The University Of Sheffield.

Explainability in advanced manufacturing: leveraging the interpretability of multi-criteria decision making and neutrosophic logic

Hesham Ali Hasan Ali Hasan Yusuf

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

in the

Computational Intelligence Systems Laboratory
Automatic Control & Systems Engineering
The University of Sheffield

November 2022

Declaration of Authorship

I, Hesham Ali Hasan Ali Hasan Yusuf, declare that this thesis titled, “Explainability in advanced manufacturing: leveraging the interpretability of multi-criteria decision making and neutrosophic logic” and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: **Hesham Ali Hasan Ali Hasan Yusuf**

Date: **22-Nov-2022**

“The only true wisdom is in knowing you know nothing.”

Socrates

Explainability in advanced manufacturing: leveraging the interpretability of multi-criteria decision making and neutrosophic logic

Abstract

Interpretability has been a vital aspect of modelling since the emergence of machine learning. Despite this, the dominance of non-transparent models due to their arguable superior performance meant that interpretable transparent models were seldom used, especially in data-driven applications. Interpretability is key to reaching explainability. Thus, interpretable models are the most promising way to achieve this. In this thesis, a new class of interpretable models (multi-criteria decision making (MCDM)) is investigated, for the first time, in a series of industrial and academic applications toward achieving explanation. The MCDM model is shown to achieve enhanced interpretability following its extension with fuzzy logic. The Fuzzy-MCDM model's *interpretability* enables the generation of output explanations. Consequently, the model's interpretability is improved further by introducing neutrosophic logic. The proposed models are applied to benchmark and industrial pipe inspection datasets. The experimental results demonstrate the framework's capability for generating meaningful explanations; while maintaining good performance. The purpose of explaining a model's result is to pave the way for broad adoption in fields where a decision's *accountability* and *transparency* are paramount due to the high stakes of the decisions. These areas include biomedical, aviation, nuclear and advanced manufacturing. Machine learning adoption is lacking in high stake areas due to the lack of explanation, an obstacle preventing the acquisition of trust from the experts. The thesis describes how an interpretable modelling framework is adapted to reduce the performance trade-off often attributed to transparent models while exploiting the advantages to generate useful *explanatory* information - a clear advantage over opaque models.

Acknowledgements

I would like to acknowledge The Welding Institute Ltd. and National Structural Integrity Research Centre for hosting my PhD. This endeavour would not have been possible without the financial support of the Lloyds Register Foundation and the University of Sheffield's Department for Automatic Control & System Engineering. I would like to extend my sincere thanks to the PolyTest project team for their support.

I could not have undertaken this journey without my supervisory committee. I would like to express my most profound appreciation to Prof. George Panoutsos for his invaluable support and guidance.

I am deeply indebted to Dr Kai Yang for his comprehensive support and guidance at The Welding Institute Ltd.

I am also thankful to Dr Channa Nageswaran for his mentorship and direction.

Special thanks to Hannah Stedman for her encouragement and support.

I am also grateful to Hab for her tremendous support throughout the writing up.

Thanks should also go to Hazem for his generous words of encouragement.

Many thanks to *Lunch Buddies* for organising all those lunches and tea breaks. Special thanks to Vats, Islam, Javid, Raheem and Asad for all the good times at River Lane.

I would like to extend my sincere thanks to the examination committee, Prof. Mohammad Osman Tokhi and Prof. Mahdi Mahfouf, for their generous constructive feedback and support.

I am also thankful for the support of my family. Despite my family not knowing much about what I do, they did not hesitate to express a strong belief in my abilities. I would not have a chance of accomplishing this feat without their tremendous unconditional and unwavering support.

Contents

List of Figures

List of Tables

List of Abbreviations

A-Scan	amplitude-scan
ACC	accuracy
AI	artificial intelligence
ANN	artificial neural networks
ATOVIC	amended fused TOPSIS-VIKOR for classification
B-Scan	brightness-scan
BF	butt-fusion
DAU	data acquisition unit
DT	decision trees
EF	electro-fusion
FIS	fuzzy inference system
FLS	fuzzy logic set
FL	fuzzy logic
FMR	F-measure
GCLM	gray level co-occurrence matrix
GDPR	general data protection regulation
HDPE	high-density polyethylene
ID	inner diameter
KNN	k nearest neighour

LIME	local interpretable model-agnostic explanations
MCDM	multi criteria decision making
MF	Membership function
ML	machine learning
MMFCs	mel frequency cepstral coefficients
NDT	non-destructive testing
NIS	neutrosophic inference system
NL	neutrosophic logic
OD	outer diameter
PAUT	phased array ultrasonic testing
PD	Parkinson's disease
SOA	state of the art
TNR	true negative rate
TOPSIS	technique for order of preference by similarity to ideal solution
TPR	true positive rate
UCI	University of California, Irvine
UT	ultrasonic testing
VIKOR	VIseKriterijumska Optimizacija I Kompromisno Resenje (translation: Multicriteria Optimization and Compromise Solution)
XAI	explainable artificial intelligence

For my family…

1 Introduction

The level of understanding of tools and techniques has a considerable impact on ensuring their safe and efficient use. This especially applies to AI, the prevalence of which is growing at an unprecedented pace. The International Data Corporation (IDC) forecasts that global AI spending to exceed $500 billion by 2023 [1]. The popularity of opaque AI, in particular, has uncovered a considerable limitation; a lack of transparency and, in turn, explainability. Pivotal to the driving forces of AI lies a dominant methodology; robust[1] machine learning (ML) algorithms that are perceived as standard data-in-data-out black box systems.

As a result, some researchers have tried to address this by prescribing the only solution available for black box models: post-hoc interpretability. Post-interpretability, for short, is an interpretation methodology that is designed to make black box models *explainable*. The key phrase is "designed to". Although post-interpretability has shown potential, the mathematical proof for this is far from formal. This is because post-interpretability seeks to model a black box model with no regard for its structure.

Therefore, post-interpretability is merely a band-aid solution to a deep-rooted issue in opaque ML: the lack of intrinsic interpretability. Simply introducing interpretability to the ML model comes with challenges. Most recent research efforts have focused on black box models. Hence, inherently interpretable methodologies have yet to develop to the same level.

The social implications of non-transparent modelling can be far-reaching. Rudin et al. [2] has warned that post-hoc interpretability is doing more harm than good. The researchers give an example of how a black box model denied a prisoner's parole

[1] robust: powerful ML techniques such as deep learning, and support vector machines and Naive Bayes

in the United States criminal justice system [2]. It is a landmark example of how undetected bias in modelling can lead to a detrimental impact on human life and potential society at large.

Groups of researchers have argued for model-based interpretability, as an alternative to post-hoc [2], [3]. Model-interpretability is an intrinsic property present in transparent models. However, opting for model-based interpretability comes with its own set of challenges.

Most notably, maintaining the model-based interpretability of transparent models comes at the cost of limiting the model's dimensionality. Thus, the solution of just adding more parameters or layers to improve performance is not a viable option; the route for appropriating deep learning models for more complex problems.

The implementation of AI has undoubtedly improved the lives of many. Nonetheless, this comes at the cost for an unlucky few, as in the case of the denied parole described earlier.

The same analogy applies to advanced manufacturing, where undetected defects can lead to potentially catastrophic incidents. For instance, when defects are left undetected in commercial jets in service, a potential consequence is a fatal crash. Notably, there are cases where cracks appear in blades long before their designed operation time [4], [5].

1.1 Transparency

Transparency is a double-edged sword, as could be a variety of things. As a concept, it is discussed extensively in a variety of different domains such as politics, media and philosophy. In 2022, the world values information more than anything else, so based on this, is non-transparency even an option? Some researchers argue that excess transparency can lead to dire unintended consequences. For instance, McGivern and Fischer [6] describe a multitude of potential practical and social implications, such as doctors operating defensively and rising levels of blame. In software publishing, open-source is occasionally considered a security risk, especially by organisations in critical domains such as nuclear or finance [7]–[9].

German philosopher Byung-Chul Han has criticised the extent to which society is compelled to be transparent in a "totalitarian" manner dispelling important ideals such as trust and the right to secrecy at a societal and political level [10]. In AI, transparency is considered a "nice to have" in many applications. For instance, AI used to unlock smartphones does not need to explain why it could not detect your face. Countless other applications exist where a lack of transparency is not particularly problematic, e.g. personalised advertising or search engines.

Nonetheless, transparency is perceived as a force for good. For many, increased transparency is a prime driver of accountability and openness. It is vital that the public is well-informed. Transparent media can provide much-needed clarity on certain topics in an informative and educative manner. In medical consultations, a transparent patient-clinician relationship does wonders for medical diagnosis, trust and compliance [11]. In cyber security, unresolved vulnerabilities in software and services are publicised, as part of the Common Vulnerabilities and Exposures system, as an effort of resolution and prevention. In logistics, customers are now provided with detailed tracking information related to a shipment.

Transparency in AI is similarly discussed. Lipton [12] explains that a few humans, if any, are interpretable, and thus, engineers could be asking too much in the interpretability of models. However, this could be true for some applications. It is clear that explainability is vital to AI's success in safety-critical areas such as healthcare [13]. Ruden et al. [2] recommends employing inherently interpretable methodology in high stake application; to avoid the pitfalls of so-called explainable black box models.

In AI, transparency is widely agreed on to be a beneficial feature. However, researchers still disagree on what explainability is and how it can be attained.

1.2 Problem statement

In decision-making, the consequence of a wrongful decision varies widely. A common conundrum for some is deciding on what to have for dinner on the takeout night. Not enjoying your meal would be one of the risks when you happen to select a mediocre restaurant or dish. User ratings, for instance, serve as a decision support tool to assist customers in selecting a good restaurant. The consequence of the wrong decision, however, is often minimal and short-lived.

By comparison, detecting defects in safety-critical parts is a different story. Inspection processes of critical parts are highly regulated. Therefore, the implementation of new tools, techniques and methodologies is scrutinised carefully. This level of precaution acts as a protective barrier against the severe consequence of wrongful decision-making.

For this reason, more care has to be taken when designing decision support solutions for high stake applications. One of the key requirements is documentation and justification. In the absence of automated models, the experts are tasked to justify all decisions they make in a systematic manner.

Therefore, providing means for *justification* is tantamount to any decision support tool for non-destructive testing (NDT) inspection. The idea of explanation is to provide insight into why or how a decision was arrived at. Thus, explainability is likely to serve as a promising starting point for justification support.

A review of current techniques for explanation revealed that two main categories exist: model-based or post-hoc [12]. Although post-hoc has been seen to provide an explanation of robust models, it is inconclusive whether the descriptive accuracy is of an adequate level. However, model-based interpretability is expected to provide better explanation capabilities. In terms of classification performance, the literature points to a potential trade-off when opting from intrinsically interpretable models. However, the theory of a trade-off has been discounted by prominent researchers in the area [2].

When there is no trade-off, users of interpretable models are likely to benefit from more insightful explanations. An explanation that is inferred rather than predicted. Interpretable models possess a transparent structure, such as the internal variables and parameters.

A research gap worth addressing is the lack of dependable data-driven explainable modelling frameworks for classification. This gap has driven designers to opt for non-explainable data-driven models. Although these models provide satisfactory performance, their explanatory capability is not suitable for high stake applications.

Consequently, a strong descriptive accuracy seeks to pave the way for fully observable models. A model that indicates to its designers and users why it is behaving the way it is. This seems like a given. However, this is not true for many prominent methodologies such as deep learning and Bayesian modelling.

1.3 Research aims and objectives

Multi-Criteria Decision Making is a set of modelling techniques that provide decision-making support based on a set of often conflicting criteria. By design, multi criteria decision making (MCDM) mimics how a human was to decide between a set of alternatives. MCDM is a sub-discipline of Operations Research - a study of how organisations manage their operations effectively and efficiently.

The advantage of MCDM is its structure. Techniques such as TOPSIS enable the use of human-understandable criteria to support decision-making. As a result, TOPSIS satisfies model interpretability guidelines and hence, is considered a viable candidate for an explainable modelling framework.

However, literature has seldom shown investigations of MCDM for classification - a prominent gap. MCDM's interpretable nature was seen as a promising starting point for developing an explainable data-driven framework for classification.

An explanation is a key to supporting high stake decision-making, where experts are often tasked with producing well-supported justification. Employing non-explainable methodologies suffers from the crucial limitation of requiring the expert to perform the classification process manually. Thus, any non-explainable model would not provide a tangible benefit to the inspection process without providing means for decision justification.

The project aim is to develop a data-driven interpretable classification framework, capable of producing meaningful explanation.

1.3.1 Objectives

A.1 Investigate and propose a ML modelling framework that enhances the accuracy of existing state-of-the-art interpretable MCDM methods.

  A.1.1 Identify literature gaps and investigate effective methods of constructing interpretable models as part of MCDM methods.

  A.1.2 Build computational frameworks that utilises systematically expert knowledge in the model construction/tuning; demonstrate how this knowledge enhances MCDM.

B.1 Propose a framework to make use of the interpretability by

  B.1.1 Design and implement algorithms and metrics to enhance tracing and analysing decision-making in linguistic form within MCDM.

  B.1.2 Utilise the information framework in B.1.1, to enhance traceability and understanding of decisions in MCDM in linguistic form.

C.1 To apply frameworks in A, B to Autonomous Defect Recognition (ADR)

  C.1.1 The ADR shall be semi-autonomous; performing all tasks for defect recognition semi-autonomously with a degree of human intervention.

  C.1.2 The ADR shall be fully autonomous; performing all tasks for defect recognition fully autonomously with no or minimal human interaction.

1.4 Achievements and contributions

1.4.1 MCDM as a classifier

MCDM's inherent transparency makes it a promising candidate for interpretable modelling. However, as a sub-discipline of operations research, it has been tailored more for dealing with supporting operational decision-making. Example situations include supplier selection or deciding which database package to use. Instead of features, MCDM often utilises human-understandable criteria to extract relevant decision supporting information. For instance, ranking techniques would produce a rank for each choice in a list of feasible alternatives. The ranks are determined based on various model parameters such as criteria and weights.

Literature seldom demonstrates MCDM-based methodologies being explored or investigated for classification [14]. Nevertheless, Baccour [14] illustrated areas where an MCDM-based classifier can achieve satisfactory classification performance. In **Chapter 3**, MCDM-based classifiers are compared with state-of-the-art classification techniques. The proposed data-driven classifiers are demonstrated to provide satisfactory performance compared to conventional techniques.

1.4.2 Fuzzy-MCDM classifiers

Mamdani-type fuzzy logic (FL) can be an inherently transparent branch of fuzzy theory. Mamdani can be constructed using expert knowledge rather than pure data-driven fitting. Therefore, Mamdani-type FL is considered suitable for maximising opportunities for model-based interpretability.

Pure-MCDM classifiers lack algorithmic transparency in their final classification stage. MCDM was extended with a fuzzy inference system (FIS) to replace the final classification process. The resulting Fuzzy-MCDM provides more insight into the decision by summarising the MCDM measures concisely using a single output.

The Fuzzy-MCDM's output is a *continuous* fuzzy class output ranging from 0 to 1. The distance of the fuzzy class output from the classification threshold indicates the classification's confidence.

Extending MCDM with FL had a minimal, if any, impact on performance. The increased interpretability of Fuzzy-MCDM is the rationale for its use.

1.4.3 MCDM classifier explanation framework

As will be discussed in further detail, interpretability and explainability are distinct concepts. Although all explainable models are interpretable, the opposite is not true. In Chapter 4, an explanation framework is proposed for MCDM-based classifiers.

The framework presents graphical and textual explanations to the user. Graphical explanations illustrate the impact of individual features on the classification result. Meanwhile, textual explanation seeks to describe the state of different sub-components of the classifier. A concluding linguistic statement explains whether the two sub-models are in agreement.

The proposed design acts as a starting point toward devising a universal explanation framework for explaining MCDM classifiers.

1.4.4 Neutrosophic-MCDM classifiers

The FL component enhances transparency noticeably. Nevertheless, the fuzzy class output does not present counterfactual and indeterminate information separately. Rather, all the information is summarised in a single output. This benefits the cause of simplicity.

However, for the sake of further explainability, counterfactual information would need to be distinguished from the factual. Neutrosophic Logic (NL) is a generalisation of FL that handles *falsity* and *indeterminacy* in addition to degree of truth. This level of granularity in the processing of information favours the explainability objective.

The Neutrosophic-MCDM classifier maintains a data-driven structure and classification performance while providing a breakdown of the class output into the three neutrosophic components: truth, indeterminacy and falsity.

1.4.5 Ultrasonic pipe inspection

Effective high-density polyethylene (HDPE) pipe inspection is vital to their safety in several key industries such as water and gas transport. Failing to detect defects can lead to catastrophic incidents. In Chapters 5 and 6, a real-world pipe weld dataset is used to test the proposed techniques and methodologies.

Defects and beads (a key indication) are classified with more than 85% overall accuracy across the model types. Insight on how the classifiers decide on a certain way is generated using the explanation framework. The insights provide an indication of the most influential features.

The use of MCDM for ultrasonic inspection of pipe welds provided decision support with insights. This is only possible by using inherently interpretable frameworks in conjunction with explanation capabilities.

1.4.6 Conference publications

- Published: H. Yusuf and G. Panoutsos, "Multi-criteria decision making using Fuzzy Logic and ATOVIC with application to manufacturing," in 2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Glasgow: IEEE Inc., Jul. 2020, pp. 1–7.
- Published: H. Yusuf, K. Yang, and G. Panoutsos, "Fuzzy Multi-Criteria Decision-Making: Example of an Explainable Classification Framework," in UK Workshop on Computational Intelligence 2021, Aberystwyth, Wales: Springer, Cham, 2021, pp. 15–26.
- Published: H. Yusuf, K. Yang, and G. Panoutsos, "Improving the Explainability of Multi-Criteria Decision-Making using Neutrosophic Logic," in UK Workshop on Computational Intelligence 2022, Sheffield, UK: Springer, Cham, 2022.
- Planned: H. Yusuf, K. Yang and G. Panoutsos, "Neutrosophic Logic as an Explanation Framework for Multi-Criteria Decision-Making".

1.5 Thesis outline

The thesis structure is outlined in Figure 1.1. Background on key topics of relevance to this project is presented in **Chapter 2**. The main topics include interpretability/explainability, AI/ML and inspection in advanced manufacturing. The chapter provides an overview of explainable artificial intelligence (XAI). A summary of key ideas and theories is presented.

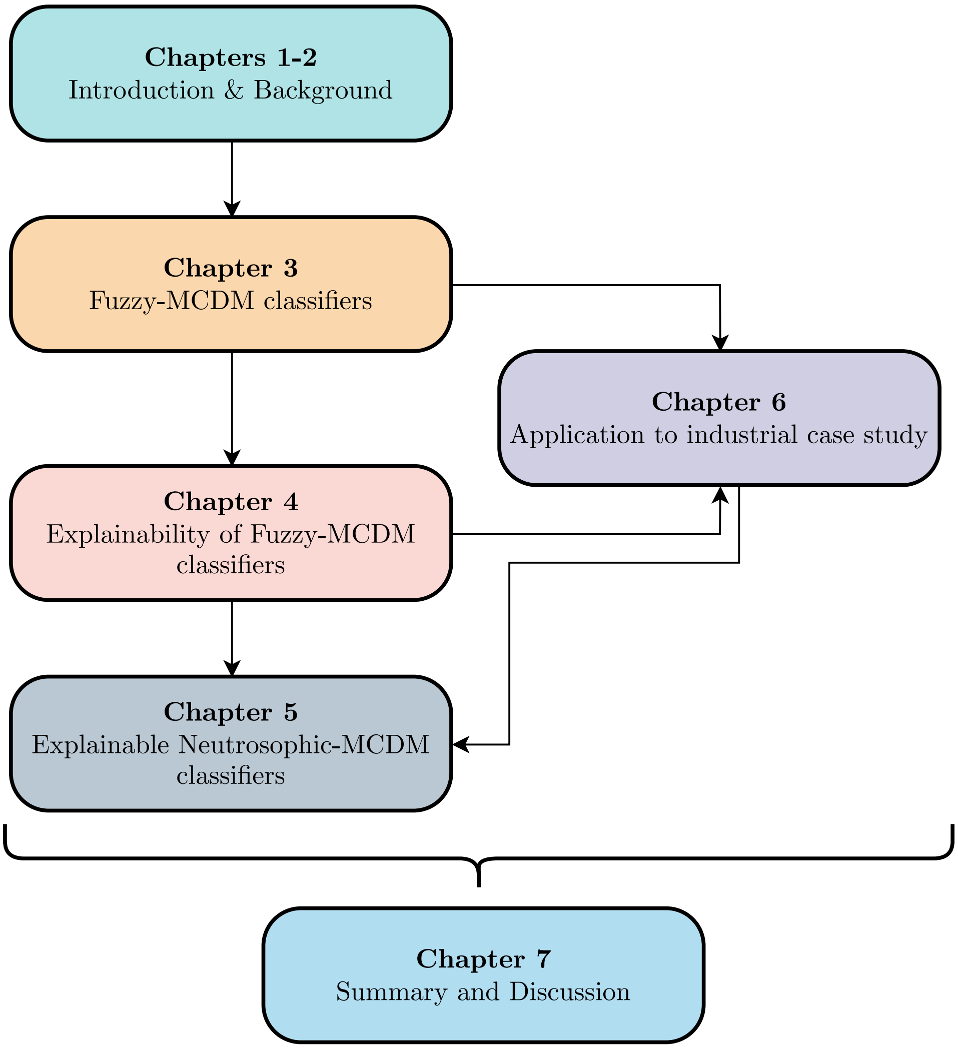

Chapters 1-2
Introduction & Background
Chapter 3
Fuzzy-MCDM classifiers
Chapter 6
Application to industrial case study
Chapter 4
Explainability of Fuzzy-MCDM classifiers
Chapter 5
Explainable Neutrosophic-MCDM classifiers
Chapter 7
Summary and Discussion

Figure 1.1: Thesis outline

Chapters 3-6 are the core contribution chapters. For the benefit of the reader, this thesis uses a recursive contribution-based structure. Thus, each core chapter includes an introduction, literature review, methodology, results and discussion.

More detailed literature reviews are presented in each of the core chapters. The rationale behind MCDM's use is presented as an introduction of Chapter 3. Notable examples of MCDM's success are presented. Consequently, Fuzzy-MCDM classifiers are proposed as an extension to the original amended fused TOPSIS-VIKOR for classification (ATOVIC) classifier; first proposed by Baccour in 2018 [14]. The proposed framework seeks to enable absolute data-driven fitting while maintaining the same level of interpretability. This was achieved by employing a transparent approach to fitting, as explained in detail in Section 3.3. A set of five benchmark datasets are used as test data for assessing and comparing the proposed frameworks with state-of-the-art classifiers. The results indicate a variable trade-off of performance, if any. The suitability of the trade-off would depend largely on the application and its size.

The reasoning for developing Fuzzy-MCDM is more apparent in Chapter 4. In this chapter, an explanation framework is implemented to generate textual and graphical explanations. The former is in the form of linguistic statements designed to present factual and counterfactual explanations to the user. Three statements are generated for the two opposite sub-models and the overall model. On the other hand, a graphical explanation illustrates in a bar graph individual feature impact. MCDM's *decomposable* allows for graphical explanation to be generated for each sub-model and class. This results in a set of four graphical explanation plots, providing a glimpse of the state of the different model components to the user.

The insight produced by the explanation framework in **Chapter 4** presents a concrete example of how MCDM's transparency can be exploited. In **Chapter 5**, interpretability is taken a step further by the introduction of a Neutrosophic-MCDM classifier. In Fuzzy-MCDM, the inference system output is in the form of the degree of truth representing the positive class. This allows for a concise representation of factual, counterfactual and indeterminate information. Nevertheless, counterfactual and indeterminate information is particularly important for the sake of explanation.

A neutrosophic logic set is designed to represent truth, indeterminacy and falsity virtually independently. To paint a clearer picture, factual, counterfactual and indeterminate information can be represented by the truth, falsity and indeterminacy components, respectively. The explanation framework is extended to graphically illustrate the neutrosophic outputs for each sub-model. By doing so, insight is grouped into three areas of interest: factual, counterfactual and indeterminate.

The performance results indicate no significant performance trade-off attributed to the increased interpretability of NL. Thus, Neutrosophic-MCDM is seen as an improvement over the initial MCDM and Fuzzy-MCDM classifiers.

The proposed modelling frameworks were investigated in a practical setting by use of an industrial case study. The area is the ultrasonic inspection of HDPE pipelines. The datasets acquired are of 110mm HDPE pipe welds.

As a means of assessing the frameworks' resilience and applicability, each contributed technique is tested on two objectives of the dataset: detecting defects or beads - a key indication.

Chapter 6 starts with a discussion of the importance of interpretability for inspection. It highlights how *accountability* and *compliance* are driving the need of XAI in safety-critical industry. Consequently, the defect and bead datasets are used to compare the performance of the frameworks proposed.

The results demonstrate the power of *MCDM*-based classifiers in extracting meaningful insight into the decision-making process. In a similar way to the benchmark datasets, the MCDM classifiers generated graphical and textual explanations. The explanations indicated why a decision went a certain way based on the feature values and sub-models' states.

Similarly to the benchmark datasets, the performance compared to conventional classification techniques is somewhat lacking. However, the minimal drop in performance, if existent, is a price worth paying for the extended explanation capabilities. A valuable asset for increasingly regulated and safety-critical industries.

The thesis is concluded with **Chapter 7** where conclusions are summarised, and future work is proposed.

2 Background: interpretability, explainability and manufacturing

2.1 Introduction

In machine learning, interest has been growing for interpretable and explainable methodologies. In this chapter, an overview of interpretability and explainability is presented. Key background concepts are described such as sources and types of interpretability. Consequently, an overview of interpretable and explainable classification frameworks is presented. The chapter is concluded with some background on how the need for interpretable ML fits into the world of advanced manufacturing.

2.2 Interpretability and explainability

In this section, an overview of interpretability and explainability in ML is presented. The key topics to be discussed in this chapter are shown in Figure 2.1. Human's reliance on ML models sheds light on the downfalls of many popular methodologies. The applicability of robust data-driven models has driven the adoption of deep and complex techniques faster than simpler transparent models. As a result, research and practice have highlighted the implications of using opaque models [2].

2.2.1 Interpretability

Most recently, there has been a growing interest in model interpretability. As a concept, interpretability has no universal mathematical definition [15]. Although

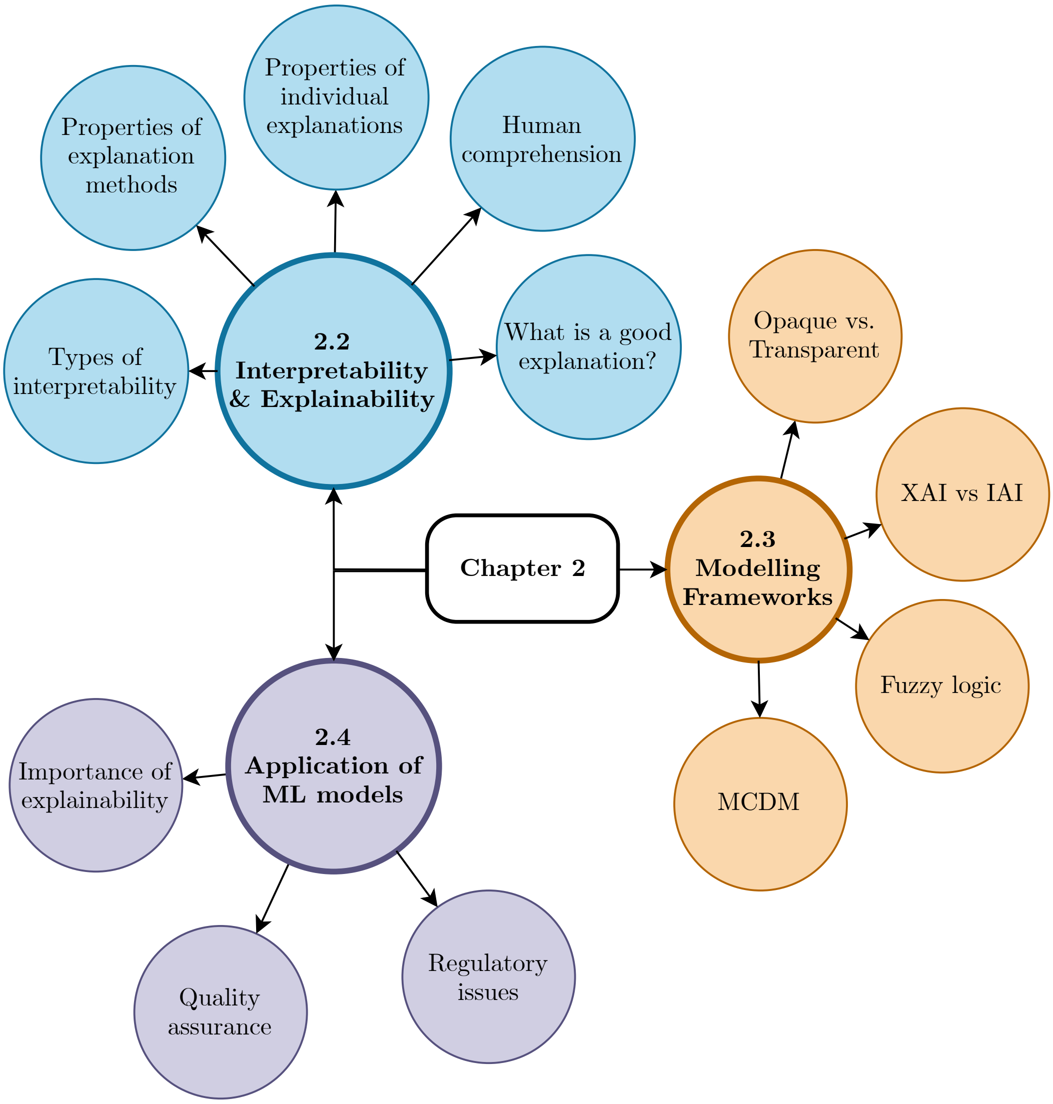

Properties of explanation methods
Properties of individual explanations
Human comprehension
Types of interpretability
2.2 Interpretability & Explainability
What is a good explanation?
Opaque vs. Transparent
XAI vs IAI
Chapter 2
2.3 Modelling Frameworks
Fuzzy logic
MCDM
2.4 Application of ML models
Importance of explainability
Quality assurance
Regulatory issues

Figure 2.1: Chapter 2: mind map of general topics and concepts

attempts have been made to quantify interpretability, they have often been limited to a certain domain or model type [16].

Nonetheless, researchers have suggested broader non-mathematical definitions. For instance, a definition by Kim et al. [17] states that: "**Interpretability** is the degree to which a human can consistently predict the model's result". This view of interpretability hinges on the *predictability* of the model to a human. In contrast, Miller's [18] definition below indicates the importance of understanding the cause.

> "**Interpretability** is the degree to which a human can understand the cause of a decision."

Despite the differing definitions, Murdoch et al. [3] state that *model interpretability* is

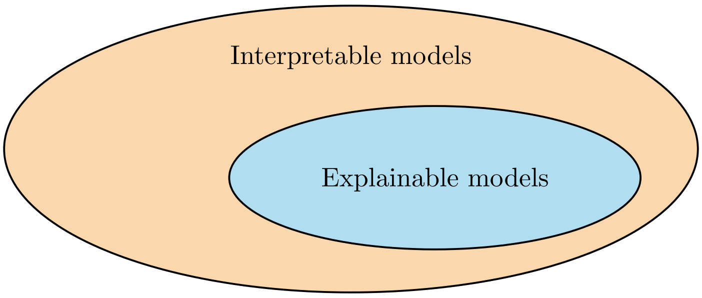


Figure 2.2: The relationship between interpretable and explainable illustrated in Venn diagram

nonetheless a beneficial blanket term that refers to methods of extracting meaningful information about a model or its data.

Based on the preliminary definitions, interpretability is argued to be vital for further success of ML's adoption. One of the most prominent obstacles to ML's wider implementation remains to be a lack of *explainable* solutions in practice.

Although explainability is a term often used interchangeably with interpretability, it is also used to distinguish between two distinct concepts. Interpretability refers to the property of the modelling framework being used. The term refers to how much of its design allows for it to be interpreted, i.e. understood. However, opting for an interpretable framework does not guarantee that the *interpretable* information is *explainable*. In other words, a model must be interpretable to have a chance of being explainable; however this is not necessarily sufficient (Figure 2.2).

In some literature, the two terms interpretable AI and explainable AI are used to distinguish between methodologies that aim to provide model-based versus post-hoc explanation, respectively [19]. In this thesis, explainability is discussed as a property of an interpretable model-based model.

The degree of explainability relies on several factors not limited to the modelling framework. Another component of explainability is the human perceiving it. No matter how many experts claim a model is explainable, it can only be considered so if it is believed to be by an actual user, the same user who is expected to use the system.

On that account, explainability is widely believed to be a concept even harder to research than interpretability. This is because interpretability depends solely on the model at hand, while explainability is more susceptible to subjectivity. Nonetheless, there are widely accepted concepts on what *explainability* aims to tackle.

The aim of explainability is to provide comprehensible information to a user in order to facilitate their understanding of a model's result. By doing so, the user could appreciate, to a suitable extent, how or why a model decided a certain way, ideally, to a level that enables the user to trust a model's outcome.

2.2.2 Sources of interpretable information

Interpretable information can be components of the model that have the potential to provide additional insight into the model's result. The insight can then be used to explain the model's result and algorithm - assuming that the model is *explainable*.

The quantity of interpretable information varies depending on the model's type. For instance, the opaque nature of deep learning model means input and output data could be the only real interpretation sources. In contrast, a more traditionally transparent model such as decision trees (DT) has more interpretable information extracted from various regions of the tree - assuming the data and thresholds are human-understandable.

More formally, two categories divide the types of interpretability: post-hoc or model-based. The former analyses the model's result along with the inputs to attempt to *predict* an explanation. It is often model-agnostic, i.e. does not require a certain model type to function.

Meanwhile, model-based, as the name suggests, uses transparent models to provide additional interpretable information extracted from the model's internal structure or based on an understanding of its transparent algorithm. Model-based uses more information indicating the model's inner workings. Hence it is seen as more likely to enable a more representative explanation. For this reason, interpretable models have more potential as providing *usable* explanation.

However, model-based explanation's reliance on interpretable models, which are inherently simpler, means it is associated with inferior performance. Therefore, several researchers argue that adopting model-agnostic approaches to explanation is far more promising than living with the performance trade-off of transparent models.

Meanwhile, another group of researchers urge experts to opt for interpretable models for applications where the stakes are high [20]. Moreover, the researchers suggest that the performance trade-off often associated with interpretable models is not a mathematically proven phenomenon. Although this statement is factual, one can say the same about the fact that the performance trade-off is, by comparison, also not mathematically disproven.

The obstacle to concluding this somewhat controversial topic is the ability to *define* interpretability mathematically. Attempts have been made to provide interpretability metrics for certain model types, such as fuzzy logic [16]. However, the metrics are far from being a universally accepted definition [15], [21].

Suppose this metric can be replicated for different model types. It could pave the way for coming up with a universal definition for model interpretability.

The involvement of the human factor in interpretability and thus, explainability means a test is arguably more practically possible than a *metric*. This is arguable from a mathematical standpoint where humans are considerably unpredictable.

2.2.3 Explainability

Given the interpretability, challenge is done and dusted. A second challenge emerges - explainability. When a model is explainable, a framework can be developed to generate *explanation*.

An explanation has several properties, as described by Robnik-Sikonja et al. in 2018 [22]. The properties can be used to assess an explanation's effectiveness based on the methods used and the explanation itself.

Similar to interpretability, model explainability is still considered in its infancy as a concept. At the time of writing, researchers are striving to define what explainability is, assess its existence and implement it. However, explainability's dependence on interpretability adds another obstacle to its advancement.

The properties of explanation methods are suggested by Robnik-Sikonja et al. as:

- Expressive power
- Translucency
- Portability
- Algorithmic complexity

Expressive power relates to what explanation can be produced based on the explanation method utilised. For instance, natural language would have more expressive power than decision trees.

Moreover, translucency specifies to what extent the model's parameters are used for explanation. For a black-box approach to explanation, only input-output data is used since model parameters are not useful. In this case, the explanation method would have no translucency. Conversely, interpretable models allow for the development of *translucent* explanation methods.

The portability of explanation methods means their applicability to a wider variety of machine learning models. It is always advantageous to design a method that works in a more generous selection of alternatives. However, increased flexibility means treating the model as a black box, thereby limiting the method's translucency.

Algorithmic complexity describes how computationally intensive the method for generating the explanation is. The method's complexity could be an issue when computation time is vital for the application.

Robotnik-Sikonja et al. go on to describe the properties of explanations themselves.

- Accuracy: refers to how well the explanation reflects unobserved uncertainties. This does not apply to black-box models where only fidelity is used.
- Fidelity: refers to how representative the explanation is of the black-box model. Fidelity is important to assess specifically for black-box models because of their opaque nature.
- Consistency: refers to whether a consistent result is produced when the explanation method is used with different models.

- Stability: also tests consistency, however, across instances of the same model. This confirms whether the explanation is *stable* i.e. dependable.
- Comprehensibility: refers to the extent to which humans can *comprehend* the explanations. An important piece to solving the puzzle of explanation is the human on the receiving end. Human comprehension and understanding vary widely. Hence, comprehensibility is considered of the most challenging aspects of explanation. It is also the *whole point* of explanation.
- Certainty: refers to how well explanation can represent *true* model certainty.
- Degree of Importance: refers to whether the explanation reflects which factors influenced a decision within a model accurately.
- Novelty: refers to the ability of the explanation in handling *novel* data not present in the training dataset in a way that indicates the data's peculiarity.
- Representativeness: how well can the explanations cater for different components of the model, i.e. does it explain all sub-structures of the model or a small part.

2.2.4 Model transparency

Model transparency defines whether a model's structure is interpretable. For instance, deep learning models lack transparency because of their increasingly complex structure. Meanwhile, the much simpler linear regression models enable a user to understand fully how a prediction was calculated just by looking at the coefficients m and c.

$$y = mx + c \tag{2.1}$$

One of the challenges of attempting to explain more complex black-box models is having to deal with high dimensionality - a case where many features are likely to impact a decision. Attempting to explain black-box models using their internal parameters is not considered practically possible. Therefore, a model-agnostic approach to interpretability was seen as the most promising way forward for black-box models.

One of the ways researchers have tried to address this is by tapping into *local interpretability* using local surrogate methods. Local surrogate is a post-hoc explanation approach that uses external interpretable models; the local models attempt to mimic the main, often black box, model's classification process [23]. Consequently, explanatory models can be used to produce explanations more easily because of their transparent nature. Post-hoc explanators are often built using interpretable modelling methodologies such as decision trees, rules, or linear regressions [23].

The main advantage of post-hoc is the ability to independently develop the ML model and its explanator. Thus, the ML model can be switched out if it turns out to be unsuitable, without the need to redesign the explanator.

Ribeiro et al. [24] proposed what turned out to be one of the most well-known local surrogate methods - local interpretable model-agnostic explanations (LIME). LIME's well-defined structure enabled its implementation in widely utilised ML programming languages such as Python and R.

LIME provided a straightforward way to *retrofit* pretty much *any* ML with interpretability. This was a major selling point at the time since most implementations opted for robust black-box models.

Post-hoc interpretability provided an easy way out for experts seeking to maintain their use of black-box models. LIME was able to provide much-needed insight on opaque models - providing users with vital explanatory information.

Although many model-agnostic interpretability methods exist, they all share a common weakness; their reliance on an external explanatory model. In spite of providing added simplicity, the explanatory model's independence from the main model means its explanation is not tied to the model's inner workings. Therefore, what the explanator provides in terms of explanation is merely a prediction.

For this reason, it becomes more of a challenge for post-hoc explanators to adhere to the justifiably stringent laws, and regulations present in critical industries such as aerospace, oil & gas, nuclear and biomedical [15].

The individual explanation properties described earlier, such as stability, fidelity and consistency, are more difficult to quantify and, in turn, maintain. Ironically, a post-hoc explanation can, in many instances, work. However, there is limited understanding on the inner workings. It creates a compounded problem where two entities are now not fully interpretable: the model and its explanator. On the other hand, opting for an inherently transparent model allows for *inferred* explanation. A type of explanation that has a higher chance of being accurate and understandable.

2.2.5 What is a good explanation?

As described above, certain properties can be used to assess explanation [18]. Although these guidelines could be considered satisfactory from an engineering perspective, they fall short of addressing *explanation* from a *human* standpoint.

Miller et al. [18] suggest that explanation is the answer to a logical why-question. Such examples of why-questions can be:

- Why is the manufactured part defective?
- Why was a mortgage application rejected?
- Why is there a forecast for rain?

The questions urge so-called "everyday" explanations that are straightforward to provide by an expert in the field. For instance, the first example can be explained by citing the indications that reveal a defect. On the other hand, counterfactual can be provided by describing the reasons the manufactured part is not considered *healthy*. This contrastive form of explanation is the preferred option for humans [25]. This is because it appears to single out the reason a decision went a certain way.

Therefore, it makes sense to design an explanation that highlights only a few factors affecting the output. When numerous reasons are provided in the form of explanation, the human is not able to appreciate the most important.

In other words, the comprehensibility of an explanation depends strongly on the complexity of the explanation itself. Hence, ensuring the explanation's simplicity goes a long way toward making it comprehensible.

2.3 Machine learning classification frameworks

ML frameworks for classification can be divided into two main categories: transparent and opaque. Transparent frameworks enable the construction of inherently interpretable models where the internal structure is accessible with the potential for being explainable. Lipton suggests that a model's transparency be assessed using three criteria: *simulatability*, *decomposability* and *algorithmic transparency* [12].

Simulatability, Lipton explains, is the degree to which a model can be simulated using pen and paper by a human. Lipton believes the more *simulatable* a model is, the more interpretable it likely is. For instance, linear regression models can be easily simulated using pen and paper by calculating the linear equations representing the model - simple two-operation arithmetic. A more complex model could be a K-Means model with a large number of features, e.g. 1000. Although this is not considered large by industry-standard, it is much too large to be simulatable on pen and paper. Similarly, deep ANNs often have a large parameter set that cannot possibly be simulated manually.

Decomposability refers to whether the model's structure can be *intuitively* visualised as a collection of compartments, where each compartment has a well-understood definable function. For instance, a system of hierarchical fuzzy logic set (FLS) can be decomposed in terms of logic sets. The behaviour of each compartment, a FLS, is defined by its rules and membership functions. Likewise, a decision tree model could be decomposed in terms of its branches.

The remaining factor to consider is algorithmic transparency. This refers to how understandable and transparent the optimisation algorithm is for a model. For example, although an algorithm such as gradient descent is considered human-understandable and transparent. Pairing this algorithm with a complex deep neural network means it will lose its *transparency*. In spite of the foundational theory being well-understood, the high degree of dimensionality prevents a human from *peeking* into the algorithm. Conversely, when the same algorithm is applied to a linear regression model, the human can even go to the extent of computing the gradient descent optimisation on pen and paper. Hence, from a practical point of view, the lack of algorithmic transparency is perceived to impact an overall model's interpretability.

2.3.1 Explainable-AI

The need for explainable AI is evident in the growing research interest surrounding it (see Figure 2.3). According to data from Scopus, the literature on interpretable and explainable AI has been increasing for the last five years. Researchers from around the world aim to address the knowledge gap of XAI.

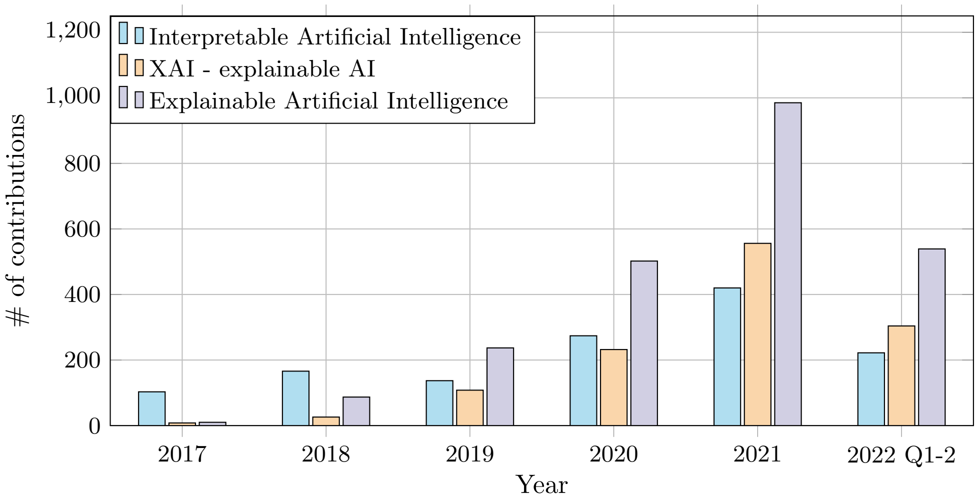


Figure 2.3: A bar graph illustrating the growth of contributions published with key words in the legend related to interpretability and explainability of AI. The data was downloaded from Scopus® on 30th June 2022. Based on the methodology used by Arrieta et al. [19].

Explainability methods in AI span a wide variety of categories from model-agnostic to model-specific, model-based to post-hoc and local to global. It is vital to understand the merits and limitations of each approach to ensure the selection of the suitable one.

The suitable flavour of explainability is sometimes dictated by the complexity of the dataset. For example, a dataset with a large number of features means a complex model may be required to address it. If so, model-based would be out of the picture. Moreover, global interpretability is less likely to be applicable, and a model-agnostic model could be the easiest way forward.

In contrast, when a dataset uses a small number of human-understandable features, it becomes easier to access model-specific, model-based and global interpretability. This is considered a near-ideal state for maximising the explanation's quality. The diagrams in Figures 2.4 aim to depict the worst versus best case scenarios of interpretability and how they lead to predicted versus inferred interpretability, respectively.

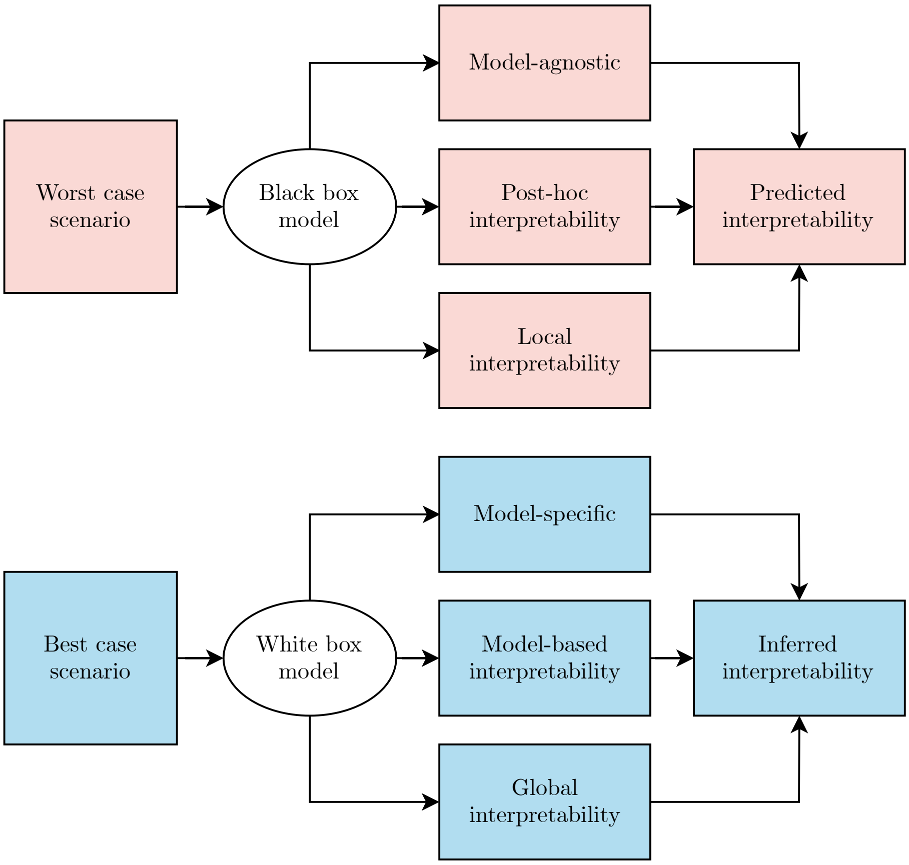


Figure 2.4: A set of block diagrams illustrating how choosing between black and white box models can affect interpretability in different ways. It depicts a case where the best and worst case scenarios are realised when using a white and black box model respectively.

The importance of the *explanation's quality* varies according to the model's intended domain. Sudjianto et al. [26] stress the importance of developing inherently interpretable models for highly regulated sectors such as healthcare.

Although, post-hoc interpretability is expected to be able to cater for numerous non-critical applications. These areas have a comparatively minimal consequence of inaccuracies and biases. For example, when an ML algorithm fails to detect the face of a user trying to unlock their phone - the consequence is that they have to log in manually. When the smartphone provides a wrongful explanation as to why it has failed to unlock - the consequence is similarly low.

Conversely, when a ML model is used to support a court of law in making its decision, it is a totally different story. In this situation, a wrong decision could very likely affect the livelihood of an individual [27].

As a result, a group of researchers have warned against attempting to explain black-box models [2]. They justify their stance by citing a concern that explained black-box models, according to theory, are more likely to produce inaccurate explanations because of the manner in which interpretable is extracted. This refers to the fact that post-hoc analysis operates independently with the model's internals even where the method is model-specific. For this reason, it is seen by the researchers as a way of masking the issue into a bigger problem [2].

The most recent XAI review uncovered a significantly larger focus on post-hoc methods. The reason for this could be the more prominent prevalence of black-box models. XAI researchers justify opting for black-box models for their perceived superior performance [3], [19], [28]–[30].

Meanwhile, supporters of inherently interpretable models have gone as far as stating that the trade-off is a 'myth' [2]. They have supported this statement by pointing to numerous cases where interpretable models could perform as well as deep learning. Nonetheless, the researchers admit that some domains remain where deep learning is necessary and justified.

The takeaway is that inherently interpretable models should be used more often than they are being used. More specifically, in areas where the importance of *proper* interpretability outweighs the *convenience* of fitting a black-box model. This is given that the performance drop is non-existent.

Depending on post-hoc analysis for interpretability leads to what is referred to as lower *descriptive accuracy* - a metric linked to how representative explanation is [3], [30]. Inferior descriptive accuracy is particularly an issue in high stake applications, where the wrong explanation can deter further investigation leading to a compounded problem.

Nevertheless, some black box models have been investigated in high stake applications despite the risks [2], [31], [32]. Reasons quoted by researchers often mention the unbeatable performance of black-box models [3], [19], [28]–[30].

2.3.2 Interpretable models

As discussed previously, research interest XAI has been growing for more than a decade [19]. More detailed reviews reveal information on key topics, techniques and methodologies currently being explored. This section aims to summarise the key findings and highlights information relevant to this project - model-based interpretability.

Two approaches exist within the model-based: interpretable models and hybrid models. The former uses a combination of interpretable and non-interpretable components. The category of hybrid models is distinct from post-hoc explanation, where an interpretable is used merely as an explanator. For hybrid model-based interpretability, both components (interpretable and non-interpretable) take part in the model execution.

On the other hand, a pure interpretable approach is possible with several standalone or hybrid interpretable models. Several interpretable modelling techniques exist, as listed in Figure 2.5.

The benefit of interpretable models is being able to leverage their transparency. Model transparency allows the direct extraction of interpretable information to deduce and infer a factual and counterfactual explanation.

Interpretable models are intrinsically understandable and transparent. For instance, DT can be visualised and understood intuitively by the average user. The criteria of model transparency (simulatability, decomposability and algorithmic transparency) are demonstratable to an extent not possible with conventional black-box models such as deep neural networks or support vector machines.

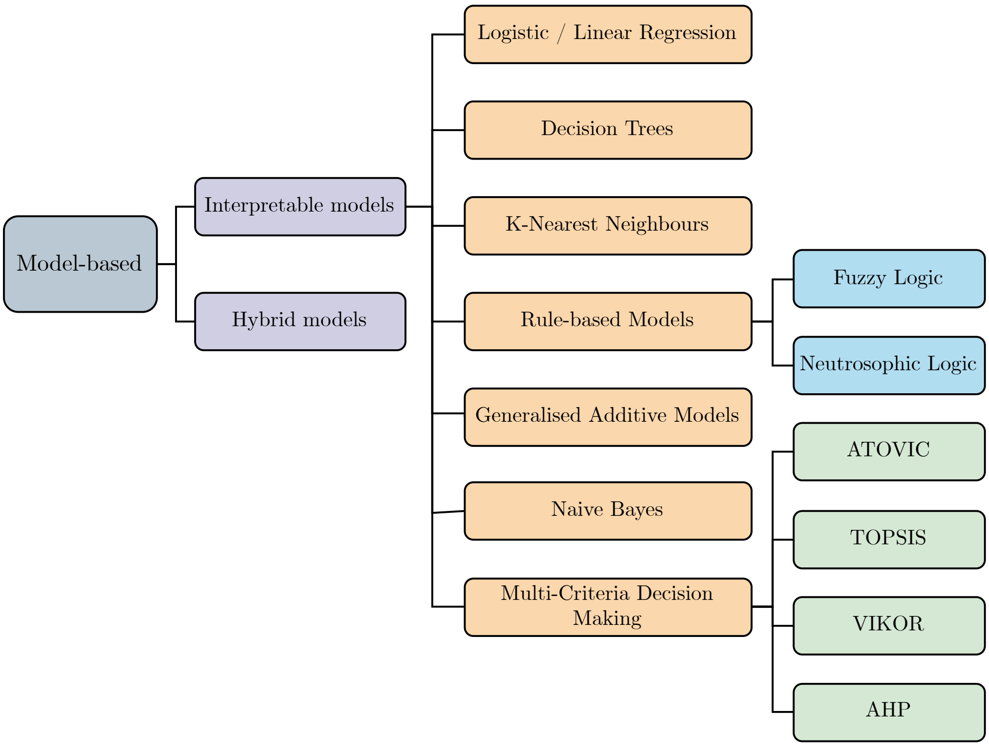


Figure 2.5: A diagram illustrating model-based interpretability methodologies and techniques.

Moreover, rule-based models such as FL allow for the same level of interpretability. When FL was first proposed by Lotfi Zadeh in, it provided a method for designing classifiers and controllers with transparency not provided by conventional models and theory at the time.

Zadeh's FL was proposed a long time before XAI was coined as a research area. Nonetheless, it is estimated that FL literature accounts for a third of XAI contributions.

Alonso et al.'s [33] analysis revealed a clear separation of fuzzy interpretability from the main cluster of XAI. In other words, inter-citations between fuzzy-related literature and non-fuzzy AI are relatively low. Therefore, Alonso et al. [33] recommend the importance of closer collaboration between the different disciplines within XAI and more specifically, interpretable FL. The researchers justify this by citing early fuzzy works on harnessing transparency and interpretability as early as 1999 and the fact that more than a quarter of XAI literature is related to FL techniques.

MCDM is a set of techniques capable of providing decision support based on an input of criteria representing the alternatives. MCDM is seen as an area where different decision theories have been materialised into mathematical frameworks for providing decision support.

MCDM is used in several areas to facilitate decision making, particularly when it involves choosing between several alternatives. Notable applications include supplier selection, data engine and software selector.

In spite of MCDM's inherent simplicity, it has seldom been explored for classification [14]. However, it has been combined with ML techniques such as FL as a means for further optimising and strengthening its core application - alternative selection.

In MCDM lies a hidden strength, its inherent transparency and human understandability. An MCDM technique such as TOPSIS uses a rather simple structure where a decision matrix is constructed from the alternatives criteria (features). Consequently, weights are assigned to indicate the importance of each criterion numerically. Simple arithmetic is then used to determine a rank for each alternative based on the criteria and ranks.

TOPSIS can be easily applied to an everyday problems such as selecting which dish soup to use. Criteria can be as understandable as cost, availability and smell. The importance of the criteria (weights) are assigned subject to the user's preference.

2.4 Advanced manufacturing

Factories around the world have witnessed an unforeseen advancement in efficiency and throughput following the advent of digital automation [34]. As a result, the staggeringly high production rate of these factories meant quality assurance became a more obvious issue.

Despite the success rates of some of these factories being over 95%, the volume of production meant that the remaining 5% could be significant. Thus, all production lines became associated with a relatively high criticality level.

For instance, a small deviation from the production parameters could spell a considerable loss of product. Therefore, robust controllers were installed to closely monitor and control key performance indicators.

In spite of this, certain production processes result in uncertainty still not fully understood. Thus, unpredictability becomes an issue which is addressed using higher-level forms of control such as Model Predictive Control (MPC). These approaches were made possible by the collection of data directly from the factory manufacturing lines and using it to optimise the process further.

Similarly, in engineering factories, quality assurance is of paramount importance. This is especially case for critical manufacturing infrastructures such as pipelines and jet engine parts. Failure to detect defects in parts destined for safety-critical applications could potentially lead to catastrophic consequences such as loss of life, environmental disasters and damage to high-value assets.

Appropriately, domains where the consequences are high need to be regulated to guard against these hazards. The regulations instantiated are designed to enforce scrutiny on changes to operating procedures in these domains. In doing so, risks of adopting new technologies can be *managed*.

One of the notable is IEC 61508[1] - a standard for ensuring the functional safety of electric, electronic and programmable safety systems. Safety systems are specialised control systems tasked with bringing a safety-critical process to rest should it becomes dangerously unstable. It is used in conjunction with an operational control system as a safeguard against the process deviating too far from the safe operating zone.

In safety-critical industries such as oil & gas, an unstable process can spell all sorts of costly incidents such as unplanned shutdowns, loss of assets or even fatalities.

[1] International Electrotechnical Commission 61508: Functional safety of electrical/electronic/programmable electronic safety-related systems

Safety systems minimise this risk in the arena of the control system.

Similarly, the use of AI/ML in these fields is subject to the same stringent regulations. Although, our understanding of XAI is yet to mature to the same level as other well-established areas such as safety system design.

Based on previous experience, it is clear *compliance* is one of the first hurdles in the adoption of AI/ML in any regulated industry. By comparison, black-box models have seen unprecedented prevalence in self-regulated or non-regulated industries. This is arguably at the cost of proper validation [2].

The European Union (EU) have highlighted the importance of explanation in their general data protection regulation (GDPR) legislation. The regulation stressed that users are *guaranteed* the right to request an explanation for decisions made using an automated algorithm that processes their personal information. This could apply to processes such as automated credit card applications.

As consumers have experienced, automated rejections for credit are seldom explained. It is unclear whether the algorithms are explainable since often these models are propriety, thus, confidential. As a consequence of their opacity, consumers are faced with rejection without being able to find the reason or set of reasons.

Meanwhile, in advanced manufacturing and specifically quality assurance (QA), an explanation is even more essential. When safety-critical parts are manufactured, it is important not to miss defects because if faulty parts are installed, they are more than likely to cause catastrophic consequences. If ML techniques are to be employed to automate QA, they have to be able to provide a suitable explanation to the NDT expert; since not providing an explanation that justifies the decision leaves the expert in a situation where they have to come up with their own justification and explanation. This defies one of the main purposes of decision support automation - reducing manual labour.

More primitive testing methods are destructive in nature such as tensile strength. The issue with destructive testing is that it ends up destroying the part, which causes waste. Also, destructive testing only tests a small sample of the products produced. Therefore, some defects could potentially pass through the screening.

Alternatively, NDT approaches were explored. Various types of NDT technologies exist, and applicability depends on the properties of the part that needs to be tested.

An example of a successful NDT method is Guided Wave Testing (GWT). It has been used successfully in the detection of corrosion in pipelines. GWT works by dispersing acoustic waves in along the length of a pipeline. The advantage of GWT vs UT is its ability to detect corrosion over tens of meters versus the vicinity of the probe.

The industrial scope of this project is to investigate ML techniques for the classification of pipe weld defects. The material of the pipes is HDPE. Plastic provides a special challenge in NDT; ultrasonic waves do not travel well in a material of this type. As a result, a defect's contrast would be more difficult to distinguish. Nonetheless, newer methods phased array ultrasonic testing (PAUT) have enabled higher resolution images.

However, the larger dataset means a more tedious job for the NDT expert. For instance, a single weld requires the expert to inspect and analyse hundreds of images manually. The time required to inspect the hundreds of newly installed welds in a major project using plastic pipes is considerable. Moreover, the task's repetitive and cumbersome nature is seen as a perfect recipe for human error.

Thus, there has been an interest in developing automated NDT - methodologies for automating the data analysis process. The aim is to reduce the amount of human intervention required by an NDT expert.

2.5 Summary

In summary, there are two types of interpretability: model-based and post-hoc. It is clear that model-based provided a better basis for generating better explanation. However, this type interpretability relies on the use of an often simpler transparent model. As a result, performance is likely to be impacted. The importance of explainability in advanced manufacturing was highlighted. A key area being high stake decision making.

MCDM and fuzzy logic are highlighted as possessing a promising potential for interpretable modelling. In the next chapter, Fuzzy-MCDM methodologies are explored for data-driven classification by comparing them with key state-of-the-art classifiers. A starting point for investigating the research gap of model-based explainable modelling.

3 Fuzzy-MCDM: interpretable ML for classification

Fuzzy logic has been a prominent research area in the field of interpretable machine learning. The versatility of FLSs has facilitated their adaptation to a wide array of problems such as control, prediction and classification. Some of the most recent interest in FLSs is excited by the fact that Mamdani-type sets have the potential for interpretability and explainability. Multi-Criteria Decision Making (MCDM) is a set of techniques capable of providing decision support based on an input of criteria representing the alternatives. Similarly, MCDM is highly interpretable by nature due to its human-understandable components and parameters. However, a pure MCDM model would lack the ML rigour required to be able to produce high-performing models. Therefore, in this chapter, a study is presented where MCDM is extended with fuzzy logic to define four variations of MCDM-based classifiers. Consequently, the classifiers are evaluated with five benchmark datasets to compare their performance to state-of-art models and other MCDM techniques. The results demonstrate how fuzzy-MCDM classifiers can provide comparable performance to SOA classifiers for certain problems, paving the way for adapting fuzzy-MCDM as a basis for an explainable data-driven classifier.

3.1 Introduction

Fuzzy-multi-criteria-decision-making (fuzzy-MCDM) is a prominent research area within the field of MCDM. As described in Chapter 2, MCDM was not initially intended as a classifier. However, its combination with ML techniques has prompted interest in its development as a potential data-driven classifier [14], [35], [36]. Its inherent interpretability and simplicity are one of the main motivations for this interest.

In this chapter, the proposed frameworks are compared with other well-established ML classifiers. By doing so, MCDM-based is assessed as a potential viable alternative to e.g. DT and Support Vector Machines (SVM), by determining key trade-offs, if any, exist. In addition, a comparative analysis is presented on MCDM techniques, including TOPSIS, VIKOR and, ATOVIC.

3.2 Interpretable ML methodologies

As described in the literature review, MCDM is a set of computational and mathematical techniques for assessing a set of alternatives, based on often conflicting criteria consisting of various costs and benefits [37]. MCDM are normally used for decision-making in various fields where single or multiple alternatives are selected from several, based on a well-defined set of criteria and weights. Due to the interpretable nature of MCDM techniques, there has been interest in their use in classification [14], [36].

3.2.1 Amended fused TOPSIS-VIKOR for classification: overview

Amended fused TOPSIS-VIKOR for Classification (ATOVIC) is an MCDM-based classification framework initially designed by Leila Baccour in 2018 [14]. ATOVIC is one of a few applications of MCDM-based methodologies for classification [14], [36]. The TOPSIS-VIKOR framework, ATOVIC, is a version of MCDM that is tailored for classification. This was achieved by treating the *features* as *criteria* and utilising a sub-model for each class present in the data. The sub-models *rank* the data separately and consequently used to classify the data; the rank essentially acts as a distance. MCDM methods are constructed using the following steps:

(a) Construction of the decision matrix

(b) Determining the weights of the criteria

(c) Setting the ideal solutions (positive and negative)

(d) The ideal solution is selected as the alternative closest to the positive ideal solution (PIS) and farthest from the negative solution (NIS)

The distinction between the two ranking methods is that TOPSIS implements normalisation for steps 1 and 2, while VIKOR does not; furthermore, TOPSIS uses just the Euclidean distances to compare alternatives. Meanwhile, VIKOR combines the usage of the Chebyshev, Euclidean and the weighted sum of both distances. Baccour [14] has demonstrated the potential of ATOVIC through a comprehensive set of comparative analyses. In the literature, the writer showed how their proposed method could be used as an MCDM-based approach to classification. The analyses contain cases where ATOVIC has performed as good as or better than SOA classification techniques. Nonetheless, there were cases where ATOVIC did not perform ideally. ATOVIC's structure meant its inability to handle non-numerical data such as text, images, or categorical attributes. This limitation was evident in a chess dataset performance result where it had a low score compared to the other models. The fitting process of the model is performed in a single iteration, i.e., no iterative *training* was used. Omitting training simplified the construction of the model, improving its interpretability, but it could have a negative impact on performance, as demonstrated by Baccour [14]. A point worth pointing out is relying on expert knowledge in certain aspects of model construction means Baccour's version of ATOVIC is not purely data-driven; this becomes an issue for applications where expert knowledge simply does not exist.

3.2.2 State-of-art ML classification frameworks

ML classification frameworks vary in transparency and performance. Striking a balance between the two can be a challenge, especially with the rising popularity of high-performing complex black box models. Black box models have an internal structure that is not human-understandable [38]. For example, **SVM!** (**SVM!**), Naive Bayes (NB) and Artificial Neural Networks (ANNs) are widely used black box classification frameworks for their resilience and robust performance. The techniques are optimised for data-driven training and provide satisfactory performance when paired with good-quality datasets.

However, black box models still have a major limitation - the lack of transparency in the model's structure preventing a *direct* explanation of a model's output. As an alternative, an indirect post-hoc explanation can be generated by the use of an external model for interpretation - often referred to as interpreter or explanator [39]–[43]. Post-hoc interpretability can be model-agnostic, i.e., compatible for all models types. Therefore, many researchers opt for post-hoc to avoid the potential performance trade-off often associated with transparent models [44]–[49]. Turner claims that the performance trade-off associated with transparent models is not justifiable, even if it is minimal [44]. This rests on the possibility that post-interpretability would develop to a level suitable for attaining meaningful and accurate explanations. Moreover, Ribeiro et al. [50] argue that transparent models are merely for insights and are not suitable for practical use because of their considerably inferior performance. The claim is problematic because interpretable models have been shown to perform adequately in certain situations [14], [35], [51]. In some applications, clear-cut explainability is essential for the model to be used, which can only be achieved by directing tapping into the model's internal parameters. On that account, attempting to interpret the model by an external *model* inherits a guaranteed *gap* between the two models. The gap cannot be perceived nor eliminated due to the opaque nature of black box models. For this reason, other post-hoc researchers have advised that transparent models should be used where performance is satisfactory to exploit their benefits for direct interpretation [48].

Explaining a model's result or decision is crucial for applications where the stake is high. Providing decision-support while omitting explanation restricts the expert from being able to fully rely on the model and, in turn, the building of *trust*. Gille et al. [52] suggest the adoption of ML models is dependent upon *trust*. The researchers highlight the implications of black box models, such as hampering the implementation of ML models in practice; transparent models were cited as a viable alternative. One of the drawbacks of black box ML models is the shallow understanding designers have access to on what the parameters/output mean, how they have been decided, and any hidden biases a model could have. This results in cases where models were inadvertently trained with significant bias [53]–[55]. An occurrence that cannot be overlooked for safety-critical applications where undetected biases could potentially lead to catastrophic consequences. For instance, if an ML model was implemented as a decision support tool, any bias present in the model could instil a bias in an expert's reasoning, possibly pushing them towards the wrong decision. The consequence of the wrong decision can vary widely depending on the application, so this needs to be taken into account when designing a model.

Based on the risks, Ruden warned of the dire consequences rush implementations of black box ML could have, with reference to real-world examples [20]. This included a case where a person was mistakenly denied parole by a black box model, which was a grave consequence to have to suffer [27]. Ruden urged ML designers to opt for white box models where possible and suggested the trade-off of performance is simply non-existent. Moreover, the decision's consequences should align with the transparency required in the model. For instance, a non-transparent model would suffice in low stake situations such as personalised advertising or email spam filters. By comparison, in healthcare, a wrongful diagnosis could cost a patient their health, making transparency an important requirement of a model in healthcare. Ruden's recommendations are sensible given the scale of the damage opaque models can cause when used inaptly.

3.2.3 FLSs as an interpretable ML framework

FLSs have often been used in the field of classification, as detailed in Chapter 2. FLSs types and many techniques span a wide variety of applications since its inception by Lotfi A. Zadeh in 1965 [56]. The main two categories of Type-1 FLSs are Sugeno and Mamdani-type. While the former relies on optimisation algorithms for generating its rule-base, the latter is formulated based on expert knowledge. The distinction in how the two types are constructed dictates how transparent the model is, with the Mamdani-type being the more interpretable of the two. Nonetheless, Sugeno-type is often favoured for its superior performance. However, its lack of transparency is problematic, prompting researchers such as Pekaslan, in [57], to propose an optimisation framework for Sugeno-type models while maximising interpretability. Pekaslan demonstrates how their new optimisation methodology maintained interpretability by constraining the search area to a *desired region* - maximising a set of fuzzy metrics. The methodology's ability to generate visually interpretable membership functions (MFs) is a step in the right direction; nevertheless, Sugeno as a fuzzy framework results in non-transparent output MFs and uses a complex optimisation algorithm which not necessarily user-understandable.

Therefore, Mamdani-type FL is being investigated and developed extensively for interpretable modelling because of its transparent nature [33], [58]–[61]. A comprehensive bibliometric analysis by Alonso et al. [33] revealed that explainable-AI (XAI) research widely varied with different clusters working in different directions. Although FL accounted for almost a third of XAI literature, highly cited FL literature was separated from the other XAI segments in terms of co-citations. This shows a lack of collaboration across the different XAI research areas. Alonso et al. suggest more inter-XAI research collaborations could be fruitful in accelerating progress [33].

To summarise, FL is considered interpretable because of its human-comprehensible structure. Current XAI research trends indicate considerable interest in developing fuzzy frameworks for interpretability and explainability. Hence, this makes FL a promising candidate for interpretable modelling.

3.2.4 Model interpretability for FLSs and MCDM

Interpretability is an important aspect of ML modelling as it paves the way for XAI. A more detailed review of interpretability is provided in Chapter 2. In this part, key concepts of interest are described. Model interpretability is divided into two main categories: model-based and post-hoc. As the name suggests, the former employs the model's components to provide interpretability while the latter makes use of an external model interpreter. Although post-hoc implementations of interpretability could provide a satisfactory explanation in certain applications, the precision of model-based interpretability is desirable in high stake applications.

Despite the considerable interest in XAI and the recognition of interpretability as a goal post for a meaningful explanation, *interpretability* as a concept is yet to be defined universally [61]. Instead, the concept of interpretability remains to be perceived as a broad concept representing the starting point of how XAI can be realised. As researchers investigate interpretability, it remains to be seen whether an absolute definition applicable to all models can be attained. Alonso et al. explain that interpretability has been difficult to formally define because of its dependence on "two heterogeneous entities" referring to the model and the user [61]. This statement highlights an important variable which is interpretability's dependence on the human user - a *variable* often overlooked by researchers. Nonetheless, the theoretical concepts attempting to define interpretability are a useful starting point for developing *potentially* interpretable frameworks.

As described in detail in Chapter 2, Lipton introduces several interpretability criteria such as *simulatability*, *decomposability* and *transparency* [12]. The concepts are broad, thus, providing a good basis for understanding how to assess the level of interpretability of a model regardless of its type. MCDM fits the criteria devised by Lipton; hence, it is considered a feasible alternative to state-of-the-art classifiers. MCDM lacks the ability to process uncertainty motivating its extension with FL [62]–[64]. Moreover, the FL extension is also an opportunity to improve the MCDM's performance.

FL interpretability relies on the transparency of its individual components, including the membership functions (MFs) and the rule-base. Fuzzy interpretation aims to design a model that is understood from two main perspectives: the model designer and the user. This could be achieved by limiting the number of MFs to a level the human mind can comprehend, as advised by Gacto et al. [16]. Psychologically speaking, this limit has been defined as 7 ± 2; dubbed as a guide for "our capacity for processing information" [65]. Gacto et al. stress that reducing the number of MFs alone is insufficient and suggest minimising the overlap between MFs is essential in maintaining their interpretability. The two metrics, the number and overlap of MFs, are also applicable to fuzzy partitions.

For Mamdani-type rule-based systems (RBS), the metrics can be manipulated by the model designer. Therefore, interpretability can be achieved by taking the metrics into account when designing the FIS structure; by avoiding the use of numerous fuzzy partitions, MFs and rules.

3.2.5 Why MCDM?

The rapid advancement of AI resulted in its unforeseen implementation in numerous fields, including but not limited to manufacturing, finance, biomedical and legal [66]–[69]. The use of AI came with an array of benefits, such as reducing costs and improving efficiency. Despite this, implications arose due to the prevalence of black-box models. The reduced transparency of black-box models meant interpretability, and explainability was lacking. As a result, instances of unexplained misclassifications have had grave consequences affecting the lives of individuals, as in the criminal law field [20]. The implications prompted international regulatory authorities and research institutions to stress the importance of explainable AI [70]–[72].

For example, the EU's General Data Protection Regulation (GDPR) law was predominantly intended to protect individuals' right to protect their data [70]. However, the law dictates that users of whom data was processed automatically have the right to obtain 'meaningful information about the logic involved' in processing said data. Although this does not explicitly signal the need for XAI, it is clear that systems utilising ML to perform automated decisions should now transition to transparent methods to facilitate explainability functionality. Furthermore, a report by the European Commission's AI-HLEG[1] suggests *transparency* as a key requirement for trustworthy AI; where *transparency* in this particular sense entails "traceability, explainability and communication".

[1] High-level expert group on artificial intelligence

Conversely, the UK government, in its 'Guidelines for AI procurement' document [71] was more specific in its approach, advising procurers to 'avoid' black box models. The justification is that white box models allow users to benefit from interpretability which results in models that are more likely to be sustainable by future AI vendors. In spite of the clearer stance, the document remains a *guideline* thus, it is less likely to result in many companies adopting white-box models.

The other obstacle to adopting white-box models, such as MCDM, is their limited versatility and applicability. If improved, MCDM models serve as a promising route to XAI. Their model-based interpretability allows designers to focus on explainability - the ultimate aim. MCDM's structure makes use of multiple weighted criteria while, a set of ideal solutions define what the criteria should be *ideally*. This method mimics how a human takes a decision when presented with several alternatives. A prime example of MCDM is supplier selection, where a company has to decide between a list of suppliers. This is often decided based on criteria such as delivery speed, cost and payment terms. Although humans would not necessarily quantify the criteria numerically in a practical setting, they would still have a sense of how *good* or *bad* each of the criteria is. The weights represent how important each of the different criteria is to the decision. Similarly, humans would associate an importance level to different factors in a decision.

MCDM's mimicking of the human's decision-making thought process allows for a promising research route. This is because the model will be able to do more than just classify data; it will also be able to demonstrate to an expert how a decision was made in a manner similar to their decision process. This can provide an expert with the peace of mind knowing that they fully understand the *reasoning* behind a decision, building the *trust* in the model [73].

3.3 Proposing fuzzy-MCDM-based classifiers

3.3.1 MCDM for classification

As described in Section 2. MCDM is a set of techniques capable of providing decision support based on an input of criteria representing the alternatives. MCDM has seldom been explored for classification. However, some studies show its success. For instance, Baccour proposed ATOVIC as one of the first MCDM-based classifiers [14].

ATOVIC is a semi-data-driven MCDM classification framework proposed by Leila Baccour in 2018 [14]. The model is constructed by defining a sub-MCDM model for each class in the dataset. For instance, in binary classification, a sub-model is constructed for each class; this involves setting all the model parameters, as will be described in detail below.

Before model construction, the data is divided into a reference and testing dataset (X^r and X^t respectively). Consecutively, normalised versions of the datasets are represented by Θ^r and Θ^t, respectively.

Constructing the model entails obtaining the following:

(a) Normalised reference dataset Θ^r

(b) Feature weights w^r

(c) Cost and benefit feature classification C, B

(d) Positive and negative ideal solutions f^+, f^-

(e) Measure extreme values $Q^\pm, R^\pm, S^\pm$

$$h^r_{j\,p} = \sqrt{\sum_{i=1}^{m^r} (x^r_{ij_p})^2} \tag{3.1}$$

where $x^r_{ij_p}$ is the non-normalised term for object i, feature j and, class p; m^r is the number of instances in the reference dataset X^r.

$$\theta^r_{ij\,p} = \frac{x^r_{ij}}{h^r_{ij\,p}} \tag{3.2}$$

Normalisation is performed separately for each feature by calculation of a normalisation factor $h^r_{j\,p}$. The factor is determined using feature data from the reference dataset X^r using Equation (3.1); this is done for features $j = 1$ to j_n and classes $p = 1$ to k. The normalised term $\theta^r_{ij\,p}$ is obtained using (3.2).

$$w_j^r = \frac{\sigma_j^r}{\sum_{j=1}^{n} \sigma_j^r} \tag{3.3}$$

where σ_j^r is the standard deviation of feature j.

After normalising the reference dataset X^r, the feature weights w_j^r are calculated using Equation (3.3). This method of weighting relies on the features variance having a positive correlation with their respective impact to the classification result [74].

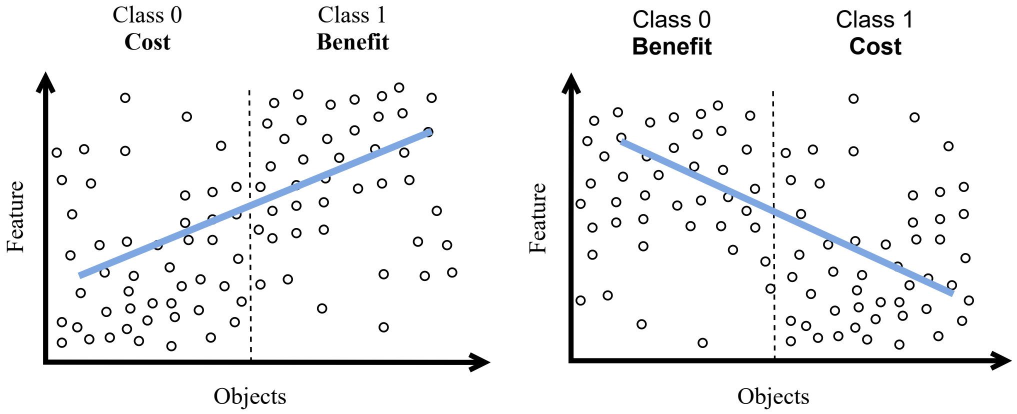


Figure 3.1: The relationship between costs and benefits for binary datasets

To be able to determine the ideal solutions, features have to be assigned as *costs* C or *benefits* B for each class. A feature that is a *cost* for a particular class, becomes larger for records pertaining to the class and, becomes smaller if not a member of the class, as illustrated in Figure 3.1.

$$\begin{aligned} f_p^+ &= \{\theta_1^{r^+}, \theta_2^{r^+}, \ldots, \theta_n^{r^+}\} \\ &= \Big\{(max_i \theta_{ij_p}^r / j \in B), (min_i \theta_{ij_p}^r / j \in C)\Big\} \end{aligned} \tag{3.4}$$

$$\begin{aligned} f_p^- &= \{\theta_1^{r^-}, \theta_2^{r^-}, \ldots, \theta_n^{r^-}\} \\ &= \Big\{(min_i \theta_{ij_p}^r / j \in B), (max_i \theta_{ij_p}^r / j \in C)\Big\} \end{aligned} \tag{3.5}$$

where f_p^+ and f_p^- are the positive and negative ideal solutions, respectively; $\theta_n^{r^+}$ and

θ_n^{r-} is the positive and negative ideal solution for feature n; $\theta^r_{ij_p}/j$ is the term for object i, feature j and class p, from the normalised reference dataset. The ideal solutions are calculated for features $j = 1$ to j_n, classes $p = 1$ to k.

Baccour suggested expert knowledge be used, where available, to determine C and B [14]. Nonetheless, no method was given in case expert knowledge was absent, implying the designer must have a certain level of expert knowledge about the dataset to be able to construct the ATOVIC model. When the *costs* and *benefits* are determined, the ideal solutions are simply calculated from the normalised features θ_n as defined in Equations (3.4-3.5).

$$S_{c_i} = \sum_{j=1}^{n} w_j^r * (f_{ij_c}^+ - \theta_{ij_c})/(f_{ij_c}^+ - f_{ij_c}^-),\ S_{c_i} \in [0,1] \tag{3.6}$$

$$R_{c_i} = \max_j \left[w_j^r * (f_{ij_c}^+ - \theta_{ij_c})/(f_{ij_c}^+ - f_{ij_c}^-) \right],\ R_{c_i} \in [0,1] \tag{3.7}$$

$$Q_{c_i} = \rho \frac{S_{c_i} - S_c^-}{S_c^+ - S_c^-} + (1-\rho)\frac{R_{c_i} - R_c^-}{R_c^+ - R_c^-},\ \rho \in [0,1] \tag{3.8}$$

where i is the row (record), j is the feature (input), ρ is the weighting parameter and;

$$S_c^+ = \max(S_c^r) \tag{3.9}$$

$$S_c^- = \min(S_c^r) \tag{3.10}$$

$$R_c^+ = \max(R_c^r) \tag{3.11}$$

$$R_c^- = \min(R_c^r) \tag{3.12}$$

where (3.9-3.12) are defined for classes $c \in [1..k]$, where k is the number of classes.

The ideal solutions are utilised to calculate the different distance measures S, R and Q; Manhattan (Equation (3.6)), Chebyshev (Equation (3.7)) and the normalised weighted sum of both (Equation (3.8)), respectively. The measures S and R are normalised using *linear scaling* to determine the weighted normalised sum Q by using Equations (3.8-3.12). To be able to determine the extremes of S and R; they are calculated from the reference dataset Θ^r for classes $c \in [1..k]$.

In order to classify data using ATOVIC, the minimums of the measures are used to determine which class they point to (Figure 3.2). Since there are three measures, there would be three results for each object. Hence, a series of logical operations are executed to narrow down the result based on the three results. The following steps are explained for binary and multi-classification applications. The first step is to assess whether there is a dominant class within the set of results. If yes, it will be used to determine the classification result, as illustrated in Figure 3.2. In cases where a dominant class does not exist, the minimums will be assessed to determine whether there is consensus between min_Q and min_R, or min_Q and min_S.

3.3.2 Improvements to ATOVIC

The original version of ATOVIC, as defined by Baccour in 2018, has its limitations [14]. The model was mostly data-driven. However, it relies on expert knowledge to determine whether each feature is a *cost* or *benefit*. This prevents ATOVIC from being widely explored and implemented for problems where such expert knowledge does not exist.

$$r_j = \frac{\sum_{i=1}^{m}(x_{ij} - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{m}(x_{ij} - \bar{x})^2}\sqrt{\sum_{i=1}^{m}(y_i - \bar{y})^2} \tag{3.13}$$

where x is the value for row (object) i and feature j. The second variable y is the labelled class for object i; and m is the number of objects.

$$\text{Class 1: } \texttt{IF } r_j > 0 \texttt{ THEN: } j \texttt{ is Cost; } \text{ELSE: } j \texttt{ is Benefit} \tag{3.14}$$

$$\text{Class 2: } \texttt{IF } r_j > 0 \texttt{ THEN: } j \texttt{ is Benefit; } \text{ELSE: } j \texttt{ is Cost} \tag{3.15}$$

As an improvement to ATOVIC, a method for classifying features as *costs* or *benefits* was developed to eliminate the requirement for expert knowledge. The method uses the Pearson correlation coefficient (see Equation (3.13)) as a means of determining costs and benefits. The correlation r_j of each feature j with the labelled class y is calculated. The correlation's polarity is used to classify features as a cost or benefit, as defined in Equations (3.14-3.15).

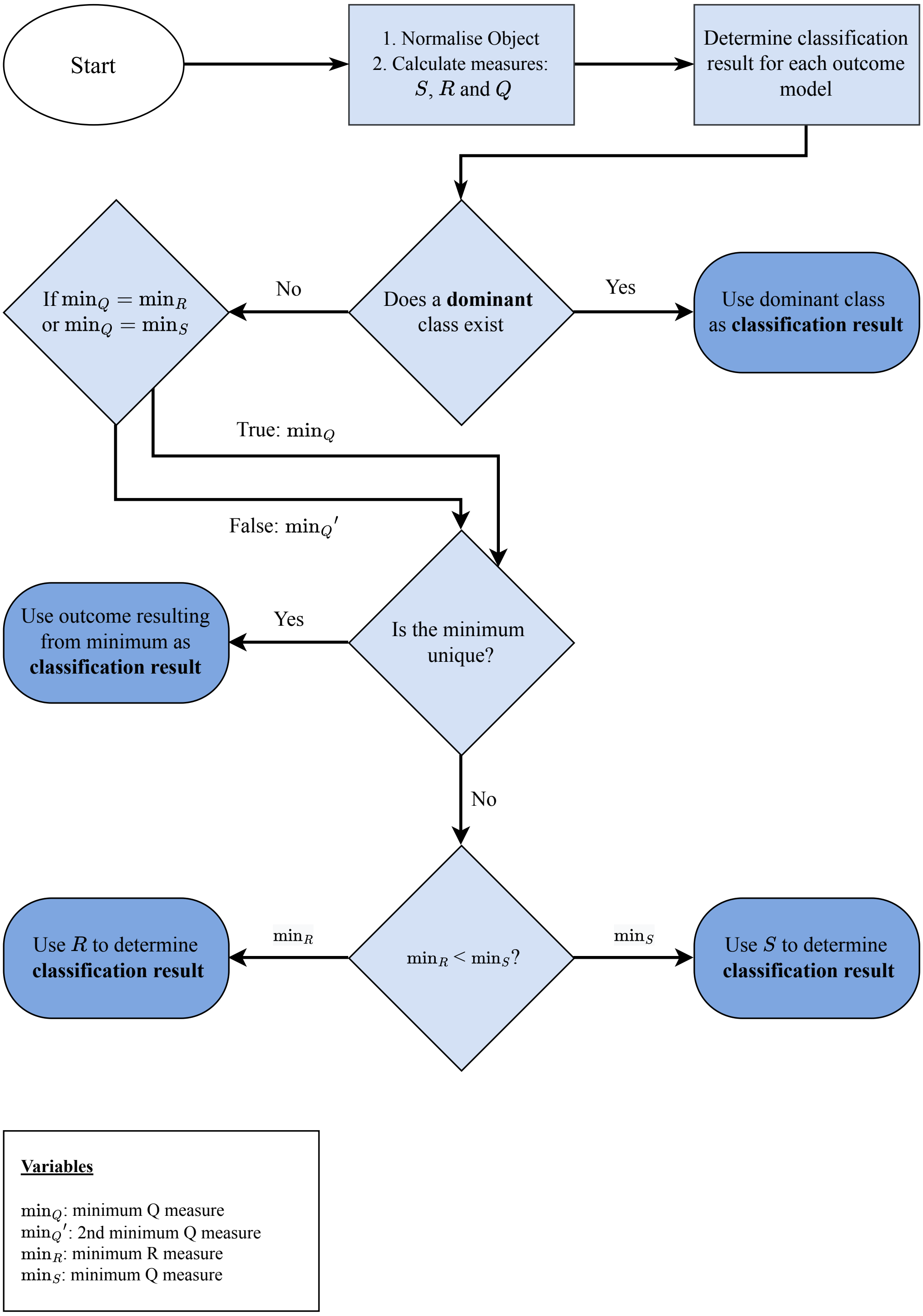

Start
1. Normalise Object
2. Calculate measures: S, R and Q
Determine classification result for each outcome model
Does a **dominant** class exist
Yes
Use dominant class as **classification result**
No
If $\min_Q = \min_R$ or $\min_Q = \min_S$
True: $\min_Q$
False: $\min_Q{}'$
Is the minimum unique?
Yes
Use outcome resulting from minimum as **classification result**
No
$\min_R < \min_S$?
$\min_R$
Use R to determine **classification result**
$\min_S$
Use S to determine **classification result**
Variables
$\min_Q$: minimum Q measure
$\min_Q{}'$: 2nd minimum Q measure
$\min_R$: minimum R measure
$\min_S$: minimum Q measure

Figure 3.2: ATOVIC classification based on the measures S and R.

$$w_j = \frac{r_j}{\sum_{j=1}^{n} r_j} \tag{3.16}$$

where r_j is Pearson coefficient (3.13) and, n is the number of features.

The weight calculation method was also modified, in Equation (3.16), to use the correlation coefficient as opposed to the standard deviation in the original version of ATOVIC. Standard deviation as a weight estimator could be satisfactory for certain types of datasets. However, the correlation coefficient extracts the statistical relationship between each feature and the class; hence, it is expected to enable superior performance.

3.3.3 Extension to fuzzy-MCDM

Fuzzy-MCDM is a class of MCDM extended with fuzzy logic sets (FLSs) for enhanced performance, adaptability and functionality. This can be achieved by incorporating different aspects of FLSs, such as fuzzy numbers or fuzzy inference systems. Fuzzy-ATOVIC is a version of ATOVIC which has been extended with FL Sets. In the first iteration of fuzzy-ATOVIC, it was augmented with a FIS to replace the final classification steps in ATOVIC. The rationale behind replacing condition statements with a FIS is to improve its interpretability. Although the condition statements provide interpretability, the FIS provides a fuzzy output which does a better job of reflecting how certain the result is - a key advantage of Fuzzy-ATOVIC versus ATOVIC.

The FIS takes in the distance measure as an input and provides a fuzzy class output ranging from 0 to 1. The distance measure can be either of the three distance measures: Q, R and S. The fuzzy output is defuzzified to determine the crispy class by the use of a threshold, as illustrated in Figure 3.3.

The FIS rules are formulated based on an understanding of how the measure reflects similarity and, whether the two sub-models are in agreement, as illustrated in Figure 3.4. The measures from both models are assessed as HI or LO - where HI is larger than 0.5 and LO is lower than 0.5. This is suitable because the measures are unity normalised. Consequently, when both models are in agreement, the sub-models

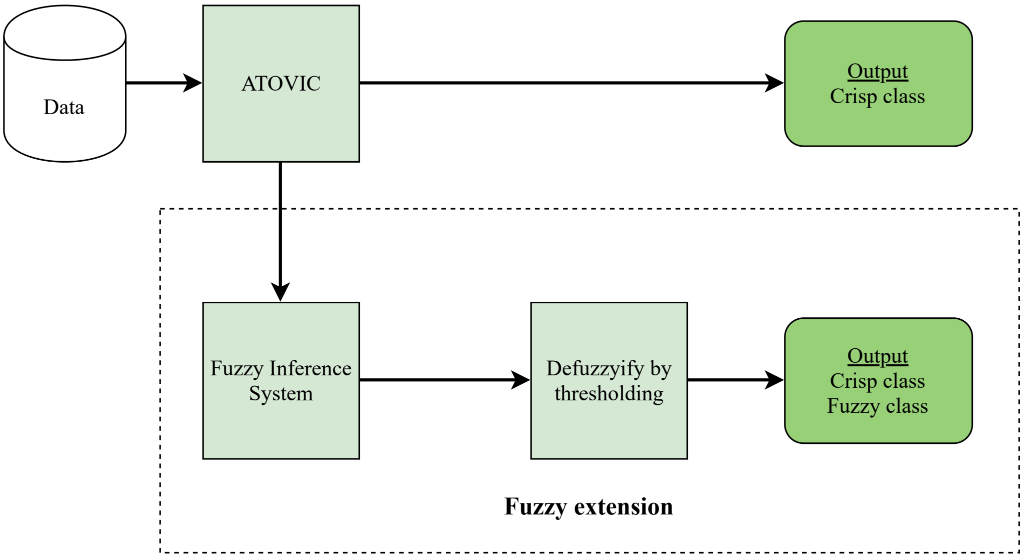


Figure 3.3: High-level structure of a Fuzzy-MCDM model

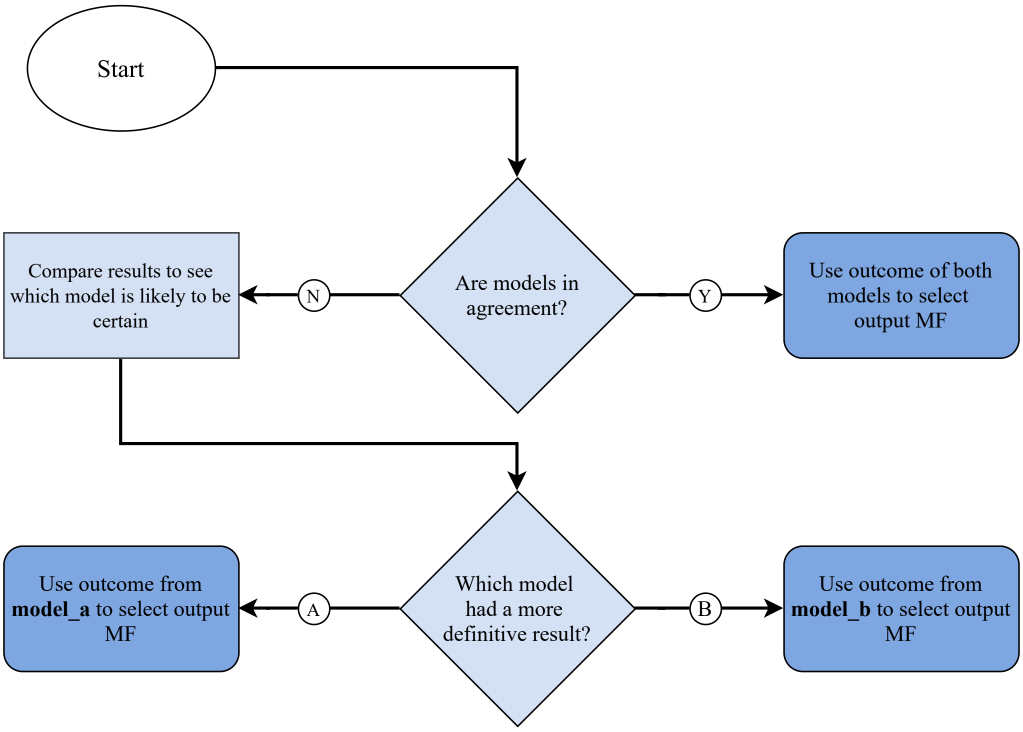

Start
Are models in agreement?
N
Y
Compare results to see which model is likely to be certain
Use outcome of both models to select output MF
Which model had a more definitive result?
A
B
Use outcome from model_a to select output MF
Use outcome from model_b to select output MF

Figure 3.4: The process of formulating FIS rules to process a single measure (Q, R or S)

Table 3.1: FIS structure: input-output MFs; where M represents either of the measures Q, R or S.

Inputs	MFs	min	max
ΔD_1	2	0	1
ΔD_2	2	0	1
n_D	2	-1	1
Output	2	0	1

outcomes are used as a basis for the decision. Meanwhile, when the sub-models are in conflict, the more definitive result from either sub-model is used. The rules are formulated such that the more definitive model is determined numerically, using n_D, as defined by Equation (3.18). Since n_D is a difference, its polarity can be used by the FIS to discern which sub-model's decision is more definitive. The MFs have been designed intuitively such that a negative n_D would mean the negative sub-model is more definitive; conversely, a positive n_D means the positive sub-model is more definitive.

(a) D_c uses two MFs: `class_1` (HI), `class_2` (LO)

(b) n_D uses two MFs: `positive` (positive outcome model is used for decision), `negative` (negative outcome model is used for decision)

(c) Output: two MFs: `class_1`, `class_2`

$$\Delta D_c = d_{s,2} - d_{s,1} \tag{3.17}$$

where s is the sub-model number; D represents one of the three measures: Q, R or S; $d_{c,1}$ is the distance to class one and, $d_{c,2}$ is the distance to class two. Values for ΔD_c are computed for $c = 1 \ldots k$ - where k is the number of classes.

$$n_D = |\Delta D_2| - |\Delta D_1| \tag{3.18}$$

The measures have an inverse relationship with similarity, similar to a distance; as

Table 3.2: Fuzzy-MCDM classifiers: overview

Model	Measures	Normalisation	FIS Type
Fuzzy-ATOVIC	Q, R, S	Eigenvalue	Hierarchical
Fuzzy-TOPSIS	S	Eigenvalue	Singular
Fuzzy-VIKOR - A	Q, R, S	Linear	Hierarchical
Fuzzy-VIKOR - B	Q, R, S	None	Hierarchical

the measure falls, the similarity is higher. Each sub-model provides two values of each measure - one for each class. Thus, the two sub-models provide four $D_{s,c}$ values, where s is the sub-model number, and c is the class number.

ATOVIC as a framework uses three measures Q, R and S. Hence, a set of three FISs is used in parallel to process the measures resulting in a Fuzzy Class Output from each, as illustrated in Figure 3.5. The decision is an aggregation of the three FIS outputs; the same applies to VIKOR. However, TOPSIS relies on a single measure (measure S); therefore, a single FIS is sufficient for classification. The difference in the ways classification occurs for the different MCDM methods could have an effect on performance, as will be investigated in the next section. The differences include the measures, normalisation method and FIS Types, as shown in Table 3.2.

3.4 Assessing fuzzy-MCDM-based classifiers

The aim of this section is to evaluate the *proposed* set of MCDM-based classifiers against the standards expected from state-of-the-art classifiers. The section starts with an introduction to the benchmark datasets used and the rationale as to why they were selected. Consequently, three comparative studies are presented assessing Fuzzy-MCDM frameworks. The first study compares Fuzzy-MCDM to a set of state-of-the-art classification frameworks. Next, all Fuzzy-MCDM methods: Fuzzy-ATOVIC, Fuzzy-TOPSIS and Fuzzy-VIKOR, are compared against each other. Finally, the section is concluded with a comparison between ATOVIC and Fuzzy-ATOVIC to demonstrate the effect of the fuzzy extension on performance. The results are analysed with the purpose of assessing Fuzzy-MCDM feasibility as an interpretable data-driven framework for classification.

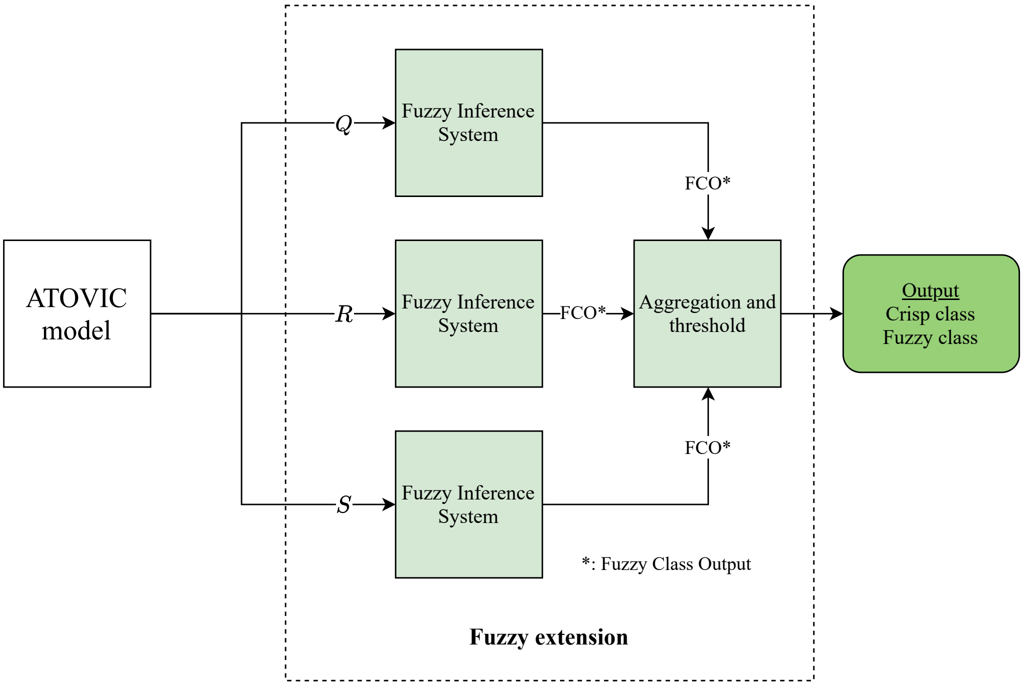


Figure 3.5: High-level structure of fuzzy component for Fuzzy-ATOVIC

3.4.1 Benchmark datasets

In this chapter, benchmark datasets are utilised to investigate the performance of various types of models against the proposed Fuzzy-MCDM-based classification techniques, as defined in Section 3.3. The datasets were retrieved from public repositories: KEEL-dataset [2] [75], and the University of California's (UCI) Machine Learning Repository [76]. In this section, the structure of the datasets and the rationale behind their selection are described.

The proposed models are being investigated from interpretability so it was important to be able to assess how the human-understandability of features can impact this. Thus, a combination of human-understandable and non-understandable datasets were selected. Also, the datasets comprised a range to dimensionalities in order to assess the proposed model's performance and interpretability in relation to dataset complexity.

As data moves to the cloud in many fields, the availability of datasets is improved, to the delight of data-hungry ML scientists around the world. The medical field is

[2] KEEL-dataset: Knowledge Extraction based on Evolutionary Training dataset repository

Table 3.3: Parkinson's disease speech dataset: attribute information

Attribute	Type	Count
ID	Integer	1
Gender	Category	1
Baseline Features	Real	21
Intensity Parameters	Real	3
Formant Frequencies	Real	4
Bandwidth Parameters	Real	4
Vocal Fold	Real	22
Mel Frequency Cepstral Coefficients (MFCCs)	Real	84
Wavelet Features	Real	182
Tunable Q-Factor Wavelet Transform (TQWT)	Real	432
	Total	754

Table 3.4: Parkinson's disease speech dataset: class information

No.	Class	Count
1	Negative	192
2	Positive	564
	Total	756

no exception, with multiple online repositories storing datasets for ML research.

Data classification is useful for *medical diagnosis*, where patients need to be classified as positive or negative. The first dataset used is for predicting whether a patient suffers from Parkinson's disease (PD), based on features extracted from a recording of the patient speaking. The dataset contains two classes (negative and positive) with significantly more instances representing positive data than negative, as shown in Table 3.4. The dataset aims to diagnose patients based on an audio recording of them speaking. Hence, the attributes include signal processing features extracted from the recordings. Moreover, the signal features aim to capture the patient's speech frequency and amplitude characteristics via a series of wavelets and transforms, as listed in Table 3.3.

For the second case study, the breast cancer dataset attributes rely on information acquired by clinicians through breast cancer screening. The attributes score various characteristics of a tumour, as listed in Table 3.5. The objective is to classify tumours as either malignant or benign, i.e. cancerous or non-cancerous, respectively. The two

Table 3.5: Breast cancer Wisconsin: attribute information

Attribute	Type	Domain
ID	Integer	1
Clump Thickness	Real	$[1, 10]$
Uniformity of Cell Size	Real	$[1, 10]$
Uniformity of Cell Shape	Real	$[1, 10]$
Marginal Adhesion	Real	$[1, 10]$
Single Epithelial Cell Size	Real	$[1, 10]$
Bare Nuclei	Real	$[1, 10]$
Bland Chromatin	Real	$[1, 10]$
Normal Nucleoli	Real	$[1, 10]$
Mitoses	Real	$[1, 10]$
	Total	754

Table 3.6: Breast cancer Wisconsin: class information

No.	Class	Count
1	Benign (non-cancerous)	458
2	Malignant (cancerous)	241
	Total	699

classes are also imbalanced. However, the negative class is over-represented in this dataset compared to the positive class for the PD dataset.

The next dataset that will be used is a Chess dataset. Any complete game of Chess has three stages from start to finish: opening, middlegame and endgame. The manner in which each player undertakes their three stages impacts the likelihood of them winning. This dataset is concerned with a certain type of endgames where the remaining pieces are the king and a rook for white, while the pieces remaining for black are a king and pawn. The attributes contain information about where the pieces are located, with a total of 36. The two classes are reasonably balanced (Table 3.7) and represent whether White can win or lose.

Table 3.7: Chess KR-vs-KP: class information

No.	Class	Count
1	Win	1669
2	Lose	1527
	Total	699

Table 3.8: Keel titanic dataset: attribute information

Attribute	Type	Domain
Class	Category	$[-1.870, 0.965]$
Age	Category	$[-0.228, 4.38]$
Sex	Category	$[-1.920, 0.521]$

Table 3.9: Kaggle titanic dataset: attribute information

Attribute	Type	Possible Values
ID	Integer	[1..891]
Ticket Class	Integer	[1..3]
Sex	Category	M, F
Age	Real	$[0, 80]$
Sibling(s)/Spouse(s)	Integer	[0..8]
Parent(s)/Children	Integer	[0..6]
Embarked	Category	C, Q, S

The last two datasets are Titanic passenger datasets for predicting passenger survival. The first Titanic dataset was retrieved from the KEEL repository and had three attributes (Table 3.8). Moreover, the second dataset contained a larger number of attributes, including further information such as the number of family members, passenger's age, and the port embarked from.

Pre-processing of the data entailed encoding categorical features into a numeric format. This was done because MCDM, in its current format, cannot handle categorical features. Consequently, ten randomised resamples were generated for each of the datasets in a 5-fold format. The k-fold cross-validation method is considered to have less bias than the traditional 1-fold approach, which employs a single set of training and testing datasets.

Table 3.10: Keel and Kaggle titanic datasets: class information

No.	Class	Keel	Kaggle
1	Survived	1490	549
2	Casualty	711	342
	Total	2201	891

Table 3.11: Parkinson disease: performance

Model	ACC (%)	F-Score (%)	TPR (%)	TNR (%)
Fuzzy-ATOVIC	78.8	70.6	84.6	61.6
Decision Trees	77.7	70.3	83.3	61.4
KNN	86.9	88.4	85.0	92.3
SVM	87.5	81.2	92.5	72.7
NB	44.2	42.9	28.3	90.7

3.4.2 Comparative analysis of MCDM-based and SOA classification frameworks

In this section, MCDM-based classifiers are compared to a set of state of the art (SOA) techniques. For simplicity, the best-performing proposed MCDM-based classification model is compared to a set of SOA techniques, including DT, K-Nearest Neighbour (KNN), Support Vector Machines (SVM) and Naive Bayes (NB).

For the PD dataset, the performance trade-off was considerable, with a performance drop of 20% when comparing the best MCDM to SVM, as shown in Table 3.11. One of the possible reasons for this is the dataset had a large number of features. MCDM's inherent simplicity is weak point for dealing with more complex problems. The inferior performance is also attributable to the MCDM's simpler training framework, which relies on a single iteration to *fit* the model, underscoring the edge of iterative *learning*. The single iteration is executed based on the steps described in Section 3.3. Moreover, certain types of real-world datasets may contain an excess of a thousand features. Thus, it may be suitable to employ feature extraction or selection to ensure MCDM-based classifiers have a better chance of performing satisfactorily.

Contrary to the first dataset, for the breast cancer case study, the MCDM-based classifier performs as good as or outperforms the SOA classifiers. When training using the breast cancer dataset (in Table 3.12), Fuzzy-ATOVIC performed marginally similar to other SOA with an accuracy of 95.6%. Since all models were able to perform well, it is clear this is a good quality dataset with clear links between the features and the different classes. Nevertheless, it serves as an example

Table 3.12: Breast cancer: performance

Model	ACC (%)	F-Score (%)	TPR (%)	TNR (%)
Fuzzy-VIKOR (no-norm)	95.9	94.7	91.6	98.1
Decision Trees	93.9	93.5	92.6	94.7
KNN	95.5	94.7	92.7	97.0
SVM	96.6	96.7	97.2	96.3
NB	96.8	96.8	97.2	96.5

Table 3.13: Keel titanic: performance

Model	ACC (%)	F-Score (%)	TPR (%)	TNR (%)
Fuzzy-ATOVIC	77.8	63.8	49.1	91.2
Decision Trees	35.2	8.9	98.7	4.9
KNN	69.1	8.1	4.2	100.0
SVM	32.3	0.0	100.0	0.0
NB	32.3	0.0	100.0	0.0

of where MCDM-based classifiers can excel and could ultimately be developed as an intrinsically *explainable* framework.

Running the same models on the KEEL Titanic dataset demonstrates a situation where an MCDM-based classifier was able to outperform SOA classification techniques. In this particular case, over-fitting caused the testing data performance to be particularly low for SVM and NB despite both models performing excellently for the training dataset. Since MCDM-based classifiers perform *fitting* rather than *learning*, they perform better for datasets prone to over-fitting.

Despite KEEL Titanic having a simpler dataset, the quality of data was lacking compared the breast cancer dataset. Therefore, the models performed consistently better using the breast cancer dataset.

A great deal of discussion surrounds the viability of interpretable models due to their perceived inferior accuracy performance. The theory that opting for interpretable models means trading off performance is frequent in the literature [77]–[80]. Despite this, other researchers have suggested the widely cited performance trade-off theory is yet to be proven [2], [81]. The results provide a promising indication that interpretable models could perform as good as less

Table 3.14: Parkinson disease: performance

Model	ACC (%)	F-Score (%)	TPR (%)	TNR (%)
Fuzzy-ATOVIC	78.8	70.6	84.6	61.6
Fuzzy-TOPSIS	78.7	70.5	84.6	61.4
Fuzzy-VIKOR (linear-norm)	76.1	72.2	78.4	69.0
Fuzzy-VIKOR (no-norm)	77.4	73.6	80.6	68.2

interpretable models. More specifically, it demonstrates MCDM-based classifiers performing comparably to SOA classifiers - an important starting point for any further investigation of MCDM as a dependable data-driven classifier. The rationale behind opting for MCDM-based classifiers as opposed to SOA interpretable techniques is the higher interpretability providing greater potential for explainability.

3.4.3 Comparative analysis of MCDM-based frameworks for classification

In this part, four types of Fuzzy-MCDM techniques (listed below) are compared against each other in order to identify the most effective for classification in terms of performance and interpretability. The use of MCDM-based classifiers was justified by the last section, where it was shown the performance trade-off is manageable for certain applications.

(a) Fuzzy-ATOVIC: a fusion of the benefits of TOPSIS and VIKOR. First introduced by Baccour in 2018 [14] and consequently extended with FL to form Fuzzy-ATOVIC.

(b) Fuzzy-TOPSIS: a FL extended version of TOPSIS tailored for classification

(c) Fuzzy-VIKOR (linear-norm): a FL extended version of VIKOR tailored for classification, using linear normalisation

(d) Fuzzy-VIKOR (no-norm): a FL extended version of VIKOR tailored for classification, with no normalisation

The PD dataset contains 754 attributes. A limitation that has been suggested for interpretable models is the claim they cannot perform well with larger datasets.

Table 3.15: Breast cancer: performance

Model	ACC (%)	F-Score (%)	TPR (%)	TNR (%)
Fuzzy-ATOVIC	93.2	89.9	82.6	98.8
Fuzzy-TOPSIS	93.2	89.9	82.6	98.8
Fuzzy-VIKOR (linear-norm)	95.8	94.6	91.5	98.1
Fuzzy-VIKOR (no-norm)	95.9	94.7	91.6	98.1

Table 3.16: Chess: performance

Model	ACC (%)	F-Score (%)	TPR (%)	TNR (%)
Fuzzy-ATOVIC	81.1	80.5	75.4	86.5
Fuzzy-TOPSIS	81.2	80.5	75.3	86.6
Fuzzy-VIKOR (linear-norm)	57.7	31.7	99.7	19.3
Fuzzy-VIKOR (no-norm)	57.4	31.0	100.0	22.9

Although 754 is not considered large by *Big Data* standards, it is nonetheless a starting point for assessing the ability of MCDM to handle larger datasets. When tested with the PD dataset, all MCDM methods achieved greater than 75% accuracy in predicting a patient, as shown in Table 3.14. The highest scoring method was Fuzzy-ATOVIC at 78.8% accuracy. Despite Fuzzy-ATOVIC performing the highest on the accuracy, Fuzzy-VIKOR (no-norm) has the highest F-Score at 73.6%, which signifies a higher balance between sensitivity and specificity (TPR and TNR, respectively). The first set of results indicates performance is largely similar. Hence, there is no clear advantage to using a specific model for the sake of accuracy levels.

Furthermore, when running the same models on the breast cancer dataset, the performance levels (in Table 3.15) indicate better overall performance for Fuzzy-VIKOR versions with and without normalisation. The trade-off for Fuzzy-ATOVIC in this instance is the lower specificity rate which is a worthy trade-off for better sensitivity - the successful detection of positive patients. In contrast, when running the models on the Chess dataset Fuzzy-ATOVIC and Fuzzy-TOPSIS performed better, as shown in Table 3.16. The Chess dataset is a non-linear problem relying on categorical attributes that are represented numerically. The structure of the data meant the method of normalisation used in Fuzzy-ATOVIC and Fuzzy-TOPSIS had a negative impact on performance.

Table 3.17: Keel Titanic: performance

Model	ACC (%)	F-Score (%)	TPR (%)	TNR (%)
Fuzzy-ATOVIC	77.8	63.8	49.1	91.2
Fuzzy-TOPSIS	77.8	63.8	49.1	91.5
Fuzzy-VIKOR (linear-norm)	77.6	63.2	48.4	91.5
Fuzzy-VIKOR (no-norm)	77.6	63.2	48.4	91.5

Table 3.18: Kaggle Titanic: performance

Model	ACC (%)	F-Score (%)	TPR (%)	TNR (%)
Fuzzy-ATOVIC	78.8	75.9	68.7	85.1
Fuzzy-TOPSIS	78.8	75.9	68.7	85.1
Fuzzy-VIKOR (linear-norm)	78.9	75.9	68.6	85.2
Fuzzy-VIKOR (no-norm)	78.8	75.9	68.6	85.2

The ideal solutions were inspected to understand how normalisation impacted performance. It revealed that linear normalisation, used for Fuzzy-VIKOR (linear-norm), caused some ideal solutions to be set as undefined values such as `Inf` or `NaN`. This rendered the ideal solutions ineffective. Furthermore, when non-normalised data (Fuzzy-VIKOR (no-norm)) was used for numerical categorical data, it resulted in ideal solutions set at arbitrary integer values. This issue was only apparent with the Chess dataset. Notably, the nature of multi-criteria frameworks dictates that normalisation is vital for managing bias because of the variability of the criteria's ranges.

Finally, the model was run on two different Titanic datasets containing attributes of passengers with the aim of predicting their survival. Interestingly, performance characteristics (in Tables 3.17-3.18) are similar, with better specificity (TNR) than sensitivity (TPR). However, the models had around 19% better sensitivity (TPR), which resulted in an improved F-Score. Furthermore, the different types of normalisation did not impact performance for the Titanic datasets.

To summarise, the performance results indicate that all four implementations of Fuzzy-MCDM-based classifiers perform similarly for all datasets, except the Chess dataset, in which case Fuzzy-ATOVIC and Fuzzy-TOPSIS had better performance. As described in Section 3.3, the different Fuzzy-MCDM classifiers differ in parameter calculation technique and not in the general structure; therefore, the level of interpretability is considered largely similar.

The expectation was that the normalisation technique was going to have considerable impact on performance for all datasets. However, the results show that the normalisation technique does not have a significant impact on performance for most datasets tested. The only exception is the Chess dataset, where performance deteriorated for the two Fuzzy-ATOVIC models because of the more suited normalisation used by Fuzzy-ATOVIC and Fuzzy-TOPSIS.

3.4.4 Comparing ATOVIC and fuzzy-ATOVIC

In this part, ATOVIC is compared to the fuzzy proposed version, Fuzzy-ATOVIC. The aim is to assess whether the proposed fuzzy extension resulted in a trade-off or improvement. Based on the theoretical link between interpretability and performance [77]–[80], it is expected the fuzzy extension could have a negative impact on classification accuracy because of the increased interpretability of the fuzzy version. However, the results yield a different story where Fuzzy-ATOVIC outperforms ATOVIC in four out of the five datasets (Table 3.19).

The difference in structure between ATOVIC and Fuzzy-ATOVIC lies mainly in the fuzzy extension. In it, a Fuzzy Inference System (FIS) is used to perform the classification, based on the ATOVIC measures Q, R and S. By processing the three measures separately in three separate FISs, it means the varying magnitudes of the measures, due to the different distancing types, do not bias the decision-making towards a specific distance type. Since ATOVIC utilises sub-models for representing each class, bias could arise when a sub-model becomes representative of more data. This could be due to a myriad of reasons that such as the feature-set, weights and ideal solutions.

In addition, the ATOVIC model performed poorly for three out of five of the datasets. Meanwhile, the Fuzzy-ATOVIC model improved accuracy and F-Measure performance across all five datasets, which indicates it is likely to be suitable for a wider variety of data problems. Moreover, the FIS adds an additional layer of interpretability, as will be demonstrated in the next chapter.

Table 3.19: ATOVIC and Fuzzy-ATOVIC performance results

	ATOVIC		Fuzzy-ATOVIC	
Model	ACC	F-Score	ACC	F-Score
Parkinson disease	58.2	60.5	78.8	70.4
Breast Cancer	68.3	14.5	93.2	89.8
Chess	60.7	30.5	81.2	80.6
Keel Titanic	77.6	63.2	75.4	68.3
Kaggle Titanic	74.5	51.2	78.8	75.9

Moreover, the theory that interpretable modelling results are a trade-off of performance is not a universal rule [2]. The results indicate a case were ATOVIC was able to perform as good as SOA classification frameworks. Therefore, the investigation of MCDM as a dependable data-driven classification framework is considered a promising research gap.

The lack of optimisation capabilities prevented MCDM-based classifiers from performing favourably with larger datasets such as the PD dataset. For this dataset, Fuzzy-ATOVIC and DT performed 10% less accurately compared to SVM models - highlighting a potential performance trade-off. Extensive literature show how interpretability forces the trade-off of performance in a number of ML models [77]–[80]. Nonetheless, the *curse of interpretability* on performance is a phenomenon still worth investigating further for the fact it is yet to be unequivocally proven [2], [81].

3.5 Summary

This chapter was the starting point for investigating a relatively new class of methodologies for interpretable data-driven classification: MCDM. Three different types of MCDM classifiers are defined: ATOVIC, VIKOR and TOPSIS. Consequently, the models were extended with FL by the use of a Fuzzy Inference System (FIS). The FIS was developed for processing the measures/distances generated by the MCDM classifiers to perform the final classification. The results demonstrate Fuzzy-MCDM's potential for performing comparably to SOA models. Examples of this include Fuzzy-ATOVIC achieving 95.9% accuracy, which was only

1% less than the best performing SOA model: Naive Bayes. Moreover, for the Keel Titanic dataset, Fuzzy-ATOVIC was the best performing at 77.8%. As expected, there are cases where MCDM classifiers are lacking, such as the PD dataset, where performance was particularly lower compared to SOA models. This highlights MCDM's limitation in dealing with larger feature sets.

Comparing the different Fuzzy-MCDM revealed performance was largely similar except for cases where certain normalisation techniques resulted in undefined values for categorical data. Hence, the recommendation is to utilise vector normalisation to avoid undefined values.

Finally, comparing ATOVIC to its fuzzy extended version demonstrated a noticeable improvement in performance and consistency of results. Therefore, Fuzzy-ATOVIC users will be able to benefit from the increased performance of FL and enhanced interpretability of ATOVIC.

Despite the benefits of MCDM-based classifiers, the methodology in its current form lacks the *training* rigour provided by state-of-the-art data-driven classifiers. Therefore, its classification performance can be quickly limited when dealing with complex datasets.

Moreover, its simple structure prevents it from handling datasets with a large number of features. As a result, the methodology's applicability is not as wide as state-of-the-art classifiers.

The main rationale for using MCDM is its inherent interpretability. If MCDM can perform as good as SOA models, it is considered to be a viable candidate for data-driven explainable-AI. The MCDM-based classifiers explainability is a gap that will be investigated in the next chapter.

4 Towards an explanation framework for Fuzzy-MCDM

Data-driven classifiers based on MCDM and fuzzy logic possess significant interpretability potential. MCDM, and in a more general sense, Decision Theory has been one of the few areas that are yet to be explored as a possible jigsaw piece to explainable AI. The area is often used to tackle a variety of problems in relation to how humans make a decision using knowledge from disciplines such as psychology, philosophy and cognition. Since the success of explainable-AI is dependent on how the human user benefits from the generated explanation, Decision Theory could prove useful in this regard as it is an area focused on how humans make decisions. MCDM are a set of modelling techniques based on the principles of Decision Theory. In this chapter, an explanation framework is introduced for the first time, as a tailored augmentation for MCDM-based classifiers. The framework uses model-based information to generate insightful explanatory output. The *insight* describes the states of internal classifier parameters and how they impact the model's decision in a manner that is similar to a human's decision making process.

4.1 Introduction

In the previous chapter, a set of Fuzzy-MCDM classifiers were proposed. They were found to be a feasible option for certain areas where they achieve satisfactory accuracy compared to state-of-the-art classifiers. This chapter explores one of the main benefits of using interpretable modelling: the potential for explainability. The chapter is started with a literature review on *interpretability and explainability*, as shown in Figure 4.1. Consequently, a comprehensive explainable framework is proposed for the first time for Fuzzy-MCDM. The framework is implemented for five benchmark datasets - some of which consist of human-understandable features. The chapter is concluded with a discussion appraising the explanation framework as an XAI solution.

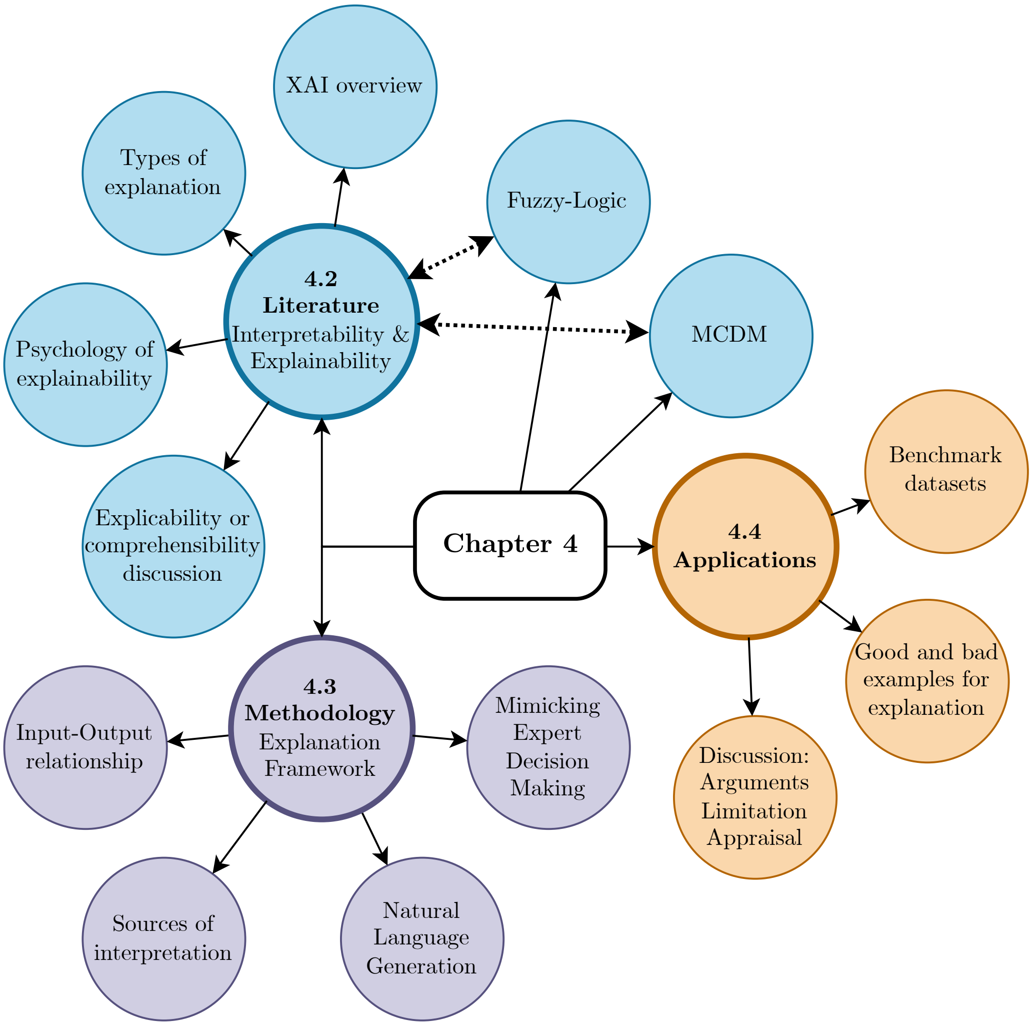

XAI overview
Types of explanation
Fuzzy-Logic
4.2 Literature Interpretability & Explainability
Psychology of explainability
MCDM
Explicability or comprehensibility discussion
Chapter 4
4.4 Applications
Benchmark datasets
Good and bad examples for explanation
4.3 Methodology Explanation Framework
Input-Output relationship
Mimicking Expert Decision Making
Discussion: Arguments Limitation Appraisal
Sources of interpretation
Natural Language Generation

Figure 4.1: Chapter 4: mind map of general topics and concepts

4.2 Explainability: an overview

4.2.1 The importance of model transparency

The term *model transparency* is often used as synonymous term for describing white box models. However, the last decade's growing interest for explainable-AI meant *model transparency* has become a *criterion* for assessing how interpretable a model is [12]. Lipton suggests three different transparency categories: decomposability, simulatability and algorithmic transparency. Decomposability is defined by Lipton as the degree to which a model can be decomposed into distinguishable components with the requirement of understanding the purpose of each. Moreover, simulatability focuses on how *intuitive* it could be to *simulate* i.e. execute the model manually with pen and paper by a human. Lipton explains how decomposable and simulatable a model is associated with its inherently interpretable. The last type of transparency, algorithmic, refers to the modelling algorithm's human-understand-ability. Although the definitions defined by Lipton are not universal, they pose a promising theory towards a complete definition of interpretability.

The importance of model transparency varies depending on preferred route to *explainability*. Explanation frameworks rely on a source to interpret information from; this can be either *model-based* or a product of *post-hoc* analysis [12], [38]. Post-hoc interpretability relies solely on the input and output data to provide model insight. Meanwhile, model-based is able to utilise internal model parameters in addition to input-output data; therefore, model-based have more potential for providing meaningful explanation inferred based on model execution. The main downside of model-based interpretability is that is it incompatible with all model types, most importantly robust complex methodologies such as deep learning. Thus, post-hoc interpretability is still opted for in situations where only a complex model can perform adequately for a given problem [3], [38], [82], [83].

However, post-hoc interpretability's main limitation lies in the manner in which it generates explanation. Since post-hoc only relies on input-output data, it is somewhat limited in what it can provide in terms of a *deep* explanation. Therefore, although post-hoc is more widely applicable, this key obstacle holds the methodology from being adopted as a dependable XAI solution. For this reason, a group of researchers urge ML designers to opt for interpretable models as a first resort for simpler problems [2], [81]. On the other hand, opposers of interpretable models argue that the performance trade-off is not worthwhile [77]–[80]. However, Rudin et al. stresses that it is not the interpretability that is responsible for the drop in performance [2]. Ruden highlights an important argument by pointing out that 'the inferiority of interpretable models is not necessarily caused by their interpretability'. The ultimate aim of explainability was always to provide a useful explanation, while avoiding significant trade-off of performance. Opting for post-hoc means attempting to improve the usefulness of the explanation while being limited by the source of the interpretation; meanwhile, in model-based useful explanation often appears more naturally while more effort has to be put in to achieve satisfactory performance.

The current state-of-the-art AI solutions are predominantly using opaque models. However, recent regulatory guidelines by the European Commission, UK government and DARPA [70], [71], [84] have prompted a growing interest in explainable-AI (XAI). The guidelines do not specify the methodology by which explainability has to be achieved and, have left it to researchers to determine. Nonetheless, the recommendations presented somewhat of a vision of what is to be expected ultimately from XAI.

The focus on XAI was motivated by the limitations observed in conventional non-explainable AI. Despite the promising performance achieved by prominent ML learning techniques, the lack of explainability prevented wider adoption. An explanation is not a functional requirement for low consequence applications such as, facial recognition in photography and, active noise cancellation in headphones. Nevertheless, explanation could be useful as an non-functional requirement for designers and users. Serious consequence problems, on the other hand, are subject to stringent protocols and regulations, which require a clear *justification* for all decisions; this makes *explanation* a functional requirement for critical applications such as medical, finance and, oil & gas.

4.2.2 The psychology of explainability

The sole purpose of XAI is presenting explanation to human users. The manner in which the *explanatory* information is received by the user depends on a host of variables such as level of expertise and intellectual ability. For example, in computer science, when designing a piece of software to be used by different entities; each entity has a different combination of duties, knowledge and experience. Therefore, the software's interface must be catered to the wide variety of requirements. Similarly, XAI as a proposition can only be truely successful if all its users are able to benefit; this is only possible when the explanation is tailored for each of the different parties [19]. The parties can include: the different users, the model designer and model manager. In some cases, the model designer and manager could be the same party.

Furthermore, an explanation is a form of language and, as cliché as it sounds, language can be subjective. This bias of subjectivity in language means although a piece of explanation is considered *comprehensible* to one human, it does not guarantee it is for any other [59]. Comprehensibility refers to the desired state where a model can be explained in a way a human can understand, i.e., grasp [19]. Comprehension in psychology is defined as the 'act of or capacity for grasping with the *intellect*' [85]. It highlights the importance of intellect in understanding explanation.

For instance, if a decision support system was designed to diagnosis pneumonia using X-ray imaging, the stakeholders involved include the patient, radiologist, clinician and model designer. Clinicians and radiologists are both tasked with interpreting image data, however, their distinct areas of expertise could affect how understandable the explanation produced by the decision support decision is. If this system was subject to GDPR law, the model's explanation would have to be understandable by the patient, i.e., the user's right to meaningful explanation [70]. To explain a model's result to an expert clinician is one thing. However, to provide the same to the average patient is a more immense challenge clinicians occasionally struggle with despite being experts in their field [86], [87].

4.2.3 Explanation methodology

Explanation frameworks can be constructed using a variety of different methodologies depending on the nature of the model to be explained. As described in the previous section, XAI can be either post-hoc or model-based, as shown in Figure 4.2. Literature on post-hoc XAI is more prominent with a higher number compared to model-based [19]. Moreover, the frameworks are classified into various categories depending on common attributes such as abstract level (local or global) and methodology (decision trees, rule-based). Post-hoc methods utilise a form of post-processing to *predict* why a model decided a certain way. With the exception of model-agnostic methods, the methodologies vary widely depending on the modelling technique they are tailored for.

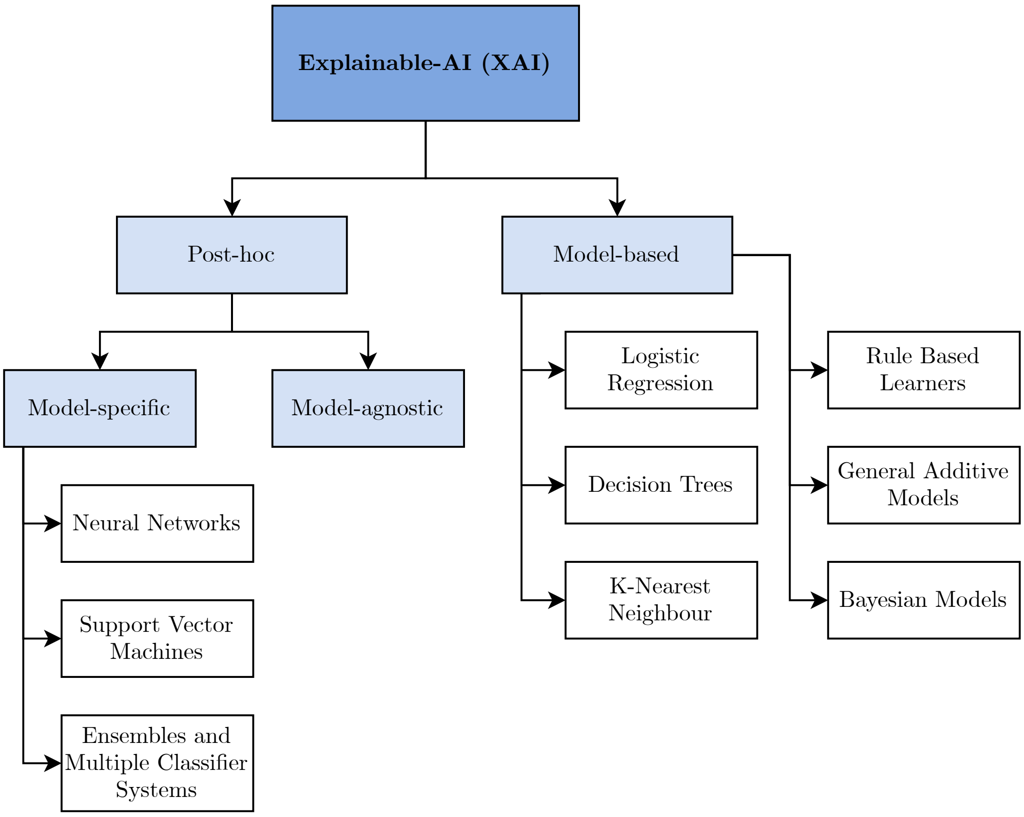


Figure 4.2: Flowchart representing the different types of explanation. It is a simplified version based on a detailed flowchart by Barredo Arrieta et al. [19].

In spite of the popularity of post-hoc interpretability, there exist several attempts at model-based explainability [17], [77], [88]–[90]. Since model-based provides direct explanation based on the model's internal parameters, the insight has the potential

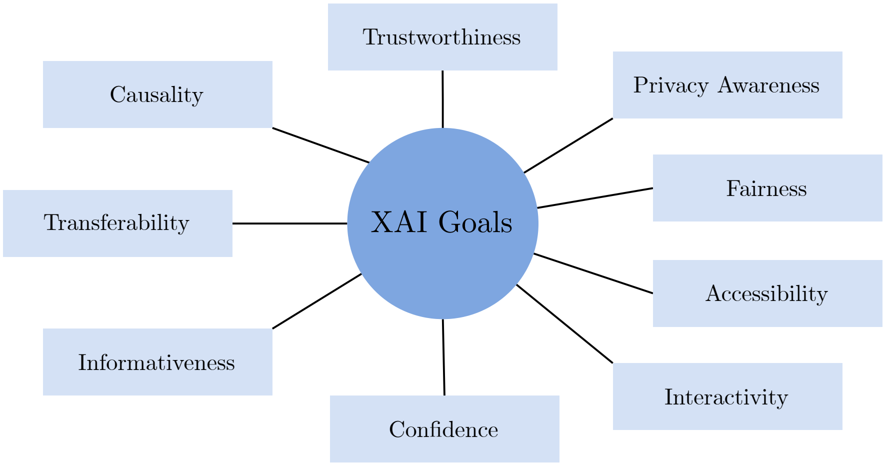


Figure 4.3: XAI goals research areas based on Barredo Arrieta et al. literature in [19].

to reflect precisely how a model decided a certain including details such as which features, rules and parameters were the root or contributing cause. Conversely, post-hoc predicted explanations are restricted by their limited source of interpretation, i.e., an opaque model's input-output data.

The Cambridge Dictionary definition of explanation is as follows: "the details or reasons that someone gives to make something clear or easy to understand" [91]. Although the dictionary definition is universal, there is a lack of agreement in the research community regarding the definition, aims and rationale for *explainable-AI* [19]. Barredo Arrieta et al. base the definition of XAI on the dictionary definition for the word *explanation* as below:

> “Given an audience, an **explainable Artificial Intelligence** produces details or reasons to make its functioning clear or easy to understand.”

In contrast, Gunning's definition emphasises additional aims of XAI, such as building trust and managing the development of AI, as below and in [92].

> “**XAI** will create a suite of machine learning techniques that enables human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners.”

Similarly, the goals of XAI are diverse (see Figure 4.3), including research areas such as trustworthiness, confidence and accessibility [19]. The set of goals is designed to tackle problems arising in the use of conventional opaque AI. The relative importance of each goal depends on the application. For instance, more priority would be given to *fairness* in a law-related application. Moreover, any problem involving users is likely to prioritise *privacy awareness*.

Explanation, when generated, can take many forms, including but not limited to textual, graphical and audial. The *medium* selected would depend on the nature of the underlying explanation. If the information presented to the user involves complex ideas and concepts, then textual would be the optimal choice. Meanwhile, numerical data could be understood better when represented graphically.

As a modelling technique, FL is considered a prominent research area for XAI with around one third of literature [33]. It is inherently interpretable when used in the Mamdani-type form. At a lower level, the model's structure can be constructed to ensure algorithmic transparency, decomposability and simulatability - interpretability's main components as proposed by Lipton [12]. Hence, FL is considered a viable candidate for model-based interpretability and explainability [58], [61], [93].

However, there has been comparatively less interest in utilising Decision Theory knowledge for explainable AI [94]. Decision Theory as defined by Oxford Dictionary as "the mathematical study of strategies for optimal decision-making between options involving different risks or expectations of gain or loss depending on the outcome". However, decision theory can be used solely as a modelling technique. A subset of Decision Theory, Multi-Criteria Decision Making (MCDM), is a widely researched methodology [95]–[97]. The nature of MCDM means it is highly interpretable as it is aimed at augmenting a human's decision making process. Nevertheless, its simple nature has long limited its applicability to data-driven applications until recently [14], [36]. It is evident that MCDM satisfies the requirements of interpretability; hence, if enhanced adequately, it could be the basis for an effective data-driven explainable framework.

4.3 Proposing an explanation framework for MCDM-based classifiers

In this section, an explanation framework is proposed for MCDM-based classifiers for the first time. The framework utilises model-based information to generate textual and graphical explanations. The former provides a series of three textual statements for explaining: a. the negative class model, b. the positive class model and c. the overall model. Meanwhile, graphical explanations provide a means for visualising the impact of the individual features on a *negative* or *positive* classification while, illustrating key model parameters.

4.3.1 Interpretable structure of MCDM-based classifiers

Multi-criteria decision making methodologies utilise various transparent structures to make decisions - enabling inherent *model-based* interpretability. In contrast to post-hoc, model-based allows for direct access to resourceful interpretable information and, in turn, more meaningful explanation. Moreover, a MCDM model uses a set of criteria linked with respective weights to signify impact on a *rank*. A rank is a continuous number used to represent how *good* or *bad* an alternative is. When a set of choices need to be compared, their ranks are computed and sorted to determine the best, in descending order. This was how MCDM was utilised in deciding between several alternatives based on a set of criteria. Nonetheless, a different approach was needed for classification where the different alternatives are *classes* [14]. This was achieved by employing sub-MCDM models, each representing a different class [14].

Fuzzy-MCDM's structure has several sources of interpretation, for which there are two categories: model training and execution. While the former uses information from model construction for interpretation, the latter uses data from the execution process. Interpretation is perceived to be useful for both the model designer and user.

The requirement for a *directly* explainable model is its model-based interpretability. As described in previous chapters, direct explainability has a better chance of providing more representative explanation [3].

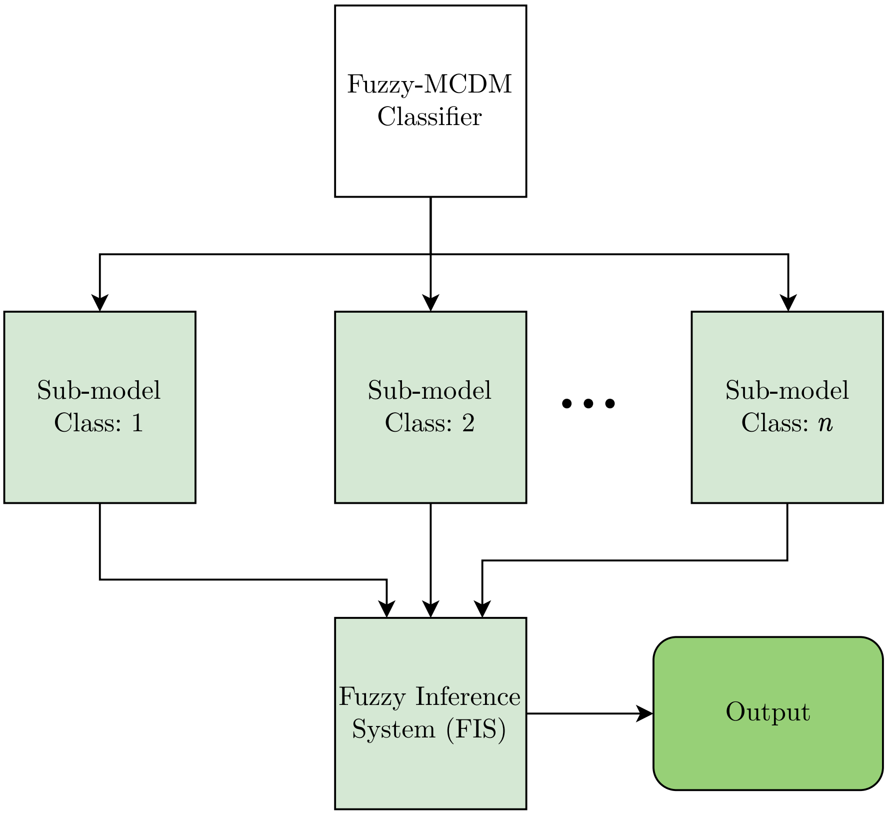


Figure 4.4: High-level abstract structure of the MCDM-classifier

Interpretable components include the model's input(s), internal structure, construction procedure and, output(s). Lipton, in [12], suggested several principles for achieving model-based interpretability defined as *simulatability*, *decomposability* and *algorithmic transparency*.

As described in detail in Chapter 3, a Fuzzy-MCDM classifier structure consists of various transparent parameters such as the weights, ideal solutions and distance measures (Figures 4.4-4.5). Moreover, the model construction procedure uses the Pearson correlation coefficient along with simple arithmetic operations to determine the ideal solutions - the only parameters required to run the model. Hence, the relative simplicity of the model's structure and construction makes Fuzzy-MCDM a promising candidate for an explainable classification framework.

4.3.2 Decision trees for explanation

A method for explaining MCDM classifiers using decision trees is introduced in this section. An explanation can be extracted by a variety of different methods. The

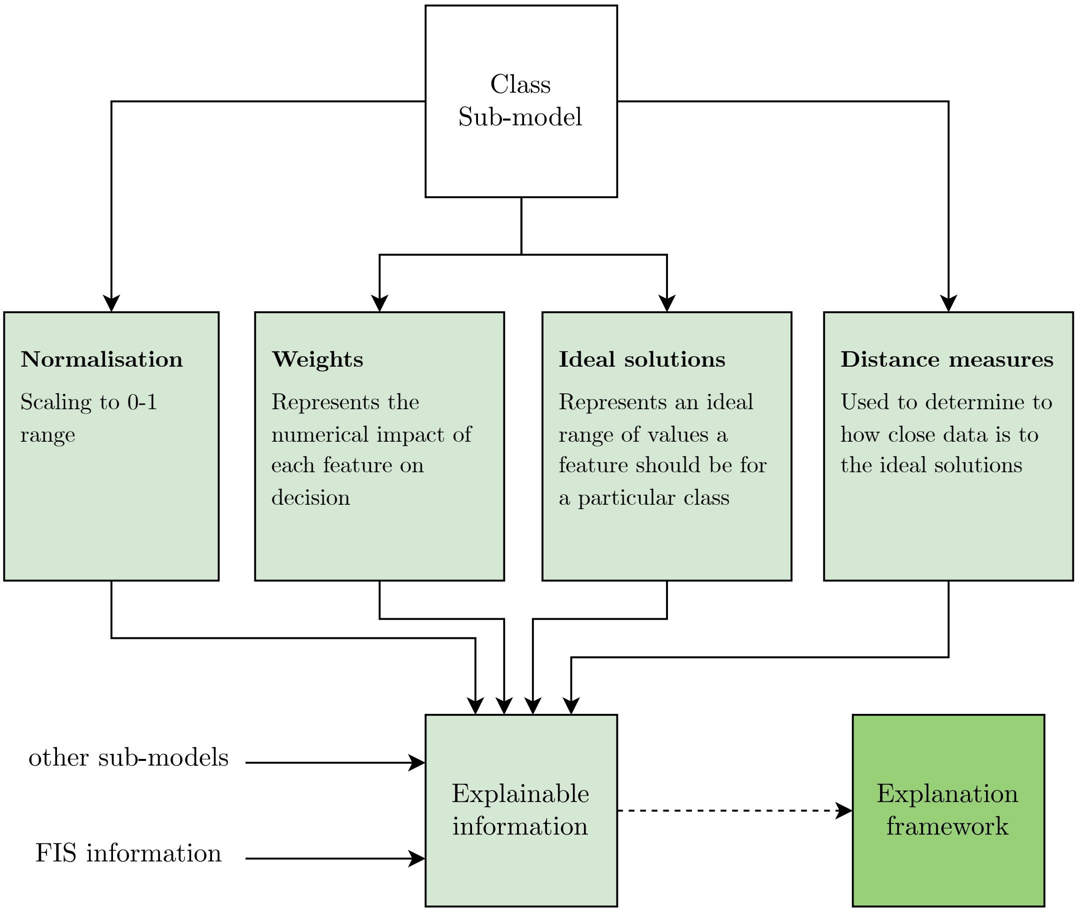

Class
Sub-model
Normalisation
Scaling to 0-1 range
Weights
Represents the numerical impact of each feature on decision
Ideal solutions
Represents an ideal range of values a feature should be for a particular class
Distance measures
Used to determine to how close data is to the ideal solutions
other sub-models
FIS information
Explainable information
Explanation framework

Figure 4.5: Diagram visualising the structure of the MCDM-classifier sub-model

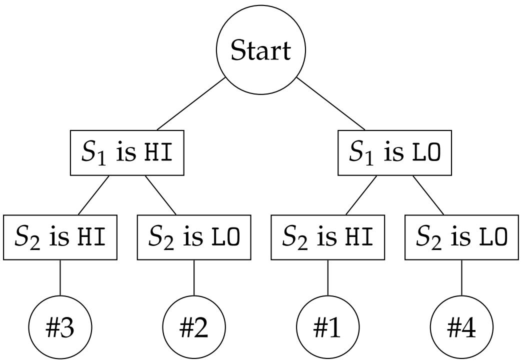


Figure 4.6: Decision tree for selecting the suitable explanatory linguistic statement

Table 4.1: Linguistic sentence templates used for explaining the classifier's sub-models

#	Sentence template
1	`<sub-model>` thinks the data is similar to `<c1>` and NOT similar to `<c2>`.
2	`<sub-model>` thinks the data is similar to `<c2>` and NOT similar to `<c1>`.
3	`<sub-model>` thinks the data is more similar to `<similar_class>` despite a low similarity for both.
4	`<sub-model>` thinks the data is more similar to the `<similar_class>` despite a high similarity for both.

method used depends on the nature of the source and, how the data will be presented. In the case of Fuzzy-MCDM classifiers, decision trees are applicable for retrieving meaningful explanations about *how* the sub-models, and, in turn, the overall model decided a certain way.

A decision tree was used to choose how to explain a model or sub-model based on pre-designed conditions, as shown in Figure 4.6. The linguistic conditions (HI and LO) represent the regions below or larger than 0.5 - the centre of the normalisation range. Four linguistic sentence templates are defined (Table 4.1). Templates #1 and #2 provide a factual and counterfactual explanation. This is for cases where the data is similar to a class while *not* similar to the opposite class. Meanwhile, templates #3 and #4 are used to explain when the data is similar or non-similar to either of the classes. The templates allow for a sentence to be generated for each sub-model for explaining *why* it decided a certain way.

Table 4.2: Linguistic sentence templates used for explaining the classifier's the overall model

#	Sentence template
1	Models are in agreement, hence the data was predicted to be `<similar_class>`
2	Models are in conflict however, the measures pointed towards a better similarity towards `<similar_class>`

Moreover, a final statement is generated to explain the overall model's decision based on the conclusions presented from the sub-models (Table 4.2). The sentence refers to the sub-models classifications by explaining whether they are in agreement or conflict. The agreement or conflict indicates the certainty.

The statements provide the user with more summarised information about the two sub-models and, how they impacted the decision. In addition, relevant numerical figures in brackets such as the distance measures and fuzzy class output, are presented in brackets. For the Fuzzy-TOPSIS classifier, a single measure S is used thus, the quantity of interpretable information is less compared to Fuzzy-ATOVIC or Fuzzy-VIKOR. This allows the cognition-space to be utilised for explaining the input(s) impact on the output, as will be demonstrated in the next section.

4.3.3 Visualisation of input impact on output

The previous section presented how linguistic statements were used to explain the outcomes of the different sub-models and overall models. In this section, a method for visualising the inputs' impact on the sub-models outcomes is introduced. The graphical explanation is generated by calculating a *score* for each feature for indicating its impact on the sub-models outcome.

$$F_{jc} = 5\left(\frac{\theta^{t} - f_{jc}^{-}}{f_{jc}^{+} - f_{jc}^{-}}\right) \tag{4.1}$$

where F_{jc} is the feature score for feature j and class c; θ_j is the normalised feature j; f_{jc}^{-} and f_{jc}^{+} are the negative and positive ideal solutions respectively. The feature score is calculated (Equation (4.1)) as a distance from the negative ideal solution, which is inversely correlated with its impact on value of the measure and in turn the outcome.

A set of scores is computed for each sub-model, representing the level of impact each feature has on the positive ideal outcome for that particular model. The score is scaled to a range of five for better readability and comparison.

Although the above method provides the user with visualisation of how close each feature is to the ideal solutions, it does so without reference to the feature weights - a key variable affecting impact.

In order to provide the user with information about the true impact a feature has on the outcome, the feature weights w, as defined in Equation (3.3), were introduced to the feature score calculation.

$$F_{j_c} = 5 \cdot w_j \left(\frac{\theta^t - f^-_{j_c}}{f^+_{j_c} - f^-_{j_c}} \right) \tag{4.2}$$

where w_j is the weight for feature j.

Consequently, the features can be plotted to aid the user's visualisation of the impact of different inputs. A horizontal bar graph is the most suitable for comparing the feature scores from several sub-models, as shown in Figure 4.7.

$$C_c = \sum_{j=a}^{b} \left[w_j \left(\frac{\theta^t - f^-_{j_c}}{f^+_{j_c} - f^-_{j_c}} \right) \right] \tag{4.3}$$

where C_c is the category impact for class c, a is the starting feature and b is the ending feature to be summed.

For datasets with a large number of features, simply plotting the feature scores would make the plots unreadable. Therefore, the features were grouped into categories by relevance to reduce the number of bars. The features are aggregated

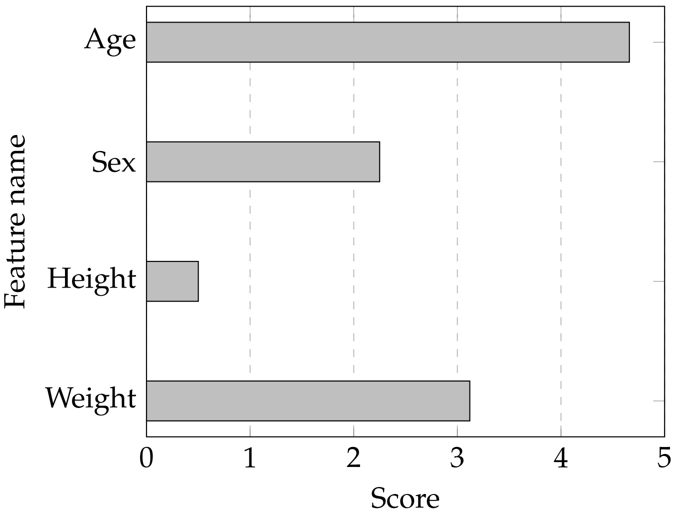


Figure 4.7: Example of how a feature score graphical explanation would look like in the form of a horizontal bar graph

by summation using Equation (4.3), but excluding the factor of five scaling. Similarly, they are plotted in a bar graph, as shown in Figure 4.7.

4.4 Assessing an explanation framework for MCDM-based classifiers

In this section, the proposed explanation framework is applied to five benchmark datasets. The explanation framework was shown to be applicable to a certain class of problems where it provided valuable precise information on which features led to a certain decision. Explanation is useful for applications where *justification* and *traceability* are prominent functional requirements.

4.4.1 Explaining human-understandable datasets

To assess the merits of the explanatory framework, a set of three datasets were used that consisted of human-understandable features. The set includes a Breast Cancer dataset and two Titanic passenger datasets from the KEEL and Kaggle repository; these datasets are the same ones used in assessing MCDM-based classifiers in the previous Chapter 3. The structure and characteristics of the datasets were described in detail in Section 3.4.1.

As described, the ultimate aim of explanation is to provide valuable insight into how the decision was made. In this part, we provide examples from the different case studies on how explanation can provide *indicative* insight into the decision. The insight attempts to explain how a decision was arrived at and, indicate decision certainty.

Breast cancer dataset

The Breast Cancer dataset offers an example where Fuzzy-ATOVIC performs well from an accuracy perspective hence, providing a glimpse of the best-case scenario for the explanation framework. As presented in Chapter 3, Fuzzy-ATOVIC achieved an overall accuracy of 93.2%. In this section, several examples of graphical and textual explanations are presented. Where an incomplete set of figures is shown for the graphical explanation examples, a full set is provided in Appendix A.

A pair of bar plots highlight graphically the level of influence each feature has on a certain classification result for each sub-model. The names of the features are listed

Table 4.3: Breast Cancer dataset: feature names

#	Name
1	Clump Thickness
2	Uniformity of Cell Size
3	Uniformity of Cell Shape
4	Marginal Adhesion
5	Single Epithelial Cell Size
6	Bare Nuclei
7	Bland Chromatin
8	Normal Nucleoli
9	Mitoses

in Table 4.3. A single bar graph illustrates the values of the measure S - the final decider used for the classification process.

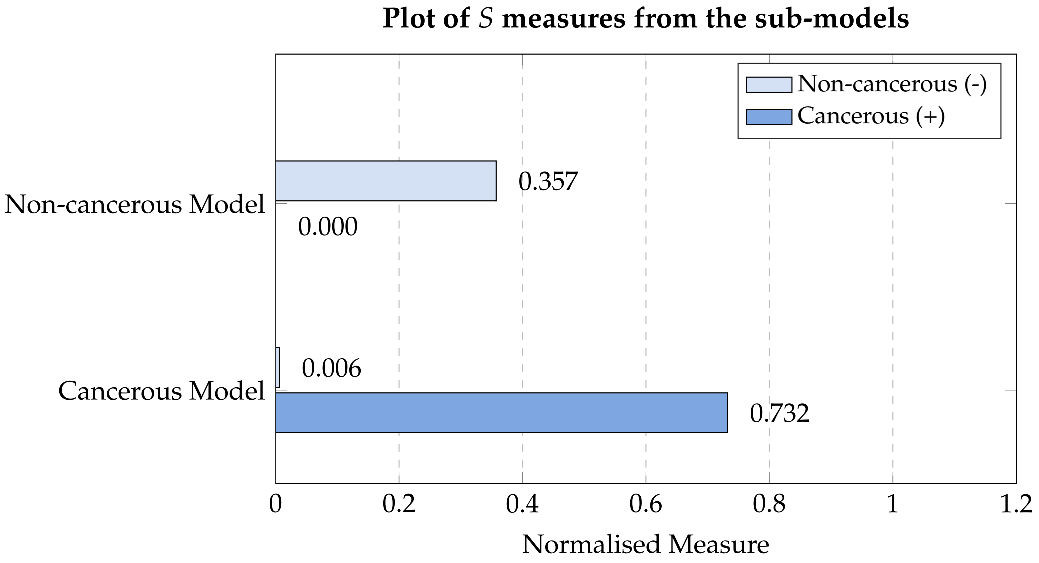


Figure 4.8: Breast cancer dataset: distance measures S from the two sub-models - example of a FN case: #1_3_098. Textual explanation: “Models are in conflict hence, it is less likely the classification is accurate." Fuzzy class: 0.36.

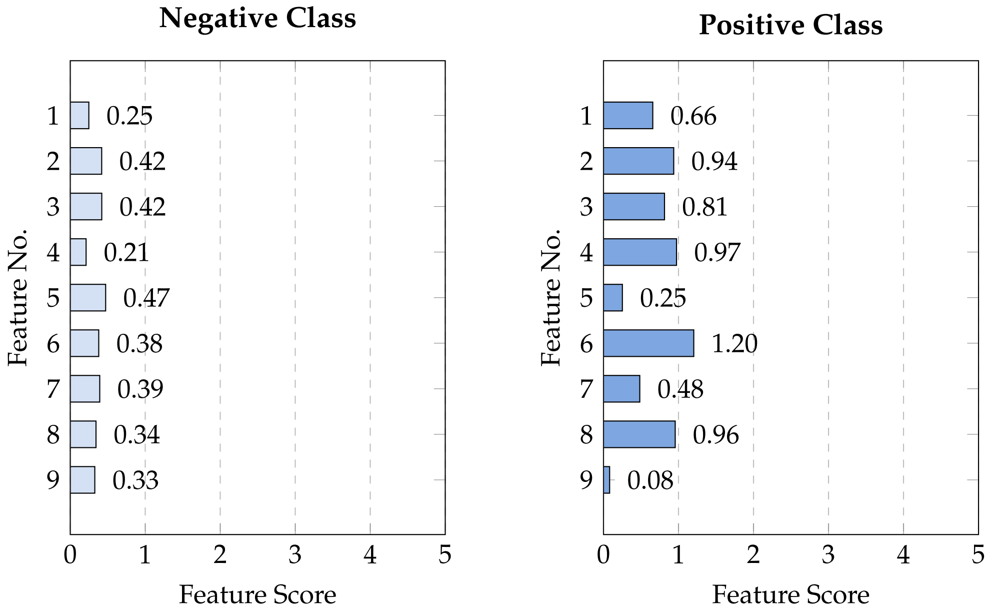


Figure 4.9: Breast cancer dataset: graphical explanation of the *non-cancerous* sub-model - example of a FN case: #1_3_98. Textual explanation: "Non-cancerous model thinks the data is more similar to Malignant (+ve) despite a high similarity for both."

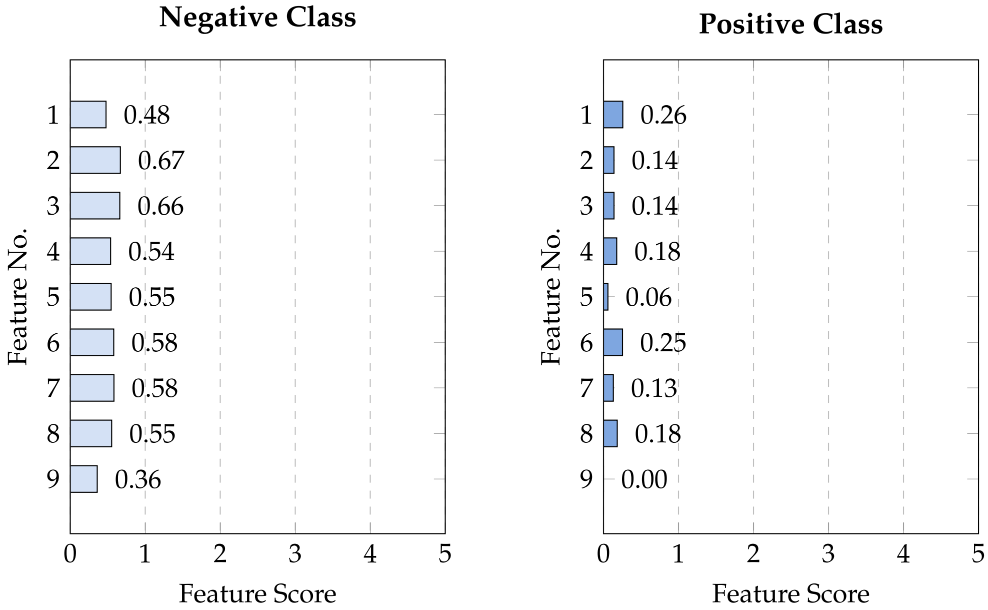


Figure 4.10: Breast cancer dataset: graphical explanation of the *cancerous* sub-model - example of a FN case: #1_3_098. Textual explanation: "Cancerous model thinks the data is similar to Benign (-ve) and NOT similar to Malignant (+ve)"

In the first example for the Breast Cancer dataset, a FN case (Figures 4.8-4.10), the graphical explanation illustrates the data being more similar to a *cancerous* classification for the *non-cancerous* sub-model; this is clear because of the higher feature scores captured, as seen in Figure 4.9. The caption of each figure includes the textual explanation associated with the data presented. Moreover, graphical explanation for the *cancerous* points the other direction with the model showing a similarity to *non-cancerous* classification and ruled out a *cancerous* decision (Figure 4.10). In this case, the overall explanation indicated that the 'models are in conflict' (in Figure 4.8) - a credible hint towards a false classification. Figure 4.8 presents the values of S for the two sub-models cancerous and non-cancerous. When the measure S is low it indicates a high similarity for that class.

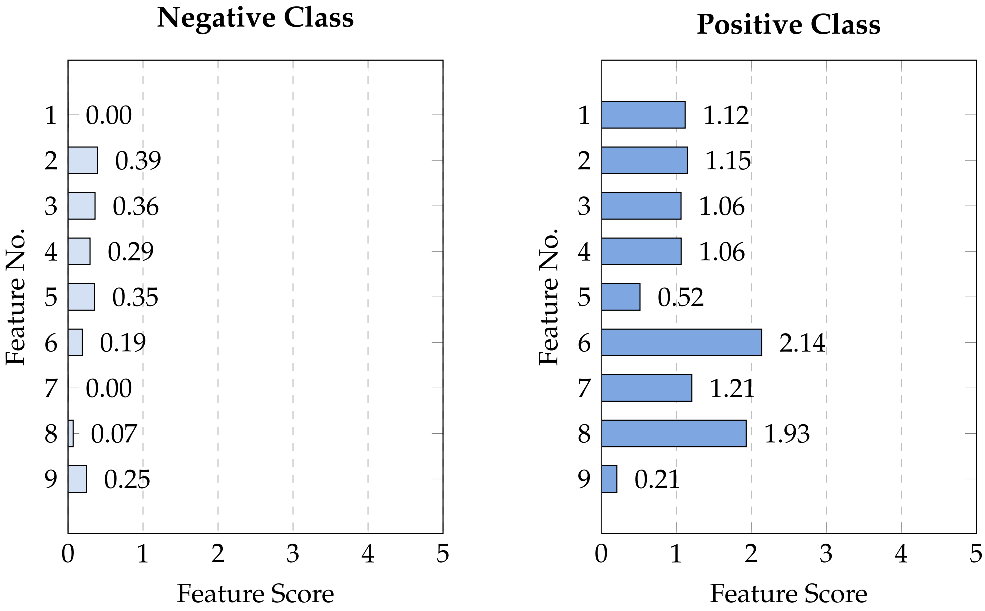


Figure 4.11: Breast cancer dataset: graphical explanation of the *non-cancerous* sub-model - example of a FP case: #1_1_21. Textual explanation: "Non-cancerous model thinks the data is similar to Malignant (+ve) and NOT similar to Benign (-ve)."

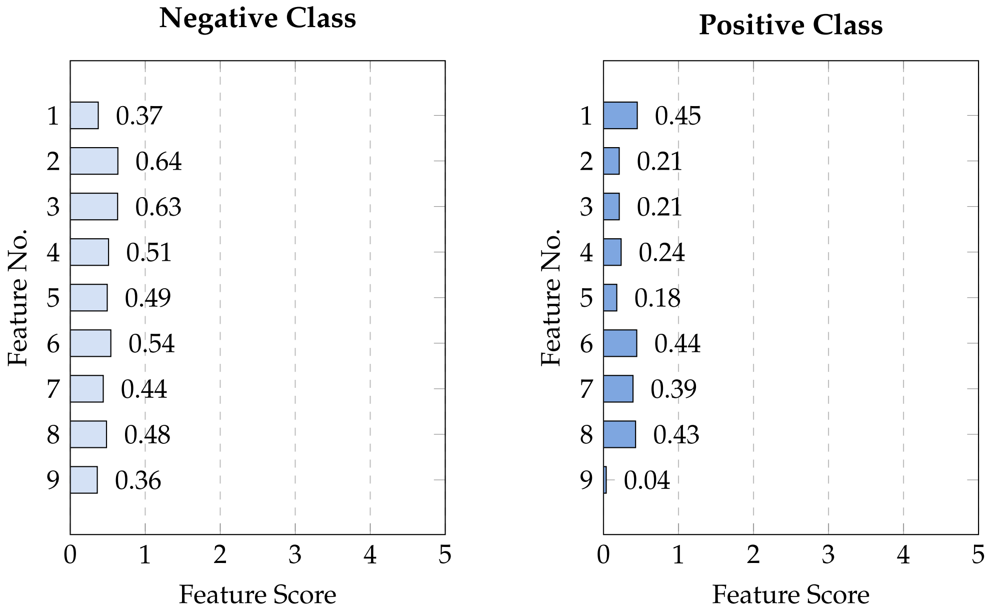


Figure 4.12: Breast cancer dataset: graphical explanation of the *cancerous* sub-model - example of a FP case: #1_1_21. Textual explanation: “Cancerous model thinks the data is more similar to the Benign (-ve) despite a high similarity for both."

Similarly, for the second example, there was a *conflict* between the sub-models; the result was a false positive classification. One of the reasons for the positive classification is the strong similarity to the positive class captured by the non-cancerous sub-model, as shown in Figure 4.11. On the other hand, the cancerous sub-model had a *weaker* negative classification because of a high similarity for both classes (Figure 4.12). In medical diagnosis, the trade-off of specificity for higher sensitivity is desirable; However, insight pointing clinicians towards a potential false classification is welcome as it can help prevent the potential risks of wrongful diagnosis leading to improper treatment.

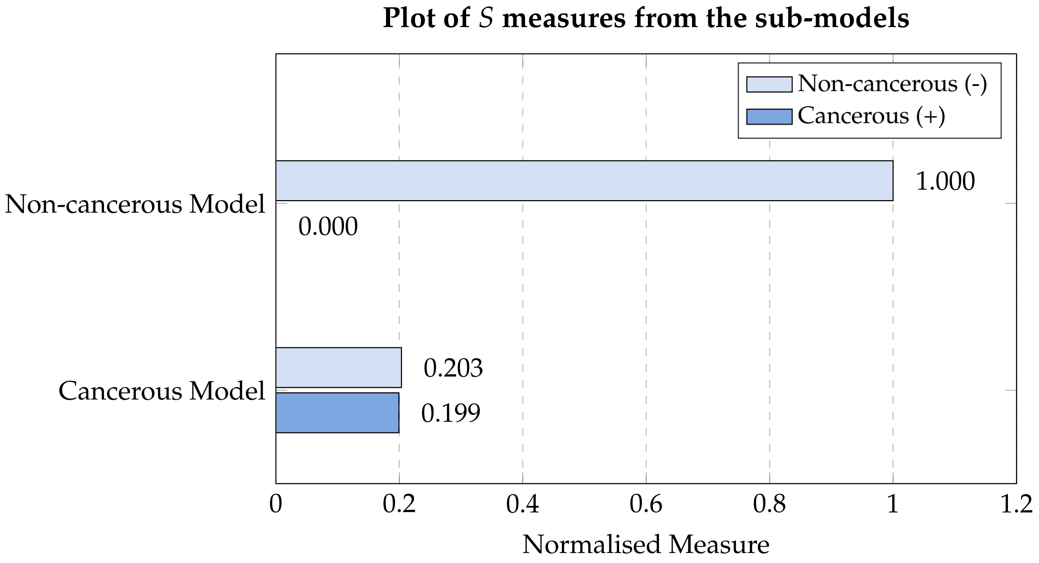


Figure 4.13: Breast cancer dataset: distance measures S from the two sub-models - example of a TP case: #1_3_93. Textual explanation: “Models are in agreement hence, it is more likely the classification is accurate." Fuzzy class: 0.80.

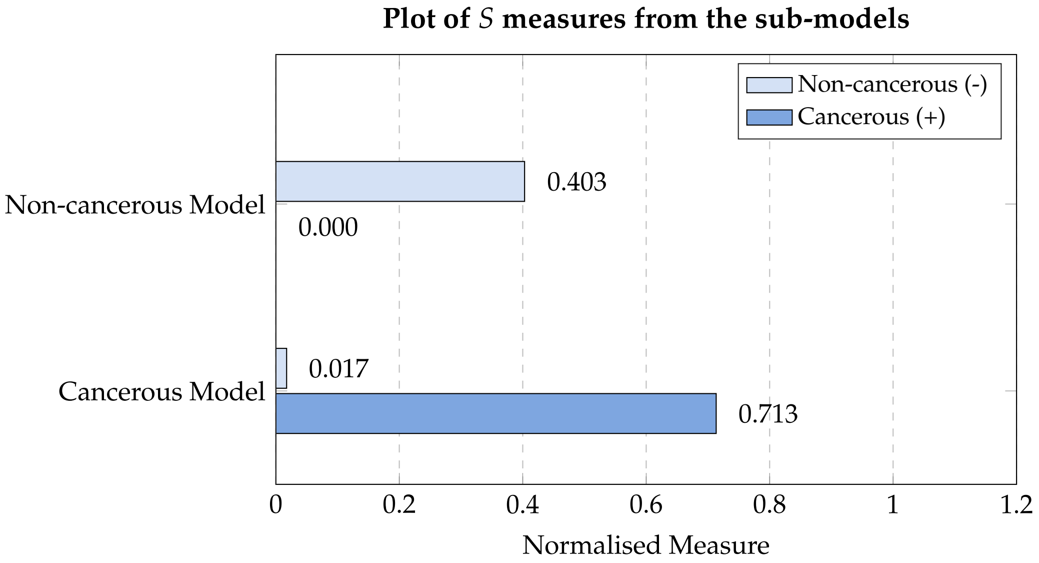


Figure 4.14: Breast cancer dataset: distance measures S from the two sub-models - example of a TN case: #1_3_63. Textual explanation: “Models are in conflict hence, it is less likely the classification is accurate." Fuzzy class: 0.39.

The last two examples demonstrate the model explaining true cases. For the TP case (Figure 4.13), the sub-models were in *agreement* where both pointed to a positive classification. In contrast, the sub-models were in conflict for the TN example (Figures 4.14); this shows that despite an accurate classification, explanation is not always indicative of this.

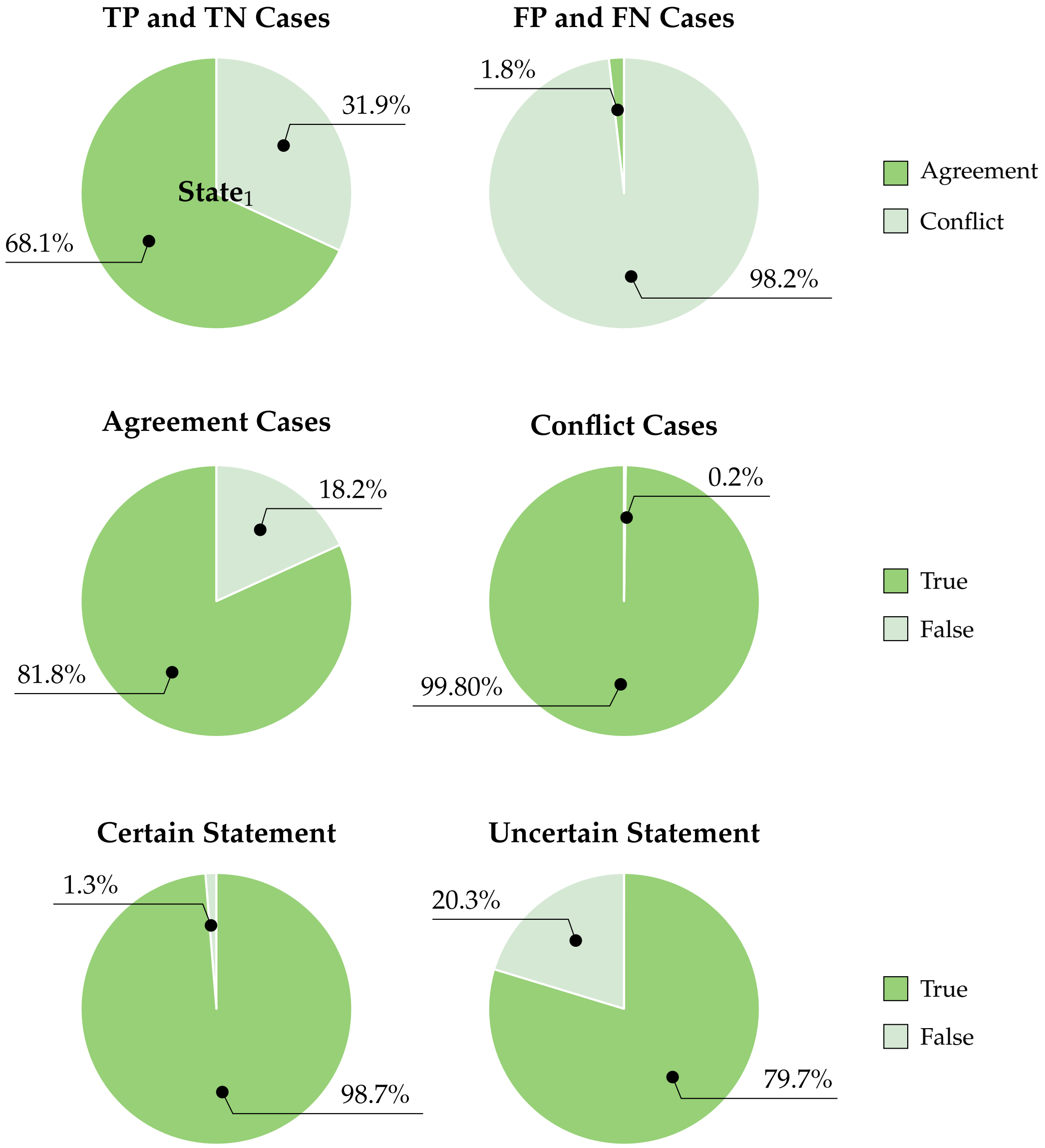


Figure 4.15: Breast cancer dataset: how *indicative* the different aspects of explanation are to a false or negative cases. Certain statements refer to sentence templates #1 and #2 while, uncertain statements refer to sentence templates #3 and #4.

To assess how indicative the explanation generated is, distribution information was gathered (in Figure 4.15) on different aspects related to result in accuracy. For the Breast Cancer dataset, true cases were more likely to result in an agreement between the two sub-models, while false cases were more likely to result in a conflict. Analysis of how the linguistic statements relate to accuracy revealed that the use of less certain statements (#3 and #4) was linked to a higher occurrence of false cases.

KEEL titanic dataset

Moreover, two Titanic datasets are explored for estimating whether a passenger is a *survivor* or *casualty*. The datasets are the same ones used in Chapter 3 to assess MCDM-based classifiers. The datasets vary in terms of the number of features utilised. The Titanic KEEL dataset only uses three features which allow for simpler graphical explanation, visualisation, and comprehension.

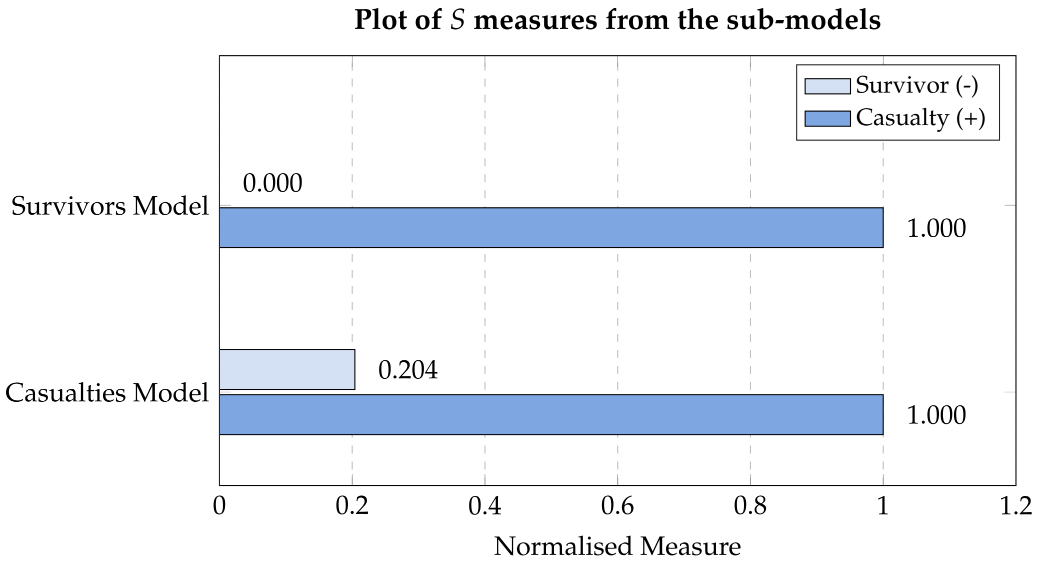


Figure 4.16: KEEL titanic dataset: distance measures S from the two sub-models - example of a FN case: #1_1_301. Textual explanation: “Models are in agreement hence, it is more likely the classification is accurate." Fuzzy class: 0.23.

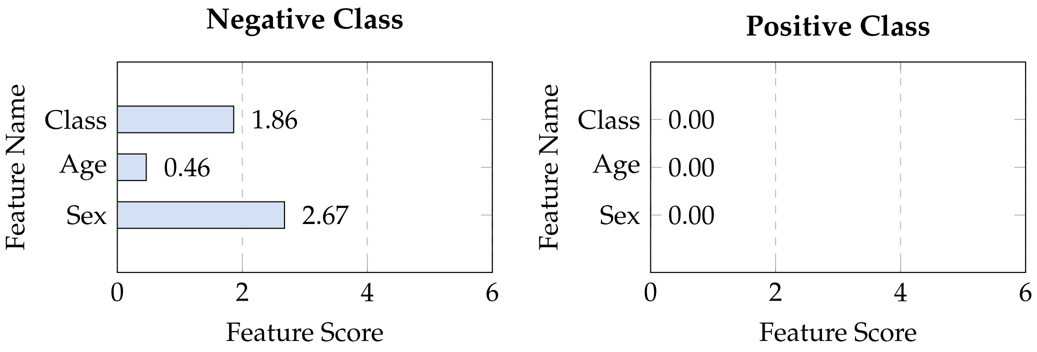


Figure 4.17: KEEL titanic dataset: graphical explanation of the *survivor* sub-model - example of a FN case: #1_1_301. Textual explanation: “Survivors model thinks the data is similar to Survivor (-ve) and NOT similar to Casualty (+ve)."

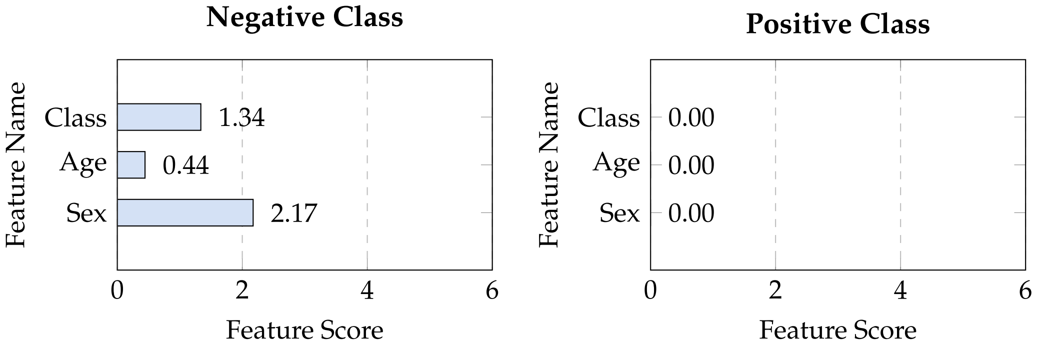


Figure 4.18: KEEL titanic dataset: graphical explanation of the *casualty* sub-model - example of a FN case: #1_1_301. Textual explanation: “Casualties model thinks the data is similar to Survivor (-ve) and NOT similar to Casualty (+ve)"

Contrary to the FN example shown for the Breast Cancer dataset, the following example resulted (Figures 4.16-4.18) in an *agreement* between the two sub-models. In this example, it is highlighted how inaccurate classification can lead misleading explanation. On the other hand, it clarifies the explanation’s prime function which is, explaining how a decision was arrived at and, not whether it was a correct or not.

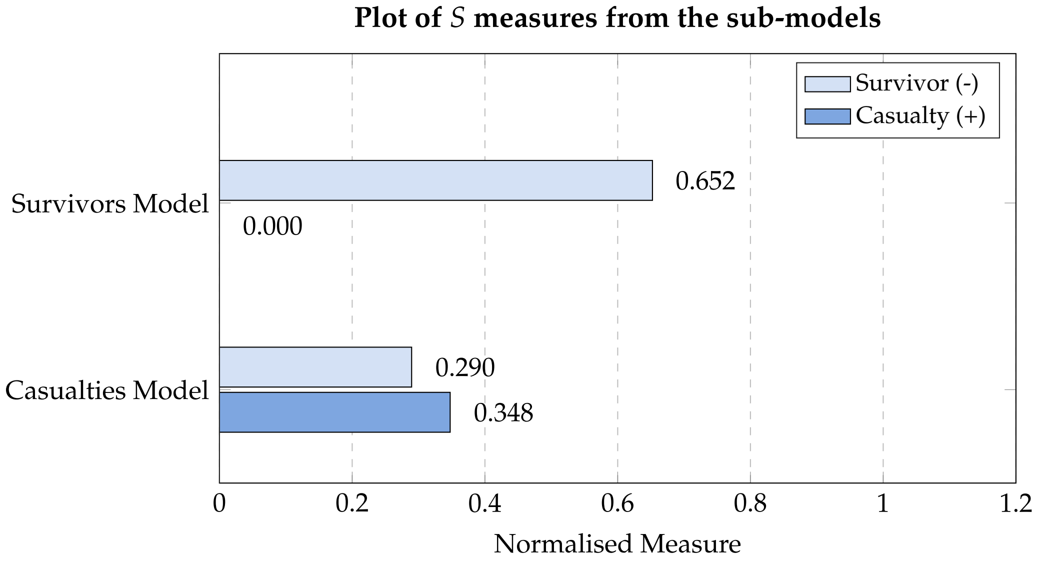


Figure 4.19: KEEL titanic dataset: distance measures S from the two sub-models - example of a FP case: #1_1_75. Textual explanation: “Models are in conflict hence, it is less likely the classification is accurate." Fuzzy class: 0.74.

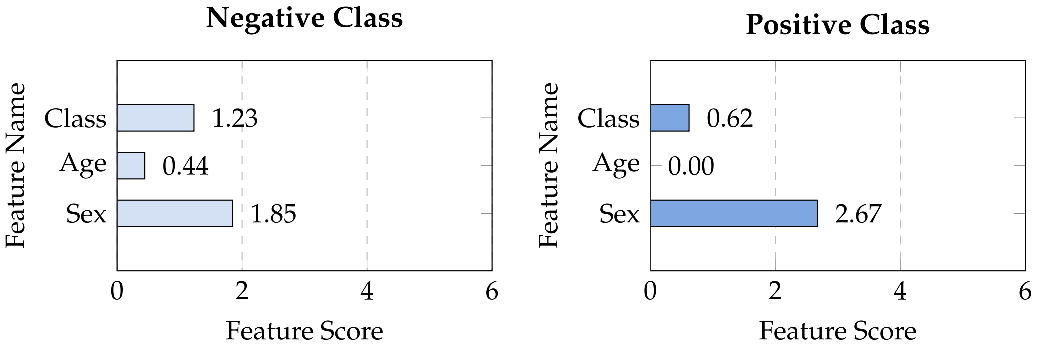


Figure 4.20: KEEL titanic dataset: graphical explanation of the *casualty* sub-model - example of a FP case: #1_1_75. Textual explanation: “Casualties model thinks the data is more similar to the Survivor (-ve) despite a high similarity for both."

Conversely, the next example (Figures 4.19-4.20) shows the sub-models in conflict. The graphical explanation highlights the main contributor for the false classification: the sex feature, which had a maximum score of five (as shown in Figure 4.20). The high impact of the feature caused the survivor sub-model to appear more certain using the criterion: the difference between the two distances. The example demonstrates a case where a single feature was pinpointed as the main cause of the classification going one way rather than the other.

The level of detail provided by the *explanation* allow the user to perceive the decision-making process transparently and in a traceable way. This can facilitate the detection of potential outliers or inconsistencies in the data.

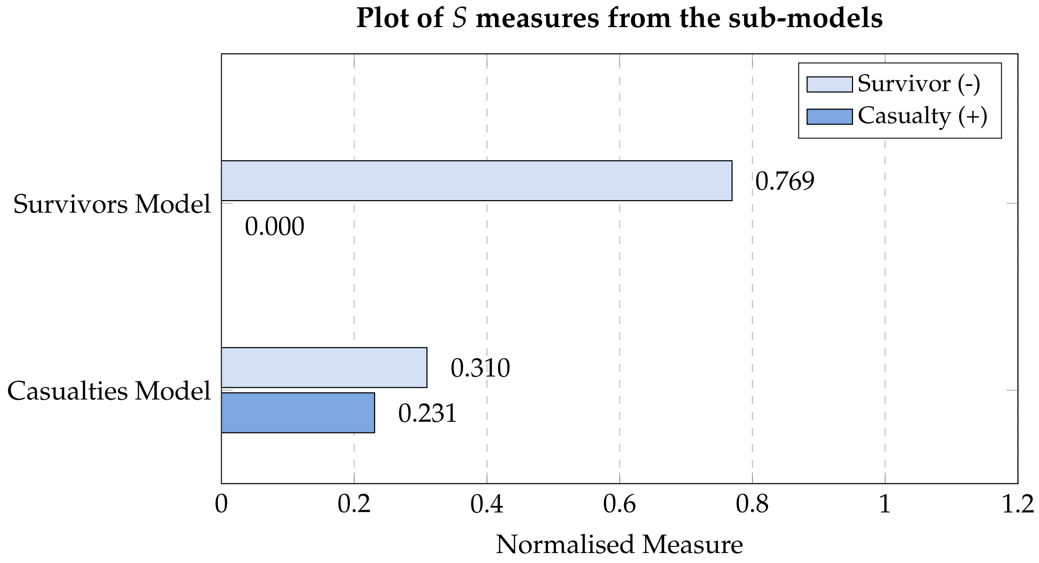


Figure 4.21: Titanic KEEL dataset: distance measures S from the two sub-models - example of a TP case: #1_1_299. Textual explanation: “Models are in agreement hence, it is more likely the classification is accurate." Fuzzy class: 0.79.

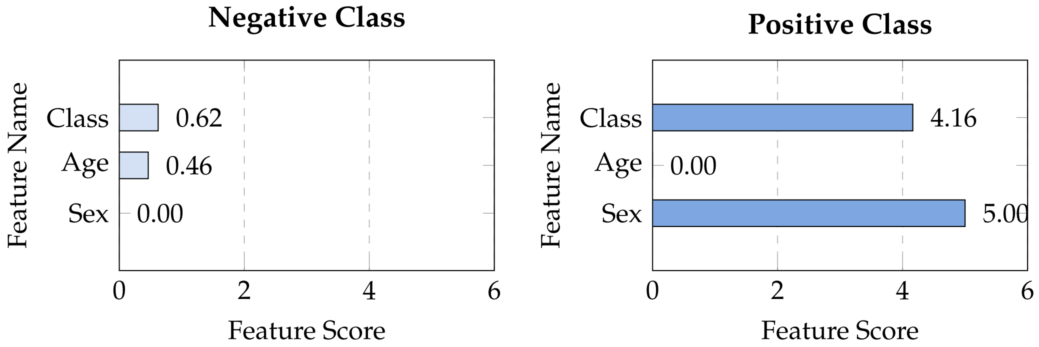


Figure 4.22: Titanic KEEL dataset: graphical explanation of the *survivor* sub-model - example of a TP case: #1_1_299. Textual explanation: “Survivors model thinks the data is similar to Casualty (+ve) and NOT similar to Survivor (-ve)."

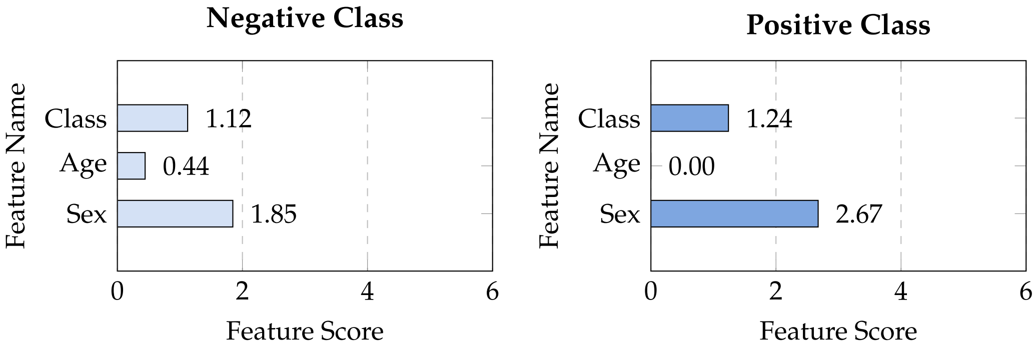


Figure 4.23: Titanic KEEL dataset: graphical explanation of the *casualty* sub-model - example of a TP case: #1_1_299. Textual explanation: "Casualties model thinks the data is more similar to the Casualty (+ve) despite a high similarity for both."

Moreover, the TP example (Figures 4.21-4.23) also demonstrate the varying degrees with which each feature value is influencing a certain decision. For the *survivor* sub-model (Figure 4.22), the Class and Sex attributes play a vital role in the positive classification. Meanwhile, a comparatively weaker decision is made by the *casualty* sub-model (Figure 4.23) due to similar feature scores for the two classes. Furthermore, the distance measures S summarise the interaction between the two models representatively (in Figure 4.21).

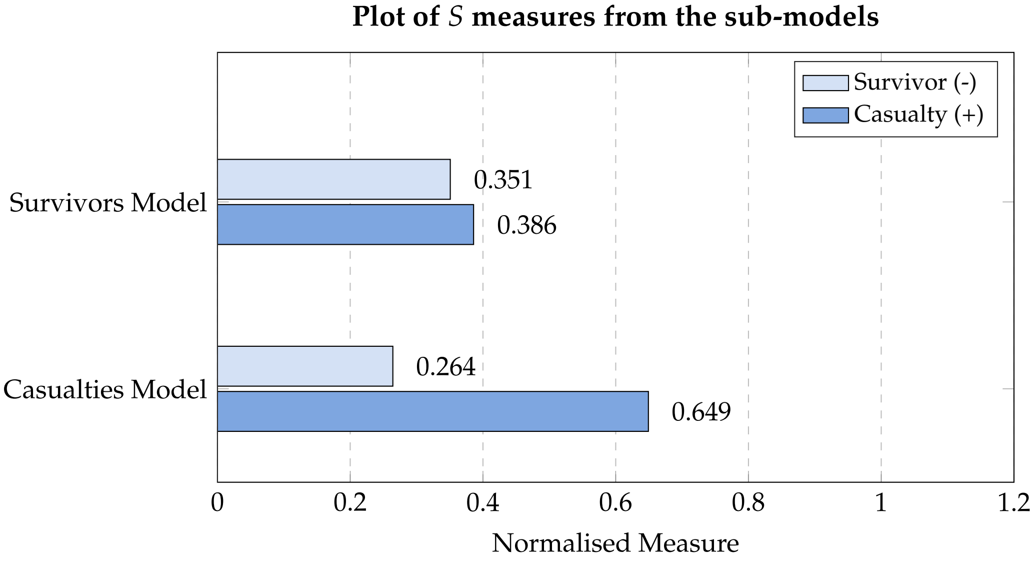


Figure 4.24: Titanic KEEL dataset: distance measures S from the two sub-models - example of a TN case: #1_1_27. Textual explanation: "Models are in agreement hence, it is more likely the classification is accurate." Fuzzy class: 0.37.

In the TP example (Figure 4.24), the sub-models are also in *agreement*. In this case the *casualty* sub-model showed a strong classification compared to the *survivor* sub-model.

Kaggle titanic dataset

Similarly to the previous titanic dataset, the Kaggle Titanic dataset aims to predict whether a passenger is a *survivor* or *casualty* based on a set of human-understandable features. However, contrary to the previous dataset, this dataset has seven features instead of three. The dataset features and structure were explained in detail in Chapter 3.

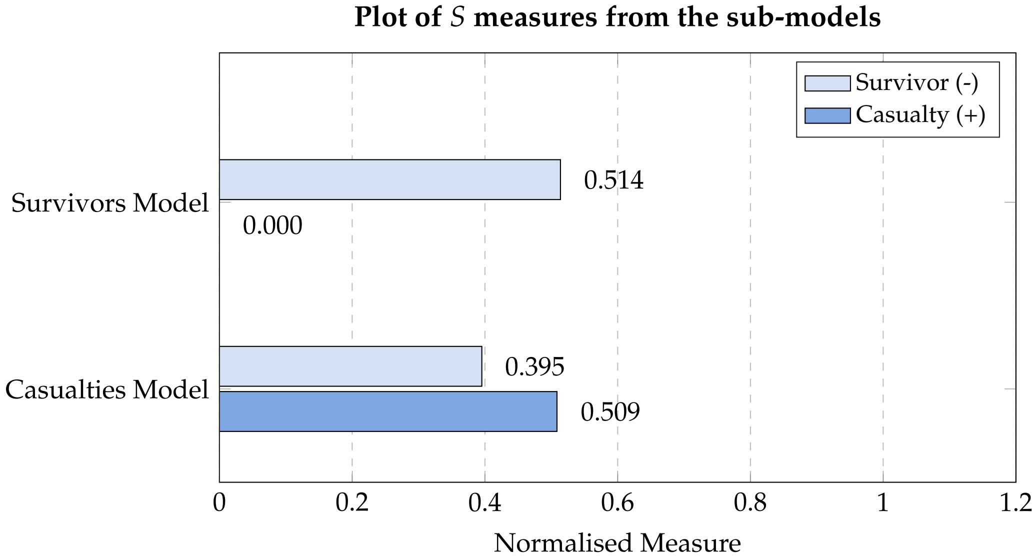


Figure 4.25: Titanic Kaggle dataset: distance measures S from the two sub-models - example of a TP case: #1_1_110. Textual explanation: “Models are in conflict hence, it is less likely the classification is accurate." Fuzzy class: 0.65.

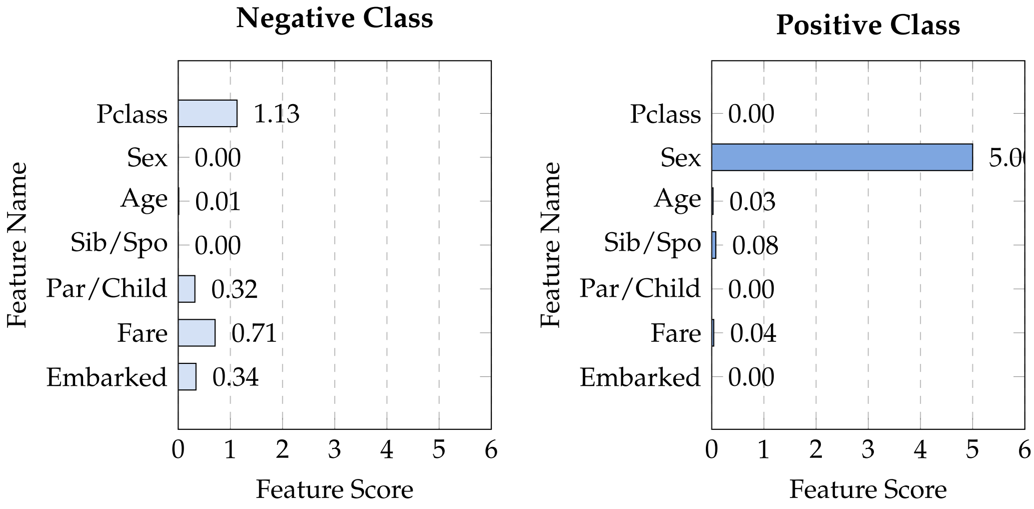


Figure 4.26: Titanic Kaggle dataset: graphical explanation of the *survivor* sub-model - example of a TP case: #1_1_110. Textual explanation: “Survivors model thinks the data is similar to Casualty (+ve) and NOT similar to Survivor (-ve)."

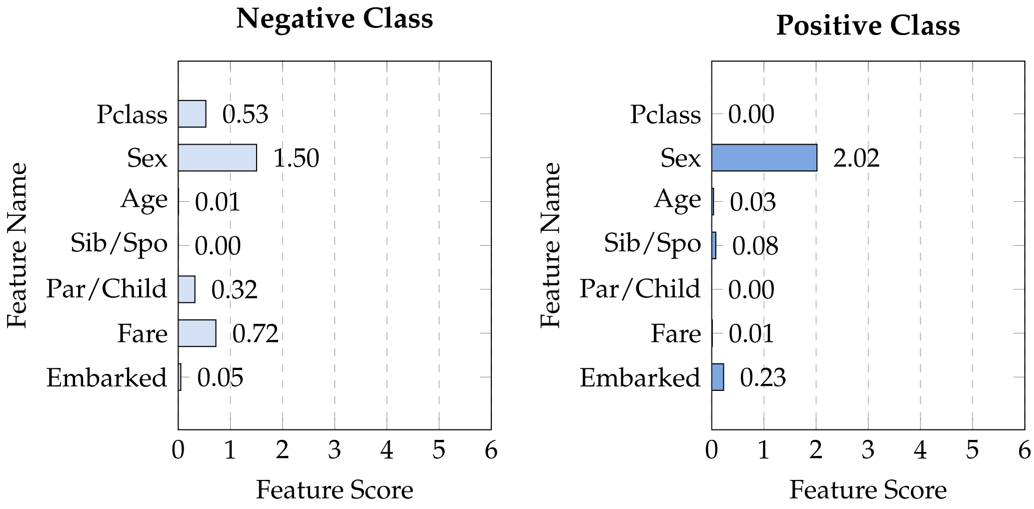


Figure 4.27: Titanic Kaggle dataset: graphical explanation of the *casualty* sub-model - example of a TP case: #1_1_110. Textual explanation: “Casualties model thinks the data is similar to Survivor (-ve) and NOT similar to Casualty (+ve)."

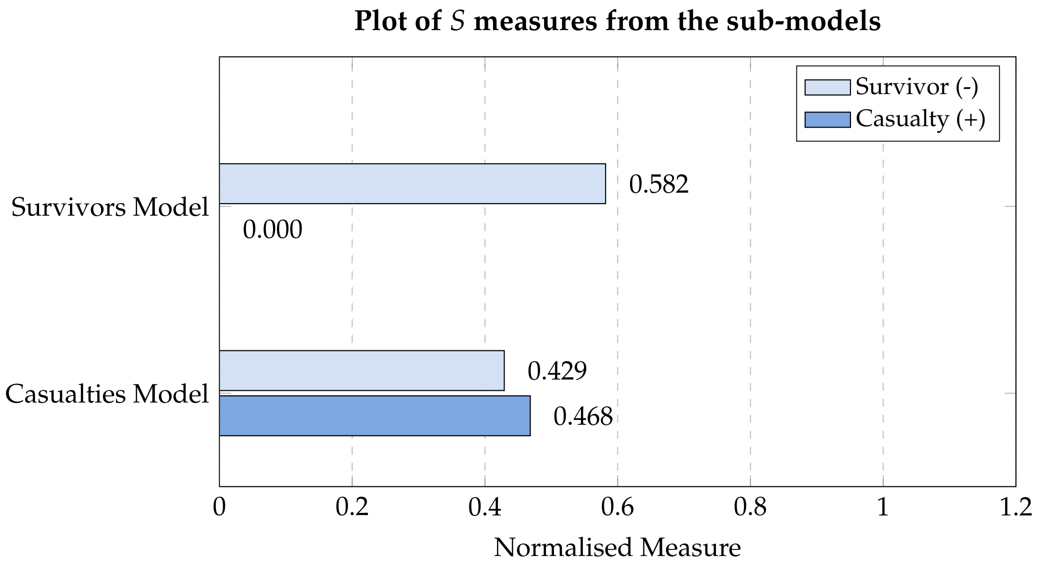


Figure 4.28: Titanic Kaggle dataset: illustration of the distance measures S from the two sub-models - example of a FP case: #1_1_4. Textual explanation: “Models are in conflict hence, it is less likely the classification is accurate." Fuzzy class: 0.72.

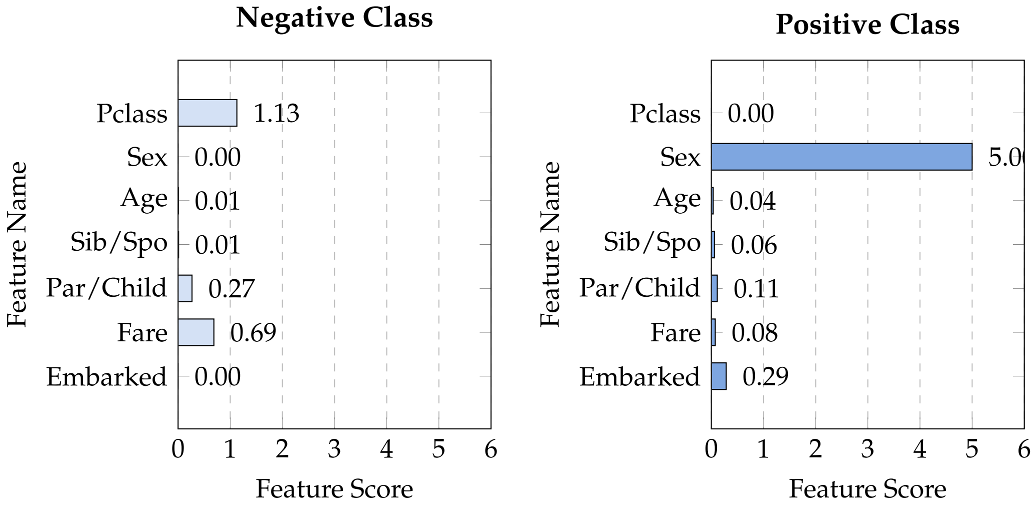


Figure 4.29: Titanic Kaggle dataset: graphical explanation of the *survivor* sub-model - example of a FP case: #1_1_4. Textual explanation: “Survivors model thinks the data is similar to Casualty (+ve) and NOT similar to Survivor (-ve)."

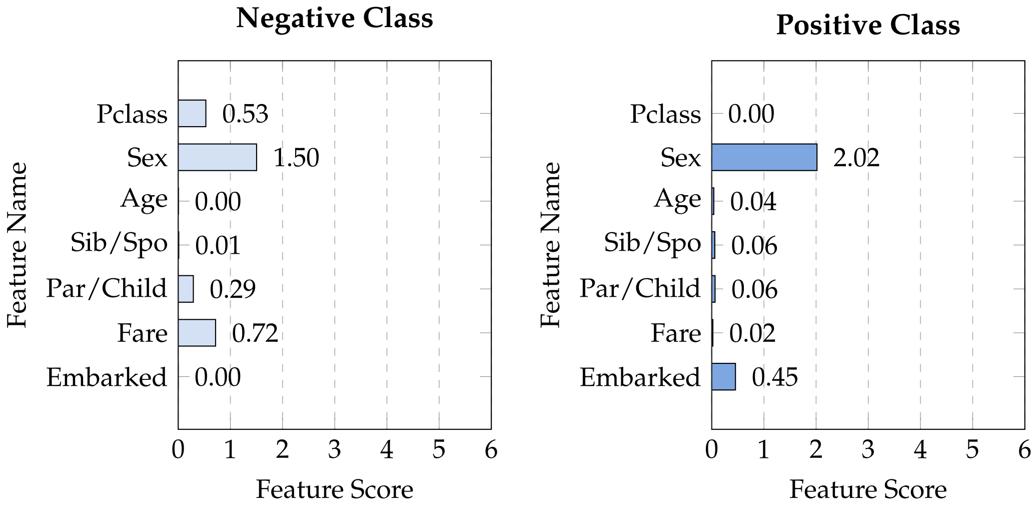


Figure 4.30: Titanic Kaggle dataset: graphical explanation of the *casualty* sub-model - example of a FP case: #1_1_4. Textual explanation: “Casualties model thinks the data is more similar to the Survivor (-ve) despite a high similarity for both."

Contrary to previous case studies, this dataset showcases instances where the explanation framework was able to pinpoint the most contributing features effectively. In the Breast Cancer dataset examples, a majority of the features often contribute to the decision, which makes it less easy to identify the *swing* feature - the feature that resulted in the decision swinging from positive to negative or vice versa. In contrast, the explanation of Kaggle Titanic dataset yielded explanation (Figures 4.25-4.30) with a clear-cut recognition of the most impactful features. In the first two examples, we can see the *Sex* feature is often the most influential by its feature score. By comparison, features such as *Age* and *Sib/Spo* often have low feature scores.

This section demonstrated that the proposed explanation framework applied to Fuzzy-ATOVIC classifiers provided significant insight into how the decision was made. This is considered particularly useful for critical applications with high stakes. It enables the user verify the model’s decision if needed without manually assessing the data. Utilising a transparent form of explanation reduces the risk of misinterpretation of the model that could be present. The limitation of the explanation framework is that it dependent upon the classifier’s accuracy performance. If the classifier performs accurately, the explanation will likely become better-informed.

4.4.2 Explaining non-human-understandable datasets

Parkinson disease dataset

Non-human understandable datasets present a unique challenge for explainability: 'how to explain the non-human understandable'. Where an incomplete set of figures is shown for the graphical explanation examples, a full set is provided in Appendix A.

The explanation framework will provide the same level of insight provided for the previous *human-understandable* datasets with one caveat; the features hold no meaning to the expected user. The Parkinson's disease datasets contain 754 features. Therefore, plotting the feature scores for all features is not likely to be comprehensible by a human user. The feature scores are aggregated into categories (as listed in Table 4.4) for better readability.

Table 4.4: Parkinson disease: feature categories and no. of features in each category

#	Name	Count
1	Gender	1
2	Baseline Features	21
3	Intensity Parameters	3
4	Formant Frequencies	4
5	Bandwidth Parameters	4
6	Vocal Fold	22
7	MFCC	84
8	Wavelet Features	182
9	TQWT Features	432

The first example from the Parkinson disease dataset (Figures 4.31-4.33) is a case where the sub-models were in conflict. The Healthy sub-model classified the data as *Sick* and, the Sick sub-model classified the data as *Healthy*. In this particular case, the classifier's processing of the measures resulted in a FN case because of the categorically *stronger* classification perceived from the Sick sub-model. Despite the difference between the measures being not that different, the Sick sub-model's slight edge caused the decision to be *Sick* rather than *Healthy*.

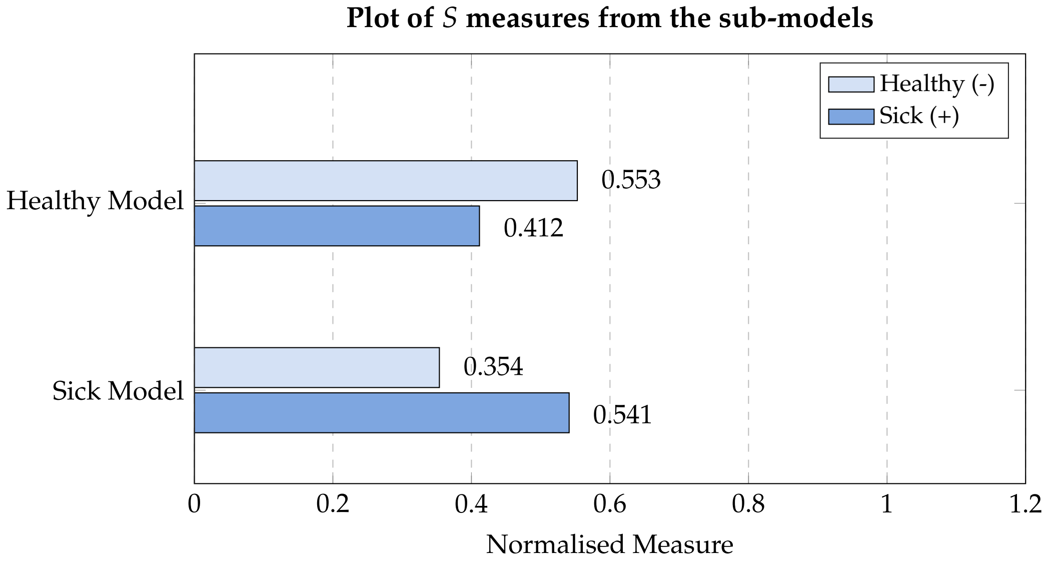


Figure 4.31: Parkinson disease dataset: distance measures S from the two sub-models - example of a FN case: #1_1_41. Textual explanation: "Models are in conflict hence, it is less likely the classification is accurate." Fuzzy class: 0.48.

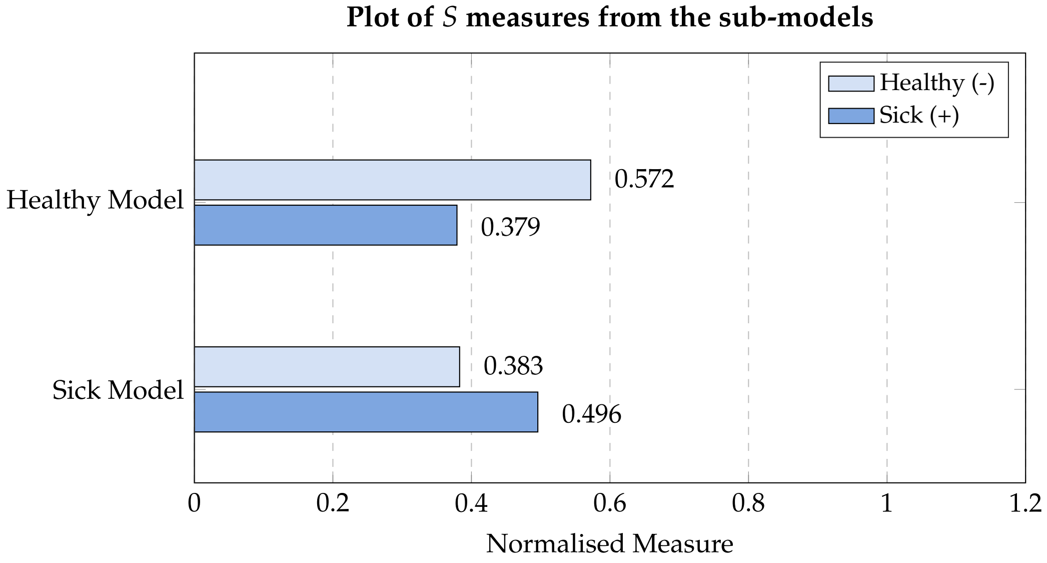


Figure 4.34: Parkinson disease dataset: distance measures S from the two sub-models - example of a TP case: #1_1_43. Textual explanation: "Models are in conflict hence, it is less likely the classification is accurate." Fuzzy class: 0.52.

Similarly, the sub-models were also in conflict for the second example - a true positive case (Figure 4.34). However, in this case, the model could find the right result despite the conflict between the two sub-models. Although the sub-models have a similar conflict, the Healthy sub-model had the edge pushing the overall model towards the correct decision, albeit marginally with a fuzzy class of 0.52 -

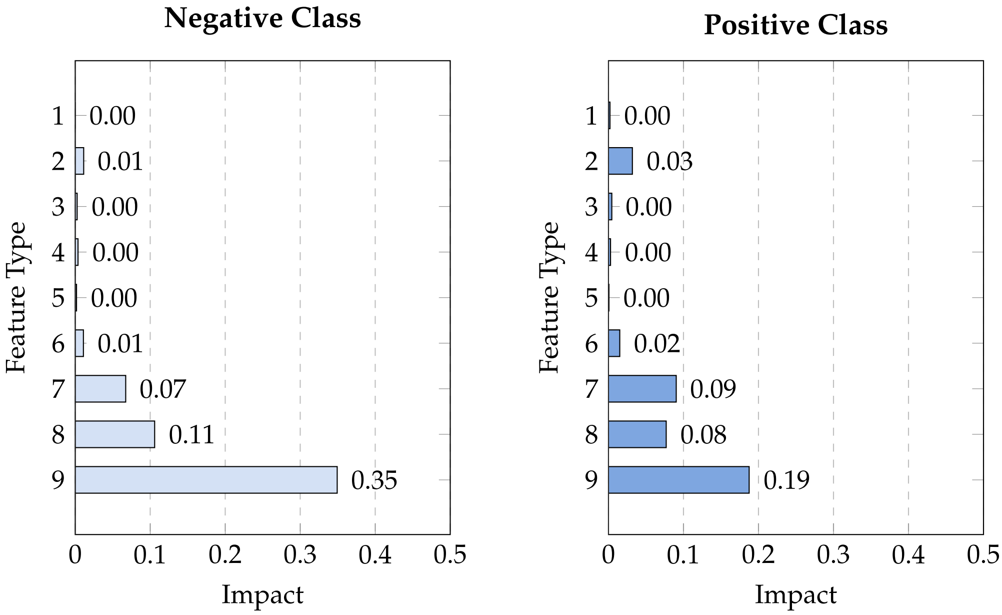


Figure 4.32: Parkinson disease dataset: graphical explanation of the *healthy* sub-model - example of a FN case: #1_1_41. Textual explanation: “Healthy model thinks the data is similar to Positive (+ve) and NOT similar to Negative (-ve)."

just 0.02 above the classification threshold.

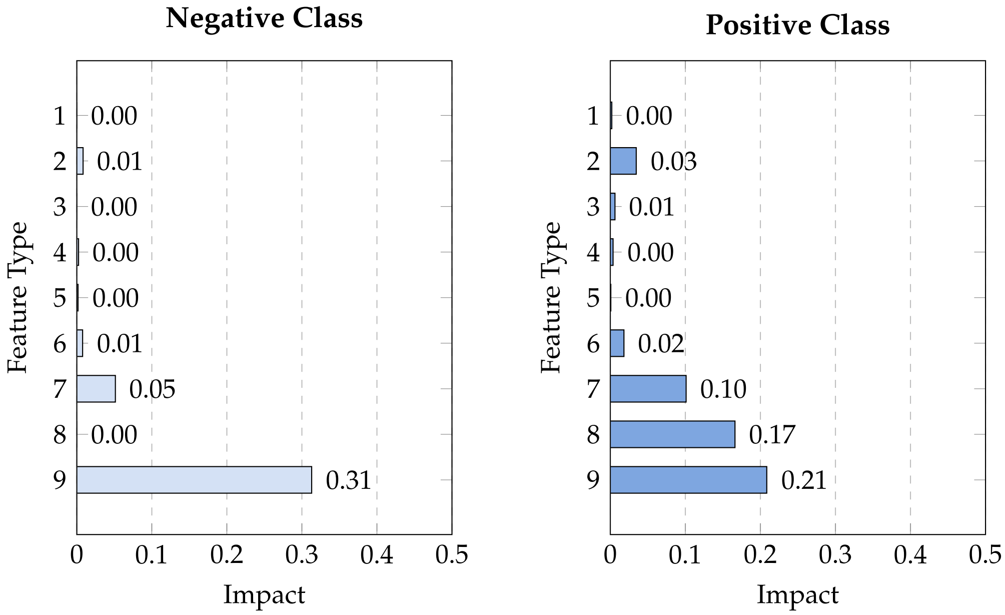


Figure 4.33: Parkinson disease dataset: graphical explanation of the *sick* sub-model - example of a FN case: #1_1_41. Textual explanation: “Sick model thinks the data is similar to Negative (-ve) and NOT similar to Positive (+ve)."

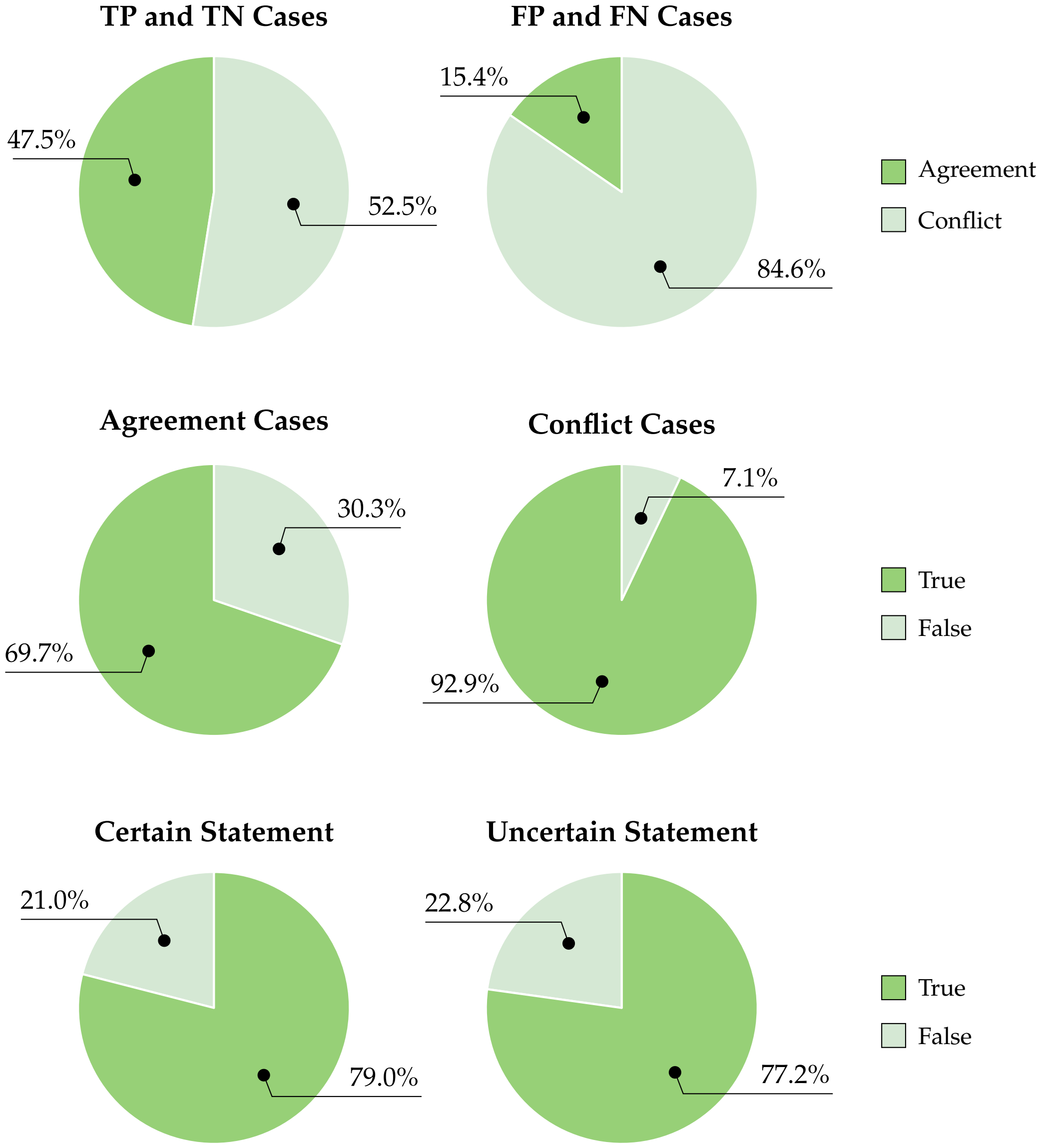


Figure 4.35: Parkinson disease dataset: how *indicative* the different

Analysing the explanation revealed (in Figure 4.35) that for each TP or TN case there is almost a 50/50 proportion of agreement or conflict between the sub-models. However, for cases that are FP or FN, the models were in conflict 84.6% of the time. When relating agreement and conflict to how accurate the model was, sub-models *conflict* was associated with a 92.9% accuracy rate compared to 69.7% for sub-models *agreement*. When comparing how accurate the models were when different linguistic explanation statements were used, the two types (certain vs uncertain) did not differ much in terms of associated accuracy: 79.0% vs 77.2% respectively.

Chess dataset

A non-understandable dataset with a smaller number of features (36) allows for a more comprehensible explanation in the case of the Chess dataset.

The feature-set consists of attributes describing vital conditions that can dictate or indicate the possibility of a *win* or *lose* for white. Despite Chess being a widely studied domain, the number of possible ways in which a game can pan out makes it a rather complex system than commonly thought. For instance, the situation in a Chess endgame where king-rook is faced with king-pawn can be decided in 209,718 possibilities [98]. The potential endings are represented by 36 conditional attributes capturing key factors supporting either outcome.

The features are considered non-understandable because only an expert in Chess theory would be able to comprehend their meaning and significance.

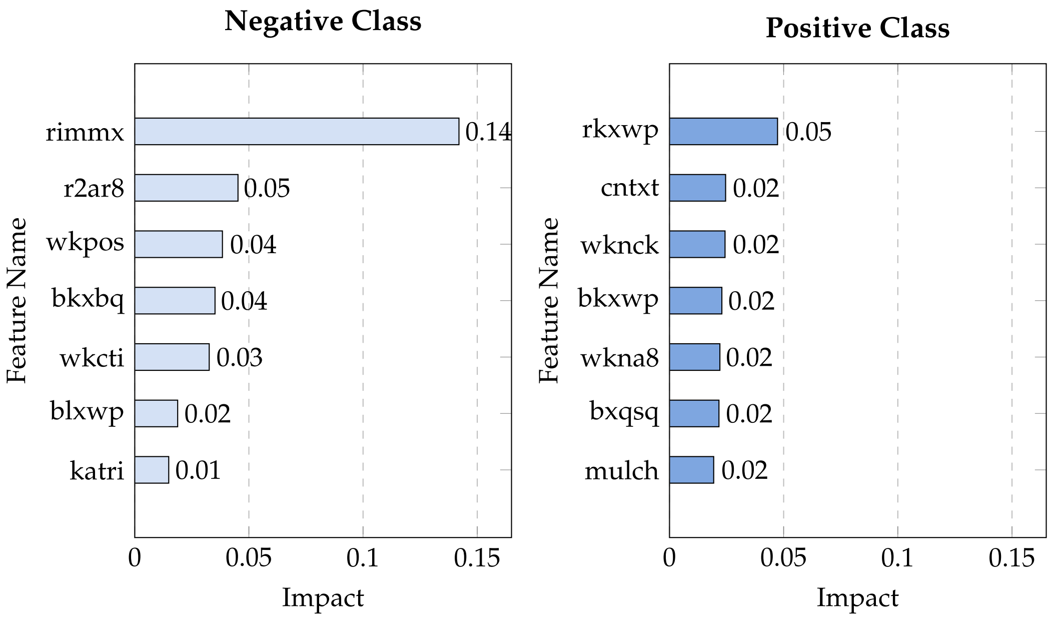


Figure 4.36: Chess dataset: graphical explanation of the *win* sub-model - example of a FN case: #1_1_339. Textual explanation: "Winners model thinks the data is more similar to the Lose (+ve) despite a high similarity for both."

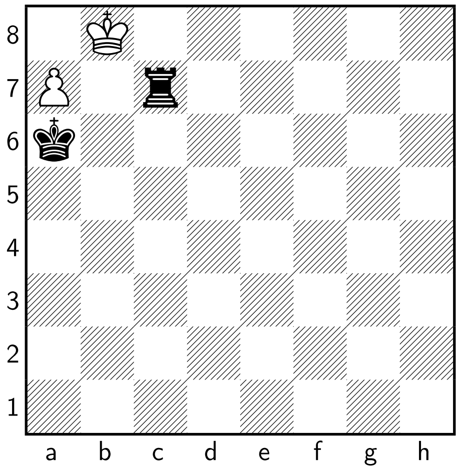


Figure 4.37: An example situation in Chess where white cannot capture the black rook safely; since the white's choice to capture the rook would most likely result in white losing its pawn. Hence, `rimmx` is false in this particular case.

The first example is a FN case (see Figure 4.36) where the `rimmx` feature had a significant impact on the negative classification of a *lose* for both sub-models. The `rimmx` feature represents whether the black rook can be captured safely, which if true, puts white at a clear advantage. Therefore, white is considered unable to win when `rimmx` is false; for example, this can arise in a situation where the black rook can only be captured by sacrificing the white pawn, as shown in Figure 4.37.

Another feature common for both sub-model negative class is `r2ar8`, which checks whether the black rook has safe access to file A or rank 8; both locations enable Black to guard against White's successful *queening* of the knight - a key milestone for White to secure a win (-ve).

Furthermore, the `wknck`, `rkxwp`, `mulch` and `wkna8` features were picked up as most impactful both sub-models. The commonality of the features between the two sub-models reveals an association detected in both sub-models.

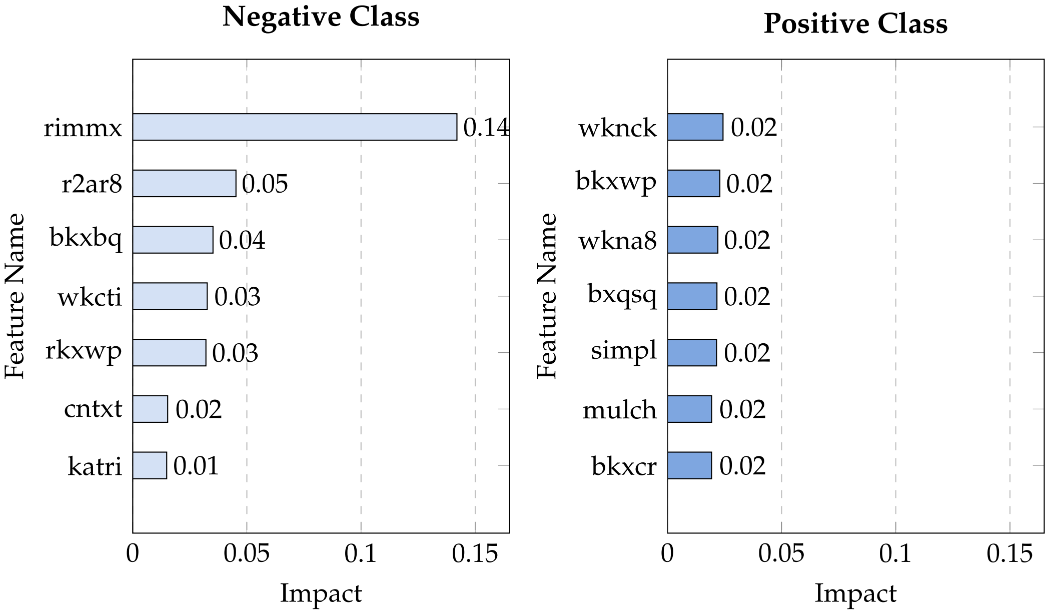


Figure 4.38: Chess dataset: graphical explanation of the *win* sub-model - example of a TN case: #1_1_2. Textual explanation: "Winners model thinks the data is more similar to the Lose (+ve) despite a high similarity for both."

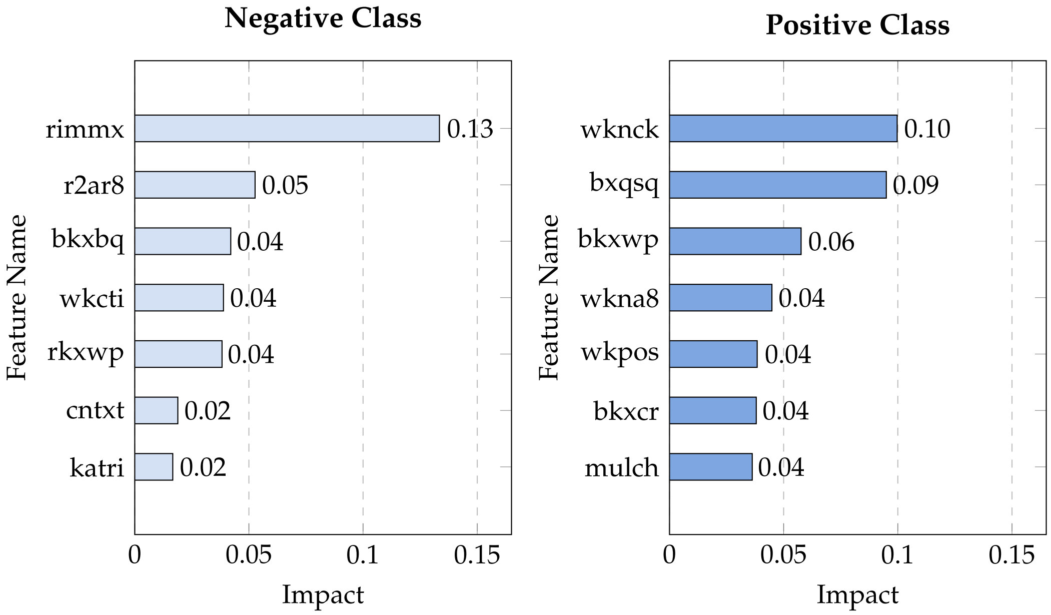


Figure 4.39: Chess dataset: graphical explanation of the *lose* sub-model - example of a TN case: #1_1_2. Textual explanation: "Losers model thinks the data is similar to Win (-ve) and NOT similar to Lose (+ve)."

For the TN case, the S measures indicate that the sub-models were also in conflict and, decided in a similar manner to the FN example. However, this case turned out to be *true*. Comparing the feature impacts (in Figures 4.38-4.39) for the sub-models from the two examples singled-out the fuzzy class as a potential indicator of whether the classification is likely to be *true* or *false*; the fuzzy class for the TN case was further away from the threshold (0.50) - at 0.37 compared to 0.41.

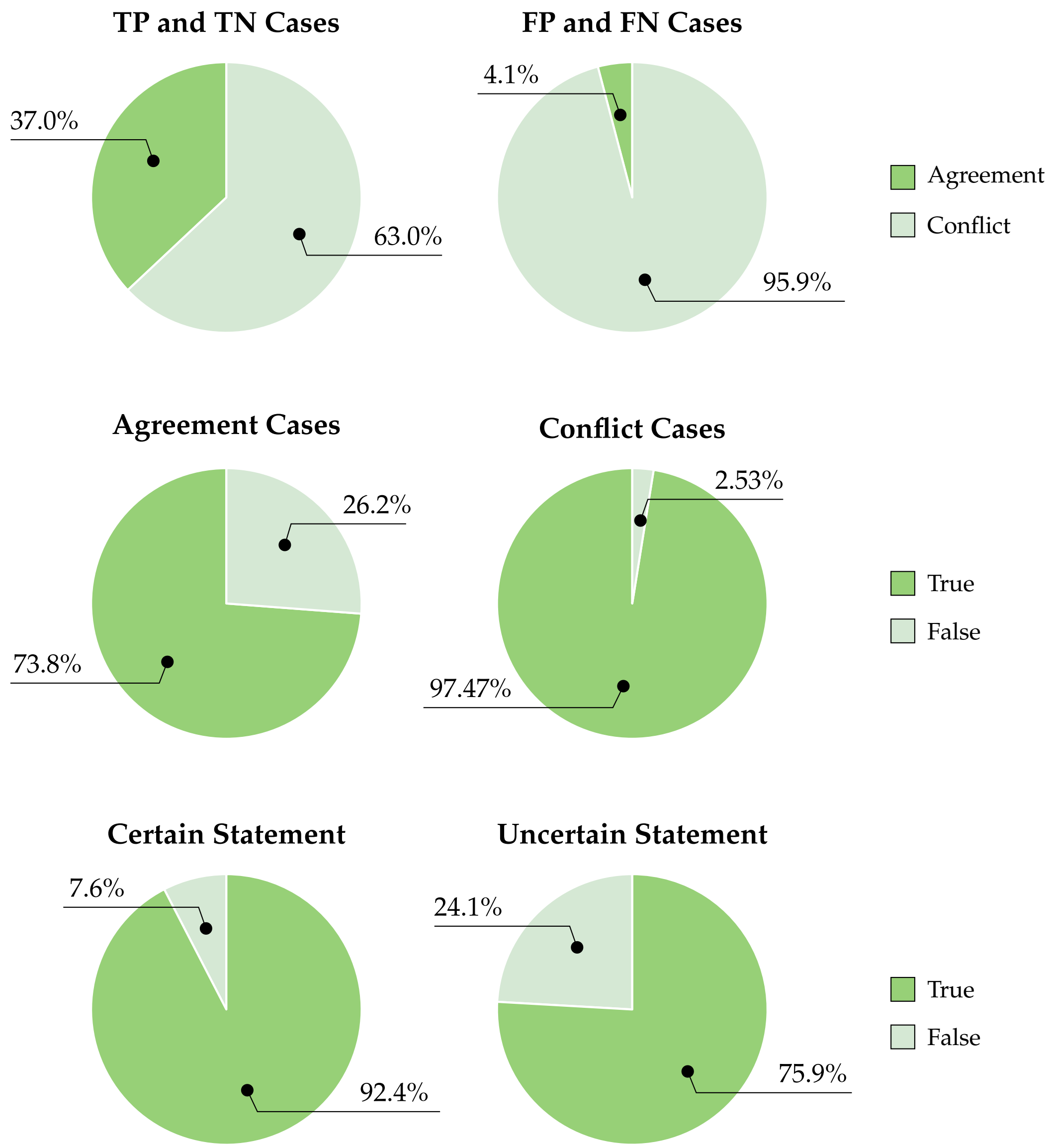


Figure 4.40: Chess dataset: how *indicative* the different aspects of explanation are to a false or negative cases. Certain statements refer to sentence templates #1 and #2 while, uncertain statements refer to sentence templates #3 and #4.

Analysing the relation between different explanation aspects revealed that TP and TN cases were more likely to result in a *conflict* between the sub-models than agreement (see Figure 4.40). However, FP and FN cases are almost certain to cause *conflict* at a rate of 95.9%. Comparing the true-vs-false distribution of cases where the sub-models were in *agreement* or *conflict* showed a higher association between *conflict* and *true* results i.e. accuracy of the result. When the models were in *agreement*, only 73.8% of the results were accurate compared to 97.5% when the models were in *conflict*. The relation between the linguistic statements and performance revealed that certain statements were associated with better accuracy (92.4% vs 75.9%).

4.5 Summary

In this chapter it was demonstrated that the proposed explanation framework developed for Fuzzy-MCDM classifiers provided valuable insight into the decision making process. Textual statements described the sub-models' decisions followed by information on whether they were in *conflict* or *agreement*. The inclusion of the fuzzy class provided a summary of the overall classification in relation to the threshold (0.5). This provided an indication for the the decision's certainty. Applying the explanation framework to human understandable datasets showcased its capability in pinpointing the key features that led to a certain decision. From a practical point of view, accurately distinguishing the data that led to a specific decision is one of the main purposes of explainable-AI.

Applying the explanation framework to non-human understandable datasets such as Parkinson's disease and Chess yielded similar results. However, the high dimensionality of the datasets meant the graphical explanation could only display a subset of the most influential features. Hence, the effectiveness of the framework is limited by the degree of comprehensibility offered by the feature-set. Although the framework was able to find the most impactful features, this information was only as useful as the user's understanding of the features. Therefore, the framework's applicability and effectiveness is dependent upon the availability of human-understandable feature-sets that adequately represent the problem at hand.

The analysis of the various explanation aspects associated with performance accuracy revealed a relation that can be potentially exploited to provide valuable additional information to the user. For instance, whether the model's are in *agreement* or *conflict* can be statistically associated with a higher probability of accuracy. In this next chapter, extending Fuzzy-MCDM's explainability is further investigated.

The explanation framework provided access to traceable insight into the decision. However, when the model was inaccurate, the explanation was often misleading. The users would benefit from indicators of classification certainty.

The feature scores illustrated the impact of each feature to the user. However, in its current representation, it is sometimes not immediately clear which smaller set of features was the tipping point for a decision. Highlighting the most impactful features in a more easily distinguishable way would make the explanation more comprehensible to the user.

5 Neutrosophic-TOPSIS for enhanced explainability

Machine Learning models are able to achieve satisfactory performance in various controlled and uncontrolled environments. In spite of this, uncertainty in the data could inadvertently cause the model to be trained incorrectly, embedding a bias in the model's logic. Some biases and inaccuracies are detected in the model's inception. Meanwhile, less obvious ones are overlooked because of the small proportion of results they affect. When the bias is detected, later on, it is often far too late to undo the harm it has caused. Neutrosophic Logic is a further generalisation of FL that provides additional interpretability using additional sets handling falsity and indeterminacy. This chapter builds on previous work; it proposed a data-driven neutrosophic-TOPSIS classifier to further enhance explainability. The proposal builds on the fuzzy-TOPSIS by using the neutrosophic components to indicate *indeterminacy* and *falsehood*. The two added components provide an extra layer of explainability that is considered useful for providing the user with a more comprehensive understanding of the decision. Performance results demonstrate no or minimal trade-off when opting for neutrosophic-TOPSIS versus fuzzy-TOPSIS. Consequently, an explanation framework is developed to process the newly added components to generate a more detailed explanation of the results.

5.1 Introduction

Data uncertainty is one of the main obstacles preventing the adoption of ML models in numerous fields. Uncertainty is, after all, part and parcel of the world. A system, natural or synthetic, is likely to exhibit at least some uncertainty during its lifetime. For instance, a car's combustion engine is known to be a *time variant* system in control engineering. This means the system's dynamics are expected to vary as the engine ages. Therefore, the engine's control loops may need to be manually tuned to maintain a satisfactory level of performance. Control theory methodology such as adaptive control has been introduced to tackle time varying systems in an autonomous manner.

Similarly, in ML modelling, tackling indeterminacy could be at the forefront of whether a particular methodology succeeds or not. This is because uncertainty, unpredictability, and inconsistencies are increasingly prevalent in real-world datasets.

For instance, the level of uncertainty varies depending on the disease being diagnosed. Despite the unprecedented advancement seen in modern medicine, it is still difficult to diagnose some diseases because of the manner in which they can be detected. Therefore, a methodology capable of handling uncertainty and inconsistency is paramount.

Similarly, in the area of advanced manufacturing, nondestructive testing methodologies for emerging materials could lack the scanning rigour required to provide clear-cut conclusions. Hence, the uncertainty in the results must be conveyed by the ML model's results. In other words, the models must be able to assess how *conclusive* a result is based on the *quality* of the data.

In the second section, a framework is proposed for using NL in conjunction with TOPSIS to construct a data-driven explainable model for classification. The neutrosophic-MCDM classifier presents indeterminacy and falsity in addition to the degree of truth.

The chapter is concluded with a set of results where the proposed framework is applied to seven datasets. Consequently, cases are presented as examples where the model was accurate or inaccurate, shedding light on how the frameworks generate explanations for datasets spanning a variety of applications and data types.

TOPSIS

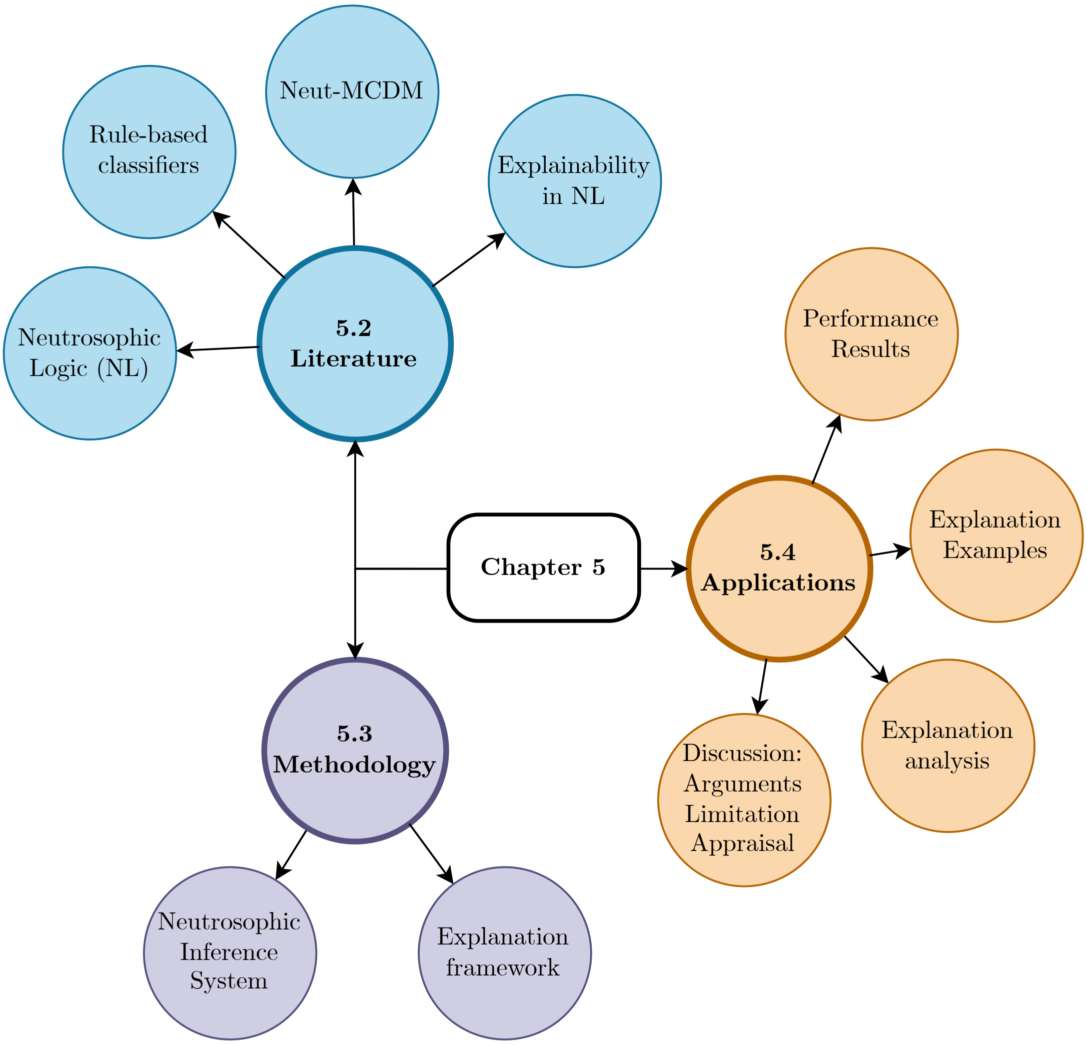


Figure 5.1: Chapter 5: mind map of general topics and concepts

5.2 Neutrosophic logic-based classifiers: an overview

NL is a generalisation of FL that represents each logical variable using an ordered triple; truth, indeterminacy and falsity. In FL, the truth value becomes a real interval [0, 1] where every value within the range represents a degree of truth t with an assumed associated degree of falsity simply defined as $1 - t$.

The falsity was seen as a limitation because of its dependence on truth. Consequently, Atassanov introduced *intuitionistic FL* which augments FL with a real set to represent falsehood f where $f \in [0, 1]$. The sentence p becomes an ordered pair as defined below.

$$t, f \in [0, 1] \tag{5.1}$$

$$v(p) = (t, f) \tag{5.2}$$

$$t + f \leq 1 \tag{5.3}$$

Atassanov's approach allows for the sum of truth and falsity (5.3) to be less than one, which could be used to represent cases where *indeterminacy* exists. The assumed value representing indeterminacy, in this case, would be $1 - (t + f)$. However, what if the indeterminacy, like falsity, would need to be independent of truth and falsity? Smarandache's neutrosophic logic sets aim to address this by defining a separate set for indeterminacy [99]. NL is based on neutrosophy - a relatively recent branch of philosophy also proposed by Smarandache. Neutrosophy is said to be based on 'ancient roots' and aims to address "the origin, nature and scope of neutralities, as well as their interactions with different ideational spectra" [100].

A neutrosophic logic set N is defined as: $N = (T, I, F) : T, I, F \supseteq [0, 1]$. Therefore, each sentence p yields $v(p) = (T, I, F)$; an ordered triple with truth, indeterminacy and falsity components. NL goes on to unrestrict the values the three components can hold such that $t + i + f \leq 3^+$. This allows for representing a variety of inconsistencies. For instance, it is possible that a sentence p a has high truth and falsity simultaneously. Moreover, indeterminacy is not reserved just for indeterminacy but also for handling all sources of unpredictability such as but not limited to uncertainty, imprecision, errors, randomness and vagueness.

The advantage of neutrosophic logic is apparent when attempting to solve complex problems, more specifically, domains where uncertainty, unpredictability and randomness are common. Previous efforts to address the challenges included the investigation of a variety of new set techniques such as intuitionistic, interval-valued and Type-2 FL. Although successful, the methods lacked the ability to process paradoxical inconsistencies. Thus, Smarandache proposed NL with the aim of dealing with a wider range of unpredictability in data [99].

5.2.1 Explainability

As described in Chapter 2, explainability is a key requirement for ML models utilised to support high stake decision-making. FL and MCDM are both considered inherently interpretable frameworks because of their transparent and decomposable structure.

Similarly, NL, as a generalisation of FL, inherits its *interpretability*. In addition to addressing truth, NL has two additional independent sets that represent indeterminacy and falsity. Therefore, the technique can be used to process indeterminate and inconsistent data while still maintaining interpretability.

FLSs use a continuous degree of membership versus the binary one used in binary sets, allowing for finer representation of data. Furthermore, intuitionistic sets provide an additional set to predict falsehood in addition to truth. Meanwhile, the higher dimensionality of NL enables it to represent a wider variety of data, as illustrated in Figure 5.2.

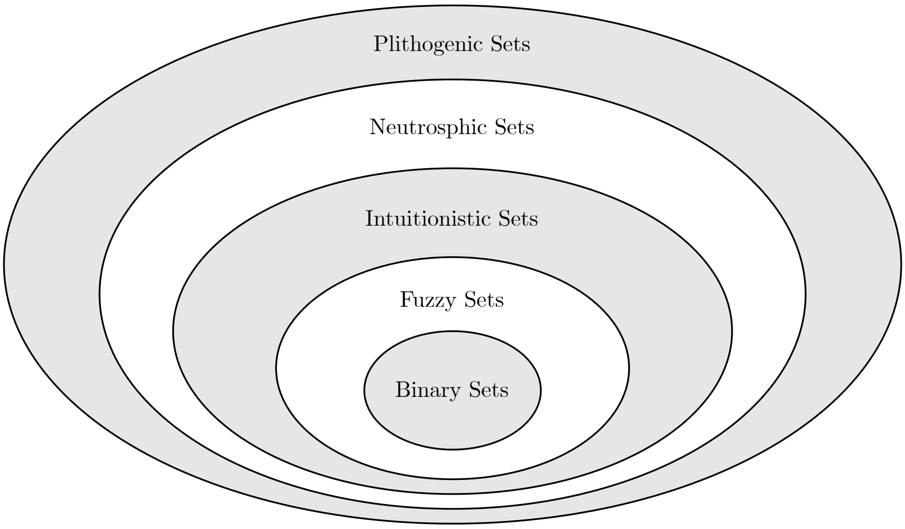


Figure 5.2: A diagram illustrating the variety of data different set types can represent

Plithogenic Sets are a higher dimensionality logic set proposed by Smarandache in 2018 [101]. Smarandache describes the newly proposed sets as a generalisation of "classical logic, fuzzy logic, intuitionistic fuzzy logic, and neutrosophic logic" [101]. NL allows for the use of three sets (truth, indeterminacy and falsehood). Meanwhile, Plithogenic Logic Sets allow for four or more sets, hence can be used to represent a wider gamut of data.

5.2.2 MCDM and neutrosophic logic

MCDM is a popular branch of decision theory often exploited for optimising *decision making*. Similarly, NL has been explored extensively for *decision making*; according to a 10-year bibliometric analysis [102].

As a result, MCDM and NL are often explored in conjunction for augmenting MCDM's capabilities [102]. Despite extensive research in the shared research area of NL-MCDM, there is a lack of investigations focusing on *classification*. Instead, a vast majority focused on MCDM-related applications such as supplier selection [103]–[111]. Moreover, MCDM and NL were combined by replacing MCDM variables with neutrosophic numbers - an approach that omits the rule-based capability of NL.

Therefore, MCDM-based NL is a research gap that could prove fruitful in exploring a data-driven explainable framework for classification. The techniques are both naturally interpretable and possess a simple model structure; thus, it is more likely a high degree of interpretability could be maintained after combining the two methodologies.

5.3 Extending TOPSIS with neutrosophic logic

In this section, a data-driven and explainable neutrosophic-TOPSIS classification framework is proposed. Contrary to previously proposed methodologies pairing Neutrosophy and TOPSIS, this framework uses NL and TOPSIS in conjunction as a *system*.

TOPSIS is extended using a neutrosophic inference system (NIS) that augments TOPSIS's classification capability while providing enhanced interpretable information via the three neutrosophic components: truth, indeterminacy and falsity. Furthermore, an explanation framework is devised to tap into the additional insight which is presented to the user in textual and graphical form.

5.3.1 Neutrosophic inference system

In previous chapters, a data-driven fuzzy-TOPSIS-based classifier was introduced. The fuzzy component was seen as an enhancement to TOPSIS's model interpretability. Meanwhile, performance has varied according to the dataset used. Furthermore, the fuzzy-TOPSIS model, paired with the explanation framework proposed in chapter 4, provided useful insight into the decision-making process. The insight enabled the user to gauge the impact features had on the sub-models and, in turn, the model overall.

The high granularity of such an explanation came at a cost; a level of detail too high to comprehend at a glance. For instance, providing a quantification for the impact of each feature could be seen as too much detail for the user. Thus, there is a need to summarise this information concisely.

The neutrosophic-TOPSIS framework is built by replacing the FL component within fuzzy-TOPSIS with a set of NIS. The inference systems process the S measures from the TOPSIS sub-models to produce the NL outputs: truth, indeterminacy and falsity.

The structure of the neutrosophic component is based on Ansari et al.'s literature [112].

Neutrosophic logic sets are a generalisation of FLSs such that each neutrosophic component is computed by means of a separate inference system, as shown in Figure 5.3. A NIS is used for each MCDM sub-model. By doing so, a pair of neutrosophic outputs are generated for each class and sub-model, providing two for each class and four in total. The number of components becomes 12, a high number in relation to the cognition limit. For this reason, the pair of outputs for the same class are aggregated to produce just six outputs.

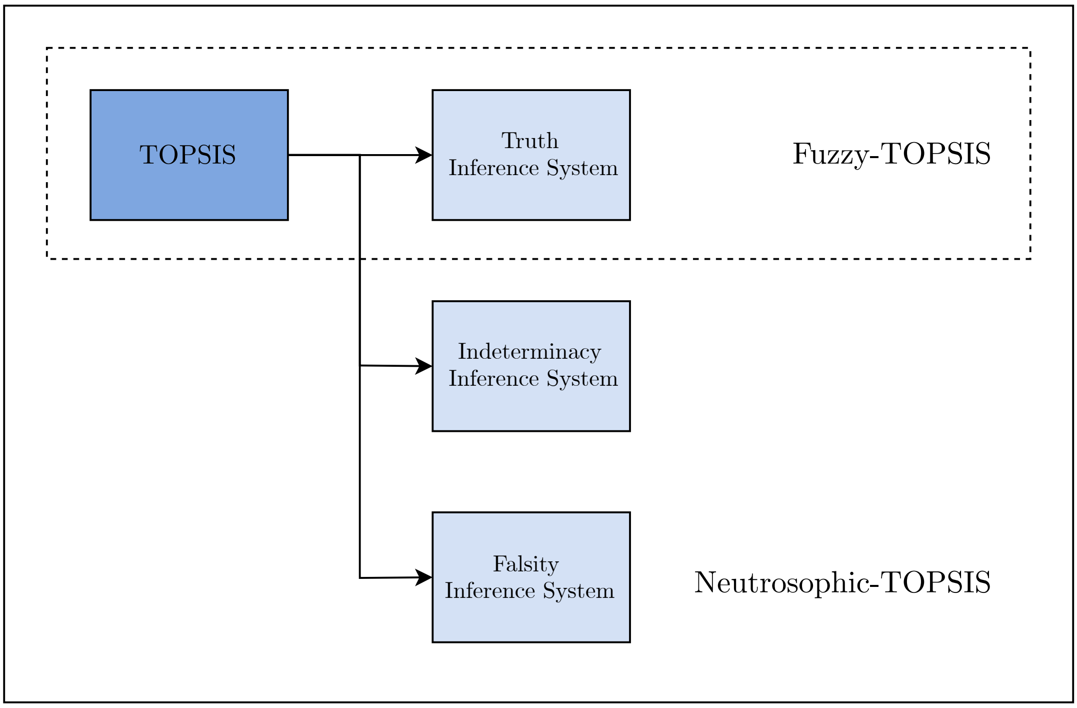


Figure 5.3: A high abstraction visualisation of the difference between fuzzy-TOPSIS and neutrosophic-TOPSIS

The class with the higher aggregated truth is considered to have a higher similarity leading to the output class. The rules of the inference system are formulated such that truth rises as the S measure drops while it is vice versa for the falsity. Meanwhile, the indeterminacy peaks when the S measures for the opposing classes are the same. When the measures are close to being the same, it is a sign the sub-model in question is indecisive. The indeterminacy is considered a useful component to provide the user with an indication as to the certainty of the decision.

Furthermore, the NL extension adopts a hierarchical structure whereby the TOPSIS sub-models' outputs are processed separately using two NL rule-based classifiers. The neutrosophic outputs are then aggregated to form a pair of outputs, one for each class. The rationale behind the design is to ensure a high level of traceability throughout the classification process. By doing so, the neutrosophic outputs are provided for each step in the classification process. This enables a great deal of transparency and decomposability, as illustrated in Figure 5.4.

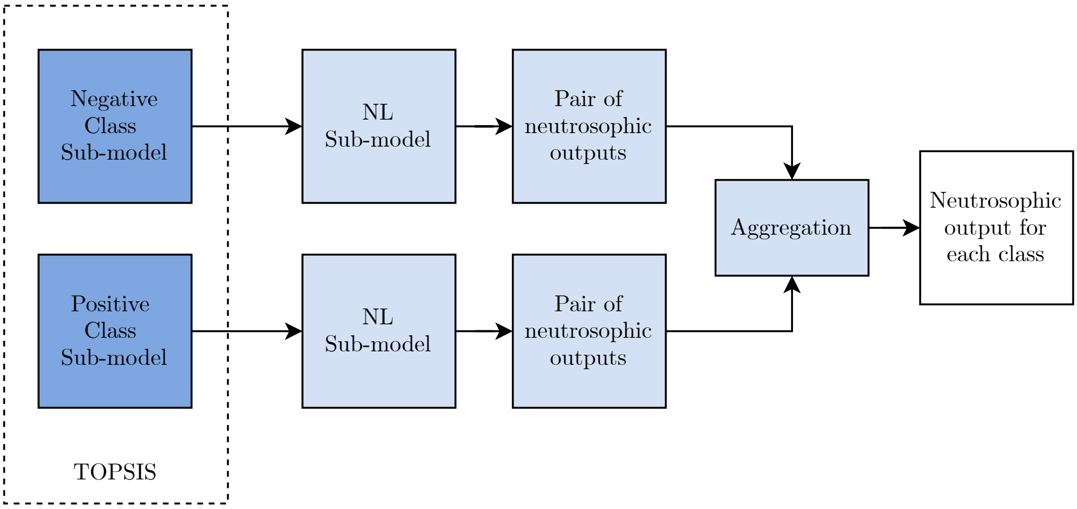

Negative Class Sub-model
Positive Class Sub-model
TOPSIS
NL Sub-model
NL Sub-model
Pair of neutrosophic outputs
Pair of neutrosophic outputs
Aggregation
Neutrosophic output for each class

Figure 5.4: An illustration of the proposed NL extension structure to the TOPSIS-based classifier. It utilised two sub-NL-models, each processing the measures output from the TOPSIS sub-models separately before final stage aggregation and classification.

The inputs to each of the NL sub-models include the following:

(a) $S_{n,1}$: the measure for the negative class; where n is the sub-model number ranging from 1 to 2.

(b) $S_{n,2}$: the measure for the positive class.

(c) ΔS: the difference between the two measures.

Table 5.1: A summary of how changes in the measure S values impact the neutrosophic outputs: truth, indeterminacy and falsity.

		Neutrosophic Output		
Measure S	TOPSIS	Truth	Indeterminacy	Falsity
0.5 to 0.0	High similarity	Rising	Dropping	Dropping
0.5 to 1.0	Low similarity	Dropping	Rising	Rising
$\Delta M \geq 0.05$	High conclusiveness	Polarity dependent	Dropping	Polarity dependent
$\Delta M < 0.05$	Low conclusiveness	Polarity dependent	Rising	Polarity dependent

$$\Delta S_c = S_{c,2} - S_{c,1} \tag{5.4}$$

Where $S_{c,1}$ is the distance to class one; $S_{c,2}$ is the distance to class two; and c is the sub-model number.

The model's dimensionality has to be kept at a minimum to ensure the inference system's interpretability. Therefore, only two input Membership function (MF)s were used for inputs 1 and 2, while three were used for input 3.

Consequently, the rules were configured for each component in a separate inference system (as illustrated previously in Figure 5.3). By doing so, the rule-based classifier maintained a high level of decomposability and simulatability - key aspects of interpretability [12].

For instance, a low S measure for a certain class signifies high similarity and, thus, high truth with low indeterminacy and falsity. In contrast, a high S measure would be vice versa, as shown in Table 5.1. Moreover, the difference between the measures impacts mainly on indeterminacy. However, truth and falsity are also impacted based on the difference's polarity. The polarity of the difference indicates which class is more similar to the other.

In line with FLS, a variety of configurations were proposed for carrying out basic connectives. For the sake of the neutrosophic extension, the conjunction connective is required. Several methods of NL conjunction exist. Rivieccio et al. summarise the key methods below [113]:

$$v(p_1 \wedge p_2) = (t_1 \cdot t_2, i_1 \cdot i_2, f_1 \cdot f_2) \tag{5.5}$$

$$v(p_1 \wedge p_2) = (\min(t_1, t_2), \min(i_1, i_2), \max(f_1, f_2)) \tag{5.6}$$

$$v(p_1 \wedge p_2) = (\min(t_1, t_2), \max(i_1, i_2), \max(f_1, f_2)) \tag{5.7}$$

Where v is the neutrosophic valuation such that $v(p_n) = (t, i, f) \in N$, and;

$$N = \{(T, I, F) : T, I, F \subseteq [0, 1]\} \tag{5.8}$$

Following aggregation using (5.6), the truth value for the opposite classes is compared to perform the final classification. The reason indeterminacy and falsity were not used was that it requires further aggregation. By doing so, it introduces an opaque layer in the classification process. Instead, they remain to provide additional insight on the decision.

5.3.2 Explanation framework

The explanation framework presents the user with textual and graphical information generated as part of the classification. Similarly to the explanation framework proposed in 4.3, this framework aims to enable the user to visualise and trace the decision-making process graphically while being provided with statements that summarise the outcomes of each sub-component of the model.

The sub-components include the TOPSIS and NL sub-models. The TOPSIS component of the model was not altered. Thus, the explanation framework was used as is for that part of the model. Meanwhile, a new set of linguistic statements were designed for the NL sub-models. Likewise, they present factual and counterfactual information in a concise manner.

Three sentence templates were used (Table 5.2) for explaining the three states of a sub-model illustrated in Figure 5.5. The templates include fields to generate the statement based on the class and sub-model being explained. In addition, the statements include neutrosophic outputs for the sub-model being explained. A

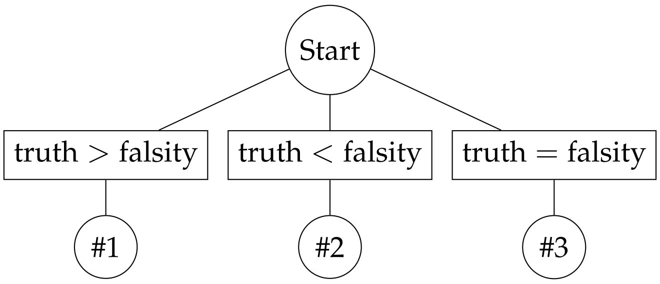


Figure 5.5: Visualisation of the conditions for selecting the sentence template.

Table 5.2: Linguistic sentence templates used for explaining the classifier's NL sub-models

#	Sentence template
1	For the `<cc>` class, the `<model>` thinks the data has a higher truth (`<t>`) relative to its falsity (`<f>`). The relative indeterminacy is `<ind>` (`<i>`).
2	For the `<cc>` class, the`<model>` thinks the data has a higher falsity (`<f>`) relative to its truth (`<t>`). The relative indeterminacy is `<ind>` (`<i>`).
3	For the `<cc>` class, the `<model>` thinks the data has a truth (`<t>`) equal to its falsity (`<f>`). The relative indeterminacy is `<ind>` (`<i>`).

descriptor was added for specifying whether indeterminacy was *less*, *average*, *more*. The descriptors are the same as the names of the MFs used in the indeterminacy inference system.

5.4 Comparative analysis of fuzzy-TOPSIS and neut-TOPSIS

The focus should be now to the experimental evidence on Neutrosophic-TOPSIS, where the proposed methodology is applied to seven different datasets. Five of the seven are the same benchmark datasets used to assess the modelling frameworks proposed in Chapters 3 and 4. The last two datasets are based on a plastic pipe inspection case study for enhancing volumetric nondestructive testing of industrial plastic pipelines.

Table 5.3: Parkinson Disease Dataset: neut-TOPSIS vs fuzzy-TOPSIS

	$\mu \pm \sigma$ (%)			
Model	ACC	FMR	TPR	TNR
fuzzy-TOPSIS	78.7 ± 3.8	70.5 ± 4.9	84.6 ± 6.5	61.4 ± 8.9
neut-TOPSIS	76.4 ± 3.4	71.3 ± 4.7	80.4 ± 5.1	64.7 ± 8.4

Table 5.4: Breast Cancer Dataset: neut-TOPSIS vs fuzzy-TOPSIS

	$\mu \pm \sigma$ (%)			
Model	ACC	FMR	TPR	TNR
fuzzy-TOPSIS	93.2 ± 2.0	89.9 ± 3.1	82.6 ± 5.1	98.8 ± 1.0
neut-TOPSIS	92.2 ± 1.8	88.0 ± 3.2	79.2 ± 5.0	99.1 ± 0.9

5.4.1 Performance: benchmark datasets

Even though performance is not the ultimate indicator for appraising explainable ML models, nonetheless, they are imperative for ruling out any significant performance trade-offs. To that effect, this section presents a comparative analysis between fuzzy-TOPSIS and Neutrosophic-TOPSIS.

The benchmark datasets are the same ones used in the previous Chapters 3 and 4. For details of dataset structure and attributes, please refer to Section 3.4.1. The neut-TOPSIS framework introduced additional components to provide a higher level of interpretability. Based on theories on the link between interpretability and performance, it was expected that neut-TOPSIS could result in a reduction in performance [3].

In line with the expectation, the accuracy (ACC) of the neut-TOPSIS was lower at 76.4% compared to 78.7% of the fuzzy-TOPSIS model (Table 5.3). The improved true negative rate (TNR) for the neut-TOPSIS model meant a lower true positive rate (TPR). Nonetheless, the neut-TOPSIS still had a slightly higher F-measure (FMR) overall. Thus, performance is perceived to be largely similar for the two frameworks for this dataset.

Similarly, for the breast cancer dataset the ACC was lower for the neut-TOPSIS model, as presented in Table 5.4. This was mainly due to the 3.4% drop in the TPR. The figures so far from the two benchmark datasets indicate a similar performance

Table 5.5: Keel Titanic Dataset: neut-TOPSIS vs fuzzy-TOPSIS

	$\mu \pm \sigma$ (%)			
Model	ACC	FMR	TPR	TNR
fuzzy-TOPSIS	77.8 ± 1.9	63.8 ± 4.1	49.1 ± 4.7	91.5 ± 1.3
neut-TOPSIS	77.6 ± 1.9	63.2 ± 4.3	48.4 ± 4.9	91.5 ± 1.3

Table 5.6: Kaggle Titanic Dataset: neut-TOPSIS vs fuzzy-TOPSIS

	$\mu \pm \sigma$ (%)			
Model	ACC	FMR	TPR	TNR
fuzzy-TOPSIS	78.8 ± 3.0	75.9 ± 4.0	68.7 ± 6.1	85.1 ± 2.5
neut-TOPSIS	78.9 ± 2.4	74.0 ± 4.6	64.6 ± 7.8	87.7 ± 4.6

between the two frameworks, despite the increased interpretability of neut-TOPSIS.

The trend is similar for the two titanic benchmark datasets (KEEL and Kaggle), where there was a slight drop in ACC and FMR performance, as presented in Tables 5.5-5.6. Likewise, the TPRs saw a reduction for the two datasets.

In contrast to the previous datasets, the neut-TOPSIS saw an improvement in all performance metrics for the chess dataset, as shown in Table 5.7. Despite the increase in performance, the standard deviation has stayed the same or improved for all metrics except the TNR.

In summary, the performance across the two different frameworks has seen a range of results. In spite of the enhanced interpretability of the neut-TOPSIS versus the fuzzy-TOPSIS framework, accuracy was satisfactorily maintained for all the benchmark datasets with at most a 2.3% drop in performance for the Parkinson's disease dataset. Meanwhile, at best was a 4.0% improvement in performance seen for the chess dataset. The results indicate a non-significant loss, if any, in performance across the two different frameworks. The results indicate

Table 5.7: Chess Dataset: neut-TOPSIS vs fuzzy-TOPSIS

	$\mu \pm \sigma$ (%)			
Model	ACC	FMR	TPR	TNR
fuzzy-TOPSIS	81.2 ± 1.8	80.5 ± 1.9	75.3 ± 2.8	86.6 ± 2.4
neut-TOPSIS	85.2 ± 1.8	85.0 ± 1.8	82.9 ± 2.3	87.2 ± 2.6

neut-TOPSIS's viability as an alternative to the fuzzy-TOPSIS classifier proposed in Chapter 3.

5.4.2 Explaining benchmark datasets

Before proceeding to examine the explanation examples, it is important to point out that non-satisfactory performance can affect how meaningful the generated explanation can be. The explanation examples seek to document the successes as well as the failures of the model for each of the five benchmark datasets.

Breast cancer

The breast cancer dataset is an example where the model performed particularly well. Examples of the model explanation in this section represent a best-case scenario example.

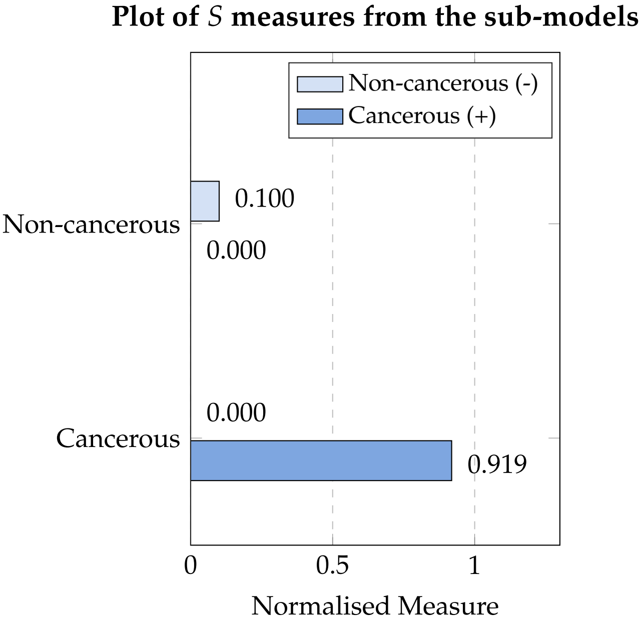
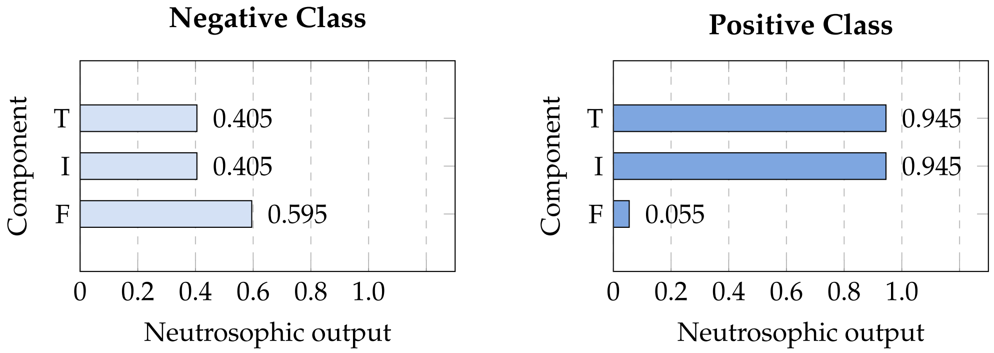


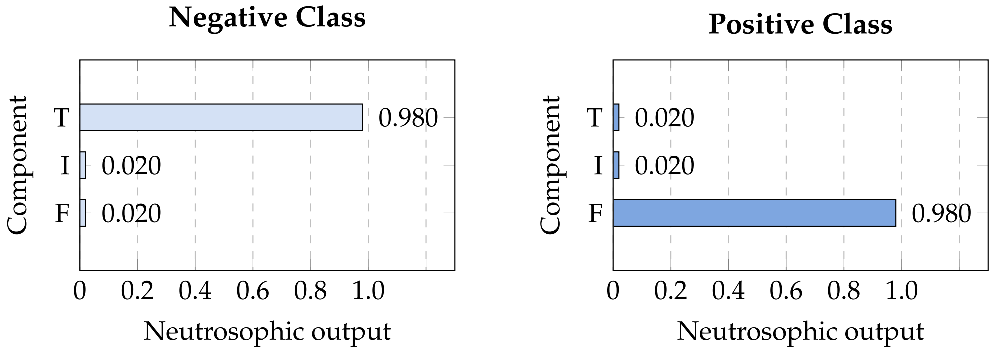
Figure 5.6: Breast cancer dataset: distance measures S from the two sub-models - example of a TN case: #5_2_61. Textual explanation: "Models are in conflict hence, it is less likely the classification is accurate. The data was classified as *Non-cancerous (-)* by the model."



(a)

(b)

Figure 5.7: Breast cancer dataset: Graphical explanation of the *Non-cancerous (-)* NL sub-model - example of a TN case: #5_2_61. (a) Explanation: "For the Non-cancerous (-) class, the Non-cancerous Model thinks the data has a higher truth (0.945) relative to its falsity (0.055). The relative indetermency is low (0.405).". (b) Explanation: "For the Cancerous (+) class, the Non-cancerous Model thinks the data has a higher falsity (0.595) relative to its truth (0.405). The relative indetermency is high (0.945)."



(a)

(b)

Figure 5.8: Breast cancer dataset: Graphical explanation of the *Cancerous (+)* NL sub-model - example of a TN case: #5_2_61. (a) Explanation: "For the Non-cancerous (-) class, the Non-cancerous Model thinks the data has a higher truth (0.980) relative to its falsity (0.020). The relative indetermency is low (0.020).". (b) Explanation: "For the Cancerous (+) class, the Non-cancerous Model thinks the data has a higher falsity (0.980) relative to its truth (0.020). The relative indetermency is low (0.020)."

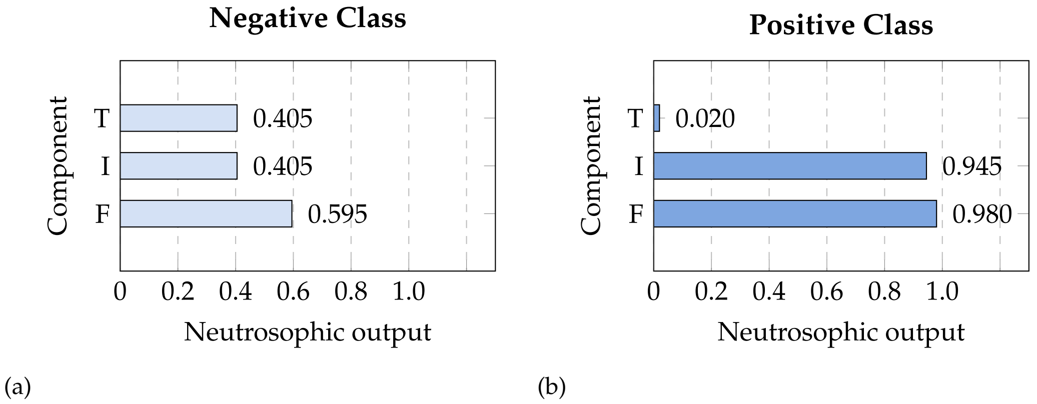


Figure 5.9: Breast cancer dataset: Graphical explanation of the overall NL sub-model - example of a TN case: #5_2_61. Overall explanation: "Based on the overall sub-model's output the data was more similar to Non-cancerous (-) compared to Cancerous (+)." (a) Explanation: "For the Non-cancerous (-) class, the overall sub-model thinks the data has a higher truth (0.945) relative to its falsity (0.055). The relative indetermency is low (0.405).". (b) Explanation: "For the Cancerous (+) class, the overall sub-model thinks the data has a higher falsity (0.980) relative to its truth (0.020). The relative indetermency is high (0.945)."

Firstly, a TN case is presented in which the MCDM sub-models were in conflict, as shown in Figures 5.6-5.9.

The conflict meant that the sub-models decided on different classes. In this case, the negative sub-model pointed towards a higher similarity to the positive class, as shown in Figure 5.7. Meanwhile, the positive sub-model pointed towards a higher similarity to the negative class (Figure 5.8).

The pairs of neutrosophic outputs from the sub-models are then aggregated to produce a pair of outputs representing the final decision.

Overall, the negative class had a higher truth (0.405) compared to the low truth for the positive class, hence the negative classification. Similarly, the falsity components indicated the same information with a higher falsity (0.980) for the positive class compared to the negative class (0.595).

Overall, the positive class sub-model's indeterminacy was relatively high (0.980) due to the small difference between the S measures, as shown in Figure 5.6. Although only the truth component is used for the decision, the additional NL components provide a quantifiable indication of the indeterminacy and falsity associated with the decision.

The following examples will show just the overall aggregated neutrosophic output. Please refer to Appendix B for an extended copy of all the examples shown.

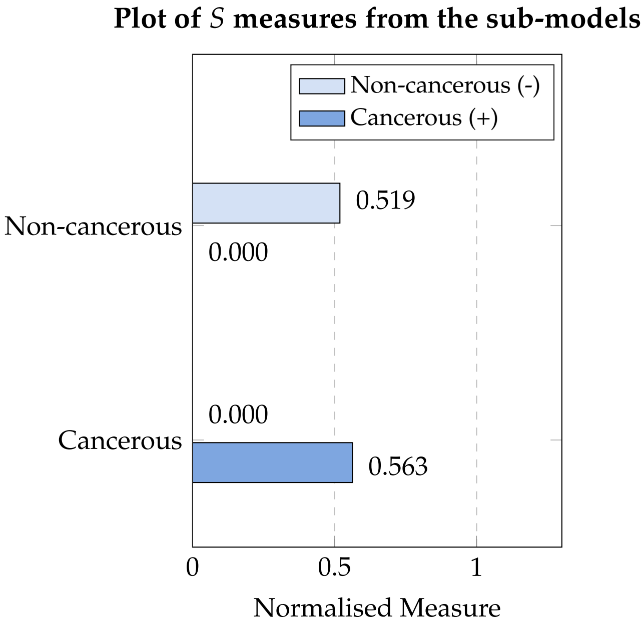
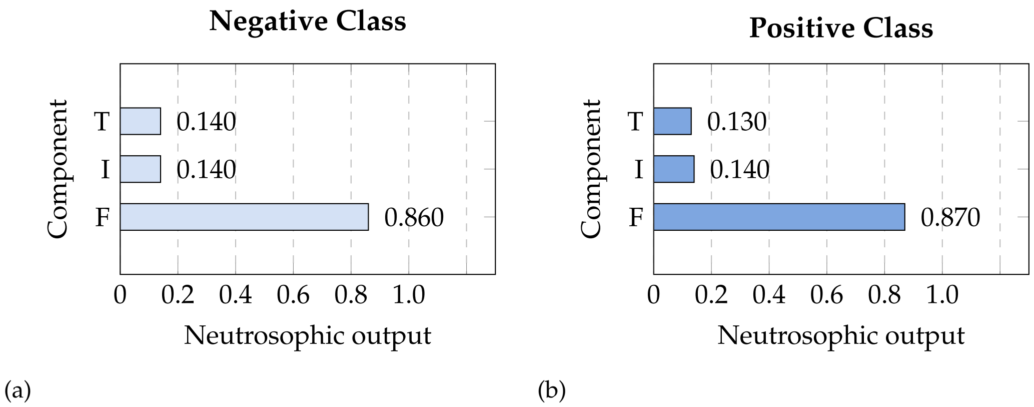


Figure 5.10: Breast Cancer Dataset: distance measures S from the two sub-models - example of a FN case: #5_3_104. Textual explanation: "Models are in conflict hence, it is less likely the classification is accurate. The data was classified as *Non-cancerous (-)* by the model."

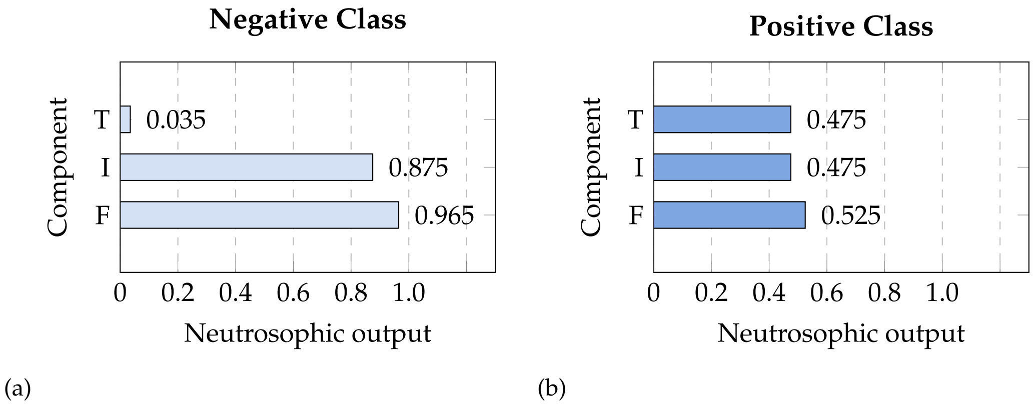


(a)

(b)

Figure 5.11: Breast cancer dataset: graphical explanation of the overall NL sub-model - example of a FN case: #5_3_104. Overall explanation: "Based on the overall sub-model's output the data was more similar to Non-cancerous (-) compared to Cancerous (+)." (a) Explanation: "For the Non-cancerous (-) class, the overall sub-model thinks the data has a higher falsity (0.860) relative to its truth (0.140). The relative indetermency is low (0.140).". (b) Explanation: "For the Cancerous (+) class, the overall sub-model thinks the data has a higher falsity (0.870) relative to its truth (0.130). The relative indetermency is low (0.140)."

Similarly, the sub-models are in conflict for this FN example, as shown in Figures 5.10-5.11. The NL component generated similar truth levels for the negative and positive classes, at 0.140 and 0.130, respectively. Two aspects indicate low certainty in the decision: small truth levels and high falsity for the negative class.



(a)

(b)

Figure 5.12: Breast cancer dataset: graphical explanation of the overall NL sub-model - example of a TP case: #8_5_106. Overall explanation: "Based on the overall sub-model's output the data was more similar to Cancerous (+) compared to Non-cancerous (-)." (a) Explanation: "For the Non-cancerous (-) class, the overall sub-model thinks the data has a higher falsity (0.965) relative to its truth (0.035). The relative indetermency is high (0.875).". (b) Explanation: "For the Cancerous (+) class, the overall sub-model thinks the data has a higher truth (0.875) relative to its falsity (0.500). The relative indetermency is average (0.475)."

For the previous example, the indeterminacy was similar for both classes. In this TP example, however, the positive class had a lower value of indeterminacy (see Figure 5.12). Despite the low truth for the negative class, its high indeterminacy conveys a potential misclassification. However, coupled with the positive class's higher truth makes for an *average* positive classification.

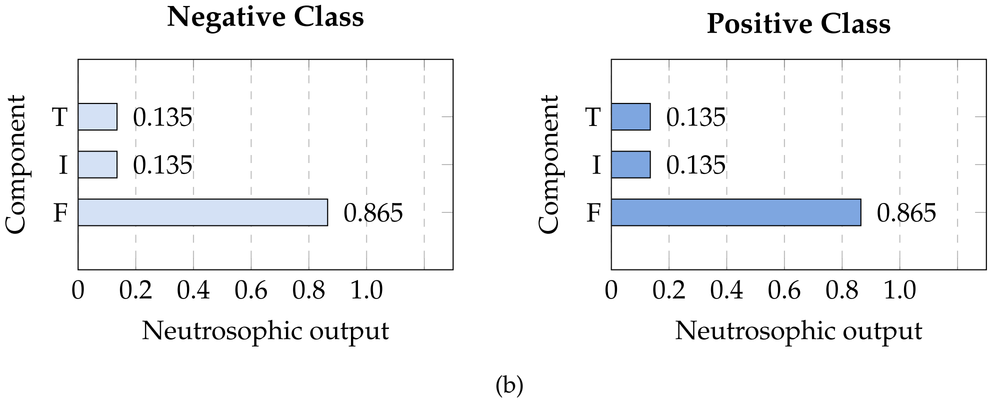


Figure 5.13: Breast Cancer Dataset: Graphical explanation of the overall NL sub-model - example of a FP case: #6_2_22. Overall explanation: "Based on the overall sub-model's output the data was more similar to Cancerous (+) compared to Non-cancerous (-)." (a) Explanation: "For the Non-cancerous (-) class, the overall sub-model thinks the data has a higher falsity (0.865) relative to its truth (0.135). The relative indetermency is low (0.135).". (b) Explanation: "For the Cancerous (+) class, the overall sub-model thinks the data has a higher falsity (0.865) relative to its truth (0.135). The relative indetermency is low (0.135)."

Moreover, this FP case had a similar output to the TP case (Figure 5.13). The distinction in the following values points to a lower certainty compared to the TP case:

- Lower positive class truth: 0.135 versus 0.475
- Lower positive class indeterminacy: 0.135 versus 0.475
- Higher negative class truth: 0.135 versus 0.035
- Lower negative class indeterminacy: 0.135 versus 0.875
- Lower negative class falsity: 0.865 versus 0.965

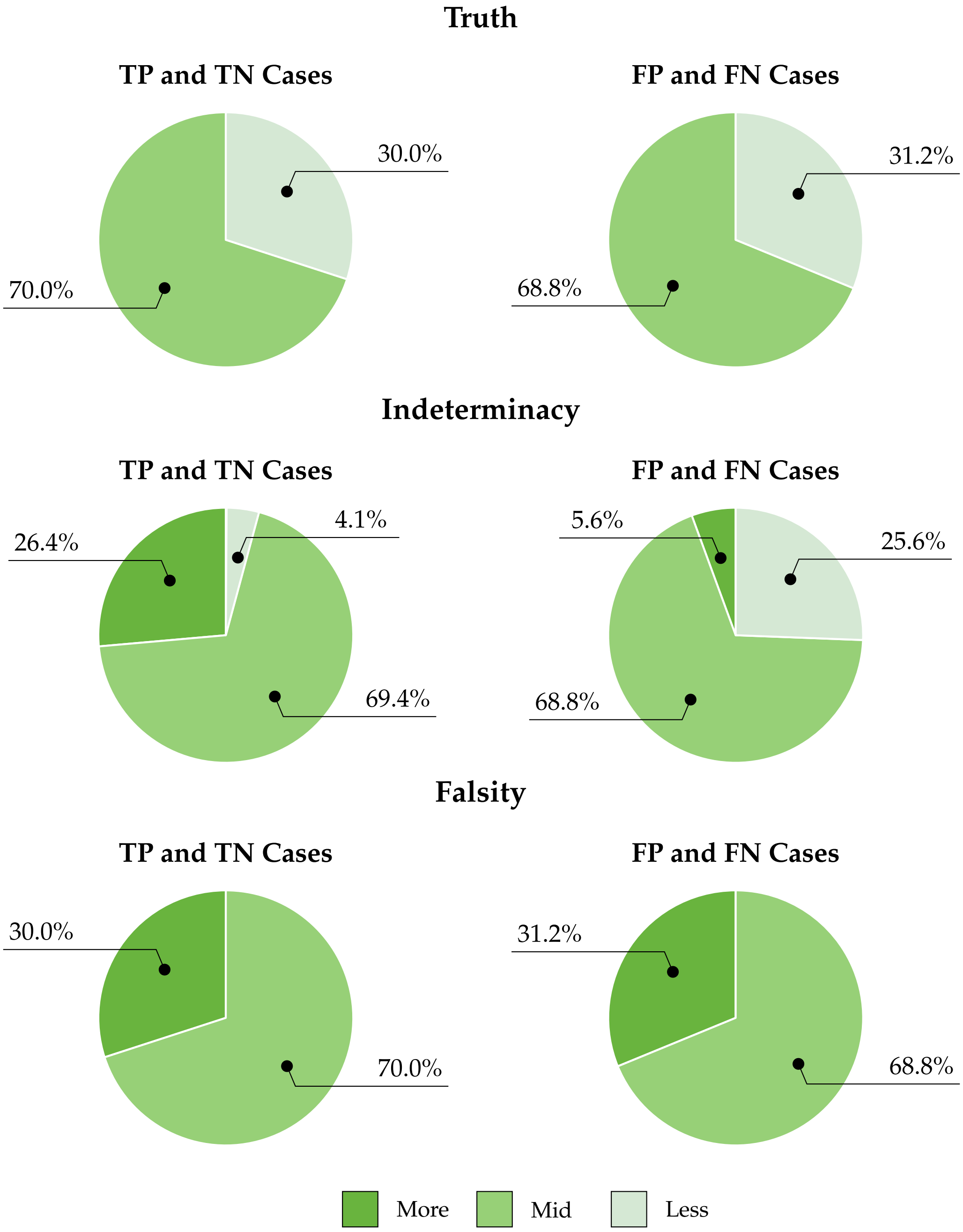


Figure 5.14: Breast Cancer Dataset: a set of pie charts presenting an analysis of how the different aspects of the neutrosophic outputs relate to accuracy performance for the negative class sub-model.

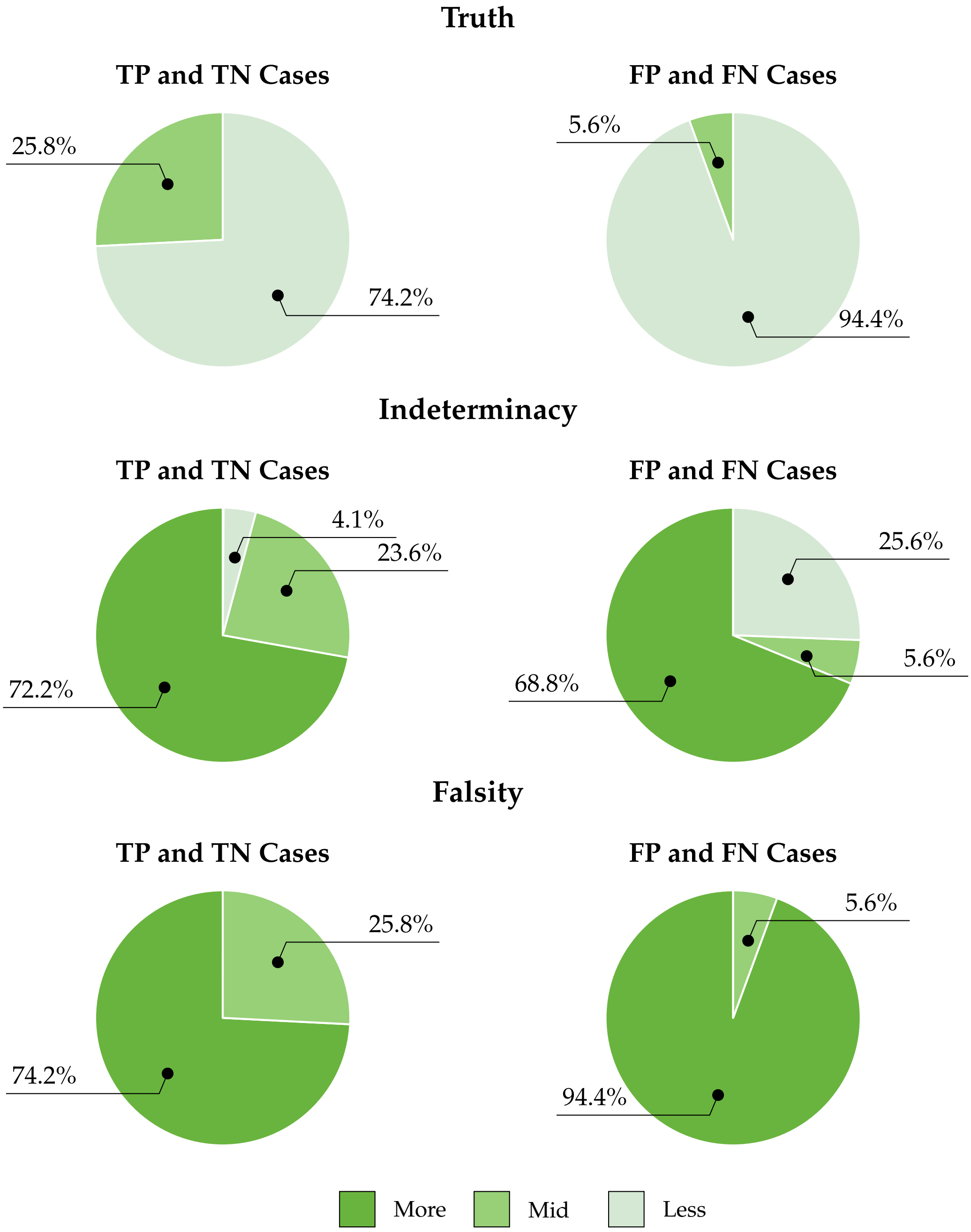


Figure 5.15: Breast Cancer Dataset: a set of pie charts presenting an analysis of how the different aspects of the neutrosophic outputs relate to accuracy performance for the positive class sub-model.

As described in section 5.3, the neutrosophic outputs (truth, indeterminacy and falsity) are calculated based on three membership functions: more, mid and less. Based on this, an analysis was run to view the relation between the output MF and performance. The analysis was run for both NL sub-models and is presented in Figures 5.14-5.15.

For the truth and falsity components, the negative class NL sub-model produced a near similar distribution across accurate and inaccurate cases (Figure 5.14). However, the indeterminacy component had a higher proportion of *more* for accurate cases compared to inaccurate (26.4% versus 5.6%) and a lower proportion of *low* (4.1% versus 25.6%). Therefore, counter-intuitively, indeterminacy is more likely to be higher values for accurate cases compared to inaccurate.

In contrast, the truth and falsity components for the positive class sub-model did not reflect the same distributions. Intuitively, accurate cases were more likely to have larger truth values and smaller falsity values. Meanwhile, the indeterminacy components produced larger values for accurate cases.

KEEL titanic

The KEEL titanic dataset uses three features to classify a passenger as a survivor or casualty.

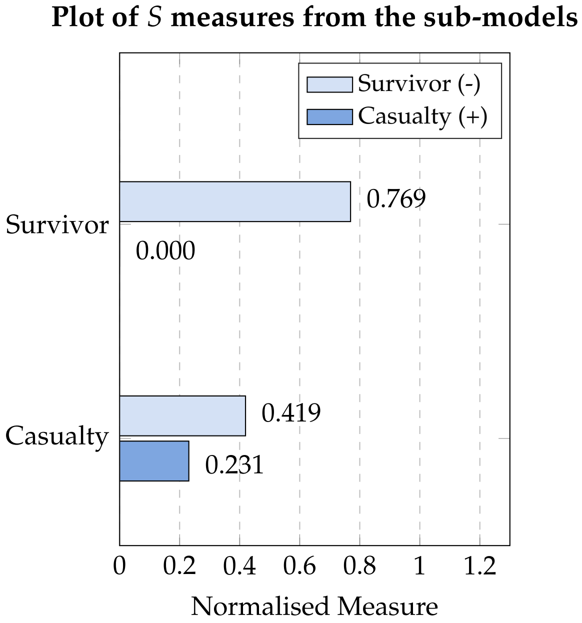


Figure 5.16: Titanic - KEEL Dataset: Illustration of the distance measures S from the two sub-models - example of a TP case: #3_4_309. Textual explanation: "Models are in agreement hence, it is more likely the classification is accurate. The data was classified as *Casualty (+)* by the model."

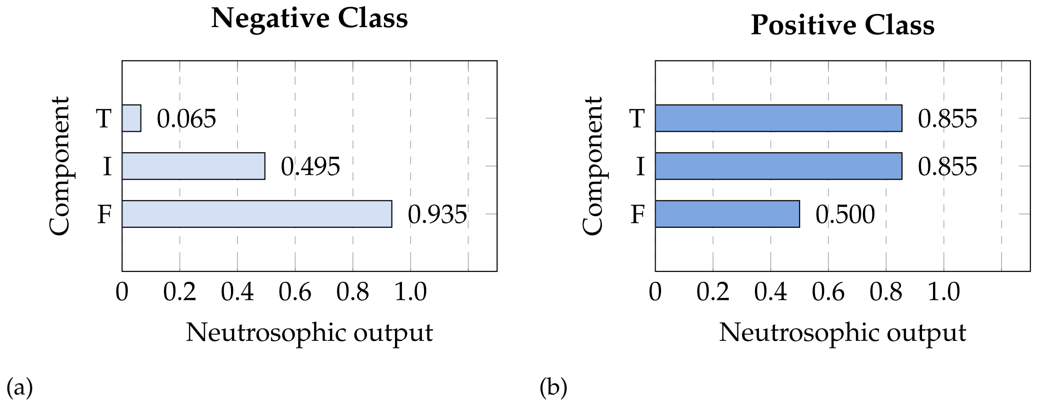


Figure 5.17: Titanic - KEEL Dataset: Graphical explanation of the overall NL sub-model - example of a TP case: #3_4_309. Overall explanation: "Based on the overall sub-model's output the data was more similar to Casualty (+) compared to Survivor (-)." (a) Explanation: "For the Survivor (-) class, the overall sub-model thinks the data has a higher falsity (0.935) relative to its truth (0.065). The relative indetermency is average (0.495).". (b) Explanation: "For the Casualty (+) class, the overall sub-model thinks the data has a higher falsity (0.505) relative to its truth (0.495). The relative indetermency is high (0.855)."

Contrary to the TN example for the breast cancer dataset, for this TP example, the MCDM sub-models are in agreement, as shown in Figures 5.16-5.17. The negative class had a significantly lower truth compared to the positive class, which led to the positive classification. Despite the high indeterminacy of the positive class, the positive class sub-model was still able to produce a truth value higher than the negative class sub-model.

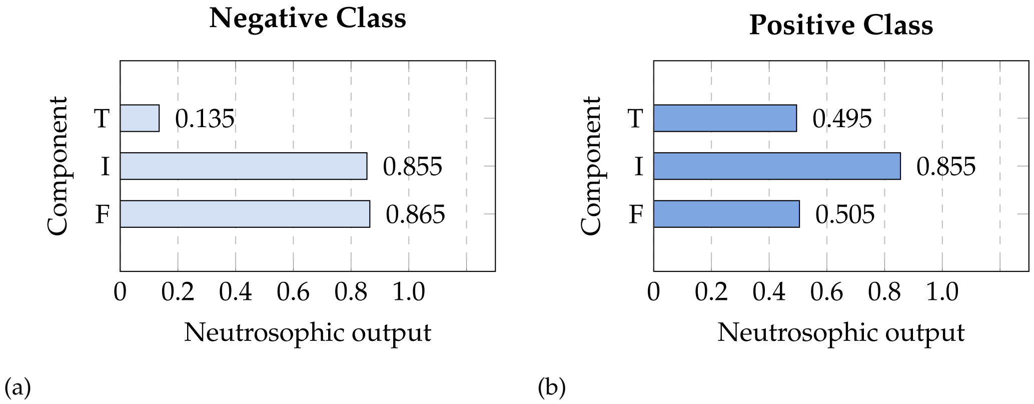


Figure 5.18: Titanic - KEEL Dataset: Graphical explanation of the overall NL sub-model - example of a FP case: #6_3_145. Overall explanation: "Based on the overall sub-model's output the data was more similar to Casualty (+) compared to Survivor (-)." (a) Explanation: "For the Survivor (-) class, the overall sub-model thinks the data has a higher falsity (0.865) relative to its truth (0.135). The relative indetermency is high (0.855).". (b) Explanation: "For the Casualty (+) class, the overall sub-model thinks the data has a higher truth (0.855) relative to its falsity (0.500). The relative indetermency is high (0.855)."

Moreover, for this second example, the model misclassified the data as positive. Looking at the neutrosophic output for the positive class reveals that the indeterminacy was high at 0.855, as shown in Figure 5.18. However, the lower truth level for the negative class caused the positive classification. The high indeterminacy in the data perceived by the negative class could be the cause of the low truth level.

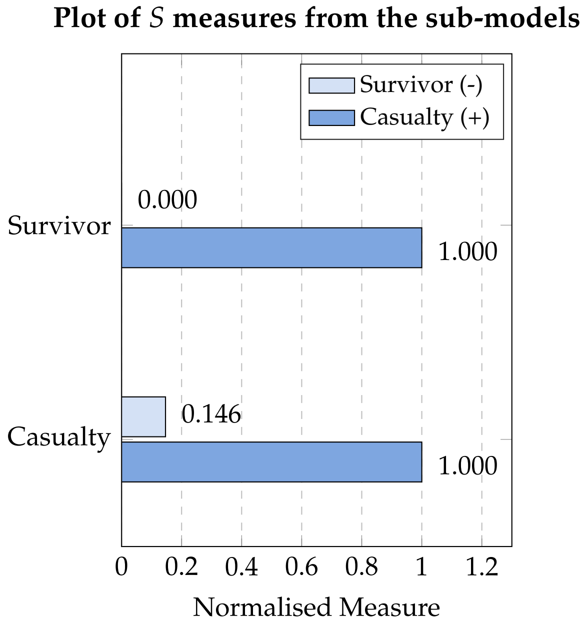
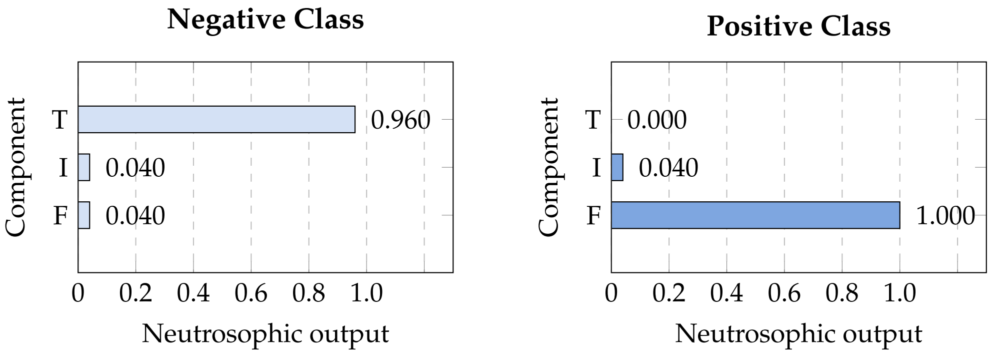


Figure 5.19: Titanic - KEEL Dataset: Illustration of the distance measures S from the two sub-models - example of a TN case: #2_2_128. Textual explanation: "Models are in agreement hence, it is more likely the classification is accurate. The data was classified as *Survivor (-)* by the model."



(a) (b)

Figure 5.20: Titanic - KEEL Dataset: Graphical explanation of the overall NL sub-model - example of a TN case: #2_2_128. Overall explanation: "Based on the overall sub-model's output the data was more similar to Survivor (-) compared to Casualty (+)." (a) Explanation: "For the Survivor (-) class, the overall sub-model thinks the data has a higher truth (0.960) relative to its falsity (0.040). The relative indetermency is low (0.040).". (b) Explanation: "For the Casualty (+) class, the overall sub-model thinks the data has a higher falsity (1.000) relative to its truth (0.000). The relative indetermency is low (0.040)."

The next example (Figures 5.19-5.20) showcases how the explanation is generated for a classification with high degree of certainty. The TN case resulted in truth and falsity values at opposite extremes for the NL sub-models, while indeterminacy maintained low values for both. The MCDM sub-models were in agreement; therefore, there is little to signify any uncertainty for this particular example.

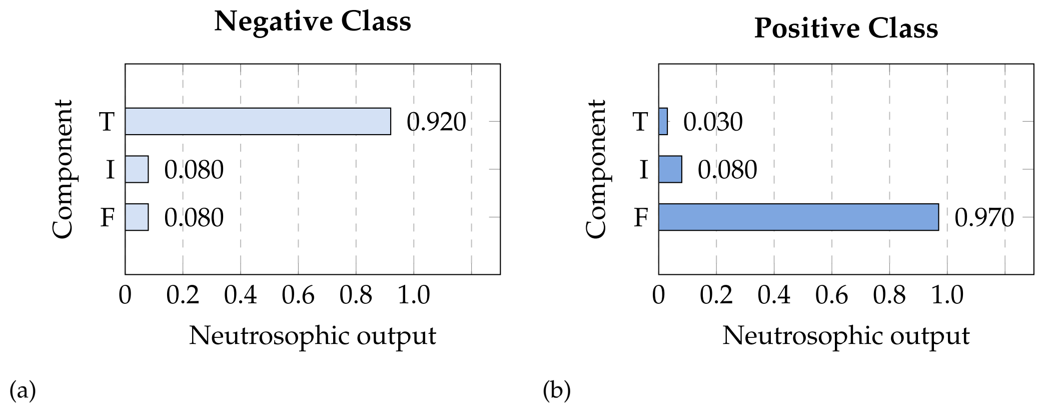
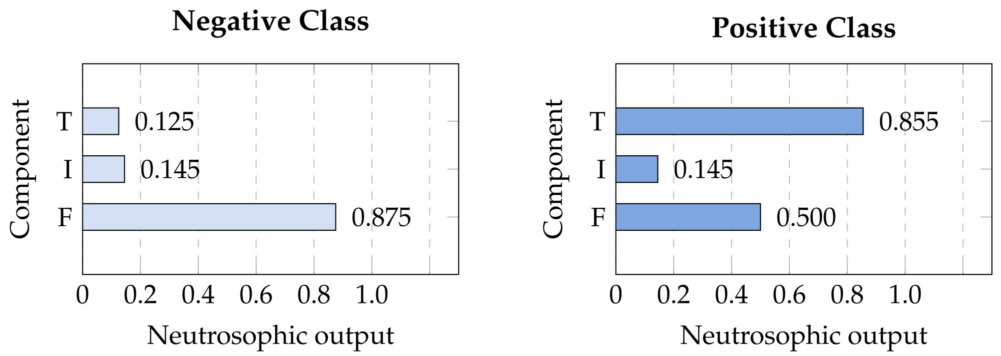


Figure 5.21: Titanic - KEEL Dataset: Graphical explanation of the overall NL sub-model - example of a FN case: #3_4_377. Overall explanation: "Based on the overall sub-model's output the data was more similar to Survivor (-) compared to Casualty (+)." (a) Explanation: "For the Survivor (-) class, the overall sub-model thinks the data has a higher truth (0.920) relative to its falsity (0.080). The relative indetermency is low (0.080).". (b) Explanation: "For the Casualty (+) class, the overall sub-model thinks the data has a higher falsity (0.970) relative to its truth (0.030). The relative indetermency is low (0.080)."

An issue arises for this FN example, where the same level of certainty is reflected in the neutrosophic output despite the inaccuracy of the model. This is an example where the explanation is misleading due to the classifier's inaccuracy, as shown in Figure 5.21.

Kaggle titanic

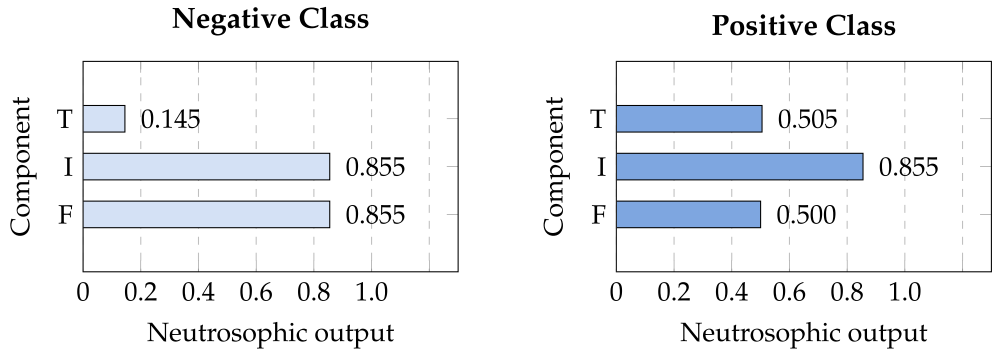
The aim of the Kaggle Titanic dataset is also to predict a passenger's survival; however, it uses seven features instead of three.



(a) (b)

Figure 5.22: Titanic - Kaggle Dataset: Graphical explanation of the overall NL sub-model - example of a TP case: #3_2_135. Overall explanation: "Based on the overall sub-model's output the data was more similar to Casualty (+) compared to Survivor (-)." (a) Explanation: "For the Survivor (-) class, the overall sub-model thinks the data has a higher falsity (0.875) relative to its truth (0.125). The relative indetermency is low (0.145).". (b) Explanation: "For the Casualty (+) class, the overall sub-model thinks the data has a higher truth (0.855) relative to its falsity (0.500). The relative indetermency is low (0.145)."

Similarly to the TP example for the KEEL dataset, the truth level is high at 0.855. However, the indeterminacy is significantly lower at 0.145 for both classes (Figure 5.22). This, combined with a low falsity, makes it a relatively certain positive classification.



(a) (b)

Figure 5.23: Titanic - Kaggle Dataset: Graphical explanation of the overall NL sub-model - example of a FP case: #4_3_38. Overall explanation: "Based on the overall sub-model's output the data was more similar to Casualty (+) compared to Survivor (-)." (a) Explanation: "For the Survivor (-) class, the overall sub-model thinks the data has a higher falsity (0.855) relative to its truth (0.145). The relative indetermency is high (0.855).". (b) Explanation: "For the Casualty (+) class, the overall sub-model thinks the data has a higher falsity (0.855) relative to its truth (0.145). The relative indetermency is high (0.855)."

By comparison, the FP example (Figure 5.23) had a lower truth for the positive class and a high indeterminacy. Both of which were indicative of less certainty in the classification.

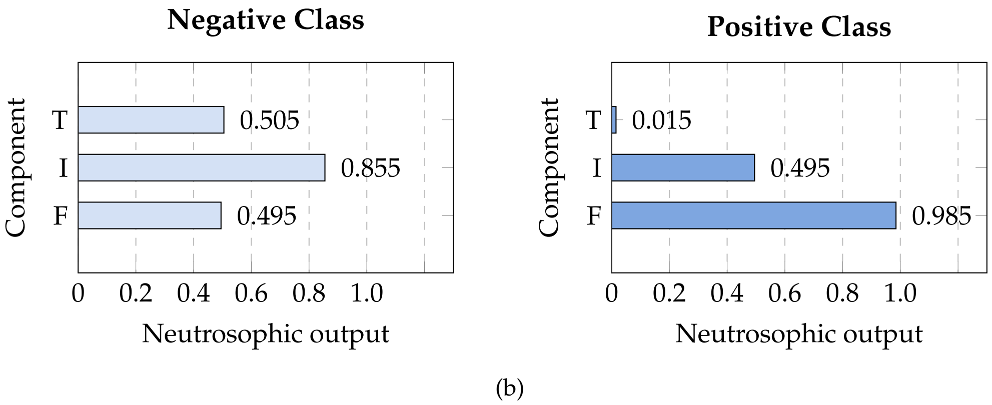


Figure 5.24: Titanic - Kaggle Dataset: Graphical explanation of the overall NL sub-model - example of a TN case: #5_4_108. Overall explanation: "Based on the overall sub-model's output the data was more similar to Survivor (-) compared to Casualty (+)." (a) Explanation: "For the Survivor (-) class, the overall sub-model thinks the data has a higher falsity (0.855) relative to its truth (0.145). The relative indetermency is high (0.855).". (b) Explanation: "For the Casualty (+) class, the overall sub-model thinks the data has a higher falsity (0.985) relative to its truth (0.015). The relative indetermency is average (0.495)."

The TN example for this (Figure 5.24) dataset was a less certain one because of the higher indeterminacy outputs for both classes compared to the KEEL dataset. In spite of this, the model was still able to accurately classify the passenger.

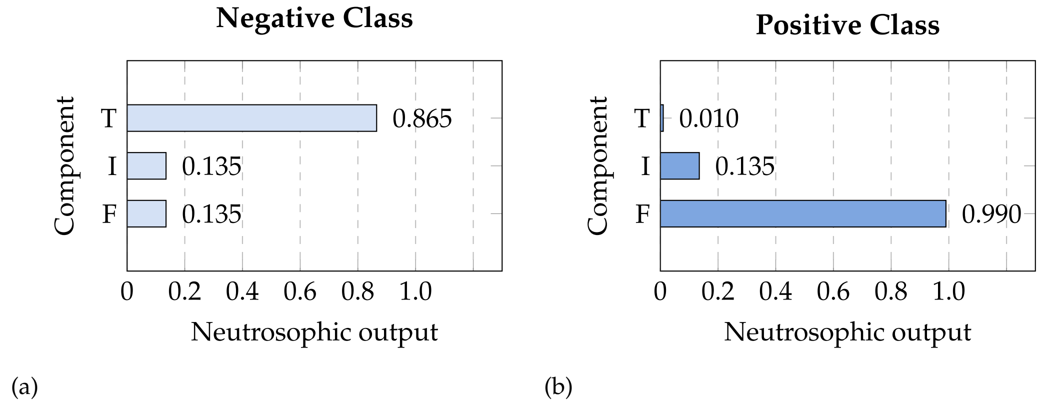


Figure 5.25: Titanic - Kaggle Dataset: Graphical explanation of the overall NL sub-model - example of a FN case: #6_1_175. Overall explanation: "Based on the overall sub-model's output the data was more similar to Survivor (-) compared to Casualty (+)." (a) Explanation: "For the Survivor (-) class, the overall sub-model thinks the data has a higher truth (0.865) relative to its falsity (0.135). The relative indetermency is low (0.135).". (b) Explanation: "For the Casualty (+) class, the overall sub-model thinks the data has a higher falsity (0.990) relative to its truth (0.010). The relative indetermency is low (0.135)."

On the other hand, for the FN example (Figure 5.25), the model misclassified the data despite the low indeterminacy. In addition, the neutrosophic outputs of the NL sub-model both pointed to the same conclusion. The MCDM model's inaccuracy resulted in the wrongful classification and misleading explanation, as seen with the FN example for the KEEL dataset. Therefore, the classifier's accuracy plays a vital role in the explanation's meaningfulness.

Parkinson disease

In this section, two cases will be compared to give an idea of whether NL output considerably varies for examples of the same type. Moreover, the explanation's *stability* is demonstrated.

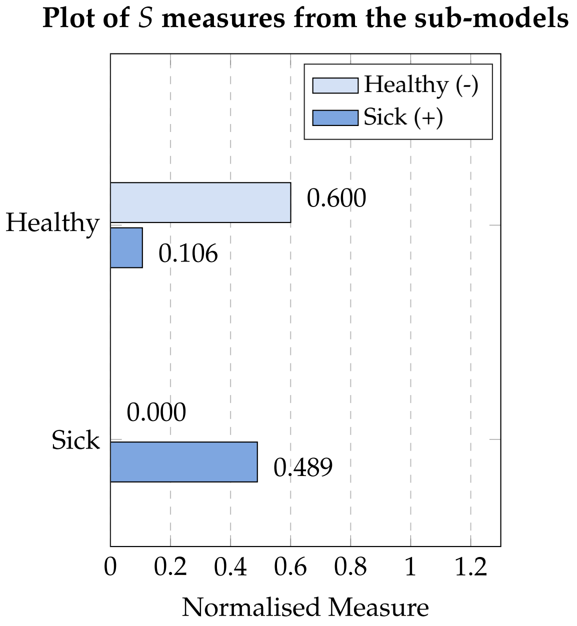


Figure 5.26: Parkinson Disease Dataset: distance measures S from the two sub-models - example of a TP case: #10_1_51. Textual explanation: “Models are in conflict hence, it is less likely the classification is accurate. The data was classified as *Sick (+)* by the model.”

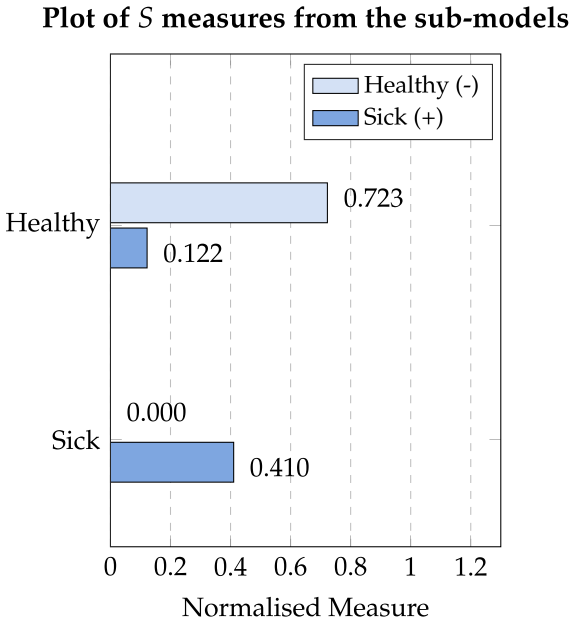
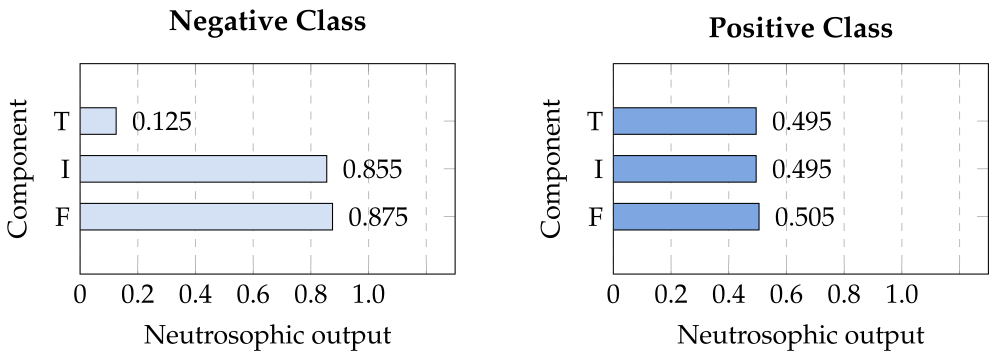


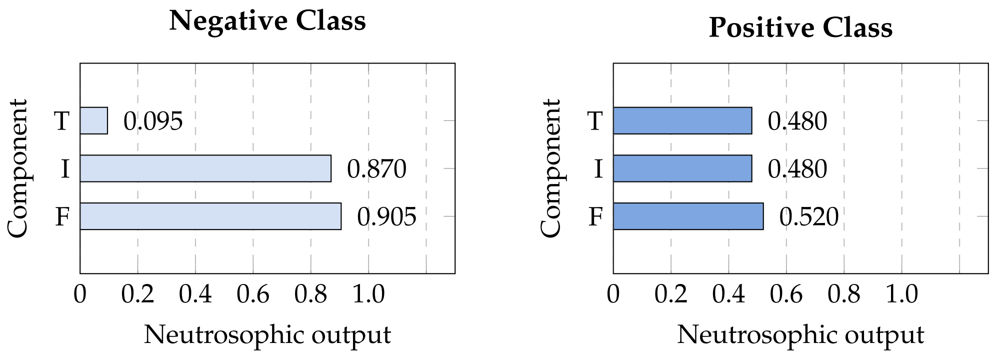
Figure 5.27: Parkinson Disease Dataset: distance measures S from the two sub-models - example of a TP case: #2_1_78. Textual explanation: “Models are in conflict hence, it is less likely the classification is accurate. The data was classified as *Sick (+)* by the model.”



(a)

(b)

Figure 5.28: Parkinson disease dataset: graphical explanation of the overall NL sub-model - example of a TP case: #10_1_51. Overall explanation: “Based on the overall sub-model’s output the data was more similar to Sick (+) compared to Healthy (-).” (a) Explanation: “For the Healthy (-) class, the overall sub-model thinks the data has a higher falsity (0.875) relative to its truth (0.125). The relative indetermency is high (0.855).”. (b) Explanation: “For the Sick (+) class, the overall sub-model thinks the data has a higher truth (0.855) relative to its falsity (0.500). The relative indetermency is average (0.495).”



(a)

(b)

Figure 5.29: Parkinson disease dataset: graphical explanation of the overall NL sub-model - example of a TP case: #2_1_78. Overall explanation: “Based on the overall sub-model’s output the data was more similar to Sick (+) compared to Healthy (-).” (a) Explanation: “For the Healthy (-) class, the overall sub-model thinks the data has a higher falsity (0.905) relative to its truth (0.095). The relative indetermency is high (0.870).”. (b) Explanation: “For the Sick (+) class, the overall sub-model thinks the data has a higher truth (0.870) relative to its falsity (0.500). The relative indetermency is average (0.480).”

The TP cases present an example where there was minimal variation in the NL outputs across the two subjects, as presented in Figures 5.28 and 5.29. This was due to the similar S measure values generated by the MCDM sub-model as shown in Figures 5.26 and 5.27.

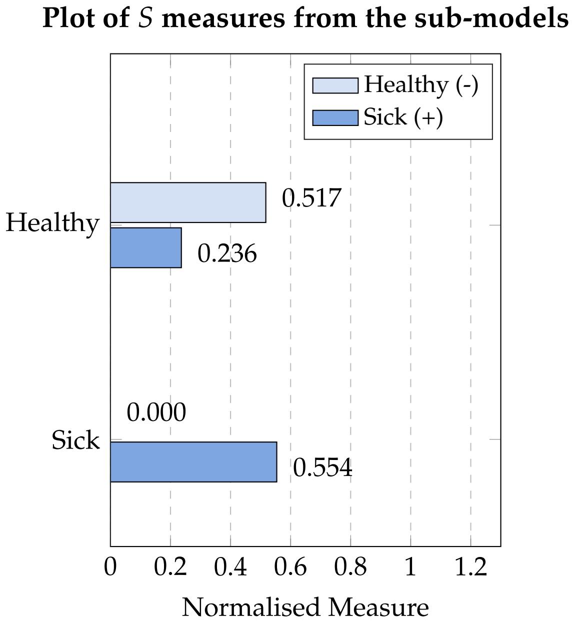


Figure 5.30: Parkinson Disease Dataset: distance measures S from the two sub-models - example of a FN case: #3_3_55. Textual explanation: "Models are in conflict hence, it is less likely the classification is accurate. The data was classified as *Healthy (-)* by the model."

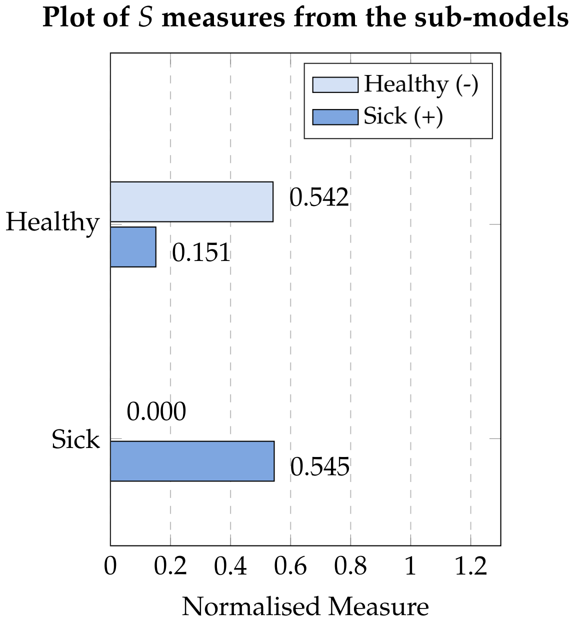
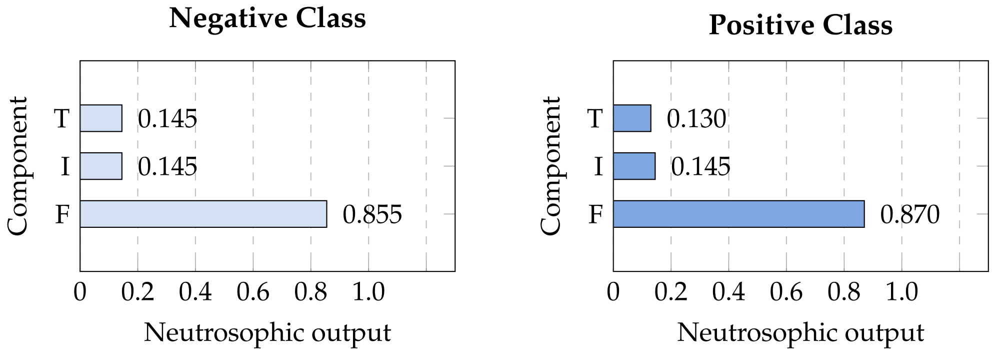
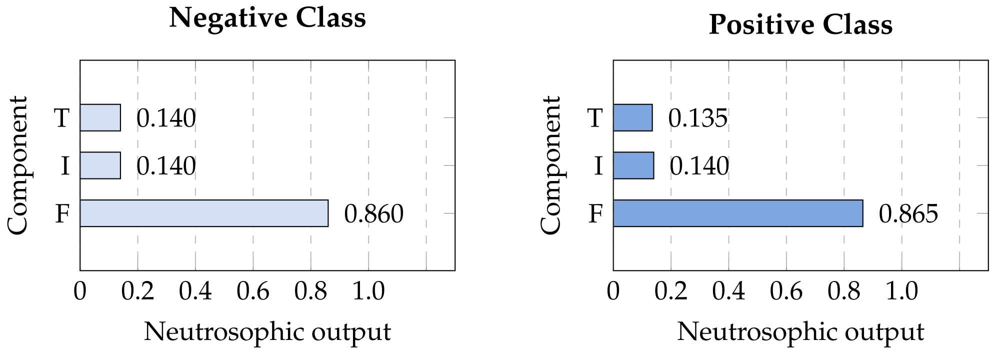


Figure 5.31: Parkinson Disease Dataset: distance measures S from the two sub-models - example of a FN case: #7_4_55. Textual explanation: "Models are in conflict hence, it is less likely the classification is accurate. The data was classified as *Healthy (-)* by the model."



(a) (b)

Figure 5.32: Parkinson disease dataset: graphical explanation of the overall NL sub-model - example of a FN case: #3_3_55. Overall explanation: “Based on the overall sub-model’s output the data was more similar to Healthy (-) compared to Sick (+).” (a) Explanation: “For the Healthy (-) class, the overall sub-model thinks the data has a higher falsity (0.855) relative to its truth (0.145). The relative indetermency is low (0.145).”. (b) Explanation: “For the Sick (+) class, the overall sub-model thinks the data has a higher falsity (0.870) relative to its truth (0.130). The relative indetermency is low (0.145).”

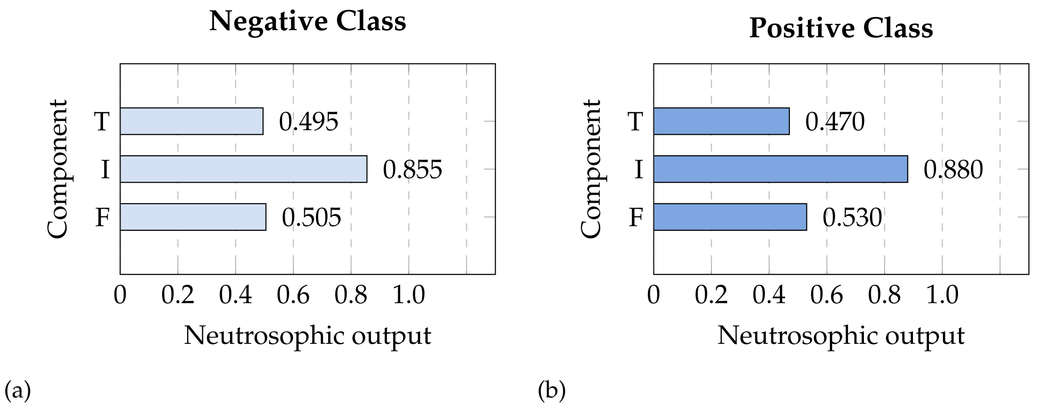


(a) (b)

Figure 5.33: Parkinson disease dataset: graphical explanation of the overall NL sub-model - example of a FN case: #7_4_55. Overall explanation: “Based on the overall sub-model’s output the data was more similar to Healthy (-) compared to Sick (+).” (a) Explanation: “For the Healthy (-) class, the overall sub-model thinks the data has a higher falsity (0.860) relative to its truth (0.140). The relative indetermency is low (0.140).”. (b) Explanation: “For the Sick (+) class, the overall sub-model thinks the data has a higher falsity (0.865) relative to its truth (0.135). The relative indetermency is low (0.140).”

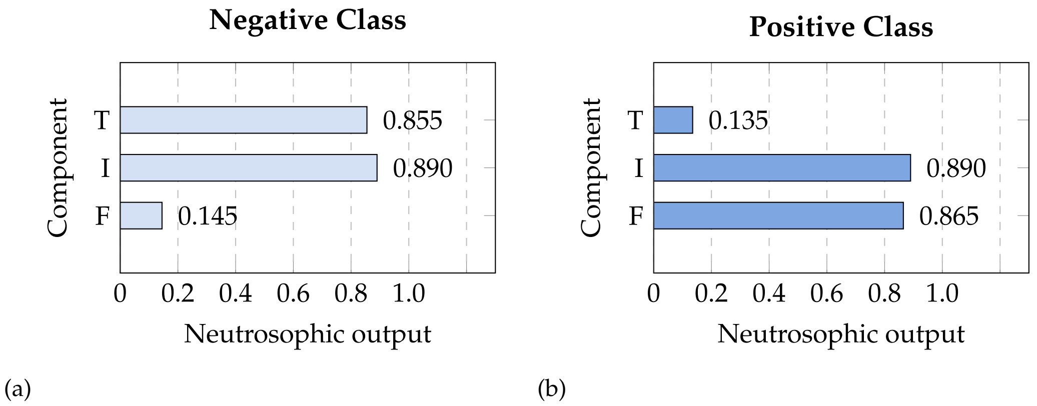
Furthermore, a similar image is portrayed for the FN examples where the NL had near identical values (Figures 5.32 and 5.33). Inspecting the S measures reveals once again this is not a mere generalisation, but, a proportional representation of the MCDM output, as shown in Figures 5.30 and 5.31.

Chess



(a) (b)

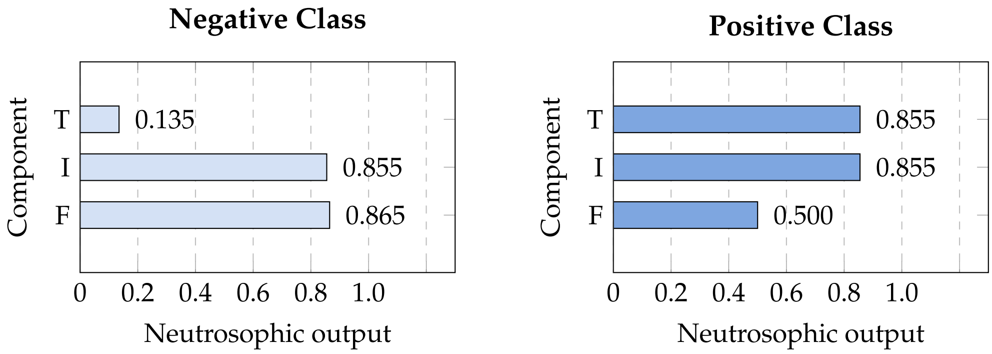
Figure 5.34: Chess dataset: graphical explanation of the overall NL sub-model - example of a TN case: #3_2_169. Overall explanation: “Based on the overall sub-model’s output the data was more similar to Win (-) compared to Lose (+).” (a) Explanation: “For the Win (-) class, the overall sub-model thinks the data has a higher falsity (0.505) relative to its truth (0.495). The relative indeterminacy is high (0.855).”. (b) Explanation: “For the Lose (+) class, the overall sub-model thinks the data has a higher falsity (0.530) relative to its truth (0.470). The relative indeterminacy is high (0.880).”



(a) (b)

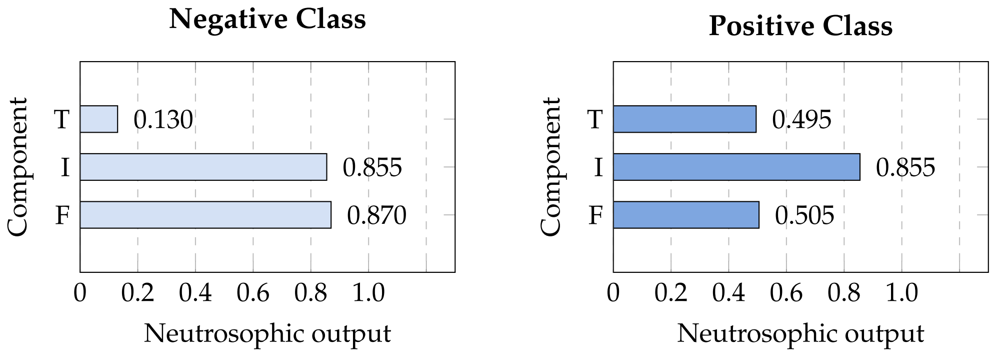
Figure 5.35: Chess dataset: graphical explanation of the overall NL sub-model - example of a TN case: #8_2_154. Overall explanation: “Based on the overall sub-model’s output the data was more similar to Win (-) compared to Lose (+).” (a) Explanation: “For the Win (-) class, the overall sub-model thinks the data has a higher truth (0.855) relative to its falsity (0.145). The relative indeterminacy is high (0.890).”. (b) Explanation: “For the Lose (+) class, the overall sub-model thinks the data has a higher falsity (0.865) relative to its truth (0.135). The relative indeterminacy is high (0.890).”

In contrast to the Parkinson disease dataset, the chess dataset resulted in distinct neutrosophic outputs for cases of the same type as presented in Figures 5.34-5.35. For the TN examples, four out of the six outputs differed. The second example had a higher certainty overall because of the higher truth for the negative class paired with a lower truth for the positive class.



(a) (b)

Figure 5.36: Chess dataset: graphical explanation of the overall NL sub-model - example of a FP case: #5_1_51. Overall explanation: "Based on the overall sub-model's output the data was more similar to Lose (+) compared to Win (-)." (a) Explanation: "For the Win (-) class, the overall sub-model thinks the data has a higher falsity (0.865) relative to its truth (0.135). The relative indeterminacy is high (0.855).". (b) Explanation: "For the Lose (+) class, the overall sub-model thinks the data has a higher truth (0.855) relative to its falsity (0.500). The relative indeterminacy is high (0.855)."



(a) (b)

Figure 5.37: Chess dataset: graphical explanation of the overall NL sub-model - example of a FP case: #5_3_200. Overall explanation: "Based on the overall sub-model's output the data was more similar to Lose (+) compared to Win (-)." (a) Explanation: "For the Win (-) class, the overall sub-model thinks the data has a higher falsity (0.870) relative to its truth (0.130). The relative indeterminacy is high (0.855).". (b) Explanation: "For the Lose (+) class, the overall sub-model thinks the data has a higher falsity (0.505) relative to its truth (0.495). The relative indeterminacy is high (0.855)."

However, only one output changed significantly for the FP examples (Figures 5.36-5.37). Therefore, confidence in the two decisions is considered relatively similar. Analysing TP examples revealed a similar pattern where outputs varied

Table 5.8: BF Bead: neut-TOPSIS vs fuzzy-TOPSIS

	$\mu \pm \sigma$ (%)			
Model	ACC	FMR	TPR	TNR
fuzzy-TOPSIS	81.3 ± 2.8	81.3 ± 2.7	81.1 ± 3.4	81.6 ± 4.0
neut-TOPSIS	79.6 ± 2.7	79.2 ± 2.6	84.6 ± 3.2	74.6 ± 3.9

Table 5.9: BF Defects: neut-TOPSIS vs fuzzy-TOPSIS

	$\mu \pm \sigma$ (%)			
Model	ACC	FMR	TPR	TNR
fuzzy-TOPSIS	87.7 ± 2.7	87.4 ± 2.7	93.2 ± 3.0	82.4 ± 4.5
neut-TOPSIS	87.6 ± 2.6	87.3 ± 2.7	93.3 ± 3.0	82.2 ± 4.6

across the same type; only for FN were the outputs nearly identical. A more in-depth analysis is required to discern whether the neutrosophic outputs provide the insight they represent. Nonetheless, it is imperative the values vary with less variability for similar cases - explanation *stability*.

5.4.3 Performance: weld datasets

In the spirit of assessing neut-TOPSIS's resilience and consistency as a data-driven classifier, the framework was investigated using the BF weld datasets: bead and defect detection. For a detailed description of the industrial case study, please refer to chapter 6.

For the bead detection (Table 5.8), the fuzzy-TOPSIS model still had better performance than neut-TOPSIS; the former achieved 0.7% and 2.1% higher ACC and FMR, respectively. This was due to the drop in TNR performance for neut-TOPSIS.

Performance for the defect dataset across the two frameworks was close to identical, with not more than a 0.1% difference for all four metrics (Table 5.9). Therefore, the absence of a performance trade-off makes neut-TOPSIS the obvious choice for this dataset.

5.4.4 Explaining: welding datasets

In this section, examples of explanation are presented for bead and defect detection datasets. Where an incomplete set of figures is shown for the graphical explanation examples, a full set is provided in Appendix B.

Bead detection

Explanation examples for these datasets include the ultrasonic image. The aim of this dataset is to classify images as either containing a bead or not. A bead is a protruding structure along the weld seam. Locating the bead indication enables NDT analysts to have an idea of the aligned of the weld in the ultrasonic image. The bead indication often appears as a high amplitude almost-horizontal ribbon.

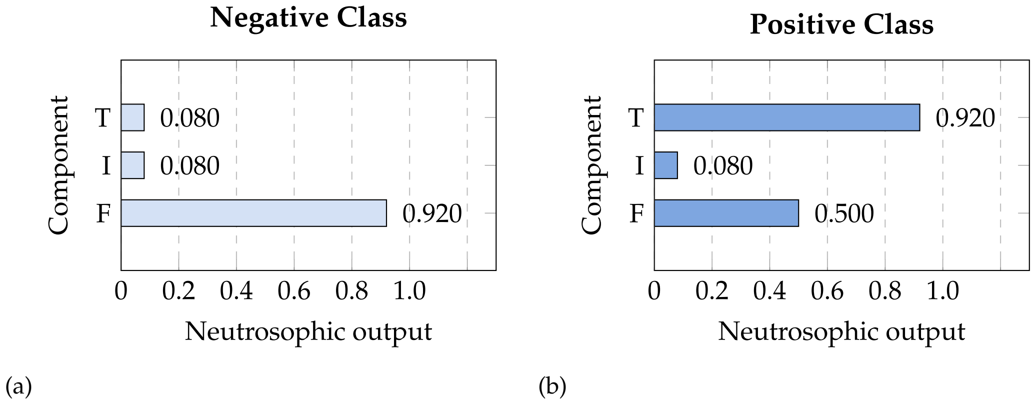
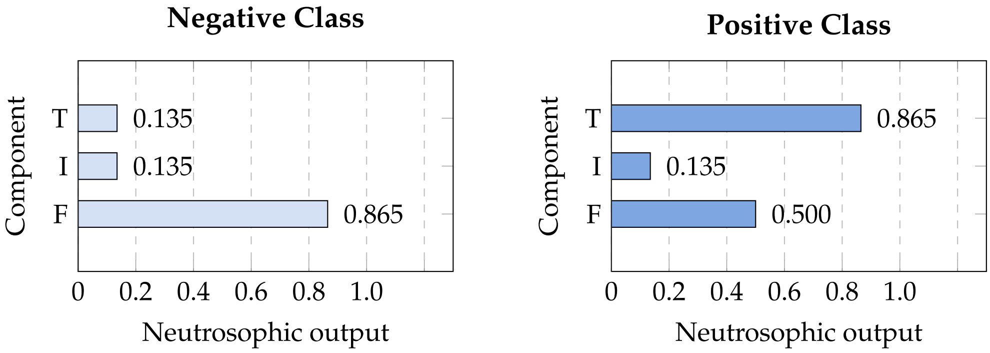
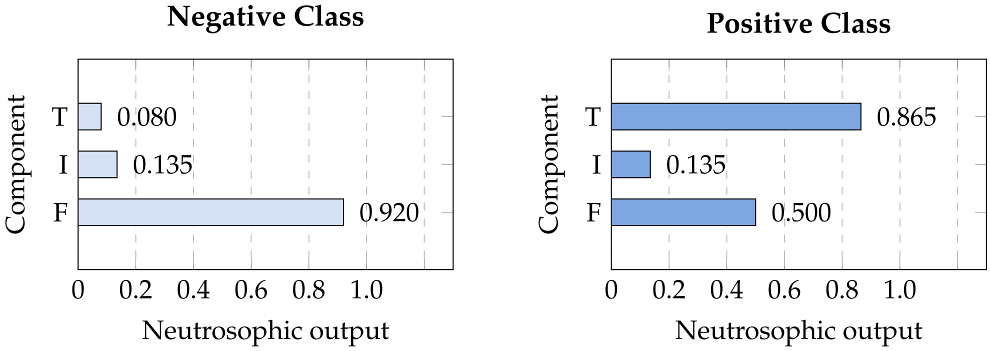


Figure 5.38: BF weld bead detection: Graphical explanation of the *No Bead* NL sub-model - example of a TP case: #1_5_175_23_45. (a) Explanation: "For the No bead (-) class, the No Bead Model thinks the data has a higher falsity (0.920) relative to its truth (0.080). The relative indetermency is low (0.080).". (b) Explanation: "For the Bead (+) class, the No Bead Model thinks the data has a higher truth (0.920) relative to its falsity (0.500). The relative indetermency is low (0.080)."



(a) (b)

Figure 5.39: BF weld bead detection: Graphical explanation of the *Bead* NL sub-model - example of a TP case: #1_5_175_23_45. (a) Explanation: "For the No bead (-) class, the No Bead Model thinks the data has a higher falsity (0.865) relative to its truth (0.135). The relative indetermency is low (0.135).". (b) Explanation: "For the Bead (+) class, the No Bead Model thinks the data has a higher truth (0.865) relative to its falsity (0.500). The relative indetermency is low (0.135)."



(a) (b)

Figure 5.40: BF weld bead detection: Graphical explanation of the overall NL sub-model - example of a TP case: #1_5_175_23_45. Overall explanation: "Based on the overall sub-model's output the data was more similar to Bead (+) compared to No bead (-)." (a) Explanation: "For the No bead (-) class, the overall sub-model thinks the data has a higher falsity (0.920) relative to its truth (0.080). The relative indetermency is low (0.135).". (b) Explanation: "For the Bead (+) class, the overall sub-model thinks the data has a higher truth (0.865) relative to its falsity (0.500). The relative indetermency is low (0.135)."

In the first example, a TP case, an extended version of the NL explanation is presented where the neutrosophic outputs are illustrated for the sub-models before (Figures 5.38-5.39) and after aggregation (Figure 5.40). The No Bead model (Figure 5.38) output shows a case where the truth is higher for the Bead class. Similarly, the Bead sub-model has the same result (Figure 5.39). As a result, following aggregation, the Bead had a higher truth overall and resulted in a strong classification.

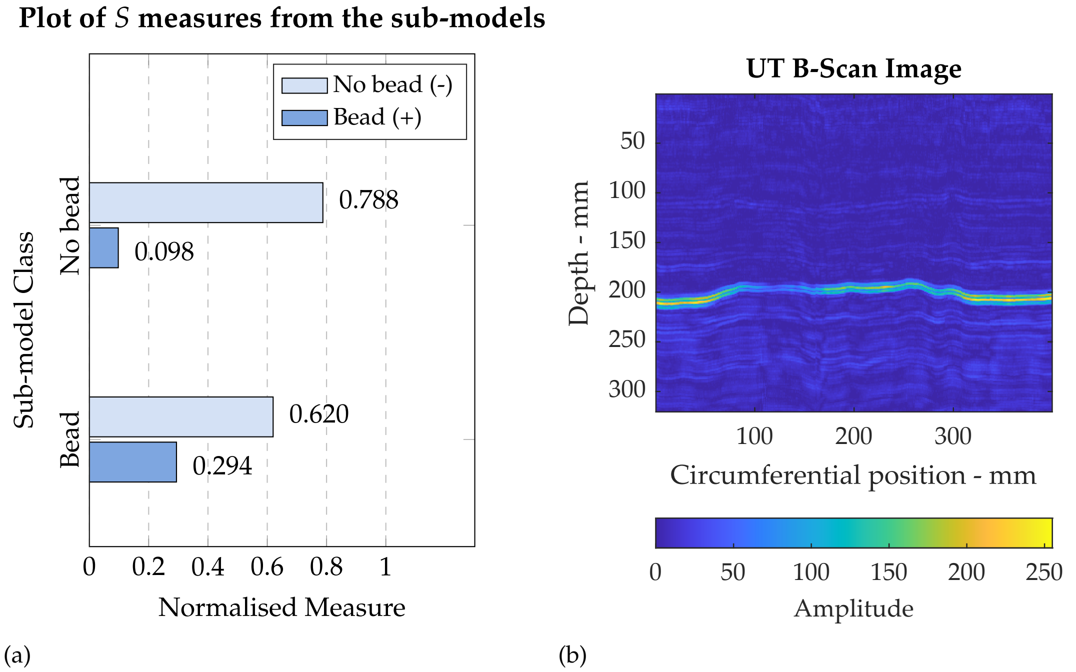
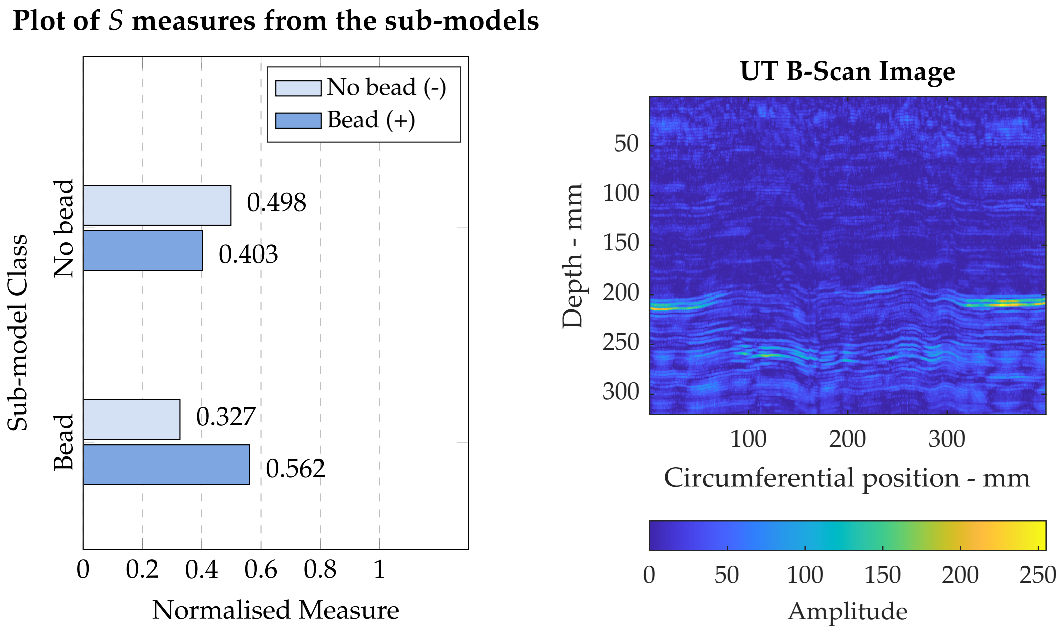


Figure 5.41: BF weld bead detection: (a) Illustration of the distance measures S from the two sub-models - example of a TP case: #1_5_175_23_45. Textual explanation: “Models are in agreement hence, it is more likely the classification is accurate.”. (b) Plot of UT image data for the same example. The image was classified as *Bead* by the model.

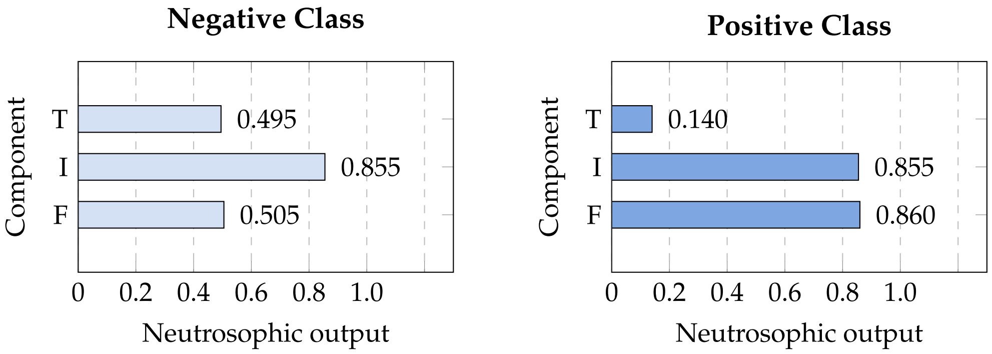
The bead is ribbon-like indication with a horizontal shape. An example of a bead indication is shown in Figure 5.41.



(a)

(b)

Figure 5.42: BF weld bead detection: (a) Illustration of the distance measures S from the two sub-models - example of a FN case: #1_2_176_23_35. Textual explanation: "Models are in conflict hence, it is less likely the classification is accurate.". (b) Plot of UT image data for the same example. The image was classified as *No Bead* by the model.



(a)

(b)

Figure 5.43: BF weld bead detection: Graphical explanation of the overall NL sub-model - example of a FN case: #1_2_176_23_35. Overall explanation: "Based on the overall sub-model's output the data was more similar to No bead (-) compared to Bead (+)." (a) Explanation: "For the No bead (-) class, the overall sub-model thinks the data has a higher falsity (0.505) relative to its truth (0.495). The relative indetermency is high (0.855).". (b) Explanation: "For the Bead (+) class, the overall sub-model thinks the data has a higher falsity (0.860) relative to its truth (0.140). The relative indetermency is high (0.855)."

In this FN example, looking at the image reveals a faint bead signal around the

200mm depth (Figure 5.42). The amplitude of the indication makes the Bead harder to discern. The MCDM sub-models were in conflict with the No Bead sub-model pointing towards Bead. Meanwhile, the Bead sub-model was pointing towards No Bead. The conflict was resolved by using the sub-model with the larger difference between its measures - the Bead sub-model. Hence, the No Bead decision, as shown in Figures 5.42-5.43.

Defect detection

Defect indications enable the detection of flaws in a pipe weld. In this section, the neut-TOPSIS model is used to classify image slices for defect detection.

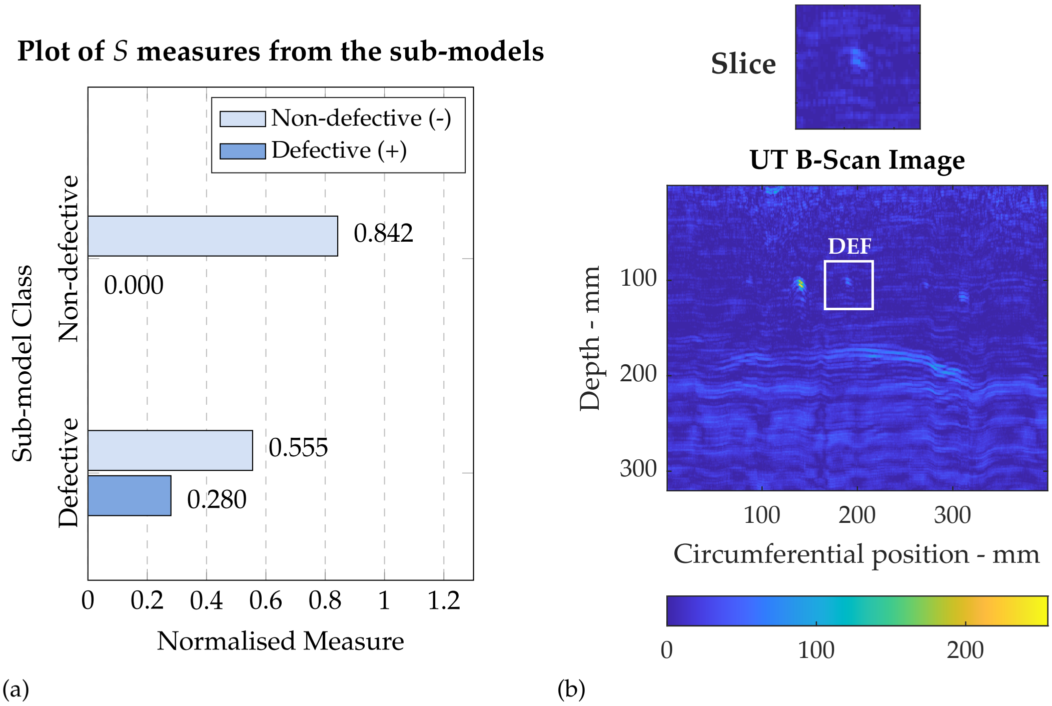


Figure 5.44: BF weld defect detection: (a) Illustration of the distance measures S from the two sub-models - example of a TP case: #2_4_114_21_58. Textual explanation: “Models are in agreement hence, it is more likely the classification is accurate.”. (b) Plot of UT image data for the same example. The image was classified as *defective* by the model.

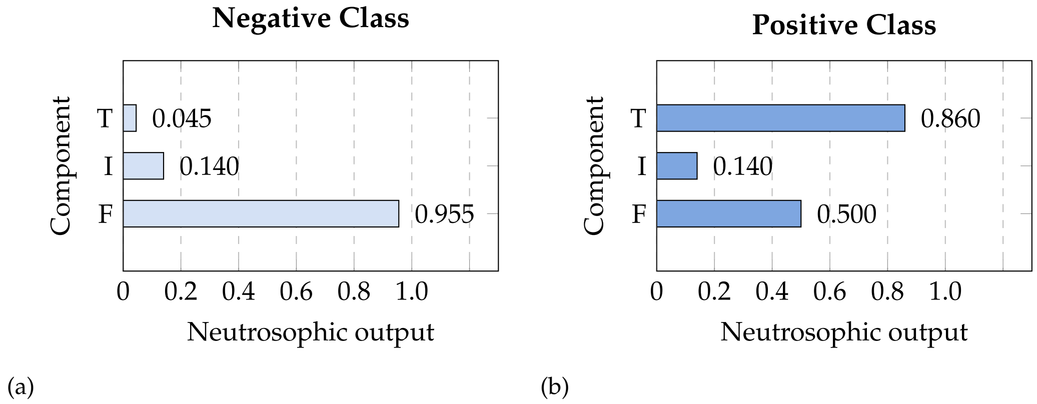
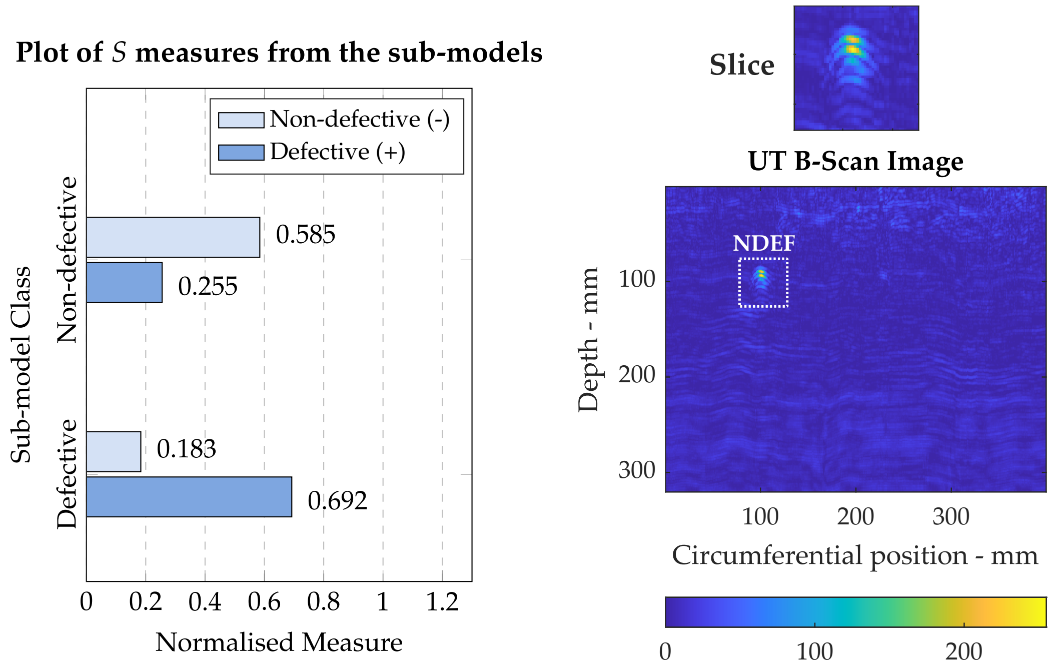


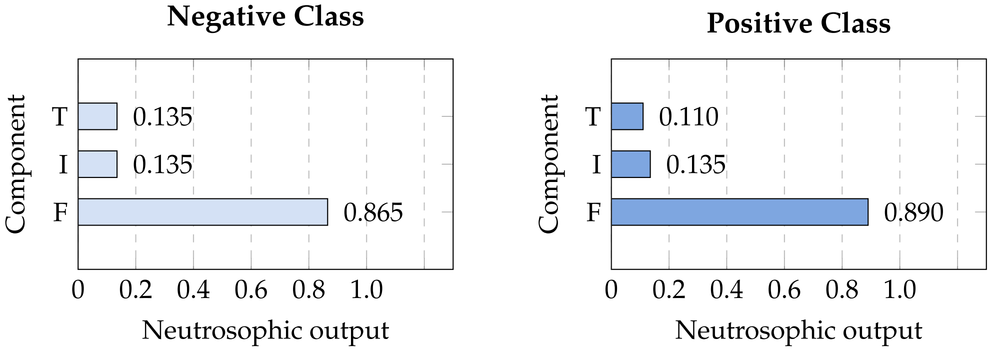
Figure 5.45: BF weld defect detection: Graphical explanation of the overall NL sub-model - example of a TP case: #2_4_114_21_58. Overall explanation: “Based on the overall sub-model’s output the data was more similar to Defective (+) compared to Non-defective (-).” (a) Explanation: “For the Non-defective (-) class, the overall sub-model thinks the data has a higher falsity (0.955) relative to its truth (0.045). The relative indetermency is low (0.140).”. (b) Explanation: “For the Defective (+) class, the overall sub-model thinks the data has a higher truth (0.860) relative to its falsity (0.500). The relative indetermency is low (0.140).”

Firstly, a TP example (Figures 5.44-5.45) demonstrates a case where the model successfully detected a defect. In this case, both NL sub-models had a high truth for the positive class and a low falsify for the negative class. Similarly, the overall NL output reflected the same characteristics; hence it is considered a high certainty positive classification.



(a) (b)

Figure 5.46: BF weld defect detection: (a) Illustration of the distance measures S from the two sub-models - example of a FN case: #8_5_107_19_61. Textual explanation: "Models are in conflict hence, it is less likely the classification is accurate.". (b) Plot of UT image data for the same example. The image was classified as *non-defective* by the model.



(a) (b)

Figure 5.47: BF weld defect detection: Graphical explanation of the overall NL sub-model - example of a FN case: #8_5_107_19_61. Overall explanation: "Based on the overall sub-model's output the data was more similar to Non-defective (-) compared to Defective (+)." (a) Explanation: "For the Non-defective (-) class, the overall sub-model thinks the data has a higher falsity (0.865) relative to its truth (0.135). The relative indetermency is low (0.135).". (b) Explanation: "For the Defective (+) class, the overall sub-model thinks the data has a higher falsity (0.890) relative to its truth (0.110). The relative indetermency is low (0.135)."

For the FN case, however, overall truth levels for both classes were low (Figure

5.47). Thus, compared to the previous example, it is indicatively, based on the explanation, less certain. Therefore, this insight could be useful for experts in discerning misclassification when presented with the image (Figure 5.46).

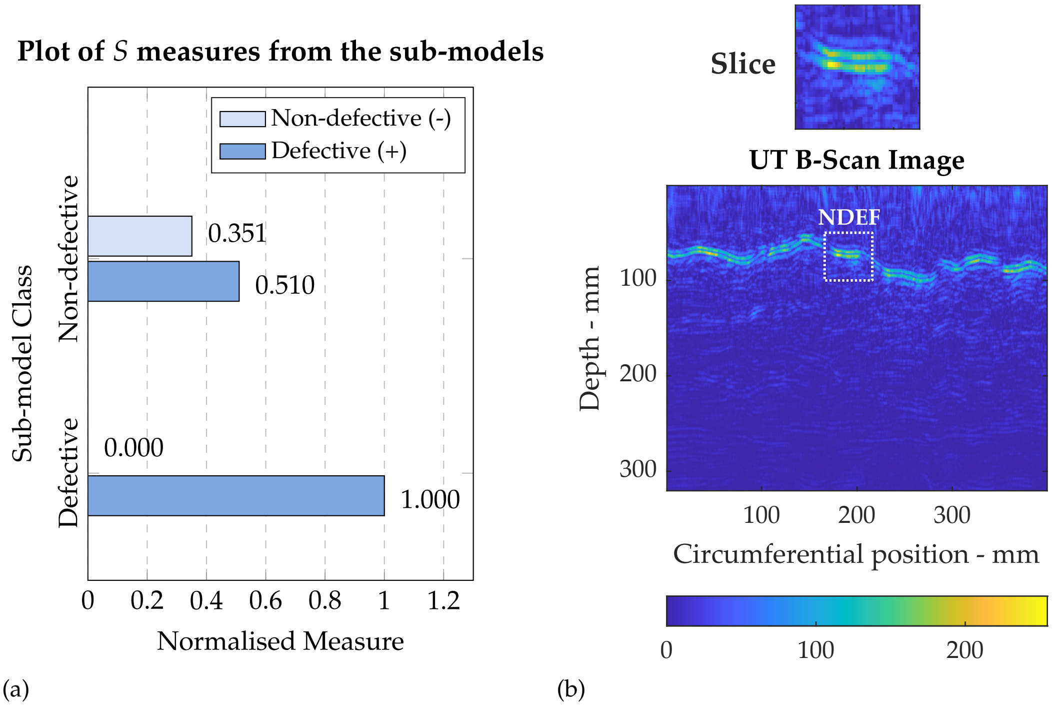


Figure 5.48: BF weld defect detection: (a) Illustration of the distance measures S from the two sub-models - example of a TN case: #5_2_25_3_8. Textual explanation: "Models are in agreement hence, it is more likely the classification is accurate.". (b) Plot of UT image data for the same example. The image was classified as *non-defective* by the model.

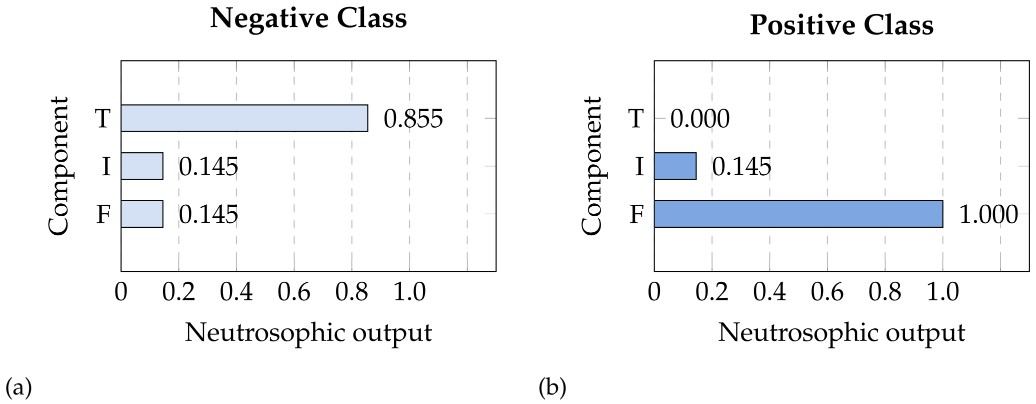
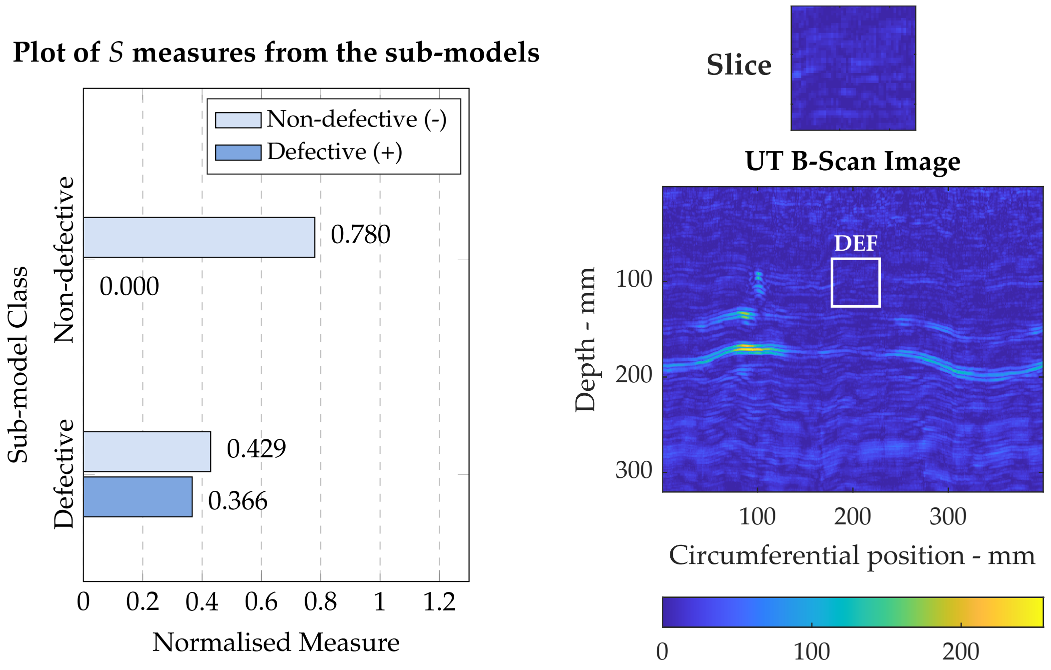


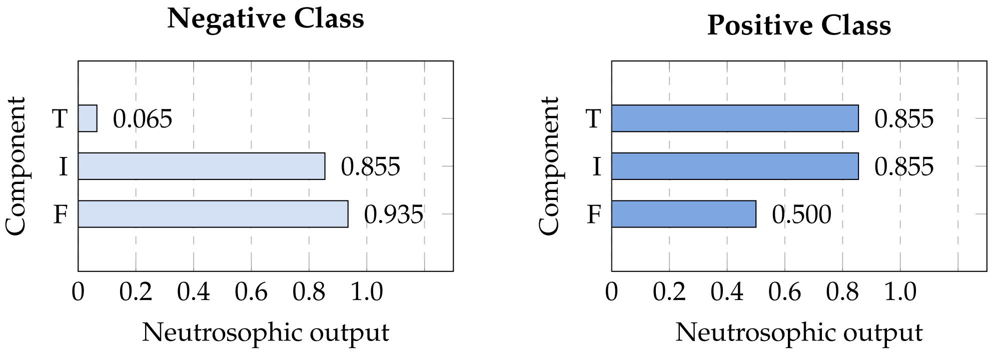
Figure 5.49: BF weld defect detection: Graphical explanation of the overall NL sub-model - example of a TN case: #5_2_25_3_8. Overall explanation: "Based on the overall sub-model's output the data was more similar to Non-defective (-) compared to Defective (+)." (a) Explanation: "For the Non-defective (-) class, the overall sub-model thinks the data has a higher truth (0.855) relative to its falsity (0.145). The relative indetermency is low (0.145).". (b) Explanation: "For the Defective (+) class, the overall sub-model thinks the data has a higher falsity (1.000) relative to its truth (0.000). The relative indetermency is low (0.145)."

In this TN case (Figures 5.48-5.49), the high amplitude indications were expected to confuse the classifier into thinking there is a defect. Although the non-defective sub-model was relatively inconclusive with S measures that are close in value, the defective sub-model could not be more conclusive with a truth level of 1.0 for the negative class.



(a) (b)

Figure 5.50: BF weld defect detection: (a) Illustration of the distance measures S from the two sub-models - example of a FP case: #7_1_40_19_49. Textual explanation: “Models are in agreement hence, it is more likely the classification is accurate.”. (b) Plot of UT image data for the same example. The image was classified as *defective* by the model.



(a) (b)

Figure 5.51: BF weld defect detection: Graphical explanation of the overall NL sub-model - example of a FP case: #7_1_40_19_49. Overall explanation: “Based on the overall sub-model’s output the data was more similar to Defective (+) compared to Non-defective (-).” (a) Explanation: “For the Non-defective (-) class, the overall sub-model thinks the data has a higher falsity (0.935) relative to its truth (0.065). The relative indetermency is high (0.855).”. (b) Explanation: “For the Defective (+) class, the overall sub-model thinks the data has a higher truth (0.855) relative to its falsity (0.500). The relative indetermency is high (0.855).”

Meanwhile, the FP case (Figures 5.50-5.51) had a less conclusive classification with indeterminacy for both sub-models at 0.855. Despite the high truth for the positive class, the falsity was average at 0.5. Also, the negative class had a low truth level and high falsity; thus, the only indication of a potential misclassification was indeterminacy. The UT image slice is considered relatively clean and has minimal if any, signs of defect indication. Therefore, this weld is likely to be picked up as inaccurate by an industry expert. It is worth noting FPs as more tolerable than FNs in the application of defect detection because of the associated consequence severity.

5.5 Summary

In this chapter, neutrosphic-TOPSIS is proposed as an improvement to fuzzy-TOPSIS. The new framework provides additional explanations based on its falsity and indeterminacy components. Despite enhanced interpretability, performance stayed largely the same for all datasets tested.

Explanation examples were presented for seven datasets. For the fuzzy-TOPSIS classifier, the explanation framework illustrated the impact of each feature on each sub-model separately. The approach presented a granular representation of impact down to the features - an explanation offering a lower level abstraction. This was useful in tracing how a decision was arrived at based on the feature values.

The Neutrosophic-TOPSIS classifier, on the other hand, summarises the state of the sub-models using its neutrosophic components: truth, indeterminacy and falsity. As a result, the user can be presented with a high-level explanation where each component is explained using the three neutrosophic components. The small number of components involved makes it easier for the user to comprehend the state of the sub-models and how they affected the overall model's decision.

Since the TOPSIS component remains unchanged, the feature-level explanations can still be provided in addition to the Neutrosophic explanations.

The NL extension expanded upon the FIS in the previous chapter by providing three components. This increased the number of insights presented to the user. As a result, the complexity of the graphs presented to the user was increased; thus, this potentially has a negative effect on the comprehensibility of the explanation.

6 Application to nondestructive testing: a pipe inspection case study

The omission of explanation from AI models is a key obstacle to their widespread adoption. Attempts have been made to solve this issue by trying to explain opaque models. Although some success has been sought, opaque models limit what is achievable since they infer insight from external models not involved in the classification process. Therefore, opting for intrinsically interpretable models is the only way to access direct explanation. In this chapter, ultrasonic testing data gathered from an industrial project on volumetric pipe inspection is processed to extract a nine-feature dataset. Consequently, an inherently explainable framework, based on Fuzzy-MCDM methodologies, is applied to the dataset and shown to achieve satisfactory performance compared to K-Means. The explanation framework is able to pinpoint the most influential features for each of the decisions in a traceable fashion. The study points to the framework's versatility and its ability to provide satisfactory performance while generating meaningful explanations aligned with the classifier's inner workings. In addition, it alludes to the possibility in the future of naturally interpretable approaches.

6.1 Introduction

The implementation of decision support systems in real-life applications depends on a host of variables. Obstacles to their adoption vary; however, a common one is the lack of transparency, traceability and understandability of such systems, all of which are related to explainability. Industrial decision support systems are no exception.

In this chapter, an application of explainable modelling is presented, using the fuzzy-TOPSIS framework, to ultrasonic pipe inspection. The framework is based on the fuzzy-TOPSIS classifier first proposed in chapter 3. In the first section, a literature review is described on advanced manufacturing and then specifically pipe inspection. Consequently, a description of the steps applied to prepare the data; pre-processing and labelling. Finally, the chapter is concluded with a presentation of the results, discussion and suggestions for future work.

6.2 Pipe inspection in advanced manufacturing

The industrial revolutions paved the way for dramatic improvements to how *manufacturing* is performed. Steam engines pumped early mechanically automated mass-producing machines at the start of the first revolution. Consequently, the second revolution brought about new inventions and technologies that increased the efficiency of how mass production is performed. Moreover, the invention of the modern steam turbine provided access to cheap electricity, which paved the way for the electrification of numerous applications. The development of electronics and, in turn, integrated circuits saw the start of a new era in industrialisation - the digital age. Although mechanical and analogue electronics enabled automation to a certain extent, digitisation made possible the emergence of several precise manufacturing processes. The boost in controllability meant highly repeatable tasks could be optimised to a level not possible with previous control technologies. For instance, the emergence of servo motors allowed for highly precise manufacturing technologies based on robotics, such as friction stir welding, additive manufacturing and robotic manipulators.

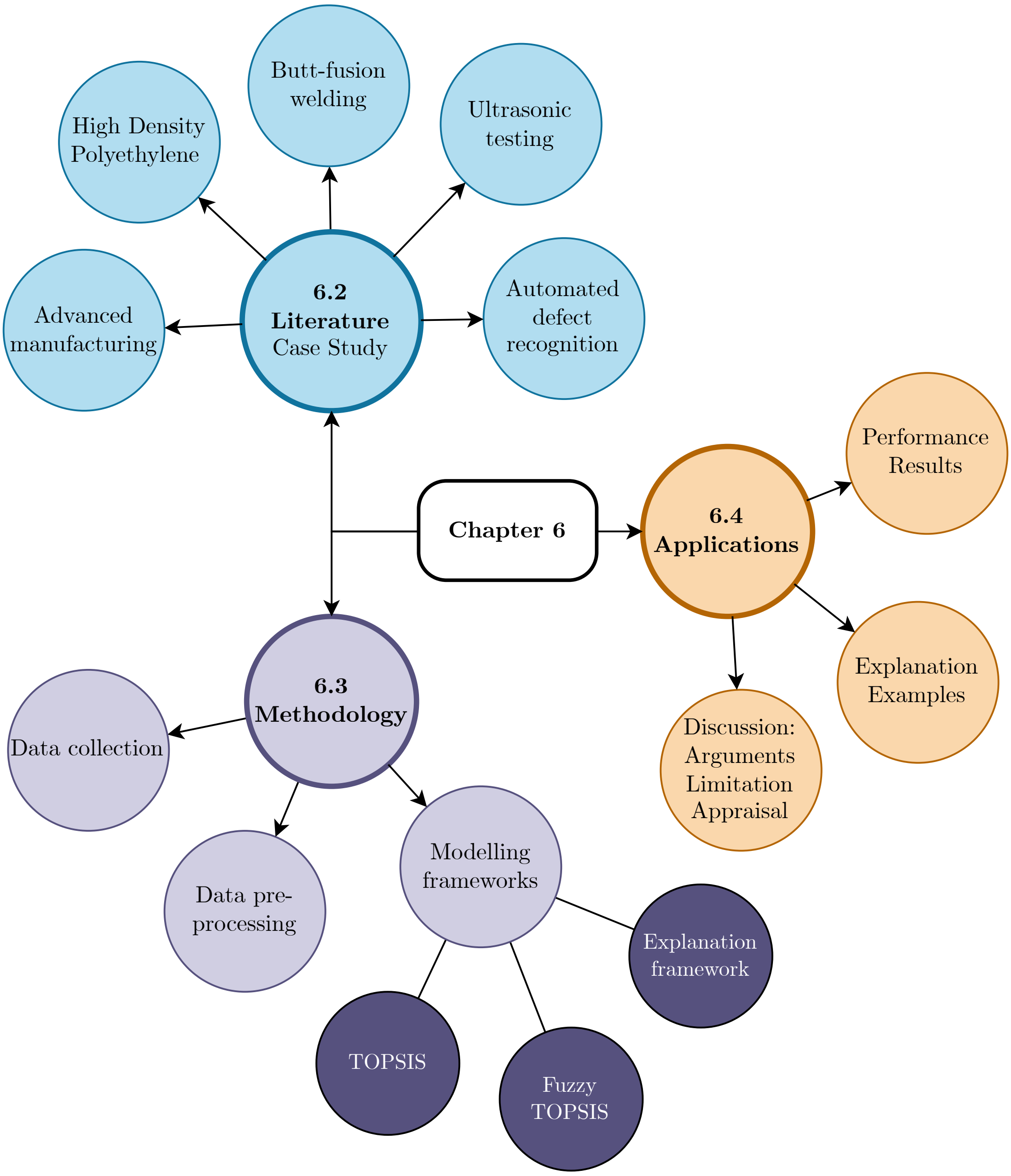


Figure 6.1: Chapter 6: mind map of general topics and concepts

The rise of automation, although vastly rewarding, came with its challenges. Despite the unimaginable level of accuracy achievable by robust control, manufacturing is everything but ideal. The increasing levels of repeatable success achieved come at a cost; less data is generated on defects because of the high rates of success. Thus, machine learning models lack the datasets required to achieve satisfactory accuracy performance. Nonetheless, Industry 4.0 initiatives are at the forefront of solving this problem. By connecting all components of the factory chain to the cloud, more data than ever before is accessible by model designers.

Buttfusion (BF) and electro-fusion (EF) welding are the most prominent thermo-fusion joining techniques used for HDPE pipe welding in industrial installations around the world. EF welding uses an external HDPE sleeve with a coil embedded; it is excited with a current that generates enough heat to mould the edges of the pipes to the sleeve and, in turn, together. In contrast, BF does not utilise any additional fillers or components; rather, it relies solely on the material of pipes. This is achieved by heating the pipe's edges and pushing them together at a pre-defined pressure. The softened edges bead up to form a tight seal once cooled.

Steel pipes have been the preferred option for critical applications for their superior tensile strength. However, HDPE provides resistance to corrosion steel cannot match among other benefits, such as strength-to-weight ratio, electrical insulation, and longer service life [114], [115]. Thus, HDPE pipes have been applied in critical areas such as water, gas and nuclear [116]; This fosters new challenges to the safety of their use. Efforts have been undertaken to optimise the welding process, such as an established procedure erected through an international standard, ISO 21307. The standard seeks to ensure the safety of welds produced. Moreover, a key requirement in producing HDPE welds is the Quality Assurance (QA) process, where NDE techniques have been designed to use PAUT.

QA is one of the main priorities of any mass-producing factory, as it goes without saying that by assuring quality, efficiency is improved. HDPE pipe welding is an area where QA is particularly vital. Experts estimate there are 700 million plastic welds in operation around the globe in various industries such as gas, mining and energy [116]. Critical industries' reliance on HDPE pipelines means defects have to be treated very seriously to ensure their safety. For this reason, there has been an industry interest in developing volumetric inspection techniques for HDPE welds [117]. The advancements in this field gave rise to a common challenge; the high dimensionality of the data generated. This makes the NDT system's current reliance on manual analysis unsustainable. As a result, ML modelling could prove useful in automating this process.

However, the introduction of modelling in any system often results in discrepancies and inaccuracies; there is minimal room for error in a high stake area such as this. In addition, the NDT expert must still be the entity to take the decision because of the absence of regulation for complete dependence on decision support systems. For this reason, explaining the classification result could prove useful for the NDT expert's task of *justifying* decisions.

There is a lack of literature that demonstrates modelling being used for automating the NDT process for plastic pipe welds [118]. Thus, there is a clear research gap that needs to be addressed. This chapter addresses the gap by demonstrating, for the first time ever, an implementation of an explainable classifier for BF plastic pipe weld defects.

6.3 Applying Fuzzy-TOPSIS classifiers to automated defect recognition

6.3.1 Data collection and conversion

UT as described in the previous section, entails a variety of distinct techniques and methodologies. PAUT is a form of UT where the scanner consists of an *array* of transducers, as opposed to a single transducer used in conventional UT systems.

In the case of UT for pipe inspection, the use of PAUT is widely used; however, the probe configuration is varied to suit the weld type. In the case of BF welds, an angle beam is favoured for its ability to focus on the region of interest - the weld seam, as illustrated in Figures 6.2-6.3.

The procedure for scanning a BF weld consists of the following steps

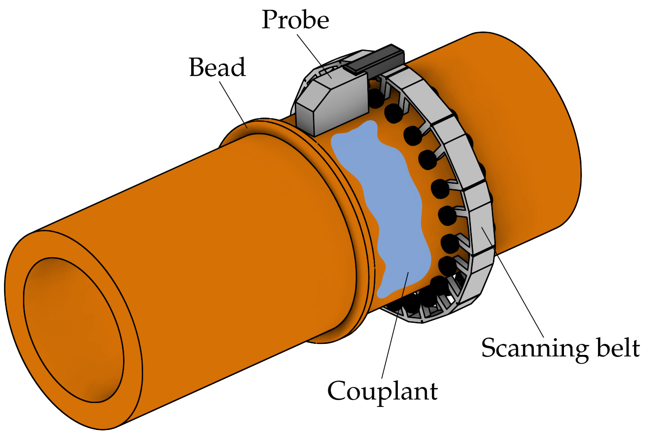


Figure 6.2: Illustration of the BF weld PAUT angle beam inspection with the use of a scanning belt to control the course of the probe. The scanning trajectory, as illustrated, is along the outer diameter of the pipe, parallel to the pipe's axis.

(a) Cleaning the pipe surface, if necessary.

(b) Installing PAUT system on the pipe; the probe and scanning belt.

(c) Applying UT coupling gel to the region of the probe is expected to be in contact with the probe. Coupling gel acts as a UT signal interface between the probe and the pipe.

(d) Initialising the data acquisition unit (DAU) with the recommended configuration suitable for the pipe material and thickness.

(e) Starting data collection mode on the DAU.

(f) Moving the scanner along the outer diameter of the pipe, as consistently as possible, with the aid of the scanning belt.

(g) Ending the data collection mode.

(h) Verifying the quality of the data collected by analysing the data manually using the DAU.

The encoder built into the scanning belt keeps track of the scanner's position along the circumference of the pipe. The DAU would then use the position information to sort the data following its scan process. Thus, providing a reference point that can be used to locate potential indications physically on the weld.

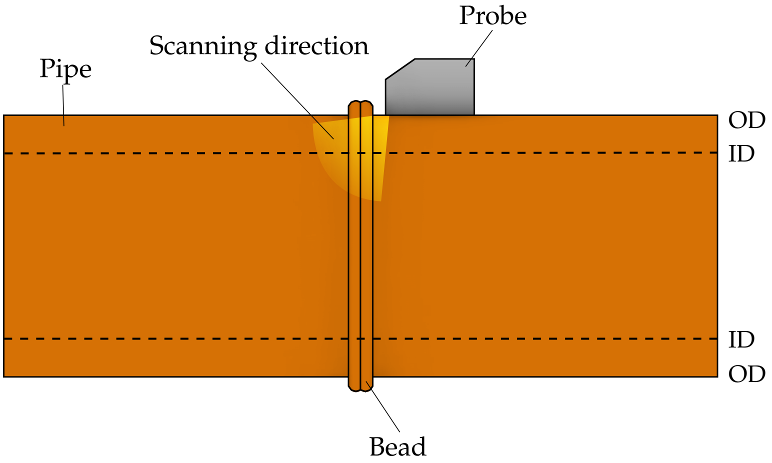


Figure 6.3: Illustration of the scanning direction in relation to the OD and ID of the pipe, in a cross-sectional view from the side. The positioning of the probe aims to capture UT waves reflected from the region of interest - the weld area. Moreover, it is worth noting that the ultrasound waves, because of their physical characteristics, do not travel past the pipe's ID; therefore, it is required to scan along the full outer circumference to be able to inspect the complete weld.

The data is saved in a proprietary format developed by Olympus - the PAUT system manufacturer. To be able to utilise the data from the system, a data extraction library[1] was utilised. The library was used to convert the data to the Matlab format. The library supports data retrieval to Matlab directly, using the code presented in Figure 6.4.

Once extracted, the data had to be converted from the raw signal format, amplitude-scan (A-Scan), to the image format utilised for BF inspection, brightness-scan (B-Scan). PAUT imagery aggregates A-Scan signal data in a manner that enables the NDT expert to analyse the data effectively and efficiently. Each A-Scan in the dataset represents an amplitude reading from a specific transducer at a certain circumferential position. Meanwhile, each B-Scan image seeks to display all A-Scans readings from a specific transducer. In BF inspection, each transducer represents scanning at a particular step angle.

$$\mathbf{B}_j = \begin{bmatrix} \mathbf{A}_{j,1} & \mathbf{A}_{j,2} & \cdots & \mathbf{A}_{j,n} \end{bmatrix} \tag{6.1}$$

Where $\mathbf{B}_j$ is the B-Scan image at angle step j, $\mathbf{A}_{j,n}$ is the A-Scan signal at angle step j and at circumference step n.

[1] Olympus NDT Data Access Library - Software Version 1.10

```
clc; clear; % clears command line and workspace

IN_DIR='C:\Users\Hesham\BFData\BF_OPD\'; %input directory
OUT_DIR='C:\Users\Hesham\BFData\BF_OPD_MAT\'; %output directory

FILENAME_MASK='*.opd'; %defines file extension to look for

%scans input directory for opd files.
[paths,names]=get_files_paths(DIR,FILENAME_MASK);

%for loop processes all .opd files detected in input directory
%and converts them into .mat format.
for file=1:length(paths)

%readRDT is a Matlab function file written by Olympus to utilise
%the Olympus \ac{NDT} Data Access Library - Software Version 1.10
DATA=readRDT(paths{file});

file_name=names{file}; %stores filename of current file

%saves the relevant variables to a .mat file by the same name
save(strcat(OUT_DIRECTORY,file_name,'.mat'),'DATA','file_name')

end
```

Figure 6.4: Matlab code for converting the data from the Olympus proprietry format (.opd) to Matlab format (.mat)

Moreover, the B-Scan image $\mathbf{B}_j$ at angle step j is defined by (6.1). Therefore, the resolution of the B-Scan image depends on the length of the A-Scan and the circumference of the pipe. The A-Scan's length is based on the scanning depth and scanning resolution.

6.3.2 Data labelling and feature extraction

After the data preparation stage, all the images were readily processed in the B-Scan format. Consequently, data labelling and feature extraction was required before any modelling. The image analysis aims to locate various key features, such as indications representing the pipe components or potential defects. Two sets of labelled data were produced for bead and defect detection, respectively.

For bead detection, the data was manually labelled in one of two classes: *Bead* or *No Bead*. Each B-Scan image was labelled to be either of these classes. A total of 2030 B-Scans were labelled from 30 weld scans.

In contrast, for the defect detection, the B-Scans were cropped into square image slices into one of two classes: *Defective* or *Non-defective*. The process was done manually to limit this project's scope to investigate the underlying interpretable modelling methodology, assuming that such a technique could be viably developed. A total of 1076 slices were labelled from 15 welds.

$$\text{Amplitude } \mu = \frac{\sum x_n}{N} \tag{6.2}$$

$$\text{Amplitude } \sigma = \sqrt{\frac{\sum(x_n - \mu)^2}{N}} \tag{6.3}$$

Where x_n is the amplitude at pixel n, and N is the total number of pixels in the image.

Following labelling, a total of nine features f_j were extracted from the labelled images. They are based on pixel amplitude, gray level co-occurrence matrix (GCLM) and edges detected. Two amplitude-based features are calculated: the mean and standard deviation (6.2, 6.3). Since high-amplitude indications are often how defects and other objects are located in the image, these features could be useful.

$$\text{GCLM Contrast} = \sum_{a,b} |a-b|^2 P(a,b) \tag{6.4}$$

$$\text{GCLM Correlation} = \sum_{a,b} \frac{(a-\mu_a)\,(a-\mu_b)\,P(a,b)}{\sigma_a \sigma_b} \tag{6.5}$$

$$\text{GCLM Energy} = \sum_{a,b} (P(a,b))^2 \tag{6.6}$$

$$\text{GCLM Homogeneity} = \sum_{a,b} \frac{P(a,b)}{1+|a-b|} \tag{6.7}$$

Where a is the first pixel; b is the adjacent pixel to be compared and; $P(a,b)$ presents the probability for gray-level pairs a and b, defined by (6.8), based on [119].

Furthermore, four GCLM based features are calculated: contrast, correlation, energy and homogeneity, as defined in (6.4-6.7), based on [119], [120]. Although the amplitude could be a promising indicator for a particular object, the image's texture feature could be used to distinguish between patterns with similar amplitude characteristics, such as *genuine indications* and *background noise*.

$$P(a,b) = \frac{x_{(a,b)}}{\sum\limits_{a,b=0}^{N} x_{(a,b)}} \tag{6.8}$$

Where $x_{(a,b)}$ is value at row a and column b.

$$\text{Mean Connected Edge Length} = \frac{\text{Total Edge Pixels}}{\text{Number of Connected Edges}} \tag{6.9}$$

The final component of the feature-set is the three edge detection-related features. The first feature provides the total number of edge pixels. Meanwhile, the second feature counts the number of total connected edges. This attempts to capture shape information and size. The third edge detection feature calculates the mean length of the connected edges detected, hence providing summarised insight about a key characteristic of the edges detected - their mean size as in (6.9).

6.3.3 Fuzzy-TOPSIS as a UT image classifier

The modelling frameworks applied in this chapter are described in detail in the previous chapters on Fuzzy-MCDM and explanation frameworks, in Sections 3.3 and 4.3.

Since performance does not differ significantly between the different Fuzzy-MCDM classifiers, fuzzy-TOPSIS was favoured for its relative simplicity. As presented in Section 3.3, fuzzy-TOPSIS utilises a simpler singular FIS as opposed to the hierarchical one used for Fuzzy-ATOVIC and Fuzzy-VIseKriterijumska Optimizacija I Kompromisno Resenje (translation: Multicriteria Optimization and Compromise Solution) (VIKOR). Thus, maximising the model's interpretability and allowing for the design of an explanation framework proposed in Section 4.3.

$$\text{Centroid } C_{j,p} = \frac{\sum_{i=1}^{N} x_{i,j_p}}{N} \tag{6.10}$$

Where $C_{j,p}$ is the centroid for feature j, for class (and cluster) p and; x_{i,j_p} is the value of feature j for row i and, class p.

As a comparison, a supervised K-Means model in a single iteration approach where a cluster was used to represent each class. The mean of each feature in each class was calculated to form the centroids, as defined in (6.10). This enables a fair assessment of fuzzy-TOPSIS as a classifier where it can be compared with a model from the same paradigm.

6.4 Experimental results

In this section, the performance results are presented for the fuzzy-TOPSIS classifier in comparison with K-means. Consequently, the benefit of opting for an inherently interpretable classifier is highlighted by demonstrating how transparency-driven explanation is exploited for meaningful and direct insight. The process is repeated across two different tasks: bead detection and defect recognition. Finally, the section is concluded by analysing how different aspects of the explanation relate to performance.

6.4.1 Bead detection

The process of manual defect recognition of BF welds entails a series of operations by the expert designed to gather as much information from the UT image. The steps would vary depending on the application. For BF, the first step is locating the bead indication. This gives an idea of where the weld is positioned in relation to the image. The aim is to save the expert time by automatically determining the images detected with beads.

The BF bead weld dataset was distributed into two classes *Bead* and *No Bead*. The classes are not balanced with significantly more cases of *No Bead*. The *No Bead* class

Table 6.1: BF weld bead dataset: class distribution before and after data balancing

No.	Class	Before	%	After	%
1	No Bead (-)	1492	73.5	538	50.0
2	Bead (+)	538	26.5	538	50.0
	Total	2030	100.0	1076	100.0

Table 6.2: BF Weld dataset: feature names

#	Name
1	Amplitude Mean
2	Amplitude SD
3	GCLM Contrast
4	GCLM Correlation
5	GCLM Energy
6	GCLM Homogeneity
7	Edge Pixels Count
8	Number of Edges
9	Average Edge Length

was under-sampled to match the sample count of the *Bead* class, as presented in Table 6.1; as a method of balancing the dataset and achieving superior performance.

The feature set contained nine features, as shown in Table 6.2. The first two features describe the amplitude characteristics of the image as this is the first criteria experts utilise to distinguish and locate indications in ultrasonic images. Secondly, GCLM features describe texture information that could be useful in discerning whether the image contains a defect. Finally, edge detection-related features (7 to 9) summarise edge data detected in the image with the aim of picking up statistical characteristics associated with edges found in the images of interest.

Performance

Contrary to expectation, the fuzzy-TOPSIS model performed favourably compared to the K-Means model with more than 5% performance improvement for both ACC and FMR, as presented in Table 6.3. Despite, K-Means having a higher TPR, its lower TNR is seen as the reason for its inferior overall performance. Therefore, for this dataset, using fuzzy-TOPSIS instead of K-Means does not result in a trade-off of performance.

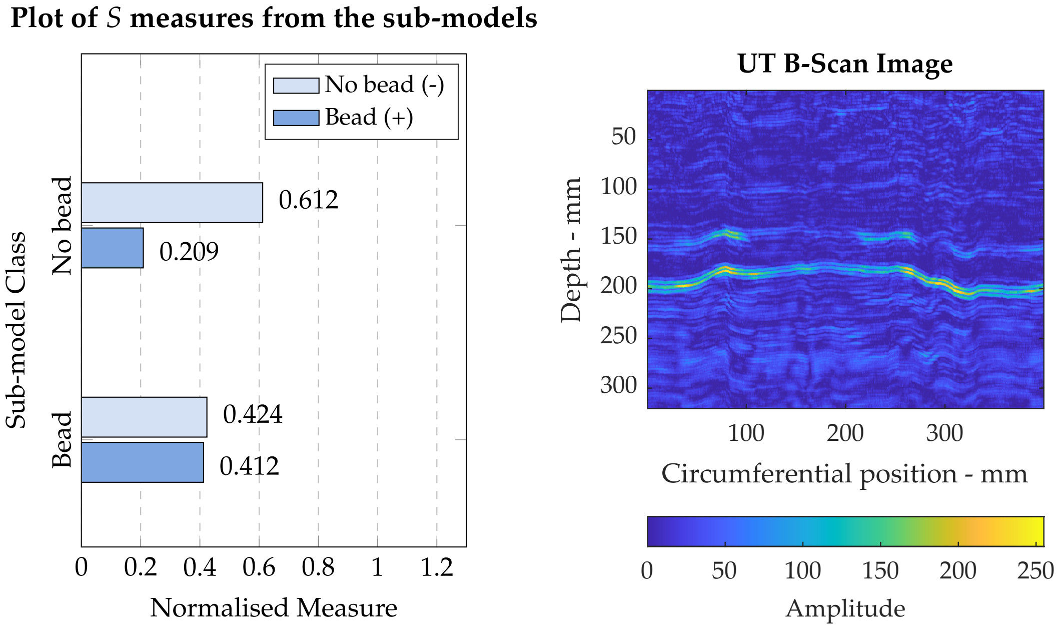
Table 6.3: BF weld bead detection: performance comparison between TOPSIS and K-Means for the testing dataset for ten randomised runs of 5-fold data

	$\mu \pm \sigma$ (%)			
Model	ACC	FMR	TPR	TNR
Fuzzy-TOPSIS	81.3 ± 2.8	81.3 ± 2.7	81.1 ± 3.4	81.6 ± 4.0
K-Means	75.9 ± 2.4	74.5 ± 3.0	85.9 ± 3.0	66.0 ± 4.8

Explanation results

The explanation framework is the key advantage of adopting an inherently interpretable framework. Below are several examples from the framework's explanation for the bead detection dataset. The bead is the region of a BF weld in-line with the weld seam - the axis at which the two pipes are joined. The nature of the BF welding technique results in what is called a bead protrusion appearing at the weld seam. Fortunately, this *bead* allows for a highly visible indication to be visible on the UT image - which provides a visual aid for experts.

The bead indication is often characterised by a higher-amplitude horizontal ribbon-like indication which is always somewhat parallel to the x-axis - the circumferential direction. This is a result of the scanning direction, which occurs along the pipe's outer circumference, i.e. in the same direction as the weld seam and bead.



(a)

(b)

Figure 6.5: BF weld bead detection: (a) Illustration of the distance measures S from the two sub-models - example of a TP case: #2_2_136_6_48. Textual explanation: “Models are in agreement hence, it is more likely the classification is accurate. Fuzzy class: 0.65”. (b) Plot of UT image data for the same example. The image was classified as *Bead* by the model.

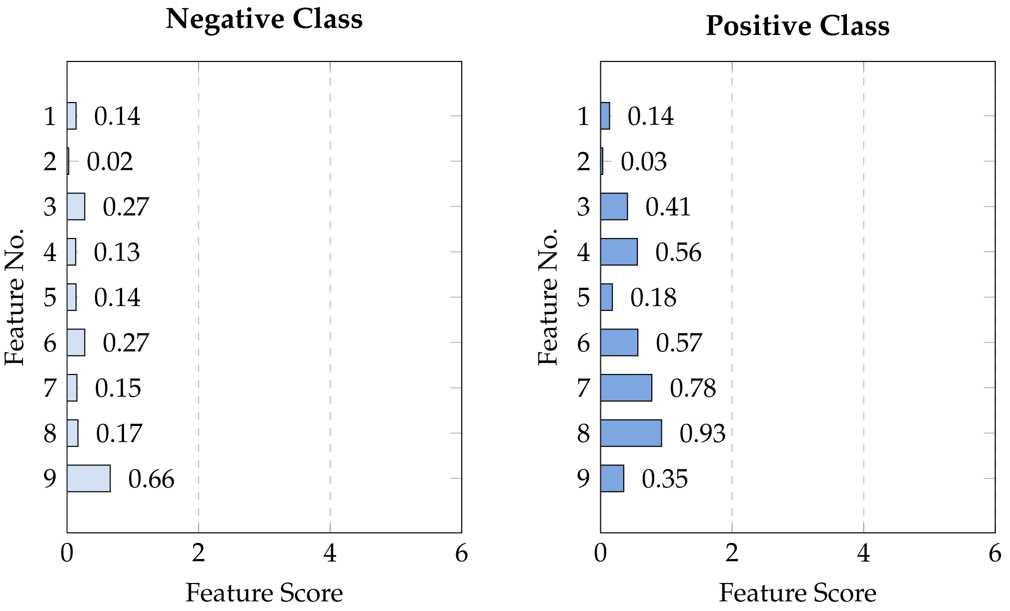


Figure 6.6: BF weld bead detection: Graphical explanation of the *No Bead* sub-model - example of a TP case: #2_2_136_6_48. Textual explanation: “No bead model thinks the data is similar to Bead (+ve) and NOT similar to No bead (-ve).”

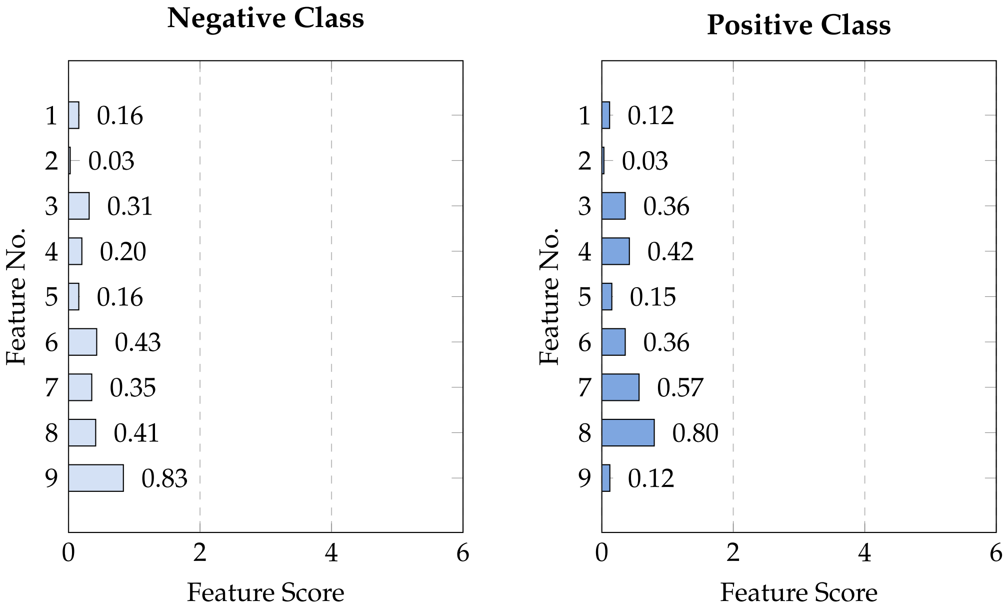
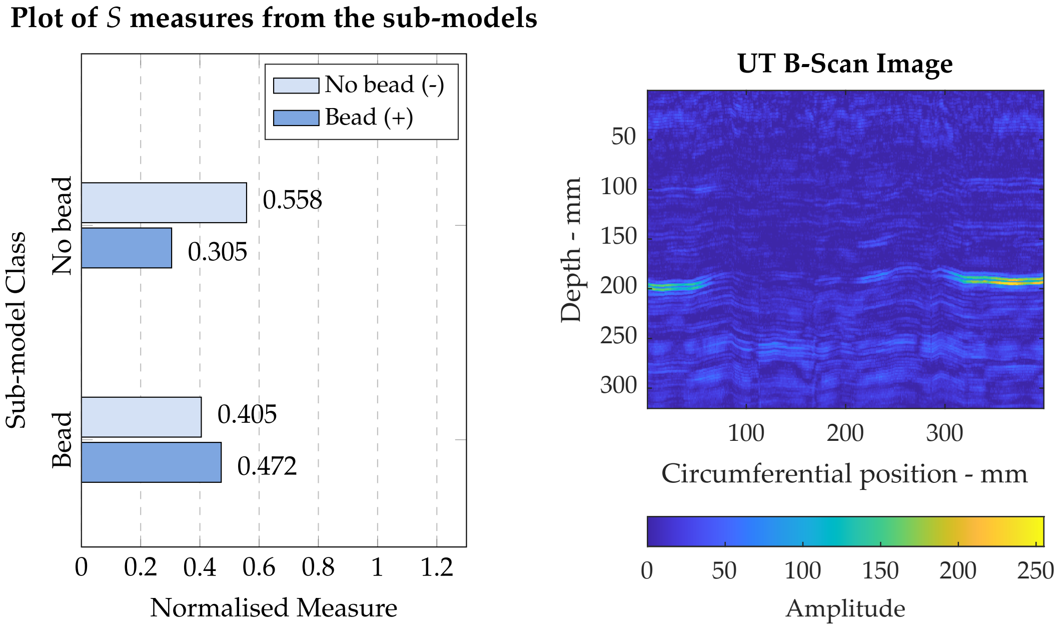


Figure 6.7: BF weld bead detection: Graphical explanation of the *Bead* sub-model - example of a TP case: #2_2_136_6_48. Textual explanation: "Bead model thinks the data is more similar to Bead (+ve) despite a high similarity for both."

We begin with two TP examples that demonstrate how different aspects of the explanation vary with how prominent the bead indication is in the UT image. The first example (Figures 6.5-6.7) illustrates a case where the bead indication is distinctly visible since it maintains high amplitude throughout - considerably distinguishable from the background noise.



(a)

(b)

Figure 6.8: BF weld bead detection: (a) Illustration of the distance measures S from the two sub-models - example of a TP case: #2_2_145_10_35. Textual explanation: "Models are in conflict hence, it is less likely the classification is accurate. Fuzzy class: 0.57". (b) Plot of UT image data for the same example. The image was classified as *Bead* by the model.

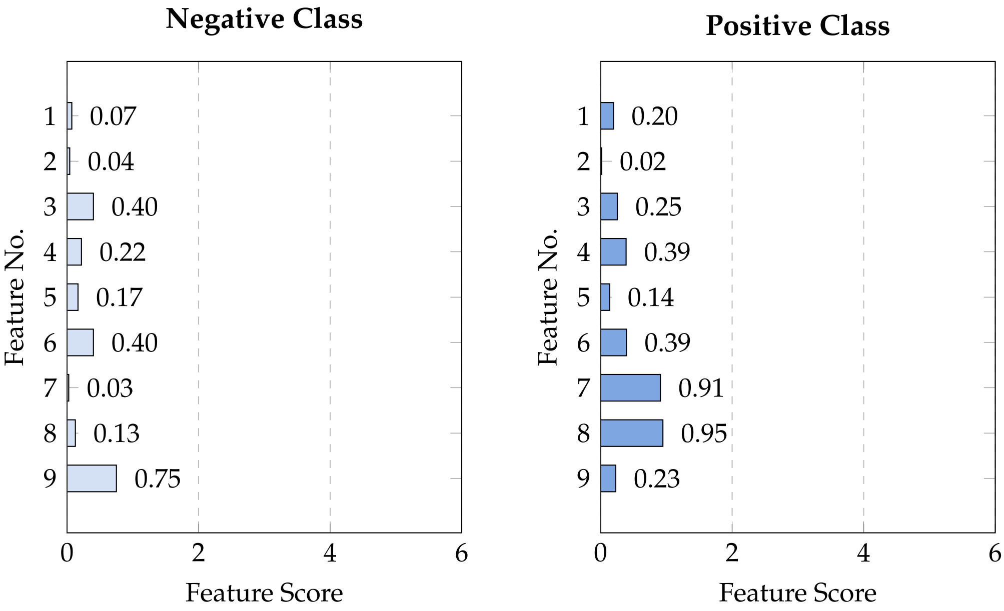


Figure 6.9: BF weld bead detection: Graphical explanation of the *No Bead* sub-model - example of a TP case: #2_2_145_10_35. Textual explanation: "No bead model thinks the data is similar to Bead (+ve) and NOT similar to No bead (-ve)."

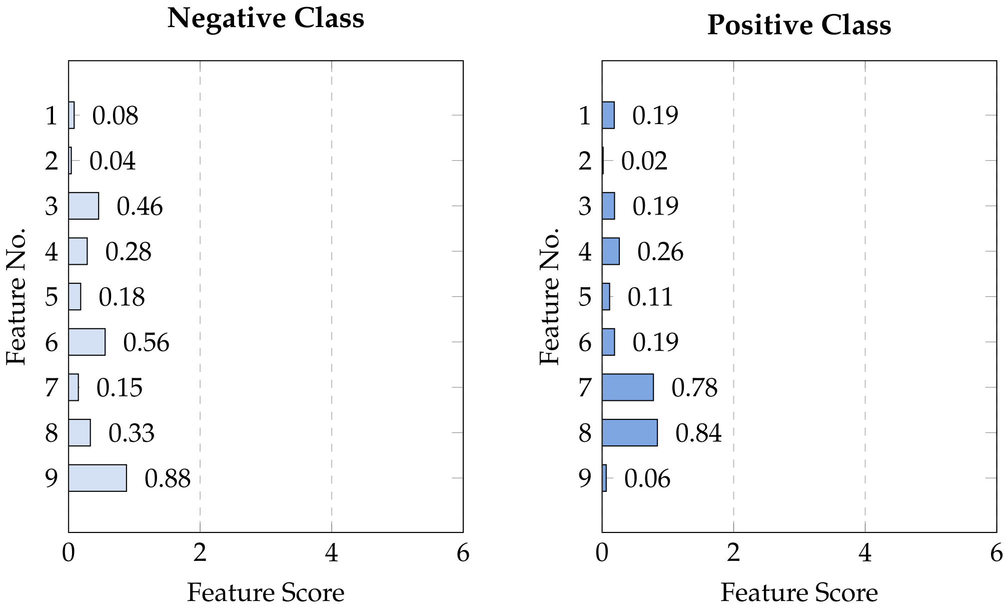


Figure 6.10: BF weld bead detection: Graphical explanation of the *Bead* sub-model - example of a TP case: #2_2_145_10_35. Textual explanation: “Bead model thinks the data is more similar to No bead (-ve) despite a high similarity for both.”

However, in the second example (Figures 6.8-6.10) the bead indication is relatively less visible between a circumferential position of 100 and 300; this is perceived by a lower amplitude in this range. The following aspects of the explanation indicate a less obvious classification:

(a) Conflict between the two sub-models versus agreement in the previous example.

(b) The fuzzy class output is closer to the 0.5 threshold with a value of 0.57 compared to 0.65 for the previous example.

(c) The *No Bead* model predicted a significantly lower similarity to the *Bead* class with a measure of 0.305 compared to 0.209 for the previous example.

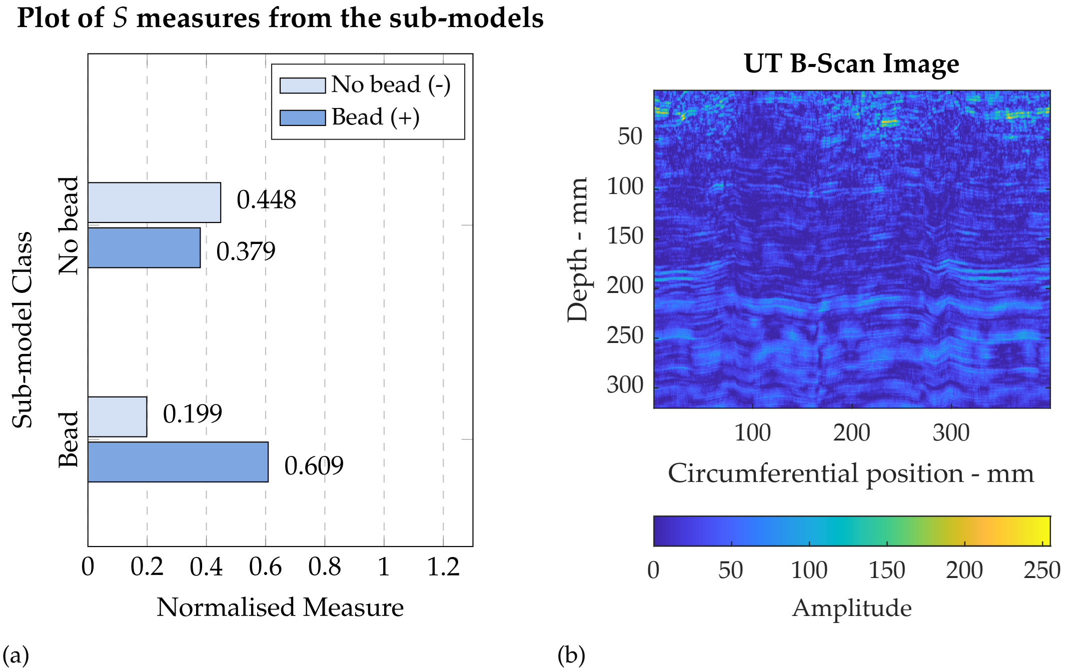


Figure 6.11: BF weld bead detection: (a) Illustration of the distance measures S from the two sub-models - example of a TN case: #2_2_8_4_60. Textual explanation: “Models are in conflict hence, it is less likely the classification is accurate. Fuzzy class: 0.37”. (b) Plot of UT image data for the same example. The image was classified as *No Bead* by the model.

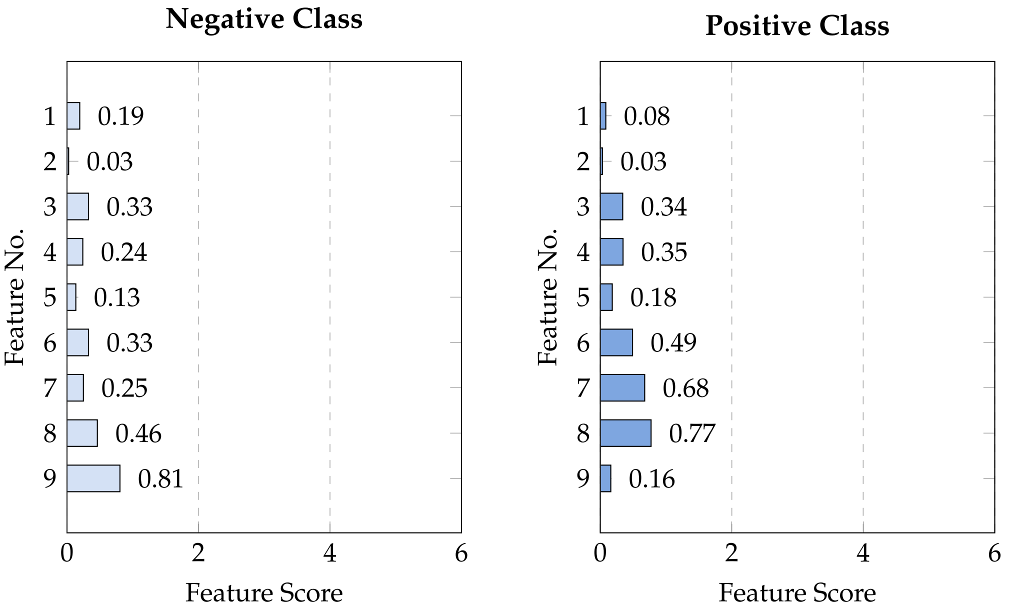


Figure 6.12: BF weld bead detection: Graphical explanation of the *No Bead* sub-model - example of a TN case: #2_2_8_4_60. Textual explanation: “No bead model thinks the data is more similar to Bead (+ve) despite a high similarity for both.”

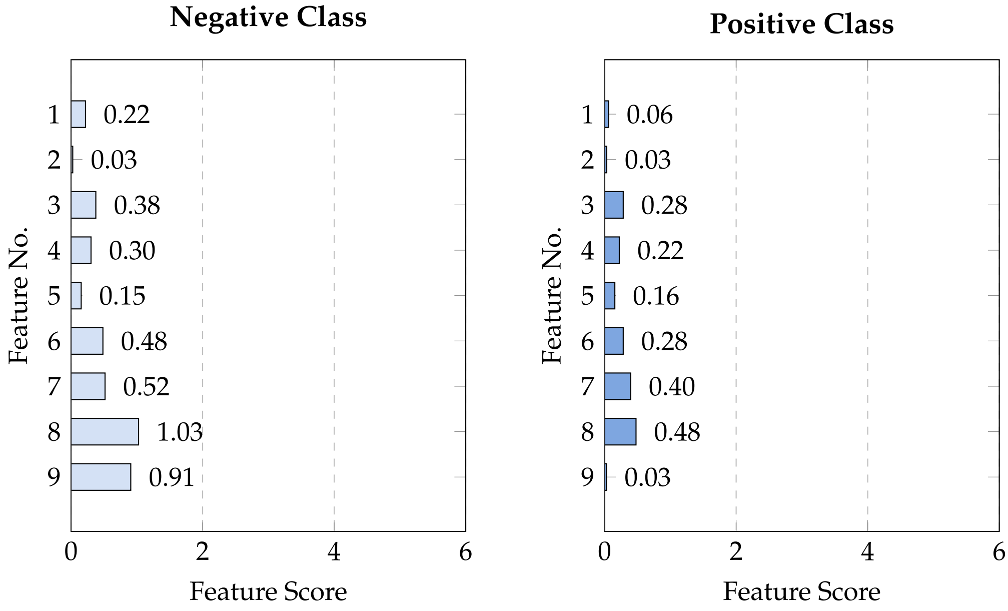
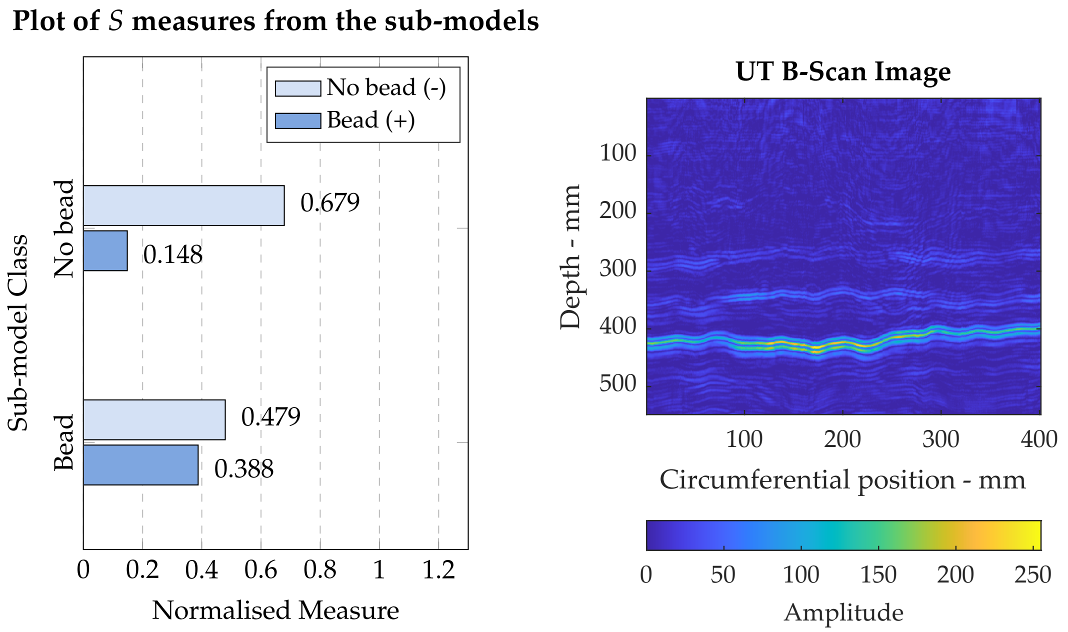


Figure 6.13: BF weld bead detection: Graphical explanation of the *Bead* sub-model - example of a TN case: #2_2_8_4_60. Textual explanation: "Bead model thinks the data is similar to No bead (-ve) and NOT similar to Bead (+ve)."

Moreover, the TN example (in Figures 6.11-6.13) shows an instance where the image lacks a prominent bead signal thus, was successfully classified as *No Bead*. The sub-models were in *conflict* which indicates potential uncertainty in the decision. After inspecting the image, in Figure 6.11, a faint but distinguishable indication in the same position as a potential bead indication is observed; this might explain the lack of agreement between the sub-models. Although the *No Bead* sub-model decided for *Bead*, the *Bead* sub-model's stronger decision towards *No Bead* was sufficient to reach the correct outcome. Furthermore, the edge-related features were the most impactful for the *Bead* sub-model's decision, as shown in Figure 6.13. Notably, such indicative information could be deemed useful to experts for understanding the model's decision and/or coming up with a justification.



(a)

(b)

Figure 6.14: BF weld bead detection: (a) Illustration of the distance measures S from the two sub-models - example of a FP case: #2_2_91_33_19. Textual explanation: “Models are in agreement hence, it is more likely the classification is accurate. Fuzzy class: 0.67”. (b) Plot of UT image data for the same example. The image was classified as *Bead* by the model.

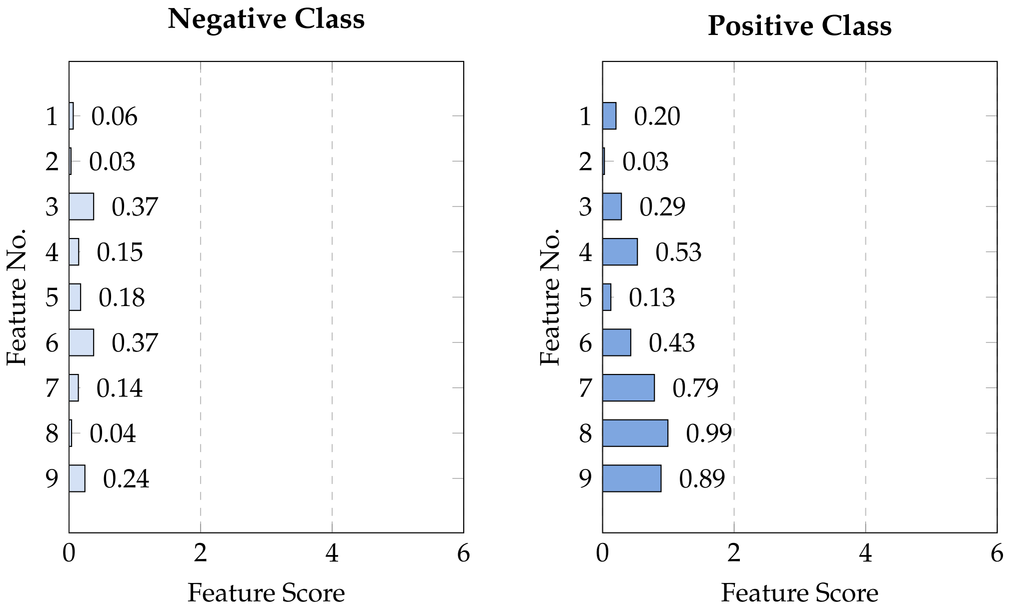


Figure 6.15: BF weld bead detection: Graphical explanation of the *No Bead* sub-model - example of a FP case: #2_2_91_33_19. Textual explanation: “No bead model thinks the data is similar to Bead (+ve) and NOT similar to No bead (-ve).”

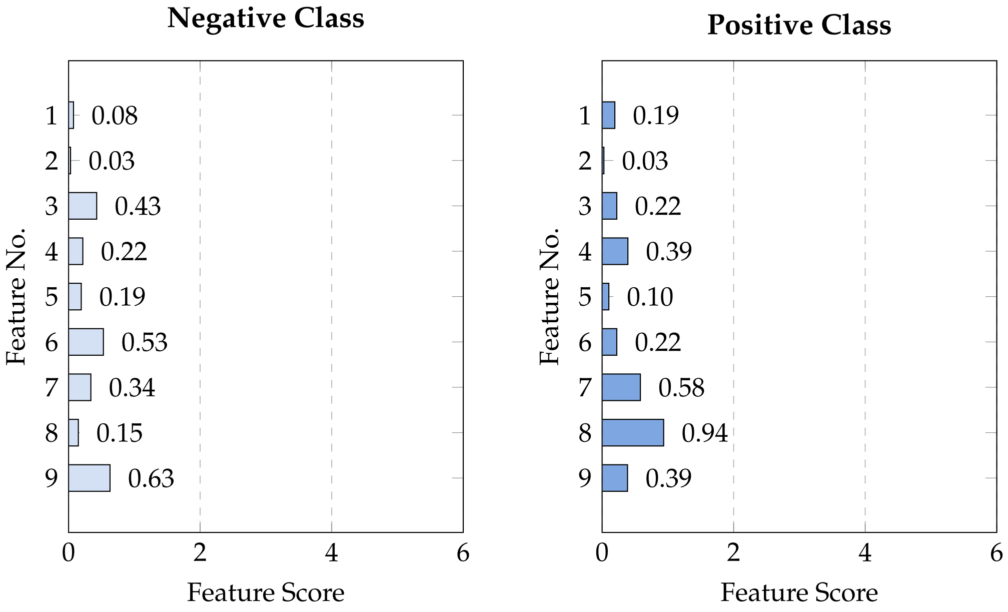
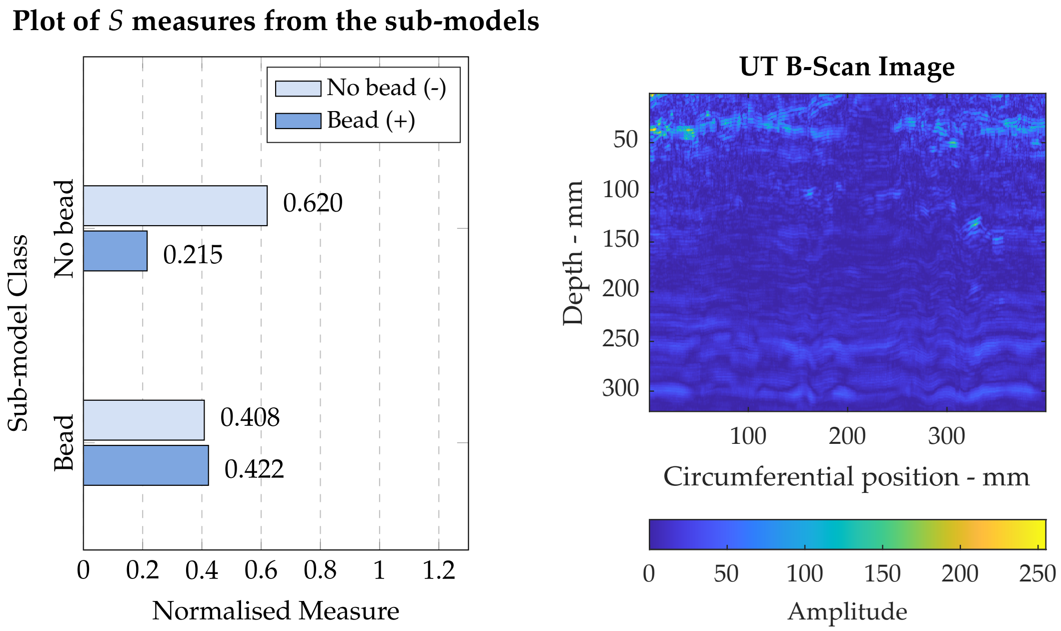


Figure 6.16: BF weld bead detection: Graphical explanation of the *Bead* sub-model - example of a FP case: #2_2_91_33_19. Textual explanation: "Bead model thinks the data is more similar to Bead (+ve) despite a high similarity for both."

Similar to the TP examples, this pair of FP cases shows another example of how explanation describes the level of certainty associated with the classification. Both cases result in an incorrect classification, however, the first example (in Figures 6.14-6.16) contains several bead-like indications which seemed to have tricked the model into misclassifying the image as *Bead*. This is confirmed by observing the edge-related features (7-9), which have high scores across the two sub-models.



(a)

(b)

Figure 6.17: BF weld bead detection: (a) Illustration of the distance measures S from the two sub-models - example of a FP case: #2_2_16_8_57. Textual explanation: “Models are in conflict hence, it is less likely the classification is accurate. Fuzzy class: 0.65”. (b) Plot of UT image data for the same example. The image was classified as *Bead* by the model.

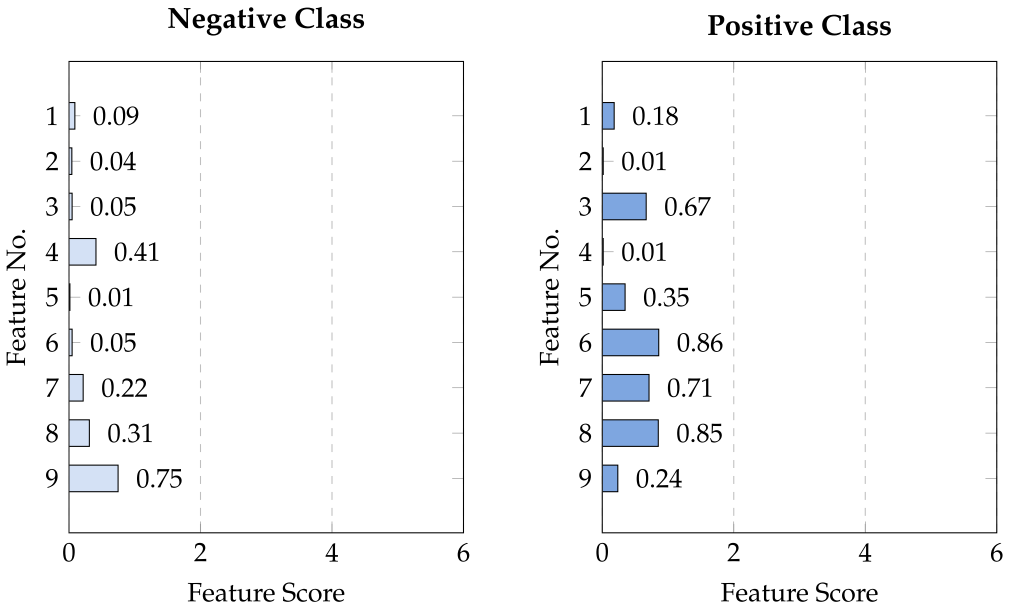


Figure 6.18: BF weld bead detection: Graphical explanation of the *No Bead* sub-model - example of a FP case: #2_2_16_8_57. Textual explanation: “No bead model thinks the data is similar to Bead (+ve) and NOT similar to No bead (-ve).”

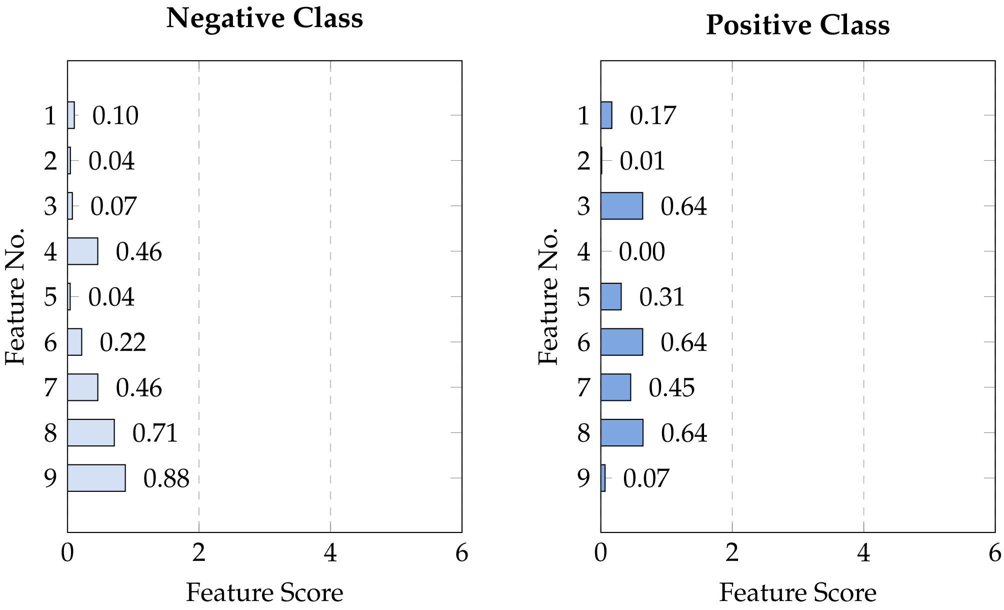


Figure 6.19: BF weld bead detection: Graphical explanation of the *Bead* sub-model - example of a FP case: #2_2_16_8_57. Textual explanation: "Bead model thinks the data is more similar to No bead (-ve) despite a high similarity for both."

In contrast, for the second example, the image (in Figure 6.17) does not contain any visible bead-like indications. Furthermore, in spite of a conflict between the sub-models, the classification was incorrect. The deciding component behind the inaccurate decision was the *No Bead* sub-model because of its stronger positive classification, as shown in Figure 6.17. The significant background noise in the top area of the image may be the cause of *No Bead* sub-model perceiving high similarity to *bead*.

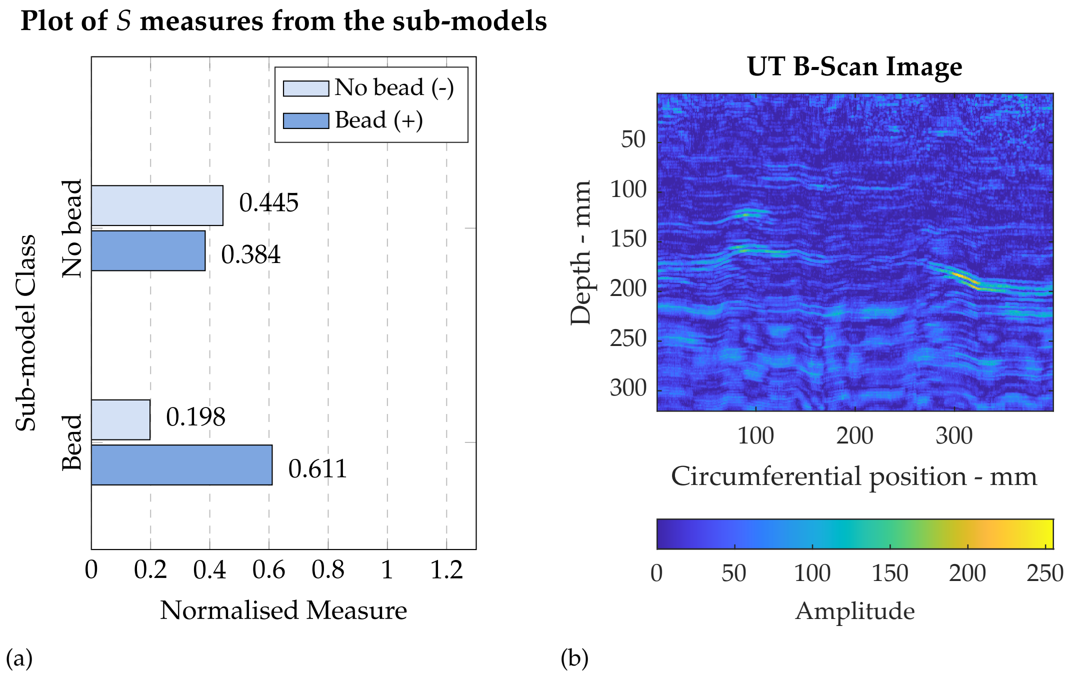


Figure 6.20: BF weld bead detection: (a) Illustration of the distance measures S from the two sub-models - example of a FN case: #2_2_151_15_51. Textual explanation: “Models are in conflict hence, it is less likely the classification is accurate. Fuzzy class: 0.37”. (b) Plot of UT image data for the same example. The image was classified as *No Bead* by the model.

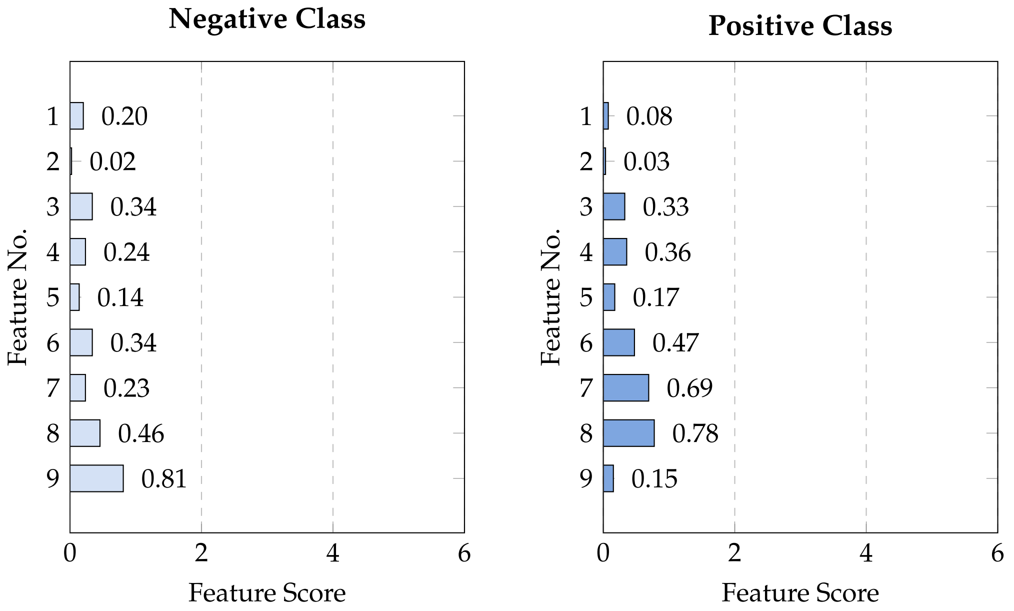


Figure 6.21: BF weld bead detection: Graphical explanation of the *No Bead* sub-model - example of a FN case: #2_2_151_15_51. Textual explanation: “No bead model thinks the data is more similar to Bead (+ve) despite a high similarity for both.”

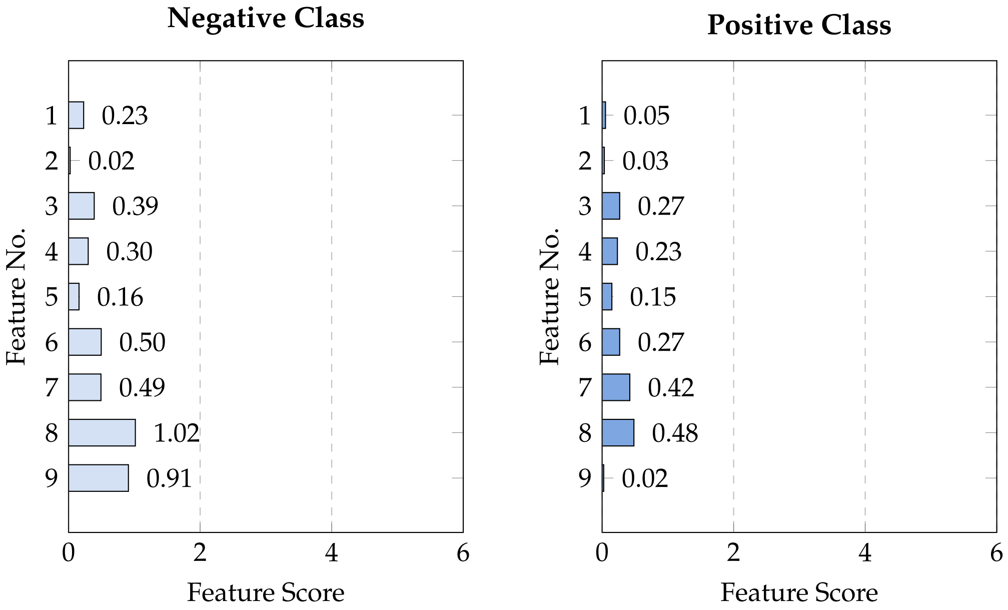


Figure 6.22: BF weld bead detection: Graphical explanation of the *Bead* sub-model - example of a FN case: #2_2_151_15_51. Textual explanation: "Bead model thinks the data is similar to No bead (-ve) and NOT similar to Bead (+ve)."

In this example (Figures 6.20-6.22), the model was not able to detect the bead indication, although it was distinguishable when looking at the image, as shown in Figure 6.20. It would have been the correct decision if the No Bead sub-model had been decided. However, the Bead sub-model's stronger decision towards No Bead was the deciding factor. Similar to some previous examples, the edge-related features (7-9) had the greatest influence on the Bead sub-model's decision, as shown in Figure 6.22. FN cases are more problematic, particularly for cases where classification models are primarily relied upon.

6.4.2 Defect recognition

The next vital step of UT inspection of BF welds is being able to recognise defects. However, experts recognise defects in a similar way to beads by looking at various regions of the images where high amplitude indications occur. Hence, for this dataset, a slice of the image is classified instead of the image as a whole - mimicking an expert's technique.

The BF defect weld dataset was distributed into two classes *defective* and *non-defective*. The classes are relatively balanced, with around the same number of

Table 6.4: BF weld defect dataset: class distribution

No.	Class	Count	%
1	Non-defective (-)	427	51.1
2	Defective (+)	408	48.9
	Total	835	100

Table 6.5: BF weld defect recognition: performance comparison between TOPSIS and K-means for the testing dataset for ten randomised runs of 5-fold data

	$\mu \pm \sigma$ (%)			
Model	ACC	FMR	TPR	TNR
Fuzzy-TOPSIS	87.7 ± 2.7	87.4 ± 2.7	93.2 ± 3.0	82.4 ± 4.5
K-Means	84.1 ± 2.7	82.9 ± 3.1	94.7 ± 2.4	73.9 ± 5.0

cases in both classes, as presented in Table 6.4. Similarly, the dataset contained the same nine features as the bead detection dataset, listed in Table 6.2.

Performance

The fuzzy-TOPSIS classifier used two sub-models to represent the two classes: defective vs non-defective. As a performance assessment, it was compared with a K Means classifier with two clusters in an attempt to gauge its accuracy on a model with similar capability. There was an expectation of seeing a performance drop for the fuzzy-TOPSIS model since it is not well-established as a classification methodology. Nonetheless, the performance was similar between the two models, with fuzzy-TOPSIS actually performing 3.6% more accurate, as presented in Figure 6.5. The specificity of fuzzy-TOPSIS was also considerably better at 82.4% compared to 73.9%; this was at an apparent cost of a 1.5% drop in sensitivity.

Explanation results

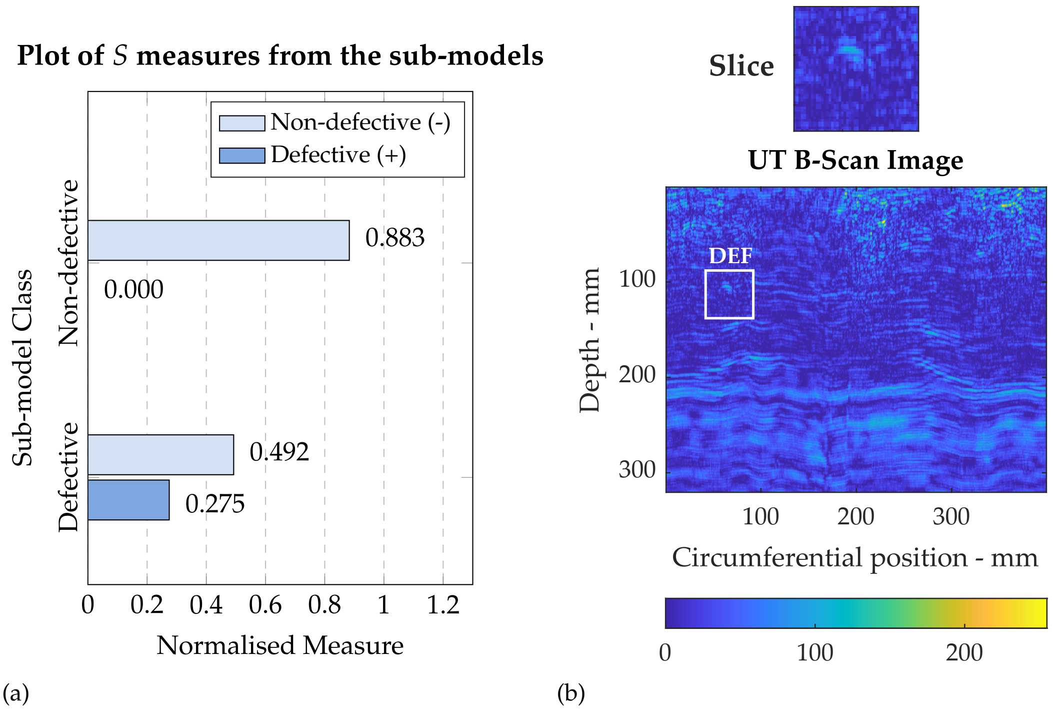


Figure 6.23: BF weld defect detection: (a) Illustration of the distance measures S from the two sub-models - example of a TP case: #2_2_97_2_57. Textual explanation: “Models are in agreement hence, it is more likely the classification is accurate. Fuzzy class: 0.78”. (b) Plot of UT image data for the same example. The image was classified as *defective* by the model.

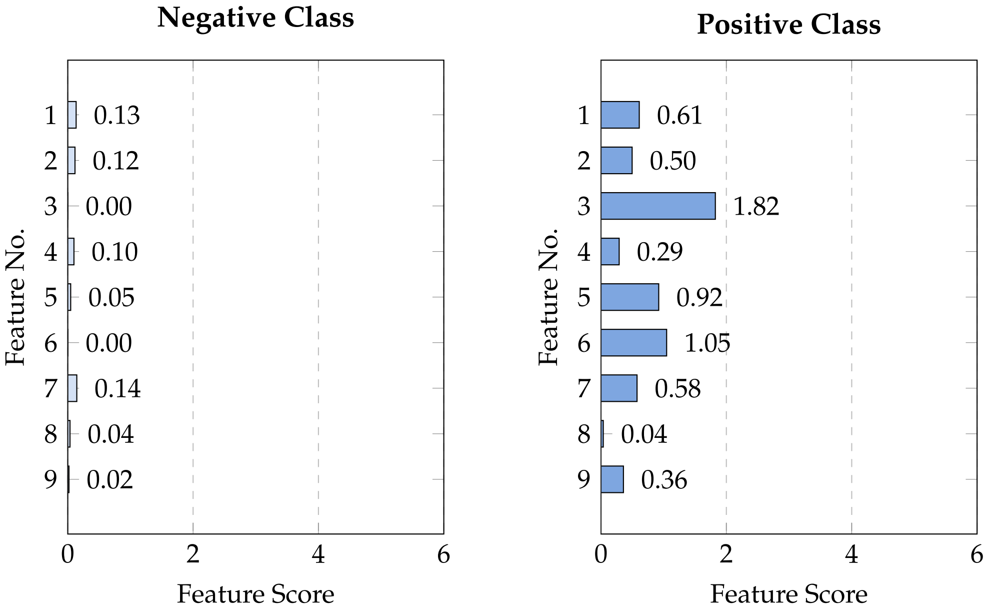


Figure 6.24: BF weld defect detection: Graphical explanation of the *Non-defective* sub-model - example of a TP case: #2_2_97_2_57. Textual explanation: “Non-defective model thinks the data is similar to Defective (+ve) and NOT similar to Non-defective (-ve).”

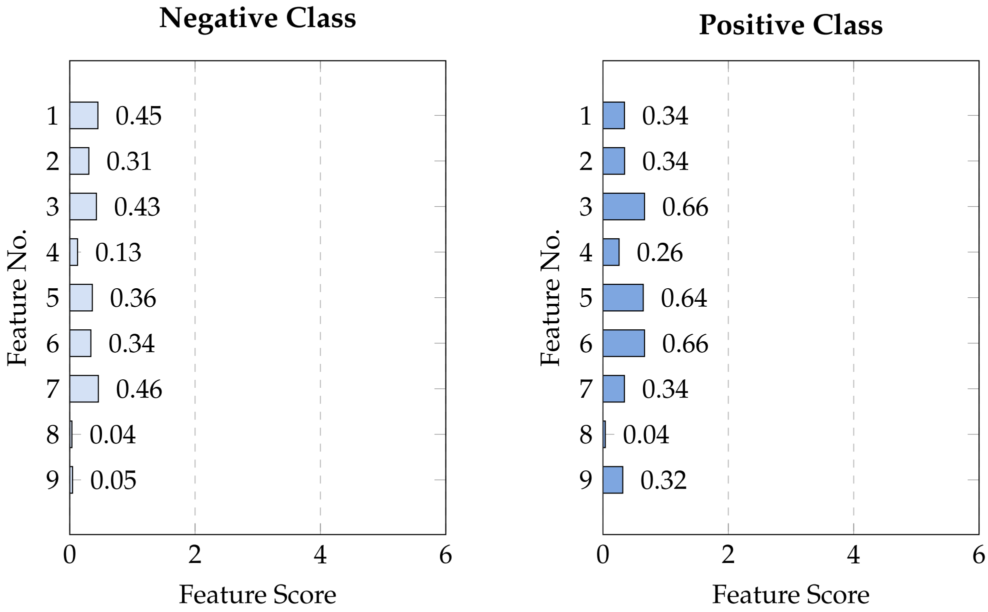
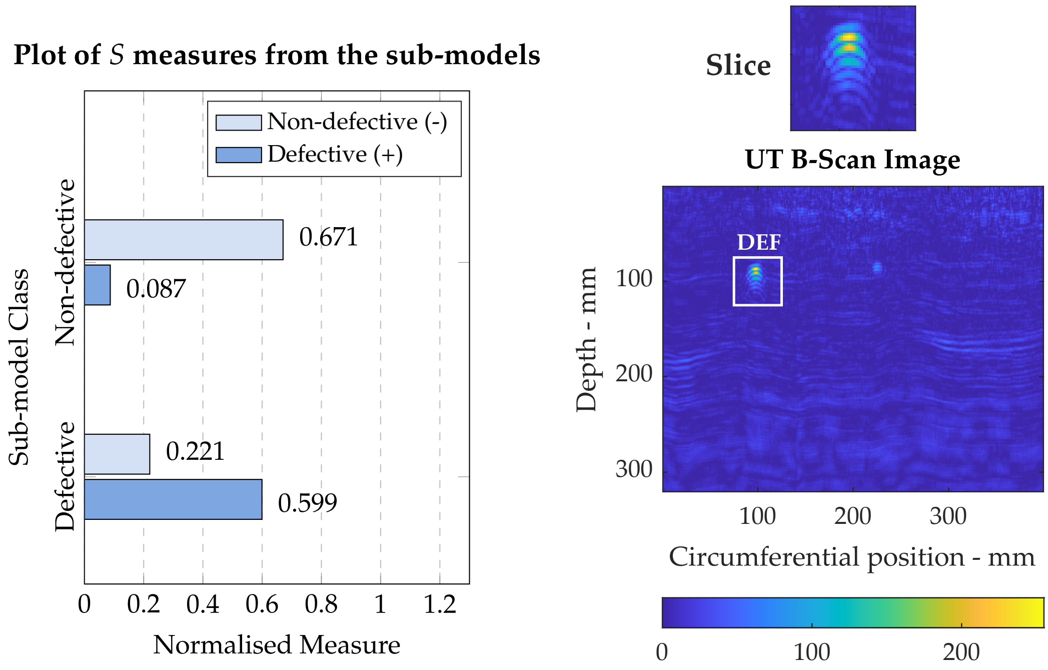


Figure 6.25: BF weld defect detection: Graphical explanation of the *Defective* sub-model - example of a TP case: #2_2_97_2_57. Textual explanation: "Defective model thinks the data is more similar to Defective (+ve) despite a high similarity for both."

The indications visible on the UT images vary in amplitude depending on how prominent the defects are relative to the ultrasonic transducer. In the first example, a TP case, the defect is not particularly obvious (as shown in Figure 6.23). Nonetheless, the sub-models were in agreement, both of which classified the image slice as *defective*. Inspecting the feature scores for the two sub-models (in Figures 6.24-6.25) revealed that feature #3 and other GCLM features had values closer to the positive ideal solutions for the positive class.



(a) (b)

Figure 6.26: BF weld defect detection: (a) Illustration of the distance measures S from the two sub-models - example of a TP case: #2_2_113_20_59. Textual explanation: "Models are in conflict hence, it is less likely the classification is accurate. Fuzzy class: 0.58". (b) Plot of UT image data for the same example. The image was classified as *defective* by the model.

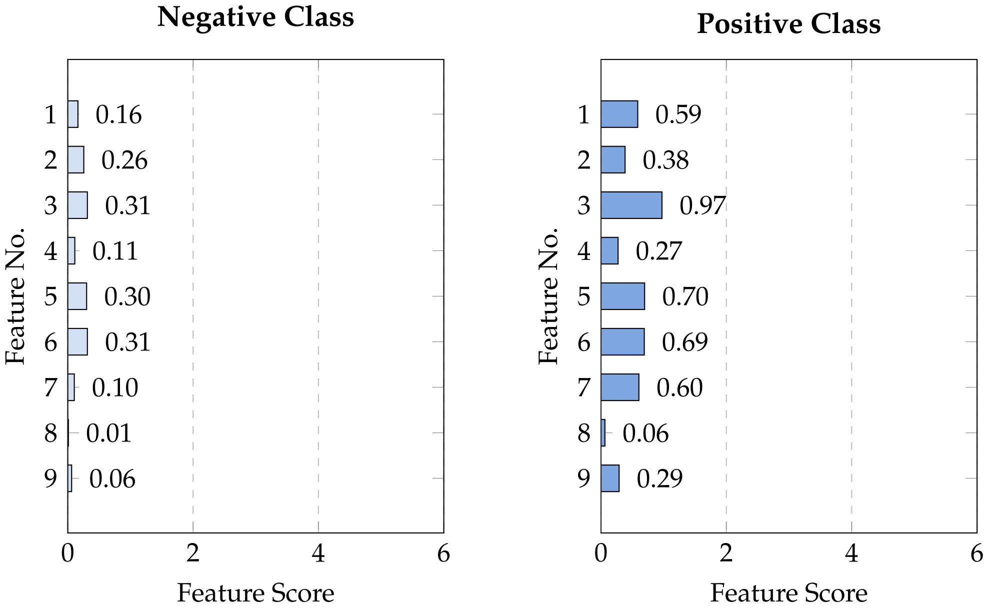


Figure 6.27: BF weld defect detection: Graphical explanation of the *Non-defective* sub-model - example of a TP case: #2_2_113_20_59. Textual explanation: "Non-defective model thinks the data is similar to Defective (+ve) and NOT similar to Non-defective (-ve)."

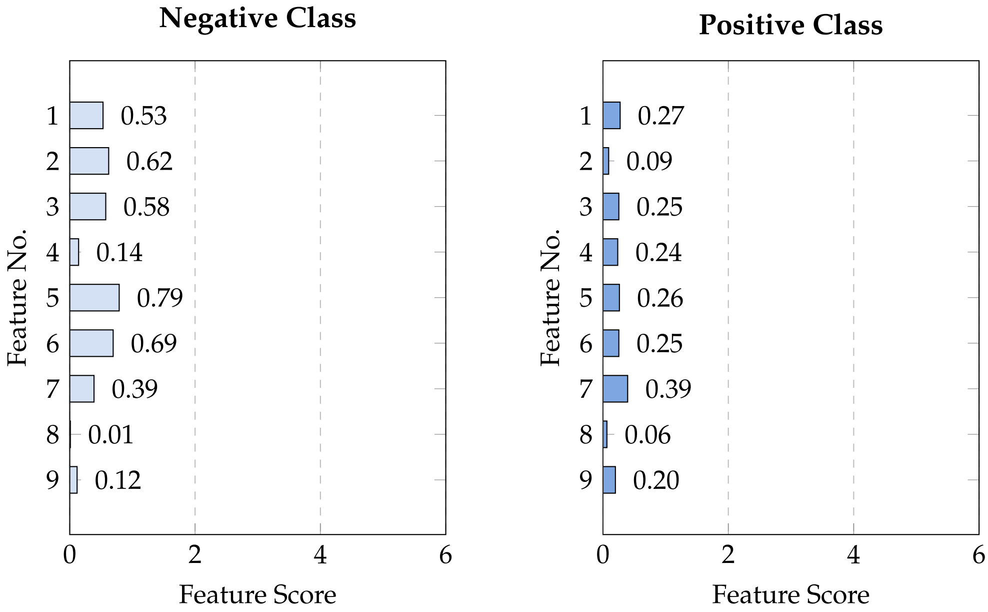
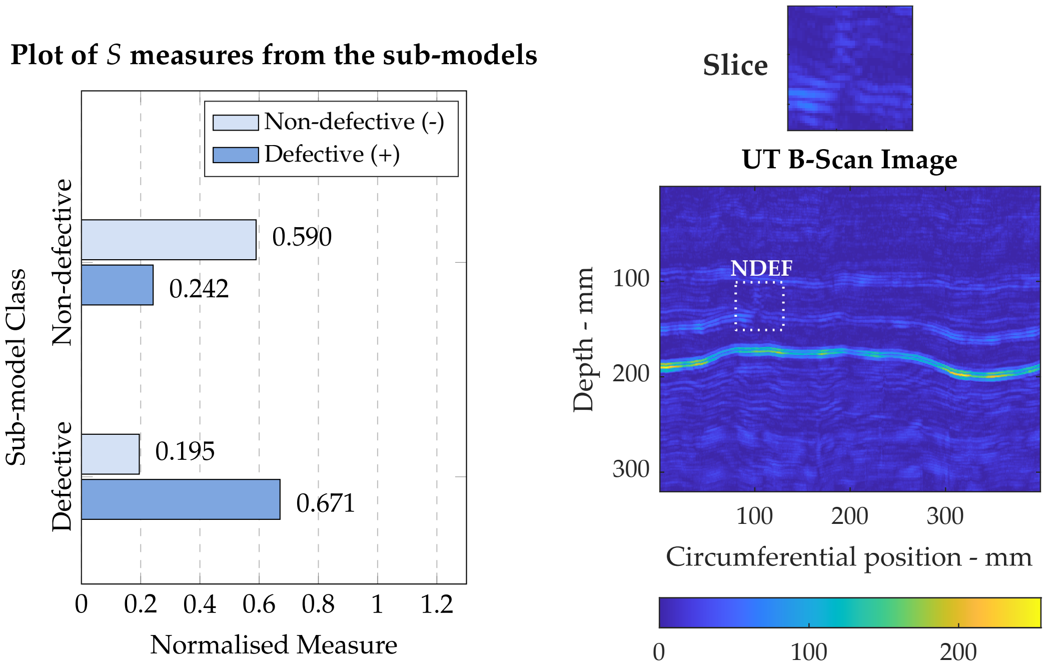


Figure 6.28: BF weld defect detection: Graphical explanation of the *Defective* sub-model - example of a TP case: #2_2_113_20_59. Textual explanation: "Defective model thinks the data is similar to Non-defective (-ve) and NOT similar to Defective (+ve)."

Moreover, the second TP example demonstrates a case where the two sub-models are in conflict, where the non-defective model predicts the slice to be defective, while the defective model predicts it to be non-defective (as shown in Figure 6.26). Despite the conflict, the slice was correctly classified as *defective*. Inspecting the feature scores (in Figures 6.27-6.28) highlighted the key criteria affecting the decision. More importantly, the feature scores for the non-defective model provided the most meaningful information because of its pivotal role in this decision. A large number of the features had a score higher than 0.5, which for this dataset is considered *impactful*. Most notable features affecting the decision include the GCLM features (specifically GCLM Contrast (#3)), edge pixels count (#7) and mean amplitude (#1).



(a) (b)

Figure 6.29: BF weld defect detection: (a) Illustration of the distance measures S from the two sub-models - example of a TN case: #2_2_41_19_39. Textual explanation: "Models are in conflict hence, it is less likely the classification is accurate. Fuzzy class: 0.45". (b) Plot of UT image data for the same example. The image was classified as *non-defective* by the model.

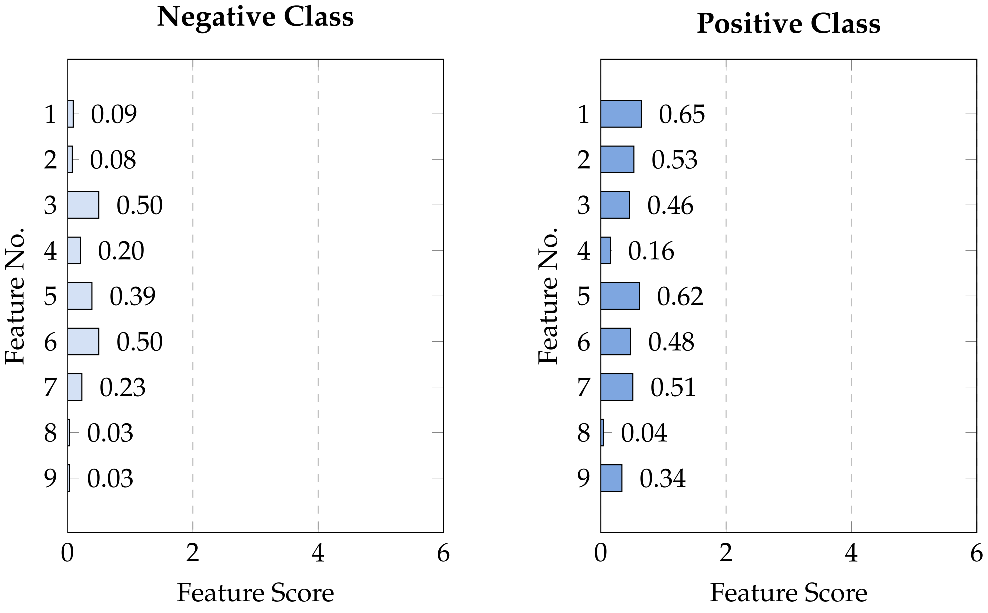


Figure 6.30: BF weld defect detection: Graphical explanation of the *Non-defective* sub-model - example of a TN case: #2_2_41_19_39. Textual explanation: "Non-defective model thinks the data is similar to Defective (+ve) and NOT similar to Non-defective (-ve)."

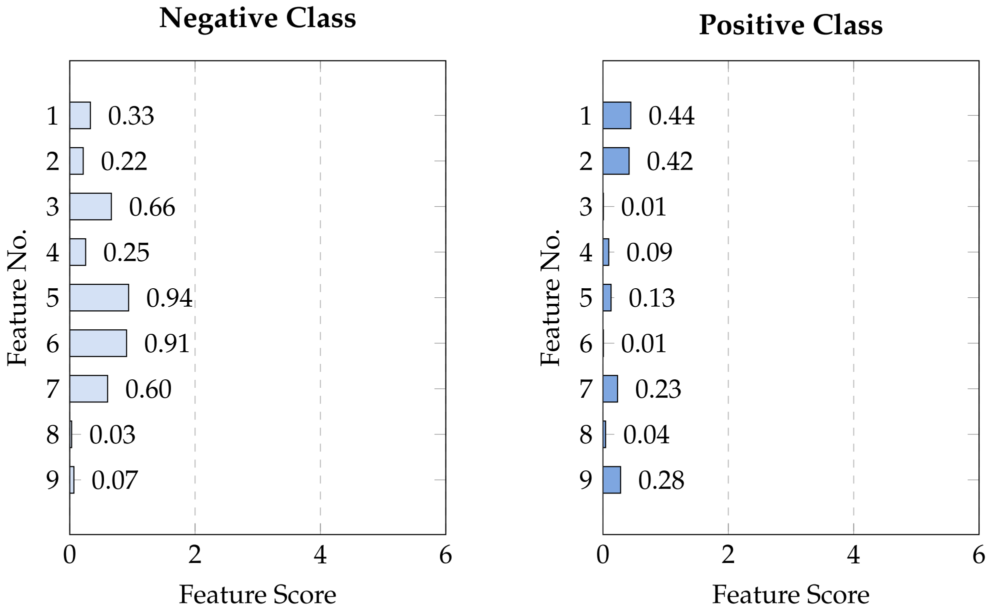


Figure 6.31: BF weld defect detection: Graphical explanation of the *Defective* sub-model - example of a TN case: #2_2_41_19_39. Textual explanation: "Defective model thinks the data is similar to Non-defective (-ve) and NOT similar to Defective (+ve)."

Meanwhile, for non-defective slices, there are examples where negative cases are less obvious than others, i.e. containing high amplitude indications that have a similar shape and size to defects. The first example (Figures 6.29-6.31) demonstrates how the model correctly classifies a negative case. The model was in a similar conflict to the previous TP example. However, in this instance, the defective model's measures indicated a higher certainty; thus, the image was classified as *non-defective*. Based on the feature scores, features 5 and 6 (GCLM Energy and Homogeneity) had a significant impact on the decision for the defective sub-model, thus, the overall model.

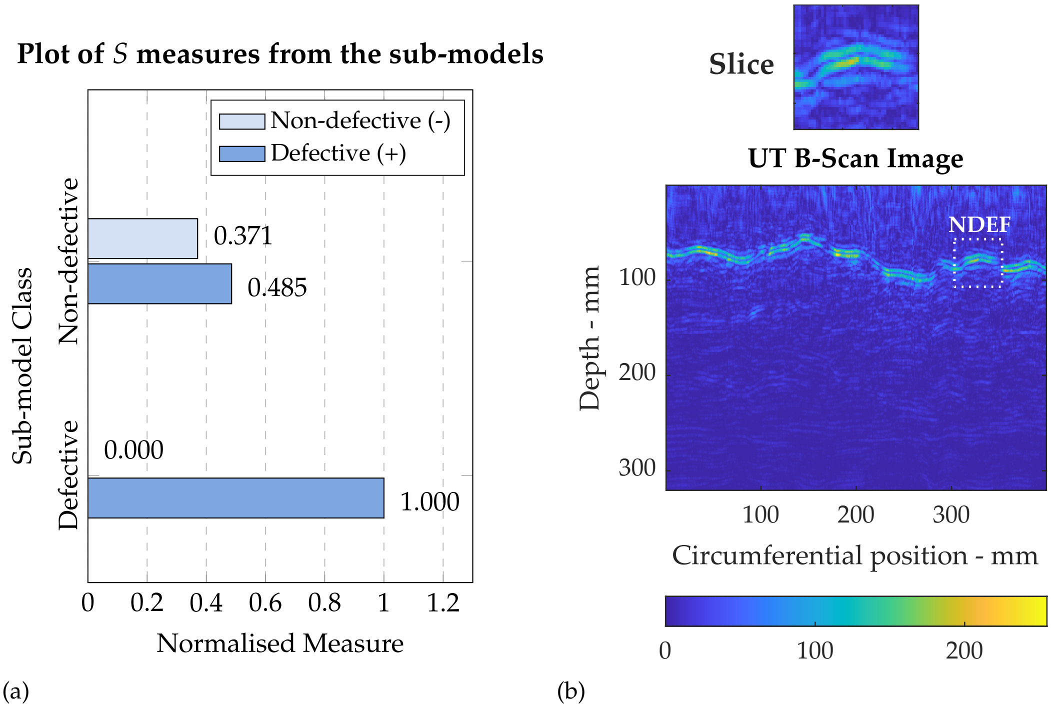


Figure 6.32: BF weld defect detection: (a) Illustration of the distance measures S from the two sub-models - example of a TN case: #2_2_22_3_8. Textual explanation: “Models are in agreement hence, it is more likely the classification is accurate. Fuzzy class: 0.2”. (b) Plot of UT image data for the same example. The image was classified as *non-defective* by the model.

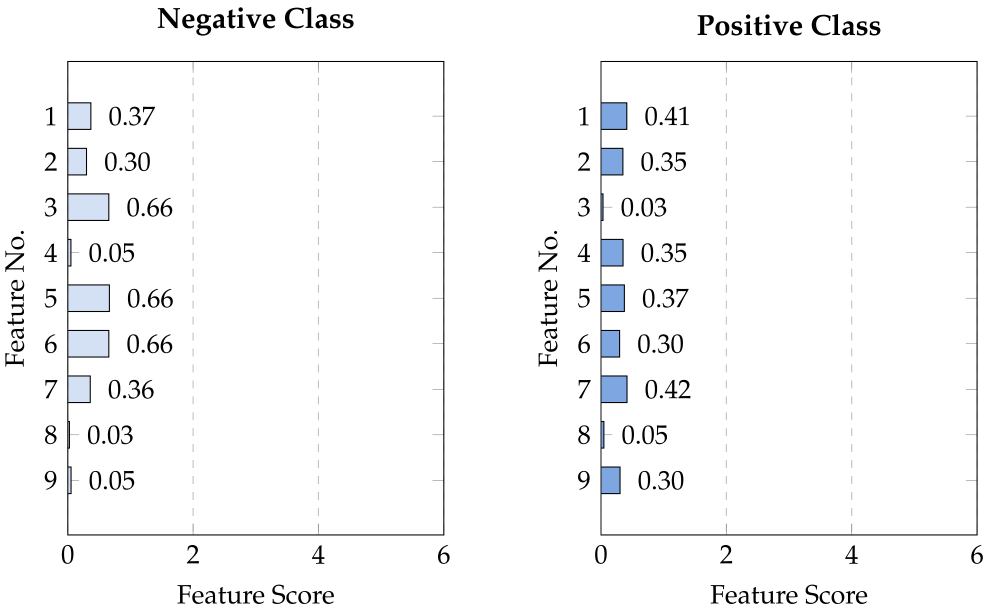


Figure 6.33: BF weld defect detection: Graphical explanation of the *Non-defective* sub-model - example of a TN case: #2_2_22_3_8. Textual explanation: “Non-defective model thinks the data is more similar to Non-defective (-ve) despite a high similarity for both.”

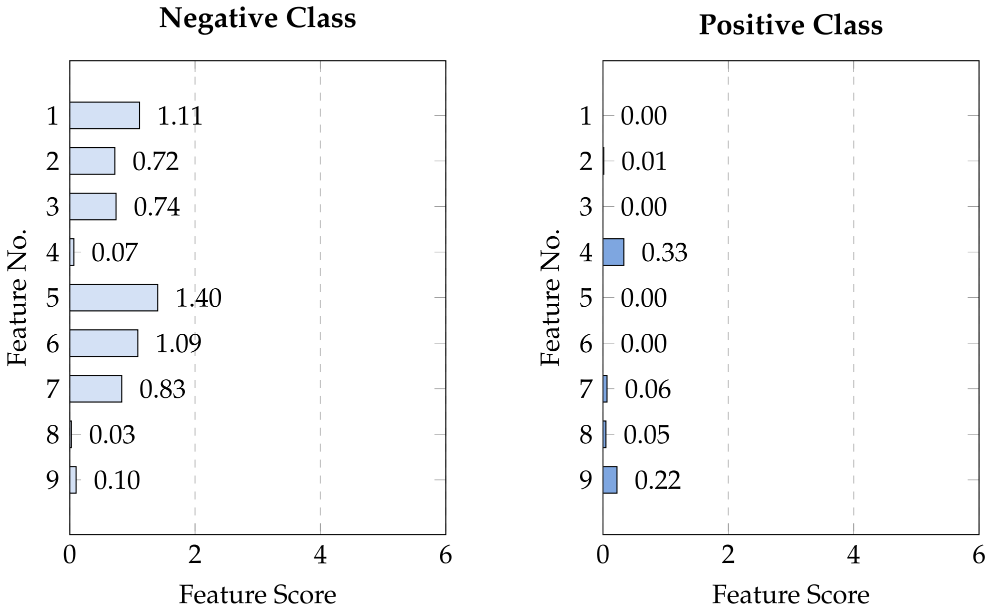


Figure 6.34: BF weld defect detection: Graphical explanation of the *Defective* sub-model - example of a TN case: #2_2_22_3_8. Textual explanation: "Defective model thinks the data is similar to Non-defective (-ve) and NOT similar to Defective (+ve)."

Furthermore, the second TN example (Figures 6.32-6.34) demonstrates how the model deals with a more challenging negative case, for which, the image contains high amplitude indications similar in shape and size to a defect. Contrary to the previous example, the models agree that the image slice is *non-defective*. In addition, as expected since the feature scores are high for the negative class for both the sub-models. As a result, the values of the measures for the non-defective class were *low*, as summarised by a low fuzzy class output of 0.2. The *strong* classification provides insight to the user with regard to the potential accuracy of the result. In this example, the insight aligns with the correctness of the classification.

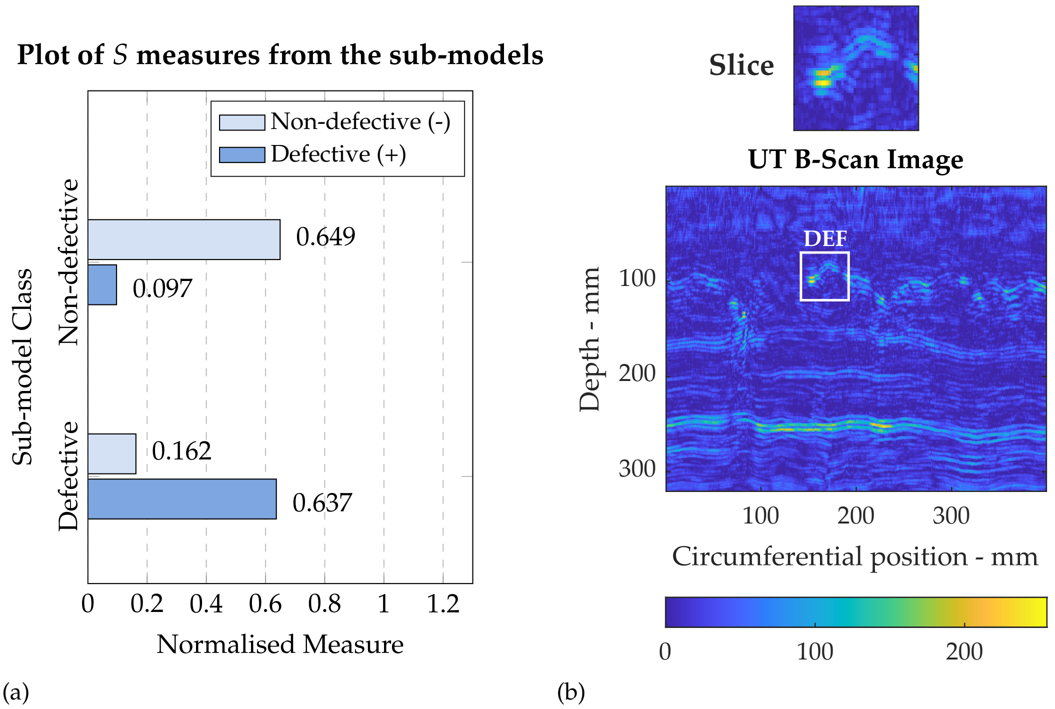


Figure 6.35: BF weld defect detection: (a) Illustration of the distance measures S from the two sub-models - example of a FP case: #2_2_13_2_19. Textual explanation: "Models are in conflict hence, it is less likely the classification is accurate. Fuzzy class: 0.53". (b) Plot of UT image data for the same example. The image was classified as *defective* by the model.

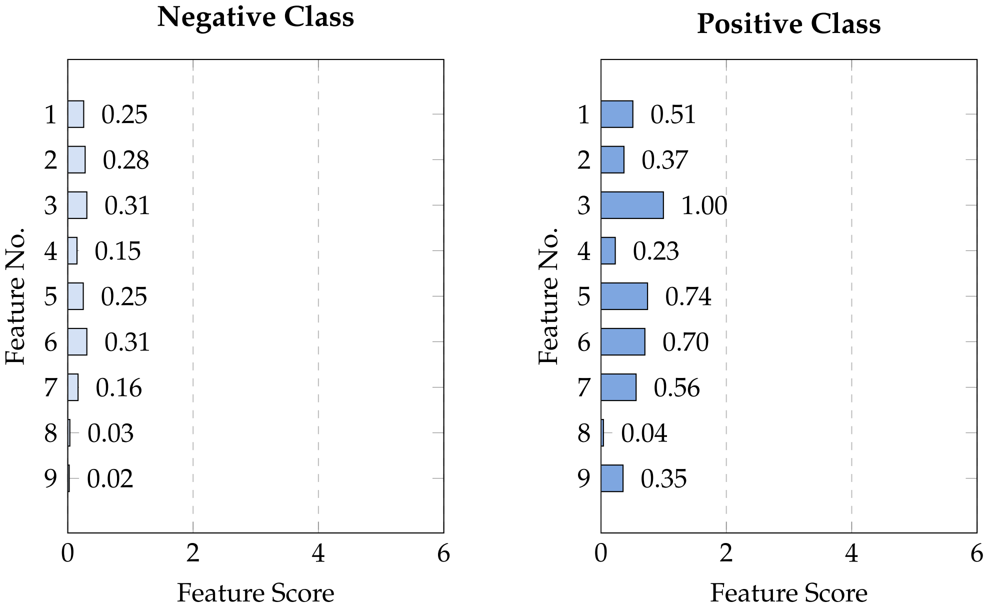


Figure 6.36: BF weld defect detection: Graphical explanation of the *Non-defective* sub-model - example of a FP case: #2_2_13_2_19. Textual explanation: "Non-defective model thinks the data is similar to Defective (+ve) and NOT similar to Non-defective (-ve)."

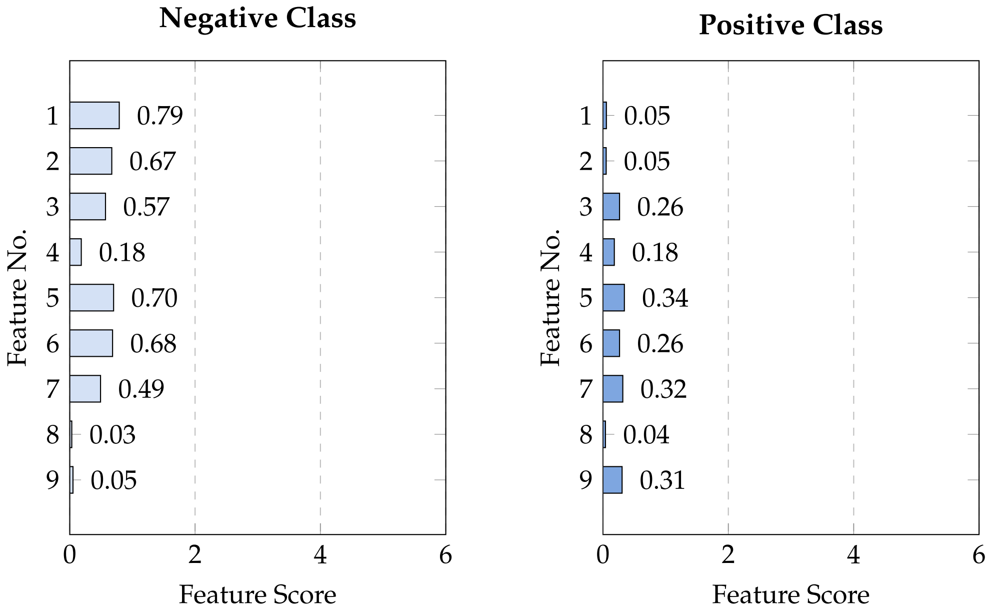
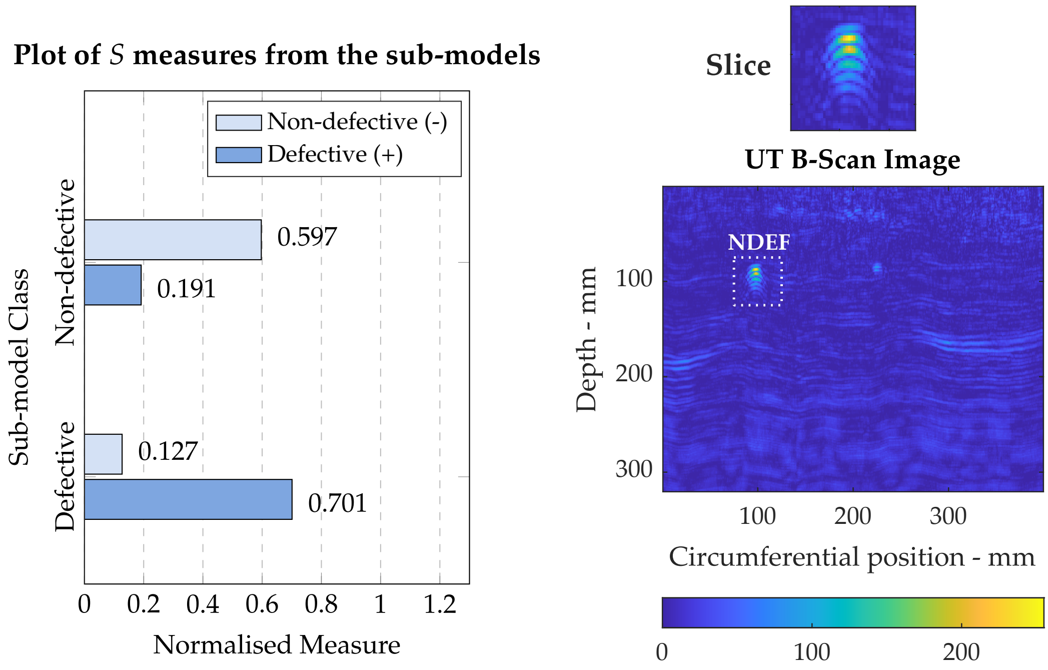


Figure 6.37: BF weld defect detection: Graphical explanation of the *Defective* sub-model - example of a FP case: #2_2_13_2_19. Textual explanation: "Defective model thinks the data is similar to Non-defective (-ve) and NOT similar to Defective (+ve)."

In cases where the model was inaccurate in its prediction, explanation insight could play a vital role in providing some indication of potential uncertainty. There are two types of false cases: negative or positive. Although FP cases require more work to assess the case manually, they are less problematic in advanced manufacturing compared to FN cases, where critical defects could be potentially overlooked. An example of an FP case in Figures 6.35-6.37, demonstrates how high amplitude indications with similar characteristics can fool the model into predicting the image slice as a defect. Despite the inaccurate classification, various indicators point to a lower level of *certainty*. For instance, the fuzzy class has a value relatively close to the threshold at 0.53.



(a)

(b)

Figure 6.38: BF weld defect detection: (a) Illustration of the distance measures S from the two sub-models - example of a FN case: #2_2_111_20_56. Textual explanation: "Models are in conflict hence, it is less likely the classification is accurate. Fuzzy class: 0.44". (b) Plot of UT image data for the same example. The image was classified as *non-defective* by the model.

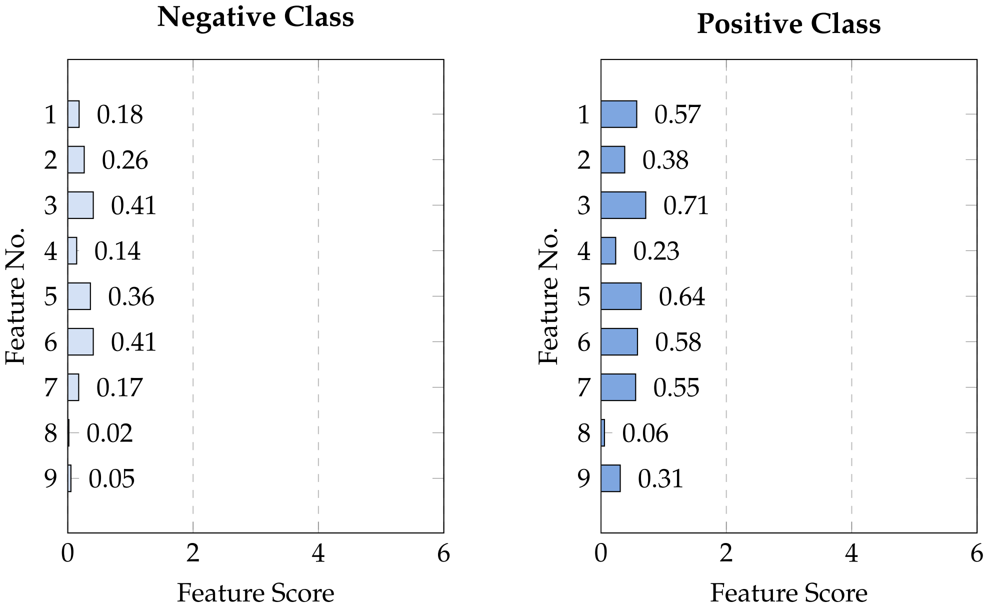


Figure 6.39: BF weld defect detection: Graphical explanation of the *Non-defective* sub-model - example of a FN case: #2_2_111_20_56. Textual explanation: "Non-defective model thinks the data is similar to Defective (+ve) and NOT similar to Non-defective (-ve)."

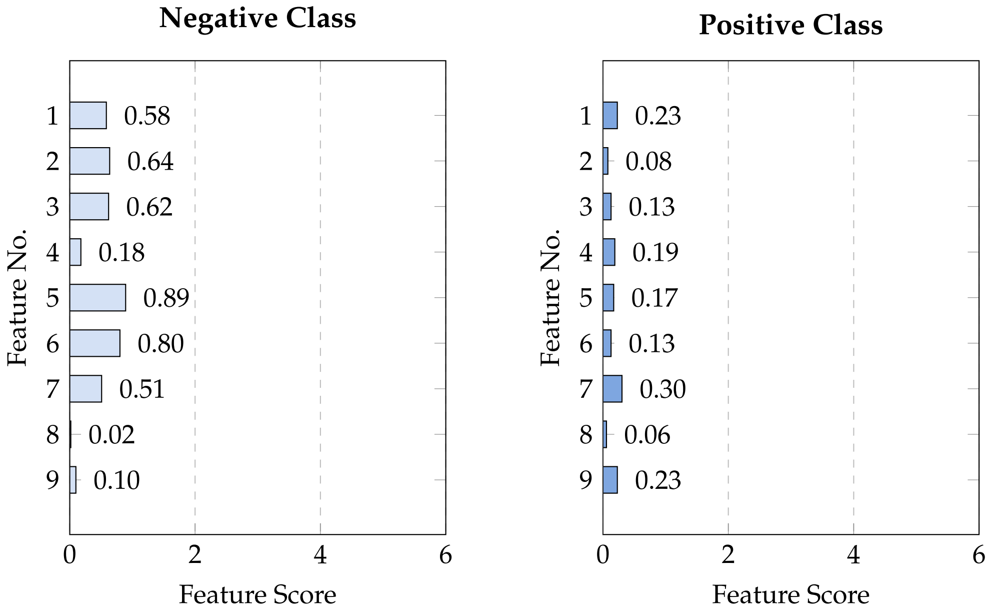


Figure 6.40: BF weld defect detection: Graphical explanation of the *Defective* sub-model - example of a FN case: #2_2_111_20_56. Textual explanation: "Defective model thinks the data is similar to Non-defective (-ve) and NOT similar to Defective (+ve)."

Similarly, the FN example (Figures 6.38-6.40) has seen a conflict between the sub-models and fuzzy class close to the threshold, as shown in Figure 6.38. The culprit behind this inaccuracy is apparent by examining the feature scores for the deciding sub-model - defective, in Figure 6.40. Despite the slice image containing a highly visible defect-shaped indication, its associated criteria resemble values of a typical image that does not contain any defects. Features related to the amplitude, GCLM and edges all had a significant impact. Even though the non-defective model still classified the image as a defect, the weaker relative classification meant the defective model was the deciding model.

6.4.3 Explanation analysis

Naturally, different aspects of the explanation are statistically associated with how accurate the model's prediction is. In this part, we present an analysis of how the aspects relate to performance for the two areas: bead and defect recognition.

For the bead dataset, analysis of true-vs-false cases, in Figure 6.41, TP and TN cases are more likely to result in an *agreement* between the two sub-models (65.2%), while FP and FN cases are more likely to result in a *conflict* (74.7%). Meanwhile, an analysis

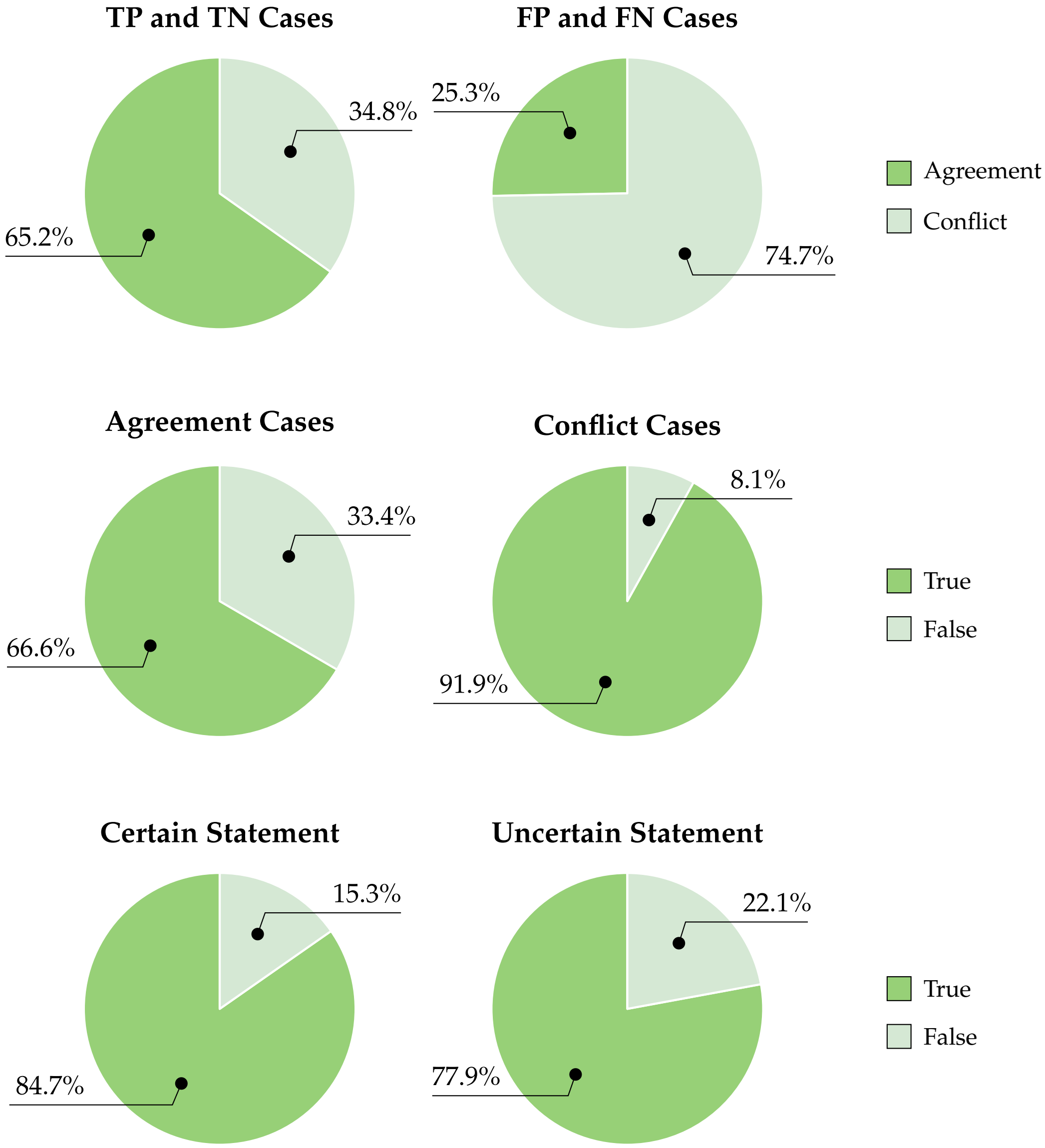


Figure 6.41: Fuzzy-TOPSIS BF Bead Weld Dataset: a set of pie charts are used to illustrate how *indicative* the different aspects of explanation are to false or negative cases.

of agreement-vs-conflict illustrated that a higher proportion of cases are likely to be true when the sub-models are in conflict versus agreement - a counter-intuitive association. Comparatively, the certainty of the linguistic statement had an intuitive association, where images described with *certain* statements by at least one of the sub-models are more likely to be *true*. In contrast, for the defect dataset, images described with *uncertain* statements were more likely to be true. Nonetheless, the defect dataset statistics follow a similar trend for the remaining two aspects of the explanation, as shown in Figure 6.42.

The correlation analysis could prove useful in tuning the linguistic terms to match the associated likeliness of accuracy. By doing so, the user would be presented with an explanation that is more likely to be meaningful, insightful and relevant.

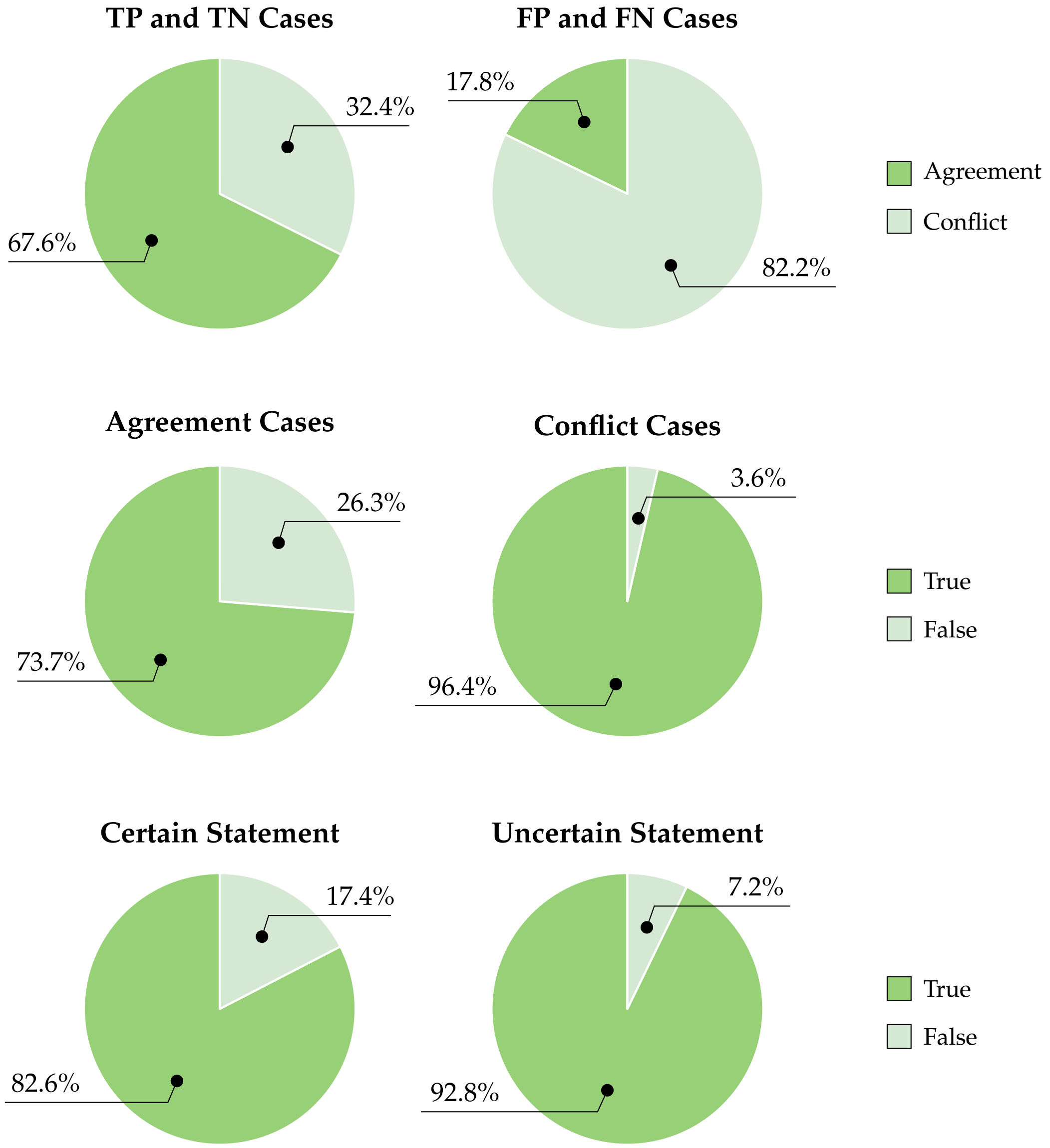


Figure 6.42: Fuzzy-TOPSIS BF Defect Weld Dataset: a set of pie charts is used to illustrate how *indicative* the different aspects of an explanation are to false or negative cases.

Fuzzy-TOPSIS as a newly proposed data-driven explainable framework allows for generating a direct explanation that not only provides much-needed insight into the model output but a means for representative traceability. The main cited con preventing the adoption of model-based interpretability is the suggestion that interpretable models lack performance rigour [77]–[80]. In spite of this, Ruden et al. argue that this trade-off is yet to be proven and stresses that the risks outweigh the loss of performance for high stake applications [2]. This is particularly insightful since a lack of explanation could potentially contribute to increasing the risk of wrongful decision-making, especially in areas where justification is vital for decision-making.

Moreover, following the regulation of plastic pipes, they have been adopted in several safety-critical applications such as water and gas transportation. For this reason, BF pipe weld inspection is considered a high stake application because of the associated potential consequences. When a weld is scanned, it results in hundreds of images that are normally analysed manually by an NDT expert. Automated NDT utilising non-explainable frameworks provides the expert with a list of potential indications lacking any associated explanation. Hence, requires the expert's effort to reanalyse the data manually to provide the justification documentation required. This is where fuzzy-TOPSIS as a framework has been demonstrated to be able to do; present the expert with all the subject matter *criteria* in a format designed to eventually obviate the need for manual analysis, i.e. reaching *autonomous* defect recognition.

6.5 Summary

Applying Fuzzy-TOPSIS classifiers to a real-world industrial problem of BF weld inspection demonstrated once again a case where the performance trade-off is insignificant or nonexistent compared to when using a purpose-built classification framework - K-Means. Moreover, the explanation framework was shown to generate a similar explanation to the ones seen for the benchmark datasets explored in previous chapters. The modelling framework's data-driven nature allowed transferability to this dataset with no modification. The evidence from this study points to the Fuzzy-TOPSIS effectiveness as a data-driven *explained* framework.

In spite of this, the explanation's dependence on model accuracy means that the insight can occasionally be misleading. Strategies to enhance the performance might involve:

- Improving the Fuzzy-TOPSIS fitting methodology

- Developing a framework for exploiting statistical associations to extract more aligned explanations
- Exploring methodologies of optimising data selection to better suit inherently interpretable models

Moreover, the current framework has only been explored for binary classification. It might be worth investigating the avenue for scalability in a multi-class problem, which could prove useful for tackling more complex classification problems. This could be potentially achieved by expanding all the individual model components (FIS, TOPSIS and explanation) to incorporate the support of several classes. Multi-class support is expected to lead to an increase in the model's dimensionality. Dimensionality is a key metric for maintaining interpretability. A reasonable approach to tackle this issue could be to utilise data aggregation techniques to streamline the data into a dimensionality that is *human-comprehensible*.

Another possible area of future research would be to investigate why interpretable models are seldom investigated in the industry despite the growing interest from academia and governmental organisations [70]–[72], [84], [121].

The advantages of applying an interpretable classification framework in advanced manufacturing provide opportunities for accessing meaningful insight to assist users and model designers.

Despite the benefits, MCDM-based classifiers are not expected to achieve comparable performance when paired with complex datasets. Performance is not the sole goal. However, it is considered paramount to any modelling methodology's success.

7 Conclusions and Recommendations for Future Work

7.1 Summary

Contrary to popular belief, interpretable models can achieve satisfactory performance compared to opaque models, given a chance [2]. An example of interpretable models performing satisfactorily had been presented. MCDM and fuzzy-MCDM classifiers had been investigated using benchmark datasets.

Although performance varied widely across the different datasets, there had been cases where the MCDM-based classifier achieved similar performance to the state-of-the-art classifier.

However, the whole point of opting for MCDM had been to harness its interpretability and potential explainability. An explanation framework is proposed for the fuzzy-TOPSIS classifier.

The explanation framework taps into the model's internal parameters to extract interpretable information. The key features influencing the decision are indicated via the graphical and textual explanation generated. The decomposable components were explained by separate textual statements. The chapter demonstrated what had been possible after harnessing a transparent model for the rationale of optimising explanation.

Furthermore, it had been demonstrated that the proposed explanation framework developed for fuzzy-MCDM classifiers provided valuable insight into the decision-making process. Textual statements described the sub-models' decisions following by information on whether they were in *conflict* or *agreement*. The fuzzy class output provided a summary of the overall classification with the threshold, which indicated the decision's statistical certainty. Applying the explanation framework to human-understandable datasets showcased its capability to pinpoint the key features that led to a certain decision. From a practical point of view, accurately distinguishing the data that led to a specific decision is one of the main purposes of XAI.

Applying the explanation framework to non-human understandable datasets such as Parkinson's disease and chess yielded similar results. However, the high dimensionality of the datasets meant the graphical explanation could only display a subset of the most influential features. Hence, the framework's effectiveness is limited by the degree of comprehensibility offered by the feature-set. Although the framework had been able to find the most impactful features, this information had been only as useful as the user's understanding of the features. Therefore, the framework's applicability and effectiveness are dependent upon the availability of human-understandable feature-sets that adequately represent the problem at hand.

The analysis of the various explanation aspects associated with performance accuracy revealed a relation that can be potentially exploited to provide valuable additional information to the user. For instance, whether the models are in *agreement* or *conflict* can be statistically associated with a higher probability of accuracy.

The explanation framework provided factual and counterfactual information by presenting key variables from the two sub-models. The measures indicated how factual or counterfactual a classification had been. The FL component summarised this information into a single fuzzy class output value. Despite the fuzzy output being a continuous number, it did not provide a breakdown of the factual and counterfactual components.

Relativity presented in the form of counterfactual explanation enhanced the comprehensibility experience. Humans seek counterfactual explanations naturally [15].

NL was seen as a potential solution to this gap. As opposed to type-1 FL, NL employs three components: truth, indeterminacy and falsity. In Chapter 5, the

TOPSIS classifier is extended with a NIS. The rules are configured such that factual and counterfactual indications are reflected in the truth, indeterminacy and falsity. Similarly to the fuzzy extension, Neutrosophic-TOPSIS had been involved in the final classification.

Each classification result presented by the Neutrosophic-TOPSIS model included factual and counterfactual information summarised concisely by two figures: truth and falsity.

Performance of neutrosophic-TOPSIS had been largely similar to fuzzy-TOPSIS. Hence, it is a clear improvement over the framework initially proposed in previous chapters.

In parallel, the proposed methodologies were investigated for two real-world industrial case studies: bead and defect detection. Bead indications are used in practice to detect the position of the pipe in the UT image.

The chapter demonstrated the applicability of interpretable modelling to real-world industrial problems. The UT inspection techniques utilised for data collection aim to detect flaws in safety-critical plastic pipelines.

The application is considered high stake because misclassification could lead to *catastrophic* consequences. Catastrophic level incidents are the highest severity and could entail a loss of life among widespread consequences such as significant environmental pollution and damage to assets. The sort of consequences that cause major harm to society.

Hence, it is imperative that ML models designed to provide decision support in safety-critical areas such as these provide transparency. The purpose of transparency is to reduce the chances of hidden biases and enable experts to understand the limitations of the model more intuitively. The proposed classifiers possess a great deal of transparency in their fitting algorithm and execution process.

Nonetheless, opting for Neutrosophic Logic and MCDM-based classifiers still has its limitations. Despite the benefits, the methodology in its current form lacks the *training* rigour achievable through state-of-the-art classifiers. Therefore, its classification performance can be quickly limited when dealing with complex datasets. In addition, its simple structure prevents it from handling datasets with many features; an example is the Parkinson's disease dataset. As a result, the methodology's applicability is not as wide as state-of-the-art classifiers.

The feature scores illustrated the impact of each feature to the user. However, in its current representation, it is sometimes not immediately clear which smaller set of features had been the tipping point for a decision. Highlighting the most impactful features in a more easily distinguishable way would make the explanation more comprehensible to the user [15].

The NL extension expanded upon the FIS in the previous chapter by providing three components. This increased the number of insights presented to the user. As a result, the complexity of the graphs presented to the user had been increased; thus, this potentially has a negative effect on the comprehensibility of the explanation.

Despite the benefits, MCDM-based classifiers are not expected to achieve comparable performance when paired with complex datasets. Performance is not the sole goal. However, it is considered paramount to any modelling methodology's success.

7.2 Conclusion

The aim of the project set at the start was to develop a data-driven interpretable classification framework, capable of producing meaningful explanation. The aim had been based on the contributions in this thesis achieving objective A of designing an MCDM-based - an interpretable framework for classification. Furthermore, the explanation framework proposed had addressed objective B. The development of an explanation framework for MCDM-based classifiers.

Finally, the frameworks were applied to ADR as per objective C. However, full autonomy was not achievable as part of this project. This entailed designing a robotic pipe inspection device which was not part of the scope the project.

7.3 Future Work

In the new proposed frameworks lie several areas of improvement. In this section, we provide suggestions of how some key limitations could be addressed.

7.3.1 Performance and dimensionality

The performance of the MCDM-based classifiers had been satisfactory for low dimensionality datasets. However, it is expected the classifiers would not perform adequately when presented with complex high dimensionality datasets such as image data because of their simple structure.

When the proposed classifiers are fitted with a large feature set, the single iteration fitting algorithm may not be adequate. Exploration of a learning algorithm is vital to the classifier's success in modelling larger or complex datasets. However, it is recommended to undertake caution when devising the algorithm to avoid introducing significant algorithmic opacity.

7.3.2 Interpretability and explainability metrics

The explanation frameworks proposed provided a proof-of-concept implementation. The frameworks presented internal model parameters in a manner that enabled the visualisation of feature impact. Moreover, the textual explanation presented key factual and counterfactual information. The explanations painted a complete picture of the different components of the model with little regard for human comprehension.

The human component of explanation effectiveness must be explored more extensively to assess the framework's suitability to be implemented in practice. Coming up with an explanation metric that can be used for optimisation is ideal. However, the explanation's subjective nature is the main obstacle to this.

Therefore, the first step is to develop an interpretability metric for MCDM classifiers. The metric will serve as an optimisation parameter that can be used for a potential training algorithm.

7.3.3 Multi-class support

The proposed models serve the purpose of experimentation using binary datasets. However, in practice exist areas where multi-class support is needed. Developing the proposed models for handling more than two classes would be useful in widening its applicability.

Implementing a multi-class MCDM model means the introduction of additional parameters through sub-models. The extension is expected to yield further challenges for interpretability. Moreover, an optimisation algorithm for learning becomes more vital to performance success.

References

[1] M. Shirer, "IDC Forecasts Companies to Increase Spend on AI Solutions by 19.6% in 2022," International Data Corporation, Tech. Rep., Feb. 2022.

[2] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, May 2019. DOI: 10.1038/s42256-019-0048-x.

[3] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, "Definitions, methods, and applications in interpretable machine learning," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 116, no. 44, pp. 22 071–22 080, Oct. 2019. DOI: 10.1073/PNAS.1900654116.

[4] *FAA Statement on Pratt & Whitney Engine Emergency Airworthiness Directive | Federal Aviation Administration*, Feb. 2021.

[5] *Engines of 787 in 2019 Rome incident had dozens of cracked blades: inquiry | News | Flight Global*, Jan. 2022.

[6] G. McGivern and M. Fischer, "Medical regulation, spectacular transparency and the blame business," *Journal of health organization and management*, vol. 24, no. 6, pp. 597–610, Nov. 2010. DOI: 10.1108/14777261011088683.

[7] Y. Li, L. Ma, L. Shen, J. Lv, and P. Zhang, "Open source software security vulnerability detection based on dynamic behavior features," *PLOS ONE*, vol. 14, no. 8, e0221530, Aug. 2019. DOI: 10.1371/journal.pone.0221530.

[8] G. Schryen and R. Kadura, "Open source vs. closed source software," in *Proceedings of the 2009 ACM symposium on Applied Computing - SAC '09*, New York, New York, USA: ACM Press, 2009, pp. 2016–2023, ISBN: 9781605581668. DOI: 10.1145/1529282.1529731.

[9] S. F. Wen, M. Kianpour, and S. Kowalski, "An empirical study of security culture in open source software communities," in *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2019*, Association for Computing Machinery, Inc, Aug. 2019, pp. 863–870, ISBN: 9781450368681. DOI: 10.1145/3341161.3343520.

[10] S. Kraft, *Widerrede - Klarheit schaffen (translation: Contradiction - create clarity)*, 2012.

[11] *Transparency in healthcare*.

[12] Z. C. Lipton, "The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery," *Queue*, vol. 16, no. 3, pp. 31–57, Jun. 2018. DOI: 10.1145/3236386.3241340.

[13] A. F. Markus, J. A. Kors, and P. R. Rijnbeek, "The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies," *Journal of Biomedical Informatics*, vol. 113, p. 103 655, Jan. 2021. DOI: 10.1016/J.JBI.2020.103655.

[14] L. Baccour, "Amended fused TOPSIS-VIKOR for classification (ATOVIC) applied to some UCI data sets," *Expert Systems with Applications*, vol. 99, pp. 115–125, Jun. 2018. DOI: 10.1016/j.eswa.2018.01.025.

[15] Christoph Molnar, *Interpretable Machine Learning*. Morrisville, NC, USA: Lulu.com, 2020, ISBN: 9780244768522.

[16] M. J. Gacto, R. Alcalá, and F. Herrera, "Interpretability of linguistic fuzzy rule-based systems: An overview of interpretability measures," *Information Sciences*, vol. 181, no. 20, pp. 4340–4360, Oct. 2011. DOI: 10.1016/J.INS.2011.02.021.

[17] B. Kim, R. Khanna, and O. Koyejo, "Examples Are Not Enough, Learn to Criticize! Criticism for Interpretability," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, ser. NIPS'16, Red Hook, NY, USA: Curran Associates Inc., 2016, pp. 2288–2296, ISBN: 9781510838819. DOI: 10.5555/3157096.3157352.

[18] T. Miller, “Explanation in artificial intelligence: Insights from the social sciences,” *Artificial Intelligence*, vol. 267, Feb. 2019. DOI: 10.1016/j.artint.2018.07.007.

[19] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, *et al.*, “Explainable Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI,” *Information Fusion*, vol. 58, pp. 82–115, Jun. 2020. DOI: 10.1016/j.inffus.2019.12.012.

[20] C. Rudin and J. Radin, “Why Are We Using Black Box Models in AI When We Don’t Need To? A Lesson From An Explainable AI Competition,” *Harvard Data Science Review*, vol. 1, no. 2, Nov. 2019. DOI: 10.1162/99608f92.5a8a3a3d.

[21] A. Bibal and B. Frénay, “Interpretability of Machine Learning Models and Representations: an Introduction,” in *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, Apr. 2016.

[22] M. Robnik-Šikonja and M. Bohanec, “Perturbation-Based Explanations of Prediction Models,” in *Human and Machine Learning*, Cham: Springer, 2018, pp. 159–175. DOI: 10.1007/978-3-319-90403-0_9.

[23] R. Poyiadzi, X. Renard, T. Laugel, R. Santos-Rodriguez, and M. Detyniecki, “On the overlooked issue of defining explanation objectives for local-surrogate explainers,” Jun. 2021. DOI: 10.48550/arxiv.2106.05810.

[24] M. T. Ribeiro, S. Singh, and C. Guestrin, “"Why Should I Trust You?": : Explaining the Predictions of Any Classifier,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA: ACM, Aug. 2016, pp. 1135–1144, ISBN: 9781450342322. DOI: 10.1145/2939672.2939778.

[25] P. Lipton, “Contrastive Explanation,” *Royal Institute of Philosophy Supplements*, vol. 27, pp. 247–266, Mar. 1990. DOI: 10.1017/S1358246100005130.

[26] A. Sudjianto and A. Zhang, “Designing Inherently Interpretable Machine Learning Models,” *ArXiv preprint*, Nov. 2021. DOI: 10.48550/arxiv.2111.01743.

[27] R. Wexler, *When a computer program keeps you in jail: how computers are harming criminal justice*, Jun. 2017.

[28] D. Martens, J. Vanthienen, W. Verbeke, and B. Baesens, "Performance of classification models from a user perspective," *Decision Support Systems*, vol. 51, no. 4, pp. 782–793, Nov. 2011. DOI: 10.1016/J.DSS.2011.01.013.

[29] F. K. Dosilovic, M. Brcic, and N. Hlupic, "Explainable artificial intelligence: A survey," in *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2018 - Proceedings*, Institute of Electrical and Electronics Engineers Inc., Jun. 2018, pp. 210–215, ISBN: 9789532330977. DOI: 10.23919/MIPRO.2018.8400040.

[30] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, "Machine Learning Interpretability: A Survey on Methods and Metrics," *Electronics*, vol. 8, no. 8, p. 832, Jul. 2019. DOI: 10.3390/electronics8080832.

[31] R. Elshawi, M. H. Al-Mallah, and S. Sakr, "On the interpretability of machine learning-based model for predicting hypertension," *BMC Medical Informatics and Decision Making*, vol. 19, no. 1, pp. 1–32, Jul. 2019. DOI: 10.1186/S12911-019-0874-0/FIGURES/48.

[32] P. R. Magesh, R. D. Myloth, and R. J. Tom, "An Explainable Machine Learning Model for Early Detection of Parkinson's Disease using LIME on DaTSCAN Imagery," *Computers in Biology and Medicine*, vol. 126, p. 104 041, Nov. 2020. DOI: 10.1016/J.COMPBIOMED.2020.104041.

[33] J. M. Alonso, C. Castiello, and C. Mencar, "A Bibliometric Analysis of the Explainable Artificial Intelligence Research Field," *Communications in Computer and Information Science*, vol. 853, pp. 3–15, Jun. 2018. DOI: 10.1007/978-3-319-91473-2_1.

[34] E. Westkämper, "Digital Manufacturing In The Global Era," in *Digital Enterprise Technology*, Boston, MA: Springer US, 2007, pp. 3–14. DOI: 10.1007/978-0-387-49864-5_1.

[35] H. Yusuf and G. Panoutsos, "Multi-criteria decision making using Fuzzy Logic and ATOVIC with application to manufacturing," in *2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, Glasgow: Institute of Electrical and Electronics Engineers Inc., Jul. 2020, pp. 1–7, ISBN: 9781728169323. DOI: 10.1109/FUZZ48607.2020.9177772.

[36] A. Piegat and W. Sałabun, "Comparative Analysis of MCDM Methods for Assessing the Severity of Chronic Liver Disease," in *ICAISC 2015: Artificial Intelligence and Soft Computing*, L. Rutkowski, M. Korytkowski, R. Scherer, R. Tadeusiewicz, L. A. Zadeh, and J. M. Zurada, Eds., Cham: Springer International Publishing, 2015, pp. 228–238, ISBN: 978-3-319-19324-3. DOI: 10.1007/978-3-319-19324-3_21.

[37] S. D. Pohekar and M. Ramachandran, "Application of multi-criteria decision making to sustainable energy planning - A review," *Renewable and Sustainable Energy Reviews*, vol. 8, no. 4, pp. 365–381, Aug. 2004. DOI: 10.1016/j.rser.2003.12.007.

[38] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A Survey of Methods for Explaining Black Box Models," *ACM Computing Surveys*, vol. 51, no. 5, Jan. 2019. DOI: 10.1145/3236009.

[39] U. Johansson and L. Niklasson, "Evolving decision trees using oracle guides," in *2009 IEEE Symposium on Computational Intelligence and Data Mining, CIDM 2009 - Proceedings*, 2009, pp. 238–244. DOI: 10.1109/CIDM.2009.4938655.

[40] J. Zhang, Y. Wang, P. Molino, L. Li, and D. S. Ebert, "Manifold: A Model-Agnostic Framework for Interpretation and Diagnosis of Machine Learning Models," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 1, Jan. 2019. DOI: 10.1109/TVCG.2018.2864499.

[41] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to data mining*. Pearson Education India, 2006.

[42] R. Krishnan, G. Sivakumar, and P. Bhattacharya, "Extracting decision trees from trained neural networks," *Pattern Recognition*, vol. 32, no. 12, Dec. 1999. DOI: 10.1016/S0031-3203(98)00181-2.

[43] O. Boz, "Extracting decision trees from trained neural networks," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '02*, New York, New York, USA: ACM Press, 2002, ISBN: 158113567X. DOI: 10.1145/775047.775113.

[44] R. Turner, "A Model Explanation System: Latest Updates and Extensions," *Black Box Learning and Inference (NIPS Workshop)*, Jun. 2015.

[45] S. Krishnan and E. Wu, "PALM: Machine Learning Explanations For Iterative Debugging," in *Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics*, New York, NY, USA: ACM, May 2017, ISBN: 9781450350297. DOI: 10.1145/3077257.3077271.

[46] M. T. Ribeiro, S. Singh, and C. Guestrin, "Nothing Else Matters: Model-Agnostic Explanations By Identifying Prediction Invariance," in *30th Conference on Neural Information Processing Systems (NIPS 2016)*, Barcelona, Nov. 2016.

[47] D. Baehrens, S. Harmeling, M. Kawanabe, K. Hansen Khansen, and C. Edward Rasmussen, "How to Explain Individual Classification Decisions Timon Schroeter," *Journal of Machine Learning Research*, vol. 11, pp. 1803–1831, 2010. DOI: 10.5555/1756006.

[48] A. Henelius, K. Puolamäki, H. Boström, L. Asker, and P. Papapetrou, "A peek into the black box: exploring classifiers by randomization," *Data Mining and Knowledge Discovery 2014 28:5*, vol. 28, no. 5, pp. 1503–1529, Jul. 2014. DOI: 10.1007/S10618-014-0368-8.

[49] P. Adler, C. Falk, S. A. Friedler, *et al.*, "Auditing black-box models for indirect influence," *Knowledge and Information Systems 2017 54:1*, vol. 54, no. 1, pp. 95–122, Oct. 2017. DOI: 10.1007/S10115-017-1116-3.

[50] M. T. Ribeiro, S. Singh, and C. Guestrin, "Model-Agnostic Interpretability of Machine Learning," Tech. Rep., 2016.

[51] M. Velez, P. Jamshidi, N. Siegmund, S. Apel, and C. Kastner, "White-Box Analysis over Machine Learning: Modeling Performance of Configurable Systems," in *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*, Institute of Electrical and Electronics Engineers (IEEE), May 2021, pp. 1072–1084. DOI: 10.1109/ICSE43902.2021.00100.

[52] F. Gille, A. Jobin, and M. Ienca, "What we talk about when we talk about trust: Theory of trust for AI in healthcare," *Intelligence-Based Medicine*, vol. 1-2, p. 100 001, Nov. 2020. DOI: 10.1016/J.IBMED.2020.100001.

[53] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A Survey on Bias and Fairness in Machine Learning," *ACM Computing Surveys*, vol. 54, no. 6, pp. 1–35, Jul. 2021. DOI: 10.1145/3457607.

[54] D. Danks and A. J. London, “Algorithmic bias in autonomous systems,” in *IJCAI International Joint Conference on Artificial Intelligence*, vol. 0, International Joint Conferences on Artificial Intelligence, 2017, pp. 4691–4697. DOI: 10.24963/IJCAI.2017/654.

[55] Y.-R. Baeza, “Bias on the web,” *Communications of the ACM*, vol. 61, no. 6, pp. 54–61, May 2018. DOI: 10.1145/3209581.

[56] L. A. Zadeh, “Fuzzy sets,” *Information and Control*, vol. 3, pp. 338–353, 1965.

[57] D. Pekaslan, C. Chen, C. Wagner, and J. M. Garibaldi, “Performance and Interpretability in Fuzzy Logic Systems – Can We Have Both?” *Communications in Computer and Information Science*, vol. 1237, pp. 571–584, Jun. 2020. DOI: 10.1007/978-3-030-50146-4_42.

[58] B. Bouchon-Meunier, A. Laurent, and M. Lesot, “XAI: a natural application domain for fuzzy set theory,” in *Women in Computational Intelligence*, A. E. Smith, Ed., 1st ed., Cham, Switzerland: Springer, 2021, ch. 2, ISBN: 978-3-030-79091-2.

[59] C. Marsala and B. Bouchon-Meunier, “Fuzzy data mining and management of interpretable and subjective information,” *Fuzzy Sets and Systems*, vol. 281, no. C, pp. 252–259, Dec. 2015. DOI: 10.1016/J.FSS.2015.08.021.

[60] J. Casillas, O. Cordón, F. Herrera, and L. Magdalena, “Interpretability Improvements to Find the Balance Interpretability-Accuracy in Fuzzy Modeling: An Overview,” in *Interpretability Issues in Fuzzy Modeling*, J. Casillas, O. Cordón, F. Herrera, and L. Magdalena, Eds., vol. 128, Springer, Berlin, Heidelberg, 2003, ch. 1, pp. 3–22. DOI: 10.1007/978-3-540-37057-4_1.

[61] J. M. Alonso, C. Castiello, and C. Mencar, “Interpretability of Fuzzy Systems: Current Research Trends and Prospects,” *Springer Handbook of Computational Intelligence*, pp. 219–237, Jan. 2015. DOI: 10.1007/978-3-662-43505-2_14.

[62] N. Ghorui, A. Ghosh, S. P. Mondal, *et al.*, “Identification of dominant risk factor involved in spread of COVID-19 using hesitant fuzzy MCDM methodology,” *Results in Physics*, vol. 21, p. 103 811, Feb. 2021. DOI: 10.1016/J.RINP.2020.103811.

[63] T. C. Chu and Y. Lin, “An extension to fuzzy MCDM,” *Computers & Mathematics with Applications*, vol. 57, no. 3, pp. 445–454, Feb. 2009. DOI: 10.1016/J.CAMWA.2008.10.076.

[64] C. Carlsson and R. Fullér, “Fuzzy multiple criteria decision making: Recent developments,” *Fuzzy Sets and Systems*, vol. 78, no. 2, pp. 139–153, Mar. 1996. DOI: 10.1016/0165-0114(95)00165-4.

[65] G. A. Miller, “The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information,” *Psychological Review*, vol. 101, no. 2, pp. 343–352, 1994. DOI: 10.1037/0033-295X.101.2.343.

[66] B. h. Li, B. c. Hou, W. t. Yu, X. b. Lu, and C. w. Yang, “Applications of artificial intelligence in intelligent manufacturing: a review,” *Frontiers of Information Technology and Electronic Engineering*, vol. 18, no. 1, pp. 86–96, Jan. 2017. DOI: 10.1631/FITEE.1601885.

[67] J. W. Goodell, S. Kumar, W. M. Lim, and D. Pattnaik, “Artificial intelligence and machine learning in finance: Identifying foundations, themes, and research clusters from bibliometric analysis,” *Journal of Behavioral and Experimental Finance*, vol. 32, Dec. 2021. DOI: 10.1016/J.JBEF.2021.100577.

[68] F. Jiang, Y. Jiang, H. Zhi, *et al.*, “Artificial intelligence in healthcare: Past, present and future,” *Stroke and Vascular Neurology*, vol. 2, no. 4, pp. 230–243, Dec. 2017. DOI: 10.1136/SVN-2017-000101.

[69] H. Prakken and G. Sartor, “Law and logic: A review from an argumentation perspective,” *Artificial Intelligence*, vol. 227, pp. 214–245, Jul. 2015. DOI: 10.1016/J.ARTINT.2015.06.005.

[70] “Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Da,” *Official Journal of the European Union*, pp. 1–88, May 2016.

[71] *Guidelines for AI procurement - GOV.UK*, Jun. 2020.

[72] M. Turek, *Explainable Artificial Intelligence*, 2021.

[73] O. Gillath, T. Ai, M. Branicky, S. Keshmiri, R. Davison, and R. Spaulding, "Attachment and trust in artificial intelligence," *Computers in Human Behavior*, vol. 115, p. 106 607, Feb. 2021. DOI: 10.1016/J.CHB.2020.106607.

[74] D. Diakoulaki, G. Mavrotas, and L. Papayannakis, "Determining objective weights in multiple criteria problems: The critic method," *Computers & Operations Research*, vol. 22, no. 7, pp. 763–770, Aug. 1995. DOI: 10.1016/0305-0548(94)00059-H.

[75] J. Alcala-Fdez, A. Fernández, J. Luengo, *et al.*, "KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework," *Journal of Multiple-Valued Logic and Soft Computing*, vol. 17, pp. 255–287, Jan. 2010.

[76] D. Dua and C. Graff, *UCI Machine Learning Repository*, 2021.

[77] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, Aug. 2015, pp. 1721–1730. DOI: 10.1145/2783258.2788613.

[78] G. Dziugaite, S. Ben-David, and D. M. Roy, "Enforcing Interpretability and its Statistical Impacts: Trade-offs between Accuracy and Interpretability," 2020.

[79] H. Lakkaraju, S. H. Bach, and J. Leskovec, "Interpretable Decision Sets: A joint framework for description and prediction," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA: ACM, Aug. 2016, ISBN: 9781450342322. DOI: 10.1145/2939672.2939874.

[80] E. Choi, M. Bahadori, J. Kulas, A. Schuetz, W. Stewart, and J. Sun, "RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism," in *Advances in Neural Information Processing Systems*, 2016, pp. 3512–3520.

[81] J. Heo, H. B. Lee, S. Kim, *et al.*, "Uncertainty-Aware Attention for Reliable Interpretation and Prediction," in *Proceedings of the 32nd International*

Conference on Neural Information Processing Systems, ser. NIPS'18, Red Hook, NY, USA: Curran Associates Inc., 2018, pp. 917–926.

[82] N. Gill, P. Hall, K. Montgomery, and N. Schmidt, "A Responsible Machine Learning Workflow with Focus on Interpretable Models, Post-hoc Explanation, and Discrimination Testing," *Information*, vol. 11, no. 3, pp. 1–32, 2020. DOI: 10.3390/info11030137.

[83] M. Moradi and M. Samwald, "Post-hoc explanation of black-box classifiers using confident itemsets," *Expert Systems with Applications*, vol. 165, p. 113 941, Mar. 2021. DOI: 10.1016/j.eswa.2020.113941.

[84] High-Level Expert Group on Artificial Intelligence, "Ethics guidelines for trustworthy AI," European Comission, Brussels, Tech. Rep., Apr. 2019, pp. 1–41.

[85] T. E. o. E. Britannica, *Comprehension*, Aug. 2011.

[86] H. H. Maung, "Diagnosis and causal explanation in psychiatry," *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, vol. 60, pp. 15–24, Dec. 2016. DOI: 10.1016/J.SHPSC.2016.09.003.

[87] J. F. Ha and N. Longnecker, "Doctor-patient communication: a review," *Ochsner Journal*, vol. 10, no. 1, pp. 38–43, 2010.

[88] B. Kim, E. L. Glassman, B. Johnson, and J. Shah, "iBCM: Interactive Bayesian Case Model Empowering Humans via Intuitive Interaction," *Computer Science*, Apr. 2015.

[89] B. Letham, C. Rudin, T. H. McCormick, and D. Madigan, "Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model," *https://doi.org/10.1214/15-AOAS848*, vol. 9, no. 3, pp. 1350–1371, Sep. 2015. DOI: 10.1214/15-AOAS848.

[90] B. Kim, C. Rudin, and J. Shah, "The Bayesian case model: A generative approach for case-based reasoning and prototype classification," *Advances in Neural Information Processing Systems*, vol. 3, no. January, pp. 1952–1960, 2014.

[91] E. Walter, *Cambridge Advanced Learner's Dictionary*. Cambridge: Cambridge University Press, 2008.

[92] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G. Z. Yang, "XAI—Explainable artificial intelligence," *Science Robotics*, vol. 4, no. 37, Dec. 2019. DOI: 10.1126/SCIROBOTICS.AAY7120.

[93] B. Bouchon-Meunier, M.-J. Lesot, and C. Marsala, "Lotfi A. Zadeh, the visionary in Explainable Artificial Intelligence," *TWMS J. Pure Applied Math*, vol. 12, no. 1, pp. 5–13, Apr. 2021.

[94] K. Främling, "Decision Theory Meets Explainable AI," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12175 LNAI, pp. 57–74, May 2020. DOI: 10.1007/978-3-030-51924-7_4.

[95] A. Mardani, A. Jusoh, K. MD Nor, Z. Khalifah, N. Zakwan, and A. Valipour, "Multiple criteria decision-making techniques and their applications – a review of the literature from 2000 to 2014," *Economic Research*, vol. 28, no. 1, pp. 516–571, Jan. 2015. DOI: 10.1080/1331677X.2015.1075139.

[96] M. Behzadian, S. Khanmohammadi Otaghsara, M. Yazdani, and J. Ignatius, "A state-of the-art survey of TOPSIS applications," *Expert Systems with Applications*, vol. 39, no. 17, pp. 13 051–13 069, Dec. 2012. DOI: 10.1016/J.ESWA.2012.05.056.

[97] J. Chai, J. N. Liu, and E. W. Ngai, "Application of decision-making techniques in supplier selection: A systematic review of literature," *Expert Systems with Applications*, vol. 40, no. 10, pp. 3872–3885, Aug. 2013. DOI: 10.1016/J.ESWA.2012.12.040.

[98] A. D. Shapiro, "The role of structured induction in expert systems," Ph.D. dissertation, The University of Edinburgh, Edinburgh, 1984, pp. 1–282.

[99] Florentin Smarandache, *A Unifying Field in Logics: Neutrosophic Logic. Neutrosophy, Neutrosophic Set, Neutrosophic Probability*. Rehoboth, NM: American Research Press, 1999.

[100] F. Smarandache, *A Unifying Field in Logics: Neutrosophic Logic. Neutrosophy, Neutrosophic Set, Neutrosophic Probability and Statistics*, 5th ed. Apr. 2006, ISBN: 978-1-59973-991-5. DOI: 10.5281/ZENODO.49174.

[101] Florentin Smarandache, *Plithogeny, Plithogenic Set, Logic, Probability, and Statistics*. arXiv, 2018. DOI: 10.48550/ARXIV.1808.03948.

[102] X. Peng and J. Dai, "A bibliometric analysis of neutrosophic set: two decades review from 1998 to 2017," *Artificial Intelligence Review 2018 53:1*, vol. 53, no. 1, pp. 199–255, Aug. 2018. DOI: 10.1007/S10462-018-9652-0.

[103] R. Şahin and A. Küçük, "On similarity and entropy of neutrosophic soft sets," *Journal of Intelligent & Fuzzy Systems*, vol. 27, no. 5, pp. 2417–2430, 2014. DOI: 10.3233/IFS-141211.

[104] P. Chi and P. Liu, " An extended TOPSIS method for the multiple attribute decision making problems based on interval neutrosophic set," *Neutrosophic Sets and Systems*, vol. 1, pp. 63–70, 2013.

[105] S. Broumi, J. Ye, and F. Smarandache, "An Extended TOPSIS Method for Multiple Attribute Decision Making based on Interval Neutrosophic Uncertain Linguistic Variables," *Neutrosophic Sets and Systems*, vol. 8, no. 1, pp. 1–31, Jan. 2015.

[106] P. Biswas, S. Pramanik, and B. C. Giri, "TOPSIS method for multi-attribute group decision-making under single-valued neutrosophic environment," *Neural Computing and Applications*, vol. 27, no. 3, pp. 727–737, Apr. 2016. DOI: 10.1007/S00521-015-1891-2/TABLES/7.

[107] P. P. Dey, S. Pramanik, and B. C. Giri, "TOPSIS for Solving Multi-Attribute Decision Making Problems under Bi-Polar Neutrosophic Environment," in *New Trends in Neutrosophic Theory and Applications*, F. Smarandache and S. Pramanik, Eds., 1st ed., arXiv, 2016, pp. 65–77.

[108] P. Dey, S. Pramanik, and B. Giri, "Generalized neutrosophic soft multi-attribute group decision making based on TOPSIS," *Critical Review*, vol. 11, pp. 41–55, 2015.

[109] S. Nădăban and S. Dzitac, "Neutrosophic TOPSIS: A general view," in *2016 6th International Conference on Computers Communications and Control, ICCCC 2016*, Institute of Electrical and Electronics Engineers Inc., Jun. 2016, pp. 250–253, ISBN: 9781509017355. DOI: 10.1109/ICCCC.2016.7496769.

[110] A. Elhassouny and F. Smarandache, "Neutrosophic-simplified-TOPSIS multi-criteria decision-making using combined simplified-TOPSIS method and neutrosophics," in *2016 IEEE International Conference on Fuzzy Systems, FUZZ-IEEE 2016*, Institute of Electrical and Electronics Engineers Inc., Nov.

2016, pp. 2468–2474, ISBN: 9781509006250. DOI: 10.1109/FUZZ-IEEE.2016.7738003.

[111] S. D. Surapati Pramanik, "GRA BASED MULTI CRITERIA DECISION MAKING IN GENERALIZED NEUTROSOPHIC SOFT SET ENVIRONMENT," *Global Journal of Engineering Science and Research Management*, vol. 3, no. 5, pp. 153–169, May 2016. DOI: 10.5281/ZENODO.53753.

[112] A. Q. Ansari, R. Biswas, and S. Aggarwal, "Neutrosophic classifier: An extension of fuzzy classifer," *Applied Soft Computing Journal*, vol. 13, no. 1, pp. 563–573, Jan. 2013. DOI: 10.1016/J.ASOC.2012.08.002.

[113] U. Rivieccio, "Neutrosophic logics: Prospects and problems," *Fuzzy Sets and Systems*, vol. 159, no. 14, pp. 1860–1868, Jul. 2008. DOI: 10.1016/J.FSS.2007.11.011.

[114] J. Zheng, Y. Zhang, D. Hou, *et al.*, "A review of nondestructive examination technology for polyethylene pipe in nuclear power plant," *Frontiers of Mechanical Engineering 2018*, vol. 13, no. 4, pp. 535–545, May 2018. DOI: 10.1007/S11465-018-0515-9.

[115] M. Troughton, "Heated Tool Bonding of Plastic Pipes," *Journal of Adhesion and Interface*, vol. 21, no. 1, pp. 1–5, 2020. DOI: 10.17702/jai.2020.21.1.1.

[116] N. Thorpe, M. Acebes, D. Wylie, *et al.*, "Ultrasonic Phased Array Non-Destructive Testing and In-Service Inspection System for high integrity Polyethylene Pipe Welds with automated analysis software," in *12th ECNDT 2018*, 2018, pp. 1–9.

[117] M. Troughton and F. Hagglund, "On-site Volumetric Inspection of Butt Fusion and Electrofusion Joints in Polyethylene Pipes," *Joining Plastics Journal*, vol. 10, pp. 1–8, 2016.

[118] D. Hou, W. Guo, and J. Zheng, "A Method of Automatic Defect Recognition for Phased Array Ultrasonic Inspection of Polythene Electro-Fusion Joints," in *ASME 2015 Pressure Vessels and Piping Conference*, vol. 5, American Society of Mechanical Engineers Digital Collection, Nov. 2015, pp. 1–19, ISBN: 9780791856987. DOI: 10.1115/PVP2015-45397.

[119] A.-S. Faleh Jassem, “Automated Generation of Metadata for Mining Image and Text Data,” Ph.D. dissertation, George Mason University, 2006.

[120] R. Hudec and M. Benco, “Novel Method for Color Textures Features Extraction Based on GLCM,” *Radioengineering*, vol. 16, no. 4, 2007.

[121] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, “Explainable AI: A Review of Machine Learning Interpretability Methods,” *Entropy*, vol. 23, no. 1, pp. 1–45, 2020. DOI: 10.3390/e23010018.

A Complete set of explanation examples for chapter 4

A.1 Breast cancer dataset

A.1.1 TP Example: Case #1_3_93

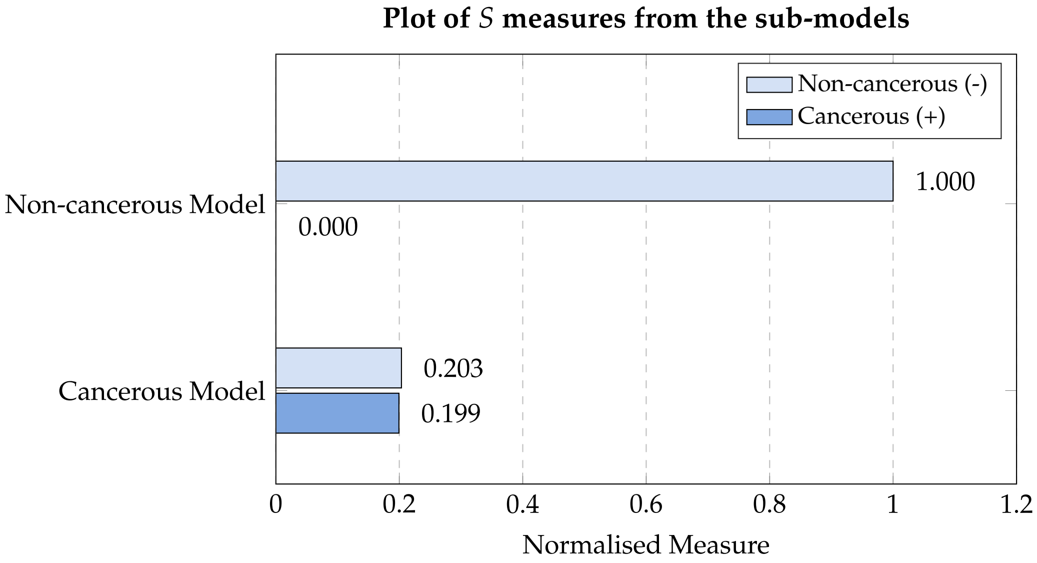


Figure A.1: Breast Cancer Dataset: Illustration of the distance measures S from the two sub-models - example of a TP case: #1_3_93. Textual explanation: "Models are in agreement hence, it is more likely the classification is accurate." Fuzzy class: 0.80.

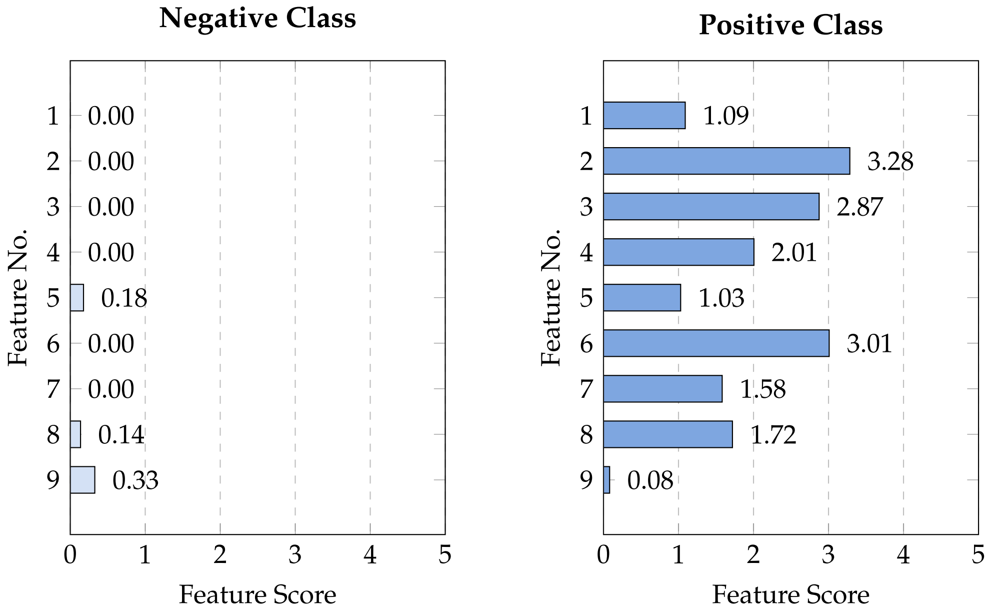


Figure A.2: Breast Cancer Dataset: Graphical explanation of the *non-cancerous* sub-model - example of a TP case: #1_3_93. Textual explanation: “Non-cancerous model thinks the data is similar to Malignant (+ve) and NOT similar to Benign (-ve)."

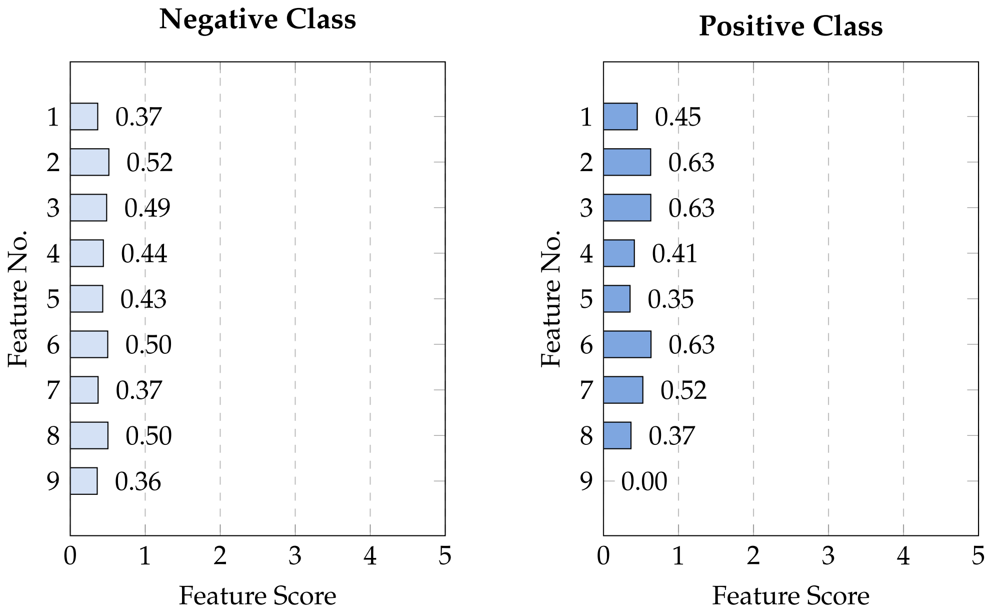


Figure A.3: Breast Cancer Dataset: Graphical explanation of the *cancerous* sub-model - example of a TP case: #1_3_93. Textual explanation: “Cancerous model thinks the data is more similar to the Malignant (+ve) despite a high similarity for both"

A.1.2 TN Example: Case #1_3_63

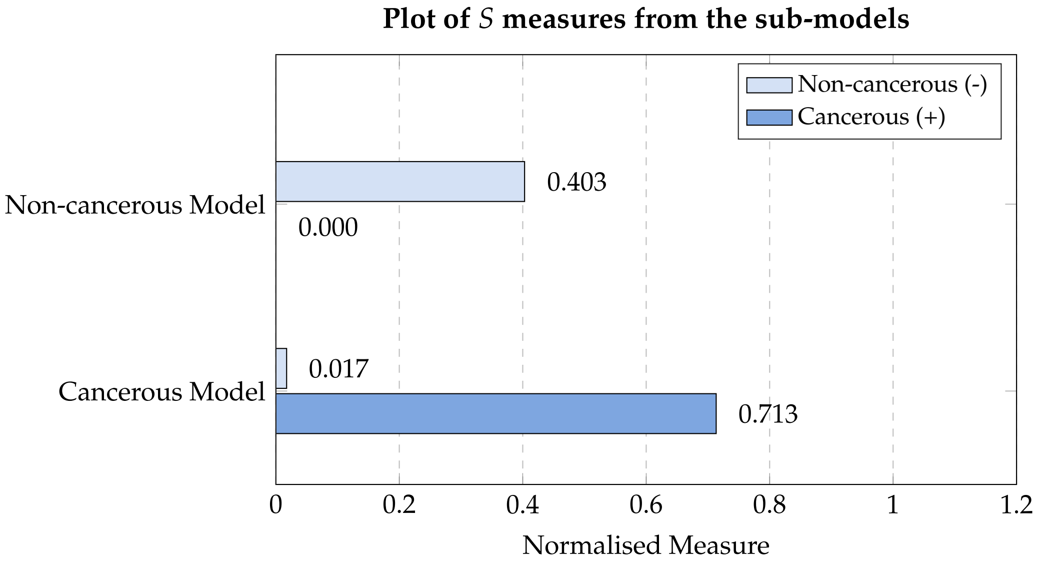


Figure A.4: Breast Cancer Dataset: Illustration of the distance measures S from the two sub-models - example of a TN case: #1_3_63. Textual explanation: "Models are in conflict hence, it is less likely the classification is accurate." Fuzzy class: 0.39.

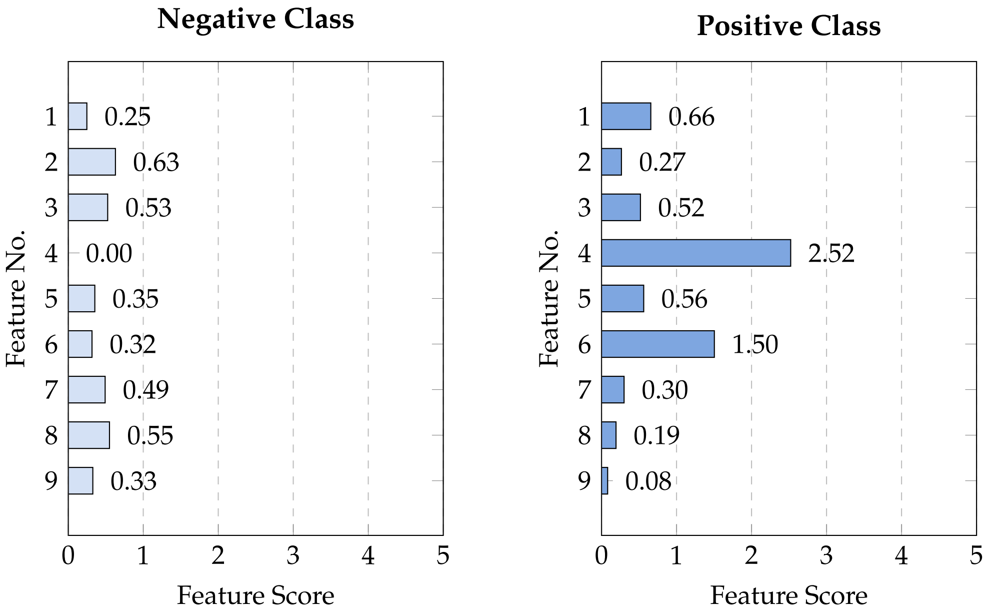


Figure A.5: Breast Cancer Dataset: Graphical explanation of the *non-cancerous* sub-model - example of a TN case: #1_3_63. Textual explanation: "Non-cancerous model thinks the data is more similar to the Malignant (+ve) despite a high similarity for both."

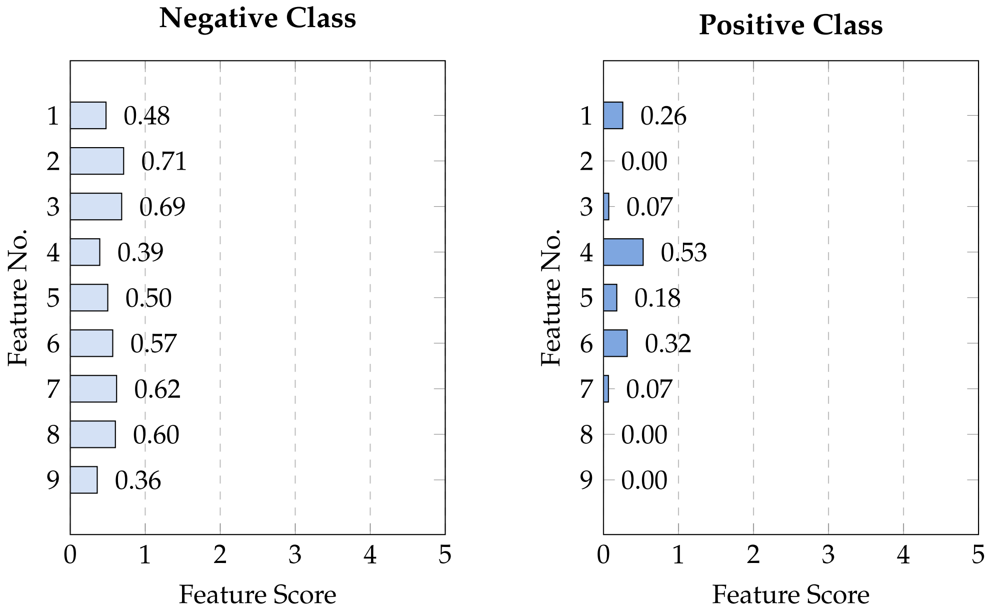


Figure A.6: Breast Cancer Dataset: Graphical explanation of the *cancerous* sub-model - example of a TN case: #1_3_63. Textual explanation: “Cancerous model thinks the data is similar to Benign (-ve) and NOT similar to Malignant (+ve)."

A.2 KEEL titanic dataset

A.2.1 FP Example: Case #1_1_75

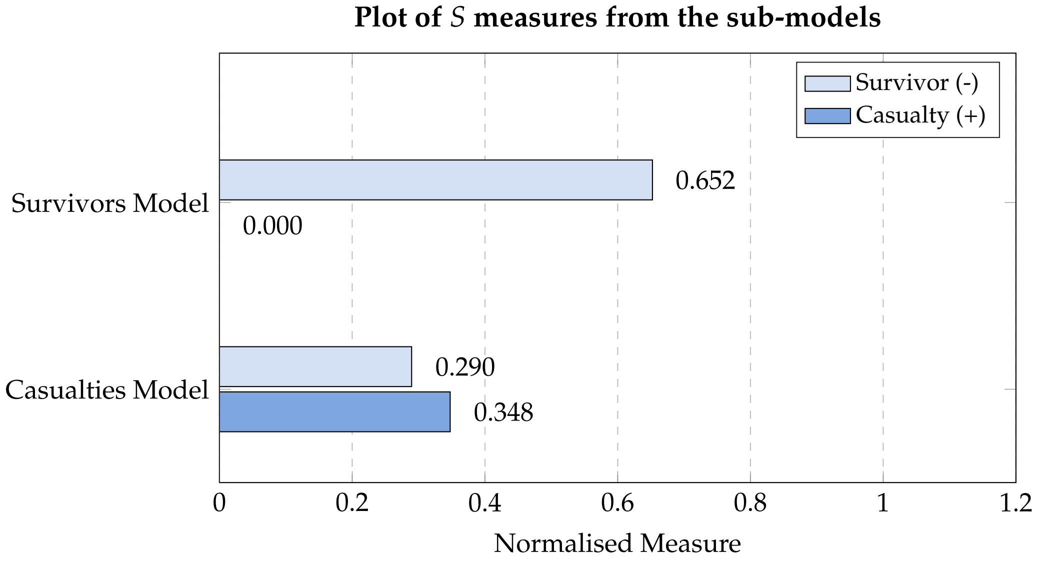


Figure A.7: KEEL Titanic Dataset: Illustration of the distance measures S from the two sub-models - example of a FP case: #1_1_75. Textual explanation: “Models are in conflict hence, it is less likely the classification is accurate." Fuzzy class: 0.74.

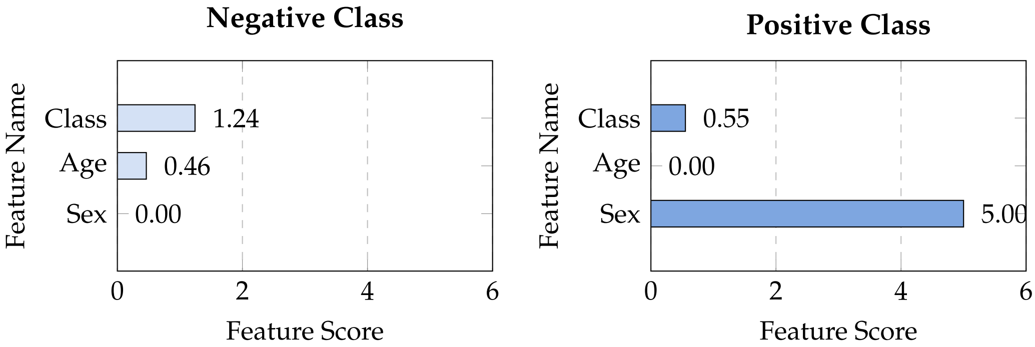


Figure A.8: KEEL Titanic Dataset: Graphical explanation of the *survivor* sub-model - example of a FP case: #1_1_75. Textual explanation: “Survivors model thinks the data is similar to Casualty (+ve) and NOT similar to Survivor (-ve)."

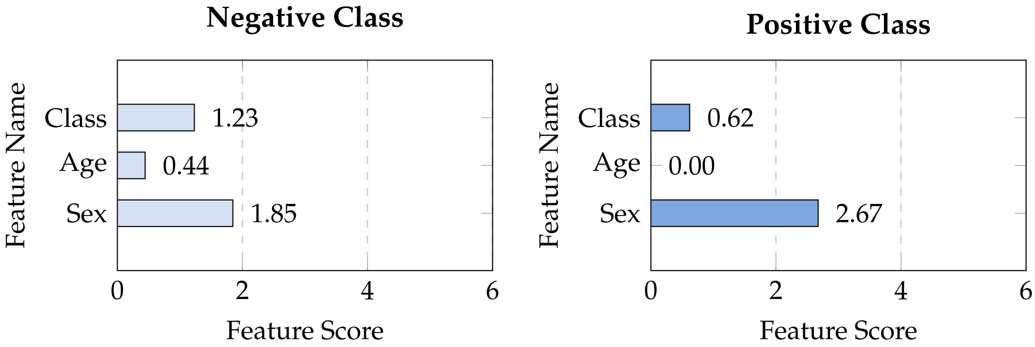


Figure A.9: KEEL Titanic Dataset: Graphical explanation of the *casualty* sub-model - example of a FP case: #1_1_75. Textual explanation: “Casualties model thinks the data is more similar to the Survivor (-ve) despite a high similarity for both."

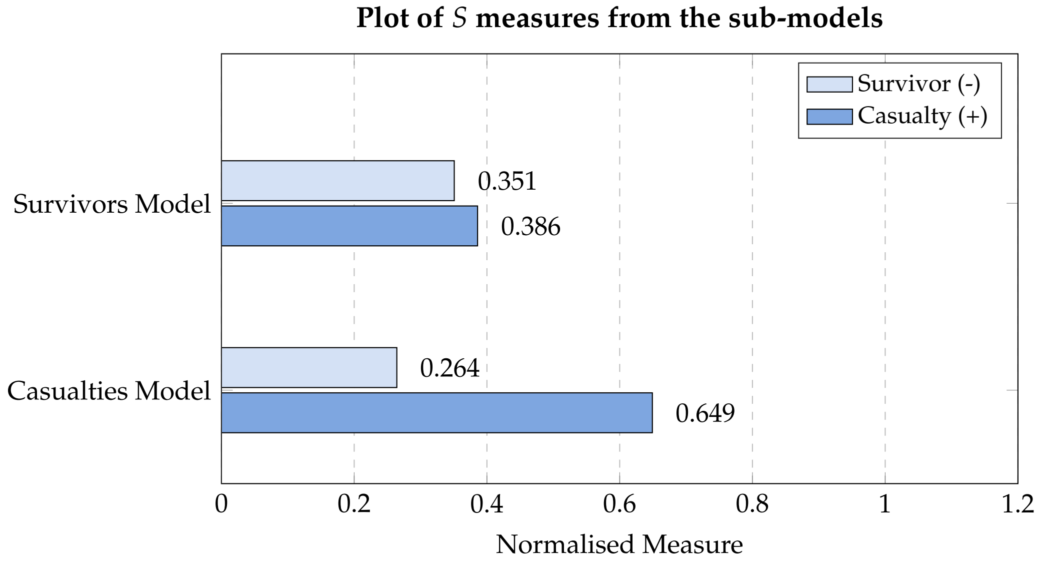


Figure A.10: Titanic KEEL Dataset: Illustration of the distance measures S from the two sub-models - example of a TN case: #1_1_27. Textual explanation: "Models are in agreement hence, it is more likely the classification is accurate." Fuzzy class: 0.37.

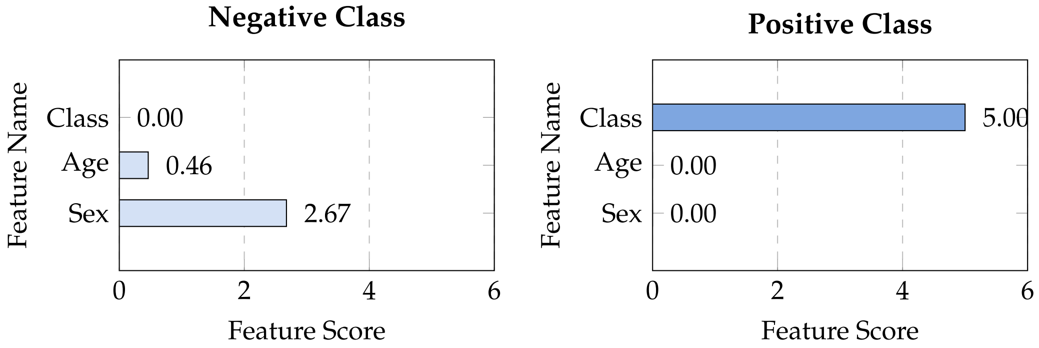


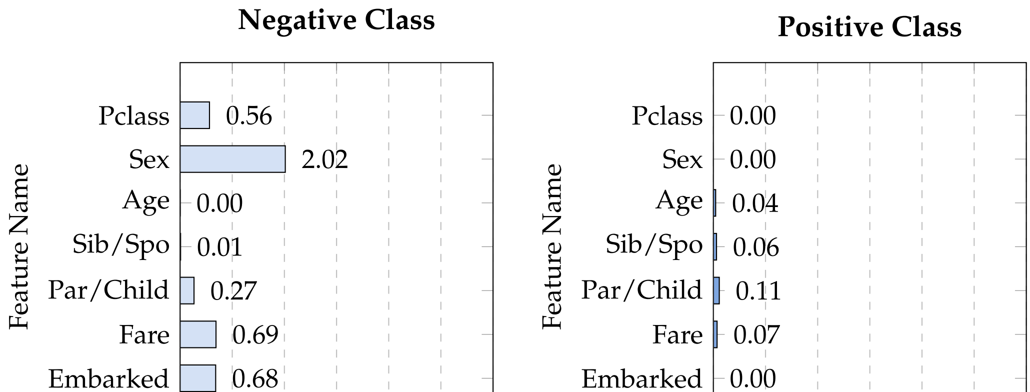
Figure A.11: Titanic KEEL Dataset: Graphical explanation of the *survivor* sub-model - example of a TN case: #1_1_27. Textual explanation: "Survivors model thinks the data is more similar to the Survivor (-ve) despite a high similarity for both."

A.2.2 TN Example: Case #1_1_27

A.2.3 Explanation analysis

A.3 Kaggle titanic dataset

A.3.1 FN Example: Case #1_1_111



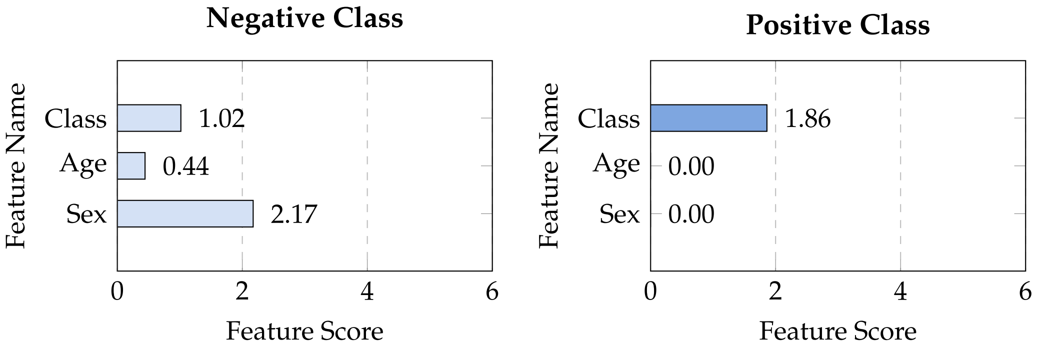


Figure A.12: Titanic KEEL Dataset: Graphical explanation of the *casualty* sub-model - example of a TN case: #1_1_27. Textual explanation: “Casualties model thinks the data is similar to Survivor (-ve) and NOT similar to Casualty (+ve)."

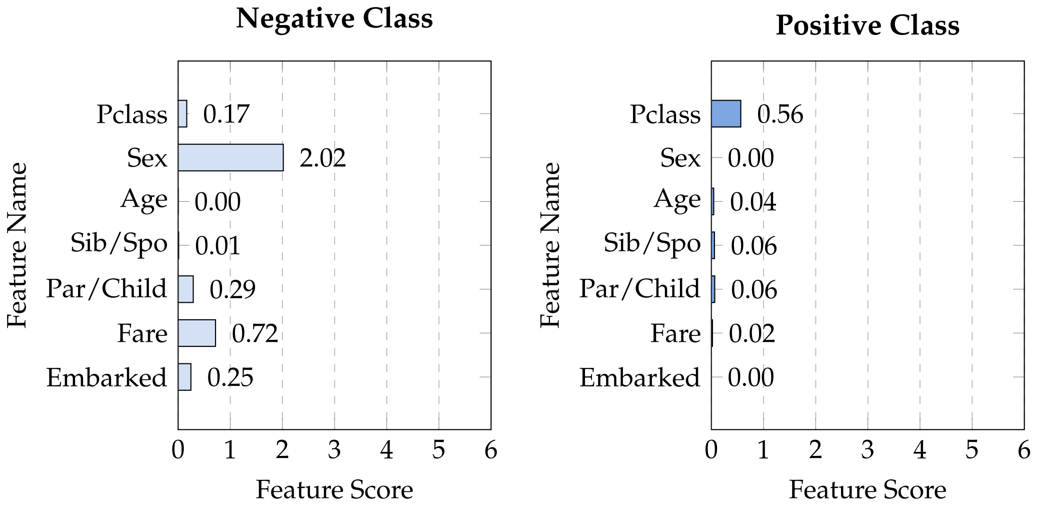


Figure A.15: Titanic Kaggle dataset: graphical explanation of the *casualty* sub-model - example of a FN case: #1_1_111. Textual explanation: “Casualties model thinks the data is similar to Survivor (-ve) and NOT similar to Casualty (+ve)."

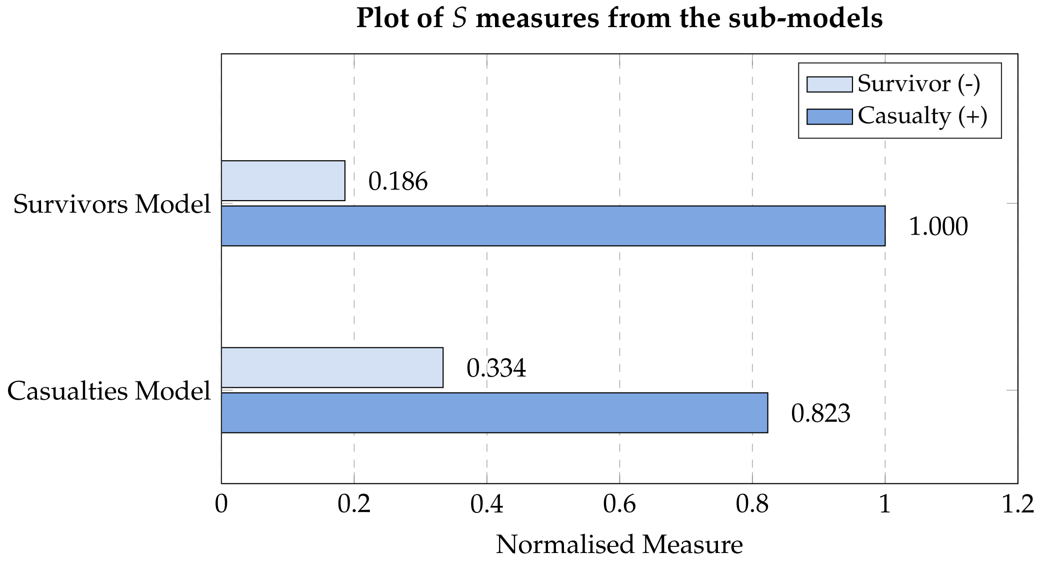


Figure A.16: Titanic Kaggle dataset: distance measures S from the two sub-models - example of a FN case: #1_1_111. Textual explanation: “Models are in agreement hence, it is more likely the classification is accurate. Fuzzy class: 0.32"

A.3.2 TN Example: Case #1_1_1

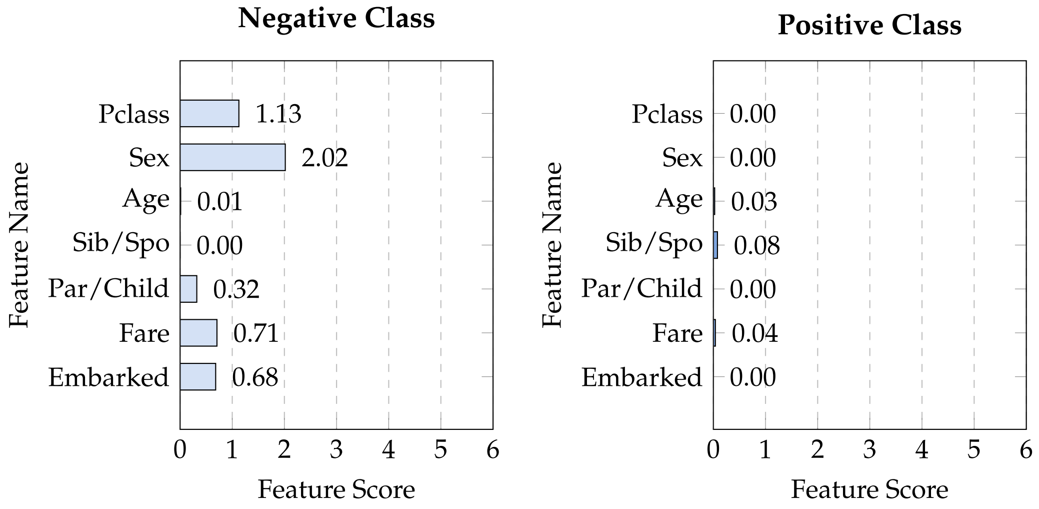


Figure A.17: Titanic Kaggle dataset: graphical explanation of the *survivor* sub-model - example of a TN case: #1_1_1. Textual explanation: “Survivors model thinks the data is similar to Survivor (-ve) and NOT similar to Casualty (+ve)."

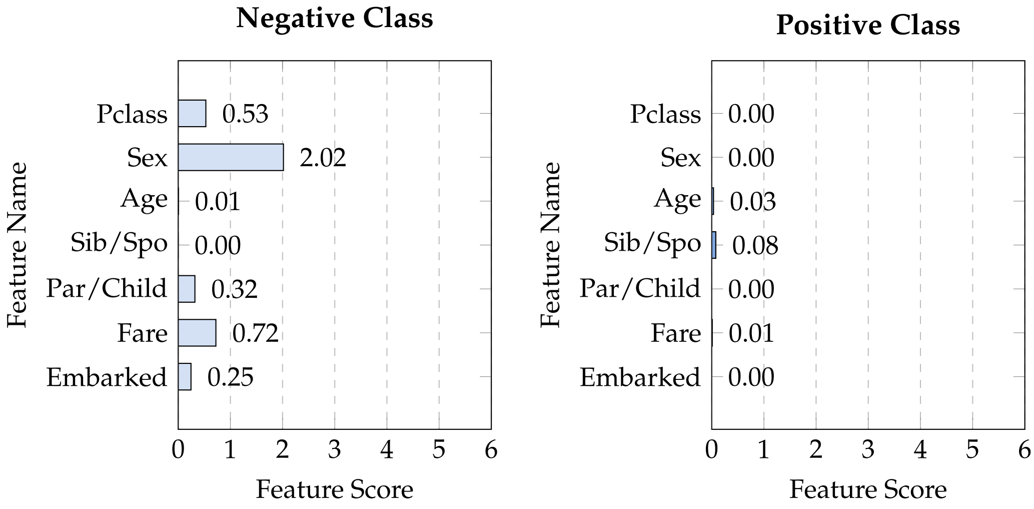


Figure A.18: Titanic Kaggle dataset: graphical explanation of the *casualty* sub-model - example of a TN case: #1_1_1. Textual explanation: “Casualties model thinks the data is similar to Survivor (-ve) and NOT similar to Casualty (+ve)."

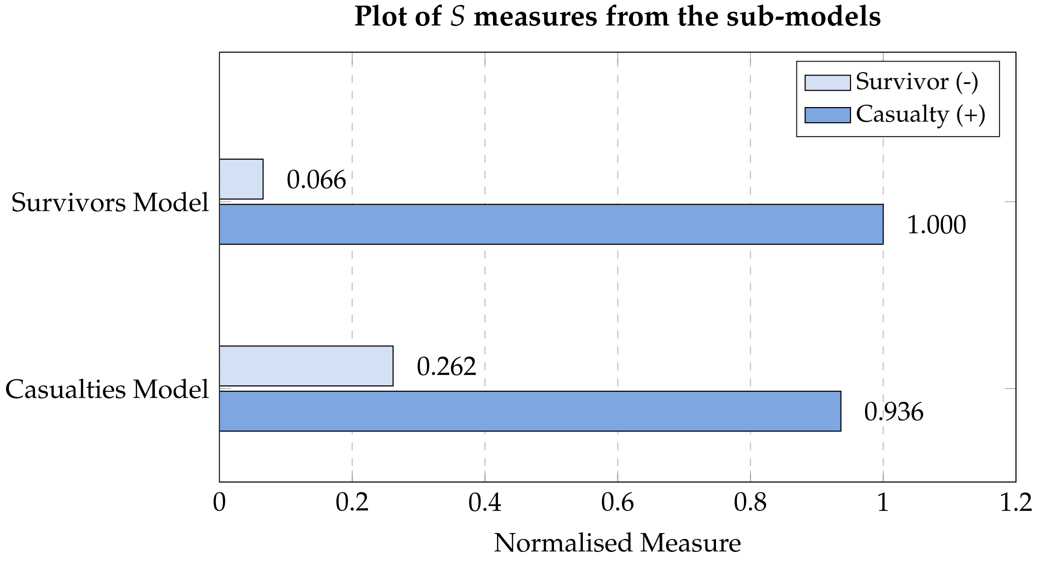


Figure A.19: Titanic Kaggle dataset: illustration of the distance measures S from the two sub-models - example of a TN case: #1_1_1. Textual explanation: “Models are in agreement hence, it is more likely the classification is accurate. Fuzzy class: 0.24"

A.3.3 Explanation analsis

A.4 Parkinson's disease dataset

A.4.1 TP Example: Case #1_1_43

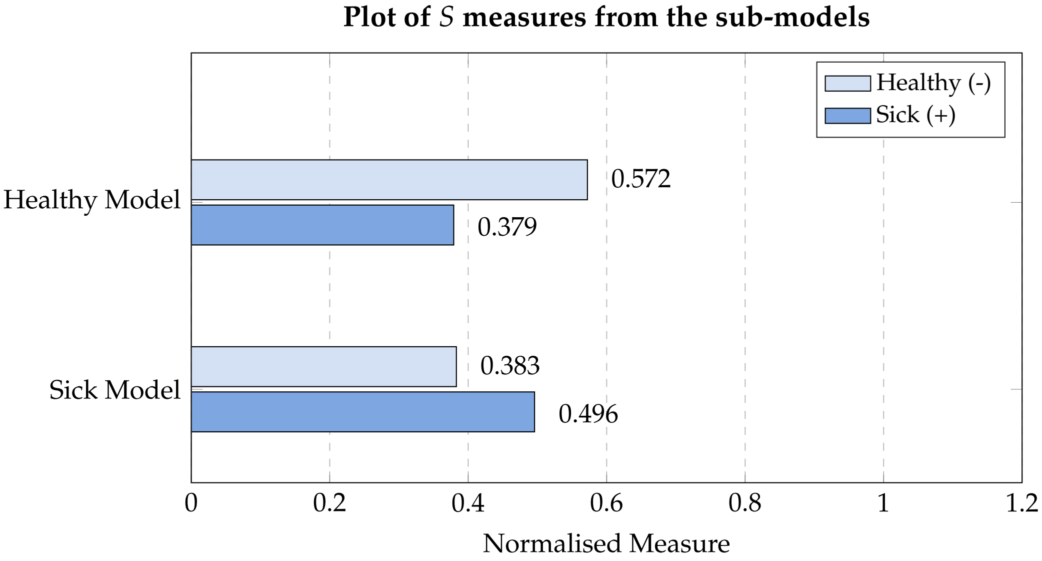


Figure A.21: Parkinson disease dataset: distance measures S from the two sub-models - example of a TP case: #1_1_43. Textual explanation: "Models are in conflict hence, it is less likely the classification is accurate." Fuzzy class: 0.52.

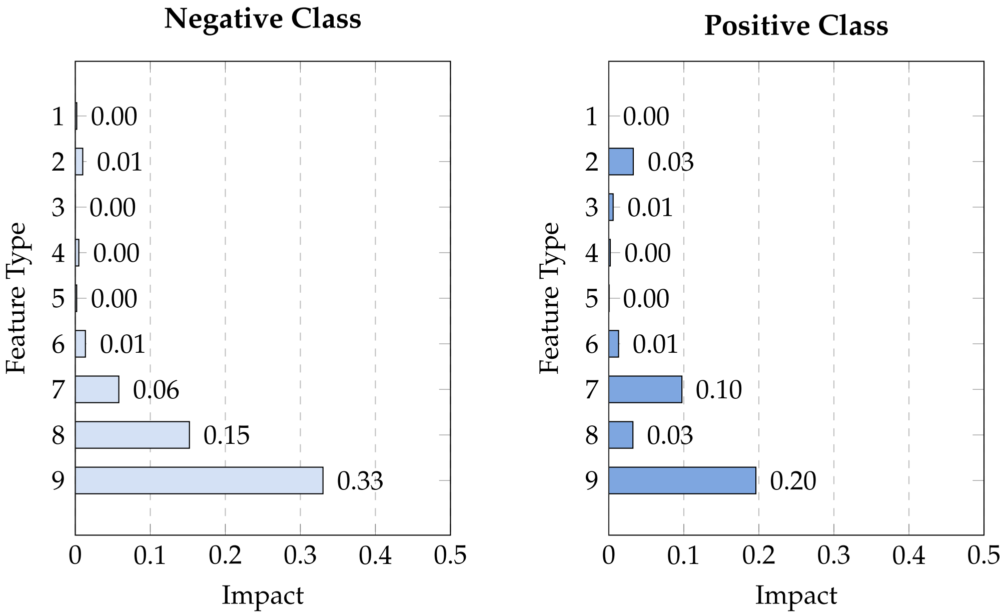


Figure A.22: Parkinson disease dataset: graphical explanation of the *healthy* sub-model - example of a TP case: #1_1_43. Textual explanation: "Healthy model thinks the data is similar to Positive (+ve) and NOT similar to Negative (-ve)."

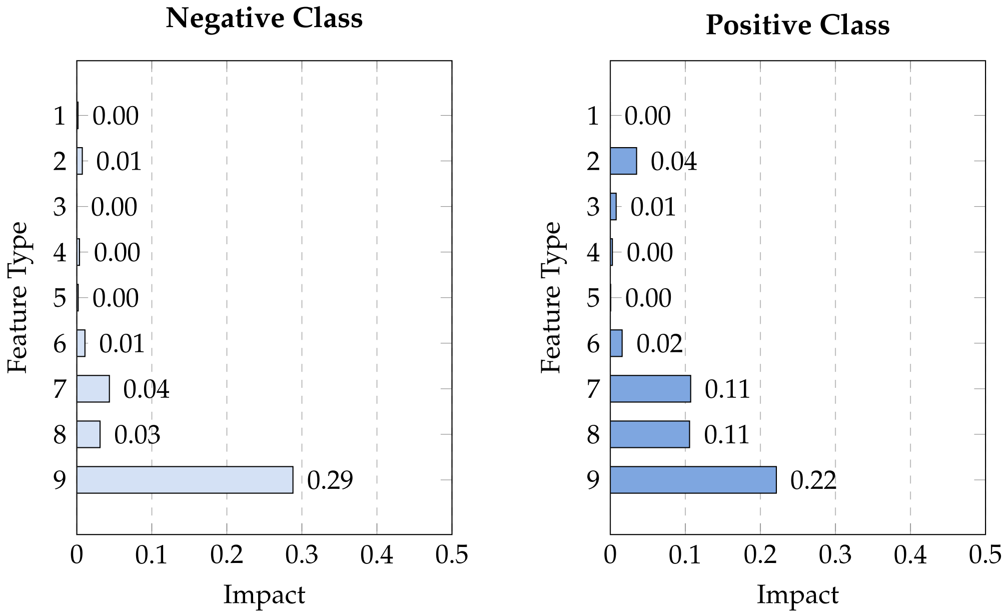


Figure A.23: Parkinson disease dataset: graphical explanation of the *sick* sub-model - example of a TP case: #1_1_43. Textual explanation: "Sick model thinks the data is more similar to the Negative (-ve) despite a high similarity for both."

A.5 Chess dataset

A.5.1 TN Example: Case #1_1_339

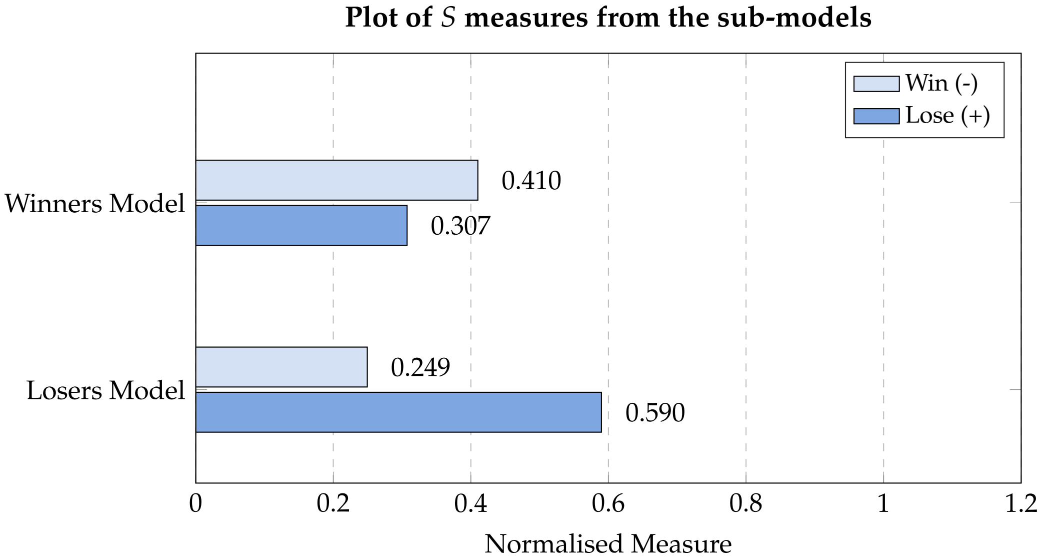


Figure A.24: Chess dataset: distance measures S from the two sub-models - example of a FN case: #1_1_339. Textual explanation: "Models are in conflict hence, it is less likely the classification is accurate." Fuzzy class: 0.41

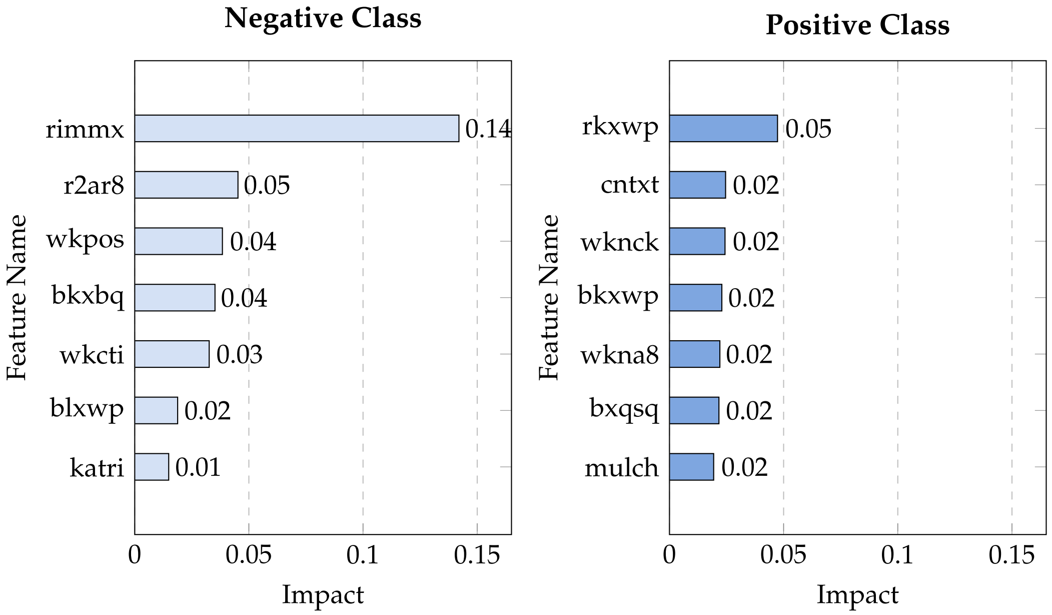


Figure A.25: Chess dataset: graphical explanation of the *win* sub-model - example of a FN case: #1_1_339. Textual explanation: "Winners model thinks the data is more similar to the Lose (+ve) despite a high similarity for both."

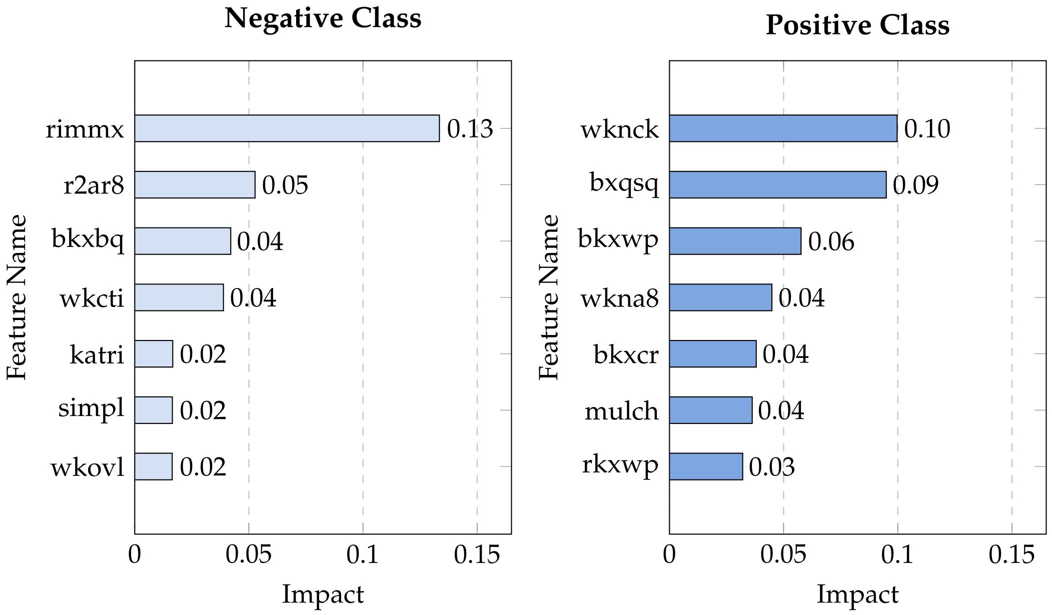


Figure A.26: Chess dataset: graphical explanation of the *lose* sub-model - example of a FN case: #1_1_339. Textual explanation: "Losers model thinks the data is similar to Win (-ve) and NOT similar to Lose (+ve)."

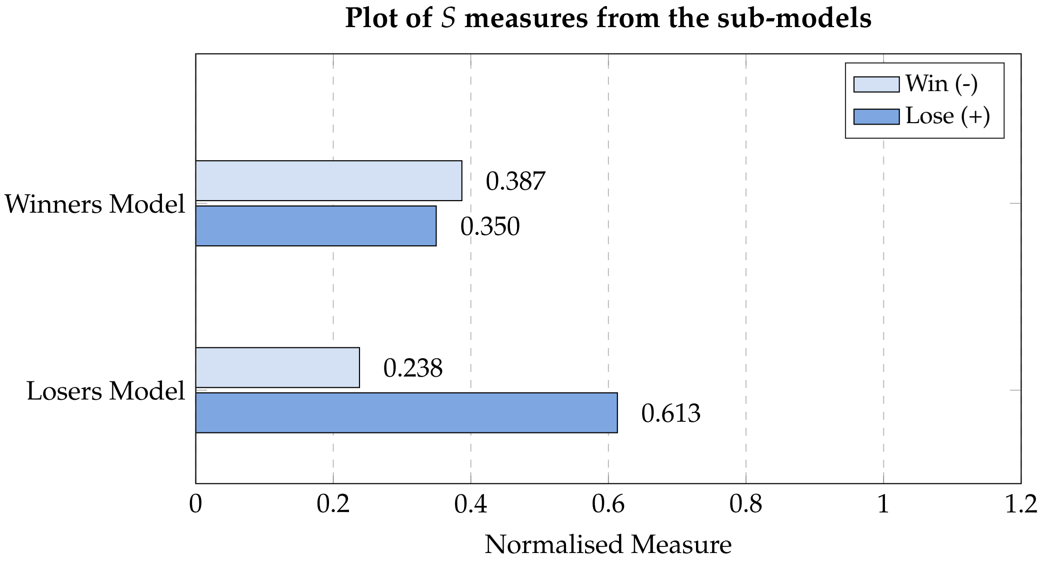


Figure A.27: Chess Dataset: Illustration of the distance measures S from the two sub-models - example of a TN case: #1_1_2. Textual explanation: “Models are in conflict hence, it is less likely the classification is accurate." Fuzzy class: 0.37

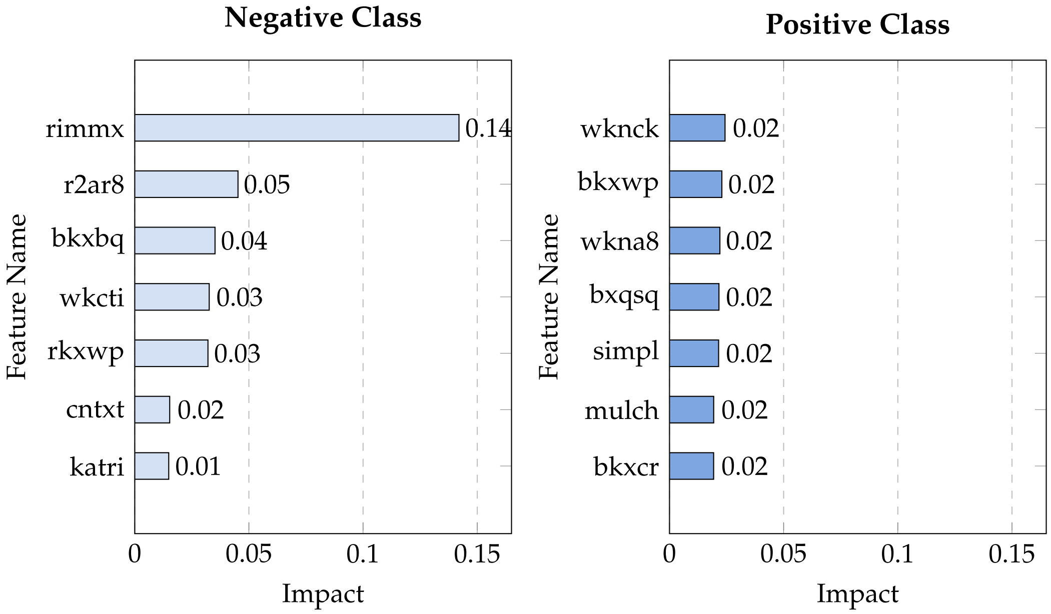


Figure A.28: Chess Dataset: Graphical explanation of the *win* sub-model - example of a TN case: #1_1_2. Textual explanation: “Winners model thinks the data is more similar to the Lose (+ve) despite a high similarity for both."

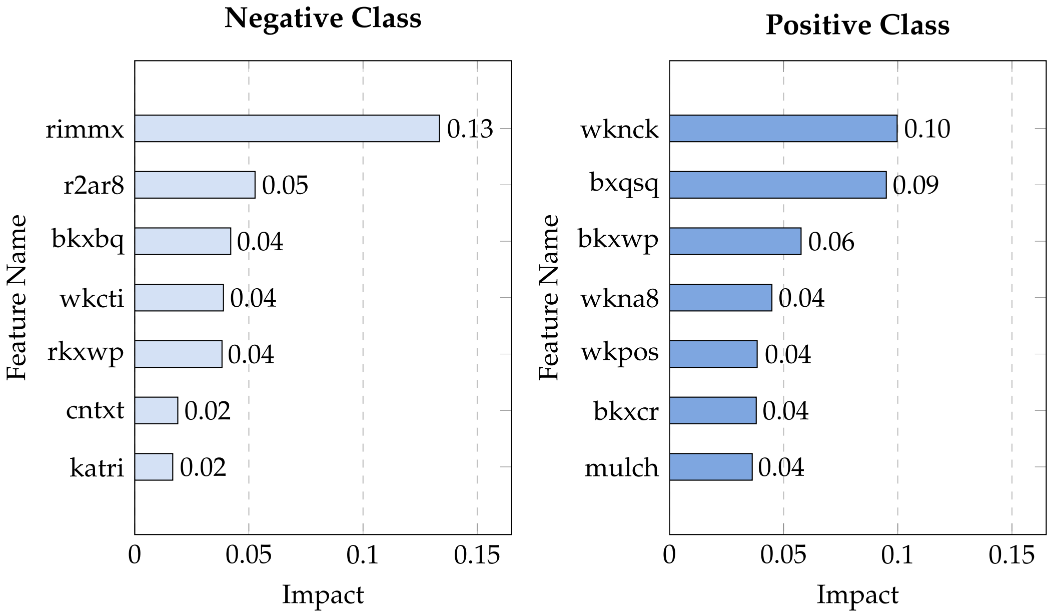


Figure A.29: Chess Dataset: Graphical explanation of the *lose* sub-model - example of a TN case: #1_1_2. Textual explanation: “Losers model thinks the data is similar to Win (-ve) and NOT similar to Lose (+ve)."

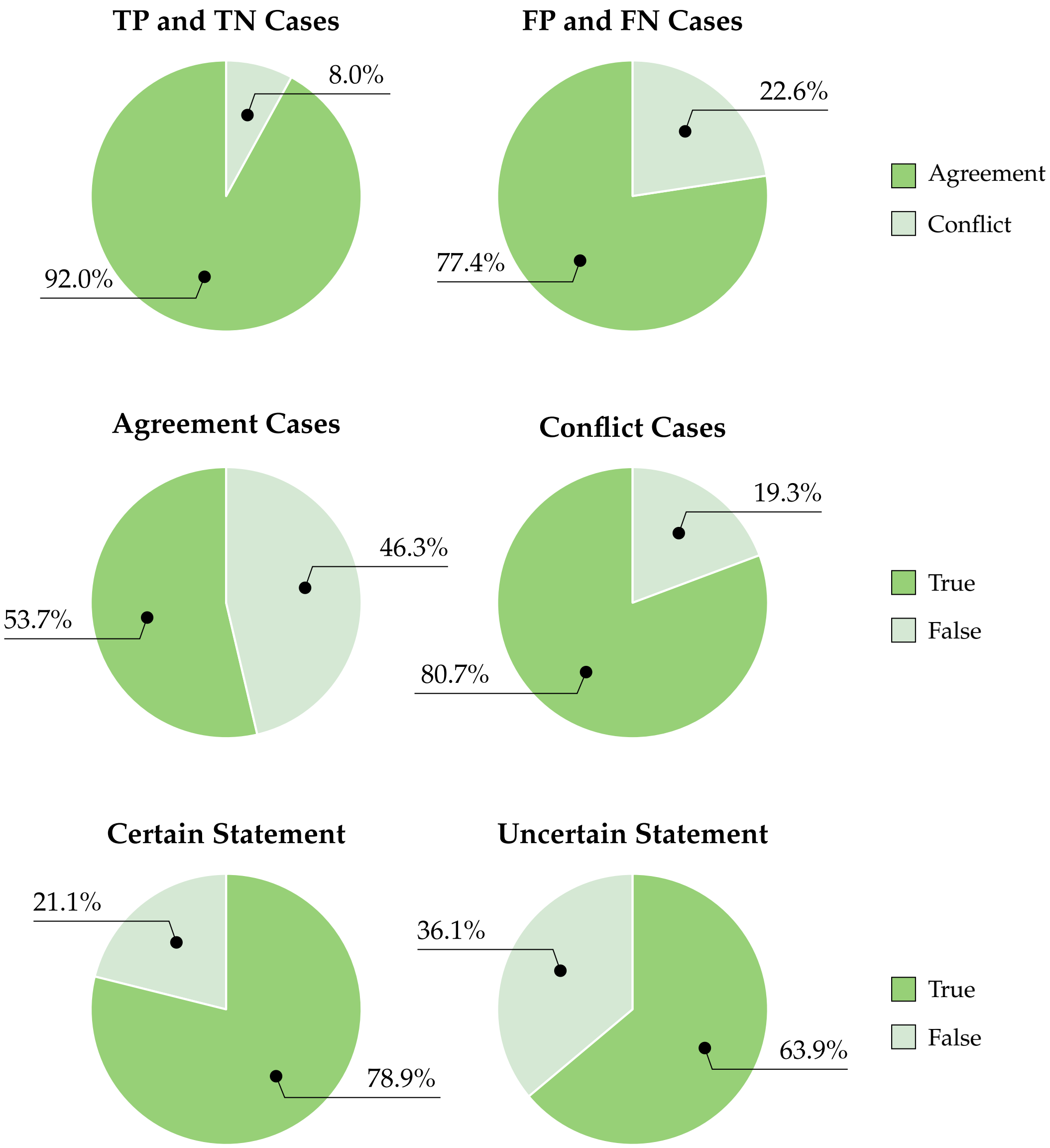


Figure A.13: Titanic KEEL Dataset: a set of pie charts are used to illustrate how *indicative* the different aspects of explanation are to a false or negative cases. Certain statements refer to sentence templates #1 and #2 while, uncertain statements refer to sentence templates #3 and #4.

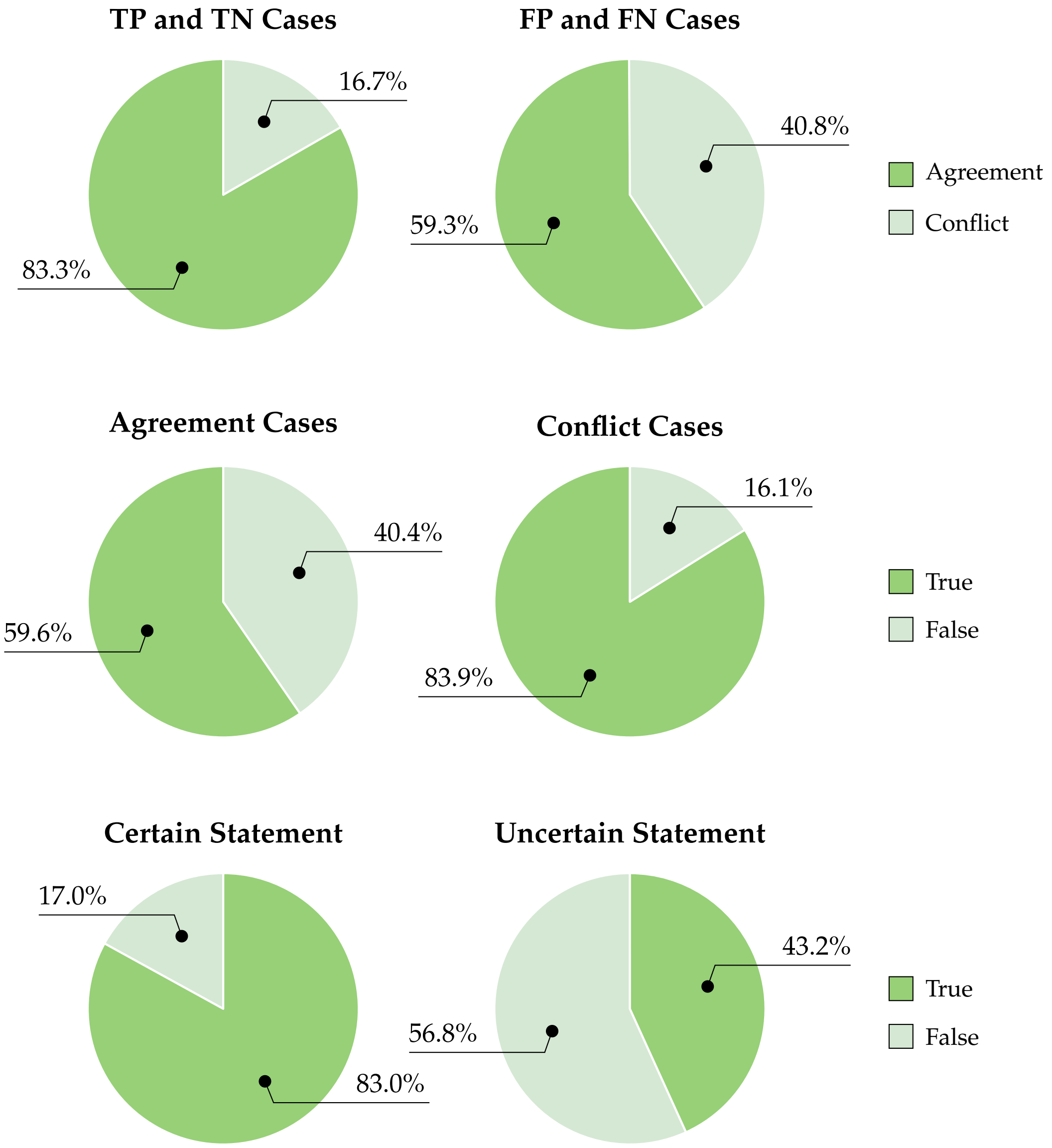


Figure A.20: Titanic Kaggle Dataset: a set of pie charts are used to illustrate how *indicative* the different aspects of explanation are to a false or negative cases. Certain statements refer to sentence templates #1 and #2 while, uncertain statements refer to sentence templates #3 and #4.

B Complete set of explanation examples for chapter 5

B.1 Breast cancer dataset

B.1.1 TP example: #8_5_106

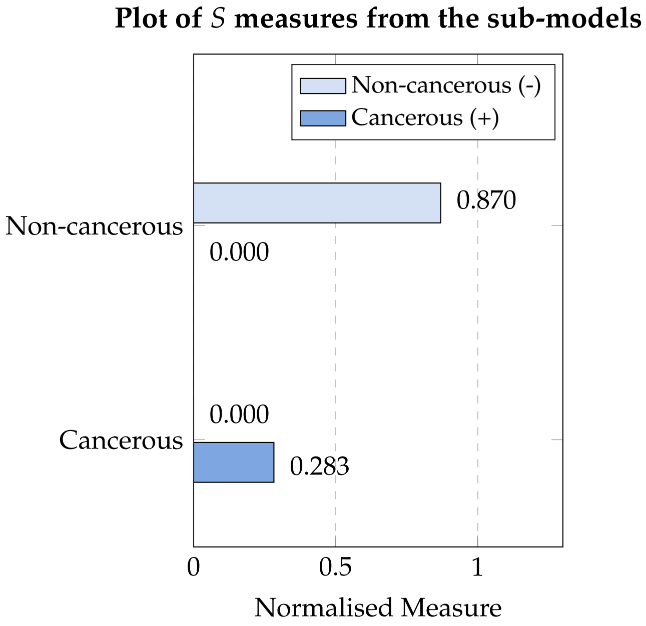


Figure B.1: Breast Cancer Dataset: distance measures S from the two sub-models - example of a TP case: #8_5_106. Textual explanation: "Models are in conflict hence, it is less likely the classification is accurate. The data was classified as *Cancerous (+)* by the model."

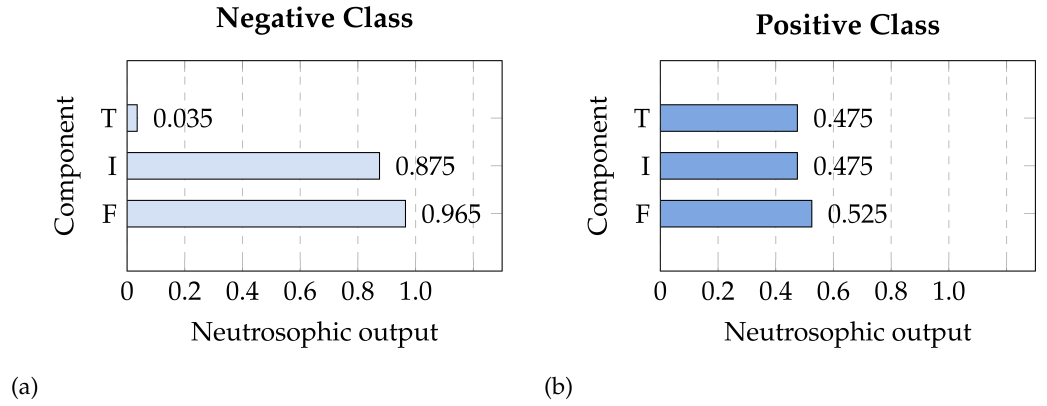


Figure B.2: Breast cancer dataset: graphical explanation of the overall NL sub-model - example of a TP case: #8_5_106. Overall explanation: "Based on the overall sub-model's output the data was more similar to Cancerous (+) compared to Non-cancerous (-)." (a) Explanation: "For the Non-cancerous (-) class, the overall sub-model thinks the data has a higher falsity (0.965) relative to its truth (0.035). The relative indetermency is high (0.875).". (b) Explanation: "For the Cancerous (+) class, the overall sub-model thinks the data has a higher truth (0.875) relative to its falsity (0.500). The relative indetermency is average (0.475)."

B.1.2 FP example: #6_2_22

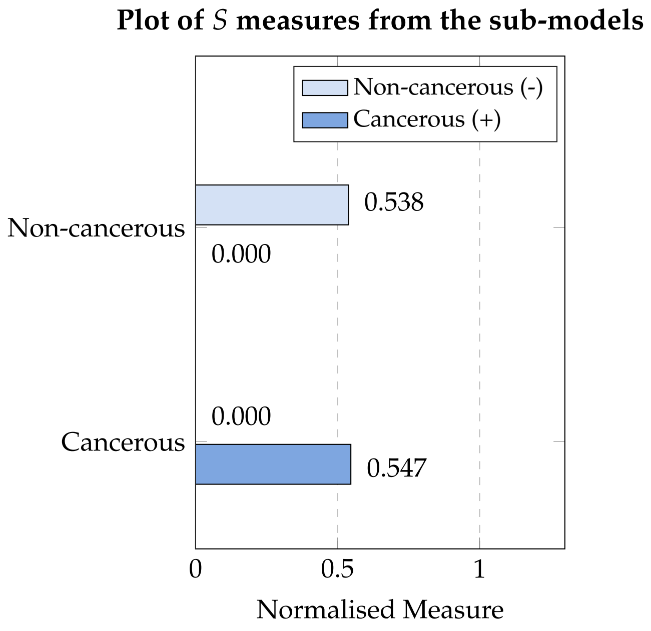
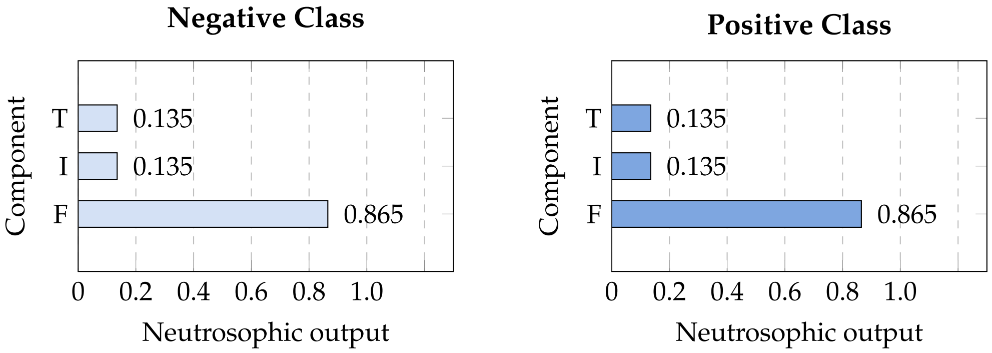


Figure B.3: Breast Cancer Dataset: distance measures S from the two sub-models - example of a FP case: #6_2_22. Textual explanation: "Models are in conflict hence, it is less likely the classification is accurate. The data was classified as *Cancerous (+)* by the model."



(a)

(b)

Figure B.4: Breast Cancer Dataset: Graphical explanation of the overall NL sub-model - example of a FP case: #6_2_22. Overall explanation: “Based on the overall sub-model’s output the data was more similar to Cancerous (+) compared to Non-cancerous (-).” (a) Explanation: “For the Non-cancerous (-) class, the overall sub-model thinks the data has a higher falsity (0.865) relative to its truth (0.135). The relative indetermency is low (0.135).”. (b) Explanation: “For the Cancerous (+) class, the overall sub-model thinks the data has a higher falsity (0.865) relative to its truth (0.135). The relative indetermency is low (0.135).”

B.2 KEEL titanic dataset

B.2.1 FP example: #6_3_145

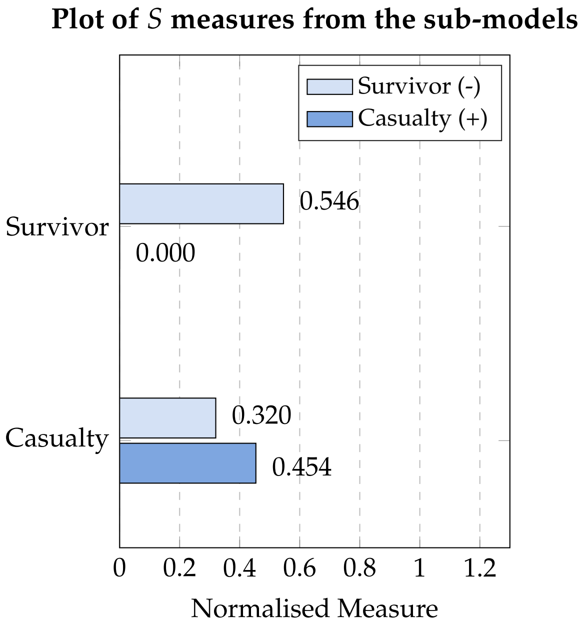
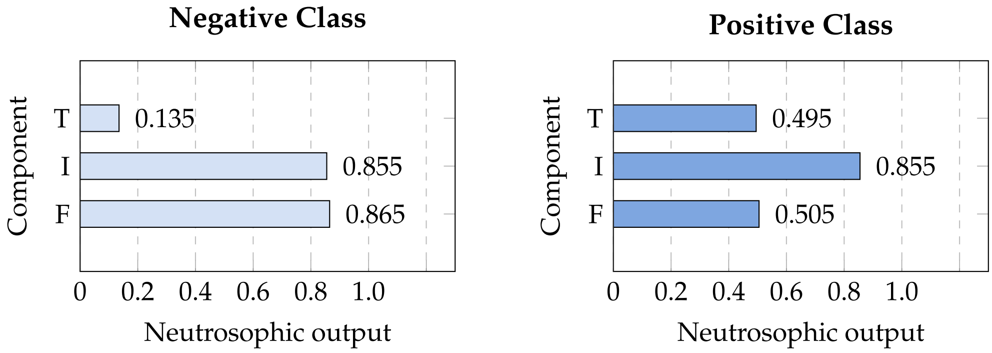


Figure B.5: Titanic - KEEL Dataset: Illustration of the distance measures S from the two sub-models - example of a FP case: #6_3_145. Textual explanation: “Models are in conflict hence, it is less likely the classification is accurate. The data was classified as *Casualty (+)* by the model.”

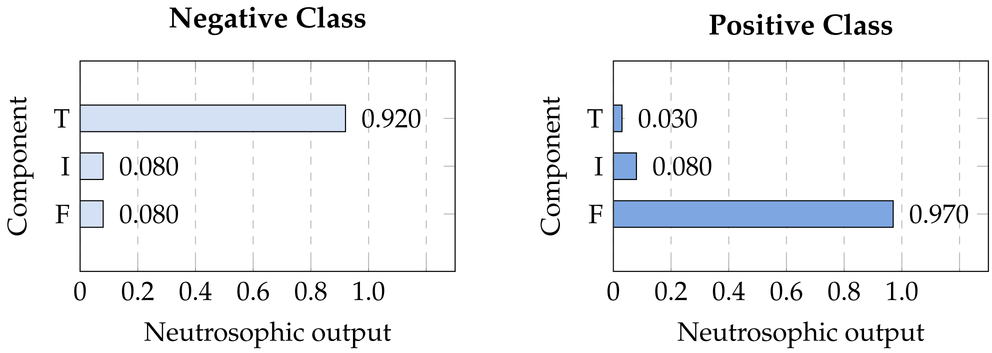


(a)

(b)

Figure B.6: Titanic - KEEL Dataset: Graphical explanation of the overall NL sub-model - example of a FP case: #6_3_145. Overall explanation: “Based on the overall sub-model’s output the data was more similar to Casualty (+) compared to Survivor (-).” (a) Explanation: “For the Survivor (-) class, the overall sub-model thinks the data has a higher falsity (0.865) relative to its truth (0.135). The relative indetermency is high (0.855).”. (b) Explanation: “For the Casualty (+) class, the overall sub-model thinks the data has a higher truth (0.855) relative to its falsity (0.500). The relative indetermency is high (0.855).”

B.2.2 FN example: #3_4_377



(a)

(b)

Figure B.7: Titanic - KEEL Dataset: Graphical explanation of the overall NL sub-model - example of a FN case: #3_4_377. Overall explanation: “Based on the overall sub-model’s output the data was more similar to Survivor (-) compared to Casualty (+).” (a) Explanation: “For the Survivor (-) class, the overall sub-model thinks the data has a higher truth (0.920) relative to its falsity (0.080). The relative indetermency is low (0.080).”. (b) Explanation: “For the Casualty (+) class, the overall sub-model thinks the data has a higher falsity (0.970) relative to its truth (0.030). The relative indetermency is low (0.080).”

B.3 Kaggle titanic dataset

B.3.1 TP example: #3_2_135

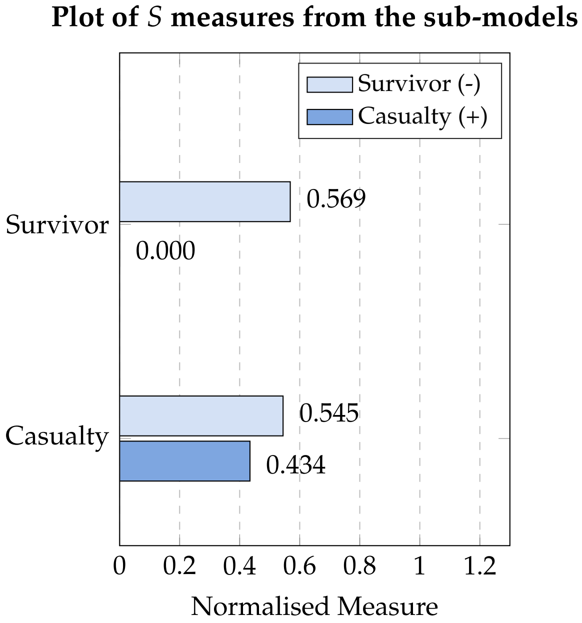


Figure B.8: Titanic - Kaggle Dataset: Illustration of the distance measures S from the two sub-models - example of a TP case: #3_2_135. Textual explanation: "Models are in agreement hence, it is more likely the classification is accurate. The data was classified as *Casualty (+)* by the model."

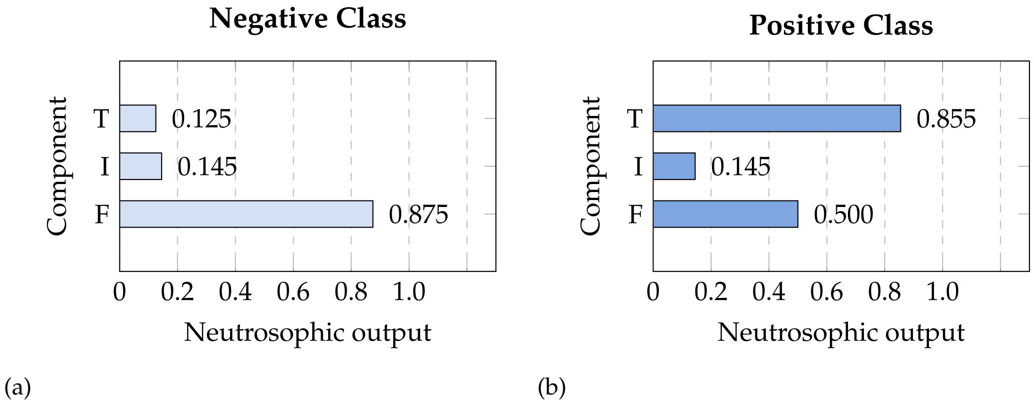


Figure B.9: Titanic - Kaggle Dataset: Graphical explanation of the overall NL sub-model - example of a TP case: #3_2_135. Overall explanation: "Based on the overall sub-model's output the data was more similar to Casualty (+) compared to Survivor (-)." (a) Explanation: "For the Survivor (-) class, the overall sub-model thinks the data has a higher falsity (0.875) relative to its truth (0.125). The relative indetermency is low (0.145).". (b) Explanation: "For the Casualty (+) class, the overall sub-model thinks the data has a higher truth (0.855) relative to its falsity (0.500). The relative indetermency is low (0.145)."

B.3.2 FP example: #4_3_38

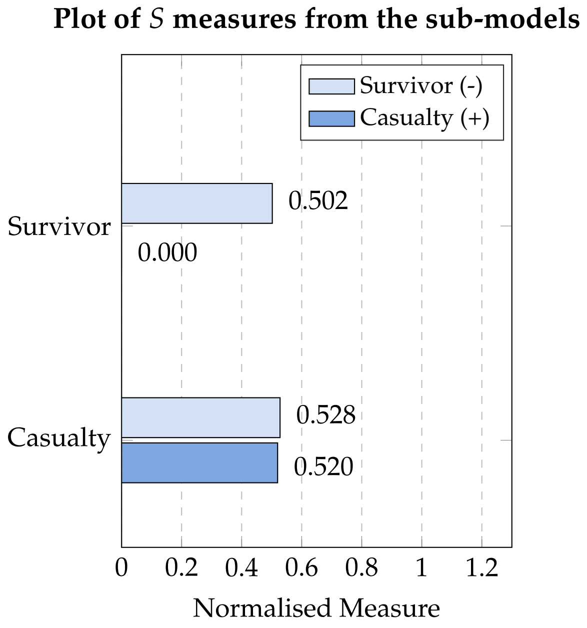


Figure B.10: Titanic - Kaggle Dataset: Illustration of the distance measures S from the two sub-models - example of a FP case: #4_3_38. Textual explanation: “Models are in agreement hence, it is more likely the classification is accurate. The data was classified as *Casualty (+)* by the model.”

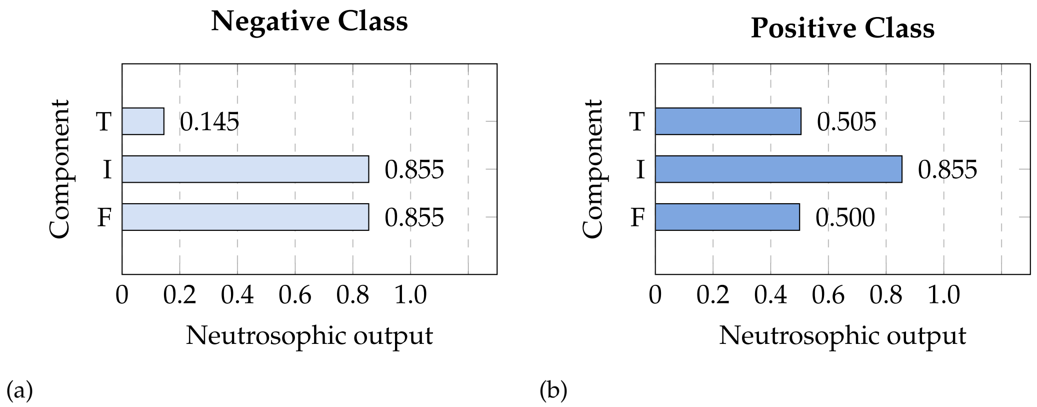


Figure B.11: Titanic - Kaggle Dataset: Graphical explanation of the overall NL sub-model - example of a FP case: #4_3_38. Overall explanation: “Based on the overall sub-model’s output the data was more similar to Casualty (+) compared to Survivor (-).” (a) Explanation: “For the Survivor (-) class, the overall sub-model thinks the data has a higher falsity (0.855) relative to its truth (0.145). The relative indetermency is high (0.855).”. (b) Explanation: “For the Casualty (+) class, the overall sub-model thinks the data has a higher falsity (0.855) relative to its truth (0.145). The relative indetermency is high (0.855).”

B.3.3 TN example: #5_4_108

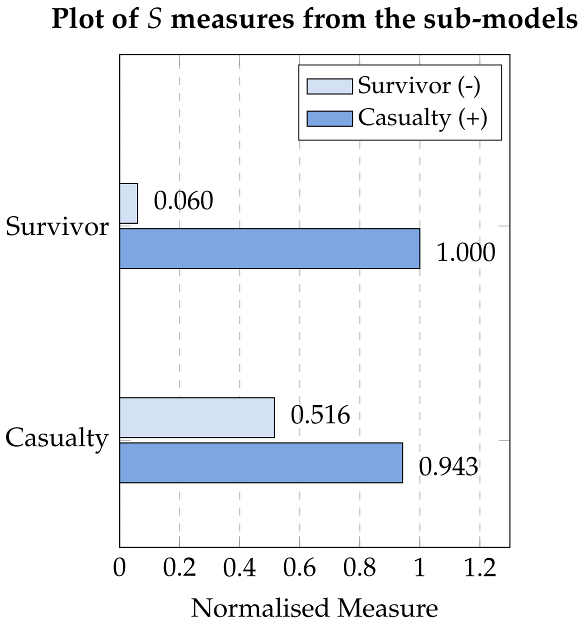


Figure B.12: Titanic - Kaggle Dataset: Illustration of the distance measures S from the two sub-models - example of a TN case: #5_4_108. Textual explanation: “Models are in agreement hence, it is more likely the classification is accurate. The data was classified as *Survivor (-)* by the model.”

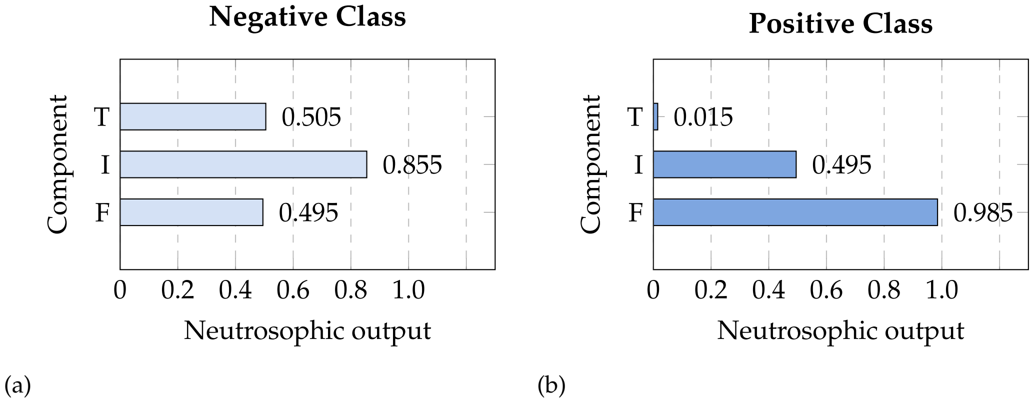


Figure B.13: Titanic - Kaggle Dataset: Graphical explanation of the overall NL sub-model - example of a TN case: #5_4_108. Overall explanation: “Based on the overall sub-model’s output the data was more similar to Survivor (-) compared to Casualty (+).” (a) Explanation: “For the Survivor (-) class, the overall sub-model thinks the data has a higher falsity (0.855) relative to its truth (0.145). The relative indetermency is high (0.855).”. (b) Explanation: “For the Casualty (+) class, the overall sub-model thinks the data has a higher falsity (0.985) relative to its truth (0.015). The relative indetermency is average (0.495).”

B.3.4 FN example: #6_1_175

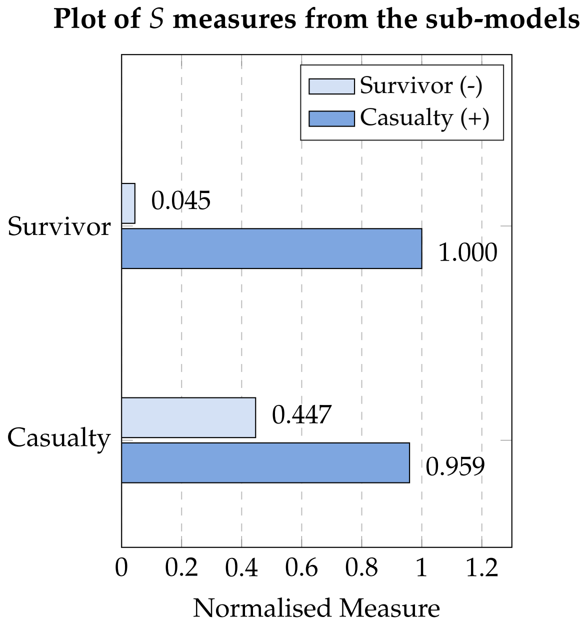


Figure B.14: Titanic - Kaggle Dataset: Illustration of the distance measures S from the two sub-models - example of a FN case: #6_1_175. Textual explanation: “Models are in agreement hence, it is more likely the classification is accurate. The data was classified as *Survivor (-)* by the model.”

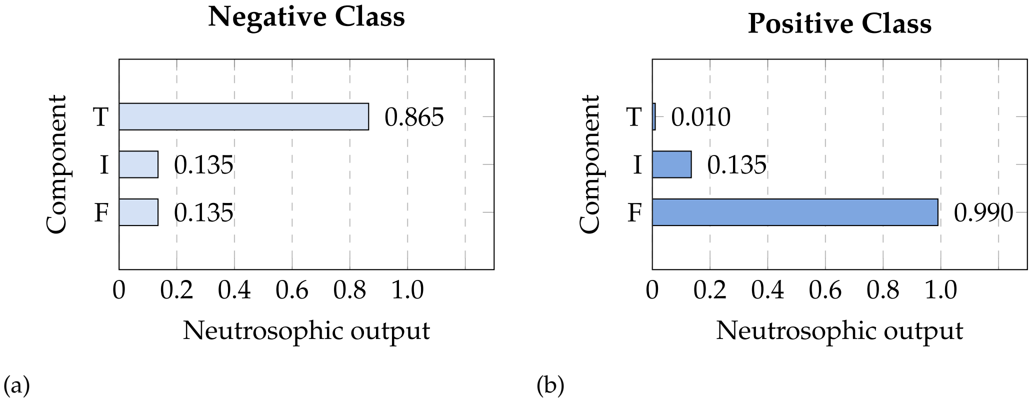


Figure B.15: Titanic - Kaggle Dataset: Graphical explanation of the overall NL sub-model - example of a FN case: #6_1_175. Overall explanation: “Based on the overall sub-model’s output the data was more similar to Survivor (-) compared to Casualty (+).” (a) Explanation: “For the Survivor (-) class, the overall sub-model thinks the data has a higher truth (0.865) relative to its falsity (0.135). The relative indetermency is low (0.135).”. (b) Explanation: “For the Casualty (+) class, the overall sub-model thinks the data has a higher falsity (0.990) relative to its truth (0.010). The relative indetermency is low (0.135).”

B.4 Chess dataset

B.4.1 First TN example: #3_2_169

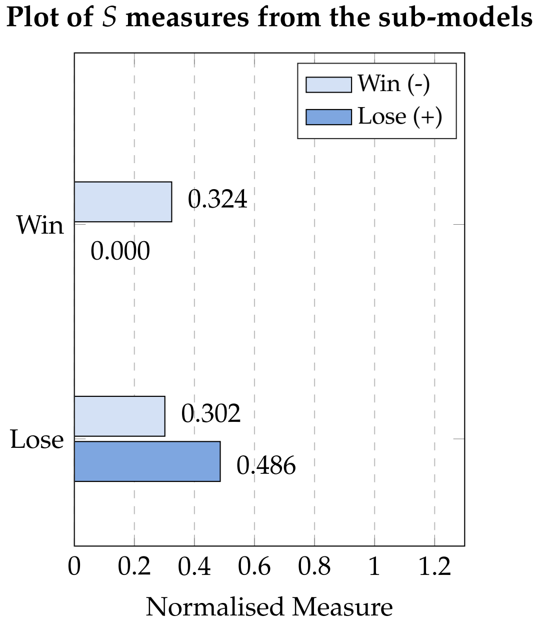


Figure B.16: Chess Dataset: distance measures S from the two sub-models - example of a TN case: #3_2_169. Textual explanation: "Models are in conflict hence, it is less likely the classification is accurate. The data was classified as *Win (-)* by the model."

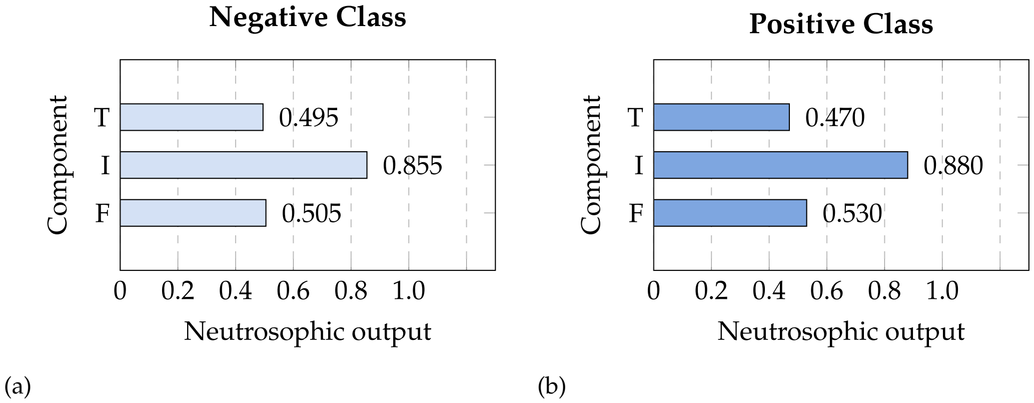


Figure B.17: Chess dataset: graphical explanation of the overall NL sub-model - example of a TN case: #3_2_169. Overall explanation: "Based on the overall sub-model's output the data was more similar to Win (-) compared to Lose (+)." (a) Explanation: "For the Win (-) class, the overall sub-model thinks the data has a higher falsity (0.505) relative to its truth (0.495). The relative indeterminacy is high (0.855).". (b) Explanation: "For the Lose (+) class, the overall sub-model thinks the data has a higher falsity (0.530) relative to its truth (0.470). The relative indeterminacy is high (0.880)."

B.4.2 Second TN example: #8_2_154

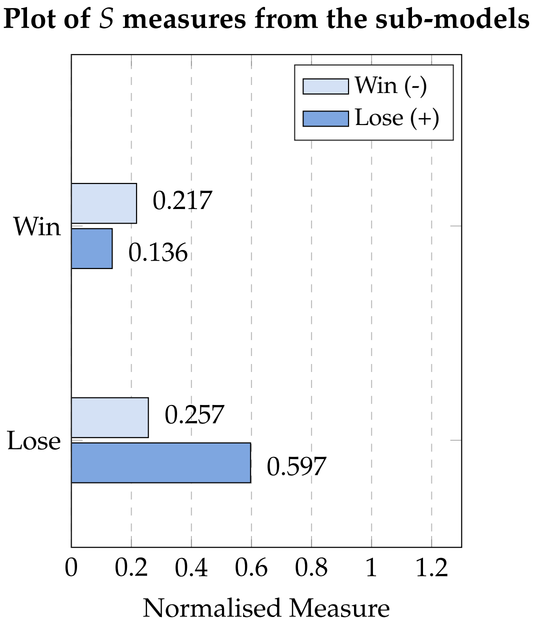


Figure B.18: Chess Dataset: distance measures S from the two sub-models - example of a TN case: #8_2_154. Textual explanation: "Models are in conflict hence, it is less likely the classification is accurate. The data was classified as *Win (-)* by the model."

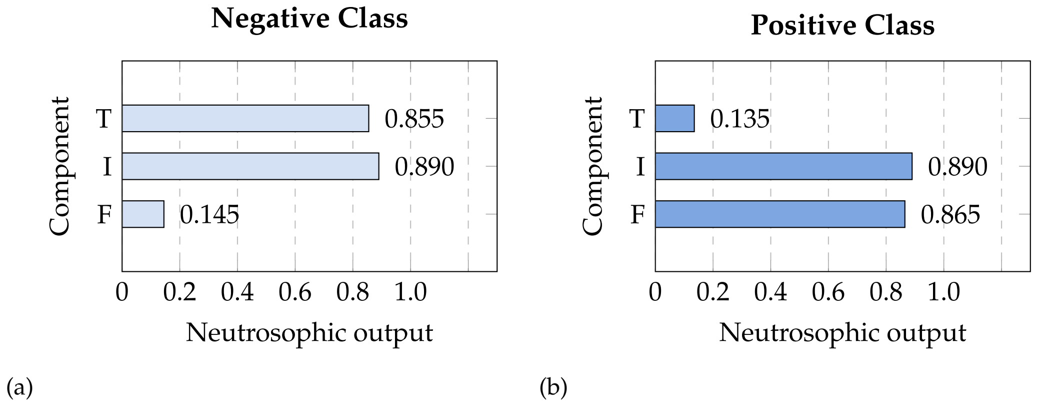


Figure B.19: Chess dataset: graphical explanation of the overall NL sub-model - example of a TN case: #8_2_154. Overall explanation: "Based on the overall sub-model's output the data was more similar to Win (-) compared to Lose (+)." (a) Explanation: "For the Win (-) class, the overall sub-model thinks the data has a higher truth (0.855) relative to its falsity (0.145). The relative indeterminacy is high (0.890).". (b) Explanation: "For the Lose (+) class, the overall sub-model thinks the data has a higher falsity (0.865) relative to its truth (0.135). The relative indeterminacy is high (0.890)."

B.4.3 First FP example: #5_1_51

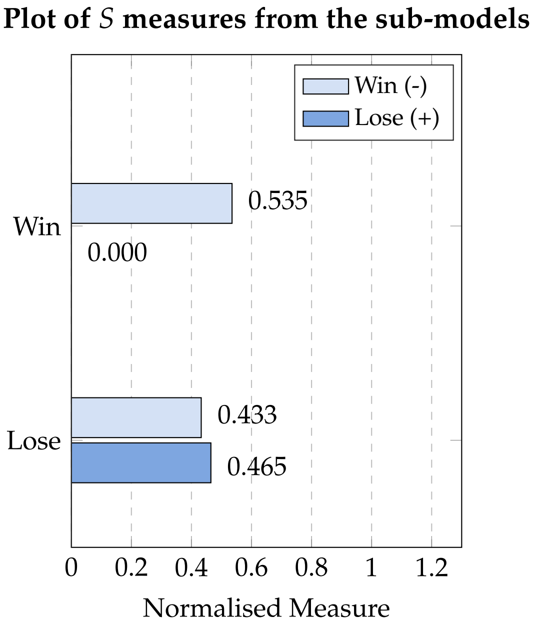


Figure B.20: Chess Dataset: distance measures S from the two sub-models - example of a FP case: #5_1_51. Textual explanation: "Models are in conflict hence, it is less likely the classification is accurate. The data was classified as *Lose (+)* by the model."

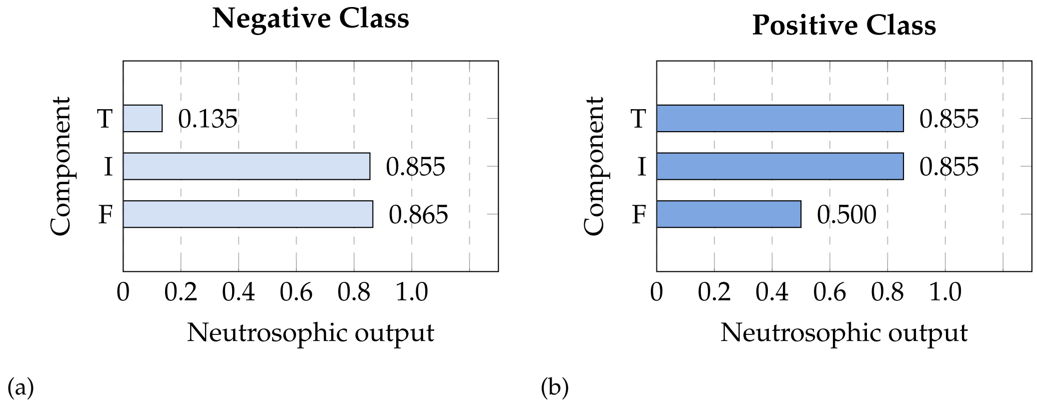


Figure B.21: Chess dataset: graphical explanation of the overall NL sub-model - example of a FP case: #5_1_51. Overall explanation: "Based on the overall sub-model's output the data was more similar to Lose (+) compared to Win (-)." (a) Explanation: "For the Win (-) class, the overall sub-model thinks the data has a higher falsity (0.865) relative to its truth (0.135). The relative indeterminacy is high (0.855).". (b) Explanation: "For the Lose (+) class, the overall sub-model thinks the data has a higher truth (0.855) relative to its falsity (0.500). The relative indeterminacy is high (0.855)."

B.4.4 Second FP example: #5_3_200

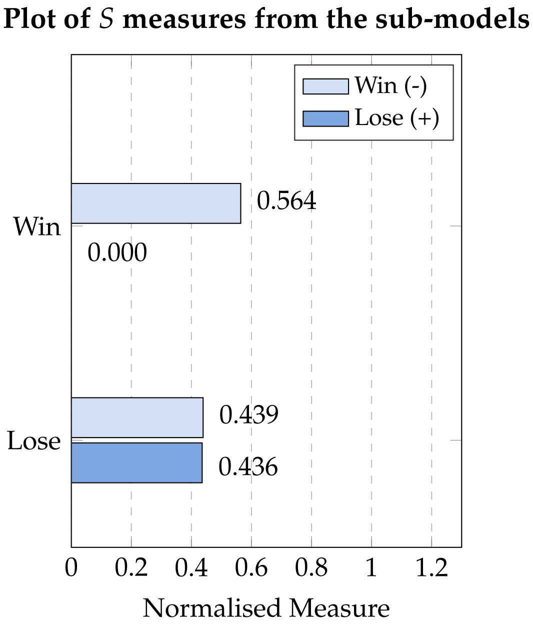


Figure B.22: Chess Dataset: distance measures S from the two sub-models - example of a FP case: #5_3_200. Textual explanation: "Models are in agreement hence, it is more likely the classification is accurate. The data was classified as *Lose (+)* by the model."

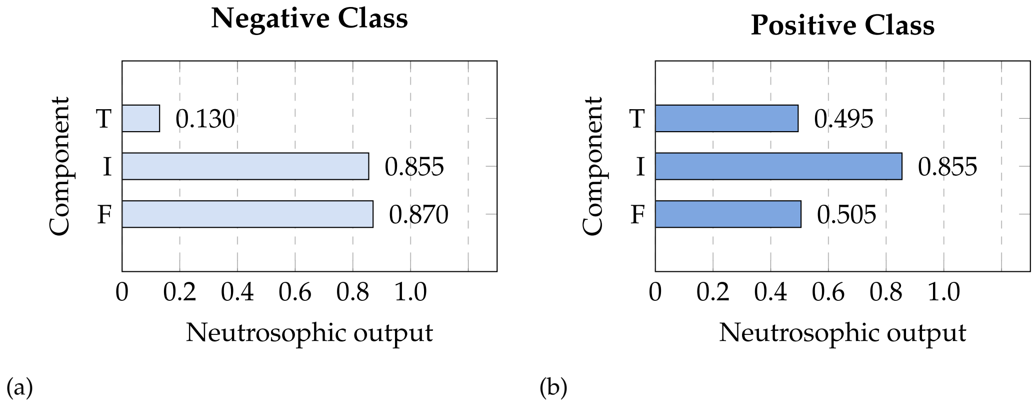
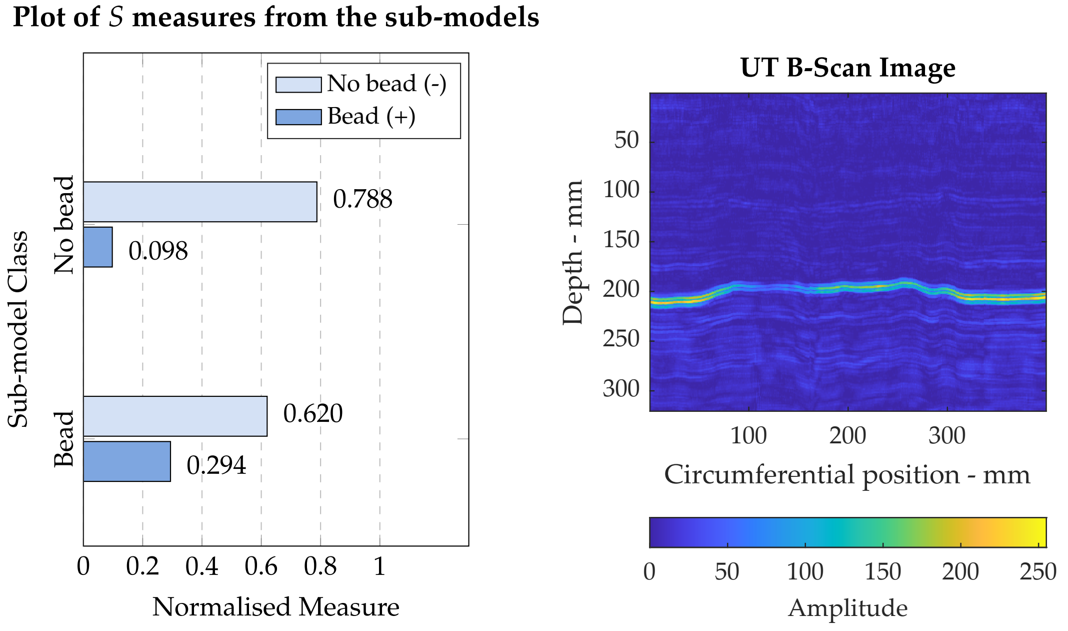


Figure B.23: Chess dataset: graphical explanation of the overall NL sub-model - example of a FP case: #5_3_200. Overall explanation: "Based on the overall sub-model's output the data was more similar to Lose (+) compared to Win (-)." (a) Explanation: "For the Win (-) class, the overall sub-model thinks the data has a higher falsity (0.870) relative to its truth (0.130). The relative indeterminacy is high (0.855).". (b) Explanation: "For the Lose (+) class, the overall sub-model thinks the data has a higher falsity (0.505) relative to its truth (0.495). The relative indeterminacy is high (0.855)."

B.5 Butt-fusion weld bead dataset

B.5.1 TP example: #1_5_175



(a)

(b)

Figure B.24: BF weld bead detection: (a) Illustration of the distance measures S from the two sub-models - example of a TP case: #1_5_175_23_45. Textual explanation: “Models are in agreement hence, it is more likely the classification is accurate.”. (b) Plot of UT image data for the same example. The image was classified as *Bead* by the model.

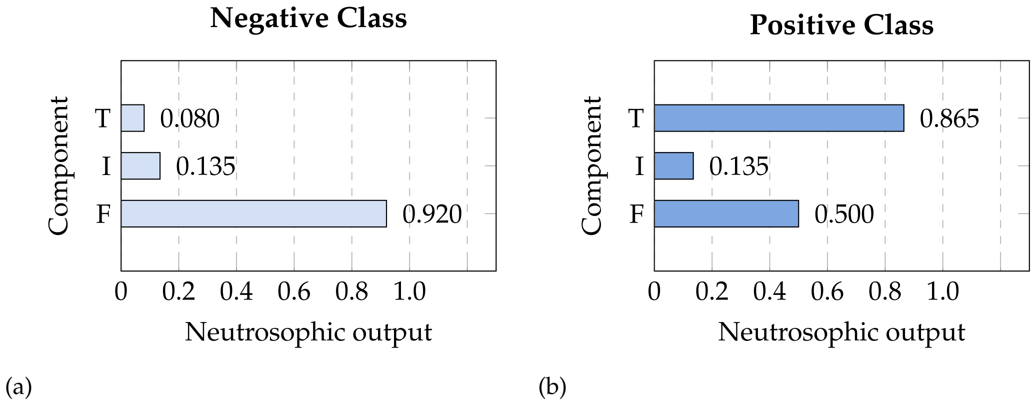


Figure B.25: BF weld bead detection: Graphical explanation of the overall NL sub-model - example of a TP case: #1_5_175_23_45. Overall explanation: “Based on the overall sub-model’s output the data was more similar to Bead (+) compared to No bead (-).” (a) Explanation: “For the No bead (-) class, the overall sub-model thinks the data has a higher falsity (0.920) relative to its truth (0.080). The relative indetermency is low (0.135).”. (b) Explanation: “For the Bead (+) class, the overall sub-model thinks the data has a higher truth (0.865) relative to its falsity (0.500). The relative indetermency is low (0.135).”

B.6 Butt-fusion weld defects dataset

B.6.1 TP example: #2_4_114

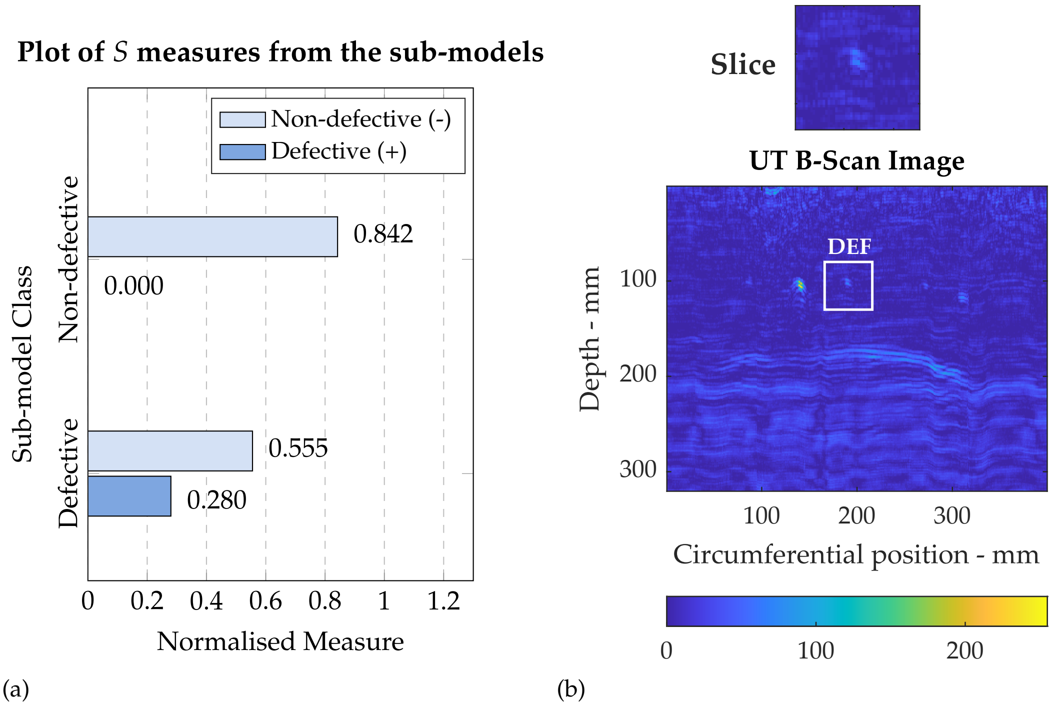
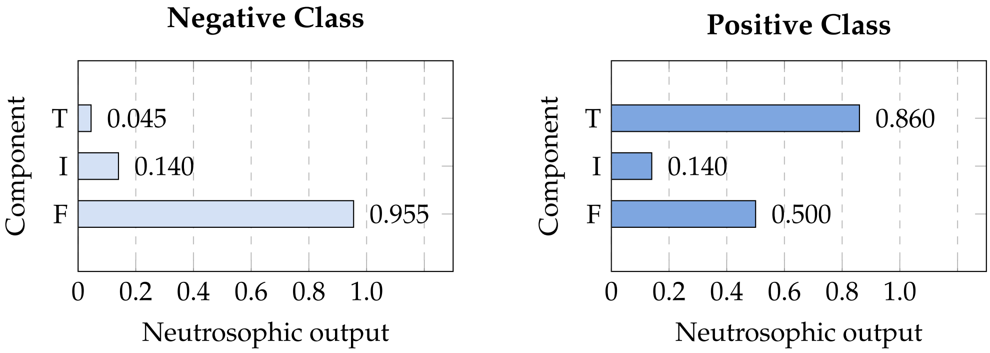


Figure B.26: BF weld defect detection: (a) Illustration of the distance measures S from the two sub-models - example of a TP case: #2_4_114_21_58. Textual explanation: “Models are in agreement hence, it is more likely the classification is accurate.”. (b) Plot of UT image data for the same example. The image was classified as *defective* by the model.



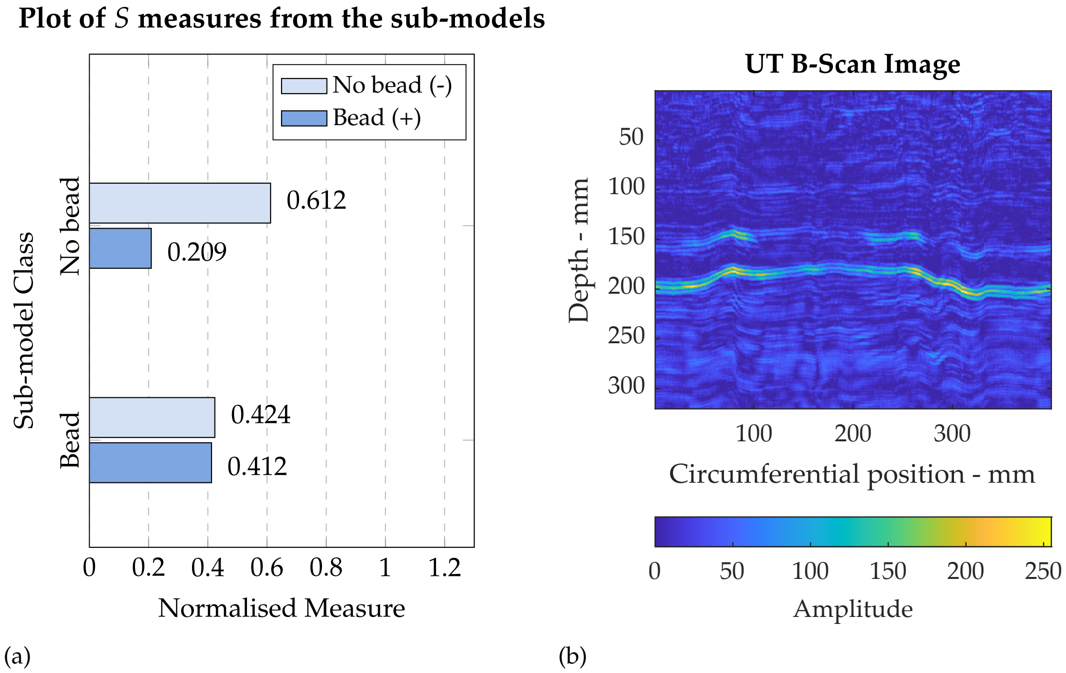
(a) (b)

Figure B.27: BF weld defect detection: Graphical explanation of the overall NL sub-model - example of a TP case: #2_4_114_21_58. Overall explanation: "Based on the overall sub-model's output the data was more similar to Defective (+) compared to Non-defective (-)." (a) Explanation: "For the Non-defective (-) class, the overall sub-model thinks the data has a higher falsity (0.955) relative to its truth (0.045). The relative indetermency is low (0.140).". (b) Explanation: "For the Defective (+) class, the overall sub-model thinks the data has a higher truth (0.860) relative to its falsity (0.500). The relative indetermency is low (0.140)."

C Complete set of explanation examples for chapter 6

C.1 Bead detection: explanation examples

C.1.1 First TP Example: Case #2_2_136_6_48



(a)

(b)

Figure C.1: BF weld bead detection: (a) Illustration of the distance measures S from the two sub-models - example of a TP case: #2_2_136_6_48. Textual explanation: “Models are in agreement hence, it is more likely the classification is accurate. Fuzzy class: 0.65”. (b) Plot of UT image data for the same example. The image was classified as *Bead* by the model.

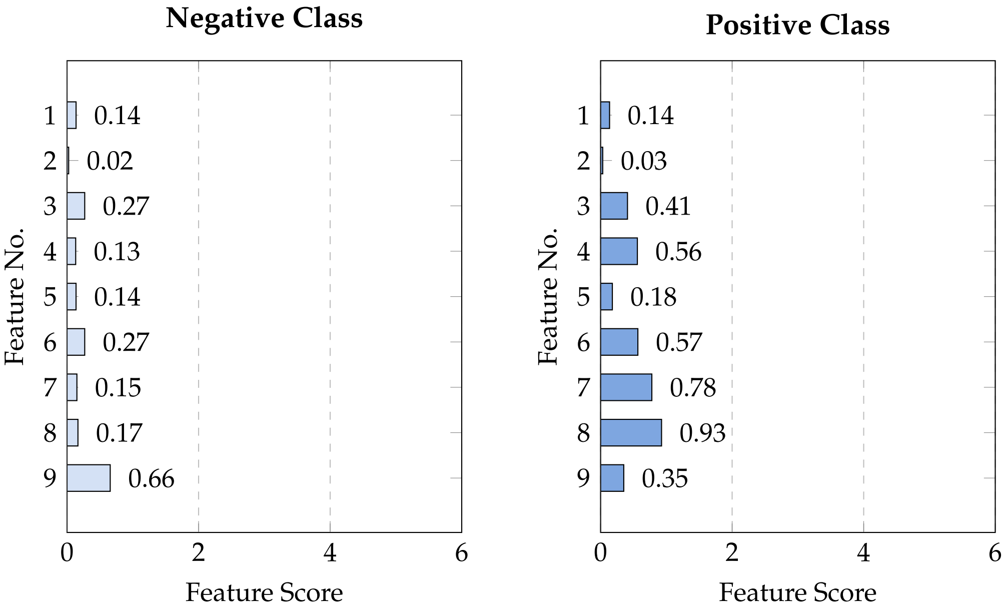


Figure C.2: BF weld bead detection: Graphical explanation of the *No Bead* sub-model - example of a TP case: #2_2_136_6_48. Textual explanation: "No bead model thinks the data is similar to Bead (+ve) and NOT similar to No bead (-ve)."

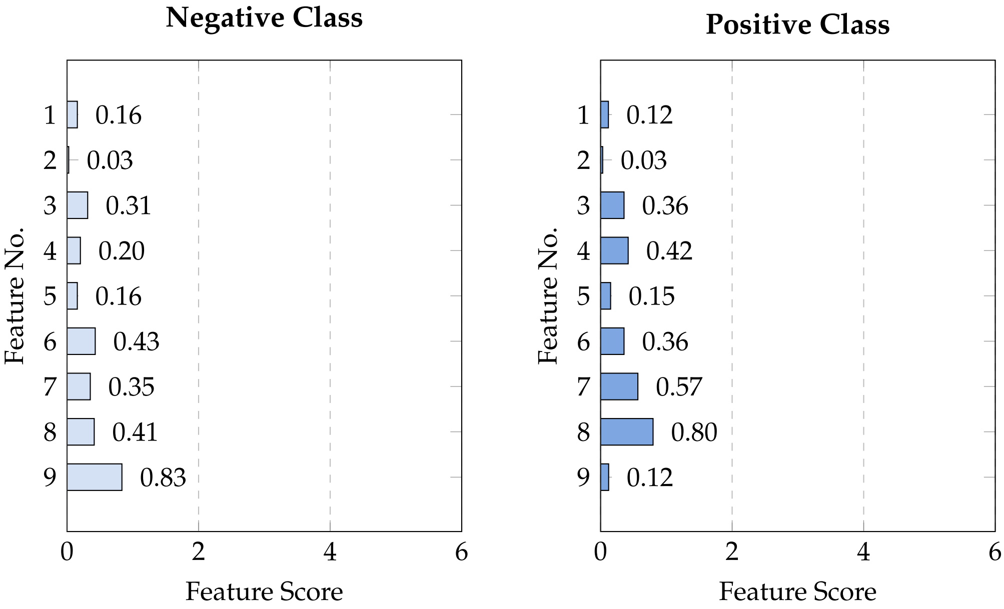
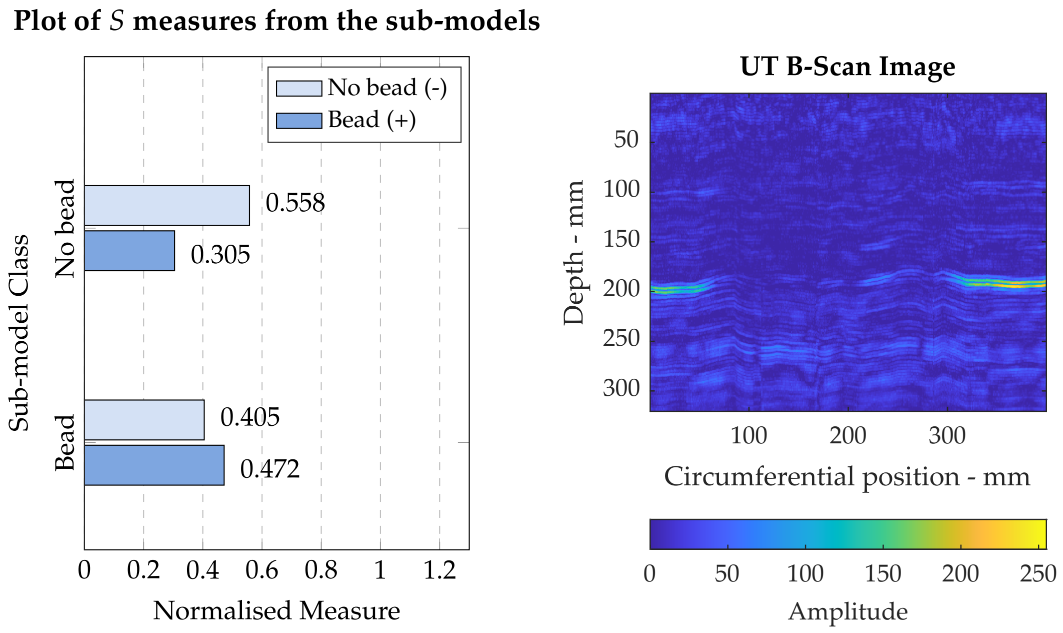


Figure C.3: BF weld bead detection: Graphical explanation of the *Bead* sub-model - example of a TP case: #2_2_136_6_48. Textual explanation: "Bead model thinks the data is more similar to Bead (+ve) despite a high similarity for both."

C.1.2 Second TP Example: #2_2_145_10_35



(a)

(b)

Figure C.4: BF weld bead detection: (a) Illustration of the distance measures S from the two sub-models - example of a TP case: #2_2_145_10_35. Textual explanation: "Models are in conflict hence, it is less likely the classification is accurate. Fuzzy class: 0.57". (b) Plot of UT image data for the same example. The image was classified as *Bead* by the model.

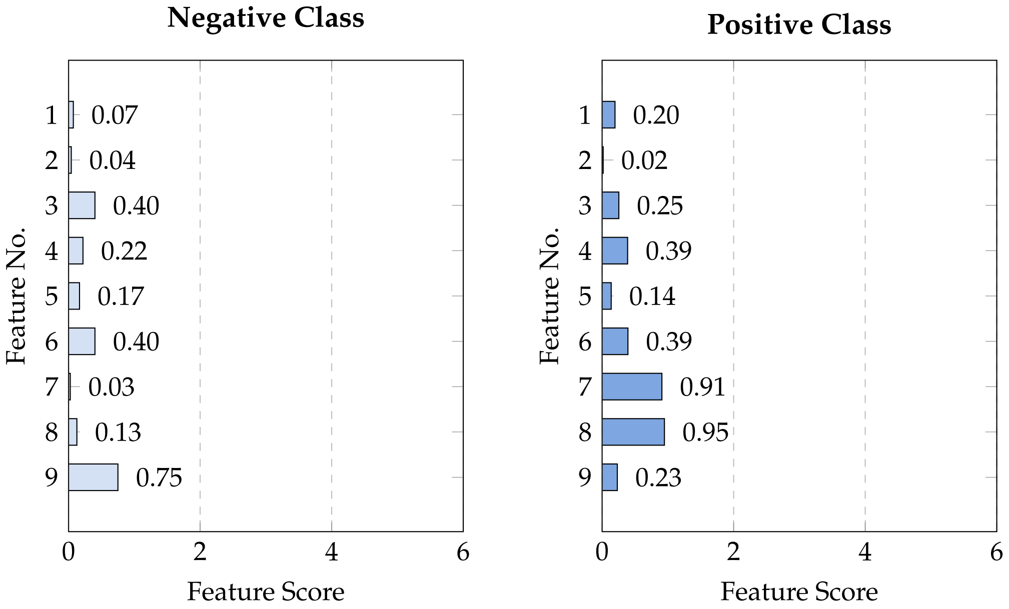


Figure C.5: BF weld bead detection: Graphical explanation of the *No Bead* sub-model - example of a TP case: #2_2_145_10_35. Textual explanation: "No bead model thinks the data is similar to Bead (+ve) and NOT similar to No bead (-ve)."

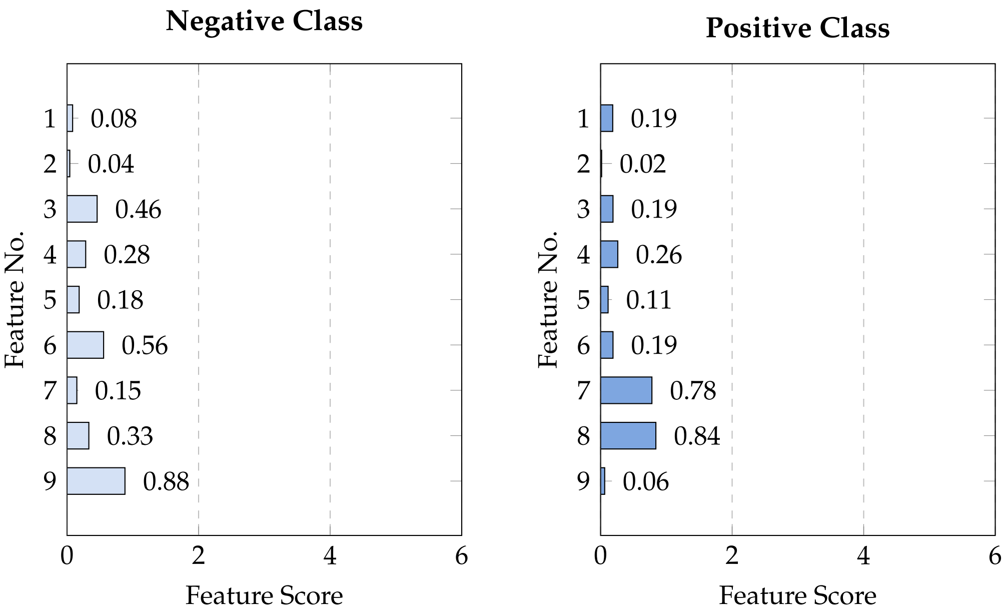


Figure C.6: BF weld bead detection: Graphical explanation of the *Bead* sub-model - example of a TP case: #2_2_145_10_35. Textual explanation: “Bead model thinks the data is more similar to No bead (-ve) despite a high similarity for both.”

C.1.3 First FP Example: Case #2_2_91_33_19

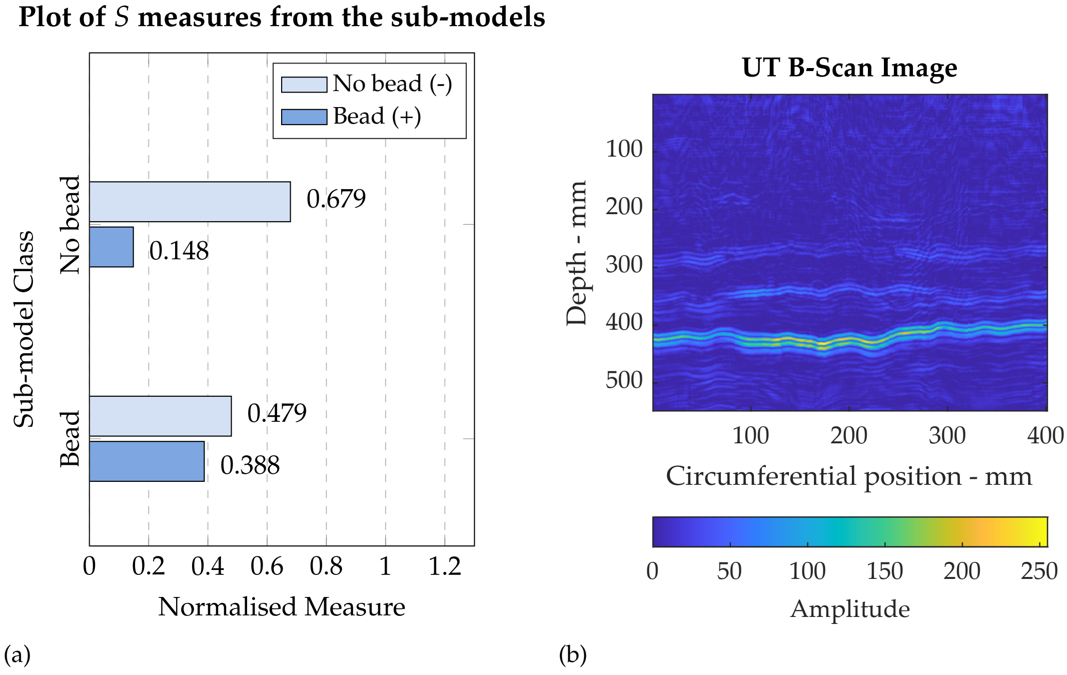


Figure C.7: BF weld bead detection: (a) Illustration of the distance measures *S* from the two sub-models - example of a FP case: #2_2_91_33_19. Textual explanation: “Models are in agreement hence, it is more likely the classification is accurate. Fuzzy class: 0.67”. (b) Plot of UT image data for the same example. The image was classified as *Bead* by the model.

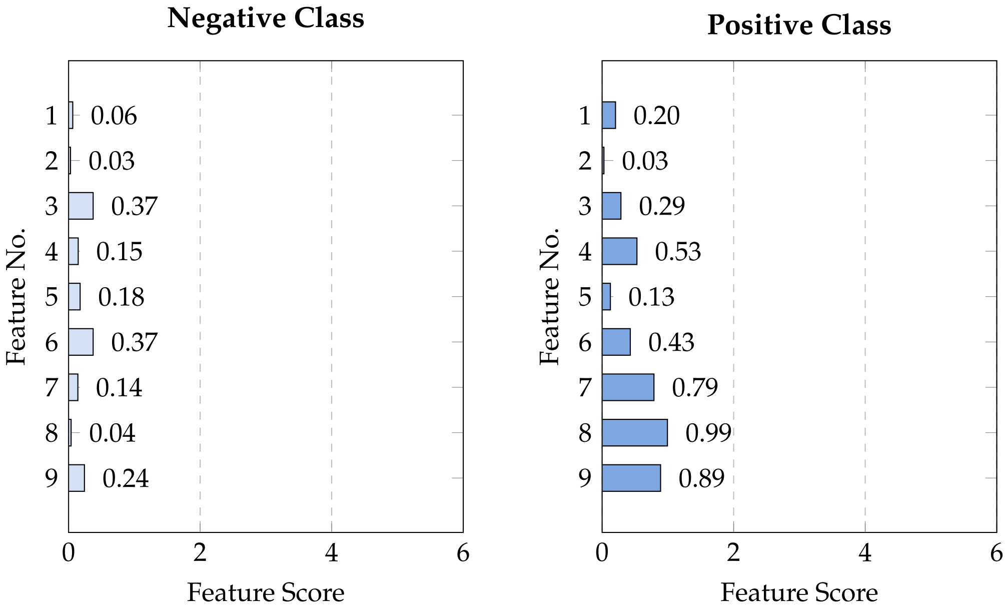


Figure C.8: BF weld bead detection: Graphical explanation of the *No Bead* sub-model - example of a FP case: #2_2_91_33_19. Textual explanation: "No bead model thinks the data is similar to Bead (+ve) and NOT similar to No bead (-ve)."

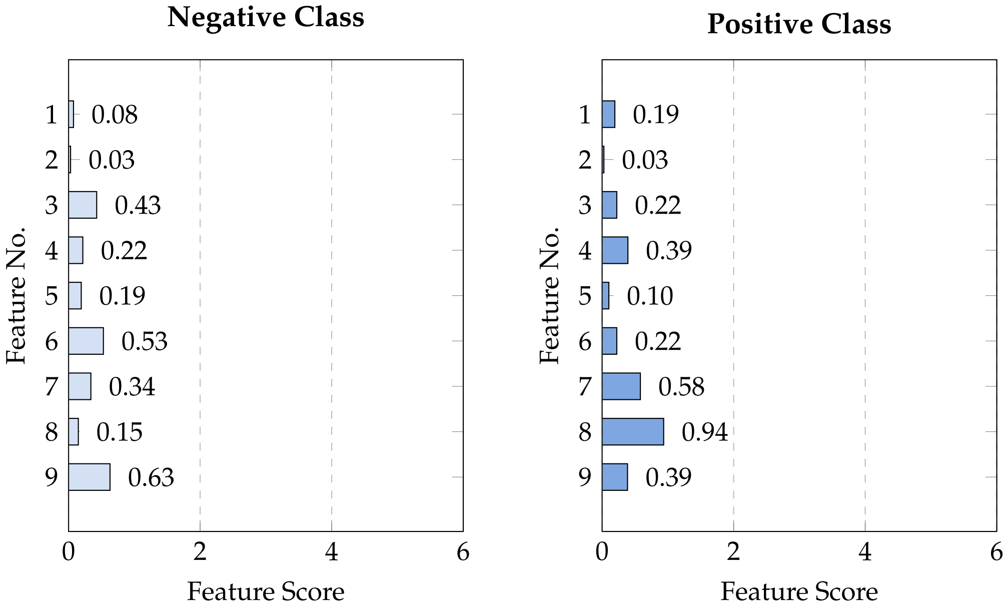
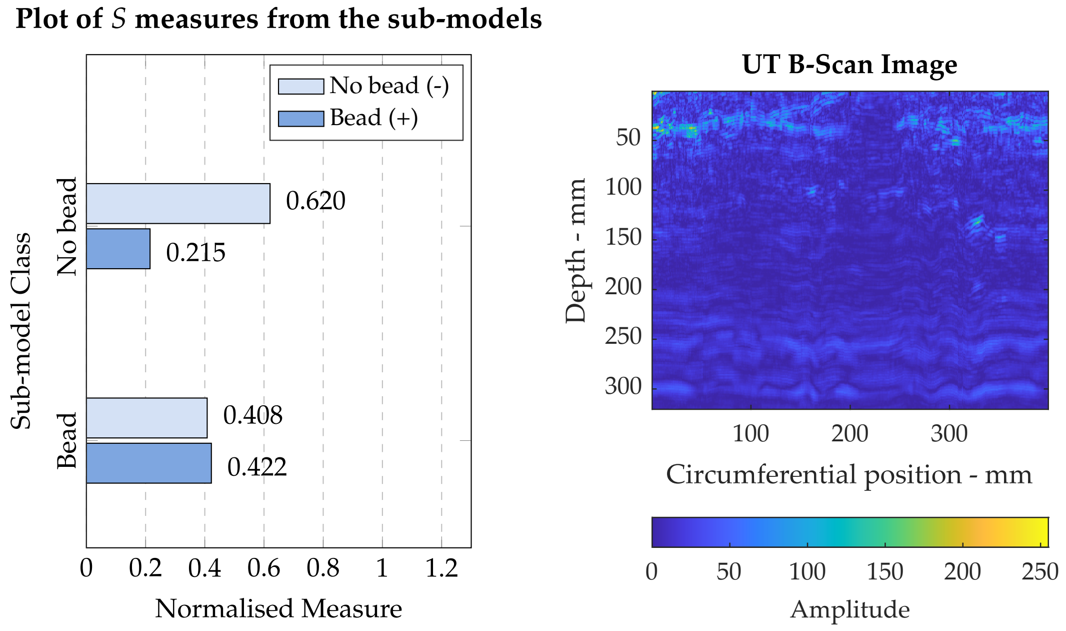


Figure C.9: BF weld bead detection: Graphical explanation of the *Bead* sub-model - example of a FP case: #2_2_91_33_19. Textual explanation: "Bead model thinks the data is more similar to Bead (+ve) despite a high similarity for both."

C.1.4 Second FP Example: Case #2_2_16_8_5



(a)

(b)

Figure C.10: BF weld bead detection: (a) Illustration of the distance measures S from the two sub-models - example of a FP case: #2_2_16_8_57. Textual explanation: "Models are in conflict hence, it is less likely the classification is accurate. Fuzzy class: 0.65". (b) Plot of UT image data for the same example. The image was classified as *Bead* by the model.

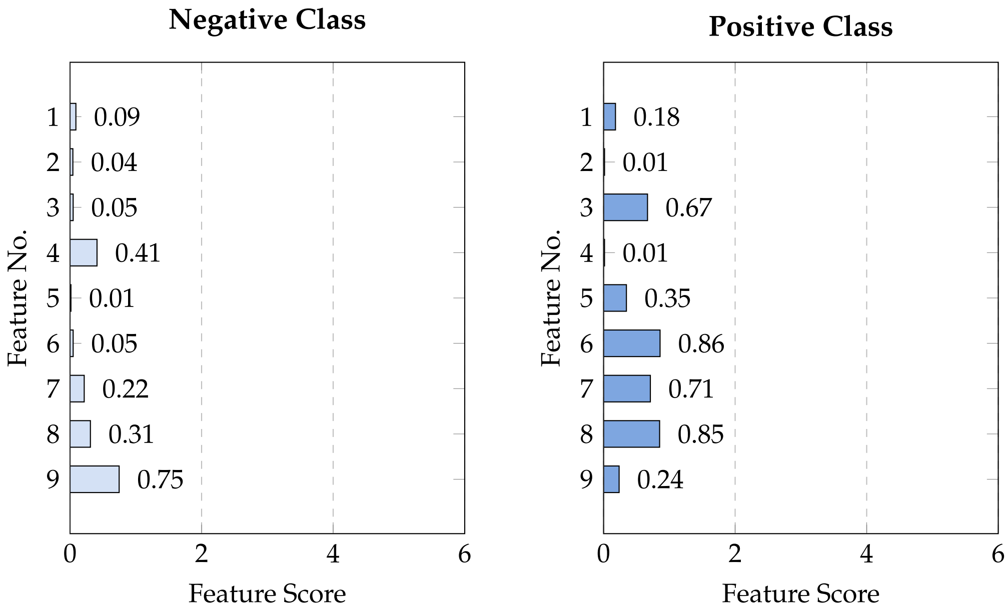


Figure C.11: BF weld bead detection: Graphical explanation of the *No Bead* sub-model - example of a FP case: #2_2_16_8_57. Textual explanation: "No bead model thinks the data is similar to Bead (+ve) and NOT similar to No bead (-ve)."

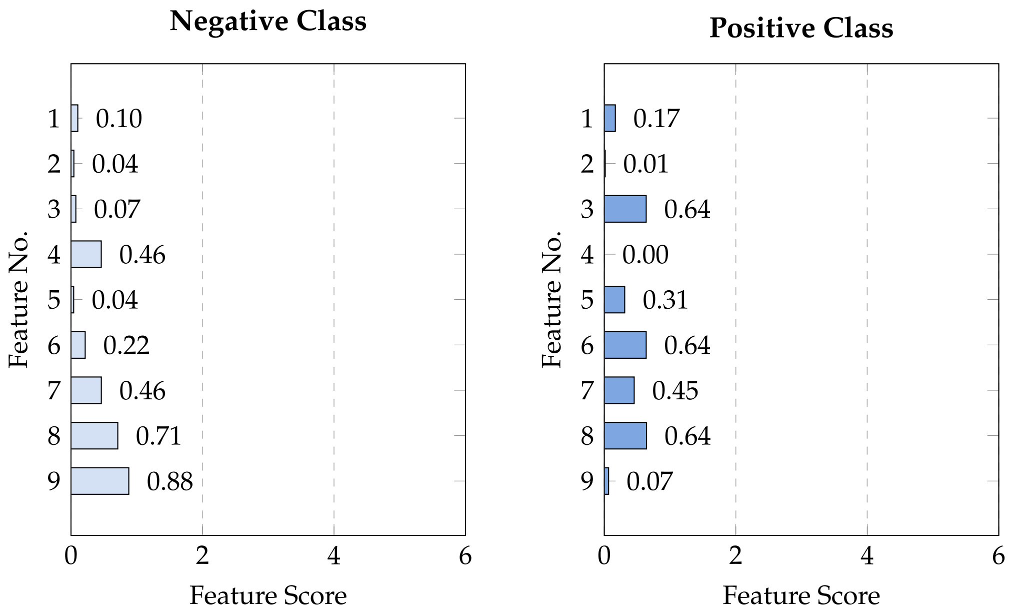


Figure C.12: BF weld bead detection: Graphical explanation of the *Bead* sub-model - example of a FP case: #2_2_16_8_57. Textual explanation: "Bead model thinks the data is more similar to No bead (-ve) despite a high similarity for both."

C.1.5 FN Example: #2_2_151_15_51

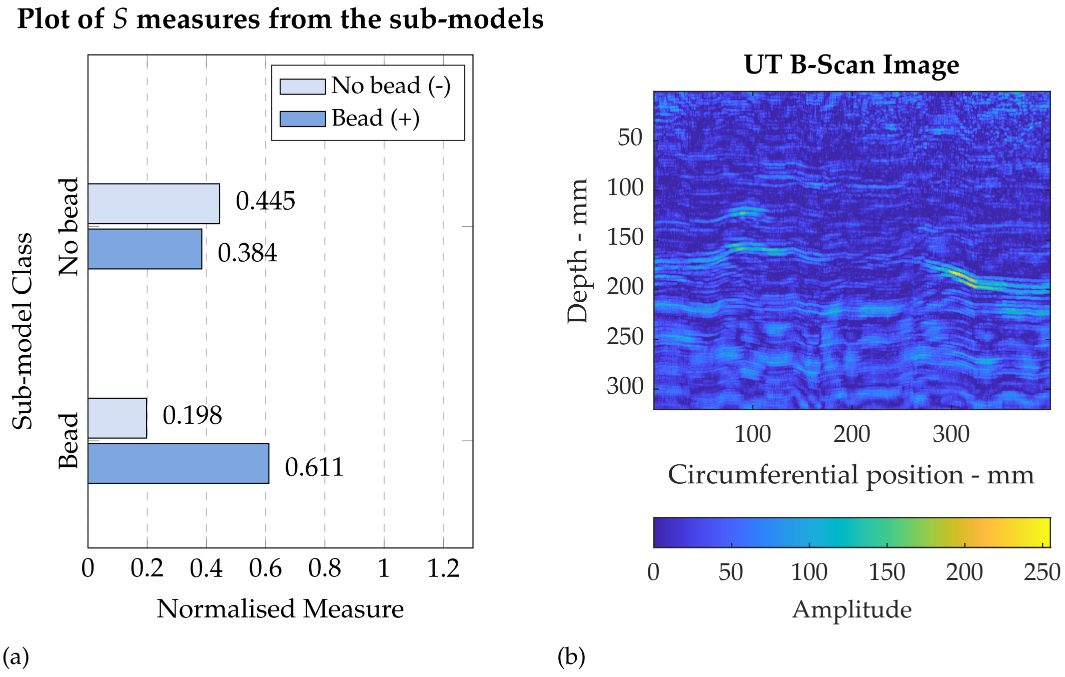


Figure C.13: BF weld bead detection: (a) Illustration of the distance measures S from the two sub-models - example of a FN case: #2_2_151_15_51. Textual explanation: "Models are in conflict hence, it is less likely the classification is accurate. Fuzzy class: 0.37". (b) Plot of UT image data for the same example. The image was classified as *No Bead* by the model.

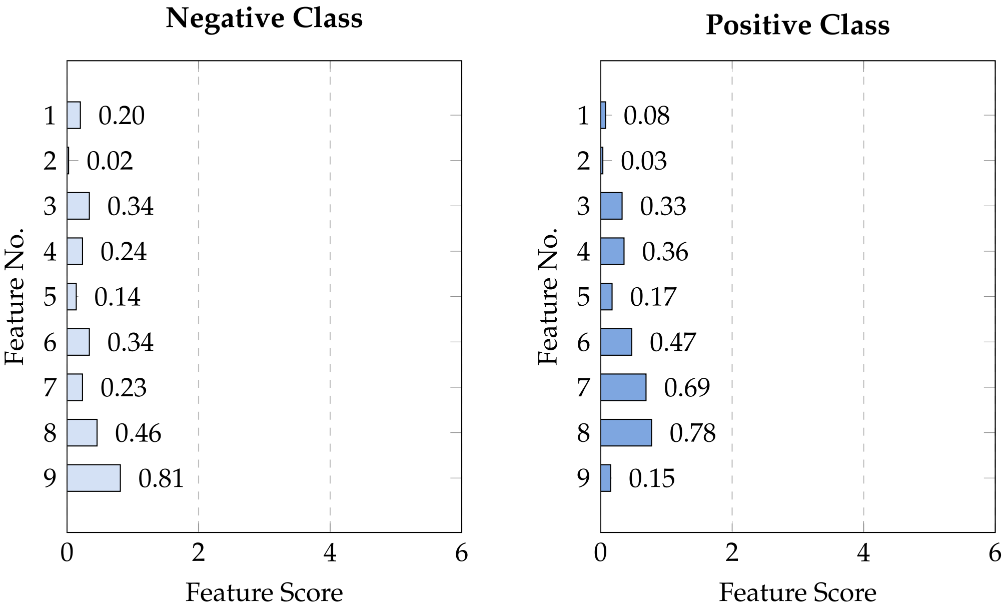


Figure C.14: BF weld bead detection: Graphical explanation of the *No Bead* sub-model - example of a FN case: #2_2_151_15_51. Textual explanation: “No bead model thinks the data is more similar to Bead (+ve) despite a high similarity for both.”

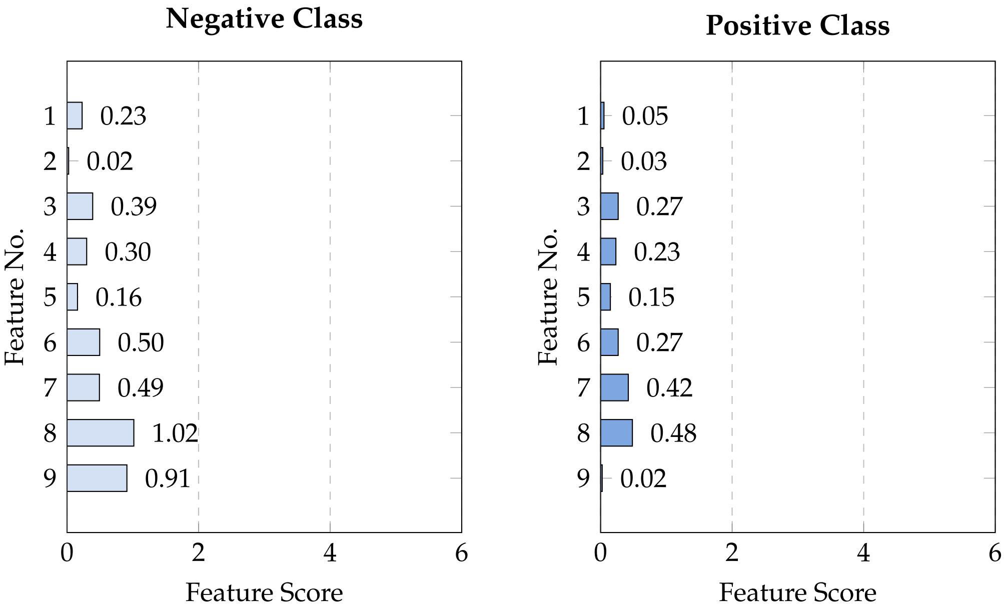


Figure C.15: BF weld bead detection: Graphical explanation of the *Bead* sub-model - example of a FN case: #2_2_151_15_51. Textual explanation: “Bead model thinks the data is similar to No bead (-ve) and NOT similar to Bead (+ve).”

C.2 Defect recognition: explanation examples

C.2.1 FP Example: #2_2_13_2_1

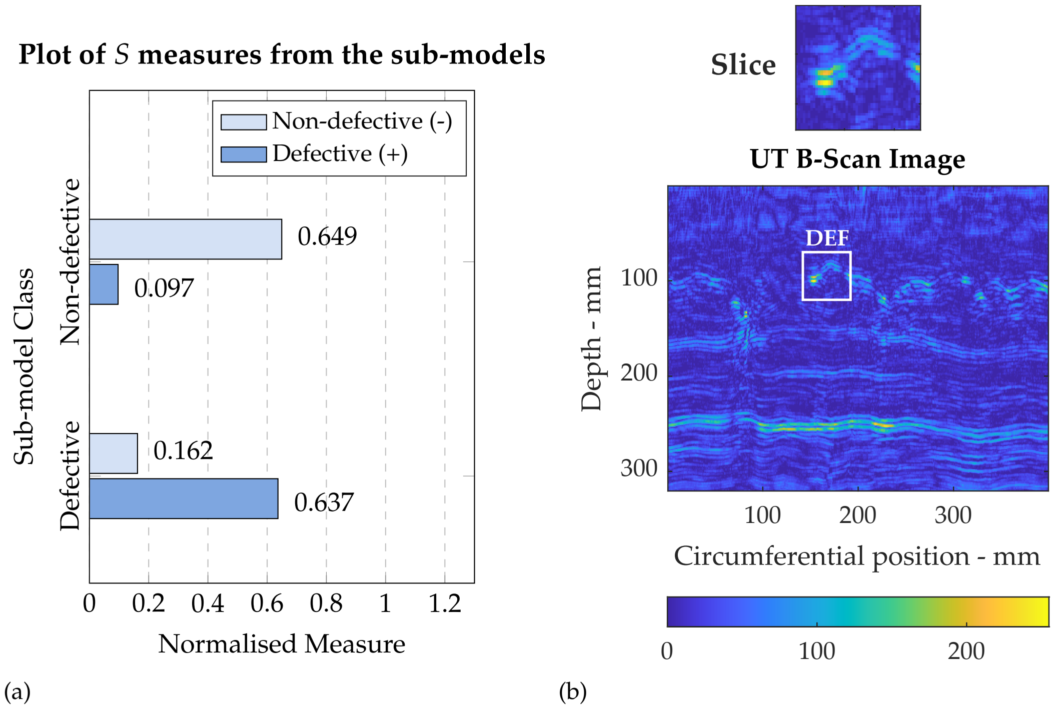


Figure C.16: BF weld defect detection: (a) Illustration of the distance measures S from the two sub-models - example of a FP case: #2_2_13_2_19. Textual explanation: “Models are in conflict hence, it is less likely the classification is accurate. Fuzzy class: 0.53”. (b) Plot of UT image data for the same example. The image was classified as *defective* by the model.

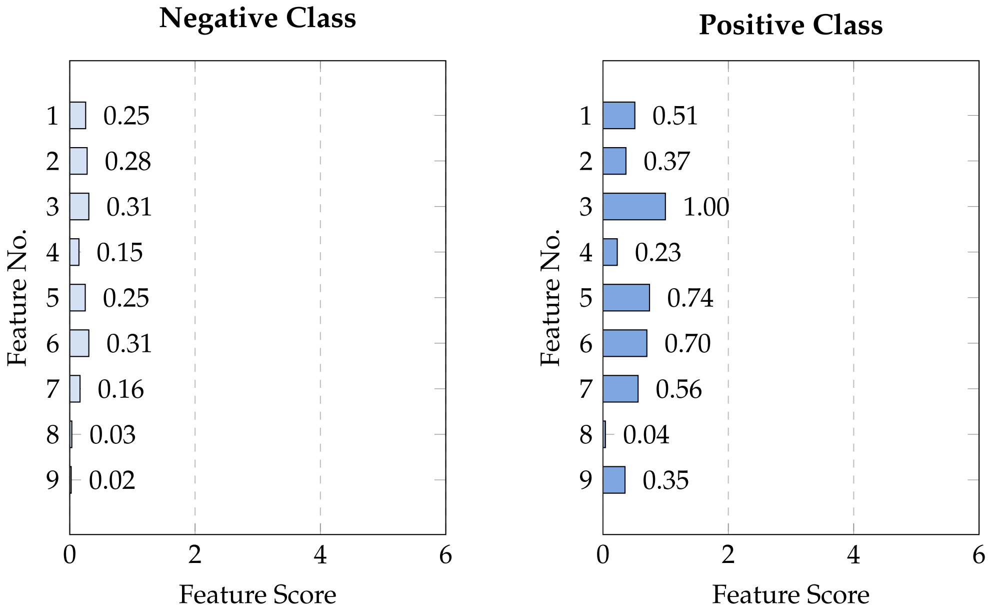


Figure C.17: BF weld defect detection: Graphical explanation of the *Non-defective* sub-model - example of a FP case: #2_2_13_2_19. Textual explanation: "Non-defective model thinks the data is similar to Defective (+ve) and NOT similar to Non-defective (-ve)."

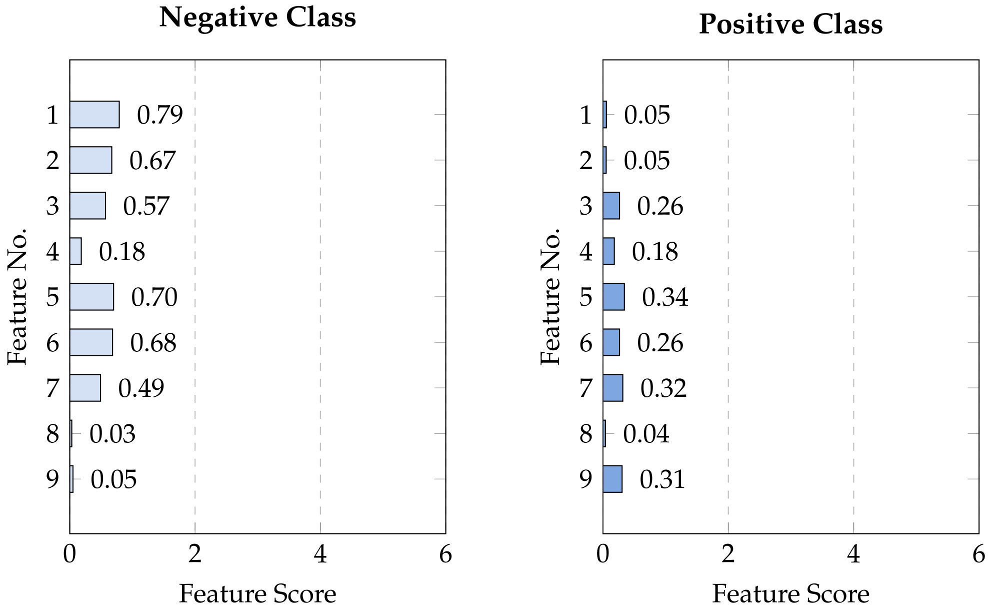
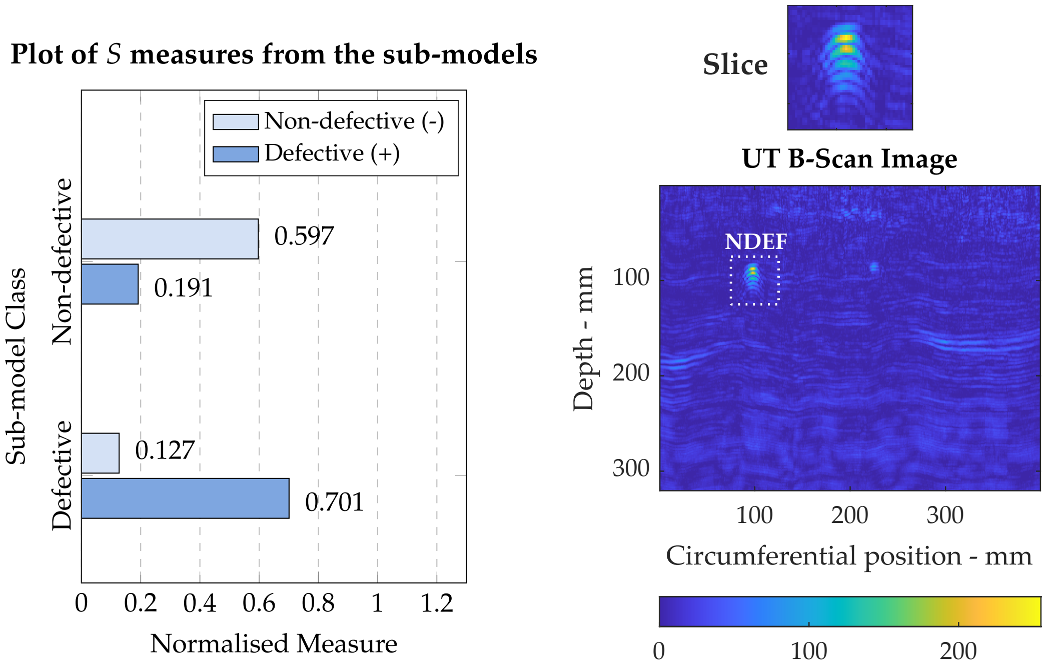


Figure C.18: BF weld defect detection: Graphical explanation of the *Defective* sub-model - example of a FP case: #2_2_13_2_19. Textual explanation: "Defective model thinks the data is similar to Non-defective (-ve) and NOT similar to Defective (+ve)."

C.2.2 FN Example: #2_2_111_20_5



(a) (b)

Figure C.19: BF weld defect detection: (a) Illustration of the distance measures S from the two sub-models - example of a FN case: #2_2_111_20_56. Textual explanation: “Models are in conflict hence, it is less likely the classification is accurate. Fuzzy class: 0.44”. (b) Plot of UT image data for the same example. The image was classified as *non-defective* by the model.

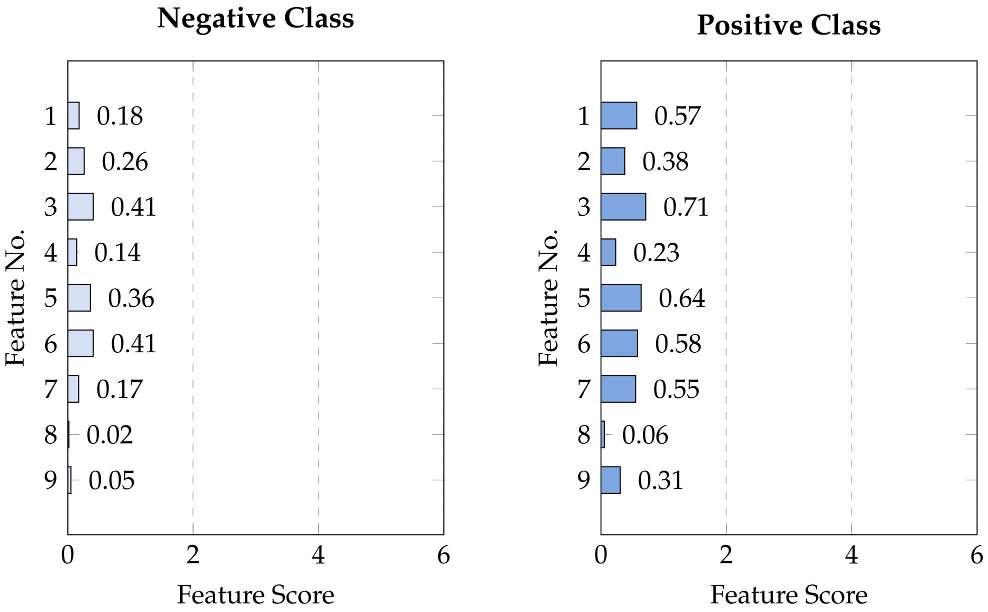


Figure C.20: BF weld defect detection: Graphical explanation of the *Non-defective* sub-model - example of a FN case: #2_2_111_20_56. Textual explanation: “Non-defective model thinks the data is similar to Defective (+ve) and NOT similar to Non-defective (-ve).”

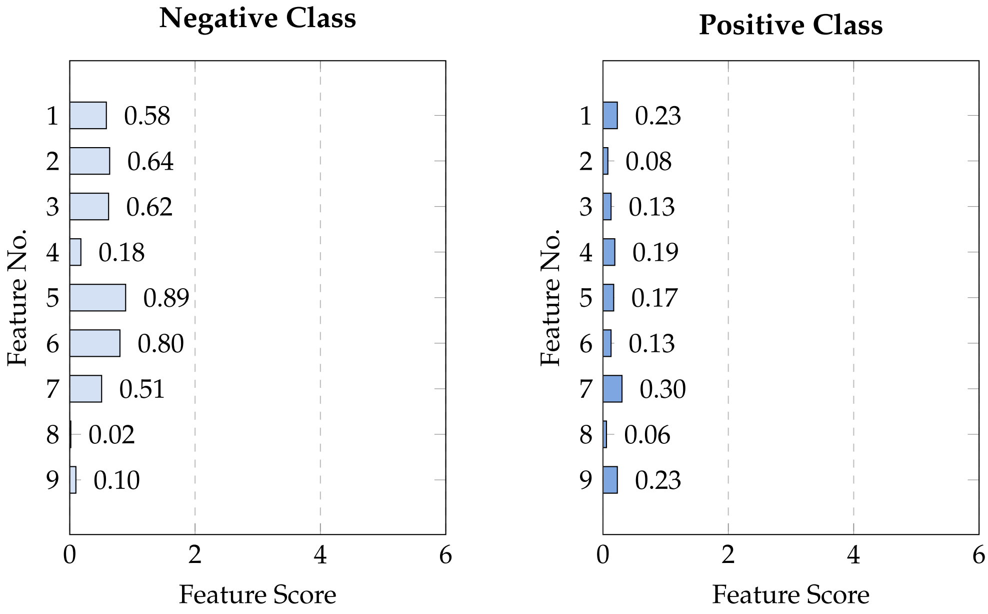


Figure C.21: BF weld defect detection: Graphical explanation of the *Defective* sub-model - example of a FN case: #2_2_111_20_56. Textual explanation: "Defective model thinks the data is similar to Non-defective (-ve) and NOT similar to Defective (+ve)."