 0\times \frac{1}{2}) = 1-0 = 1$$
$$v(\mathbf{x},\{X_1,X_2\}) - v(\mathbf{x},\{X_1\}) = (1+2\times 1\times \frac{1}{2}) - (1+2\times 0\times \frac{1}{2}) = 2-1 = 1$$
$$v(\mathbf{x},\{X_2,X_3\}) - v(\mathbf{x},\{X_3\}) = (0+2\times 1\times 1) - (0+2\times 0\times 1) = 2-0 = 2$$
$$v(\mathbf{x},\{X_1,X_2,X_3\}) - v(\mathbf{x},\{X_1,X_3\}) = (1+2\times 1\times 1) - (1+2\times 0\times 1) = 3-1 = 2$$

$$v(\mathbf{x},\{X_3\}) - v(\mathbf{x},\varnothing) = (0+2\times 0\times 1) - (0+2\times 0\times \frac{1}{2}) = 0-0 = 0$$
$$v(\mathbf{x},\{X_1,X_3\}) - v(\mathbf{x},\{X_1\}) = (1+2\times 0\times 1) - (1+2\times 0\times \frac{1}{2}) = 1-1 = 0$$
$$v(\mathbf{x},\{X_2,X_3\}) - v(\mathbf{x},\{X_2\}) = (0+2\times 1\times 1) - (0+2\times 1\times \frac{1}{2}) = 2-0 = 1$$
$$v(\mathbf{x},\{X_1,X_2,X_3\}) - v(\mathbf{x},\{X_1,X_2\}) = (1+2\times 1\times 1) - (1+2\times 1\times \frac{1}{2}) = 3-2 = 1$$

Which gives the Shapley attribution vector of $\phi_{X_1} = 1$, $\phi_{X_2} = \frac{3}{2}$, $\phi_{X_3} = \frac{1}{2}$.

APPENDIX B

APPENDIX: AUMANN-SHAPLEY CALCULATIONS

This appendix details the full calculation of the Aumann-shapley value for Example 6.4, and Example 6.5 in Chapter 6.

B.1 Calculation of the Aumann-Shapley for Example 6.4

To calculate the Aumann-Shapley value for Example 6.4 we first take the partial derivative of the function f in each variable giving the following equations.

$$f_1 = \ln(2) X_3 2^{X_1+X_2} \tag{B.1}$$

$$f_2 = \ln(2) X_3 2^{X_1+X_2} \tag{B.2}$$

$$f_3 = 2^{X_1+X_2} \tag{B.3}$$

To integrate the partial derivative along the line segment represented by the start point $(0,0,0)$ and $(1,1,1)$ we use the vector representation of a line segment [144] which starts at $\mathbf{x}$ and finishes at $\mathbf{x}'$. The vector representation $r(t)$ of a line segment where t lies on the interval $[0,1]$ [144] is defined as,

$$r(t) = (1-t)\mathbf{x} + t(\mathbf{x}') \tag{B.4}$$

Substituting in our initial $\mathbf{x} = (0,0,0)$ and final $\mathbf{x}' = (1,1,1)$ values for Example 6.4 gives the representation (t,t,t). Substituting t for each X_1, X_2, X_3 from Equations B.1, B.2 and B.3 and then integrating gives the following definite integrals in each variable:

$$X_1 : \int_0^1 \ln(2) t 2^{2t} = \left[\frac{(2\ln(2)t - 1)2^{2(t-1)}}{\ln(2)} \right]_0^1 \tag{B.5}$$

$$X_2 : \int_0^1 \ln(2) t 2^{2t} = \left[\frac{(2\ln(2)t - 1)2^{2(t-1)}}{\ln(2)} \right]_0^1 \tag{B.6}$$

$$X_3 : \int_0^1 2^{2t} = \left[\frac{2^{2t-1}}{\ln(2)} \right]_0^1 \tag{B.7}$$

Finally, taking the definite integral between the limits $t = 1$ and $t = 0$ gives the values for X_1, X_2,

$$\frac{(2\ln(2)1 - 1)2^{2(1-1)}}{\ln(2)} - \frac{(2\ln(2)0 - 1)2^{2(0-1)}}{\ln(2)} = 0.56 - -0.36, \tag{B.8}$$

and for X_3,

$$X_1 = \frac{2^{2-1}}{\ln(2)} - \frac{2^{-1}}{\ln(2)} = 2.885 - 0.72. \tag{B.9}$$

We obtain the Aumann-shapley attribution for X_1, X_2, X_3 given $f(\mathbf{x})$ and $f(\mathbf{x}')$ in Example 6.4 as $\mathbf{z} = (0.92, 0.92, 2.16)$. As the Aumann-Shapley value is calculated via the partial derivative of the function with respect to each individual variable we can see why the Aumann-Shapley value is robust to the reformulation of the function f, f' from Example 6.4 as the partial derivative of X_3 is constant in both formulations. This essentially encapsulates the fact that the Aumann-Shapley value considers change in variables as they happen over time rather than consecutively. To see this, for f', our partial derivatives of the modified function and variable set $\mathbf{X} = (X_1, X_3)$ are as follows:

$$f_1' = \ln(2) X_3 2^{2X_1 + 1}, \tag{B.10}$$

$$f_3' = 2^{2X_1}. \tag{B.11}$$

Given the vector representation of the line segment between endpoints $(0,0)$ and $(1,1)$, $r(t) = (t,t)$ generates the resulting attribution, $\mathbf{x} = (1.84, 2.16)$.

B.2 Calculation of the Aumann-Shapley for Example 6.5

To show how the Aumann-Shapley value satisfies Differential Attribution Desideratum 2 we apply it to Example 6.5. Applying the Aumann-Shapley value to f first generates the following partial derivatives.

(B.12) $$f_1 = X_3$$

(B.13) $$f_2 = X_3$$

(B.14) $$f_3 = X_1 + X_2$$

We then obtain the following vector representation of the line segment with endpoints at $\mathbf{x} = (0,0,0)$ and $\mathbf{x}' = (1,2,-1)$

(B.15) $$r(t) = (1-t)\mathbf{x} + t(\mathbf{x}') = (t, 2t, -t).$$

Substituting our vector representation of the line segment into the partial derivatives and integrating gives the definite integrals,

(B.16) $$X_1 : \int_0^1 t = \left[-\frac{1}{2}t^2\right][_0^1$$

(B.17) $$X_2 : \int_0^1 t = \left[-\frac{1}{2}t^2\right][_0^1$$

(B.18) $$\int_0^1 3t = \left[\frac{3}{2}t^2\right]_0^1$$

Finally, taking the definite integral between the limits $t = 1$ and $t = 0$ gives the values, for X_1, X_2, $-\frac{1}{2}$, and for X_3, $\frac{3}{2}$. Multiplying these with the change in each variable $x_i - x_i'$ results in the Aumann-Shapley attribution vector $\mathbf{z} = (-0.5, -1, -1.5)$. We note that for this function which is multi-linear, the Aumann-Shapley value coincides with the Shapley value. However, when applying the Aumann-Shapley value to the modified function f', we obtain the following partial derivatives.

(B.19) $$f_1' = 2X_1X_3$$

(B.20) $$f_2' = X_3$$

(B.21) $$f_3' = X_1^2 + X_2$$

Which, given the same vector representation of the line segment as before $r(t) = (t, 2t, -t)$ gives the definite integrals,

(B.22) $$X_1 : \int_0^1 -2t^2 = \left[-\frac{2}{3}t^3\right]_0^1$$

(B.23) $$X_2 : \int_0^1 -t = \left[-\frac{1}{2}t^2\right]_0^1$$

(B.24) $$X_3 : \int_0^1 t^2 + 2t = \left[\frac{1}{3}t^3 + t^2\right]_0^1$$

Given the above definite integrals, the Aumann-Shapley value gives the following attribution $\{-\frac{2}{3}, -1, -\frac{4}{3}\}$ which, when compared to the attribution of f, has increased the attribution awarded to X_1 in accordance with its greater impact on the function as determined by its partial derivative, capturing the behaviour of the function over the entire attribution region despite the same outcome at the endpoints.

BIBLIOGRAPHY

[1] K. AAS, M. JULLUM, AND A. LØLAND, *Explaining individual predictions when features are dependent: More accurate approximations to shapley values*, Artificial Intelligence, 298 (2021), p. 103502.

[2] R. P. ADAMS AND D. J. MACKAY, *Bayesian online changepoint detection*, arXiv preprint arXiv:0710.3742, (2007).

[3] C. AGARWAL AND A. NGUYEN, *Explaining image classifiers by removing input features using generative models*, in Proceedings of the Asian Conference on Computer Vision, 2020.

[4] A. ALLAM, S. FEUERRIEGEL, M. REBHAN, AND M. KRAUTHAMMER, *Analyzing patient trajectories with artificial intelligence*, Journal of medical internet research, 23 (2021), p. e29812.

[5] D. ALVAREZ MELIS AND T. JAAKKOLA, *Towards robust interpretability with self-explaining neural networks*, Advances in neural information processing systems, 31 (2018).

[6] D. ALVAREZ-MELIS AND T. S. JAAKKOLA, *On the robustness of interpretability methods*, arXiv preprint arXiv:1806.08049, (2018).

[7] L. AMGOUD AND J. BEN-NAIM, *Axiomatic foundations of explainability*, in 31st International Joint Conference on Artificial Intelligence (IJCAI 2022), 2022.

[8] S. AMINIKHANGHAHI AND D. J. COOK, *A survey of methods for time series change point detection*, Knowledge and information systems, 51 (2017), pp. 339–367.

[9] R. ANDERSON, *The rashomon effect and communication*, Canadian Journal of Communication, 41 (2016), pp. 249–270.

[10] J. ANGWIN, J. LARSON, S. MATTU, AND L. KIRCHNER, *Machine bias: there's software used across the country to predict future criminals. and it's biased against blacks. propublica 2016*, 2016.

[11] S. ANJOMSHOAE, A. NAJJAR, D. CALVARESI, AND K. FRÄMLING, *Explainable agents and robots: Results from a systematic literature review*, in Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, International Foundation for Autonomous Agents and Multiagent Systems, 2019, pp. 1078–1088.

[12] R. T. P. ANONYMITY, *Comparing the fairness of two popular solution concepts of coalition games: Shapley value and nucleolus*.

[13] A. B. ARRIETA, N. DÍAZ-RODRÍGUEZ, J. DEL SER, A. BENNETOT, S. TABIK, A. BARBADO, S. GARCÍA, S. GIL-LÓPEZ, D. MOLINA, R. BENJAMINS, ET AL., *Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai*, arXiv preprint arXiv:1910.10045, (2019).

[14] E. ATES, B. AKSAR, V. J. LEUNG, AND A. K. COSKUN, *Counterfactual explanations for machine learning on multivariate time series data*, arXiv preprint arXiv:2008.10781, (2020).

[15] R. AUMANN AND L. SHAPLEY, *Values of non-atomic games i: The axiomatic approach*, tech. rep., HEBREW UNIV JERUSALEM (ISRAEL) DEPT OF MATHEMATICS, 1968.

[16] R. J. AUMANN AND L. S. SHAPLEY, *Values of non-atomic games*, Princeton University Press, 2015.

[17] L. AURET AND C. ALDRICH, *Interpretation of nonlinear relationships between process variables by use of random forests*, Minerals Engineering, 35 (2012), pp. 27–42.

[18] J. BA AND R. CARUANA, *Do deep nets really need to be deep?*, in Advances in neural information processing systems, 2014, pp. 2654–2662.

[19] E. BAREINBOIM, J. D. CORREA, D. IBELING, AND T. ICARD, *On pearl's hierarchy and the foundations of causal inference*, in Probabilistic and causal inference: the works of judea pearl, 2022, pp. 507–556.

[20] A. BELL, I. SOLANO-KAMAIKO, O. NOV, AND J. STOYANOVICH, *It's just not that simple: an empirical study of the accuracy-explainability trade-off in machine learning for public policy*, in 2022 ACM Conference on Fairness, Accountability, and Transparency, 2022, pp. 248–266.

[21] R. BELLMAN AND R. KALABA, *On adaptive control processes*, IRE Transactions on Automatic Control, 4 (1959), pp. 1–9.

[22] S. BLOEMHEUVEL, J. VAN DEN HOOGEN, D. JOZINOVIĆ, A. MICHELINI, AND M. ATZMUELLER, *Graph neural networks for multivariate time series regression with application to seismic data*, International Journal of Data Science and Analytics, (2022), pp. 1–16.

[23] J. F. BRISSY, *Computers in organizations: The (white) magic of the black box*, Organizational symbolism, (1990), pp. 225–236.

[24] E. BRYNJOLFSSON, D. ROCK, AND C. SYVERSON, *Artificial intelligence and the modern productivity paradox: A clash of expectations and statistics*, in The economics of artificial intelligence: An agenda, University of Chicago Press, 2018, pp. 23–57.

[25] S. BUBECK, V. CHANDRASEKARAN, R. ELDAN, J. GEHRKE, E. HORVITZ, E. KAMAR, P. LEE, Y. T. LEE, Y. LI, S. LUNDBERG, ET AL., *Sparks of artificial general intelligence: Early experiments with gpt-4*, arXiv preprint arXiv:2303.12712, (2023).

[26] H. BUHRMAN AND R. DE WOLF, *Complexity measures and decision tree complexity: a survey*, Theoretical Computer Science, 288 (2002), pp. 21–43.

[27] C. CASTRO, *What's wrong with machine bias*, Ergo, an Open Access Journal of Philosophy, 6 (2019).

[28] E. CAUER, W. MATHIS, AND R. PAULI, *Life and work of wilhelm cauer (1900 1945)*, in Proc. 14th Int. Symp. Mathematical Theory of Networks and Systems, MTNS, 2000, pp. 1–10.

[29] A. CHADDAD, J. PENG, J. XU, AND A. BOURIDANE, *Survey of explainable ai techniques in healthcare*, Sensors, 23 (2023), p. 634.

[30] L. CHANG, *Pyearth: A lightweight python package for earth science*, 2022.

[31] A. CHARTSIAS, T. JOYCE, G. PAPANASTASIOU, S. SEMPLE, M. WILLIAMS, D. E. NEWBY, R. DHARMAKUMAR, AND S. A. TSAFTARIS, *Disentangled representation learning in cardiac image analysis*, Medical image analysis, 58 (2019), p. 101535.

[32] C. CHATFIELD AND H. XING, *The analysis of time series: an introduction with R*, CRC press, 2019.

[33] J. CHEN, Y. LI, X. WU, Y. LIANG, AND S. JHA, *Robust out-of-distribution detection for neural networks*, arXiv preprint arXiv:2003.09711, (2020).

[34] M. CHEN, W. DU, Y. TANG, Y. JIN, AND G. G. YEN, *A decomposition method for both additively and non-additively separable problems*, IEEE Transactions on Evolutionary Computation, (2022).

[35] Y. CHEN, E. KEOGH, B. HU, N. BEGUM, A. BAGNALL, A. MUEEN, AND G. BATISTA, *The ucr time series classification archive*, July 2015.
www.cs.ucr.edu/~eamonn/time_series_data/.

[36] E. CHO, *The social credit system: Not just another chinese idiosyncrasy*, Journal of Public and International Affairs, (2020).

[37] M. CHRYSANTHOU, *Transparency and selfhood:: Utopia and the informed body*, Social Science & Medicine, 54 (2002), pp. 469–479.

[38] R. COOPER, *The post-modern state and the world order*, Demos, 2000.

[39] J. CRABBÉ AND M. VAN DER SCHAAR, *Explaining time series predictions with dynamic masks*, in International Conference on Machine Learning, PMLR, 2021, pp. 2166–2177.

[40] M. CRAVEN AND J. SHAVLIK, *Extracting tree-structured representations of trained networks*, Advances in neural information processing systems, 8 (1995).

[41] R. F. CRESPO, *Causality, teleology and explanation in social sciences*, (2016).

[42] P. DABKOWSKI AND Y. GAL, *Real time image saliency for black box classifiers*, Advances in neural information processing systems, 30 (2017).

[43] S. DANDL, C. MOLNAR, M. BINDER, AND B. BISCHL, *Multi-objective counterfactual explanations*, in Parallel Problem Solving from Nature–PPSN XVI: 16th International Conference, PPSN 2020, Leiden, The Netherlands, September 5-9, 2020, Proceedings, Part I, Springer, 2020, pp. 448–469.

[44] A. DAS AND P. RAD, *Opportunities and challenges in explainable artificial intelligence (xai): A survey*, arXiv preprint arXiv:2006.11371, (2020).

[45] A. DATTA, S. SEN, AND Y. ZICK, *Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems*, in 2016 IEEE symposium on security and privacy (SP), IEEE, 2016, pp. 598–617.

[46] T. DAVENPORT, J. LOUCKS, AND D. SCHATSKY, *Bullish on the business value of cognitive, leaders in cognitive and ai weigh in on what's working and what's next, the 2017 deloitte state of cognitive survey, deloitte white paper, dostęp sierpień 2018*, 2017.

[47] J. DE ABREU FONTES, M. J. ANZANELLO, J. B. G. DE BRITO, G. B. BUCCO, F. S. FOGLIATTO, AND F. PUGLIA, *Combining wavelength importance ranking to the random forest classifier to analyze multiclass spectral data*, Forensic Science International, (2021), p. 110998.

[48] A. DEMPSTER, F. PETITJEAN, AND G. I. WEBB, *Rocket: exceptionally fast and accurate time series classification using random convolutional kernels*, Data Mining and Knowledge Discovery, 34 (2020), pp. 1454–1495.

[49] H. DENG, G. RUNGER, E. TUV, AND M. VLADIMIR, *A time series forest for classification and feature extraction*, Information Sciences, 239 (2013), pp. 142–153.

[50] A. DHURANDHAR, P.-Y. CHEN, R. LUSS, C.-C. TU, P. TING, K. SHANMUGAM, AND P. DAS, *Explanations based on the missing: Towards contrastive explanations with pertinent negatives*, in Advances in neural information processing systems, 2018, pp. 592–603.

[51] F. DI MARTINO AND F. DELMASTRO, *Explainable ai for clinical and remote health applications: a survey on tabular and time series data*, Artificial Intelligence Review, (2022), pp. 1–55.

[52] J. DIEBER AND S. KIRRANE, *Why model why? assessing the strengths and limitations of lime*, arXiv preprint arXiv:2012.00093, (2020).

[53] I. S. M. F. N. V. P. R. R. A. S. A. S. M. S. K. T. S. T. A.-S. DIVYA BALASUBRAMANIAN, KAI HOU YIP, *Communicating high-street bakery sales predictions using counterfactual explanations*, Data Study Group, Alan Turing Institute, (2021).

[54] A.-K. DOMBROWSKI, M. ALBER, C. ANDERS, M. ACKERMANN, K.-R. MÜLLER, AND P. KESSEL, *Explanations can be manipulated and geometry is to blame*, Advances in neural information processing systems, 32 (2019).

[55] D. DORAN, S. SCHULZ, AND T. BESOLD, *What does explainable ai really mean? a new conceptualization of perspectives, 2017*, arXiv preprint arXiv:1710.00794.

[56] F. DOSHI-VELEZ AND B. KIM, *Towards a rigorous science of interpretable machine learning*, arXiv preprint arXiv:1702.08608, (2017).

[57] A. B. DOWNEY, *A novel changepoint detection algorithm*, arXiv preprint arXiv:0812.1237, (2008).

[58] N. A. ERNST AND G. BAVOTA, *Ai-driven development is here: Should you worry?*, IEEE Software, 39 (2022), pp. 106–110.

[59] L. FLORIDI, *Ai as agency without intelligence: On chatgpt, large language models, and other generative models*, Philosophy & Technology, 36 (2023), p. 15.

[60] R. C. FONG AND A. VEDALDI, *Interpretable explanations of black boxes by meaningful perturbation*, in Proceedings of the IEEE international conference on computer vision, 2017, pp. 3429–3437.

[61] D. FRANCOIS, V. WERTZ, M. VERLEYSEN, ET AL., *About the locality of kernels in high-dimensional spaces*, in International Symposium on Applied Stochastic Models and Data Analysis, Citeseer, 2005, pp. 238–245.

[62] J. H. FRIEDMAN, *Multivariate adaptive regression splines*, The annals of statistics, 19 (1991), pp. 1–67.

[63] ——, *Greedy function approximation: a gradient boosting machine*, Annals of statistics, (2001), pp. 1189–1232.

[64] C. FRYE, C. ROWAT, AND I. FEIGE, *Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability*, Advances in Neural Information Processing Systems, 33 (2020), pp. 1229–1239.

[65] D. FRYER, I. STRÜMKE, AND H. NGUYEN, *Shapley values for feature selection: The good, the bad, and the axioms*, IEEE Access, 9 (2021), pp. 144352–144360.

[66] D. GARREAU AND S. ARLOT, *Consistent change-point detection with kernels*, Electronic Journal of Statistics, 12 (2018), pp. 4440–4486.

[67] D. GARREAU AND D. MARDAOUI, *What does lime really see in images?*, arXiv preprint arXiv:2102.06307, (2021).

[68] D. GATELY, *Sharing the gains from regional cooperation: A game theoretic application to planning investment in electric power*, International Economic Review, (1974), pp. 195–208.

[69] A. GEVAERT, A.-J. ROUSSEAU, T. BECKER, D. VALKENBORG, T. DE BIE, AND Y. SAEYS, *Evaluating feature attribution methods in the image domain*, arXiv preprint arXiv:2202.12270, (2022).

[70] S. GHARGHABI, Y. DING, C.-C. M. YEH, K. KAMGAR, L. ULANOVA, AND E. KEOGH, *Matrix profile viii: domain agnostic online semantic segmentation at superhuman performance levels*, in 2017 IEEE international conference on data mining (ICDM), IEEE, 2017, pp. 117–126.

[71] M. GHASSEMI, L. OAKDEN-RAYNER, AND A. L. BEAM, *The false hope of current approaches to explainable artificial intelligence in health care*, The Lancet Digital Health, 3 (2021), pp. e745–e750.

[72] R. P. GILLES AND L. MALLOZZI, *Generalised gately values of cooperative games*, arXiv preprint arXiv:2208.10189, (2022).

[73] T. GÓRECKI AND P. PIASECKI, *A comprehensive comparison of distance measures for time series classification*, in Stochastic Models, Statistics and Their Applications: Dresden, Germany, March 2019 14, Springer, 2019, pp. 409–428.

[74] A. GRAMEGNA AND P. GIUDICI, *Shap and lime: an evaluation of discriminative power in credit risk*, Frontiers in Artificial Intelligence, 4 (2021), p. 752558.

[75] H. P. GRICE, *Logic and conversation*, in Speech acts, Brill, 1975, pp. 41–58.

[76] R. GUIDOTTI, A. MONREALE, S. RUGGIERI, D. PEDRESCHI, F. TURINI, AND F. GIANNOTTI, *Local rule-based explanations of black box decision systems*, arXiv preprint arXiv:1805.10820, (2018).

[77] M. GUILLEMÉ, V. MASSON, L. ROZÉ, AND A. TERMIER, *Agnostic local explanation for time series classification*, in 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI), IEEE, 2019, pp. 432–439.

[78] H. HAGRAS, *Toward human-understandable, explainable ai*, Computer, 51 (2018), pp. 28–36.

[79] O. HAIMANKO, *Partially symmetric values*, Mathematics of Operations Research, 25 (2000), pp. 573–590.

[80] J. Y. HALPERN, *A modification of the halpern-pearl definition of causality*, arXiv preprint arXiv:1505.00162, (2015).

[81] J. Y. HALPERN AND J. PEARL, *Causes and explanations: A structural-model approach. part i: Causes*, The British journal for the philosophy of science, (2005).

[82] J. Y. HALPERN AND J. PEARLY, *Causes and explanations: A structural-model approach—part i: Causes*, (2001).

[83] K. HAO, *An ai saw a cropped photo of aoc, it autocompleted her wearing a bikini*, 2021.

[84] J. HAREL, C. KOCH, AND P. PERONA, *Graph-based visual saliency*, in Advances in neural information processing systems, 2007, pp. 545–552.

[85] W. HEAVEN, *Geoffrey hinton tells us why he is now scared of the tech he helped build*, MIT Technology Review, (2023).

[86] W. D. HEAVEN, *Hundreds of ai tools have been built to catch covid. none of them helped*, MIT Technology Review. Retrieved October, 6 (2021), p. 2021.

[87] T. HESKES, E. SIJBEN, I. G. BUCUR, AND T. CLAASSEN, *Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models*, Advances in neural information processing systems, 33 (2020), pp. 4778–4789.

[88] D. J. HILTON, *Conversational processes and causal explanation.*, Psychological Bulletin, 107 (1990), p. 65.

[89] G. HINTON, L. DENG, D. YU, G. E. DAHL, A.-R. MOHAMED, N. JAITLY, A. SENIOR, V. VANHOUCKE, P. NGUYEN, T. N. SAINATH, ET AL., *Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups*, IEEE Signal processing magazine, 29 (2012), pp. 82–97.

[90] C. R. HITCHCOCK, *The mishap at reichenbach fall: Singular vs. general causation*, Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition, 78 (1995), pp. 257–291.

[91] G. HOOKER, L. MENTCH, AND S. ZHOU, *Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance*, Statistics and Computing, 31 (2021), pp. 1–16.

[92] A. HUMAYUN, F. LI, AND J. M. REHG, *The middle child problem: Revisiting parametric min-cut and seeds for object proposals*, in Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1600–1608.

[93] H. ISMAIL FAWAZ, G. FORESTIER, J. WEBER, L. IDOUMGHAR, AND P.-A. MULLER, *Deep learning for time series classification: a review*, Data mining and knowledge discovery, 33 (2019), pp. 917–963.

[94] T. ITO, K. OCHIAI, AND Y. FUKAZAWA, *C-lime: A consistency-oriented lime for time-series health-risk predictions*, in Knowledge Management and Acquisition for Intelligent Systems: 17th Pacific Rim Knowledge Acquisition Workshop, PKAW 2020, Yokohama, Japan, January 7–8, 2021, Proceedings 17, Springer, 2021, pp. 58–69.

[95] D. JANZING, L. MINORICS, AND P. BLÖBAUM, *Feature relevance quantification in explainable ai: A causal problem*, in International Conference on Artificial Intelligence and Statistics, PMLR, 2020, pp. 2907–2916.

[96] Y. JIA, J. BAILEY, K. RAMAMOHANARAO, C. LECKIE, AND M. E. HOULE, *Improving the quality of explanations with local embedding perturbations*, in Proceedings of the 25th ACM SIGKDD International conference on knowledge discovery & Data Mining, 2019, pp. 875–884.

[97] U. JOHANSSON, L. NIKLASSON, AND R. KÖNIG, *Accuracy vs. comprehensibility in data mining models*, in Proceedings of the seventh international conference on information fusion, vol. 1, 2004, pp. 295–300.

[98] A. E. JOHNSON, T. J. POLLARD, L. SHEN, L.-W. H. LEHMAN, M. FENG, M. GHASSEMI, B. MOODY, P. SZOLOVITS, L. ANTHONY CELI, AND R. G. MARK, *Mimic-iii, a freely accessible critical care database*, Scientific data, 3 (2016), pp. 1–9.

[99] J. R. JOSEPHSON AND S. G. JOSEPHSON, *Abductive inference: Computation, philosophy, technology*, Cambridge University Press, 1996.

[100] F. KARIM, S. MAJUMDAR, AND H. DARABI, *Insights into lstm fully convolutional networks for time series classification*, IEEE Access, 7 (2019), pp. 67718–67725.

[101] E. KASNECI, K. SESSLER, S. KÜCHEMANN, M. BANNERT, D. DEMENTIEVA, F. FISCHER, U. GASSER, G. GROH, S. GÜNNEMANN, E. HÜLLERMEIER, ET AL., *Chatgpt for good? on opportunities and challenges of large language models for education*, Learning and Individual Differences, 103 (2023), p. 102274.

[102] M. T. KEANE, E. M. KENNY, E. DELANEY, AND B. SMYTH, *If only we had better counterfactual explanations: Five key deficits to rectify in the evaluation of counterfactual xai techniques*, arXiv preprint arXiv:2103.01035, (2021).

[103] C. J. KELLY, A. KARTHIKESALINGAM, M. SULEYMAN, G. CORRADO, AND D. KING, *Key challenges for delivering clinical impact with artificial intelligence*, BMC medicine, 17 (2019), p. 195.

[104] E. KEOGH, L. WEI, X. XI, S. LONARDI, J. SHIEH, AND S. SIROWY, *Intelligent icons: Integrating lite-weight data mining and visualization into gui operating systems*, in Sixth International Conference on Data Mining (ICDM'06), IEEE, 2006, pp. 912–916.

[105] P. KIDGER, J. MORRILL, AND T. LYONS, *Generalised interpretable shapelets for irregular time series*, arXiv preprint arXiv:2005.13948, (2020).

[106] B. KIM, R. KHANNA, AND O. O. KOYEJO, *Examples are not enough, learn to criticize! criticism for interpretability*, in Advances in neural information processing systems, 2016, pp. 2280–2288.

[107] T. W. KIM, *Explainable artificial intelligence (xai), the goodness criteria and the grasp-ability test*, arXiv preprint arXiv:1810.09598, (2018).

[108] B. V. KINI AND C. C. SEKHAR, *Large margin mixture of ar models for time series classification*, Applied Soft Computing, 13 (2013), pp. 361–371.

[109] D. S. KIRSCHEN, *Demand-side view of electricity markets*, IEEE Transactions on power systems, 18 (2003), pp. 520–527.

[110] P. W. KOH AND P. LIANG, *Understanding black-box predictions via influence functions*, in International conference on machine learning, PMLR, 2017, pp. 1885–1894.

[111] N. KOKHLIKYAN, V. MIGLANI, M. MARTIN, E. WANG, B. ALSALLAKH, J. REYNOLDS, A. MELNIKOV, N. KLIUSHKINA, C. ARAYA, S. YAN, AND O. REBLITZ-RICHARDSON, *Captum: A unified and generic model interpretability library for pytorch*, 2020.

[112] R. KOMMIYA MOTHILAL, D. MAHAJAN, C. TAN, AND A. SHARMA, *Towards unifying feature attribution and counterfactual explanations: Different means to the same end*, in Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, 2021, pp. 652–663.

[113] M. KOMOROWSKI, L. A. CELI, O. BADAWI, A. C. GORDON, AND A. A. FAISAL, *The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care*, Nature medicine, 24 (2018), pp. 1716–1720.

[114] R. KONIG, U. JOHANSSON, AND L. NIKLASSON, *G-rex: A versatile framework for evolutionary data mining*, in 2008 IEEE International Conference on Data Mining Workshops, IEEE, 2008, pp. 971–974.

[115] A. KRIZHEVSKY, I. SUTSKEVER, AND G. E. HINTON, *Imagenet classification with deep convolutional neural networks*, Advances in neural information processing systems, 25 (2012).

[116] I. KUMAR, C. SCHEIDEGGER, S. VENKATASUBRAMANIAN, AND S. FRIEDLER, *Shapley residuals: Quantifying the limits of the shapley value for explanations*, Advances in Neural Information Processing Systems, 34 (2021), pp. 26598–26608.

[117] I. E. KUMAR, S. VENKATASUBRAMANIAN, C. SCHEIDEGGER, AND S. FRIEDLER, *Problems with shapley-value-based explanations as feature importance measures*, in International Conference on Machine Learning, PMLR, 2020, pp. 5491–5500.

[118] N. B. KUMARAKULASINGHE, T. BLOMBERG, J. LIU, A. S. LEAO, AND P. PAPAPETROU, *Evaluating local interpretable model-agnostic explanations on clinical machine learning classification models*, in 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS), IEEE, 2020, pp. 7–12.

[119] I. LAGE, E. CHEN, J. HE, M. NARAYANAN, B. KIM, S. GERSHMAN, AND F. DOSHI-VELEZ, *An evaluation of the human-interpretability of explanation*, arXiv preprint arXiv:1902.00006, (2019).

[120] P. LAI, N. PHAN, H. HU, A. BADETI, D. NEWMAN, AND D. DOU, *Ontology-based interpretable machine learning for textual data*, in 2020 International Joint Conference on Neural Networks (IJCNN), IEEE, 2020, pp. 1–10.

[121] J. LAMONT, *Distributive justice*, Routledge, 2017.

[122] M. LANGER, D. OSTER, T. SPEITH, H. HERMANNS, L. KÄSTNER, E. SCHMIDT, A. SESING, AND K. BAUM, *What do we want from explainable artificial intelligence (xai)?–a stakeholder perspective on xai and a conceptual model guiding interdisciplinary xai research*, Artificial Intelligence, 296 (2021), p. 103473.

[123] J. LARSON, S. MATTU, L. KIRCHNER, AND J. ANGWIN, *How we analyzed the compas recidivism algorithm*, ProPublica (5 2016), 9 (2016), pp. 3–3.

[124] T. LAUGEL, X. RENARD, M.-J. LESOT, C. MARSALA, AND M. DETYNIECKI, *Defining locality for surrogates in post-hoc interpretablity*, arXiv preprint arXiv:1806.07498, (2018).

[125] M. LENG AND M. PARLAR, *Analytic solution for the nucleolus of a three-player cooperative game*, Naval Research Logistics (NRL), 57 (2010), pp. 667–672.

[126] S. LEVIN AND J. C. WONG, *Self-driving uber kills arizona woman in first fatal crash involving pedestrian*, The Guardian, 19 (2018).

[127] H. LI, J. LI, X. GUAN, B. LIANG, Y. LAI, AND X. LUO, *Research on overfitting of deep learning*, in 2019 15th international conference on computational intelligence and security (CIS), IEEE, 2019, pp. 78–81.

[128] D. LIBEN-NOWELL, A. SHARP, T. WEXLER, AND K. WOODS, *Computing shapley value in supermodular coalitional games*, in Computing and Combinatorics: 18th Annual International Conference, COCOON 2012, Sydney, Australia, August 20-22, 2012. Proceedings 18, Springer, 2012, pp. 568–579.

[129] B. LIM AND S. ZOHREN, *Time-series forecasting with deep learning: a survey*, Philosophical Transactions of the Royal Society A, 379 (2021), p. 20200209.

[130] B. LIM, S. ZOHREN, AND S. ROBERTS, *Recurrent neural filters: Learning independent bayesian filtering steps for time series prediction*, in 2020 International Joint Conference on Neural Networks (IJCNN), IEEE, 2020, pp. 1–8.

[131] S. LIPOVETSKY AND M. CONKLIN, *Analysis of regression in game theory approach*, Applied Stochastic Models in Business and Industry, 17 (2001), pp. 319–330.

[132] P. LIPTON, *Contrastive explanation*, Royal Institute of Philosophy Supplements, 27 (1990), pp. 247–266.

[133] Z. C. LIPTON, *The mythos of model interpretability*, Queue, 16 (2018), pp. 31–57.

[134] G. LIU, J. ZHANG, A. B. CHAN, AND J. H. HSIAO, *Human attention-guided explainable artificial intelligence for computer vision models*, arXiv preprint arXiv:2305.03601, (2023).

[135] M. LÖNING, A. BAGNALL, S. GANESH, V. KAZAKOV, J. LINES, AND F. J. KIRÁLY, *sktime: A unified interface for machine learning with time series*, arXiv preprint arXiv:1909.07872, (2019).

[136] S. M. LUNDBERG, G. ERION, H. CHEN, A. DEGRAVE, J. M. PRUTKIN, B. NAIR, R. KATZ, J. HIMMELFARB, N. BANSAL, AND S.-I. LEE, *From local explanations to global understanding with explainable ai for trees*, Nature Machine Intelligence, 2 (2020), pp. 2522–5839.

[137] S. M. LUNDBERG, G. G. ERION, AND S.-I. LEE, *Consistent individualized feature attribution for tree ensembles*, arXiv preprint arXiv:1802.03888, (2018).

[138] S. M. LUNDBERG AND S.-I. LEE, *A unified approach to interpreting model predictions*, in Advances in neural information processing systems, 2017, pp. 4765–4774.

[139] ——, *A unified approach to interpreting model predictions*, in Advances in Neural Information Processing Systems 30, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds., Curran Associates, Inc., 2017, pp. 4765–4774.

[140] X. MA, F. LIU, Y. QI, X. WANG, L. LI, L. JIAO, M. YIN, AND M. GONG, *A multiobjective evolutionary algorithm based on decision variable analyses for multiobjective optimization problems with large-scale variables*, IEEE Transactions on Evolutionary Computation, 20 (2015), pp. 275–298.

[141] P. MADUMAL, *Explainable agency in intelligent agents: Doctoral consortium*, in Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, International Foundation for Autonomous Agents and Multiagent Systems, 2019, pp. 2432–2434.

[142] J. MAO, C. GAN, P. KOHLI, J. B. TENENBAUM, AND J. WU, *The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision*, arXiv preprint arXiv:1904.12584, (2019).

[143] D. MARDAOUI AND D. GARREAU, *An analysis of lime for text data*, in International Conference on Artificial Intelligence and Statistics, PMLR, 2021, pp. 3493–3501.

[144] J. E. MARSDEN AND A. TROMBA, *Vector calculus*, Macmillan, 2003.

[145] R. L. MASON AND J. C. YOUNG, *Autocorrelation in multivariate processes*, in Statistical Process Monitoring and Optimization, CRC Press, 1999, pp. 243–260.

[146] L. MERRICK AND A. TALY, *The explanation game: Explaining machine learning models using shapley values*, in International Cross-Domain Conference for Machine Learning and Knowledge Extraction, Springer, 2020, pp. 17–38.

[147] T. MILLER, *Explanation in artificial intelligence: Insights from the social sciences*, Artificial Intelligence, 267 (2019), pp. 1–38.

[148] ——, *Contrastive explanation: A structural-model approach*, The Knowledge Engineering Review, 36 (2021), p. e14.

[149] T. MILLER, P. HOWE, AND L. SONENBERG, *Explainable ai: Beware of inmates running the asylum or: How i learnt to stop worrying and love the social and behavioural sciences*, arXiv preprint arXiv:1712.00547, (2017).

[150] B. MITTELSTADT, C. RUSSELL, AND S. WACHTER, *Explaining explanations in ai*, in Proceedings of the conference on fairness, accountability, and transparency, ACM, 2019, pp. 279–288.

[151] K. E. MOKHTARI, B. P. HIGDON, AND A. BAŞAR, *Interpreting financial time series with shap values*, in Proceedings of the 29th Annual International Conference on Computer Science and Software Engineering, 2019, pp. 166–172.

[152] C. MOLNAR, *Interpretable machine learning*, Lulu. com, 2019.

[153] F. MUJKANOVIC, V. DOSKOČ, M. SCHIRNECK, P. SCHÄFER, AND T. FRIEDRICH, *timexplain–a framework for explaining the predictions of time series classifiers*, arXiv preprint arXiv:2007.07606, (2020).

[154] N. NARODYTSKA, A. SHROTRI, K. S. MEEL, A. IGNATIEV, AND J. MARQUES-SILVA, *Assessing heuristic machine learning explanations with model counting*, in Theory and Applications of Satisfiability Testing–SAT 2019: 22nd International Conference, SAT 2019, Lisbon, Portugal, July 9–12, 2019, Proceedings 22, Springer, 2019, pp. 267–278.

[155] M. NAUTA, J. TRIENES, S. PATHAK, E. NGUYEN, M. PETERS, Y. SCHMITT, J. SCHLÖTTERER, M. VAN KEULEN, AND C. SEIFERT, *From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai*, ACM Computing Surveys, (2022).

[156] I. NEVES, D. FOLGADO, S. SANTOS, M. BARANDAS, A. CAMPAGNER, L. RONZIO, F. CABITZA, AND H. GAMBOA, *Interpretable heartbeat classification using local model-agnostic explanations on ecgs*, Computers in Biology and Medicine, 133 (2021), p. 104393.

[157] A. NIELSEN, *Practical time series analysis: Prediction with statistics and machine learning*, O'Reilly Media, 2019.

[158] M. N. OMIDVAR, X. LI, Y. MEI, AND X. YAO, *Cooperative co-evolution with differential grouping for large scale optimization*, IEEE Transactions on evolutionary computation, 18 (2013), pp. 378–393.

[159] G. OWEN, *Multilinear extensions of games*, Management Science, 18 (1972), pp. 64–79.

[160] I. PALATNIK DE SOUSA, M. MARIA BERNARDES REBUZZI VELLASCO, AND E. COSTA DA SILVA, *Local interpretable model-agnostic explanations for classification of lymph node metastases*, Sensors, 19 (2019), p. 2969.

[161] E. PARISOTTO, A.-R. MOHAMED, R. SINGH, L. LI, D. ZHOU, AND P. KOHLI, *Neuro-symbolic program synthesis*, arXiv preprint arXiv:1611.01855, (2016).

[162] J. PEARL, *Direct and indirect effects*, in Probabilistic and causal inference: The works of Judea Pearl, 2022, pp. 373–392.

[163] ——, *Probabilities of causation: three counterfactual interpretations and their identification*, in Probabilistic and Causal Inference: The Works of Judea Pearl, 2022, pp. 317–372.

[164] J. PEARL AND D. MACKENZIE, *The book of why: the new science of cause and effect*, Basic books, 2018.

[165] F. PEDREGOSA, G. VAROQUAUX, A. GRAMFORT, V. MICHEL, B. THIRION, O. GRISEL, M. BLONDEL, P. PRETTENHOFER, R. WEISS, V. DUBOURG, J. VANDERPLAS, A. PASSOS, D. COURNAPEAU, M. BRUCHER, M. PERROT, AND E. DUCHESNAY, *Scikit-learn: Machine learning in Python*, Journal of Machine Learning Research, 12 (2011), pp. 2825–2830.

[166] T. PENZEL, G. B. MOODY, R. G. MARK, A. L. GOLDBERGER, AND J. H. PETER, *The apnea-ecg database*, in Computers in Cardiology 2000. Vol. 27 (Cat. 00CH37163), IEEE, 2000, pp. 255–258.

[167] F. PILLING AND P. COULTON, *Forget the singularity, its mundane artificial intelligence that should be our immediate concern*, The Design Journal, 22 (2019), pp. 1135–1146.

[168] S. PIRES, S. MATHUR, R. A. GARCÍA, J. BALLOT, D. STELLO, AND K. SATO, *Gap interpolation by inpainting methods: Application to ground and space-based asteroseismic data*, Astronomy & Astrophysics, 574 (2015), p. A18.

[169] S. PURUSHOTHAM, C. MENG, Z. CHE, AND Y. LIU, *Benchmarking deep learning models on large healthcare datasets*, Journal of biomedical informatics, 83 (2018), pp. 112–134.

[170] C. QI, Y. WANG, W. WU, AND X. WANG, *Short-term predictions and lime-based rule extraction for standard and poor's index*, in Data Science: 6th International Conference of Pioneering Computer Scientists, Engineers and Educators, ICPCSEE 2020, Taiyuan, China, September 18-21, 2020, Proceedings, Part II 6, Springer, 2020, pp. 329–343.

[171] Z. QI, S. KHORRAM, AND F. LI, *Visualizing deep networks by optimizing with integrated gradients.*, in CVPR Workshops, vol. 2, 2019, pp. 1–4.

[172] C. A. RATANAMAHATANA AND E. KEOGH, *Making time-series classification more accurate using learned constraints*, in Proceedings of the 2004 SIAM international conference on data mining, SIAM, 2004, pp. 11–22.

[173] G. D. P. REGULATION, *Art. 22 gdpr. automated individual decision-making, including profiling*, Intersoft Consulting, https://gdpr-info. eu/art-22-gdpr, (2020).

[174] M. T. RIBEIRO, S. SINGH, AND C. GUESTRIN, *" why should i trust you?" explaining the predictions of any classifier*, in Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144.

[175] ——, *Model-agnostic interpretability of machine learning*, arXiv preprint arXiv:1606.05386, (2016).

[176] M. ROBNIK-ŠIKONJA AND I. KONONENKO, *Explaining classifications for individual instances*, IEEE Transactions on Knowledge and Data Engineering, 20 (2008), pp. 589–600.

[177] A. ROSENFELD, *Better metrics for evaluating explainable artificial intelligence*, in Proceedings of the 20th international conference on autonomous agents and multiagent systems, 2021, pp. 45–50.

[178] J. ROŽANEC, E. TRAJKOVA, K. KENDA, B. FORTUNA, AND D. MLADENIĆ, *Explaining bad forecasts in global time series models*, Applied Sciences, 11 (2021), p. 9243.

[179] C. RUDIN, *Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead*, Nature Machine Intelligence, 1 (2019), pp. 206–215.

[180] A. SAADALLAH, *Explainable adaptation of time series forecasting*.

[181] S. SAITO, E. CHUA, N. CAPEL, AND R. HU, *Improving lime robustness with smarter locality sampling*, arXiv preprint arXiv:2006.12302, (2020).

[182] D. SALINAS, V. FLUNKERT, J. GASTHAUS, AND T. JANUSCHOWSKI, *Deepar: Probabilistic forecasting with autoregressive recurrent networks*, International Journal of Forecasting, 36 (2020), pp. 1181–1191.

[183] R. SALUJA, A. MALHI, S. KNAPIČ, K. FRÄMLING, AND C. CAVDAR, *Towards a rigorous evaluation of explainability for multivariate time series*, arXiv preprint arXiv:2104.04075, (2021).

[184] W. SAMEK, A. BINDER, G. MONTAVON, S. LAPUSCHKIN, AND K.-R. MÜLLER, *Evaluating the visualization of what a deep neural network has learned*, IEEE transactions on neural networks and learning systems, 28 (2016), pp. 2660–2673.

[185] U. SCHLEGEL, H. ARNOUT, M. EL-ASSADY, D. OELKE, AND D. A. KEIM, *Towards a rigorous evaluation of xai methods on time series*, in 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), IEEE, 2019, pp. 4197–4201.

[186] M. SCHMITZ, S. SODERLAND, R. BART, O. ETZIONI, ET AL., *Open language learning for information extraction*, in Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning, 2012, pp. 523–534.

[187] M. SETZU, R. GUIDOTTI, A. MONREALE, F. TURINI, D. PEDRESCHI, AND F. GIANNOTTI, *Glocalx-from local to global explanations of black box ai models*, Artificial Intelligence, 294 (2021), p. 103457.

[188] L. S. SHAPLEY, *A value for n-person games*, Classics in game theory, 69 (1997).

[189] A. SHRIKUMAR, P. GREENSIDE, AND A. KUNDAJE, *Learning important features through propagating activation differences*, in International conference on machine learning, PMLR, 2017, pp. 3145–3153.

[190] H. A. SIMON, *What is an "explanation" of behavior?*, Psychological science, 3 (1992), pp. 150–161.

[191] T. SIVILL, *Ethical and statistical considerations in models of moral judgments*, Frontiers in Robotics and AI, 6 (2019), p. 39.

[192] T. SIVILL AND P. FLACH, *Limesegment: Meaningful, realistic time series explanations*, in International Conference on Artificial Intelligence and Statistics, PMLR, 2022, pp. 3418–3433.

[193] ——, *Shapley sets: Feature attribution via recursive function decomposition*, arXiv preprint arXiv:2307.01777, (2023).

[194] D. SLACK, S. HILGARD, E. JIA, S. SINGH, AND H. LAKKARAJU, *Fooling lime and shap: Adversarial attacks on post hoc explanation methods*, in Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 2020, pp. 180–186.

[195] K. SOKOL AND P. FLACH, *Explainability fact sheets: a framework for systematic assessment of explainable approaches*, in Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020, pp. 56–67.

[196] ——, *Interpretable representations in explainable ai: From theory to practice*, arXiv preprint arXiv:2008.07007, (2020).

[197] ——, *One explanation does not fit all: The promise of interactive explanations for machine learning transparency*, arXiv preprint arXiv:2001.09734, (2020).

[198] K. SOKOL, A. HEPBURN, R. SANTOS-RODRIGUEZ, AND P. FLACH, *blimey: surrogate prediction explanations beyond lime*, arXiv preprint arXiv:1910.13016, (2019).

[199] J. STAUDACHER AND J. ANWANDER, *Conditions for the uniqueness of the gately point for cooperative games*, arXiv preprint arXiv:1901.01485, (2019).

[200] E. ŠTRUMBELJ AND I. KONONENKO, *Explaining prediction models and individual predictions with feature contributions*, Knowledge and information systems, 41 (2014), pp. 647–665.

[201] P. STURMFELS, S. LUNDBERG, AND S.-I. LEE, *Visualizing the impact of feature attribution baselines*, Distill, 5 (2020), p. e22.

[202] Y. SUN, M. KIRLEY, AND S. K. HALGAMUGE, *Extended differential grouping for large scale global optimization with direct and indirect variable interactions*, in Proceedings of the 2015 annual conference on genetic and evolutionary computation, 2015, pp. 313–320.

[203] Y. SUN, M. KIRLEY, AND S. K. HALGAMUGE, *A recursive decomposition method for large scale continuous optimization*, IEEE Transactions on Evolutionary Computation, 22 (2017), pp. 647–661.

[204] Y. SUN AND M. SUNDARARAJAN, *Axiomatic attribution for multilinear functions*, in Proceedings of the 12th ACM conference on Electronic commerce, 2011, pp. 177–178.

[205] M. SUNDARARAJAN AND A. NAJMI, *The many shapley values for model explanation*, in International conference on machine learning, PMLR, 2020, pp. 9269–9278.

[206] M. SUNDARARAJAN, A. TALY, AND Q. YAN, *Axiomatic attribution for deep networks*, in International conference on machine learning, PMLR, 2017, pp. 3319–3328.

[207] M. TAJGARDOON, M. J. SAMAYAMUTHU, L. CALZONI, AND S. VISWESWARAN, *Patient-specific explanations for predictions of clinical outcomes*, ACI open, 3 (2019), pp. e88–e97.

[208] P. TAMAGNINI, J. KRAUSE, A. DASGUPTA, AND E. BERTINI, *Interpreting black-box classifiers using instance-level visual explanations*, in Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics, ACM, 2017, p. 6.

[209] S. TESO AND K. KERSTING, *Explanatory interactive machine learning*, in Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, 2019, pp. 239–245.

[210] W. THOMSON, *The shapley value, a crown jewel of cooperative game theory*, in Handbook of the Shapley Value, Chapman and Hall/CRC, 2019, pp. 1–15.

[211] H.-C. THORSEN-MEYER, A. B. NIELSEN, A. P. NIELSEN, B. S. KAAS-HANSEN, P. TOFT, J. SCHIERBECK, T. STRØM, P. J. CHMURA, M. HEIMANN, L. DYBDAHL, ET AL., *Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records*, The Lancet Digital Health, 2 (2020), pp. e179–e191.

[212] S. H. TIJS AND T. S. DRIESSEN, *Game theory and cost allocation problems*, Management science, 32 (1986), pp. 1015–1028.

[213] S. TONEKABONI, S. JOSHI, M. D. MCCRADDEN, AND A. GOLDENBERG, *What clinicians want: contextualizing explainable machine learning for clinical end use*, in Machine learning for healthcare conference, PMLR, 2019, pp. 359–380.

[214] M. VILLANI, J. LOCKHART, AND D. MAGAZZENI, *Feature importance for time series data: Improving kernelshap*, arXiv preprint arXiv:2210.02176, (2022).

[215] G. VISANI, E. BAGLI, F. CHESANI, A. POLUZZI, AND D. CAPUZZO, *Statistical stability indices for lime: obtaining reliable explanations for machine learning models*, Journal of the Operational Research Society, (2020), pp. 1–11.

[216] S. WACHTER, B. MITTELSTADT, AND C. RUSSELL, *Counterfactual explanations without opening the black box: Automated decisions and the gdpr*, Harv. JL & Tech., 31 (2017), p. 841.

[217] D. WALTON, *Dialogical models of explanation.*, ExaCt, 2007 (2007), pp. 1–9.

[218] T. WANG, Z. BUÇINCA, AND Z. MA, *Learning interpretable fair representations*, tech. rep., Technical report, Technical report, Harvard University, 2021.

[219] Z. WANG, Y. HUANG, D. SONG, L. MA, AND T. ZHANG, *Deepseer: Interactive rnn explanation and debugging via state abstraction*, in Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, 2023, pp. 1–20.

[220] Z. WANG, W. YAN, AND T. OATES, *Time series classification from scratch with deep neural networks: A strong baseline*, in 2017 International joint conference on neural networks (IJCNN), IEEE, 2017, pp. 1578–1585.

[221] A. L. WASHINGTON, *How to argue with an algorithm: Lessons from the compas-propublica debate*, Colo. Tech. LJ, 17 (2018), p. 131.

[222] D. WATSON, *The rhetoric and reality of anthropomorphism in artificial intelligence*, Minds and Machines, 29 (2019), pp. 417–440.

[223] Y. WEI, M.-C. CHANG, Y. YING, S. N. LIM, AND S. LYU, *Explain black-box image classifications using superpixel-based interpretation*, in 2018 24th International Conference on Pattern Recognition (ICPR), IEEE, 2018, pp. 1640–1645.

[224] L. WEIDINGER, J. MELLOR, M. RAUH, C. GRIFFIN, J. UESATO, P.-S. HUANG, M. CHENG, M. GLAESE, B. BALLE, A. KASIRZADEH, ET AL., *Ethical and social risks of harm from language models*, arXiv preprint arXiv:2112.04359, (2021).

[225] S. WISDOM, T. POWERS, J. PITTON, AND L. ATLAS, *Interpretable recurrent neural networks using sequential sparse recovery*, arXiv preprint arXiv:1611.07252, (2016).

[226] O. YALCIN, X. FAN, AND S. LIU, *Evaluating the correctness of explainable ai algorithms for classification*, arXiv preprint arXiv:2105.09740, (2021).

[227] F. YANG, Q. SUN, H. JIN, AND Z. ZHOU, *Superpixel segmentation with fully convolutional networks*, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 13964–13973.

[228] P. J. YOUNG, M. SAXENA, R. BEASLEY, R. BELLOMO, M. BAILEY, D. PILCHER, S. FINFER, D. HARRISON, J. MYBURGH, AND K. ROWAN, *Early peak temperature and mortality in critically ill patients with or without infection*, Intensive care medicine, 38 (2012), pp. 437–444.

[229] J. YU, Z. LIN, J. YANG, X. SHEN, X. LU, AND T. S. HUANG, *Generative image inpainting with contextual attention*, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 5505–5514.

[230] C. ZEDNIK AND H. BOELSEN, *Scientific exploration and explainable artificial intelligence*, Minds and Machines, 32 (2022), pp. 219–239.

[231] M. D. ZEILER AND R. FERGUS, *Visualizing and understanding convolutional networks*, in Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13, Springer, 2014, pp. 818–833.

[232] Q. ZHANG, Y. NIAN WU, AND S.-C. ZHU, *Interpretable convolutional neural networks*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8827–8836.

[233] X. ZHANG, Y. GAO, J. LIN, AND C.-T. LU, *Tapnet: Multivariate time series classification with attentional prototypical network*, in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, 2020, pp. 6845–6852.

[234] Y. ZHANG AND X. CHEN, *Explainable recommendation: A survey and new perspectives*, arXiv preprint arXiv:1804.11192, (2018).

[235] F. ZHAO, Q. HUANG, AND W. GAO, *Image matching by normalized cross-correlation*, in 2006 IEEE international conference on acoustics speech and signal processing proceedings, vol. 2, IEEE, 2006, pp. II–II.

[236] L. ZHAO, *Rewards*, (2014).

[237] Y. ZOU, R. V. DONNER, N. MARWAN, J. F. DONGES, AND J. KURTHS, *Complex network approaches to nonlinear time series analysis*, Physics Reports, 787 (2019), pp. 1–97.