Imperial College London

MEng Individual Project

Imperial College London

Department of Computing

Arguing with physical activity data from a real-world wearable clinical trial with patients with a primary brain tumour

Author:
Adam Gould

Supervisor:
Prof. Francesca Toni

Second Marker:
Dr. Francesco Belardinelli

June 19, 2023

Abstract

The objective of this project is to leverage Abstract Argumentation for Case-Based Reasoning (AA-CBR) to support clinical decision-making. The BrainWear study aims to assess the feasibility of its innovative use of wearable accelerometers to capture real-time Physical Activity (PA) data over intermittently collected Patient Reported Outcome (PRO) measures, for the assessment of patient Health-Related Quality of Life (HRQoL). This project proposes novel uses of AA-CBR as a transparent machine learning model to assess the utility of PA data from the BrainWear study, focusing on a patient-centred and transparent approach.

Objectives: Firstly, we aim to develop novel applications of AA-CBR to predict the status of patient disease with PRO and PA data. This involves developing new processes to characterise real-world medical data into interpretable representations, as well as evaluating the performance of argumentation models at this prediction task. Furthermore, we aim to assess the trade-off between constructing clinically interpretable models, achieving high performance, and the associated effort required for data characterisations. Additionally, we seek to determine if PA data can supplement or replace PRO measures and investigate the use of argumentation to support our conclusions. This will involve constructing innovative methods to identify conflicts between the two measures. We aim to create novel variants of AA-CBR that reduce the burden of feature characterisation and can handle time-based representations. By experimenting with these diverse AA-CBR variants, we seek to identify the most suitable approach for clinical decision-making.

Results: Our findings demonstrate that AA-CBR models are effective at accurately predicting the status of patient disease utilising PA and PRO data. We show that our patient-centred approach is able to characterise complex data, handle missing values and act with clinical caution. Furthermore, we show that model performance is better when utilising solely features characterised from PA data. Additionally, we present a methodology for identifying conflicts between the measures, thus allowing us to identify when individuals deviate from the population trend. We find that the best model for these predictions utilises a time component that relates to previous assessments of patient disease. Moreover, we demonstrate that our characterisations of the PA and PRO data and the explanations generated by AA-CBR models are interpretable and clinically relevant. We introduce novel AA-CBR models that reduce the burden of characterisation by being value-oriented rather than set-based characterisations found in the existing literature. Lastly, we lay the groundwork for future research into neural network-based AA-CBR models.

Acknowledgements

I would like to thank my supervisor, Professor Francesca Toni for her guidance, support, and enthusiasm throughout this entire project. I am extremely grateful to Dr Seema Dadhania for her clinical expertise, direction, and invaluable help. I'd also like to thank Guilherme Paulino Passos for his insightful explanations and helpful suggestions and Adam Dejl for his dedication and interest in my work.

I would like to thank Leah Redmond and Noor Sawhney for making fun of my Microsoft Word looking graphs.

I want to thank my family and friends for supporting me throughout my whole education.

Contents

Chapter 1

Introduction

Clinical decision-making, the process of selecting the most appropriate course of action for a patient's care, is a complex task requiring healthcare professionals to make choices based on limited information [1]. The emergence of wearable devices has introduced a novel source of longitudinal physical activity data, which holds great potential in healthcare settings. Unlike conventional approaches which rely on intermittent data collection, wearables provide continuous and real-time monitoring of an individual's physical activity [2, 3]. To assess the utility of this new data, we propose innovative uses of argumentation [4, 5] to build explainable AI models, based on real-world datasets from the BrainWear study [6, 7]. Our study presents the first Machine Learning analysis of this multi-modal data. By incorporating transparency, this approach effectively addresses reliability, ethical, and regulatory concerns surrounding medical data. The primary objective of our research is to leverage argumentation to facilitate better patient-care decision-making, by providing interpretable lines of reasoning in the prediction of disease progression.

1.1 Context

BrainWear is a clinical study that seeks to improve the management of patient Health-Related Quality of Life (HRQoL) during oncological treatment for brain tumours. High-Grade Glioma (HGG), the most common and aggressive type of brain tumour, afflicts patients in this study, and those with the most advanced progression have a median survival time of only 15 months when undergoing treatment [8]. Given these circumstances, it is crucial that HRQoL is managed effectively [9, 10].

The study collected physical activity (PA) data from patients using wrist-worn accelerometers to assess the feasibility of obtaining measures of HRQoL in real-time. This is in contrast to traditional methods that relied on periodic assessments when clinicians met with patients. These assessments are exposed to bias as they are based on patient reported outcomes (PROs) which require a subjective interpretation by clinicians. Furthermore, patient HRQoL is summarised into performance status (PS) scores which may not capture changes in HRQoL over time precisely. For example, the ECOG Performance Status scale ranges from a score of 0, meaning the patient is able to carry on with their pre-disease activity as normal, to 5, meaning the patient has died [11]. These scales can fail to capture the nuances in HRQoL and the specific experiences of a patient. The limitations associated with periodic assessments suggest the need for a more comprehensive and patient-centred approach to clinical assessments that considers a broader range of factors beyond PROs and performance status scores. Thus, more objective measures, like PA, aim to improve the reliability of HRQoL assessments and aid clinical decision-making.

Nonetheless, PA data is complex. Patients wore the accelerometers over the course of months whilst participants of the study. As a result, each patient generates a time series of PA according to different distributions that may change as the patient progresses through their cancer journey. Additionally, patients did not wear the accelerometers for the complete duration of the study, leading to gaps in the data that must be managed. Whilst the initial review of the data showed correlations between PA data and PROs, this was based on an aggregate of the data across multiple patients. Clinical decisions about an individual must look at the data on a per-patient basis and take into account conflicts between observed measurements and patient reported metrics. Hence, we propose a patient-centred approach utilising argumentation that can characterise the data collected and generate lines of reasoning that support clinical decisions.

Computational argumentation is an explainable AI (XAI) method focusing on how to structure and evaluate arguments that can be used to reason about known outcomes and make predictions or recommendations [5]. A

key principle of computational argumentation is computing which arguments are to be accepted and which do not hold up to scrutiny. Argumentation-based models are easily interpretable and generate explanations about the outcomes they argue in favour of. As a result, we use novel computational argumentation approaches to support clinical decision-making. Research has shown that clinical decision-making can be flawed, leading to sub-optimal patient outcomes, increased healthcare costs and loss of life [1, 12]. As such, there is a need for these automated approaches.

However, the adoption of black-box machine learning (ML) in healthcare raises concerns about reliability, ethics, and regulatory issues. The use of AI models in healthcare could perpetuate biases due to their nature of learning only from the data they are provided and without wider context [13]. As such, there has not been widespread adoption in clinical settings. One main reason for this is that decisions based on black-box model predictions, such as neural networks, are difficult to make as there isn't an understanding of why those predictions exist or what biases are present [14]. Moreover, patients need to be reassured that approaches to their treatment are optimal and that they can trust their doctors' recommendations [15]. There's also a legal expectation under GDPR for a "right to explanation" [16]. The concerns raised by this lack of transparency can be addressed by argumentation.

1.2 Objectives

The primary objective of this project is to utilise Artificial Argumentation for Case-Based Reasoning (AA-CBR) [17] to predict disease progression utilising PRO and PA data. This will require developing processes for characterising the complex data into interpretable representations and evaluating the performance of the argumentation models at this prediction task. Figure 1.1 showcases an example argumentation model we aim to generate, where for a focus case we predict that the patient's disease is stable and justify it by arguing which previous cases are relevant to the new case.

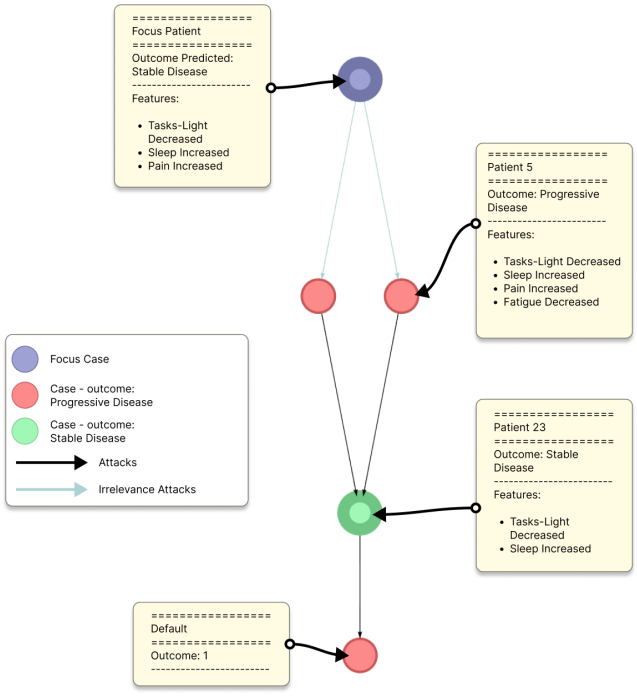


Figure 1.1: An example argumentation model with clinical data

Moreover, as this is a novel use of real-world medical data with argumentation, we aim to assess the models and data characterisation by evaluating the trade-off between building clinically interpretable models, achieving high performance and the expended effort required for these characterisations.

Furthermore, we propose using argumentation models to assess whether PA data is suitable to supplement or replace PRO measures. We aim to use the explanations derived from the argumentation model to identify clinically relevant attributes of the data. We strive to develop a novel methodology for identifying conflicts in the data between the PA data and PRO measures that can provide clinicians with additional context in their decision-making process. These original approaches to patient-centred lines of reasoning generated by argumentation will allow for the usefulness of the new data to be evaluated effectively.

1.3 Contributions

1. This study is the first to use Machine Learning methods to assess the utility of the Physical Activity data of the BrainWear study, which is novel in its use of wearable data for healthcare applications.
2. We introduce the use AA-CBR with real-world medical data to classify progressive disease. We show how interpretable models can be used to reason about new and complex data
3. We propose approaches for characterising complex PA data and PRO measures. We transform raw data into interpretable representations suitable for the AA-CBR models. We evaluate this approach showing that it can find clinically significant features in the data and utilise them for high-performing models.
4. We build on the AA-CBR literature, showcasing how to construct AA-CBR models with medical data. We extend this to create AA-CBR with Dynamic Features that can support representations of features that change over time. Additionally, we propose novel methods of AA-CBR that do not require set-based characterisations but are instead oriented on the values of the data points directly.
5. We lay the foundation for neural network-based argumentation approaches that automatically learn how to characterise the medical data without requiring a characterisation pipeline.
6. The study assesses the performance of the developed AA-CBR models in predicting disease progression with the BrainWear data. We compare the models utilising solely PA data, solely PRO measures and a combination of both PA and PRO data. This evaluation provides insights into the effectiveness of argumentation-based approaches and the clinical significance of PA data. We show that the models are comparable to or outperform baseline models and that utilising AA-CBR with Dynamic Features is particularly effective.
7. We evaluate the clinical significance of the models, reviewing the explanations generated by AA-CBR and identifying key insights from the patient-centred approach. We consider how the models can benefit clinicians' decision-making processes and compare the explanations generated against a decision-tree baseline.
8. We showcase a novel method for identifying conflicts in the data, highlighting an example where known trends in the population do not apply to individuals and laying the groundwork for utilising this method for deeper clinical analysis of the data.

1.4 Outline of Report

In Chapter 2, we outline the necessary Background and relevant research for this report. Chapter 3 describes the ethical considerations of utilising medical data for this research project. Chapter 4 is a review of the relevant data provided by the BrainWear study and details how the data is pre-processed. We explain how we will experiment with and tune the models under consideration in Chapter 5. Each model under consideration is detail in Chapter 6 and a final evaluation of all models is conducted in Chapter 7. We conclude in Chapter 8 outlining the key results of this study.

Chapter 2

Background

2.1 Explainable AI

Explainable AI (XAI) models are a form of AI that are transparent and interpretable for humans. In healthcare, XAI offers solutions to identifying bias and providing lines of reasoning that can support clinical decision-making. The use of machine learning (ML) techniques can quickly analyse large amounts of data and provide recommendations for patient care potentially improving the efficiency and accuracy of clinicians [18]. This could improve patient outcomes, reduce the costs of patient care and increase clinicians' productivity by allowing them to shift their focus to administering care rather than investigating causes of illness and searching for optimal treatments.

For example, a rule-based model unexpectedly learned from given data that patients with asthma had a lower risk of dying from pneumonia [19]. This is a correct assumption to make from the data but misses the context that these patients were often given greater care due to admission to an Intensive Care Unit (ICU). As this is a transparent model, researchers were able to see that this bias was present in their models and training data which would not be possible with opaque, black-box methods.

We provide the necessary detail of two XAI methods, decision trees and k-Nearest Neighbor for later comparison against argumentation.

2.1.1 Decision Trees

Decision Trees are an XAI model used for classification or regression tasks [20]. A decision tree represents a series of boolean choices that can be used to determine the output for a given input. The model can be used with categorical or real-valued data. Decision Tree Learning algorithms, as a supervised learning approach, construct a decision tree using labelled training data. The resulting decision tree learned is capable of approximating the classification or regression function. The tree structure and boolean decisions are simple to interpret and follow as a line of reasoning. However, the decision boundaries may appear arbitrary and are highly dependent on the algorithm used to for constructing the decision tree. Furthermore, decision trees have a tendency to overfit the training data and do not scale well to large datasets.

2.1.2 k-Nearest Neighbor

K-Nearest Neighbor is a supervised learning approach for classification and regression based on assigning the output put for a new data point to the same output or an aggregate output as the k-nearest data points [21]. A distance metric and the value of k have to be decided on. A common distance metric is the Euclidean distance, for data point $A = [x_1, y_1, z_1, \ldots]$ and $B = [x_2, y_2, z_2, \ldots]$

$$d(A, B) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2 + \ldots}$$

These models have no training time however at run-time they can take longer to generate an output as the model compares the input data point to every known data point. k-Nearest Neighbor models are easy to interpret albeit the reasoning for a given output is simplistic and not very expressive.

2.2 Artificial Argumentation

When presented with complicated, conflicting and fragmented information, drawing conclusions from the data is a challenging endeavour. One tool that humans employ is argumentational reasoning, comparing arguments that attack or defend each other in order to make sense of the information presented. Explanations that humans use to make sense of complex information can be seen as argumentative [22]. This same form of reasoning can be applied computationally given a model of argumentation that can effectively represent the underlying intentions. Fundamentally, an argument is in the form of claims and counterclaims that refute each other. To represent this computationally, researchers have identified 5 main layers of argumentation that must be addressed [5]:

- Structural Layer - describes how arguments are characterised. These could be represented using natural language, logical formulas, sets of features, etc.
- Relational Layer - describes the relationships between arguments such as which support or attack which
- Dialogical Layer - describes how arguments can be represented as dialogues for easier explanation
- Assessment Layer - describes which arguments are acceptable and to what degree
- Rhetorical Layer - describes how the objectives of an argumentation model can be designed for the audience

It is important to note that there is not a clear dividing boundary between these layers, nor a one-to-one translation between them and the models that this project will evaluate and build. However, by viewing aspects of argumentation through the lens of these five layers, we can identify where they apply in existing models and reason about how they affect the model. In this subsection, we will begin by examining an abstract view of argumentation and how to assess and compute acceptable arguments.

2.2.1 Abstract Argumentation Frameworks

Dung presents Abstract Argumentation (AA) as a general framework for computational argumentation [4]. In this framework, a claim can be accepted if all counterarguments to it are successfully refuted.

Example 1. A proponent for surgery to treat cancer states the argument "surgery is the best option for treating cancer". An opponent counters with the argument "surgery can have serious side effects". The proponent of surgery then responds "the benefits of surgery outweigh the risks of side effects". If the opponent has no further counters, the original statement is successfully defended and the proponent's arguments are accepted.

As the number of arguments grow and diverge, determining which arguments to accept becomes more difficult and a formal notion of acceptable arguments needs to be defined. To address this, Dung proposes representing arguments as abstract entities, independent of their structure, and defining a single binary relation between arguments to indicate which arguments attack which. This forms the basis of an Abstract Argumentation Framework, which is represented as the pair of a set of arguments, $Args$, and an attacks relation, $\rightsquigarrow$:

Definition 1 (Argumentation Framework adapted from [4])**.**

$$AF = \langle Args, \rightsquigarrow \rangle$$

We can define useful properties to describe an argumentation framework:

- For arguments, $a, b \in Args$, we say that
  - if $a \rightsquigarrow b$, then we say that a *attacks* b.
  - If for $a \in Args$ there is no $c \in Args$ with $c \rightsquigarrow a$, then a is *unattacked*
- For sets of arguments $E, E' \subseteq Args$ and an argument $b \in Args$, we say that:
  - E *attacks* b, denoted $E \rightsquigarrow b$, if $\exists a \in E$ with $a \rightsquigarrow b$
  - E *attacks* E', denoted $E \rightsquigarrow E'$, if $\exists b \in E'$ with $E \rightsquigarrow b$
  - E is *conflict-free* if $E \not\rightsquigarrow E$
  - E *defends* $a \in Args$ if for all $b \rightsquigarrow a$ it holds that $E \rightsquigarrow b$
  - E is *admissible* if $E \not\rightsquigarrow E$ and E *defends* all $a \in E$

Example 2. Following example 1, we can take a to represent the statement "surgery is the best option for treating cancer", b to represent "surgery can have serious side effects", c to represent "the benefits of surgery outweigh the risks of side effects".

We therefore have $Args = \{a, b, c\}, b \rightsquigarrow a$ and $c \rightsquigarrow b$.

We say that b *attacks* a and as c *attacks* b, we can say that c *defends* a. Additionally, as no argument attacks c, we have that c is *unattacked*.

Arguments and their relations can be represented graphically, where arguments are nodes and edges are attack relations.

Example 3. Consider a set of abstract arguments, $Args = \{a, b, c\}$, an argumentation framework can be represented as in Figure 2.1

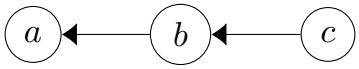


Figure 2.1: A graphical representation of an argumentation graph. The nodes represent arguments and the arrows represent attacks

Dung proposes semantics defining which sets of arguments in an argumentation framework are to be accepted. These sets are called an extension. We focus on grounded semantics, which defines the set of arguments that are successfully defended by unattacked or other defended arguments and thus, can be accepted. This can be computed by an iterative bottom-up approach starting by adding the unattacked arguments to the grounded set, $\mathbb{G}$, and then adding all the arguments that $\mathbb{G}$ defends repeatedly until $\mathbb{G}$ no longer changes.

Definition 2 (Grounded Extension adapted from [4, 17])**.** $\mathbb{G} = \bigcup_{i \geqslant 0} G_i$, where G_0 is the set of unattacked arguments and $\forall i \geqslant 0$, G_{i+1} is the set of all arguments that G_i defends.

The intuition behind grounded semantics is that the unattacked arguments are accepted by default and then we accept any arguments that are defended by these arguments. Using the Argumentation Framework in Figure 2.1, argument c is unattacked so we add this to G_0. This set only defends one argument, so a is added to G_1. As no other arguments are defended by G_1, we have the grounded extension $\mathbb{G} = \{a, c\}$.

Abstract Argumentation is an important tool in computational argumentation due to its generality in representation. It provides a foundation for extension in which the structure of arguments can be defined, argument relationships can be constructed, and dialogical arguments can be extracted. It is a versatile and powerful tool for reasoning about complex and conflicting information.

2.2.2 Abstract Argumentation for Case-Based Reasoning

Abstract Argumentation for Case-Based Reasoning (AA-CBR) [17] is a methodology inspired by case-based reasoning that defines the structure of arguments as cases.

Definition 3 (Case as defined in [17])**.** A case is a pair (X, o), with a set of features, $X \subseteq \mathbb{F}$, where $\mathbb{F}$ is an arbitrary set of features and o is one of two outcomes, $o \in \{+, -\}$.

This representation is analogous to the labelled data point representation in other machine learning disciplines. Note that the terms *case* and *argument* will be used interchangeably when referring to AA-CBR.

In Case-Based Reasoning, previously acquired knowledge is used to learn the outcome of new cases [23]. For AA-CBR, we represent this knowledge in our case base.

Definition 4 (Case Base adapted from [24])**.** Previously acquired knowledge is represented as a case base. For AA-CBR:

- A *case base* is a finite set $CB \subseteq \wp(\mathbb{F}) \times \{+, -\}$ of cases.
- A set of cases, CB, is *coherent* if for $(X, o_x), (Y, o_Y) \in CB$, if $X = Y$, then $o_X = o_Y$.

A case *attacks* another if it has a different outcome and has features more specific and relevant features. As cases are characterised by sets, we can define specificity in terms of the subset relation. A case will attack other cases that have a subset of its features. A concision condition is also enforced to ensure that cases only attack cases that are "as near as possible" i.e. subset-minimal.

Definition 5 (Attacks relation as defined in [17])**.** For $(X, o_X), (Y, o_Y) \in CB$, it holds that $(X, o_X) \rightsquigarrow (Y, o_Y)$ iff

1. $o_X \neq o_Y$, (different outcomes)
2. $Y \subsetneq X$ (specificity)
3. $\nexists (Z, o_X) \in CB$ with $Y \subsetneq Z \subsetneq X$ (concision)

The case base can be used to determine the outcome of a new case $\phi = (F_\phi, o_\phi)$, that is not in the case base.

Example 4. Extending Example 1 arguing about surgery as a form of treatment, consider a patient for whom we need to reason about if they may experience negative side effects after surgery (represented by the outcome +). The patient's cancer was detected early (feature A), the patient is otherwise healthy (feature B), is older than 65 (feature C) and has a tumour that is located in a position that is difficult to operate on (feature D). The new case would be represented as $\phi = (\{A, B, C, D\}, o_\phi)$, where the outcome, o_ϕ, represents that we do not currently know the outcome for the new case. There is also the additional condition that a patient has previously had surgery (feature E), which is not relevant to this new patient but is to a patient in the case base.

The case base represents previous cases about patients' surgeries with an outcome, +, representing if they had experienced QoL-impacting side effects and − otherwise. Consider the case base:

- $C_1 = (\{B\}, -)$
- $C_2 = (\{B, C\}, +)$
- $C_3 = (\{A, B, C\}, -)$
- $C_4 = (\{B, D\}, +)$
- $C_5 = (\{B, E\}, +)$

The case base can be visualised in Figure 2.2 using the attacks relation from Definition 5. We could intuitively think about the case base as rules. C_1 states that any patient who is otherwise healthy (B) should not have negative side effects. However, C_5 is an *exception* to this rule as it is more *specific* and has a different outcome, any patient who also has previously had surgery (E) and is otherwise healthy should expect to have side effects. We need to determine which rules apply to the new case.

Applying the grounded semantics to this case base, the arguments that we accept are $\{C_3, C_4, C_5\}$. This could provide us with any intuition on how to assign the outcome for the new case, ϕ. The outcome of the new case could be assigned the same outcome as the case that is *nearest* to it.

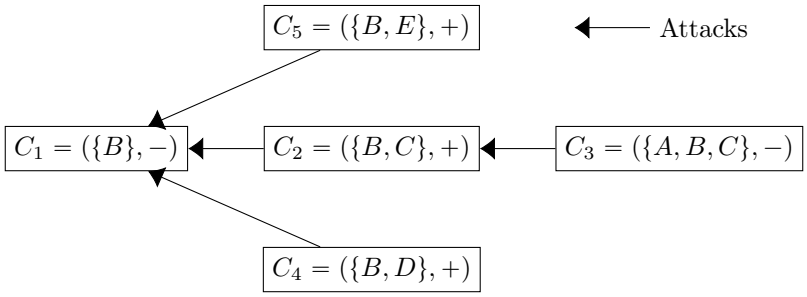


Figure 2.2: A graphical interpretation of the attacks between arguments in the case base of example 4

Definition 6 (Nearest as defined in [17])**.** For a case base CB and a new case $\phi = (F_\phi, o_\phi)$, a past case $(X, o_X) \in CB$ is nearest to the new case if $X \subseteq F_\phi$, and there is no $(Y, o_Y) \in CB$ such that $Y \subseteq F_\phi$ and $X \subsetneq Y$.

However, in this example, both C_3 and C_4 are nearest to the new case and have different outcomes. A patient, C_4, who is otherwise healthy (B) but has a tumour that is hard to operate on (D) has shown to have side effects from surgery however another patient, C_3 that is otherwise healthy (B), over 65 (C) and whose cancer was detected early (A) was shown to not have side effects from surgery - so the challenge is how to determine if the new patient, who has all of these features, could experience side effects.

AA-CBR solves this issue by introducing the *default case* (also referred to as the *default argument*). The default case is represented as $(\varnothing, \delta)$ where δ is the default outcome. The default outcome is what is expected when lacking information to make a decision for a given case and δ is set to a value based on the context. The default case can be attacked by arguments in the case base following the attacks relation in definition 5. In this example, we assume the patient will experience serious side effects so as to act sceptically when lacking information, thus, $\delta = +$ and the default argument is $(\varnothing, +)$.

The default argument is added to the argumentation framework. If it is accepted when the grounded set is computed, then the new case can be assigned to the default outcome. The arguments in the case base would have to reason that the default argument no longer holds if the new case is to be assigned an outcome that is not the default. However, there may be some cases in the case base that contain features that the new case does not have. These arguments would be irrelevant to the new case, so the new case can be added to the argumentation

framework attacking irrelevant arguments. Essentially, irrelevance means that the new case eliminates rules that do not apply to it.

Definition 7 (Irrelevance attacks as defined in [17])**.** The attacks relation is extended such that new cases attack irrelevant arguments: For $(Y, o_Y) \in CB$, $(F_\phi, o_\phi) \rightsquigarrow (Y, o_Y)$ holds iff $Y \nsubseteq F_\phi$.

Definition 8 (AA-CBR Framework as defined in [17])**.** We define the Argumentation Framework with respect to a given case base CB, a default outcome $\delta \in \{+, -\}$ and a new case ϕ as $(Args, \rightsquigarrow)$ satisfying:

- $Args = CB \cup \{(F_\phi, o_\phi)\} \cup \{(\varnothing, \delta)\}$
- For $(X, o_X), (Y, o_Y) \in CB$, it holds that $(X, o_X) \rightsquigarrow (Y, o_Y)$ iff (Definition 5)
  1. $o_X \neq o_Y$, (different outcomes)
  2. $Y \subsetneq X$ (specificity)
  3. $\nexists (Z, o_X) \in CB$ with $Y \subsetneq Z \subsetneq X$ (concision)
- For $(Y, o_Y) \in CB$, $(F_\phi, o_\phi) \rightsquigarrow (Y, o_Y)$ holds iff $Y \nsubseteq F_\phi$ (Definition 7)

Figure 2.3 shows the complete argumentation framework for this example.

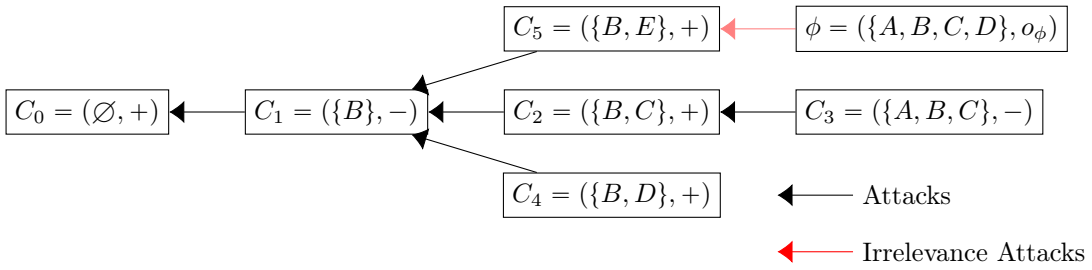


Figure 2.3: Argumentation Framework corresponding to example 4

If the default argument is accepted, then the new case is assigned the default outcome. The grounded set for this AF is $\{\phi, C_3, C_4, C_0\}$. As the default case is in the grounded set, it is an accepted argument and therefore we can assign the outcome of the new case to the default outcome, so $o_\phi = \delta = +$, i.e the new patient may expect side effects. The intuition is that the new case attacks any arguments that are irrelevant to it, automatically disqualifying them from being accepted and in the grounded set. In this example, ϕ attacks C_5 as arguments regarding having had previous surgery (feature E) are not relevant to the new case. The remaining arguments reason about whether the default argument should be accepted for the new case or not. In this example, we see it should be, so we can assign the default outcome to the new case. If case C_4 were removed from the case base, then the grounded set would be $\{\phi, C_3, C_1\}$ and as the default case is not accepted, we would expect that the new case would not experience side effects, outcome $o_\phi = -$.

Notably, AA-CBR can also be used to explain cases where the outcome is already known. Instead of new cases, we can selected focus cases which are to be explained. They can be treated the same as new cases and the argumentation framework can be built in the same way but the goal of argumentation changes from predicting an outcome to explaining it. Methods of presenting explanations from argumentation are detailed in Section 2.2.5.

Using sets to characterise cases is very expressive and works well for representing binary features that are either present or not. Nonetheless, how features are defined in this argument representation is sensitive. Consider, for example, if patients' sex was important to the outcome. If sex is represented by a single feature, say F for female, then cases with the feature F can attack cases that do not have that feature (which we may assume are cases related to males or those whose sex we do not presently know) but not the other way around. This results from the specificity condition in the definition of attacks. In essence, we have that knowing a case is about a female provides more specific information and this can act as a reason for a certain outcome. This could be desired if the patient being female has a greater effect on the outcome than if the patient is not female. For example, consider a domain arguing about outcomes related to breast cancer which disproportionately affects females [25]. Although, in other situations, this may not be the desired reasoning approach. If instead sex was represented by using M for male and F for female, a case with the feature M will not be able to attack a case with the feature F or vice versa. This could be useful when we want separate lines of reasoning about disjoint subsets of the total dataset if for example arguing in a domain about outcomes associated with testicular cancer and ovarian cancer.

On the contrary, this may not be useful when features like sex could impact the outcome but reasoning with the entire dataset is intended. For example, studies show that males may respond worse to treatment for glioblastoma than females [26], but separating the dataset by sex or only representing sex with the single feature M, could lead to missing insights about features that are more important than sex. This shows the importance of the representation of arguments and the attacks relation and how careful one must be when characterising appropriate features. Section 2.2.3 will look at methods of automatically characterising features on datasets. These issues may mean that the AA-CBR methodology unaltered is not suitable in certain contexts.

2.2.3 Argumentation Pipelines

How arguments are structured and related to one another is crucial for building a model that can represent the underlying data and effectively make arguments for the target audience. When presented with large real-world datasets, manually encoding data into cases becomes infeasible. Data-Empowered Argumentation (DEAr) is a paradigm for generating arguments based on real-world data [27].

Generalising AA-CBR

Firstly, AA-CBR can be generalised further, where instead of structuring arguments as cases that have sets of features, they can be represented by any characterisation. The attack relation does not have to be defined in terms of subsets, but any relation that allows for cases that are more specific to attack relevant cases that are less specific and for new cases to attack irrelevant cases. By defining these relations, we can create a partial order of the cases in the case base and of the new case. Selecting which partial order and irrelevance relation to use is key to defining an AA-CBR-based framework. This allows for a general structural layer and relational layer with the goal that it can more freely represent any underlying data.

Definition 9 (General AA-CBR as defined in [27])**.** Let $\mathcal{D}$ be a finite dataset consisting of labelled data points dp_i, each of the form (C_i, o_i) with C_i a characterisation of the data point and $o_i \in \mathbb{O}$, $\mathbb{O} = \{\delta, \bar{\delta}\}$ with δ the default outcome. Let dp_U be an unlabelled data point of the form C_U with C_U a characterisation. Finally, let $\succcurlyeq$ be a partial order over $\mathcal{D} \cup \{dp_U\}$ and $\prec$ a notion of irrelevance. Then an argumentation debate mined from $\mathcal{D} \cup \{dp_U\}$ is an abstract argumentation framework $(Args, \rightsquigarrow)$ with

- $Args = \mathcal{D} \cup \{(C_\delta, \delta)\} \cup \{dp_U\}$, for C_δ a characterisation of the default argument (C_δ, δ);
- for $(X, o_X), (Y, o_y) \in (D) \cup \{(C_\delta, \delta)\}$, it holds that $(X, o_X) \rightsquigarrow (Y, o_Y)$ iff
  1. $o_x \neq o_Y$, and
  2. either $X \succ Y$ and $\nexists (Z, o_X) \in (D) \cup \{(C_\delta, \delta)\}$ with $X \succ Z \succ Y$;
  3. or $X = Y$
- for $(Y, o_Y) \in (D) \cup \{(C_\delta, \delta)\}$, it holds that $dp_U \rightsquigarrow (Y, o_Y)$ iff $dp_U \prec (Y, o_Y)$.

We could therefore define the AA-CBR as characterised with sets in terms of Definition 9. We select a default argument, a partial order over the cases in the case base and an irrelevance relation as:

- $(C_\delta, \delta) = (\varnothing, \delta)$
- $\succcurlyeq = \supseteq$
- $\prec = \not\supseteq = \not\supseteq$

Specificity and Exceptionality

A condition of the attacks relation is the attacking case is relevant and more specific than the attacked case. When data points are characterised with sets of features, an attacking case is "more specific" than another when it has a superset of features. This is an intuitive notion. However, for characterisations that do not use sets, the attacking case might not contain more features than the attacked case but, for example, the attacking case might have values that are all larger than the attacked case. The attacking case could be considered an "exception" the attacked case. For this reason, the terms specificity and exceptionality are used interchangeably.

Pipeline

Secondly, the DEAr paradigm utilised a pipeline detailing steps required for argumentation on real-world data. These steps are:

1. Characterisation Extractor

2. Argumentation Debate Miner
3. Argumentation Framework
    - Predictor
    - Explainer

The characterisation extractor transforms the underlying input data into the cases that can be used with the variant of AA-CBR selected. These cases are then inputted into the Argumentation Debate Miner, along with the relations for comparing the information of cases, the irrelevance relation and the default case. A focus case or new case will also have to be identified. The Argumentation Debate Miner then outputs an argumentation framework that can be used to generate a prediction for a new case and explain the prediction or explain a known outcome of a focus case.

An example of the DEAr pipeline in practice is presented with the Artificial Neural Networks with Argumentation (ANNA) methodology [28]. In this case, a labelled dataset of examples of mushrooms was used to predict which were edible or poisonous. The feature set was made up of 126 binary features and so each example was suitable to be represented as a case in AA-CBR. However, arguing with many features can make the explanations generated more complex and many features in the dataset may not be salient for the classification. Furthermore, the dataset may not be coherent so finding which arguments to be accepted may not work.

The ANNA methodology, therefore, presents using a neural network autoencoder to do feature selection, reducing the size of the feature set used in the argumentation framework and can enforce coherence. By creating an autoencoder with a single hidden layer, the weights of the input layer can be inspected. These weights correspond to the importance of each individual feature in reconstructing the data. Once feature importance is identified, the most important features can be selected to use in the models. The autoencoder is the characterisation extractor. Then AA-CBR can be used as defined in section 2.2.2 with the output from the autoencoder. The feature selection using an autoencoder is not an intrinsically transparent model but if the accuracy of the AA-CBR predictor using the selected features is high and we can generate explanations for those predictions with AA-CBR, then we can be confident that the selected features are important for the outcome.

This pipeline is extremely versatile, providing a strong foundation for how to adapt a dataset to use argumentation. The generality of the pipeline is very useful for applying it to new contexts and different types of data. However, the characterisation extractor needs to be suited to the data and must ensure the goal of transparency and interpretability is maintained. This means that when developing an argumentation pipeline, one must be careful in choosing an appropriate characterisation extractor that does not introduce bias and allows for explanations to be understood by the intended audience.

2.2.4 AA-CBR extended with Stages

One context in which AA-CBR falters is representing data with features that can change over time. Thus an extension to AA-CBR has been created that adds stages to cases to reason about cases at different time periods [24]. We have termed this AA-CBR extended with Stages.

Definition 10 (Stages as defined in [24])**.** Stages are represented using sequences:

- Let $\mathbb{S} = \langle s_1, .., s_n \rangle$, with $n \geqslant 1$ be a finite sequence.
- Let $\langle\rangle$ denote the empty sequence
- A subsequence is of the form $\langle s_1, .., s_m \rangle$ where $m \leqslant n$
- A binary relation, initial subsequence $\sqsubseteq$, over subsequences of $\mathbb{S}$ is defined for $S, S' \in \mathbb{S}$, as $S' \sqsubseteq S$ iff either
    - $S' = \langle s_1, ..., s_k \rangle$ and $S = \langle s_1, ..., s_m \rangle$ and $k \leqslant m \leqslant n$; or
    - $S' = \langle\rangle$
- For sequences $S, S' \in \mathbb{S}$, the proper initial subsequence is $S' \sqsubsetneq S$ iff $S' \sqsubseteq S$ and $S \not\sqsubseteq S'$

This method then extends cases in definition 3 to be a triple, $C = (F, S, o)$ where S represents a subsequence of $\mathbb{S}$. The default case would be represented as $(\varnothing, \langle\rangle, \delta)$. Cases attack other cases prioritised first on features, as in the unmodified AA-CBR and then if features are the same, cases that occur later attack cases that occur earlier. The definition of an argumentation framework is as follows:

Definition 11 (AA-CBR with Stages adapted from [24])**.** The AF corresponding to a case base CB, a default outcome $\delta \in \{+, -\}$ and a new case (F_ϕ, S_ϕ, o_ϕ), is $(Args_\phi, \rightsquigarrow_\phi)$ satisfying the following conditions:

- $Args = CB \cup \{(\varnothing, \langle\rangle, \delta)\}$

- $(\varnothing, \langle\rangle, \delta)$ is called the default argument
- For $(F, S, o), (F', S', o') \in Args$, it holds that $(F, S, o) \rightsquigarrow (F', S', o')$ iff
  1. $o \neq o'$, and (different outcomes)
  2. either
     (a) $F' \subsetneq F$, and (specificity)
     (b) $\nexists (F^*, S^*, o) \in CB$ with (concision)
        - either $F' \subsetneq F^* \subsetneq F$,
        - or $F^* = F$ and $S^* \subsetneq S$
        - or $F' = F^*$ and $S' \subsetneq S^* \sqsubseteq S$
  3. or
     (a) $F' = F$ and $S' \subsetneq S$ and (advance)
     (b) $\nexists (F, S^*, o) \in CB$ with $S' \subsetneq S^* \subsetneq S$. (proximity)
- $Args_\phi = Args \cup \{F_\phi, S_\phi, o_\phi\}$
- $\rightsquigarrow_\phi = \rightsquigarrow \cup \big\{\big((F_\phi, S_\phi, o_\phi), (F, S, o)\big) : (F, S, o) \in Args \text{ and } (\text{either } F \not\subseteq F_\phi \text{ or } S \not\sqsubseteq S_\phi)\big\}$.

The difference with concision between AA-CBR and this extension is that now stages are considered and so we must ensure that attacking cases are nearest to the attacked case prioritised by features and then by stages. Additionally, irrelevance has been changed such that new cases attack irrelevant arguments that either do not have a subset of their features or are currently at a later stage.

Example 5. Consider a new example in a similar domain to Example 4, arguing about if a patient needs intervention after surgery due to experiencing side effects, with an outcome of + representing that they do need intervention and - that no intervention is needed. In the patient's case they have seen a recent decline in their mobility and energy levels (feature A), have experienced a loss of appetite and weight loss (feature B), but are experiencing less pain and discomfort (feature C) and, there's an improved appearance of the affected area (feature D). We must also consider that patients in the case base may have had an infection after surgery (feature E) but this is not relevant for our new patient. So the total feature set is $\mathbb{F} = \{A, B, C, D, E\}$

Stages can be represented as the number of weeks since surgery.
$\mathbb{S} = \langle w_1, w_2, w_3, w_4, w_5, w_6\rangle$, where $w_i = i$ weeks since surgery.

The default argument is $C_0 = (\varnothing, \langle\rangle, +)$. The new case is in their 5th week of recovery after surgery so would be represented as $\phi = (\{A, B, C, D\}, \langle w_1, w_2, w_3, w_4, w_5\rangle, o_\phi)$. The case base is:

- $C_1 = (\{D\}, \langle w_1\rangle, -)$
- $C_2 = (\{B, C, D\}, \langle w_1, w_2, w_3\rangle, +)$
- $C_3 = (\{C\}, \langle w_1, w_2, w_3, w_4\rangle, -)$
- $C_4 = (\{B, C, E\}, \langle w_1\rangle, +)$
- $C_5 = (\{B, C, D\}, \langle w_1, w_2, w_3, w_4, w_5, w_6\rangle, -)$
- $C_6 = (\{A, D\}, \langle w_1, w_2\rangle, +)$
- $C_7 = (\{B, C, D, E\}, \langle w_1, w_2\rangle, -)$

The argumentation framework for example 5 is represented in Figure 2.4. We can see how stages adapt the framework to represent dynamic features. For example, argument C_5 attacks C_2 despite having the same set of features because case C_5 has progressed further and the outcome has changed. This could suggest that features become less important to the decision of intervention if they present themselves later on or that there is some feature that is not represented in our feature set that is causing the change in outcome between C_2 and C_5. Additionally, we can see that C_5 is an irrelevant argument to the new case, ϕ as it occurs at a later stage.

The grounded semantics can be computed in the same way as AA-CBR. For this AF, we have $\mathbb{G} = \{\phi, C_6, C_2, C_0\}$ so the default argument is accepted and therefore we can argue that the new case does need intervention, so outcome +. This shows one way the AA-CBR can be extended to take into account dynamic features. For our domain, the accelerometer data in the BrainWear study is a time series and so being able to represent how features change over time and compare cases at different time scales may be helpful.

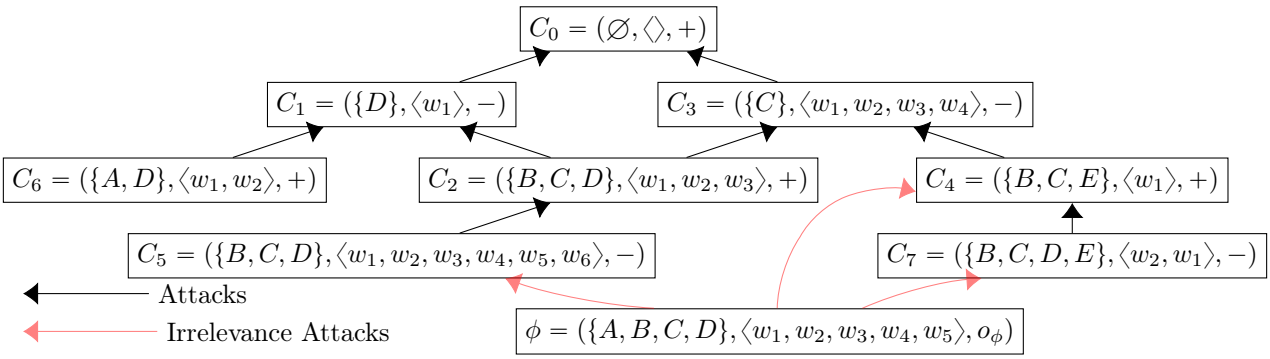


Figure 2.4: Argumentation Framework corresponding to example 5

2.2.5 Argumentative Explanations

The argumentation framework is an interpretable AI method due to its transparency and ability to be easily visualised and understood as an argument graph. This is useful for comprehending how predictions are made. However, as the number of arguments grows, visualising the entire graph hurts its readability and restricting the graph to specific sections is only useful if they are relevant to the outcome of the focus case. The framework could instead by used to generate explanations of outcomes predicted for new cases or known outcomes for focus cases. We look at two methods of explanation by argumentation: dispute trees, used to visualise dialogical explanations of an outcome and excess features, used to explain which features led to a change in outcome.

Dispute Trees

An argument can be thought of as a dialogue between two agents, one that debates in favour of a certain outcome and another to the contrary. In AA-CBR, the argument would begin with the default case, $(\varnothing, \delta)$ and ends when there are no arguments left in this dispute. This can be represented in a dispute tree, which shows the arguments that can be formed between a winner, W, who argues in favour of the focus case's outcome and a loser, L, arguing against the outcome of the focus case. The winner will make the final argument, as it is unattacked, they win the dispute. A dispute tree, $\mathcal{T}_1$ for Example 5, is shown in Figure 2.5.

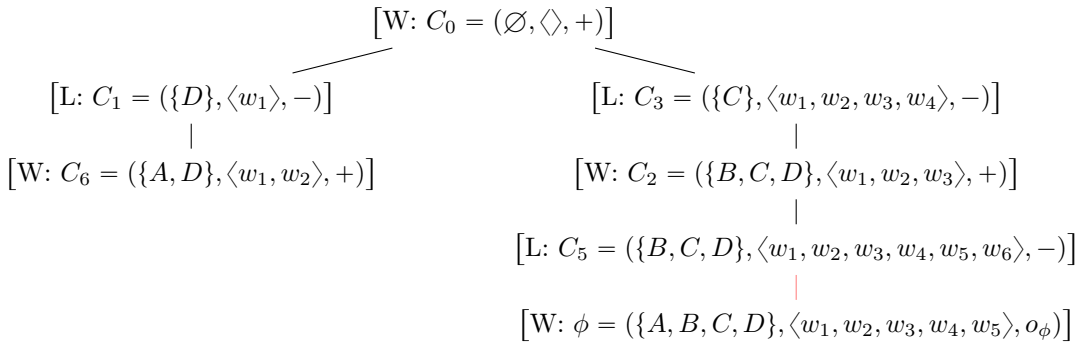


Figure 2.5: Dispute Tree $\mathcal{T}_1$ generated from Argumentation Framework in Example 5.

The dispute tree is a dialogical explanation that can be understood in words. Firstly, W, claims that by default, the new patient should be given early intervention. L then counters this claim by stating that there was a case in which a patient showed an improved appearance of the affected area and did not need intervention. W disagrees, citing the case of a patient who also showed an improved appearance but had seen a recent decline in mobility and they did need intervention. L has no counterarguments to this claim, however, they do have another counter to the default argument stating that there was a case who had not needed intervention after showing less pain and discomfort. W counters with a case of a patient who needed intervention that exhibited weight loss in addition to less pain and discomfort and an improved appearance of the affected area. L argues with the case of another patient expressing the same symptoms and improvements that had not needed intervention. This case was at 6 weeks after surgery. W replies by stating that L's argument is irrelevant to the new case as the new case has only had 5 weeks of recovery since surgery. L has no counters and thus W wins the dispute. There are other dispute trees that could be generated from the argumentation framework as both L and W could make different arguments for some of the claims each proposes and these could provide further explanations for why W wins the dispute and the new patient should get intervention. Dispute Tree, $\mathcal{T}_2$, in Figure 2.6, shows another

such tree. A dispute tree can be defined formally as:

Definition 12 (Arbitrated Dispute Tree as defined in [24])**.** Let $AF = (Args_\phi, \rightsquigarrow_\phi)$. An arbitrated dispute tree is a tree $\mathcal{T}$ such that:

1. every node of $\mathcal{T}$ is of the form $[N : x]$ for $N \in \{W, L\}$ and $x \in Args_\phi$: the node is called N-node labelled by argument x;
2. the root of $\mathcal{T}$ is labelled by argument $(\varnothing, \langle\rangle, \delta)$ and is
   - a W-node, if $o_\phi = \delta$
   - a L-node, if $o_\phi \neq \delta$;
3. for every W-node n labelled by some $b \in Args_\phi$, and for every $c \in Args_\phi$ such that $c \rightsquigarrow_\phi b$, there exists a child of n, which is an L-node labelled by c;
4. for every L-node n labelled by some $b \in Args_\phi$, there exists exactly one child of n which is an W-node labelled by some $c \in Args_\phi$ such that $c \rightsquigarrow_\phi b$;
5. there are no other nodes in $\mathcal{T}$ except those given by 1-4.

Dispute trees are vital to understanding complex argumentation frameworks. The dialogical explanations that can be derived from the dispute trees provide a clear explanation of the outcome. However, presented without a translation into words, the dispute trees can be harder to interpret, especially once the number of features and different types of features increases. We thus also look at excess features as another method of determining why an outcome occurred.

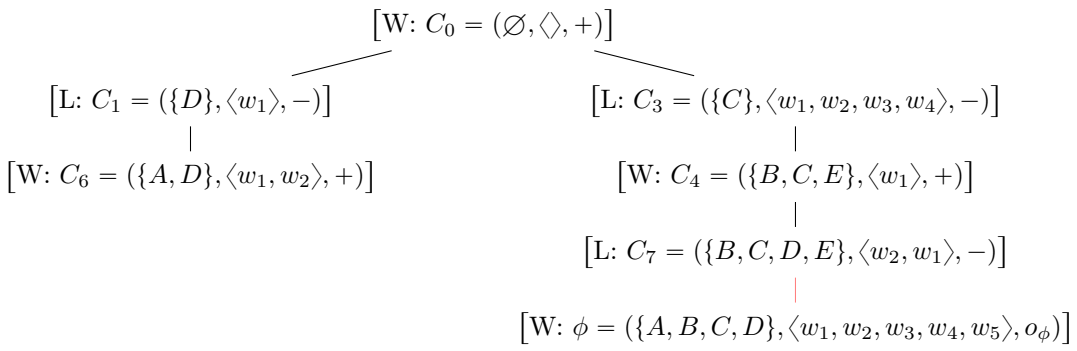


Figure 2.6: Dispute Tree, $\mathcal{T}_2$, generated from Argumentation Framework in example 5

Excess Features

Dispute trees can also be used to identify *excess features* which are features that the winner, W, identifies as making the loser's, L, claim irrelevant. In dispute tree $\mathcal{T}_2$, Figure 2.6, L attacks W with case C_7, but W replies that this case is not relevant and excess features identify that it's because the new case, ϕ did not experience an infection (feature E). This helps explains why some claims by L do not hold. Note that in $\mathcal{T}_1$, W wins by attacking C_5 which is irrelevant due to being at a later stage. A notion of excess stages could be defined if the passage of time is what causes a change in outcome. For the domain used in the literature, instead of excess stages, it is stated that changes due to time progression indicate there are unknown features gained by a case over time that caused the outcome change.

Formally, excess features can be defined as:

Definition 13 (Excess Features as defined in [24])**.** Given an arbitrated dispute tree $\mathcal{T}$, the excess features are given by the set

$\mathcal{F} = \bigcup \{F \backslash F_\phi : [W : (F_\phi, S_\phi, o_\phi)]$ is a leaf in $\mathcal{T}$ with parent $[L : (F, S, o)]\}$

If (F_ϕ, S_ϕ, o_ϕ) does not label any node in $\mathcal{T}$ then $\mathcal{F} = \varnothing$

2.2.6 Cumulative AA-CBR

A "regular" AA-CBR framework is one in which the irrelevance relation in as a negation of the partial order relation i.e. $\prec = \not\succcurlyeq$ and the default argument is always the least argument with respect to $\succcurlyeq$ in the dataset. This allows us to more easily reason about the properties of an AA-CBR framework. AA-CBR as originally proposed is regular however the AA-CBR with Stages proposed in the literature is not.

It has been shown that a regular AA-CBR has two properties, that may be undesirable: it is not cautiously monotonic and is noise intolerant [29].

Example 6. To illustrate non-monotonicity, we can utilize a similar example to the literature [29]. Consider a simple AA-CBR framework with two cases in the case base, the default $(\varnothing, +)$, where + represents the outcome that a patient requires intervention and the case $(\{A\}, -)$ where feature A signifies a patient decline in mobility. Given a new case in which we aim to classify $(\{A, B\}, ?)$, where feature B represents the patient's MRI scan shows stable disease. In this example, the default argument is attacked by $(\{A\}, -)$ which itself is unattacked. This means the default outcome is not in the grounded extension. As a result, the new case is classified with a $-$. Thus no features, except A, has any bearing on the classification, due to the limited cases in the case base. The presence of feature A alone is adequate to classify a new case as $-$. However, if $(\{A, B\}, +)$ is added to the case base, any case that contains both A and B will be classified as $+$. A alone is no longer sufficient to classify a new class.

Example 7. To demonstrate noise intolerance, consider an incoherent case base consisting of $(\varnothing, +)$, $(\{A\}, -)$ and $\{A\}, +)$. In this scenario, the default argument is not in the grounded extension so the new case $(\{A, B\}, ?)$ is classified as $-$. Curiously, despite the absence of any attacking cases within the grounded extension, the default argument remains unaccepted. This situation appears contradictory since there is no argument that effectively challenges the default argument, yet it fails to hold true. This contradiction arises as a consequence of the incoherence present within the cases comprising the case base.

These two properties prove useful to examine given that this project is working with real-world data and will have to handle the fact that the data may include incoherent cases and monotonicity may be a desired property if it improves the classification power of models built.

The proposed solution is to build a Cumulative AA-CBR, denoted cAA-CBR, designed to make AA-CBR cautiously monotonic and as a result, can handle incoherence. The main idea behind a cAA-CBR is to restrict the training dataset of the classifier to a concise subset. Intuitively, a concise subset is the minimum set required to be able to classify every input data point. Removing any data point from the concise subset removes the classifier's ability to classify that particular data point and other data points not in the concise subset. Adding any data point to the concise subset that is not present in the original dataset can increase the number of data points the classifier can classify but does not change any previous data points' classification - it is cautiously monotonic.

The literature provides an algorithm for how to restrict the dataset to the concise subset. The algorithm provided solves the issue of incoherence as for any two incoherent cases, only one will be included. For our purposes, it is enough to know that the algorithm will build a cAA-CBR from a provided regular AA-CBR and will have the properties of cautious monotonicity and noise tolerance.

2.2.7 Argumentation in Healthcare

Argumentation has been applied in healthcare in the past. For example, whilst not utilising AA-CBR, a system for aggregating evidence from clinical trials has been previously developed. This can take into account specific preferences and can argue with easy-to-follow lines of reasoning whether one treatment is preferred over another [30]. This approach can take into account outcomes from the treatments, such as mortality rates or the likelihood of cancer to aid clinicians. This shows the potential of argumentation in a healthcare setting, motivating its use for providing clinicians with additional context necessary for decision-making.

2.3 Neural Networks

A neural network is an AI model that consists of interconnected nodes that simulate biological neurons. It is a supervised learning approach that uses highly mathematical algorithms to iteratively adjust connections and weights between neurons, thus, allowing it to automatically learns how to produce a required output from a given input [31]. The input for each neuron comes from the neurons of the previous layer. Each neuron does the following calculation:

$$y = f\left(\sum_{i=1}^{n} w_i \cdot x_i + b\right)$$

where:

$$\begin{aligned} y &: \text{output of the neuron} \\ f &: \text{activation function} \\ w_i &: \text{weight of the input} \\ x_i &: \text{input to the neuron} \\ b &: \text{bias term} \\ n &: \text{number of inputs} \end{aligned}$$

The strength of neural networks comes from their ability to learn the weights and biases that enable approximations of a required function, such as classifying inputs or doing regression tasks. The learning process involves performing a forward pass through the network, measuring the error of the output compared to the true values (referred to as the loss), and backpropagating the gradient of the loss to update the weights and biases. The activation functions allow for the neural network to learn more complicated functions by introducing non-linearity. Additionally, the activation function ensures that the output of a neuron is propagated forward only if it has reached an appropriate threshold. (Refer to [31] for more details on common activation functions, loss functions and training algorithms).

Despite the impressive capabilities of neural networks, their representations are not easily interpretable. The learned parameters, the weights and biases, lack meaningful significance for individuals reviewing the model. Furthermore, the complex and highly mathematical training process makes understanding how the model learns difficult or impossible to comprehend. Explanations cannot be generated from neural networks intrinsically, only as a post-hoc explanation of the output. Whilst some visualisations or interpretations inspect individual layers of a neural network, often there is no real mapping from the representations generated by the hidden layers to any real-world features. These drawbacks are why neural networks are considered black-box models, they are inherently opaque in nature.

2.3.1 Autoencoders

Autoencoders [32] are a neural network architecture commonly used for reducing the dimensionality of the input data. These models are comprised of an encoder network and a decoder network that are trained together. The encoder portion of the autoencoder contains a final layer that outputs a transformed representation of the input data. The decoder portion of the autoencoder begins from the output of the encoder and finishes with a final layer with the same number of input features. The complete autoencoder is trained by feeding data through the encoder and then the decoder and computing the loss between the original input and the output of the network. This loss is then propagated through both the decoder and encoder. Once trained, the encoder portion of the network is able to generate a condensed representation of the original data. The decoder portion of the network is trained to reconstruct the original input from the condensed representation. Autoencoders are useful for dimensionality reduction or identifying feature importance.

Chapter 3

Ethics

The ethical implications of ML in healthcare have to be carefully considered. The sensitivity of the data being handled, fairness in patient treatment and, the accuracy and transparency of any predictions made by an ML system must be appropriately addressed. We must strictly adhere to BrainWear's data processing and ethical commitments. Therefore we review ethical considerations and how to mitigate harmful impacts.

Medical data is personal and patients must trust that their data is being handled correctly and confidentially for effective patient-clinician communication. Properly handling sensitive and personally identifiable information is legally required under GDPR [33]. As such, the data provided by the BrainWear study is a pseudonymised dataset from participants who have consented to the data being used for research purposes. No steps to de-anonymise the data will be taken. Access to the data will be restricted from third parties. Graphs and figures illustrating patient data will be augmented so as not to share patient data. Patients were assigned IDs in the study, these will be replaced with randomly assigned case numbers in the figures presented. Data will be exclusively held and processed on Imperial's Research Data Store (RDS) and High Performance Computing (HPC) systems [34]. Access to patient data on the RDS is restricted by BrainWear project owners. Any computation output must also be kept on these systems.

The provided data is only collected during the period that participants' consented and no further attempts by this project to collect more will occur, ensuring only data participants have consented to can be used. Despite the continuous and real-time nature of the data collection methods, no live-tracking of patients occurred during this project, data is provided after the fact.

All ML methods can present bias, wherein subgroups of the population are not well represented by the model compared to a larger majority of the population. This can occur based on a lack of appropriate data that under-represents minority populations, the feature set selected in the ML pipeline ignores or prioritises attributes that would better represent a certain group or the choice of outcome is more likely to favour certain groups [35]. This project uses medical records that include sex, age and disability which are protected characteristics [36], so we must ensure the models built treat participants fairly or that any unfairness and bias can be exposed.

As data is provided by the BrainWear clinical trial, there is no option to increase the size of the dataset. The data provided is not large enough to augment or balance the data based on certain characteristics. Additionally, the data does not include details about some protected characteristics such as race, so uncovering biases with regard to this is challenging. However, the project benefits from building explainable AI models that can allow those using the models to review outcomes and provide their own insights into whether the explanations generated are fair or contain bias. By increasing the transparency of the model, we allow biases to be exposed so that they can be more easily corrected.

By developing an explainable AI model, using argumentation, we aim to have a positive ethical impact by making it so clinicians could exploit the power of ML methods but without the potential impacts of black-box models. However, despite building transparent models, we must consider their interpretability and explainability. The benefit of using explanations is that they are interpretable and so clinicians can use their expertise and insight to decide if the explanations are reliable. Moreover, we will ensure that we properly evaluate our models - see Chapter 7 - comparing them against baseline models and getting feedback from clinicians. In doing so, we can improve the reliability of the frameworks developed and hope to achieve our positive impacting goal.

Chapter 4

Data

Data provided by the BrainWear study can be categorised as follows: patient characteristics, PRO questionnaire results, Physical Activity Data and Brain MRIs. We reviewed the data to ensure sufficient quality for the use of predicting disease status. Data from 79 patients have been provided. This has been restricted to focus on patients who have high-grade glioma, have provided good quality PA data, provided two or more PRO questionnaires and have had at least one MRI scan. As a result, there is data from 31 patients that meet this criteria. We restrict to high-grade glioma patients as tracking their quality of life is of particular clinical interest due to the severity of their disease. The data provided can be fit into two main categories, Patient Reported Outcomes and Physical Activity Data both of which require pre-processing to be used effectively. Additionally, classifications of MRI scans were provided to identify patient status.

4.1 Patient Reported Outcomes

PRO questionnaires are standardised instruments given to each patient at regular intervals to assess their quality of life from many different perspectives. We focus on the European Organization for the Research and Treatment of Cancer Quality of Life Questionnaire (EORTC QLQ-C30) questionnaire with the additional brain tumour-specific BN20 module [37].

EORTC QLQ-C30 gives patients 30 questions to track symptoms, quality of life and, physical and emotional functioning. QLQ-C30 was given with the BN20 module which used a further 20 questions to additionally track patient symptoms such as visual disorder, headaches, and motor dysfunction. A full breakdown of the EORTC scoring scales can be found in Table A.1. Of the 31 patients included, 245 total questionnaires were returned with the average number of EORTC questionnaires collected per patient at 7.9 questionnaires.

4.1.1 Pre-Processing

The raw scores of each of the 50 questions were provided. In order to group the data and provide clinical relevance, the scores of each question were aggregated into a relevant scale. For QLQ-C30 there are 5 different functional scales (physical functioning, role functioning, emotional functioning, cognitive functioning and social functioning), 9 different symptom scales (Including fatigue, pain and insomnia), and one global health status/QoL measure. BN20 tracked an additional 11 scales focused on additional symptoms experienced by brain tumour patients. Aggregating questions to their relevant scale was done according to the provided EORTC scoring manual [38]. For a given scale we calculate the average of the scores for the questions related to the scale and then do a linear transformation such that each scale is measured from 1-100.

4.2 Physical Activity Data

Each patient in the study was given an Axivity AX3 triaxial wearable accelerometer sampling acceleration at 100 times per second (100Hz) in the X, Y and Z axes. This provides a longitudinal time series of accelerations for every patient on the study. Whilst patients were consenting and fit enough to participate, they could wear the accelerometers continuously. However, not all patients wore the accelerometers all the time, leading to data missing at different intervals. Of the 31 patients considered, each was on the study for an average of 331 days, but only 45.34% of those days had at least 50% of usable PA data collected. This gives 4661 days of data in total to consider across all of the patients.

4.2.1 Pre-Processing

The raw data was processed according to the UK Biobank Accelerometer Analysis pipeline, providing a single average acceleration for every non-overlapping 30-second time period [39, 40, 41, 42]. The pipeline additionally uses balanced random forests followed by a Hidden Markov Model to generate a classification of the functional behaviour (sleep, sedentary, moderate, tasks-light, walking) during the 30-second epoch. Table 4.1 showcases a small example snapshot of the data after it has been pre-processed. We can visualise the data graphically in figure 4.1.

Time	Acc (mg)	Sleep	Sedentary	Moderate	Tasks-light	walking
20/09/19 15:09:00	64.9	0	1	0	0	0
20/09/19 15:09:30	25	0	1	0	0	0
20/09/19 15:10:00	30.2	0	0	0	0	1
20/09/19 15:10:30	80.1	0	0	0	0	1

Table 4.1: Example snapshot of the PA data after pre-processing

Raw data files where the wear time of the accelerometer is less than 3 days are removed as these files are not considered to be good quality wear time. It is important to note that these models were pre-trained on a subset of participants of the UK Biobank study. The UK Biobank study is comprised of patients with varying HRQoL and illnesses and thus the models were not strictly trained on patients with high-grade glioma, whose activity distribution likely follows a different distribution of healthy patients [6].

Of the 31 patients used, a total of 5,439 hours was classified as moderate activity, 52,956 hours as sedentary activity, 56,329 hours as sleeping, 2,993 hours doing light tasks and 4,831 hours as walking.

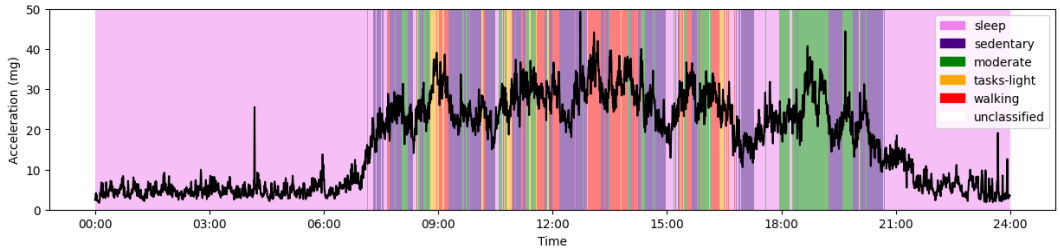


Figure 4.1: Physical activity data example of a single day with classification of functional behaviour at each 30 second epoch

4.3 Identifying Patient Status

In the study, periodic MRI scans were conducted for each of the 31 considered patients. The purpose of these scans was to track the progression of their diseases. The data from the MRI scans were consolidated into a single statistic, indicating whether a patient's disease exhibited progressive or stable characteristics. On average, each patient had three MRI outcomes available for analysis, and in total 50 scans showed an outcome of progressive disease and 38 showed stable disease.

At the time when the data was provided, it was observed that 14 patients had passed away. Their deaths represent important outcomes within the context of the study, as they highlight the impact and gravity of the disease under investigation.

The dataset consisting of the MRI scan results, along with the information on patient deaths, contributes valuable information regarding the progression of the cancers being studied. This data can be utilized to explore the predictive capabilities of PRO and PA measures with regard to patient disease and HRQoL.

Chapter 5

Model Objectives and Experiment Design

Our objective is to develop an argumentation model and pipeline that can accurately predict the status of patient disease utilising patient characteristics, physical activity data and patient reported outcome measures. This novel application of AA-CBR with real-world medical allows us to explore a range of literature-based models to find an optimal process for classification through argumentation. Additionally, we propose original models aimed at reducing the burden of characterisation extraction required by conventional AA-CBR approaches.

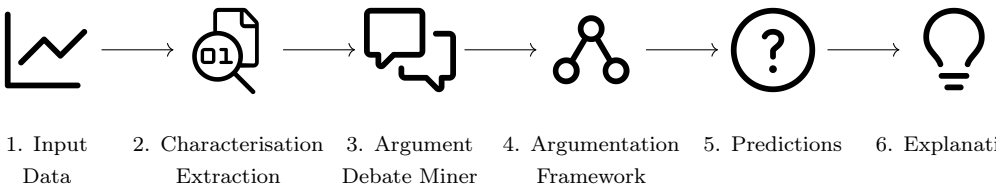


Figure 5.1: Illustration of the general argumentation pipeline. The input data undergoes characterisation extraction, followed by integration into an argumentation debate miner, yielding an argumentation framework. Subsequently, predictions are generated, and their corresponding explanations are visualized

To effectively leverage argumentation models, it is crucial to first characterise the data into cases that are clinically interpretable and contribute to the development of useful argumentation models. As a result, for each model under experimentation, we construct an argumentation pipeline that encompasses feature characterisation, the building of argumentation frameworks and, makes predictions and explanations about the state of patient disease. Figure 5.1 shows the steps involved in this pipeline.

Section 5.1 details how the data is to be used. Section 5.2 showcases the characterisation extraction methods explored. Section 5.3 explains how the models are tuned. Each argumentation model explored is based on a form of AA-CBR. Sections 6.1.1 to 6.3.2 describe the argumentation models used and highlight which characterisation extraction methods and hyperparameters perform optimally for that model.

5.1 Data Application

As the objective is to design models to predict the status of patient disease using the PA and PRO data, it is imperative to decide what aspects of the data to focus on. We must also establish a representation that is clinically interpretable, fair for comparisons and effective with an argumentation classifier.

5.1.1 Data Points

From the provided data, we can extract 110 data points where patients have completed a PRO questionnaire and have physical activity data that covers at least 50% of a time period spanning 4 weeks prior and 4 weeks after the questionnaire date. This representation was selected because the currently accepted method of assessing patient HRQoL utilises PRO measures, so we focus our experiments on the hypothesis that PA data can be used to supplement or replace existing PRO measures.

The state of the patient's disease can be measured by looking at the outcome of the next MRI scan following a questionnaire date. MRI scans show if a patient has progressive disease or stable disease. In the case where a questionnaire date has no following MRI scan, we can take the outcome as whether the patient had died or not at the time this data was collected. Thus, each data point in the model is labelled a 1 indicating that a patient either has progressive disease or has died and labelled a 0 indicating that a patient has stable disease or was alive at the time of data collection.

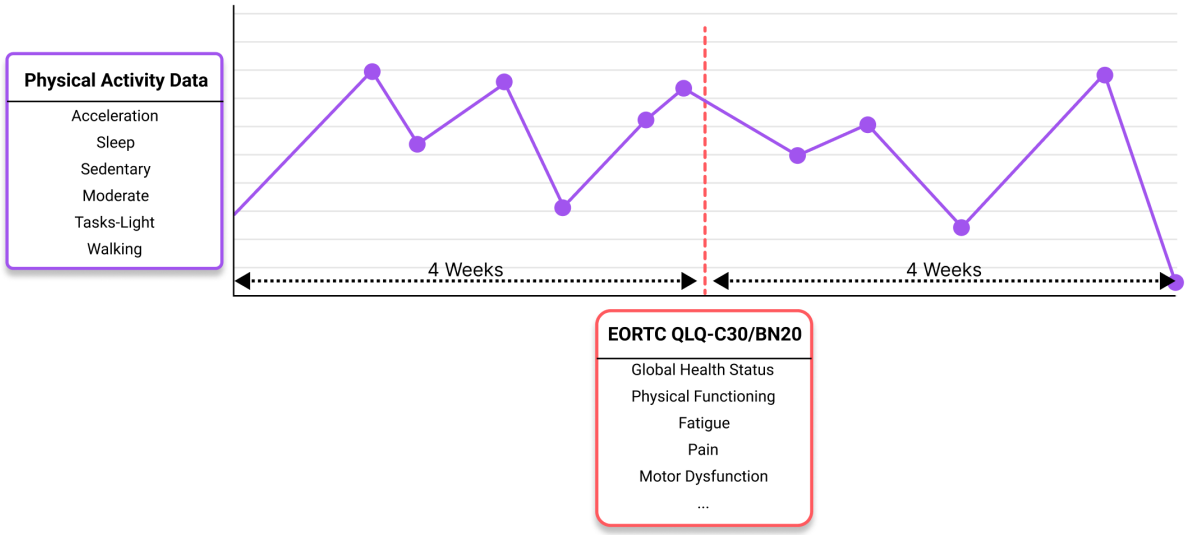


Figure 5.2: Each data point is a representation of an 8-week time series of physical activity data centered on a PRO Questionnaire

5.1.2 Data Representation

The PA data provided is a time series of accelerations (measured in mg) and a functional behaviour classification (sleep, sedentary, moderate, tasks-light, walking) at 30-second epochs. As the data is collected from 31 different patients over long periods, the distributions of the time-series data are different per patient. Additionally, as a result of patients' symptoms changing over time, for an individual patient, the distribution is not constant. This means that to fairly compare physical activity we cannot take the raw values.

To illustrate, consider patient A has a mean acceleration of 20 mg across a 1 week period and patient B has a mean acceleration of 15 mg. Superficially, one might reason that patient A is 'healthier' than patient B. However, this fails to capture changes over time and the fact that these values are generated from different distributions. The following week patient A may have a mean acceleration of 10 mg, a -50% change in acceleration whilst patient B records an acceleration of 13 mg, a -13% change in acceleration. Patient A's steeper decrease in acceleration week-on-week suggests a potentially greater cause for concern about their health compared to patient B. Hence, to enable fairer comparisons of physical activity at different time points and across patients, the PA data should be characterised as change over time rather than raw values.

For each 8-week focus period of physical activity data considered, different methods for extracting changes from PA data were considered, such as removing seasonality and extracting trend data, dividing the 8 weeks into smaller periods and encoding the full period as a series of angles between linear trend lines or using an LSTM autoencoder to learn to represent the data. These methods come with the benefit that they can accurately capture the trends in the data and can be effectively fine-tuned to work with argumentation models. However, these methods lack interpretability and intuitiveness which is incredibly important when handling medical data and working with argumentation. We have to balance the trade-off between useful clinical interpretations vs the performance of the models. This is where we consider the rhetorical layer of argumentation.

Instead, we opted to encode the PA data in the focus period by averaging the acceleration and the time spent for each of the functional behaviours and representing it as a percentage change compared to a two-week baseline. These baselines were established as the first two weeks of PA data collected from each patient.

Regarding the PRO measures, a similar approach was taken wherein change over time is a representation that is more relevant than comparing raw values. The first PRO questionnaire that a patient completed was taken as their baseline and the values from the other PROs were subsequently represented as a percentage change compared to these baselines.

5.1.3 Default Case

Each argumentation model requires a default case with a default outcome. In our study, we designate outcome 1 as the default, signifying progressive disease or death. By adopting this default outcome, the models operate under the assumption that patients exhibit progressive disease as the default condition when we are lacking information. This approach was selected due to the more damaging consequences associated with overlooking cases where patients have progressive disease. In contrast, exercising excessive caution with patients displaying stable disease is considered less detrimental. However, it is important to acknowledge that any misclassification by the model in a practical application can potentially be harmful, hence the need for rigorous evaluation.

5.1.4 Missing Values

Notably, not all cases contain values for every feature. Not all patients have 100% wear time of the given accelerometer. To ensure data quality, for each case there is accelerometer data that covers at least 50% of the 8-week time period. However, when splitting the accelerometer data into sub-periods, we may encounter a sub-period where no data covers it. For example, consider splitting the 8-week time period into two 4-week periods where there is full coverage of accelerometer data in the first period and no data in the second. Furthermore, not every PRO question was completed in full. A scale is considered unreported if less than 50% of the questions for that scale were missing. In these cases, we characterise these values as conceptually "unknown" and assign them the value 0. As a result, in our models, a lack of a feature does not necessarily mean that the case does not have that feature, but that we do not know that the case has that feature. As a reasoning system, this approach is not as robust as removing cases with missing values, which would allow us to draw more confident conclusions. Nevertheless, this is the nature of real-world data where it is difficult to comprehensively or accurately capture every facet of a given case and so we design our models to accommodate this inherent challenge. Future research could explore other strategies for handling missing values.

5.2 Characterisation Extraction for Argumentation Models

5.2.1 Thresholds and Sub-Periods

The methodology employed in AA-CBR relies on the concepts of exceptionality. This requires establishing criteria for identifying cases that are more exceptional compared to others within the case base. Each model employs various approaches to define when a case is considered more specific. Based on these conditions, we need to decide when a feature is considered "exceptional enough" to be worth representing.

As PA data and PRO measures are represented as percentage changes compared to a baseline, we can set thresholds on the change to determine if that feature should be included or not. If the magnitude of change does not exceed the threshold set, the value is interpreted as a 0. These thresholds help determine if a case should be considered more exceptional than another and we can adjust these thresholds as a hyperparameter as appropriate for the model. For instance, a recorded change in walking time of-0.1% may not be significant enough to be represented. Hence, it is necessary to set thresholds to determine when a feature qualifies as exceptional. We use two thresholds, one for the PA values and one for the PRO values. A separate threshold for each individual feature is infeasible due to the number of features and the large range of values that the thresholds can be.

Additionally, we experiment with splitting the 8-week PA data into non-overlapping sub-periods to capture trends in the PA data. Instead of representing the PA data as an average over the full 8-week period, we could instead average the PA data over two non-overlapping 4-week periods, four 2-week periods or eight 1-week periods. The number of sub-periods we select is a hyperparameter of our characterisation method.

The value of these thresholds and the number of sub-periods can be found by conducting a random hyperparameter optimisation search. For each model, we will identify the best thresholds and the number of sub-periods to use in the model that results in an optimal classification of the labelled data points.

5.2.2 Feature Selection

The selection of features plays a crucial role in AA-CBR, as it lacks an inherent method for determining the most important features of the data for the model. Therefore, it is necessary to carefully consider which features from the PA data and PRO data should be included. Regarding the PA data, we must decide whether to incorporate acceleration (mg) and determine which functional behaviours to utilise. In the case of PRO data, the selection involves narrowing down which scales to include. It is important to strike a balance between explainability and model performance when choosing the features, and so we must provide clear justifications for which features are selected.

As there are 6 possible features that can be selected for PA data, there are 2^6 possible subsets of PA features that can be used. For the EORTC QLQ-C30/BN20 questionnaire, there are 2^{26} possible subsets of features. Thus, in total, there are 3^{32} possible subsets of features that can be used. With this many possible combinations, we need a better strategy for finding which features should be used than a grid search. Three methods of feature selection were explored for each model: a method based on autoencoders, a method based on neural network classifiers, and a method of ranking features based on their inclusion in argumentation models.

1. **Autoencoder:** The autoencoder technique aims to identify the most crucial features for reconstructing the dataset (refer to the background section 2.2.3, for more details). By examining the weights of the single hidden layered autoencoder, we can assess the significance of each feature. Subsequently, these features can be sorted based on their weight values and then allowing us to determine the top features to use for optimal model performance.
2. **Neural Network Classifier:** Although the autoencoder approach is advantageous due to its task-agnostic nature, our specific objective is binary classification. Thus, our focus lies not in determining the most suitable features to reconstruct the data, but rather identifying the features that can effectively discriminate between cases with and without progressive disease. To address this, we explore an alternative method by employing a single-layered neural network trained for data classification. Similar to the autoencoder approach, we examine the weights of this network to assess feature importance.
3. **Inclusion Ranking:** The autoencoder and neural network approach are useful methods for finding which features are important for the task, however, they aren't directed at finding which features are important for argumentation. The inclusion ranking approach can be used to determine the optimal feature by randomly selecting a subset of 10 features to use in the argumentation model, utilising the model as a classifier and evaluating the model's binary cross-entropy loss. We repeat this process 400 times, recording the loss against each included feature. For each feature, we can subsequently calculate the average loss of the model when that feature is included and sort the list of features based on the ones associated with the lowest loss.

Using these methods, we can identify which features appear to be the most important for the task and then conduct a smaller grid search to find the subset of features that perform most optimally. For each model, we will state which method of feature selection found the best subset of features to use.

5.3 Hyperparameter Tuning

We divided the 110 cases randomly into a training set consisting of 60 cases and a held-out test set consisting of 50 cases. With this split, each set is large enough to be subdivided such that some cases are used in the argumentation case base and some for making predictions. The test set will not be utilised in hyperparameter tuning.

The training set is used to find the best hyper-parameters of the characterisation extraction methods. For each model, we use the training set only. Due to the limited size of the training set, we conduct a 3-fold cross-validation 10 times and then average the results. This allows us to tune the models to identify parameters that work most effectively on the training data without relying on a small validation set.

We will compare the metrics F1 Score, Accuracy, Precision and Recall in order to select the models with the best hyperparameters. We find the F1 Score, Precision and Recall taking an outcome of 1 as the positive outcome.

The hyperparameter tuning involves the consideration of the thresholds, the number of weeks to average PA data over and the feature selection in concert with each other. Adjusting one of these parameters affects the optimal value of the others. To assess the potential of PA data as a supplementary or replacement measure of patient-reported outcomes (PRO), for each model we analyse three variants: one solely utilising PA data, one solely utilising PRO data, and one incorporating both PA and PRO data.

The held-out test set is used for evaluating the model on unseen data and comparing it against each of the baseline models. More details on the evaluation methodology and an analysis of the results can be found in the Evaluation Chapter 7.

Chapter 6

Models

In this section, we propose eight AA-CBR models that can be utilised to predict the status of patient disease utilising PA and PRO data. Each model is defined utilising Definition 9. We also detail how the data is characterised for each model and we describe the hyperparameters that we used to tune the data characterisation pipeline utilising the training data set. A deeper analysis of the models' performance on the held-out test set is in Chapter 7. Appendix A lists the optimal features selected for each model.

Models are implemented building on an AA-CBR tool (https://github.com/CLArg-group/AACBR) based on the literature [17, 29, 27].

6.1 Set-Based AA-CBR Models

Each of the literature-based AA-CBR models that we explore in this section utilises sets to represent arguments and make comparisons. Consequently, it becomes necessary to characterise the PA values and PRO values of each data point using sets. Since the PRO and PA data are measured as percentage changes, we can encode each feature by labelling it as either 'Increased' or 'Decreased'.

To illustrate the set characterisation, if the hours of sleep recorded for a data point have changed by -20% relative to the baseline, we can represent this change as the feature 'Sleep_Decreased'. If a patient reports a $+30\%$ change in pain symptoms, it can be represented as the feature 'PA_Increased'. However, we also need to establish the extent to which a change is considered exceptional enough to be represented using thresholds.

Additionally, as PA data spans an 8-week period, we experiment with breaking up its representation into smaller sub-periods. Each sub-period can be represented as an independent feature in the set. For example, if we choose to split the 8 weeks into two non-overlapping sub-periods, the set can include 'Sedentary_0_Increased' and 'Sedentary_1_Decreased', if the first 4-week period recorded an increase in time spent being sedentary whilst the second 4-week period recorded a decrease. This gives us flexibility in our representation whilst still characterising the data into sets.

6.1.1 Model 1: AA-CBR

Model Description and Design

First, we explore AA-CBR as originally proposed in the literature [17]. Further details can be found in the background section 2.2.2. We define the model in terms of Definition 9:

Model 1 (AA-CBR)**.**

- Data point: (X, o) where X is a set of features and $o \in \{0, 1\}$
- Dataset: $\mathcal{D}$ = full set of data points
- Partial Order: $\succcurlyeq = \supseteq$
- Default Case: $(C_\delta, \delta) = (\varnothing, 1)$
- Irrelevance Relation: $\not\prec = \not\succcurlyeq = \not\doteq$

Hyperparameter Tuning

Table 6.1 shows the best hyperparameters. Table 6.2 shows the performance of each model, using these hyperparameters, on an average of ten 3-fold cross-validations over the training dataset.

For Model 1, we see that the model that solely incorporates PA features exhibits better performance when the threshold is set to 0 as every change demonstrates exceptional significance for achieving optimal results. In contrast, the models utilising just PRO features or a combination of PA and PRO features require much larger thresholds. This suggests that models that contain more features require stricter criteria for defining exceptionality on a per-feature basis. Furthermore, we see that dividing the 8 weeks into sub-periods does not enhance the representation of PA data. Consequently, for AA-CBR, employing the smallest possible representation, in terms of subsets, appears to lead to the best performance.

Regarding the feature selection method, when using fewer features, the autoencoder and NN classifier methods consistently identified the same set of features for inclusion. However, as the number of features increases, it becomes necessary to use a method that utilises AA-CBR in the feature selection process. Thus, inclusion ranking performs optimally. Notably, during hyperparameter tuning, the autoencoder and NN classifier methods showed greater stability, consistently ordering the features in a similar order for each usage whereas the inclusion ranking method would not. The suggested set of features identified was from one of the runs of the inclusion ranking method.

Features	**PA Threshold %**	**Number of sub-periods**	**PRO Threshold %**	**Feature Selection Method**
PA Features	0	1	N/A	Autoencoder / NN Classifier
PRO Features	N/A	N/A	60	Inclusion Ranking
PA and PRO Features	180	1	80	Inclusion Ranking

Table 6.1: AA-CBR Hyperparameters

	Accuracy	**Precision**	**Recall**	**F1**
PA Features	0.780	0.859	0.67	0.750
PRO Features	0.736	0.693	0.79	0.720
PA and PRO Features	0.713	0.684	0.8	0.732

Table 6.2: AA-CBR average performance over ten 3-fold cross-validations on the training dataset.

Graphical Representation

We visualise the argumentation framework for a given focus case in Figure 6.1 using the training data set. However, many of the nodes in the graph are inconsequential to the explanation of the outcome of the focus case. Due to the complexity of this representation, we have also chosen to create a smaller representation that only includes paths from the focus case to the default case that contain at least one node from the grounded set. This can be seen in Figure 6.2. This condensed representation resembles a dispute tree, albeit without formally labelling each node or separating the paths into distinct branches. More details on Dispute Trees can be found in Section 2.2.5. Although not as explicit, the same dialogical interpretations as those derived from dispute trees can be concluded from this representation by simply following the paths from the default node backwards to the focus case. Therefore, we have chosen this representation to include in the report, given its succinctness. It is worth noting that the case base exhibits noise and incoherence, which is a common consequence when dealing with real-world data. Fortunately, our graphical representation enables us to identify and visualise these aspects effectively.

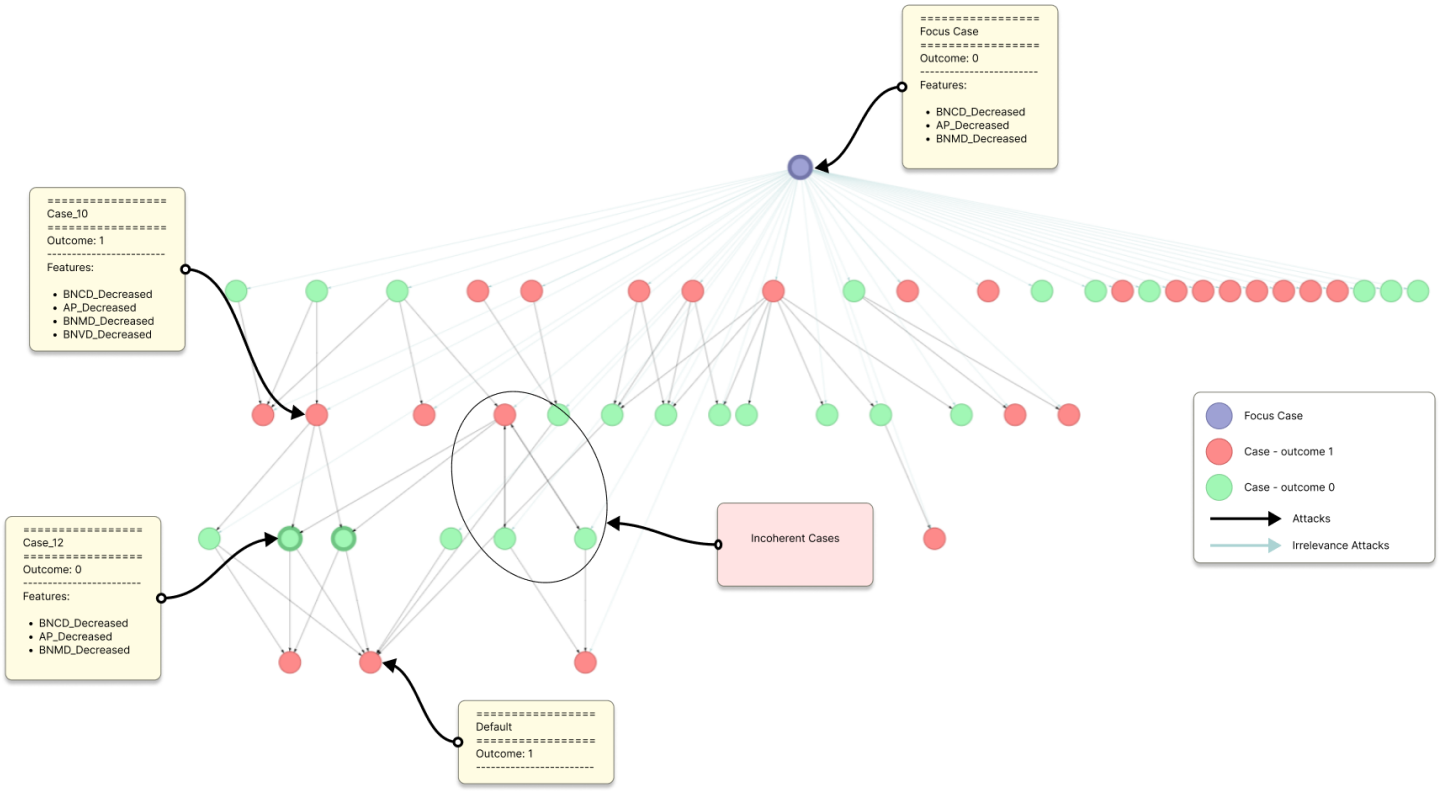


Figure 6.1: An example graphical AAF generated by Model 1

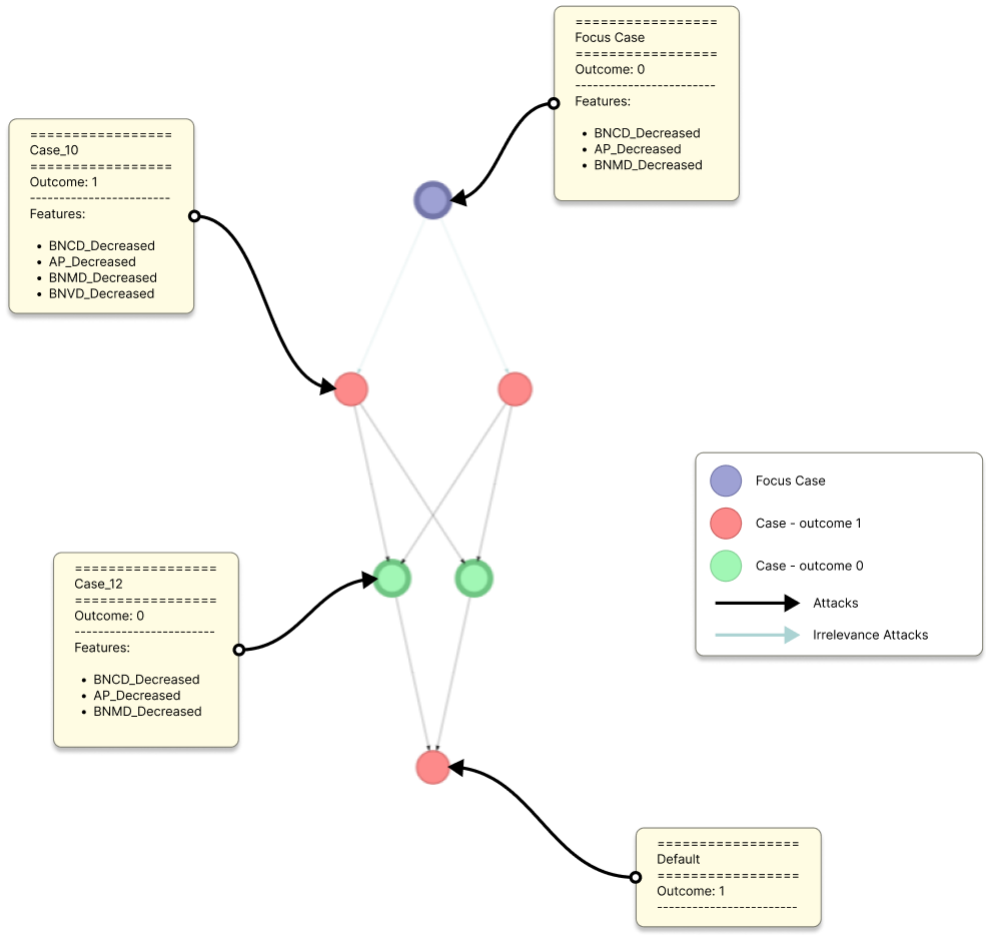


Figure 6.2: Condensed representation generated by Model 1

6.1.2 Model 2: cAA-CBR

Model Description and Design

The cAA-CBR model is defined analogously to Model 1, but the dataset is restricted to a concise dataset. Background section 2.2.6 explains cAA-CBR further. We define the model in terms of Definition 9:

Model 2 (cAA-CBR)**.**

- Data point: (X, o) where X is a set of features and $o \in \{0, 1\}$
- Dataset: $\mathcal{D}$ = concise subset of data points
- Partial Order: $\succcurlyeq = \supseteq$
- Default Case: $(C_\delta, \delta) = (\varnothing, 1)$
- Irrelevance Relation: $\not\prec = \not\succcurlyeq = \not\doteq$

We characterise each set of features in the same way as with Model 1, but we explore different hyperparameters in our character extraction method and assess if we can get better performance using cAA-CBR. The benefit of experimenting with this model is that it offers a method of handling incoherent cases. These cases are likely to occur in a real-world dataset when real values are being characterised as sets of simplified features. But as incoherence is handled with cAA-CBR, the characterisation of data points can be less strict as we aren't as concerned with encountering noise; the algorithms used for cAA-CBR will handle it for us.

Hyperparameter Tuning

Table 6.3 shows the best hyperparameters. Table 6.4 shows the performance of each model, using these hyperparameters, on an average of ten 3-fold cross-validations over the training dataset.

For the cAA-CBR Model 2, the same set of features was selected as with AA-CBR to perform optimally on the classification task. The most notable difference in the hyperparameter for cAA-CBR is the setting of the thresholds. For the model that solely utilises PA Features, setting a significantly larger threshold, at 90%, to perform optimally. This is because when exclusively utilising PA features, increasing the threshold results in more cases that are incoherent. For AA-CBR, this results in a decrease in performance but for cAA-CBR, as the algorithm contains a method for handling incoherent cases, we can see a performance increase by relaxing the definition of exceptionality. Interestingly, we see smaller thresholds for the model that contains both PA and PRO features, in this case, it is again for a similar reason that with different thresholds set compared to AA-CBR, we see incoherent cases handled differently.

Features	**PA Threshold %**	**Number of sub-periods**	**PRO Threshold %**	**Feature Selection Method**
PA Features	90	1	N/A	Autoencoder / NN Classifier
PRO Features	N/A	N/A	60	Inclusion Ranking
PA and PRO Features	100	1	60	Inclusion Ranking

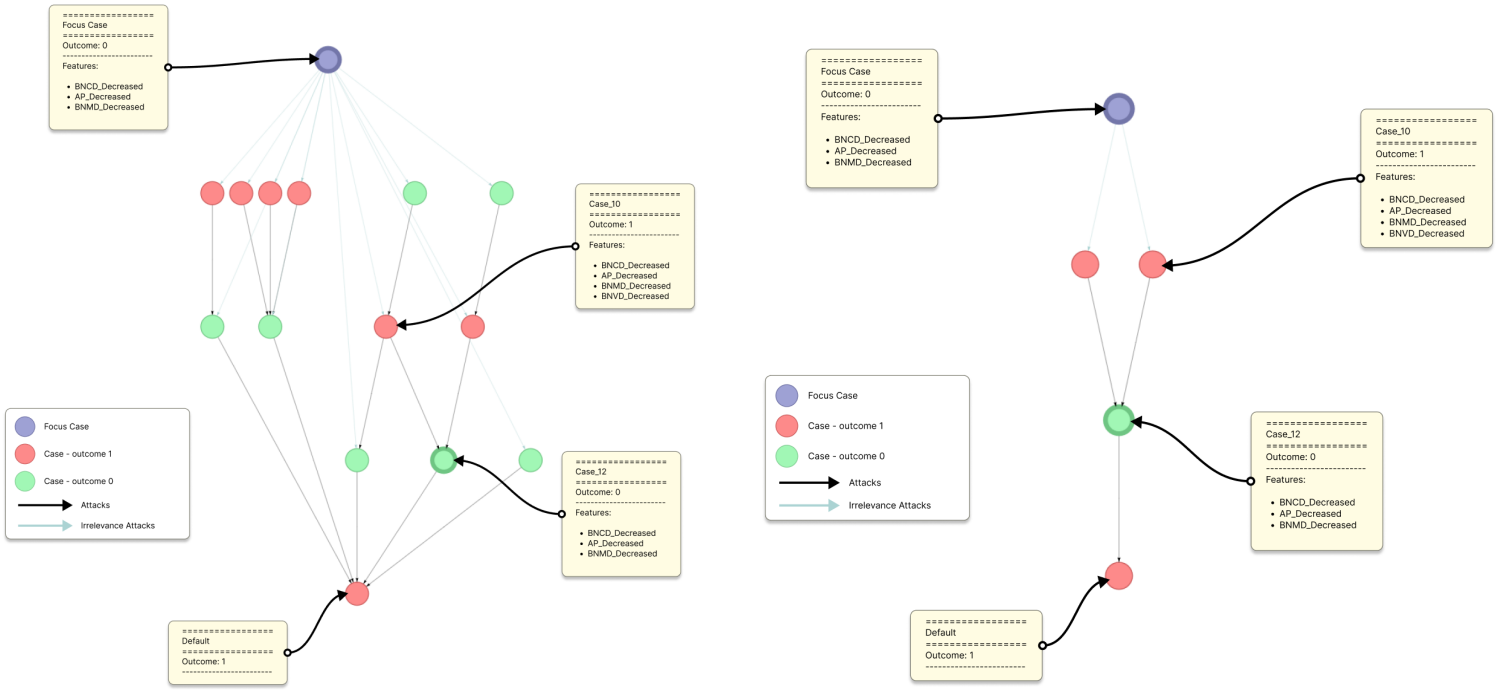
Table 6.3: cAA-CBR Hyperparameters

	Accuracy	**Precision**	**Recall**	**F1**
PA Features	0.760	0.717	0.860	0.776
PRO Features	0.748	0.711	0.810	0.736
PA and PRO Features	0.730	0.692	0.830	0.752

Table 6.4: cAA-CBR average performance over ten 3-fold cross-validations on the training dataset.

Graphical Representation

The graphical representation of the cAA-CBR in Figure 6.3(a) differs greatly compared to the graphical representations generated by Model 1. We present the representation using the same case base and focus case but we now see that there are fewer nodes in the final graph. Additionally, there are now no incoherent cases. This is a result of the algorithm that restricts the case base to a concise dataset. This offers a simpler representation that is easier to interpret as only paths from the focus case to the default are included in the graph. However, removing cases to construct the concise data set comes at the cost of other aspects of interpretability. We are no longer able to visually see all the cases that are considered irrelevant to the focus case which may hold clinical significance.



(a) Graphical AAF

(b) Condensed Representation

Figure 6.3: Graphical Representations generated by Model 2

6.1.3 Model 3: AA-CBR with Dynamic Features

Model Description and Design

AA-CBR can be extended to represent Dynamic Features. More details can be found in the Background Section 2.2.4. We can design an argumentation framework that can take into account the point in time that a case occurs. We propose a novel modification of the framework for AA-CBR with dynamic features described in the background section in which real values are used to represent time in place of discrete stages. Each data point will therefore be characterised as a set of features, similar to Model 1, along with a measure of time. It is challenging to define this model strictly using Definition 9 as the definition of specificity for this model is dependent on time value conditions. Instead of defining a partial order relation, we will explicitly state the attacks relation and specificity conditions for this model as derived from Definition 11:

Model 3 (AA-CBR with Dynamic Features)**.**

- Data point: (X, t, o) where X is a set of features, $t \in \mathbb{R}$ is a point in time and $o \in \{0, 1\}$
- Dataset: $\mathcal{D}$ = full set of data points
- For $(F, t, o), (F', t', o') \in Args$, it holds that $(F, t, o) \rightsquigarrow (F', t', o')$ iff
  1. $o \neq o'$, and (different outcomes)
  2. either
     (a) $F' \subsetneq F$, and (specificity)
     (b) $\nexists (F^*, t^*, o) \in CB$ with (concision)
     - either $F' \subsetneq F^* \subsetneq F$,
     - or $F^* = F$ and $t^* < t$
     - or $F' = F^*$ and $t' < t^* \leqslant t$
  3. or
     (a) $F' = F$ and $t' < t$ and (advance)
     (b) $\nexists (F, t^*, o) \in CB$ with $t' < t^* < t$. (proximity)
- Default Case: $(C_\delta, t_\delta, \delta) = (\varnothing, 0, 1)$
- Irrelevance Relation: $(F, t, o) \nsim (F', t', o')$ iff either $(F \nsubseteq F_\phi$ or $t < t')$

Conceptually, the amount of time that a patient has had their brain tumour could be an important feature in predicting progressive disease or stable disease. We must decide how to represent the time feature such that it can be fairly compared across different patients. Ideally, we would represent time as the number of days that the patient has had their cancer for. However, not all patients are diagnosed when their cancer first forms, so accurately measuring this isn't possible. The date of diagnosis is not provided and so cannot be used as an approximation. Instead, we experiment with the number of days each patient was in the study at the point of the questionnaire, the previous number of questionnaires completed and the number of previous times the patient has had a progressive disease outcome.

Hyperparameter Tuning

We present the results of our hyperparameter tuning, focusing on the different characterisations of time. Among these characterisations, we find that the most similar parameters to Model 1 are obtained when characterising time based on the number of days a patient spent on the study at the time they completed their questionnaire. The corresponding hyperparameters and average performance of the model on ten 3-fold cross-validations over the training dataset are shown in Table 6.5 and Table 6.6, respectively. The performance of the model on the training set results in a similar accuracy and f1-score albeit with a decrease in precision and an increase in recall. This occurs because for any questionnaire completed by a patient who has been on the study for a longer duration than when the focus case questionnaire was completed is considered irrelevant. Consequently, compared to the other versions of AA-CBR that we have looked at, there are a greater number of cases considered irrelevant for a focus case and so the default outcome is more likely to be assigned. As a result, this characterisation proves too strict, too many cases are considered irrelevant and so impedes the improvement of the model.

Alternatively, considering the number of previous questionnaires the patient has completed may be a better choice. However, we observe that the model's performance decreases with all three variants. Similar to the use of the number of days in the study, this characterisation proves overly strict. From a clinical standpoint,

this observation is reasonable, as neither of these measures are an accurate reflection of the progression of the patient's disease. Patients have their cancer for different durations and the impact of the disease varies among individuals. Furthermore, some patients were on the study for longer duration than others, often patients were taken off of the study when they were too ill to continue. As a result, patients with more days on the study or more completed questionnaires may actually be those patients that were healthier and thus able to continue on the study.

The most effective method for characterising time was counting the number of times that a patient had had a previous case with progressive disease. Intuitively, this approach aligns with the shortcomings of the previous methods. Interestingly, with this characterisation, fewer PA features, with a threshold of 0, were necessary to achieve optimal results. Only requiring the features of tasks-light and sedentary. As this model performed the most optimally, we assess only the model with time based on the number of previous progressive disease cases in the final evaluation.

Features	**PA Threshold %**	**Number of sub-periods**	**PRO Threshold %**	**Feature Selection Method**
Number of days on study				
PA Features	0	1	N/A	Autoencoder / NN Classifier
PRO Features	N/A	N/A	80	Inclusion Ranking
PA and PRO Features	180	1	60	Inclusion Ranking
Number of Previous Cases				
PA Features	70	1	N/A	Autoencoder / NN Classifier
PRO Features	N/A	N/A	80	Inclusion Ranking
PA and PRO Features	180	1	80	Inclusion Ranking
Number of Previous Progressive Disease Cases				
PA Features	0	1	N/A	NN Classifier
PRO Features	N/A	N/A	80	Inclusion Ranking
PA and PRO Features	180	1	80	Inclusion Ranking

Table 6.5: Dynamic AA-CBR Hyperparameters

	Accuracy	**Precision**	**Recall**	**F1**
Number of days on study				
PA features	0.716	0.672	0.84	0.747
PRO features	0.650	0.600	0.900	0.709
PA and PRO features	0.741	0.62	0.927	0.741
Number of Previous Cases				
PA features	0.683	0.648	0.810	0.712
PRO features	0.623	0.598	0.760	0.656
PA and PRO features	0.650	0.617	0.79	0.685
Number of Previous Progressive Disease Cases				
PA features	0.803	0.752	0.907	0.819
PRO features	0.697	0.676	0.750	0.710
PA and PRO features	0.723	0.700	0.770	0.731

Table 6.6: Dynamic AA-CBR average performance over ten 3-fold cross-validations on the training dataset.

Graphical Representation

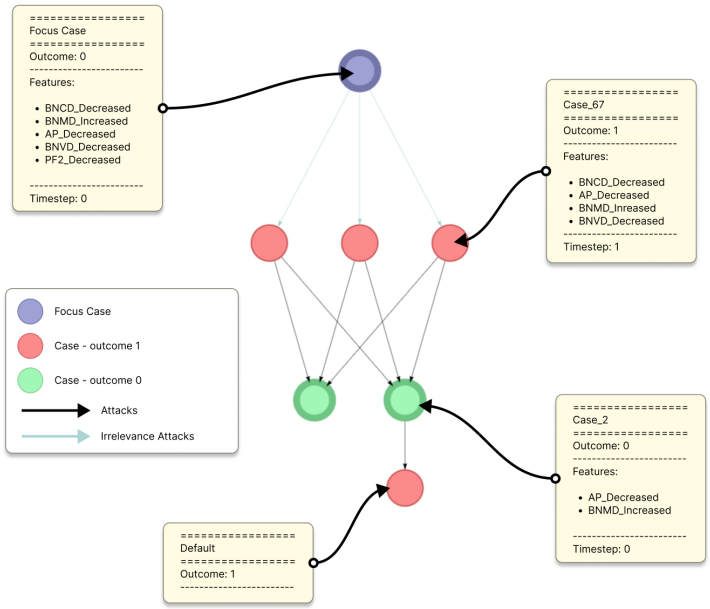


Figure 6.4: Condensed representation generated by Model 3

Figure 6.4 presents the condensed graphical representation of the Dynamic AA-CBR framework, generated by Model 3. Here we use the number of previous progressive disease cases for the time step measure. We highlight a specific path in which the use of time steps contributes to the explanation of the outcome predicted for the focus case. We can generate a dialogical explanation, akin to a Dispute Tree, as follows: The loser (L) states that the default outcome applies. Conversely, the winner (W) argues that Case_2 is an exception to the default case and applies instead. L contradicts this by stating that Case_67 is an exception to Case_2 and thus holds. W counters this by saying that Case_67 is irrelevant to the focus case, because it occurs at a later time step. Thus we see that the outcome of stable disease holds and that one such reason for this is that the focus case has not experienced progressive disease to the same extent as Case_67.

6.2 Value-Oriented AA-CBR Models

The previous models of AA-CBR that we have looked at are based on subsets of features. However, the data we have collected is more complex than the simplified characterisation of "Feature_Increased" and "Feature_Decreased". Instead, we explore novel approaches surveying a variety of argumentation models that operate directly on the values of the features. We characterise each case as (X, o) where $X = [x_1, x_2, ..., x_n]$ is a vector of values. We first explore manually defining custom partial orders that act directly on the values of each data point and then we look at a method of using neural networks to learn the partial order for us. For each partial order, we ensure that the zero vector, $\mathbf{0}$, is minimal with respect to the partial order selected - this allows us to define a default case.

6.2.1 Model 4: AA-CBR with Euclidean Norm Order

This model involves comparing the magnitude of the vectors. In terms of AA-CBR, cases characterised with vectors of a greater magnitude are considered more exceptional and attack cases with smaller magnitudes. This is analogous to Nonetheless, we still consider thresholding and feature selection such that the value of the vector magnitude is only dependent on the features that are most exceptional.

Given the cases $C_X = (X, o_X)$, $C_Y = (Y, o_Y)$, we define the Euclidean Norm order:

Definition 14 (Euclidean Norm order)**.**

- $C_X \succcurlyeq_e C_Y$ iff $||X|| \geqslant ||Y||$;

We can define the model in terms of Definition 9, selecting one of the above partial orders:

Model 4 (AA-CBR with Euclidean Norm)**.**

- Data point: (X, o) where $X = [x_1, x_2, ..., x_n]$ is a vector of values and $o \in \{0, 1\}$
- Dataset: $\mathcal{D}$ = full set of data points
- Partial Order: $\succcurlyeq = \succcurlyeq_e$
- Default Case: $(C_\delta, \delta) = (\mathbf{0}, 1)$
- Irrelevance Relation: $\nprec = \nsucc$

Hyperparameter Tuning

For the Euclidean Norm Order, we see similar hyperparameters to that of Model 1. The main difference is that for the model that solely focuses on the PA features, the acceleration feature was selected. Despite the way this method aggregates the values in the feature vector into a single score, we find that there is still a requirement feature selection and thresholding in order to achieve optimal performance and therefore we have similar a burden of feature characterisation as with the set-based methods.

Euclidean Norm Order $\succcurlyeq_e$				
Features	**PA Threshold %**	**Number of sub-periods**	**PRO Threshold %**	**Feature Selection Method**
PA Features	20	1	N/A	Inclusion Ranking
PRO Features	N/A	N/A	110	Inclusion Ranking
PA and PRO Features	180	1	100	Inclusion Ranking

Table 6.7: AA-CBR Hyperparameters with $\succcurlyeq_e$

Euclidean Norm Order $\succcurlyeq_e$				
	Accuracy	**Precision**	**Recall**	**F1**
PA features	0.680	0.674	0.700	0.682
PRO features	0.580	0.562	0.730	0.625
PA and PRO features	0.547	0.540	0.610	0.571

Table 6.8: AA-CBR with $\succcurlyeq_e$ average performance over ten 3-fold cross-validations on the training dataset.

Graphical Representation

The use of a total order results in argumentation graphs, Figure 6.5, that are hard to interpret due to the volume of the nodes and the complex paths from the default case to the focus case. The condensed representation does not improve this issue. Furthermore, using the Euclidean norm to represent cases in the order reduces the interpretability further. It becomes challenging to determine which specific features are responsible for the outcome. Additionally, we have to consider our target audience for these models. The Euclidean Norm does not provide a clinically interpretable model. Due to the total order, the visualisations demonstrate that the focus case is essentially assigned to the outcome of the case with the largest Euclidean Norm that is still smaller than the Euclidean Norm of the focus case. However, this assignment has no meaningful clinical significance.

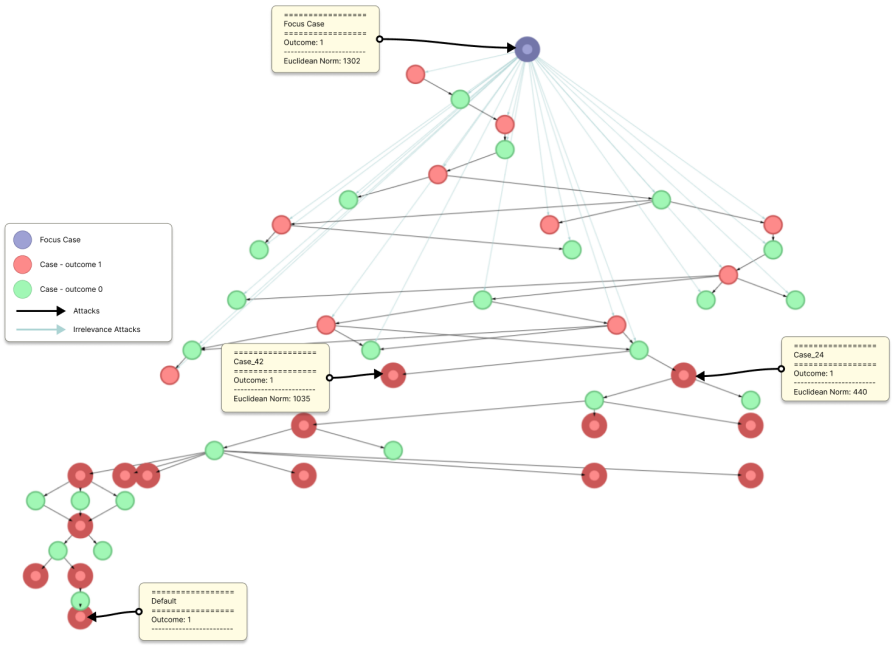


Figure 6.5: An example graphical AAF generated by Model 4

6.2.2 Model 5: AA-CBR with Absolute Product Order

The Absolute Product Order operates by comparing the magnitudes of each individual value in the vectors. Within the context of AA-CBR, cases, where the magnitude of every value in the vector is larger than another case, are considered more exceptional and attack the other cases.

Given the cases $C_X = (X, o_X)$, $C_Y = (Y, o_Y)$, we define the Absolute Product Order:

Definition 15 (Absolute Product Order)**.**

- $C_X \succcurlyeq_{apo} C_Y$ iff $\forall i \in [0, n], |X[i]| \geqslant |Y[i]|$;

We can define the model in terms of Definition 9:

Model 5 (AA-CBR with Absolute Product Order)**.**

- Data point: (X, o) where $X = [x_1, x_2, ..., x_n]$ is a vector of values and $o \in \{0, 1\}$
- Dataset: $\mathcal{D}$ = full set of data points
- Partial Order: $\succcurlyeq = \succcurlyeq_{apo}$
- Default Case: $(C_\delta, \delta) = (\mathbf{0}, 1)$
- Irrelevance Relation: $\prec = \not\succcurlyeq$

Hyperparameter Tuning

Using the Absolute Product Order, we see an improved performance of the models on the training dataset compared to the Euclidean Norm Order models. This is the first model where see an increased performance for the PA Features when characterised using more than 1 sub-period. This is likely because when using Absolute Product Order, more values are comparable as we are not concerned with the direction of change compared to the set-based AA-CBR models. However, as this approach disregards the direction of the values, it likely explains the loss in performance and reduces our ability to reason about the model output.

Absolute Product Order $\succcurlyeq_{apo}$				
Features	**PA Threshold %**	**Number of sub-periods**	**PRO Threshold %**	**Feature Selection Method**
PA Features	70	8	N/A	Inclusion Ranking
PRO Features	N/A	N/A	50	Inclusion Ranking
PA and PRO Features	180	1	80	Inclusion Ranking

Table 6.9: AA-CBR Hyperparameters with $\succcurlyeq_{apo}$

Absolute Product Order $\succcurlyeq_{apo}$				
	Accuracy	**Precision**	**Recall**	**F1**
PA features	0.647	0.5973	0.890	0.710
PRO features	0.673	0.650	0.76	0.687
PA and PRO features	0.617	0.600	0.690	0.636

Table 6.10: AA-CBR with $\succcurlyeq_{apo}$ average performance over ten 3-fold cross-validations on the training dataset.

Graphical Representation

Figure 6.6 illustrates the Absolute Product Order leads to nodes in the graph that have a larger degree on average. This shows that the arguments have more cases that they can attack compared to the previous models examined. Furthermore, many of the attacks relations aren't required for the explanation of the outcome of the focus case. As a result, these are removed in the condensed representation in Figure 6.7. Additionally, we see incoherence is prevalent with this partial order. These differences arise from the relatively relaxed nature of this partial order, as it does not take into account the direction of change for a given feature. Consequently, the generated graphs are more difficult to interpret.

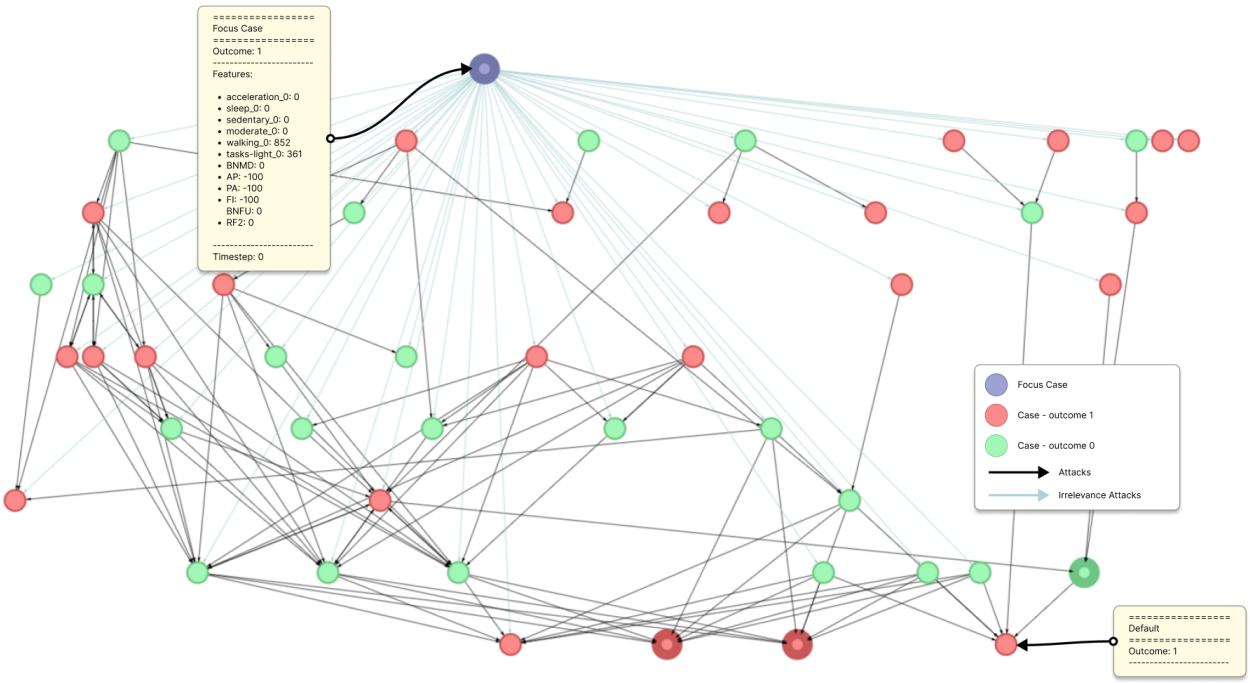


Figure 6.6: An example graphical AAF generated by Model 5

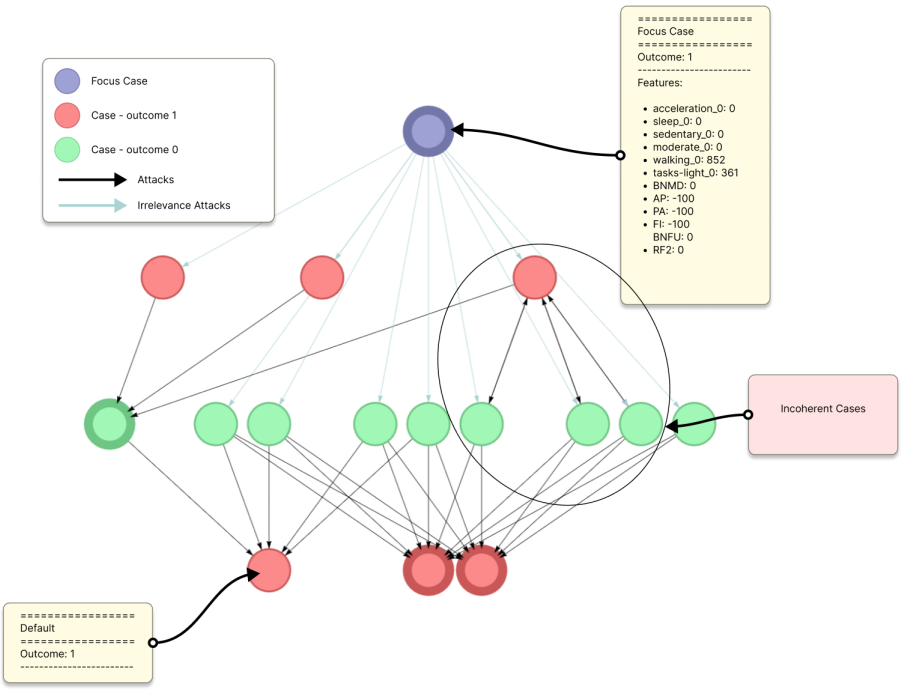


Figure 6.7: Condensed representation generated by Model 5

6.2.3 Model 6: AA-CBR with Sign and Magnitude Partial Order

The Sign and Magnitude Partial Order, which considers the signs of the values before comparing their magnitudes. If the signs are not the same for two cases, neither case attacks the other. However, when the signs are the same, the magnitude of each value is examined. Cases with larger magnitudes in every value are considered more exceptional and can attack cases with smaller values. This order doesn't consider 0 as either positive or negative so values can always be compared to 0. By incorporating the sign comparison, this partial order addresses the limitations of the Absolute Product Order. This approach is the most similar to the set-based AA-CBR as proposed in Model 1.

Given the cases $C_X = (X, o_X)$, $C_Y = (Y, o_Y)$, we define the Sign and Magnitude Order:

Definition 16 (Sign and Magnitude Partial order)**.**

- $C_X \succcurlyeq_{sm} C_Y$ iff $\forall i \in [0, n]$, $(X[i] \times Y[i] \geqslant 0 \wedge |X[i]| \geqslant |Y[i]|) \vee \neg(X[i] \times Y[i] \geqslant 0)$

We can define the model in terms of Definition 9:

Model 6 (AA-CBR with Sign and Magnitude Order)**.**

- Data point: (X, o) where $X = [x_1, x_2, ..., x_n]$ is a vector of values and $o \in \{0, 1\}$
- Dataset: $\mathcal{D}$ = full set of data points
- Partial Order: $\succcurlyeq = \succcurlyeq_{sm}$
- Default Case: $(C_\delta, \delta) = (\mathbf{0}, 1)$
- Irrelevance Relation: $\prec = \not\succcurlyeq$

Hyperparameter Tuning

As this approach is the one that most closely resembles set-based AA-CBR, we see a similar characterisation, including the same set of features that perform most optimal. This model offers more nuance in the definition of the partial order than with the set-based AA-CBR, and may be too strict hence the slight decrease in performance during training. This method performs better than with the Euclidean Norm Order and the Absolute Product Order on the 3-fold cross-validation on the training dataset.

Sign and Magnitude Order $\succcurlyeq_{sm}$				
Features	**PA Threshold %**	**Number of sub-periods**	**PRO Threshold %**	**Feature Selection Method**
PA Features	50	4	N/A	Inclusion Ranking
PRO Features	N/A	N/A	80	Inclusion Ranking
PA and PRO Features	160	1	80	Inclusion Ranking

Table 6.11: AA-CBR Hyperparameters with $\succcurlyeq_{sm}$

Sign and Magnitude Order $\succcurlyeq_{sm}$				
	Accuracy	**Precision**	**Recall**	**F1**
PA features	0.620	0.576	0.913	0.702
PRO features	0.727	0.692	0.810	0.741
PA and PRO features	0.717	0.675	0.830	0.740

Table 6.12: AA-CBR with $\succcurlyeq_{sm}$ average performance over ten 3-fold cross-validations on the training dataset.

Graphical Representation

In Figure 6.8 and Figure 6.9, we observe that the graphical representation of the Sign and Magnitude Partial Order model provides clearer explanations compared to previously proposed value-oriented models, although not as clear as set-based AA-CBR models. Additionally, incoherence is less prevalent than with the Absolute Partial Order. It is important to highlight that in using values to compare cases, generating dialogical explanations simply by looking at the condensed representation is more complicated. In contrast to the set-based AA-CBR approach that relies on the presence of features, we now need to determine which features possess larger magnitudes as the reason for the attack relation.

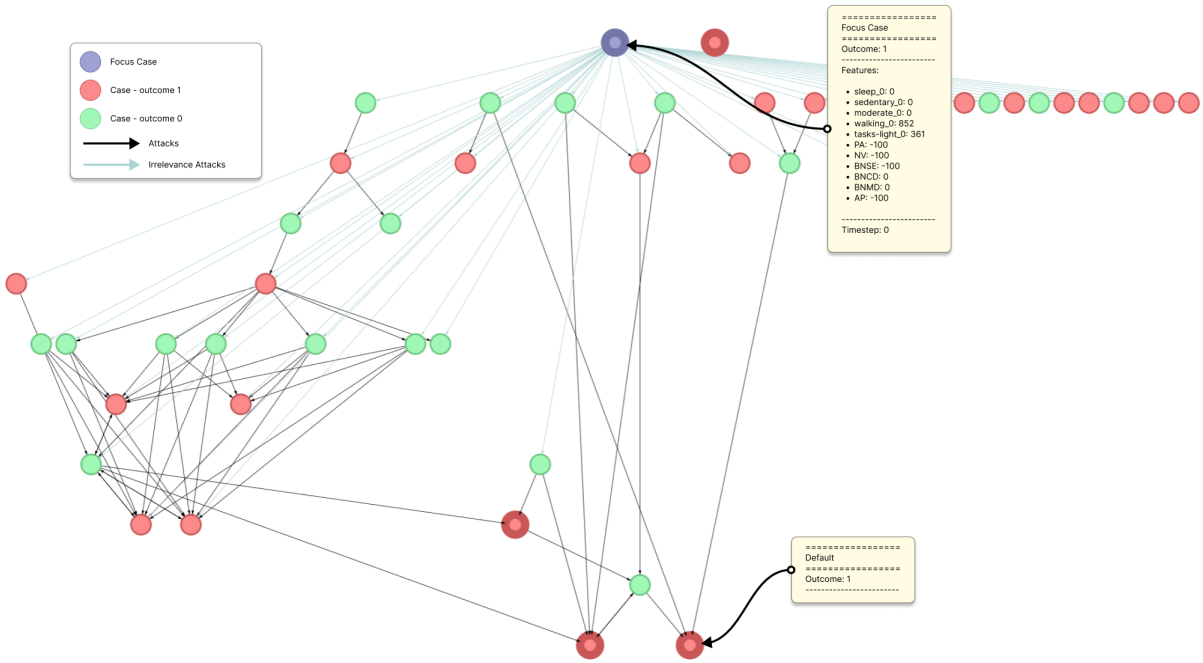


Figure 6.8: An example graphical AAF generated by Model 6

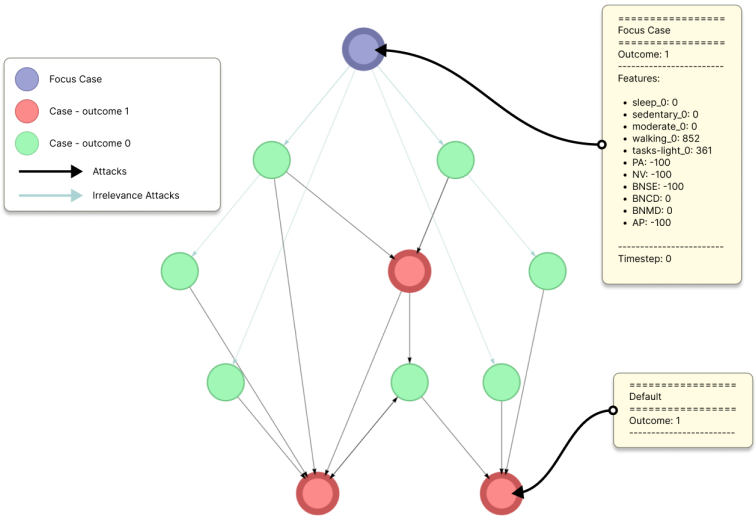


Figure 6.9: Condensed representation generated by Model 6

6.3 Neural Network Based Models

The weakness of AA-CBR compared to other machine learning models such as decision trees or neural networks is that AA-CBR has no inherent method of identifying which features are most important and considerable effort has to be placed on characterisation extraction. To address this, we explore models that utilise a neural network to learn an order over the cases. We denote these models as Neural Network AA-CBR (NN-AA-CBR).

6.3.1 Model 7: Total Ordered NN-AA-CBR

Model Description and Design

The model presented here learns a total order over the cases. This means that every case is assigned a value and that every case can be compared to every other case in the order.

Model 7 (NN-AA-CBR Total Order)**.**

- Data point: (X, o) where $X = [x_1, x_2, ..., x_n]$ is a vector of values and $o \in \{0, 1\}$
- Dataset: $\mathcal{D}$ = full set of data points
- Partial Order: $(X, o_x) \succcurlyeq (Y, o_Y)$ iff $\mathbf{NN}(X) \geqslant \mathbf{NN}(Y)$
- Default Case: $(C_\delta, \delta) = (\mathbf{0}, 1)$
- Irrelevance Relation: $\nprec = \ngtr$

Where $\mathbf{NN}(X)$ is a neural network trained to learn a total order on the cases.

Neural Networks are a supervised learning method, however, the true partial order is unknown, and so we adopt an approach to training the NN that utilises the output of the AA-CBR model. Algorithm 1 details how the model is trained. The key idea is to fit the NN-AA-CBR model to a training set and use it to make predictions on a separate validation set and update the partial order based on the results.

Algorithm 1 Neural Network Training Loop

Initialize neural network weights
Set learning rate α
Set number of epochs N
Initialize AA-CBR model
for $i \leftarrow 1$ to N **do**
    Shuffle training data
    Split the training data into a training set $(\mathbf{x}_t, \mathbf{y}_t)$ and validation set $(\mathbf{x}_v, \mathbf{y}_v)$
    Initialize empty array new_orders
    Fit AA-CBR with training set
    Predict outcomes for validation set with AA-CBR
    Compute accuracy, acc, of validation set predictions
    for each training example (x, y) in $(\mathbf{x}_v, \mathbf{y}_v)$ and corresponding prediction, $\hat{y}$ **do**
        if $\hat{y} = y$ **then**
            Set p to $\mathbf{NN}(x)$
        else if $\hat{y} = \delta$ **then**
            Set p to min($\mathbf{NN}(x)$ + $(1 - acc)$, 1)
        else
            Set p to max($\mathbf{NN}(x)$ - $(1 - acc)$, -1)
        end if
        append p to new_orders
    end for
    Compute loss $L = \text{Loss}(\mathbf{NN}(\mathbf{x}_v), new\_orders)$
    Compute gradients $\nabla_\theta L = \text{Backpropagation}(\mathbf{NN}(\mathbf{x}), new\_orders)$
    Update weights $\theta \leftarrow \theta - \alpha \nabla_\theta L$
end for
Output: Trained neural network weights

For each data point in the validation set, if the prediction aligns with the expected outcome we don't change the partial order for that particular case. However, if the prediction is incorrect we need to adjust the case's position in the total order and determine the magnitude of the adjustment. When incorrectly predicting the default outcome for a case, it signifies the default case is either unattacked or successfully defended. In such

scenarios, we aim to reduce the number of cases regarded as irrelevant to the new case, enabling more successful attacks on the default case or its defenders. On the other hand, if the default outcome is not predicted when it should indeed hold, the converse holds true.

The algorithm introduces changes in the partial order based on the accuracy of the predictions made by the validation set. The underlying concept is that if the majority of the validation set is accurately predicted, significant adjustments to the partial order are unnecessary. Conversely, if a substantial portion of the validation set is predicted incorrectly, significant adjustments are needed. To ensure manageable adjustments, we establish maximum and minimum values that the partial order can take. This prevents the magnitude of the values from becoming excessively large. For Algorithm 1, we set the maximum value to 1 and the minimum value to -1 and use a Tanh output activation.

The architecture of the Neural Network is comparable to that of the autoencoder and classifier used in feature selection and the classifier used in the evaluation. As a result, the network has a single hidden layer with 64 neurons.

Hyperparameter Tuning

As a result of using a Neural Network to learn the partial order, the need for extensive hyperparameter tuning is significantly reduced. Nonetheless, the performance of this model on the training dataset is considerably inferior to that of previous models. This implies that our training method fails to identify an optimal partial order or that the neural network used lacks the necessary capacity. However, attempts to increase the capacity of the network didn't result in an increase in performance, implying that the issue is with the training method. Furthermore, we see that model performance with the PRO features is worse than solely using PA features or using a combination of PA and PRO features. This also suggests that network capacity isn't the issue as increasing the number of features does not result in worsened performance. This does indicate that the neural network can learn which cases are more exceptional from the PA features better than the PRO features.

Features	**Number of sub-periods**
PA Features	4
PRO Features	N/A
PA and PRO Features	1

Table 6.13: NN-AA-CBR Hyperparameters

	Accuracy	**Precision**	**Recall**	**F1**
PA features	0.567	0.5496	0.720	0.623
PRO features	0.444	0.465	0.720	0.551
PA and PRO features	0.511	0.505	0.933	0.640

Table 6.14: NN-AA-CBR average performance over ten 3-fold cross-validations on the training dataset.

Graphical Representation

As with the Euclidean Norm Order (Model 4), using a total order does not contribute to useful explanations. Nonetheless, as a proof-of-concept for using a neural network to learn a partial order, we can gain valuable insights from this model. Figure 6.10 illustrates how the training method results in partial order values that are close -1 or 1, resulting in a graph structure that is less linear and with fewer layers than the graphs generated by the Euclidean Norm Order model. This outcome occurs because of the Tanh output activation function which has a larger output range for values near -1 or 1. Furthermore, the magnitude of changes applied to the position of values within the partial order during the training process is likely too large, resulting in heightened sensitivity of the partial order to the model's performance during training.

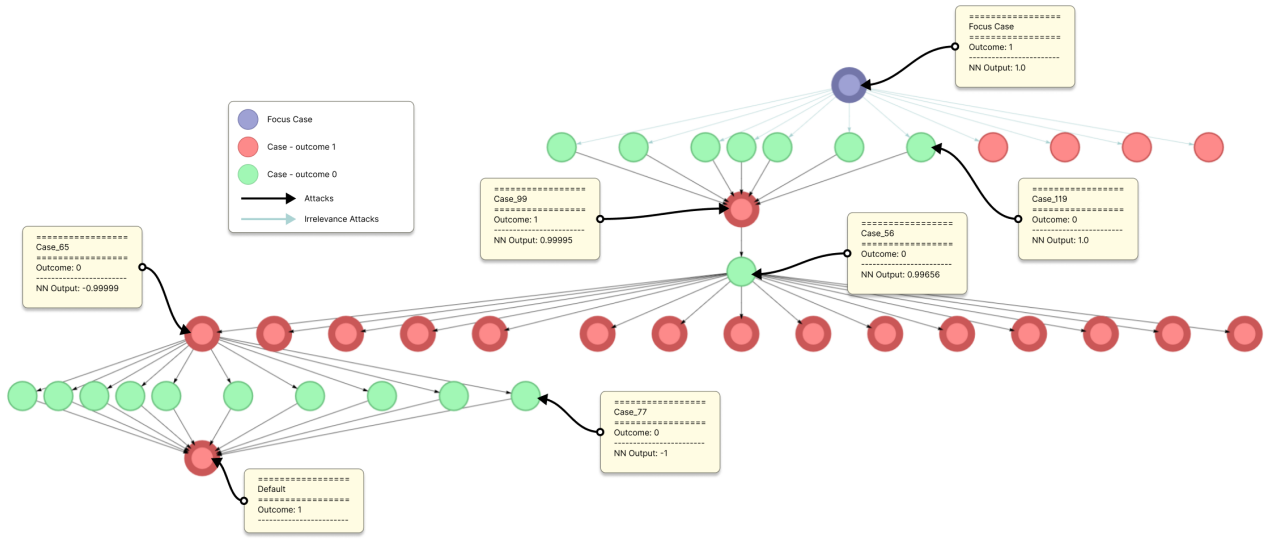


Figure 6.10: An example graphical AAF generated by Model 7

6.3.2 Model 8: Strict Partial Ordered NN-AA-CBR

Model Description and Design

As highlighted, the issue with using a total order is that the explanations generated are more difficult to interpret. We can instead modify Model 7 to learn a partial order that results in more easily interpretable explanations. Instead of using the neural network to learn a real number that's used to order the cases, instead the NN is trained to learn integers. We don't allow two cases with the same integer value to attack each other. This addresses the concerns of using a total order

Model 8 (NN-AA-CBR Strict Partial Order)**.**

- Data point: (X, o) where $X = [x_1, x_2, ..., x_n]$ is a vector of values and $o \in \{0, 1\}$
- Dataset: $\mathcal{D}$ = full set of data points
- Partial Order: $(X, o_x) \succcurlyeq (Y, o_Y)$ iff $argmax(\mathbf{NN}(X)) > argmax(\mathbf{NN}(Y))$
- Default Case: $(C_\delta, \delta) = (\mathbf{0}, 1)$
- Irrelevance Relation: $\nprec = \nsucc$

Where $\mathbf{NN}(X)$ is a neural network trained to learn a strict partial order on the cases. Algorithm 2 outlines the training process. This model has three key distinctions between compared to Model 7. Firstly, the NN is trained to output an integer output within the range 1 and the upper limit C. Secondly, the partial order utilises a strict comparison ($>$) to evaluate the outputs from the NN. Consequently, cases with identical NN outputs cannot be compared within the partial order. Finally, the training loop has been modified to accommodate these changes. The NN is now trained as a multi-class classification task where the learned 'class' represents the case's positions within the strict partial order.

Algorithm 2 Neural Network Training Loop

Initialize neural network weights
Set learning rate α
Set number of epochs N
Initialize AA-CBR model
Initialize upper limit C
for $i \leftarrow 1$ to N **do**
    Shuffle training data
    Split the training data into a training set $(\mathbf{x}_t, \mathbf{y}_t)$ and validation set $(\mathbf{x}_v, \mathbf{y}_v)$
    Initialize empty array *new_orders*
    Fit AA-CBR with training set,
    Predict outcomes for validation set with AA-CBR
    Compute accuracy, *acc*, of validation set predictions
    for each training example (x, y) in $(\mathbf{x}_v, \mathbf{y}_v)$ and corresponding prediction, $\hat{y}$ **do**
        Initialize array *expected_class* of 0s of length C
        Set *pred_class* to $argmax(\mathbf{NN}(x))$
        if $\hat{y} = y$ **then**
            Set $expected\_class[pred\_class] = 1$
        else if $\hat{y} = \delta$ **then**
            Set $excpected\_class[min(round(pred\_class + (C * (1 - acc)), C - 1)] = 1$
        else
            Set $excpected\_class[max(round(pred\_class - (C * (1 - acc)), 1)] = 1$
        end if
        append *expected_class* to *new_orders*
    end for
    Compute loss $L = \text{Loss}(argmax(\mathbf{NN}(\mathbf{x}_v)), new\_orders)$
    Compute gradients $\nabla_\theta L = \text{Backpropagation}(argmax(\mathbf{NN}(\mathbf{x})), new\_orders)$
    Update weights $\theta \leftarrow \theta - \alpha \nabla_\theta L$
end for
Output: Trained neural network weights

Hyperparameter Tuning

For our hyperparameter tuning, we compare the model when the partial order can take 10 different values, 20 different values and 30 different values. This comparison allows us to identify the granularity of the order required to achieve optimal performance. Table 6.15 and Table 6.16 show the hyperparameters and results on the training data respectively.

When using fewer values for the order, we see that we require more features representing the PA data in order to achieve optimal performance. This result is likely because as the total number of values that the order can take is relatively small, we need more features in order to identify the optimal where in the order a case should lie.

Increasing the values of the order results in better performance of the model and doesn't require the PA data to be split into sub-periods. There is a clear performance benefit to using 20 values over using 10 values or 30 values in the partial order. We see with these models, as with the previous Total-Ordered NN-AA-CBR, Model 7, that using PA features alone offers better performance on the training dataset than using PRO features alone. Adding the PA data to the PRO data increases the performance compared to just PRO measures but is still inferior to that of solely utilising PA features. If this behaviour is reflected on the test set, we could therefore conclude that when using this training method for learning a partial order with a neural network, the PA data is better for learning which cases are more exceptional. We will select the partial order with 20 values to assess in the evaluation.

Features	**Number of sub-periods**
10 Values	
PA Features	4
PRO Features	N/A
PA and PRO Features	2
20 Values	
PA Features	1
PRO Features	N/A
PA and PRO Features	1
30 Values	
PA Features	1
PRO Features	N/A
PA and PRO Features	1

Table 6.15: Strict Partial NN-AA-CBR Hyperparameters

	Accuracy	**Precision**	**Recall**	**F1**
10 Values				
PA features	0.572	0.577	0.560	0.510
PRO features	0.506	0.504	0.610	0.552
PA and PRO features	0.550	0.543	0.630	0.582
20 Values				
PA features	0.611	0.628	0.540	0.581
PRO features	0.578	0.559	0.760	0.630
PA and PRO features	0.550	0.546	0.600	0.526
30 Values				
PA features	0.589	0.585	0.620	0.571
PRO features	0.522	0.523	0.580	0.515
PA and PRO features	0.528	0.525	0.530	0.530

Table 6.16: Strict Partial NN-AA-CBR average performance over ten 3-fold cross-validations on the training dataset.

Graphical Representation

When comparing the graphical representation of the Strict Partial Ordered NN-AA-CBR, we observe that the path lengths from the default case to the focus case increase as the number of values that can be represented by the partial order increase. This is illustrated in Figure 6.11, Figure 6.12 and Figure 6.13. As a result, striking a balance between model performance with the explanations that can be generated becomes crucial. Models with a larger range of values in the partial order appear to reduce the ability to follow the explanations, as these explanations now contain more cases. Conversely, models with fewer values in the partial order perform worse, with more cases assigned to the same values. Using a neural network to learn the partial order poses a challenge in generating dialogical explanations, as the output value from the neural network lacks clinical significance. However, we could use the ordering generated by the Neural Network and then inspect the features of models to identify clinical interpretations. However, this requires additional effort from those reviewing the model.

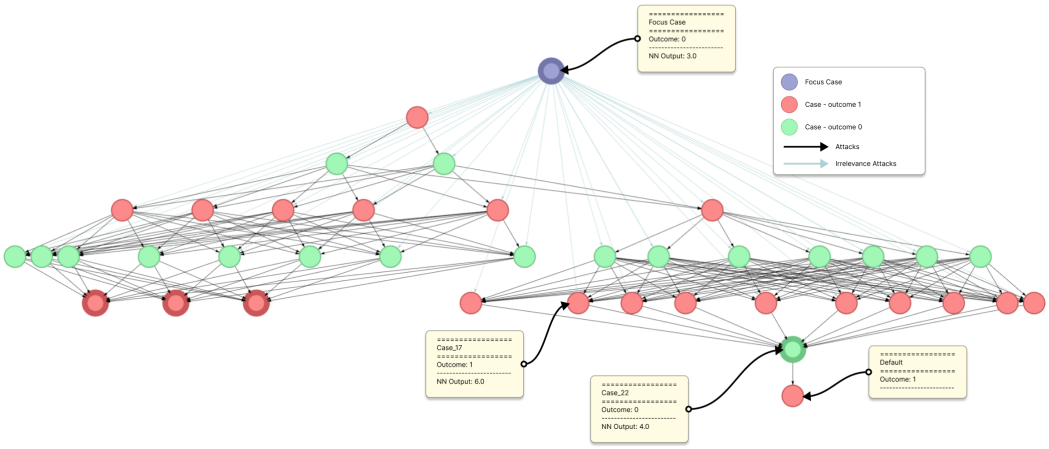


Figure 6.11: An example graphical AAF generated by Model 8 with 10 Values

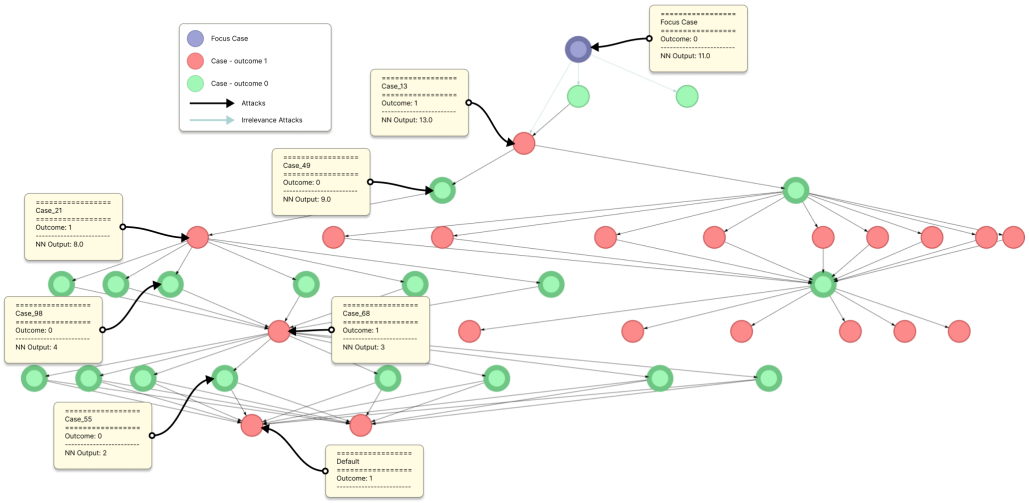


Figure 6.12: An example graphical AAF generated by Model 8 with 20 Values

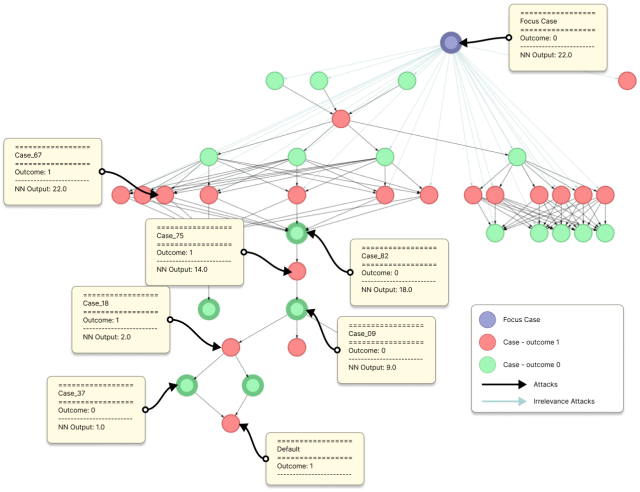


Figure 6.13: An example graphical AAF generated by Model 8 with 30 Values

Chapter 7

Evaluation

7.1 Model Evaluation Plan

The held-out test set is used for evaluating the model on unseen data and comparing it against each of the baseline models. As there are only 50 cases in the test set, we evaluate the models by conducting a 3-fold cross-validation, where for each iteration we randomly split the 50 cases such that two-thirds of the data can be used in the model case base and the other third can be used to make predictions. This cross-validation was conducted 10 times. The results presented in Section 7.3 are therefore an average of 30 runs of every model. For each argumentation model, we use the best hyperparameters for the data characterisation method as described in Chapter 6.

The AA-CBR models are evaluated using standard evaluation metrics: accuracy, precision, recall and F1 score. Accuracy assesses the proportion of correctly classified instances for both progressive disease and stable disease classifications. Precision measures the proportion of correctly predicted cases of patients having progressive disease out of all cases predicted as such. This metric indicates the model's ability to correctly identify true cases of progressive disease. Similarly, recall measures the proportion of correctly predicted cases of patients having progressive disease out of all cases that actually have progressive disease. Given the potential harm caused by misclassifying true instances of progressive disease, recall is a vital measure to identify the models' ability to not miss progressive disease classifications. Moreover, we look at F1 Score as an aggregation of precision and recall where a high F1 score is a result of both a high precision and recall, indicating the model's overall effectiveness in correctly identifying progressive disease cases. We compare these metrics against three baseline models, detailed in Section 7.2. Section 7.3 details our analysis of the model results.

Furthermore, we evaluate three variants of the models, one utilising solely PA features, one utilising solely PRO features and one incorporating both PA and PRO features. We will analyse which variants perform most optimally. Appendix A lists the specific features each model was run with.

In Section 7.4 we will examine the models from a clinical perspective, assessing what insights about the data and models can be learned from their interpretations. We will analyse the used characterisation methods in Section 7.4.1, reviewing the trade-off between interpretable characterisations, model performance and the effort required to determine these characterisations. Additionally, we evaluate explanations generated in Section 7.4.2 for the most interesting models to gain clinical insight into the data. It is important to recognise the inherent dialogical nature of the explanations and the significant challenge presented by empirical evaluation. Nonetheless, we will use the models to identify traits of the data. In Section 7.4.3, we examine how these traits are supported by the relevant research and propose a methodology to utilise the AA-CBR models to identify conflicts between the features.

7.2 Baseline Models

Each model is compared against three baseline models: A decision tree, a k-nearest neighbor (kNN) and a neural network. The same methodology for evaluating the performance of the dataset on these models is utilised as previously, a 3-fold cross-validation averaged over 10 iterations. Characterisation extraction methods will not be utilised for the baseline models. This means we are not applying thresholding, feature selection or sub-periods to the data used in the baseline models. Characterisation extraction is not used because decision trees and neural networks have the capability for feature selection and feature importance inherently. As for kNNs,

characterisation extraction is avoided to preserve the interpretability of the model, prevent information loss, and reduce the introduction of potential bias in classification results.

7.2.1 Decision Tree

Decisions Trees (see Background Section 2.1.1) were selected as a baseline model due to their capability to handle small datasets, work with missing features and produce models that can be easily interpreted. We aim to capture these desirable properties with our argumentation models. Therefore, we compare argumentation to decision trees in an effort to address limitations associated with decision trees, such as their tendency to overfit or their sensitivity to changes in training data. Decision Trees are the most similar of our baseline models to AA-CBR, making comparisons to them crucial.

7.2.2 K-Nearest Neighbor

K-Nearest Neighbor (see Background Section 2.1.2) as it is an easily interpretable model. Although a full analysis of model complexity is beyond the scope of this report, it is worth noting that a kNN has similar worse-case time complexity at the point of classification, requiring each data point to be compared to every other. This makes it a useful comparison to assess if the performance of the AA-CBR models can surpass the kNN at a comparable level of complexity whilst offering more comprehensive explanations. Hyperparameter tuning on the training dataset found that setting k to 3 with the Eucilidean distance achieved highest levels of performance.

7.2.3 Neural Network

The neural network (see Background Section 2.3) was selected as a baseline due to its powerful capabilities as a classifier. Furthermore, as we use neural networks for feature selection for certain models and have models that utilise neural networks to learn their partial order, it is appropriate to compare our models to a baseline neural network. Despite their ability to learn complex patterns, neural networks have significant drawbacks including a tendency to overfit with small datasets, an inability to handle missing values and a black-box nature reducing interpretability. Argumentation can address these limitations and so it is meaningful to compare AA-CBR to a neural network.

The neural network architecture is comprised of an input layer, a single hidden layer with 64 neurons and a Relu activation and, an output layer to a single neuron using a sigmoid output activation. The model was trained using Adam optimisation, with a learning rate of 0.01 for 15 epochs. This architecture and training process was chosen as it was found to be effective the training dataset and is consistent with the architecture used by the NN-AA-CBR models.

7.3 Model Performance Analysis

PA and PRO Features

Examining the results presented in Table 7.1, we see that the set-based models have the most optimal performance. Model 2: cAA-CBR offers performance comparable compared to Model 1: AA-CBR. As a result, we can benefit from the desirable properties of cAA-CBR, namely its handling of incoherent cases leading to clearer explanations. Moreover, we conclude that the most optimal is Model 3: AA-CBR with Dynamic Features when utilising both PA and PRO features. This model achieves the highest accuracy, 0.637, and F1 score, 0.685. In contrast, AA-CBR shows comparable performance with an accuracy of 0.619 and the same F1 score of 0.685. While AA-CBR with Dynamic Features outperforms both the kNN and neural network baselines it falls short compared to the decision tree baseline by a significant margin, which has an accuracy of 0.717 and an F1 score of 0.741. Decision trees can inherently conduct feature selection within a large feature set and identify data thresholds. Consequently, the decision tree excels at learning feature importance for the classification of progressive disease when all features are involved. This highlights the weakness of AA-CBR methodologies in general, as these models lack capabilities for characterisation extraction, thus requiring considerable effort by the model designers.

PA and PRO Features				
	Accuracy	**Precision**	**Recall**	**F1**
Baseline Models				
Decision Tree	**0.717**	**0.765**	**0.749**	**0.741**
k-Nearest Neighbour	0.420	0.506	0.578	0.510
Neural Network	0.540	0.547	0.659	0.574
AA-CBR Models				
Model 1: AA-CBR	0.619	0.640	0.763	**0.685**
Model 2: cAA-CBR	0.588	0.603	**0.824**	0.684
Model 3: AA-CBR Dynamic Features	**0.637**	0.660	0.733	**0.685**
Model 4: AA-CBR Euclidean Norm Order	0.546	0.581	0.653	0.599
Model 5: AA-CBR Absolute Product Order	0.540	**0.700**	0.403	0.480
Model 6: AA-CBR Sign and Magnitude Order	0.589	0.618	0.706	0.649
Model 7: NN-AA-CBR Total Order	0.517	0.563	0.592	0.549
Model 8: NN-AA-CBR Strict Partial Order	0.398	0.553	0.626	0.497

Table 7.1: Model Results with PA and PRO Features

PA Features				
	Accuracy	**Precision**	**Recall**	**F1**
Baseline Models				
Decision Tree	0.749	0.779	**0.783**	0.764
k-Nearest Neighbour	0.741	0.821	0.751	0.756
Neural Network	**0.761**	**0.824**	0.753	**0.770**
AA-CBR Models				
Model 1: AA-CBR	0.598	0.690	0.549	0.583
Model 2: cAA-CBR	0.701	0.688	0.857	0.760
Model 3: AA-CBR Dynamic Features	**0.758**	**0.831**	0.718	**0.763**
Model 4: AA-CBR Euclidean Norm Order	0.440	0.512	0.482	0.478
Model 5: AA-CBR Absolute Product Order	0.534	0.566	0.502	0.506
Model 6: AA-CBR Sign and Magnitude Order	0.541	0.652	0.502	0.560
Model 7: NN-AA-CBR Total Order	0.619	0.614	**0.967**	0.743
Model 8: NN-AA-CBR Strict Partial Order	0.580	0.644	0.587	0.602

Table 7.2: Model Results with PA Features

PRO Features				
	Accuracy	**Precision**	**Recall**	**F1**
Baseline Models				
Decision Tree	**0.569**	**0.621**	0.620	**0.594**
k-Nearest Neighbour	0.431	0.510	0.602	0.520
Neural Network	0.527	0.569	**0.685**	0.581
AA-CBR Models				
Model 1: AA-CBR	0.595	0.621	0.660	0.631
Model 2: cAA-CBR	0.580	0.612	0.688	0.638
Model 3: AA-CBR Dynamic Features	0.605	0.644	0.685	0.648
Model 4: AA-CBR Euclidean Norm Order	0.535	0.574	0.714	0.624
Model 5: AA-CBR Absolute Product Order	0.597	0.640	0.681	0.645
Model 6: AA-CBR Sign and Magnitude Order	**0.635**	**0.647**	**0.798**	**0.704**
Model 7: NN-AA-CBR Total Order	0.457	0.480	0.600	0.487
Model 8: NN-AA-CBR Strict Partial Order	0.561	0.592	0.764	0.638

Table 7.3: Model Results with PRO Features

Attempts to reduce the burden of feature characterisation by introducing models that operate directly on values of data points come at the expense of both model performance and interpretability. Overall, the set-based AA-CBR models outperform all of the value-oriented models. Of these value-oriented models, Model 6: AA-CBR Sign and Magnitude Order exhibits the best performance with an accuracy of 0.589 and an F1 score of 0.649. This outcome is expected, as this partial order aligns most similarly with the characterisation of the set-based models. In contrast, Model 5: AA-CBR Absolute Product Order has inferior performance. This emphasises the importance of the direction of change in the values and validates our choice for utilising direction with AA-CBR Sign and Magnitude as well as set-based models. However, we find that comparing the values with the Sign and Magnitude Order results in a more stringent definition of exceptionality compared with set-based models, resulting in the decreased performance seen. The performance of Model 4: AA-CBR Euclidean Norm Order exceed the kNN baseline, which is notable as they both rely on Euclidean distance metric for evaluating the cases. Using an AA-CBR model with Euclidean distance in these circumstances results in better performance than a kNN.

The neural network learned partial order methods show the least optimal performance. These methods do not surpass the neural network baselines. Notably, Model 7: NN-AA-CBR Total Order demonstrates better performance than Model 8: NN-AA-CBR Strict Partial Order. Using a strict partial order was motivated by the need to create a learned AA-CBR model that was more interpretable than the total order method. However, the current implementation has come at a significant performance cost, with NN-AA-CBR Strict Partial Order unable to successfully classify cases with an accuracy of less than 50%. On the other hand, NN-AA-CBR Total Order does not outperform the other total order method, AA-CBR Euclidean Norm Order which does not use a neural network. This illustrates how the learned methods are underperforming and indicates a potential for future improvement.

PA Features

The analysis of the results in Table 7.2 shows that models solely utilising PA features generally outperform models utilising both PA and PRO features or solely PRO features. Model 3: AA-CBR with Dynamic Features once again demonstrates the most optimal performance of the AA-CBR models, achieving an accuracy of 0.758 and an F1 score of 0.763. Remarkably, the model exhibits comparable performance to all of the baseline models. This is in contrast to utilising both PA and PRO features. Additionally, Model 2: cAA-CBR outperforms Model 1: AA-CBR. In this case, with the reduced number of features the likelihood of incoherence increases. Consequently, using an algorithm that can accommodate this noise leads to considerable performance gains.

Notably, the value-oriented models do not show an increase in performance when limited to PA features alone. These models each perform comparably with no clear winner between Model 5: AA-CBR Absolute Product Order and Model 6: AA-CBR Sign and Magnitude. Model 4: AA-CBR Euclidean Norm Order is outperformed by these two models. With set-based models the degree to which arguments attack one another is larger than with AA-CBR Absolute Product Order and AA-CBR Sign and Magnitude. This means that set-based models are more prone to incoherence. Consequently, for PA features using set-based models and handling the subsequent noise (as with cAA-CBR) results in far better performance than using models with a more stringent definition of exceptionality. The value-oriented models fall considerably short of the baseline models, which excel when run with just the PA features. This suggests that whilst our characterisations and partial orders explored perform poorly, potential performance gains may exist with different characterisations or partial orders.

We see that the NN-AA-CBR methods outperform the value-oriented models and demonstrate comparable performance to Model 1. We could thus infer PA data is better than PRO data for learning which cases should be considered more exceptional than others. However, our experiments do not fully support this conclusion, as it is possible that the neural networks used do not have the capacity to learn the partial order given the number of input features when combined with PRO features. Further experimentation is required to explore this aspect. Nevertheless, this result supports the potential capabilities of using a neural network to learn the partial order and motivates future work in this area. The relatively high recall of 0.967 and lower precision of 0.614 seen by NN-AA-CBR Total Order highlights the model's bias towards classifying cases with progressive disease. This is because the model utilises the default case in circumstances when it is unable to effectively classify a case which we observe to occur more often with this model.

PRO Features

The results presented in Table 7.3 showcase model performances when utilising solely PRO features. The set-based models exhibit inferior performance compared to the PA feature variants of the same models. This is to be expected given we see a similar decline in performance when comparing the baseline models used with the different features. Furthermore, the set-based models outperform the baseline models with PRO features.

Notably, the best-performing set-based model and second best-performing model overall is Model 3: AA-CBR with Dynamic Features once again, with an accuracy of 0.604 and an F1 score of 0.648. This is a considerable increase on the best-performing baseline model, the decision tree, which achieves an accuracy of 0.569 and an F1 score of 0.594. Consequently, these results suggest it is more challenging to predict progressive disease using PRO features but the characterisation extraction methods and the use of argumentation for PRO features appear to mitigate this difficulty to an extent.

Model 2: cAA-CBR demonstrates comparable performance to Model 1: AA-CBR. The removal of incoherent cases by cAA-CBR did not improve model performance. Inspecting the dataset with the characterisation extraction parameters of this model revealed that there we no incoherent cases so we do not gain the benefits from cAA-CBR that we hope to.

In contrast to the models utilising solely PA features or a combination of PA and PRO features, the value-oriented models show relatively strong performance. Particularly, Model 6: AA-CBR Sign and Magnitude have the best performance of all AA-CBR models with PRO features, with an accuracy of 0.645 and F1 score of 0.704. This is a significant increase in performance compared to the baseline models. Whilst not performing as strongly, Model 4: AA-CBR Euclidean Norm Order and Model 5: AA-CBR Absolute Product Order exceed the expectations set by the model variants with PA features or PA and PRO features. These values-oriented models use more stringent criteria to identify when a case is considered more exceptional than another, suggesting that models with PRO features perform better under these conditions.

Furthermore, Model 8: NN-AA-CBR Strict Partial Order outperforms Model 7: NN-AA-CBR Total Order. This outcome is not observed with the models that utilise solely PA features or a combination of PA and PRO features. NN-AA-CBR Strict Partial Order even outperforms the baseline models, with an accuracy of 0.561 and an F1 score of 0.638. The baseline neural network exhibited inferior performance with an accuracy of 0.527 and an F1 score of 0.581. Despite the NN-AA-CBR models' worse performance than the other AA-CBR models, these findings support our choice to use neural networks for learning partial orders and that further research in this area could prove fruitful.

Overall Performance Comparison

Overall, the results show that AA-CBR models are effective for the classification of progressive disease. The best models consistently performed comparably to or outperformed the baselines.

The PA models generally outperformed the PRO models and the models using a combination of both PA and PRO features. From a clinical perspective, this supports the conclusion that PA data can be used to supplement or replace PRO measures. This trend is observed with the baseline models and with the argumentation models, further highlighting the use of argumentation and specifically AA-CBR models as a predictive classifier used with real-world clinical data.

Models utilising solely PRO features generally show lower performance in classifying progressive disease. Firstly, we see that the performance of both AA-CBR models and the baseline models is significantly inferior to that of utilising solely PA features. Secondly, we note that more stringent conditions for identifying exceptional cases, such as AA-CBR Sign and Magnitude Order, are required to achieve better classification performance on PRO features. This requirement to achieve better performance indicates that the underlying data is harder to classify because we need to place more emphasis on identifying the most exceptional features and the more subtle differences in feature values. Conversely, better performance with less stringent conditions, as seen in models with PA features, suggests that a larger range of features and variations in the data can be accommodated, as the underlying data exhibit clearer boundaries between classes.

Model 3: AA-CBR with Dynamic Features consistently showed the best performance. This suggests that including temporal information in disease tracking is crucial to predicting future instances of progressive disease. Furthermore, we have shown that we are able to effectively leverage argumentation to include this information using an adapted version of the model from the literature. Models such as decision trees typically require handling time values in the same way as any other feature whilst neural networks employ more complicated architectures such as RNNs or LSTMs in order to separate the time component. It is significant that AA-CBR with Dynamic Features can utilise changes over time and handle the time component independently of the other features. Consequently, AA-CBR can work with longitudinal data whilst being clinically relevant, easily interpretable and high performing.

The value-oriented models showed relatively weak performance compared to the set-based models. Model 4: AA-CBR Euclidean Norm Order had the most relaxed ordering, being a total order and performed poorly across the board. Model 5 was a stricter model, comparing cases on a per-feature basis but not taking into account the direction of the feature, and outperforms the Euclidean Norm Order model but lacked the strictness required

to perform effectively with PRO features and is too strict for PA features. However, Model 6: AA-CBR Sign and Magnitude offered an appropriate balance resulting in higher model performance. This indicates that while value-oriented models reduce the burden of characterisation extraction, this reduced effort comes at the expense of model performance.

Although NN-AA-CBR models do not perform optimally, we observe that they demonstrate improved performance when utilising solely PA features. This suggests that there is potential for model improvements. Further research is required to explore neural networks in learning partial orders.

7.4 Clinical Discussion

Beyond the performance, we analyse the attributes of the models and the characterisation methods from a clinical perspective assessing the relevancy of the models, the usefulness of PA data as a new metric and evaluating the explanations generated. Furthermore, we present a methodology for identifying data conflicts.

7.4.1 Characterisation Extraction Analysis

To effectively utilise real-world data from a clinical trial in AA-CBR models, it is necessary to consider some key criteria: selecting a clinically relevant characterisation, ensuring sufficient performance of AA-CBR models in predicting progressive disease, and ensuring transparency and interpretability of AA-CBR models.

For a clinically relevant characterisation, we chose to encode the PA data and PRO data using percentage changes from the baseline. Further characterisation extraction was then conducted for each specific model. The set-based models and value-oriented models, excluding Model 4: Euclidean Norm Order offer clinically relevant characterisations. For the set-based, the characterisation involves labelling each feature as either "Increased" or "Decreased". This is a clear and easy-to-interpret representation that accurately reflects the underlying clinical representation albeit void of the details that quantify the changes. On the other hand, Model 5: Absolute Partial Order and Model 6 Sign and Magnitude Order, which compare features using their values, do not abstract the underlying data which could make them more useful for making clinical decisions but comes at the cost of added complexity. The Euclidean Norm Order method and neural network-based AA-CBR models do not use a clinically interpretable characterisation of the data for their models. For the NN-AA-CBR methods, our goal was to explore the possibility of learning the partial using more automated methods but for clinical use, further work would need to explore characterising the data in more clinically interpretable methods.

Characterisation Parameters Analysis

To ensure the sufficient performance of the AA-CBR models, we tuned the characterisation extraction methods to achieve optimal performance on a training data set, exploring feature selection, thresholding and varying representations of the PA time series.

Of our methods for selecting features, the inclusion ranking approach consistently resulted in superior performance on the training set and was the optimal feature selection method used in the majority of cases. Conversely, the autoencoder approach adapted from the literature [28] and the neural network classifier were only used when solely utilising PA features. Thus, feature selection processes that utilise the AA-CBR models as part of their methodology, such as inclusion ranking, appear to result in better performance. Furthermore, our analysis of the model results shows that it was more challenging to classify progressive disease using PRO features. Therefore, the neural network classifier approach to feature selection appears to only be effective when there the underlying data exhibit clearer class boundaries. Nevertheless, our best-performing model was Model 3: AA-CBR Dynamic Features utilising solely PA features. The features used, "Sedentary" and "Tasks-Light", were selected by the neural network classifier approach. This suggests that when working with real-world data trialling a variety of approaches for feature selection for AA-CBR is crucial for identifying the most suitable approach for the task.

We observed that for AA-CBR, using smaller sets of features to represent the data tends to be superior. For example, when using both PA and PRO features, Model 3 only had 7 out of 32 features selected in order to achieve optimal performance. Furthermore, models performed worse when the 8-week focus period was split into sub-periods. The best models represent the PA data as an average across the full 8 weeks. This prompts the need for dimensionality reduction techniques when using AA-CBR with multi-dimensional data sets.

For AA-CBR, thresholding the magnitude of the change is clearly important. We previously identified that using partial orders with more stringent criteria is important for data that is harder to classify as we need to identify the most exceptional cases for attacks. The same principle applies to thresholds, where we note that in general, models that use PRO features require thresholds to be set higher than models solely utilising PA

features. Larger thresholds help identify when features are considered more exceptional. However, our approach was relatively broad, using only two thresholds, one for all PA features and one for all PRO features. More fine-grained thresholding on individual features and different thresholds depending on the direction of change would likely lead to better model performance. As there are considerably more PRO features, using the same threshold of change for all of them discounted some features that would have otherwise had an impact on the classification. A more fine-grained approach comes at the cost of considerably more effort in hyperparameter tuning and model design. Future research should aim to identify methods of reducing this burden and finding an appropriate trade-off between model performance and tuning effort.

7.4.2 Explanations Analysis

For our analysis, we will examine the explanations generated by one set-based method, Model 3: AA-CBR Dynamic Features, one value-oriented method, Model 6 AA-CBR Sign and Magnitude Order and one neural network based method, Model 8 NN-AA-CBR Strict Partial Order. We will present a dialogical explanation of the outcome of focus cases from the test set for each model and then discuss the interpretability of the model. We present a series of decisions made by a sample decision tree for comparison.

AA-CBR Dynamic Features

Figure 7.1 illustrates the condensed representation of AA-CBR Dynamic Features. An example dialogical explanation is as follows: The Loser (L), claims that the default case applies and that the new case should be classified with progressive disease. The Winner (W), argues that Case_16 is an exception to the default with the features BNMD_Decreased (Motor Dysfunction), PA_Decreased (Pain), BNSE_Decreased (Seizures), AP_Decreased (Appetite Loss) and NV_Decreased (Nausea) and outcome of stable disease. L counters by stating that Case_45 is an exception to Case_16, presenting more specific features, namely tasks-light_0_Increased and BNCD_Increased (Communication Deficit). W wins the argument by stating that Case_45 is irrelevant to the focus case as Case_45 occurs at a later timestep, where Case_45 has previously experienced progressive disease whilst the focus case has not. As a result, the default argument does not hold and the focus case can be classified as having stable disease.

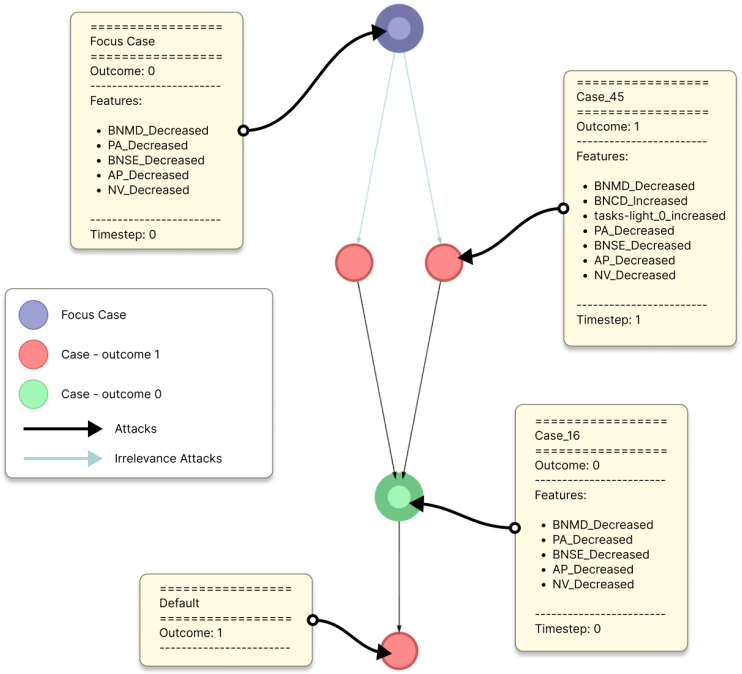


Figure 7.1: Condensed representation generated by Model 3 on the Test Set

The set-based characterisations are simple to understand and the presence of additional features by the attacking cases clearly explains why they are an exception to the case they attack. The line of reasoning provided can support clinical decision-making by providing sufficient context with respect to the features of an individual focus case.

As previously demonstrated AA-CBR with Dynamic Features is the overall best-performing model. A key insight here is that acknowledging previous instances of progressive disease is effective in classifying future instances of progressive disease or stable disease. Beyond model performance, this is evident in the explanations derived from the AA-CBR model. Dynamic Features allow for a longitudinal analysis of patients' disease progression. This means that the model can take into account the changes and developments in MRI outcomes over multiple time points, providing a more detailed understanding of a patient's cancer journey. As a result, the model can better adapt its classifications to the specific circumstances of each patient. This contributes to our goal of a patient-centred approach.

AA-CBR Sign and Magnitude

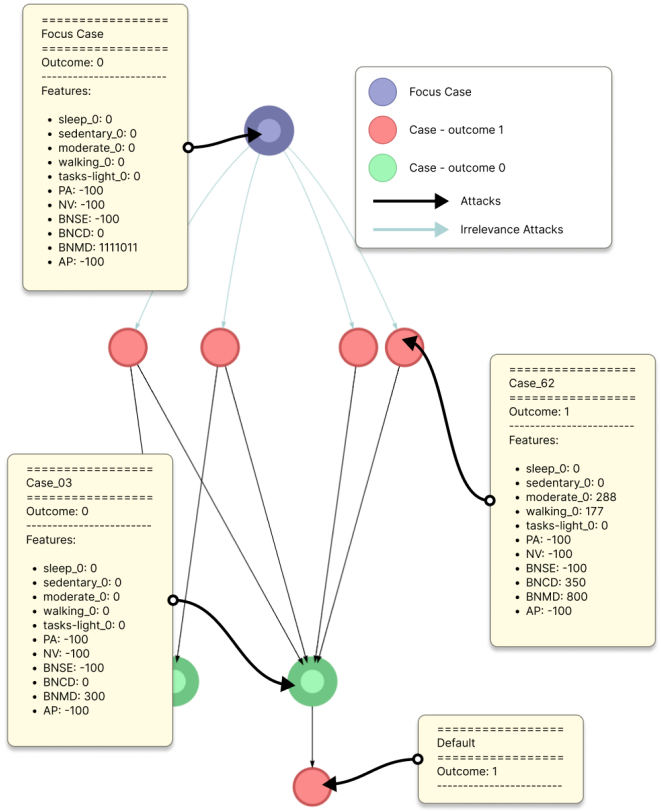


Figure 7.2: Condensed representation generated by Model 6 on the Test Set

In contrast, the dialogical explanations derived from the AA-CBR Sign and Magnitude Order contain more detail, requiring a more in-depth understanding of the partial order to effectively interpret the model. We can derive a dialogical argument from Figure 7.2. The Loser (L) claims the default case holds. The Winner (W) counters with Case_03, with the features PA: -100, NV: -100, BNSE: -100, BNMD: 300, AP: -100 and all other features set to 0. L responds with Case_62 which has features with the same direction of change but the features moderate_0: 288, walking_0: 177, BNCD: 350 and BNMD: 800 have a larger magnitude with all other features the same as with Case_03. W wins the argument by stating that Case_62 is irrelevant to the Focus Case because either Case_62 does not have the same direction of change for all features or at least one feature in Case_62 has a larger magnitude. We can see from the values that Case_62 has the features moderate_0: 288, walking_0: 288 and BNCD: 350 which all have larger magnitudes than the same features in the Focus Case (which are all 0). W's argument is unattacked and therefore the default argument does not hold.

Evidently, this explanation contains more details than the set-based method, with the specific values involved in the characterisation of the cases and a more complex relationship structure. This detail offers more insight into the underlying data compared to the set-based models benefiting clinical settings where concrete values can support reasoning. Future research will involve assessing the relevance of this model in a wider clinical setting and comparing it with the more abstract set-based models to determine which offers explanations that are more effective for clinical decision-making.

NN-AA-CBR Strict Partial Order

Compared to these previous models, the NN-AA-CBR Strict Partial Order models provide significantly less detail and lack clinical relevance with their characterisation. An example dialogical argument derived from Figure 7.3 is as follows: The Loser (L) claims the default outcome holds. The Winner (W) counters with Case_20, which has a value of 1 in the partial order. L responds with Case_84 with 2.0 in the partial order which W counters with Case_72 which has 3.0 in the partial order. L then uses Case_93 to counter further with a value of 11.0 in the partial order. W claims that Case_93 is irrelevant as it has a larger value than the focus case and thus wins the argument.

This explanation fails to provide an understanding of the underlying data, with no reasoning behind the black-box neural network's assignment. Evidently, this model is more difficult to follow and lacks interpretability. This model could not be deployed in a clinical setting, with no clear interpretation of the model and poor performance.

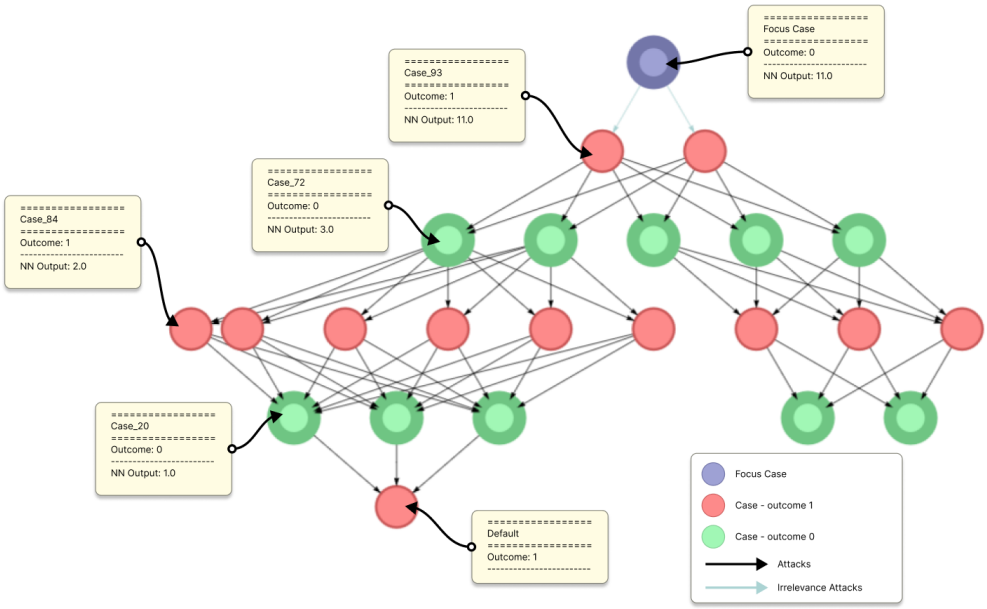


Figure 7.3: Condensed representation generated by Model 8 with 20 Values on the Test Set

Decision Tree

Decision Trees are also capable of generating interpretable models. We can compare these explanations to the ones generated by AA-CBR. We have generated a sample decision tree as shown in Figure 7.4. To illustrate the example explanation, we have limited the feature set to be the same as with AA-CBR Sign and Magnitude and maintained the same focus case.

The focus case has a BNMD value of 1111011 which is greater than -90 so we move to the right child node. Again, the BNMD value is greater than 50 so we move to the right. The value of walking is 0, which is smaller than 285.723 so we move to the left child node. The value of NV is -199 which is less than 33333233.5 so move to the left child. The value of BNMD is greater than 149.998 so move to the right child. The value of BNCD is 0 which is greater than 249.997 so we move to the right child. The value of BNMD is smaller than 27777678 so we move to the left. All leaf nodes from this point are classed as Stable disease so we classify the focus case with stable disease.

This provides a step-by-step evaluation of the features in the focus case, leading to the classification of stable disease. This explanation is easy to follow and as with the AA-CBR Sign and Magnitude method provides details about the values of the data. However, understanding the reasoning behind the ordering of the tree and the thresholds set is more complicated. The construction process is not transparent and requires a deep understanding of the underlying algorithms used to learn the decision tree. These algorithms often prioritise model performance, focusing on metrics such as information gain or probability of misclassification which may not align with clinical intuition. As a result, the decision tree may not offer valuable insights from a clinical perspective.

In contrast, AA-CBR utilises the relationships between data points to reason about the classifications. This reveals more complex patterns within the data. This is evident in the explanations generated by AA-CBR with Dynamic Features and AA-CBR Sign and Magnitude. AA-CBR explanations go beyond simply explaining why a classification has been assigned but provide insights into why the case has not been assigned the opposite classification. On the other hand, decision trees oversimplify, only providing a step-by-step series of boolean decisions. This line of reasoning presents boolean decisions without providing the context necessary for a clinical situation.

However, the overview of the training data provided by the decision tree allows for more general rules to be learned about the dataset. While complex relationships in AA-CBR provide detailed reasoning, it is more difficult to learn general rules about features in the data. Future research could explore methods of generalising the data before applying argumentation, such as by using k-means clustering to aggregate the data. This approach would mitigate overfitting and provide more general explanations that may have greater clinical significance and work across a broader range of scenarios.

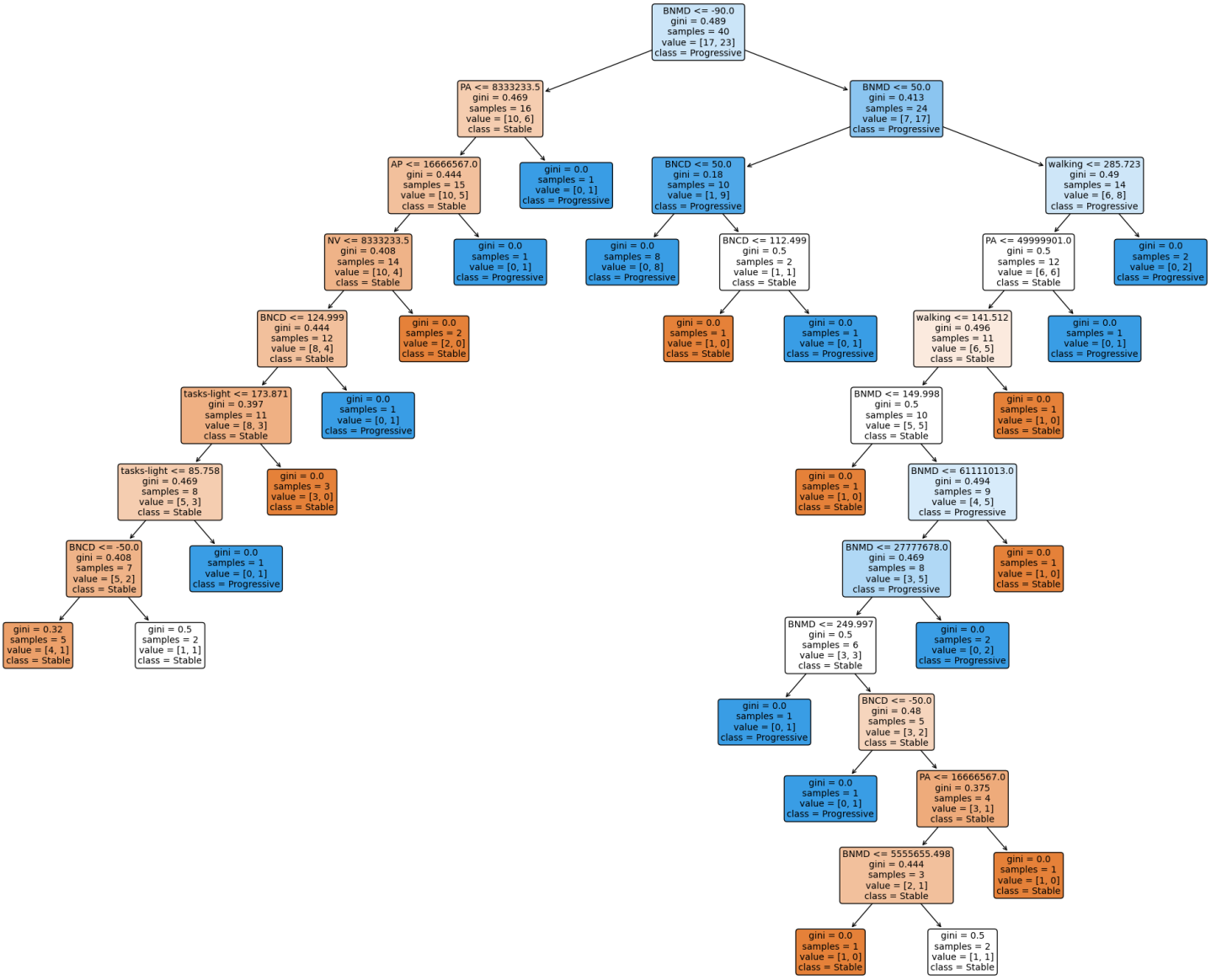


Figure 7.4: A decision tree generated on the Test Set

Patient-Centered Approach

The explanations derived from AA-CBR models have a patient-centred perspective, where the focus is on an individual's circumstances at a specific time point along their cancer journey. Rather than solely relying on aggregate population data, as with decision trees, we explain the specific outcome of the individual by comparing and contrasting their circumstances to other patients. This is a unique patient-centred machine-learning approach that provides the necessary context and detail for clinical decision-making.

7.4.3 Feature Conflicts

Feature Selection

We examine the choice of features that work most optimally for the AA-CBR models. The features selected for each model can be viewed in Appendix A. Of the PA features, the average acceleration recorded at each 30-second epoch is used the least. This is intriguing as the functional behaviours (Sleep, Sedentary, Moderate, Tasks-Light and Walking) are derived from the value of acceleration, thus we'd expect there to be a high correlation between the use of acceleration and the use of the functional behaviours. However, by encoding this acceleration into separate categories, we can consider when each class is most exceptional in the AA-CBR models. Additionally, using the functional behaviours provides a more intuitive interpretation of the acceleration data but we must note that these functional behaviours are generated from the pre-processing methods that are not likely to not be 100% accurate.

Generally, most models see optimal performance utilising all the functional behaviours and we note that Tasks-Light is used in all models. In fact, Tasks-Light was also the feature that was consistently identified as the most important feature by all feature selection methods, followed by Sedentary. We also note that the best-performing AA-CBR model was Model 3 utilising only these two features. Tasks-Light was the behaviour recorded the least across the 31 patients in the study which possibly explains why the presence of this feature is considered exceptional by the AA-CBR models and contributed to the discriminative power of PA features.

In contrast, for PRO features, we see that the most selected features are Global Health Status (QL2), Physical Functioning (PF2), Cognitive Functioning (CF), Fatigue (FA), Appetite Loss (AP), Visual Disorder (BNVD), Motor Dysfunction (BNMF) and Communication Deficit (BNCD). This is interesting because Global Health Status, Physical Functioning, Fatigue and Motor Dysfunction were shown to be clinically linked and correlate to activity levels in the BrainWear study [7]. This suggests that the feature selection methods utilised are able to identify clinically relevant features that can effectively discriminate instances of progressive and stable disease.

A Methodology for Identifying Conflicts

However, correlations identified by the BrainWear study do not exceed a Spearman's Rank correlation coefficient of larger than 0.53. This suggests that the correlations identified are not strong. To illustrate, the study identified a coefficient of 0.28 between the patient-reported symptoms of fatigue and the recorded time spent sleeping. Although this positive correlation suggests that an increase in sleep duration would be accompanied by an increase in fatigue, and vice versa, the correlation is weak. We can utilise AA-CBR models to identify cases where these features are conflicting and do not follow the expected trend.

For example, we can create a cAA-CBR model utilising the same features examined in the BrainWear study: Global Health Status, Physical Functioning, Fatigue, Future Uncertainty and Motor Dysfunction as PRO measures and the functional behaviour Sleep as the sole PA feature. We have chosen to use cAA-CBR to reduce incoherence in the model as we are using a small feature set and it can be used to identify excess features responsible for the classification. By iteratively considering each case as the focus case and the rest of the cases as the case base, we can inspect the explanations generated. Specifically, we analyse scenarios where sleep increase is a feature of the focus case but the set of excess features contains solely fatigue increase. This conflict occurs when the classification of a focus case with increased sleep is explained by the lack of increased fatigue, contrary to the positive correlation expected. There are four such scenarios we need to consider: where the focus case contains sleep increased but fatigue increased is an excess feature, the converse where the focus case contains fatigue increased but sleep increased is an excess feature and similarly for cases where both features decrease.

Using this methodology, we identified 73 cases in which the classification of progressive disease was a result of a conflict between fatigue and sleep. Table 7.4 shows the number of explanations associated with each type of conflict. Note that the total number of explanations can exceed the number of cases as there can be more than one explanation for a case's assigned outcome. Notably, the majority of explanations state that the assigned outcome is because cases have a sleep feature but do not contain the corresponding fatigue feature that would otherwise be expected. This shows how PA data clearly supplements PRO measures in the lines of reasoning generated.

Focus Case Feature	Excess Feature	No. Explanations
Sleep Decreased	Fatigue Decreased	34
Sleep Increased	Fatigue Increased	39
Fatigue Decreased	Sleep Decreased	1
Fatigue Increased	Sleep Increased	6

Table 7.4: Number of explanations with conflicts

This process can be easily scaled to support more features. Furthermore, clinicians can inspect individual cases where conflicts occur and identify groups of patients based on clinically relevant features such as age, sex, surgery type, type of treatment, duration of radiotherapy and more. This provides clinicians with additional context regarding discrepancies between patient reports and recorded physical activity data, facilitating more comprehensive decision-making in patient care. While an in-depth clinical analysis is beyond the scope of this report, we have clearly demonstrated that these AA-CBR models can identify feature conflicts, enabling detailed analysis to determine the extent to which PA data can supplement or replace PRO data.

Identifying conflicts in the data is extremely valuable. Considering that while general trends will exist within the population, individual cases may not follow these trends. Disease progression is a complex process and the correlations between symptoms and behaviours may not always follow expected patterns due to unique individual circumstances. Additionally, identifying conflicts allows clinicians to consider when beyond the collected data

is necessary. This highlights the need for personalised approaches to healthcare and showcases how AA-CBR models can be used to gain insight into complex relationships between various features and their influence on disease progression. AA-CBR, therefore, provides a valuable framework for the longitudinal monitoring of patients with PA data and PRO measures. By continuously tracking feature conflicts, clinicians can gain insights into the dynamic nature of the disease and make informed decisions regarding patient care.

7.4.4 Default Outcome and Recall

It is important to note that the selection of the default outcome needs to accurately reflect the objectives of the clinicians. Additionally, any bias associated with the selected default outcome must be acknowledged if we are to develop a transparent model. In our study, we selected the chosen default outcome to prioritise excess caution in the classifications of progressive disease.

Notably, the recall of the best-performing models is relatively high and comparable to or exceeding the baseline models. The recall of the models is clinically significant as it quantifies the models' ability to accurately classify cases of progressive disease without misclassifying them. This is particularly relevant when utilising PA and PRO features where the best-performing models exhibit a recall of greater than 0.7. In contrast, for the baseline models, only the decision tree achieves such a high recall. The high recall observed can be attributed to using progressive disease as the default outcome.

This is clinically significant as it means the models assign this more cautious outcome effectively. In contrast, the baseline models do not use a default case and so lack the same level of clinical caution as the AA-CBR models. Using the default outcome to establish pre-dispositions about the status of patient disease is an effective way to inject clinical preferences into the model. For the use case of predicting disease status, the selected pre-disposition aligns with our goals.

Chapter 8

Conclusion

8.0.1 Summary

Our study has successfully demonstrated that our novel use of AA-CBR models is effective at predicting the status of patient disease utilising PA and PRO data, exceeding baseline models. Our research is the first to utilise Machine Learning techniques to analyse the utility of Physical Activity data in the BrainWear study. Importantly, we have found that the performance of our models is superior when utilising Physical Activity data compared to relying solely on the established Patient Reported Outcomes. We have showcased how AA-CBR models offer transparency and interpretability, providing explanations that are easily understood and support a cautious patient-centred approach to healthcare.

Furthermore, we have effectively characterised real-world datasets to make them suitable for use with AA-CBR. We have also introduced the Inclusion Ranking method for feature selection for AA-CBR. Our results clearly illustrate that AA-CBR models can generate clinically significant lines of reasoning for predicting progressive disease and that this result is supported by the characteristics of the data and relevant research. Our characterisation approaches can result in models that are high-performing, interpretable and can handle limited data sets with missing values, despite the effort to fine-tune the characterisation methods.

We have demonstrated that AA-CBR with Dynamic Features represents the most effective approach, offering high performance and clear, easy-to-follow explanations. This method shows how ML techniques can evaluate data points and create lines of reasoning where time is handled as a separate feature. Furthermore, we have shown that it is effective and clinically significant to consider previous instances of progressive disease in the future prediction of patient status.

Additionally, we have introduced novel value-oriented variants of AA-CBR that can achieve good performance and provide more detailed explanations, which can provide more evidence for clinical decision-making. We identified the Sign and Magnitude partial order as the most effective value-oriented AA-CBR approach. In addition, we lay the foundations for further research into a novel approach to AA-CBR utilising neural networks to learn the partial order of the cases, automatically characterising the features requiring less effort for characterisation tuning.

Moreover, we have outlined a method to identify conflicts within the data which can be used to reason about which types of data are better and for which groups of patients. We have illustrated how this method can be used for a deeper clinical analysis of the data, identifying patients that do not follow the trends of the population and allowing for explanations as to why this is the case.

The insights of our study have significant potential for the use of AA-CBR in healthcare and more broadly the use of argumentation on complex, real-world data sets. The insights that can be drawn from AA-CBR show the potential to support clinical decision-making, which has practical implications for improving disease prediction and managing patient HRQoL.

8.0.2 Future Work

Building upon the findings of our study, several avenues for future research emerge. From a clinical perspective, a trial into the use of these techniques in a practical clinical setting, and reviewing the benefits provided to clinicians would be invaluable to verifying and refining these models. A more in-depth clinical study that utilises the data characterisation process, explanations generated and the methodology for identifying conflicts can lead to a more in-depth analysis of the utility of PA data in healthcare. Interactive tools for designing, running and visualising AA-CBR models can be designed to work in a clinical setting to aid clinical decision-making from a patient-centred approach.

Additionally, assessing the techniques utilised in our study on similar real-world data sets could enhance the support for the use of AA-CBR in healthcare. BrainWear, for instance, collected PRO questionnaires beyond the EORTC QLQ-C30/BN20 such as the Montreal Cognitive Assessment (MoCA) [43] and the Multidimensional Fatigue Inventory (MFI) [44] which we could be compared against. Further clinical research into the use of wearables in healthcare and, in particular, patients with High-Grade Gliomas, can provide larger datasets to verify our results and assess the effectiveness of the methods utilised.

Furthermore, argumentation methods that are designed to analyse conflicts more directly can be developed based on the methodology in [30]. An approach such as this could offer more insights into the conflicts existing in the data set and develop lines of reasoning that support or oppose the use of PA data.

Moreover, other variants of AA-CBR could be explored, in particular variants that utilise time components such as AA-CBR Dynamic Features. Additionally, cAA-CBR variants of the models explored could lead to performance gains and further insights into characterising data given that incoherence would be able to be handled effectively.

Exploring other methods of characterising the data such as utilising LSTMs for PA data to better capture the complex representations of time. This would have to be balanced with ensuring models remain interpretable. Experiments with clustering methods could be applied prior to fitting the AA-CBR models to reduce overfitting and increase model performance. Additionally, we characterised data as percentage changes and used thresholding to identify exceptional changes. Future work could explore other representations such as using the raw values and thresholding based on deviation from the mean.

We introduced NN-AA-CBR which utilises neural networks to learn partial orders. These showed potential in reducing the burden of data characterisation otherwise required by AA-CBR. Future research into NN-AA-CBR can explore other methods of training, utilising different types of neural networks or classifying the data points into known groups and building custom partial orders over these groups.

Appendix A

Feature Tables

Measure	Shortcode	Question(s)
QLQ-C30		
Global Health Status	QL2	29, 30
Physical Functioning	PF2	1, 2, 3, 4, 5
Role Functioning	RF2	6, 7
Emotional Functioning	EF	21, 22, 23, 24
Cognitive Functioning	CF	20, 25
Social Functioning	SF	26, 27
Fatigue	FA	10, 12, 18
Nausea and Vomiting	NV	14, 15
Pain	PA	9, 19
Dyspnoea	DY	8
Insomnia	SL	11
Appetite Loss	AP	13
Constipation	CO	16
Diarrhoea	DI	17
Financial Difficulties	FI	28
BN20		
Future Uncertainty	BNFU	31, 32, 33, 35
Visual Disorder	BNVD	36, 37, 38
Motor Dysfunction	BNMD	40, 45, 49
Communication Deficit	BNCD	41, 42, 43
Headaches	BNHA	34
Seizures	BNSE	39
Drowsiness	BNDR	44
Itching Skin	BNIS	47
Hair Loss	BNHL	46
Weakness of Legs	BNWL	48
Bladder Control	BNBC	50

Table A.1: EORTC scoring scales

PA Features	Model 1: AA-CBR	Model 2: cAA-CBR	Model 3: Dyn AA-CBR: Days on Study	Model 3: Dyn AA-CBR: Prev Cases	Model 3: Dyn AA-CBR Prev PD	Model 4: Euclidean Norm Order	Model 5: Absolute Product Order	Model 6: Sign and Magnitude Order
Acceleration						X	X	
Sleep	X	X	X	X		X	X	X
Sedentary	X	X	X	X	X	X	X	X
Moderate	X	X	X	X		X	X	X
Tasks-Light	X	X	X	X	X	X	X	X
Walking	X	X	X	X		X	X	X

Table A.2: PA Features Selected for each model

PRO Features	Model 1: AA-CBR	Model 2: cAA-CBR	Model 3: Dyn AA-CBR: Days on Study	Model 3: Dyn AA-CBR: Prev Cases	Model 3: Dyn AA-CBR Prev PD	Model 4: Euclidean Norm Order	Model 5: Absolute Product Order	Model 6: Sign and Magnitude Order
QL2	X	X	X	X	X	X		X
PF2	X	X	X	X	X	X		X
RF2							X	
EF								
CF	X	X	X	X	X	X		X
SF								
FA	X	X	X	X	X	X	X	X
NV							X	
PA							X	
DY								
SL								
AP	X	X	X	X	X	X	X	X
CO								
DI								
FI							X	
BNFU							X	
BNVD	X	X	X	X	X	X		X
BNMD	X	X	X	X	X	X	X	X
BNCD	X	X	X	X	X	X		X
BNHA								
BNSE								
BNDR								
BNIS								
BNHL								
BNWL								
BNBC								

Table A.3: PRO Features Selected for each model

PA and PRO Features	Model 1: AA-CBR	Model 2: cAA-CBR	Model 3: Dyn AA-CBR: Days on Study	Model 3: Dyn AA-CBR: Prev Cases	Model 3: Dyn AA-CBR Prev PD	Model 4: Euclidean Norm Order	Model 5: Absolute Product Order	Model 6: Sign and Magnitude Order
Acceleration							X	
Sleep			X			X	X	X
Sedentary			X			X	X	X
Moderate	X	X	X			X	X	X
Tasks-Light	X	X	X	X	X	X	X	X
Walking	X	X	X			X	X	X
QL2	X	X	X			X	X	
PF2	X	X	X			X	X	
RF2								
EF								
CF							X	
SF								
FA	X	X	X			X	X	
NV				X	X			X
PA				X	X			X
DY								
SL								
AP	X	X	X	X	X	X	X	X
CO								
DI								
FI								
BNFU								
BNVD	X	X	X			X	X	
BNMD	X	X	X	X	X	X	X	X
BNCD	X	X	X	X	X	X	X	X
BNHA								
BNSE				X	X			X
BNDR								
BNIS								
BNHL								
BNWL								
BNBC								

Table A.4: PA and PRO Features Selected for each model

Bibliography

[1] Linda T Kohn, Janet M Corrigan, Molla S Donaldson, and editors. *To Err Is Human*. National Academies Press, March 2000. doi: 10.17226/9728. URL https://doi.org/10.17226/9728.

[2] Nitin Ohri, Rafi Kabarriti, William R. Bodner, Keyur J. Mehta, Viswanathan Shankar, Balazs Halmos, Missak Haigentz, Bruce Rapkin, Chandan Guha, Shalom Kalnicki, and Madhur Garg. Continuous activity monitoring during concurrent chemoradiotherapy. *International Journal of Radiation Oncology - Biology - Physics*, 97(5):1061–1065, April 2017. doi: 10.1016/j.ijrobp.2016.12.030. URL https://doi.org/10.1016/j.ijrobp.2016.12.030.

[3] Nitin Ohri, Balazs Halmos, William R. Bodner, Haiying Cheng, Chandan Guha, Shalom Kalnicki, and Madhur Garg. Daily step counts: A new prognostic factor in locally advanced non-small cell lung cancer? *International Journal of Radiation Oncology - Biology - Physics*, 105(4):745–751, November 2019. doi: 10.1016/j.ijrobp.2019.07.055. URL https://doi.org/10.1016/j.ijrobp.2019.07.055.

[4] Phan Minh Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2):321–357, 1995. ISSN 0004-3702. doi: https://doi.org/10.1016/0004-3702(94)00041-X. URL https://www.sciencedirect.com/science/article/pii/000437029400041X.

[5] Katie Atkinson, Pietro Baroni, Massimiliano Giacomin, Anthony Hunter, Henry Prakken, Chris Reed, Guillermo Simari, Matthias Thimm, and Serena Villata. Towards artificial argumentation. *AI magazine*, 38(3):25–36, 2017.

[6] Seema Dadhania, Lillie Pakzad-Shahabi, Kerlann Le Calvez, Waqar Saleem, James Wang, Waleed Mohammed, Sanjay Mistry, and Matthew Williams. BrainWear: Longitudinal, objective assessment of physical activity in 42 high grade glioma (HGG) patients. *Neuro-Oncology*, 23(Supplement_4):iv3–iv3, October 2021. doi: 10.1093/neuonc/noab195.006. URL https://doi.org/10.1093/neuonc/noab195.006.

[7] Seema Dadhania, Lillie Pakzad-Shahabi, Sanjay Mistry, and Matt Williams. Triaxial accelerometer-measured physical activity and functional behaviours among people with high grade glioma: The BrainWear study. *PLOS ONE*, 18(5):e0285399, May 2023. doi: 10.1371/journal.pone.0285399. URL https://doi.org/10.1371/journal.pone.0285399.

[8] Roger Stupp, Warren P. Mason, Martin J. van den Bent, Michael Weller, Barbara Fisher, Martin J.B. Taphoorn, Karl Belanger, Alba A. Brandes, Christine Marosi, Ulrich Bogdahn, Jürgen Curschmann, Robert C. Janzer, Samuel K. Ludwin, Thierry Gorlia, Anouk Allgeier, Denis Lacombe, J. Gregory Cairncross, Elizabeth Eisenhauer, and René O. Mirimanoff. Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *New England Journal of Medicine*, 352(10):987–996, 2005. URL https://doi.org/10.1056/NEJMoa043330.

[9] David N Louis, Arie Perry, Pieter Wesseling, Daniel J Brat, Ian A Cree, Dominique Figarella-Branger, Cynthia Hawkins, H K Ng, Stefan M Pfister, Guido Reifenberger, Riccardo Soffietti, Andreas von Deimling, and David W Ellison. The 2021 WHO Classification of Tumors of the Central Nervous System: a summary. *Neuro-Oncology*, 23(8):1231–1251, 06 2021. ISSN 1522-8517. doi: 10.1093/neuonc/noab106. URL https://doi.org/10.1093/neuonc/noab106.

[10] Jin-xiang Cheng, Xiang Zhang, and Bo-Lin Liu. Health-related quality of life in patients with high-grade glioma. *Neuro-Oncology*, 11(1):41–50, 02 2009. ISSN 1522-8517. doi: 10.1215/15228517-2008-050. URL https://doi.org/10.1215/15228517-2008-050.

[11] Ecog performance status scale - ecog-acrin cancer research group, Jun 2022. URL https://ecog-acrin.org/resources/ecog-performance-status/.

[12] Rebecca Featherston, Laura E. Downie, Adam P. Vogel, and Karyn L. Galvin. Decision making biases

in the allied health professions: A systematic scoping review. *PLOS ONE*, 15(10):1–15, 10 2020. doi: 10.1371/journal.pone.0240716. URL https://doi.org/10.1371/journal.pone.0240716.

[13] Leo Anthony Celi, Jacqueline Cellini, Marie-Laure Charpignon, Edward Christopher Dee, Franck Dernoncourt, Rene Eber, William Greig Mitchell, Lama Moukheiber, Julian Schirmer, Julia Situ, Joseph Paguio, Joel Park, Judy Gichoya Wawira, Seth Yao, and for MIT Critical Data. Sources of bias in artificial intelligence that perpetuate healthcare disparities—a global review. *PLOS Digital Health*, 1(3):1–19, 03 2022. doi: 10.1371/journal.pdig.0000022. URL https://doi.org/10.1371/journal.pdig.0000022.

[14] Christopher J. Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*, 17(1), October 2019. doi: 10.1186/s12916-019-1426-2. URL https://doi.org/10.1186/s12916-019-1426-2.

[15] Joanne E Croker, Dawn R Swancutt, Martin J Roberts, Gary A Abel, Martin Roland, and John L Campbell. Factors affecting patients' trust and confidence in GPs: evidence from the english national GP patient survey. *BMJ Open*, 3(5):e002762, 2013. doi: 10.1136/bmjopen-2013-002762. URL https://doi.org/10.1136/bmjopen-2013-002762.

[16] Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". 2016. doi: 10.48550/ARXIV.1606.08813. URL https://arxiv.org/abs/1606.08813.

[17] Kristijonas Cyras, Ken Satoh, and Francesca Toni. Abstract argumentation for case-based reasoning. In *Fifteenth international conference on the principles of knowledge representation and reasoning*, 2016.

[18] Thomas Davenport and Ravi Kalakota. The potential for artificial intelligence in healthcare. *Future Healthcare Journal*, 6(2):94–98, June 2019. doi: 10.7861/futurehosp.6-2-94. URL https://doi.org/10.7861/futurehosp.6-2-94.

[19] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, August 2015. doi: 10.1145/2783258.2788613. URL https://doi.org/10.1145/2783258.2788613.

[20] Ethem Alpaydin. *Introduction to Machine Learning*, pages 213–238. 2014.

[21] Zhongheng Zhang. Introduction to machine learning: k-nearest neighbors. *Annals of Translational Medicine*, 4(11):218–218, June 2016. doi: 10.21037/atm.2016.03.37. URL https://doi.org/10.21037/atm.2016.03.37.

[22] Charles Antaki and Ivan Leudar. Explaining in conversation: Towards an argument model. *European Journal of Social Psychology*, 22(2):181–194, 1992.

[23] Ramon Lopez de Mantaras. Case-based reasoning. In *Machine Learning and Its Applications*, pages 127–145. Springer Berlin Heidelberg, 2001. doi: 10.1007/3-540-44673-7_6. URL https://doi.org/10.1007/3-540-44673-7_6.

[24] Kristijonas Cyras, David Birch, Yike Guo, Francesca Toni, Rajvinder Dulay, Sally Turvey, Daniel Greenberg, and Tharindi Hapuarachchi. Explanations by arbitrated argumentative dispute. *Expert Systems with Applications*, 127:141–156, 2019. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2019.03.012. URL https://www.sciencedirect.com/science/article/pii/S0957417419301654.

[25] Breast cancer statistics, May 2022. URL https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer#heading-Zero.

[26] Wei Yang, Nicole M Warrington, Sara J Taylor, Paula Whitmire, Eduardo Carrasco, Kyle W Singleton, Ningying Wu, Justin D Lathia, Michael E Berens, Albert H Kim, et al. Sex differences in gbm revealed by analysis of patient imaging, transcriptome, and survival data. *Science translational medicine*, 11(473): eaao5253, 2019.

[27] Oana Cocarascu, Andria Stylianou, Kristijonas Cyras, and Francesca Toni. Data-empowered argumentation for dialectically explainable predictions. In *ECAI 2020*, pages 2449–2456. IOS Press, 2020.

[28] Toni F Cocarasc U O, Cyras K. Explanatory predictions with artificial neural networks and argumentation. In *Proceedings of the 2ndWorkshop onExplainable Artificial Intelligence (XAI 2018)*, May 2018. URL http://hdl.handle.net/10044/1/62202.

[29] Guilherme Paulino-Passos and Francesca Toni. Monotonicity and noise-tolerance in case-based reasoning with abstract argumentation. In *Proceedings of the Eighteenth International Conference on Principles of*

Knowledge Representation and Reasoning. International Joint Conferences on Artificial Intelligence Organization, September 2021. doi: 10.24963/kr.2021/48. URL https://doi.org/10.24963/kr.2021/48.

[30] Anthony Hunter and Matthew Williams. Aggregating evidence about the positive and negative effects of treatments. *Artificial Intelligence in Medicine*, 56(3):173–190, 2012. ISSN 0933-3657. doi: https://doi.org/10.1016/j.artmed.2012.09.004. URL https://www.sciencedirect.com/science/article/pii/S0933365712001194.

[31] Wei Di. *Deep Learning Essentials*. Packt Publishing, 1st edition edition, 2018. ISBN 1-78588-036-5.

[32] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006. doi: 10.1126/science.1127647. URL https://www.science.org/doi/abs/10.1126/science.1127647.

[33] Richie Koch. What is considered personal data under the eu gdpr?, Feb 2019. URL https://gdpr.eu/eu-gdpr-personal-data/.

[34] Matthew Harvey. Imperial college research computing service, 2017. URL https://data.hpc.imperial.ac.uk/resolve/?doi=2232.

[35] Irene Chen, Fredrik D. Johansson, and David Sontag. Why is my classifier discriminatory?, 2018. URL https://arxiv.org/abs/1805.12002.

[36] Government Digital Service. Discrimination: Your rights, Mar 2015. URL https://www.gov.uk/discrimination-your-rights.

[37] Aaronson NK Ahmedzai S Bergman B Bullinger M Cull A Duez NJ Filiberti A Flechtner H Fleishman SB de Haes JCJM Kaasa S Klee MC Osoba D Razavi D Rofe PB Schraub S Sneeuw KCA Sullivan M Takeda F. The european organisation for research and treatment of cancer qlq-c30: A quality-of-life instrument for use in international clinical trials in oncology. *Journal of the National Cancer Institute*, 85:365–376, 1993.

[38] Fayers PM Aaronson NK Bjordal K Groenvold M Curran D Bottomley A. on behalf of the eortc quality of life group. the eortc qlq-c30 scoring manual (3rd edition). 2001.

[39] Rosemary Walmsley, Shing Chan, Karl Smith-Byrne, Rema Ramakrishnan, Mark Woodward, Kazem Rahimi, Terence Dwyer, Derrick Bennett, and Aiden Doherty. Reallocation of time between device-measured movement behaviours and risk of incident cardiovascular disease. *British journal of sports medicine*, 56(18):1008–1017, 2022.

[40] Aiden Doherty, Karl Smith-Byrne, Teresa Ferreira, Michael V Holmes, Chris Holmes, Sara L Pulit, and Cecilia M Lindgren. Gwas identifies 14 loci for device-measured physical activity and sleep duration. *Nature communications*, 9(1):1–8, 2018.

[41] Matthew Willetts, Sven Hollowell, Louis Aslett, Chris Holmes, and Aiden Doherty. Statistical machine learning of sleep and physical activity phenotypes from sensor data in 96,220 uk biobank participants. *Scientific reports*, 8(1):1–10, 2018.

[42] Aiden Doherty, Dan Jackson, Nils Hammerla, Thomas Plötz, Patrick Olivier, Malcolm H Granat, Tom White, Vincent T Van Hees, Michael I Trenell, Christoper G Owen, et al. Large scale population assessment of physical activity using wrist worn accelerometers: the uk biobank study. *PloS one*, 12(2):e0169649, 2017.

[43] Robert Anton Olson, Taruna Chhanabhai, and Michael McKenzie. Feasibility study of the montreal cognitive assessment (MoCA) in patients with brain metastases. *Supportive Care in Cancer*, 16(11):1273–1278, March 2008. doi: 10.1007/s00520-008-0431-3. URL https://doi.org/10.1007/s00520-008-0431-3.

[44] E.M.A. Smets, B. Garssen, B. Bonke, and J.C.J.M. De Haes. The multidimensional fatigue inventory (MFI) psychometric qualities of an instrument to assess fatigue. *Journal of Psychosomatic Research*, 39(3):315–325, April 1995. doi: 10.1016/0022-3999(94)00125-o. URL https://doi.org/10.1016/0022-3999(94)00125-o.