JÖNKÖPING UNIVERSITY
Jönköping International Business School

Doctoral Thesis

Trustworthy Explanations

Improved Decision Support Through Well-Calibrated Uncertainty Quantification

Helena Löfström

Jönköping University
Jönköping International Business School
JIBS Dissertation Series No. 159 • 2023

JÖNKÖPING UNIVERSITY
Jönköping International Business School

Doctoral Thesis

Trustworthy Explanations

Improved Decision Support Through Well-Calibrated Uncertainty Quantification

Helena Löfström

Jönköping University
Jönköping International Business School
JIBS Dissertation Series No. 159 • 2023

Doctoral Thesis in Informatics

Trustworthy Explanations: Improved Decision Support Through
Well-Calibrated Uncertainty Quantification
JIBS Dissertation Series No. 159

Published by
Jönköping International Business School, Jönköping University
P.O. Box 1026
SE-551 11 Jönköping
Tel. +46 36 10 10 00
www.ju.se

Printed by Stema Specialtryck AB, 2023

Trycksak
3041 0234

ISSN: 1403-0470
ISBN: 978-91-7914-031-1 (Printed version)
ISBN: 978-91-7914-032-8 (Online version)

"*You can never get a cup of tea large enough or a book long enough to suit me."*

C.S. Lewis

Abstract

The use of Artificial Intelligence (AI) has transformed fields like disease diagnosis and defence. Utilising sophisticated Machine Learning (ML) models, AI predicts future events based on historical data, introducing complexity that challenges understanding and decision-making. Previous research emphasizes users' difficulty discerning when to trust predictions due to model complexity, underscoring addressing model complexity and providing transparent explanations as pivotal for facilitating high-quality decisions.

Many ML models offer probability estimates for predictions, commonly used in methods providing explanations to guide users on prediction confidence. However, these probabilities often do not accurately reflect the actual distribution in the data, leading to potential user misinterpretation of prediction trustworthiness. Additionally, most explanation methods fail to convey whether the model's probability is linked to any uncertainty, further diminishing the reliability of the explanations.

Evaluating the quality of explanations for decision support is challenging, and although highlighted as essential in research, there are no benchmark criteria for comparative evaluations.

This thesis introduces an innovative explanation method that generates reliable explanations, incorporating uncertainty information supporting users in determining when to trust the model's predictions. The thesis also outlines strategies for evaluating explanation quality and facilitating comparative evaluations. Through empirical evaluations and user studies, the thesis provides practical insights to support decision-making utilising complex ML models.

Keywords - Explainable Artificial Intelligence, Interpretable Machine Learning, Decision Support Systems, Uncertainty Estimation, Explanation Methods

Sammanfattning

Användningen av Artificiell intelligens (AI) har förändrat områden som diagnosticering av sjukdomar och försvar. Genom att utnyttja sofistikerade maskininlärningsmodeller predicerar AI framtida händelser baserat på historisk data. Modellernas komplexitet resulterar samtidigt i utmanande beslutsprocesser när orsakerna till prediktionerna är svårbegripliga. Tidigare forskning pekar på användares problem att avgöra prediktioners tillförlitlighet på grund av modellkomplexitet och belyser vikten av att tillhandahålla transparenta förklaringar för att underlätta högkvalitativa beslut.

Många maskininlärningsmodeller erbjuder sannolikhetsuppskattningar för prediktionerna, vilket vanligtvis används i metoder som ger förklaringar för att vägleda användare om prediktionernas tillförlitlighet. Dessa sannolikheter återspeglar dock ofta inte de faktiska fördelningarna i Datat, vilket kan leda till att användare felaktigt tolkar prediktioner som tillförlitliga. Därutöver förmedlar de flesta förklaringsmetoder inte om prediktionernas sannolikheter är kopplade till någon osäkerhet, vilket minskar tillförlitligheten hos förklaringarna.

Att utvärdera kvaliteten på förklaringar för beslutsstöd är utmanande, och även om det har betonats som avgörande i forskning finns det inga benchmark kriterier för jämförande utvärderingar.

Denna avhandling introducerar en innovativ förklaringsmetod som genererar tillförlitliga förklaringar, inkluderande osäkerhetsinformation, för att stödja användare att avgöra när man kan lita på modellens prediktioner. Avhandlingen ger också förslag på strategier för att utvärdera kvaliteten på förklaringar och underlätta jämförande utvärderingar. Genom empiriska utvärderingar och användarstudier ger avhandlingen praktiska insikter för att stödja beslutsfattande vid användande av komplexa maskininlärningsmodeller.

List of Papers

Primary Papers:
The following papers are included in the thesis.

Paper I[1]**:** Löfström, H., Löfström, T., & Johansson, U. (2018). Interpretable Instance-based Text Classification for Social Science Research Projects. *Archives of Data Science, Series A*, 5(1).

Paper II: Löfström, H., Hammar, K., & Johansson, U. (2022). A Meta Survey of Quality Evaluation Criteria in Explanation Methods. *In Intelligent Information Systems: CAiSE Forum 2022*, Leuven, Belgium, June 6–10, 2022, Proceedings (pp. 55-63). Cham: Springer International Publishing.

Paper III: Löfström, H. (2023) On the Definition of Appropriate Trust: and the Tools that Come with it. ICDATA 2023, Las Vegas, USA, July 24-27, 2023. **UNDER PUBLISHING**

Paper IV: Löfström, H., Löfström, T., Johansson, U., & Sönströd, C. (2023). Investigating the Impact of Calibration on the Quality of Explanations. *Annals of Mathematics and Artificial Intelligence*, 1-18.

Paper V: Löfström, H., Löfström, T., Johansson, U., & Sönströd, C. (202X). Calibrated Explanations with Uncertainty Information and Counterfactuals (2023). *Expert Systems with Applications*. **RESUMBISSION UNDER REVIEW**

Related Papers:
The following paper also contribute to the thesis, although of secondary importance.

I: Johansson, U., Löfström, T., Sönströd, C., & Löfström, H. (2023). Conformal Prediction for Accuracy Guarantees in Classification with Reject Option. In Modeling Decisions for Artificial Intelligence: 20th International Conference, MDAI 2023, Umeå, Sweden, June 19–22, 2023, Proceedings (pp. 133-145). Cham: Springer Nature Switzerland.

II: Löfström, T., Löfström,H., Johansson, U., Sönströd, C., Matela, R. (202X) Calibrated Explanations for Regression with Cunjunctional Rules. **UNDER REVIEW**

Acknowledgements

It is said that it requires a village to raise a child; I would like to add that it takes a whole bunch of people to raise a PhD student. Thank you, bunch of people, you finally managed to raise the headstrong PhD student: Ulf Siegerroth, Ulf Johansson, Cecilia Sönström, Karl Hammar, Thomas Müllern, Jan Fång and George Grönwald.

Cecilia, we really have to dive into the interesting topics of tea and books more often. Perhaps over a cup of tea?

Thank you Stella and Samuel Cavallin for giving me the courage to follow my dreams wherever they may lead.

I want to add that this dissertation would never have been possible without the support of my family. Nathanael, Sam, Kristina, Ingrid, Erik, Signe and Elsa, thank you for your encouraging words and for all hours of interesting discussions.

Finally, Tuwe, you have been, as always, a steadfast pillar of support. Through all the ups, downs, and turnarounds with the thesis, you have been there. You are the best!

Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Meaning
AI	**A**rtificial **I**ntelligence
AT	**A**ppropriate **T**rust
CS	**C**omputer **S**cience
CP	**C**onformal **P**rediction
DSR	**D**esign **S**cience **R**esearch
DSS	**D**ecision **S**upport **S**ystems
DT	**D**ecision **T**heory
ECE	**E**xpected **C**alibration **E**rror
HCI	**H**uman **C**omputer **I**nteraction
IS	**I**nformation **S**ystems
ML	**M**achine **L**earning
MMR	**M**ixed **M**ethod **R**esearch
PAVA	**P**air-**A**djacent **V**iolator's **A**lgorithm
PE	**P**robability **E**stimate
RF	**R**andom **F**orest
RSS	**R**ecommendation **S**ystem**s**
SOTA	**S**tate **O**f **T**he **A**rt
VA	**V**enn-**A**bers
XAI	E**x**plainable **A**rtificial **I**ntelligence

List of Terms

Term	Description
Feature	Columns or variables in the data that describe the relevant objects in the domain of a model, e.g., the age of customers modelling for shopping behaviour or the level of BMI for modelling risk of diabetes.
Class	Subset of objects in a dataset that have something in common, e.g., dogs and cats in a dataset with animals. The name of the class is often referred to as the *class label*.
Model	The output of a ML algorithm applied to training data.
Prediction	The output from a predictive model when applied to new data. In classification, it can be presented with an estimated probability for a class label.
Explanation Method	In this thesis meaning a post hoc method applied on top of a ML model aimed at explaining the rationale of the predictions from the underlying model. The term 'method' is commonly used in IS with a different definition. However, in this thesis, if not said otherwise, the term is used as denoting an explanation method.
Feature weight	The importance of a feature, measured as the increase in the model's prediction error after permutation of a feature's values.
Calibration	How the underlying model's probability estimates correspond to the actual probabilities in the model. For a model to be well calibrated, the predicted probabilities must be matched by the observed accuracy. In other words, the predicted probability of an event should correctly correspond to the observed probability for that event.
Trust	When the user agrees with the prediction from the model.

Appropriate Trust	This thesis uses the term to signify if a user can identify correct and incorrect predictions based on the provided explanations. This is similar to the definition in the work of DARPA (Gunning & Aha, 2019) and (Yang et al., 2020), where appropriate trust is defined as the user's ability to *"[not] follow an [in]correct recommendation"*. Appropriate trust is binary for each prediction; the user can either correctly or incorrectly identify the prediction. If a user can correctly identify a noteworthy amount of predictions with a low amount of incorrect, the user is said to have high appropriate trust in the model.
AT	A metric, measuring the level of users appropriate trust (averaged over many predictions) in a set of example predictions. Also referred in the thesis as a user's performance, or accuracy.

1 Introduction

"*Unseen in the background, Fate was quietly slipping lead into the boxing glove.*"
P.G. Wodehouse, Very Good, Jeeves!

In this first chapter, the problem under study is introduced, and its significance for the research area of *explainable artificial intelligence* (XAI), is discussed. The research questions considered in this work are formulated, and the aim of the thesis is presented. This chapter also describes how the sub-questions are related to each other and to the main research question through the contributions of the papers included in this thesis. The re- search sub-questions and their connections to the contributions of this work are visualised in Figure 1.1.

1.1 Background

In the early hours of 26th September 1983, in one of the Soviet Union's missile warning bases, duty officer Stanislav Petrov was staring at a large, back-lit screen. In a later interview with the BBC, he recalled that it was entirely red, with the single word 'Launch' displayed on it[1]. The siren had started howling, indicating that the system had registered the launch of an intercontinental ballistic missile, from the United States.

Petrov was part of a well-trained team, with clear instructions to register missile strikes and report them to the Soviet military and political leadership. In the BBC interview, he said that all he had to do was to reach for the phone and report the attack, which would almost certainly trigger a retaliation. Every second he hesitated meant, they were losing valuable time. The system indicated the highest possible level of reliability. However, Petrov did hesitate; he was not convinced.

As an IT specialist, he knew that the system could make errors. He turned to the satellite radar operators, but they had not registered any missiles. His instructions were that he should base his decisions solely on the computer readouts, he recalled in the interview, and that humans were only a support service, nothing more. He doubted the system, but still calculated that the odds were about 50/50 that it was correct[2]. f he were wrong, the Soviet would be hit by nuclear missiles within minutes. If the system was wrong, the world could see its first nuclear world war. The situation was agitated between the two superpower nations[3]which must have caused the utmost stress. Nevertheless, in direct contrast to his instructions from training, he decided to report a system malfunction.

[1] www.bbc.com/news/world-europe-24280831
[2] nytimes.com/2017/09/18/world/europe/stanislav-petrov-nuclear-war-dead.html
[3] nsarchive2.gwu.edu/NSAEBB/NSAEBB426

This incident involving Petrov serves as a dramatic reminder of the catastrophic consequences that can arise from erroneous predictions. Whether an investor is contemplating a significant financial commitment, a company is striving to retain its customers, or a physician is grappling with a diagnosis, a decision based on an erroneous prediction could prove disastrous (Wynants et al., 2020).

Duty officer Petrov was using an early version of a *decision support system* (DSS), which organises and analyses the data put into the system to support data-driven decisions with high quality. DSS is an area in *Information systems* (IS), where the focus is on supporting and improving decision-making (Arnott & Pervan, 2014). There are various types of DSS, such as *expert systems*, *data analytic systems*, and *recommendation systems* (RSS) (Arnott & Pervan, 2014). The traditional assumption in the DSS literature is that if decision-makers are provided with expanded processing capabilities, they will use them to analyse problems in more depth and make better decisions (Todd & Benbasat, 1992)); however, the mere use of a DSS does not necessarily result in the reduction of human error, and could cause the creation of new classes of error, such as errors of *omission* (where events are missed when no explicit prompt is given about them by the DSS) and *commission* (where the user does what the DSS recommends, even when it contradicts their training (Skitka et al., 1999)).

Today *Artificial Intelligence* (AI)) is frequently incorporated into many types of DSS to enable high- performance predictions (e.g., medical diagnoses) from historical data (such as earlier patients). *Machine learning* (ML) algorithms are often used in AI due to their broad spectrum of potential applications (Negnevitsky, 2005). However, many of the most accurate ML algorithms behave like black boxes; the data is fed into the box, and a prediction is produced that is almost impossible to back-trace and understand (Guidotti et al., 2018; Mueller et al., 2019; Ribeiro et al., 2016). If decisions are to be made based on the predictions from such a system, the user must understand the logic behind them, i.e., they must be explained. In other words, explanations are essential for the proper use of a DSS. Without explanations, users may have difficulty in identifying and rejecting a system's recommendation when it is incorrect, and they may be reluctant to accept its advice even when it is correct (Dzindolet et al., 2003; Lacave & Diez, 2002, 2004; Martens & Foster, 2014). At the same time, users are more willing to use a system if they are given a reason for erroneous answers (Dzindolet et al., 2003). The rationale underlying methods dedicated to offering such explanations, commonly referred to as *explanation methods*, is to provide users with additional insights into the justifications behind predictions generated by complex ML models.

One common approach is to use an ML model to predict the category or class of new data samples, for example to determine whether a new patient has diabetes. This type of task, where the goal is to find the most probable class for the data, is called *classification*. ML models for classification provide probability estimates for their predictions, which serve as a guide for the user as to the quality of each prediction (Negnevitsky, 2005). Explanation methods often use these probabilities to calculate the impact, or *feature weight*,of each feature of the predictions. However, these estimated probabilities are often poorly calibrated, meaning they are not well aligned with the probabilities observed in the data (Grushka-Cockayne et al., 2017; Van Calster et al., 2019). In essence, a poorly calibrated ML model is one that fails to assign predictions with accurate probabilities. Ideally, if a model assigns a 70% probability to an event, that event should occur roughly 70%of the time; however, a poorly calibrated model consistently overestimates or underestimates the actual probabilities, leading to unreliable predictions. Conversely, a well-calibrated model accurately reflects the probabilities of the predictions. Thus, it may be challenging for the user of a poorly calibrated

ML model to identify incorrect predictions, which undermines a high level of appropriate trust in the model rather than supporting it. This alignment between estimated and actual probabilities may be crucial in terms of creating trustworthy explanations that accurately represent the model's behaviour. Ensuring that the feature weights correspond to the correct impact on the probability estimates is therefore essential in order to support appropriate trust in the model's predictions. In other words, a trustworthy explanation must have feature weights that accurately correspond to the impact on the probability estimates.

The provision of accurate and trustworthy explanations may also be critical for product recommendations and purchase decisions in an area such as digital retail, in order to establish brand loyalty and drive sales. Consider a scenario where a customer is searching for a new pair of run- ning shoes online, and is overwhelmed by the wide range of options. In a situation such as this, a transparent and comprehensive explanation of why a particular pair of shoes has been recommended can instil confidence in the customer making a purchase decision, resulting in a positive shopping experience and possibly encouraging repeat business. Conversely, if the explanation is incomplete or incorrect, it could erode the customer's trust in the retailer and deter future purchases, which underscores the importance of developing trustworthy methods for explanations in the area of digital retail

The main focus of explanation methods was initially to create correct explanations; however, over the last few years, this has changed, with a shift towards the effect on the user, i.e., whether they *trust* (agrees with) or *distrust* (do not agree with) the predictions (Adadi & Berrada, 2018; Chiou & Wong, 2010; Dzindolet et al., 2003; Hoffman et al., 2018; Mueller et al., 2019; Ribeiro et al., 2016). Trust is seen as being intimately connected to the accuracy of the model, i.e., the proportion of accurate predictions. If a model produces a high number of correct predictions, it is trustworthy, and should result in a high level of trusted predictions. Trust is a somewhat blunt instrument, however, as we do not want the user to trust an incorrect prediction. The user needs guidance in order to identify when to trust the model's predictions and when to distrust them, so that the user can gain *appropriate trust* in the model. In this context, appropriate trust refers to the accuracy the user can gain based on the information provided by the explanations. In the example given above involving Petrov, he did not accept or unquestioningly trust the system's outcome of the system; instead, he actively searched for an explanation when the system could not give him one, and based his decision on this additional information. His understanding of the system came from a combination of expert knowledge and appropriate trust, meaning that he neither overly trusted nor mistrusted it[4].

1.2 Problem formulation

As highlighted in the introduction, it is essential to support users of ML models in discerning when to trust predictions, in order to establish a high level of appropriate trust. For these users, explanations that can clarify the rationale behind a prediction are indispensable, especially when the goal is to employ these predictions for decision support. The additional information given through the use of explanation methods aims to enhance transparency and to answer the fundamental questions of why and how the model has arrived at its conclusions. Hence, the aim is to broaden the user's understanding, enabling them to identify more predictions accurately.

[4]It was later suggested that the satellite picked up the sun's reflection from the clouds, and interpreted that as a missile launch.

In addition, there is often a higher or lower degree of *uncertainty* connected with the prediction of an ML model that is not revealed by a single value feature weight; an ML model can generate a prediction with a high estimated probability, which is simultaneously highly uncertain. Neglecting uncertainty in explanations is problematic, since the user may get the impression of a model that is certain of the probability estimate, and be persuaded to accept highly uncertain predictions. Information about the uncertainty of the model is, in fact, seen as highly critical for model transparency and hence explanations (Bhatt et al., 2021). Consequently, the ability of an explanation method to support the user's appropriate trust in the system depends on a well-calibrated model for which the uncertainty can be communicated through the explanation method. Nevertheless, uncertainty information is lacking in the majority of today's *state-of-the-art* (SOTA) explanation methods (Slack et al., 2021).

An evaluation of the quality of an explanation is crucial in order to determine whether an explanation method faithfully represents the behaviour of the model and supports a level of appropriate trust. However, evaluating the quality of an explanation is a complex process that often includes subjective measurements such as curiosity or the user's satisfaction with the explanation (Hoffman et al., 2018). Several taxonomies have been created as guides as to when to use which explanation method, and various criteria have been proposed to evaluate the quality of these methods (Chromik & Schuessler, 2020). Nonetheless, the focus has been on human understanding and satisfaction, which has resulted in evaluations focusing on single methods (Linardatos et al., 2021; Murdoch et al., 2019). The use and choice of evaluation criteria have been left to the individual researcher, meaning that comparative studies of explanation quality pose a significant problem (Carvalho et al., 2019; Zhang & Chen, 2018).

The discussion above reveals several knowledge gaps in the area of explanation methods. Firstly, although uncertainty has been highlighted as a type of transparency and is seen as crucial, very few expla- nation methods have incorporated it. Secondly, despite it being commonly known that most ML models are poorly calibrated, there is an absence of research on the effects of this on explanations; and thirdly, due to the plethora of criteria with a particular focus on subjective ones, it is difficult to know how to measure whether explanations improve with changes. In other words, this thesis focuses on two areas of inquiry: the *development* of trustworthy explanations and the *evaluation* of their explanation efficiency.

1.3 Research Questions

The aim of this thesis is to provide suggestions for assessing and enhancing the quality of explanations for ML models, in order to support high-quality decisions. From the knowledge gaps and areas of inquiry identified above, a research question was formulated, as follows:
RQ: *How can explanation methods for decision support be improved in terms of trustworthiness?*

Since the main research question is wide, and focuses on both improvement and trustworthiness, it is split into two sub-questions, each one of which answers different aspects of the main research question. More specifically, the following two research questions are studied in this thesis, and together address the main research question:

- **RQ1:** *What criteria could be used to evaluate the usability of explanation methods for decision support?*
  This first sub-question involves how to measure the quality or 'usability' of explanation

methods. Here, the word 'usability' has the definition given by the *International Organization for Standardization* (ISO) standard 9241-210:2019, as the extent to which specific users can utilise a product, system, or service to achieve specific goals with effectiveness and satisfaction within a particular context of use.

In this context, the evaluation of the usability of an explanation method is closely connected with the terms defined by the ISO: *verification* (a set of activities that ensure that a system is able to accomplish its intended use, goals, and objectives) and *user experience*(UX; a person's perceptions and responses that result from the usage of a system, product or service.).

This sub-question contributes to answering the main research question with knowledge about how to evaluate whether the proposed explanation method is reliable, i.e., if it produces the desired result of well-calibrated explanations with uncertainty information. An explanation method that produces well-calibrated explanations could be said to be trustworthy; hence, this research question also contributes to answering the main question by identifying criteria that can be used to evaluate the trustworthiness of a suggested explanation method.

- **RQ2:** *How can well-calibrated uncertainty information be included in explanation methods for decision support?*
  The second sub-question focuses on increasing the trustworthiness of a model by including uncertainty information in an explanation method. As discussed in the problem formulation, another approach to improving the trustworthiness of explanation methods might be to calibrate the under- lying model. Hence, calibration is also included in this question. The research question relates both to how calibration affects explanations, and to how uncertainty information can be included in an explanation method..

1.4 Main Contributions

Two contributions were generated in the thesis when answering the sub-questions. A primary contribution in Paper V of the novel explanation method and a secondary contribution in Paper III with a well-formulated definition of a metric for appropriate trust. The development of these two contributions is possible to follow in Figure 1.1:

- **A suggestion of how to asses the user performance in an evaluation**
  The fundamental contribution from Paper I lies in emphasising the crucial point that relying solely on the subjective concept of trust as a measure of explanation quality can be misleading. A comprehensive evaluation of the quality of an explanation method should go beyond trust, and should consider potential implications such as increased misuse resulting from persuasive explanations. By highlighting this aspect, the paper calls for a more nuanced evaluation approach that takes into account additional aspects of the explanation quality beyond user trust alone. This insight is vital for developing trustworthy explanation methods. Paper II corroborates the findings from Paper I, and underscores the insufficiency of assessing explanation methods based solely on subjective trust measurements. Instead, it highlights the need to consider multiple aspects of the quality of an explanation, including aspects related to the model, the explanation itself, and the user's perspective. Moreover, the paper describes various evaluation criteria, their relationships, and the frequent lack of a clear method. If the user can identify correct and incorrect predictions, appropriate trust emerges as a particularly useful criterion for conducting comparative evaluations of explanation methods, although lacking in method. By adopting a more holistic approach to evaluation, this research enhances the current understanding of how to assess and compare the quality of explanation methods for decision support. Paper III synthesises the knowledge from the earlier papers and describes how the level of misuse, disuse, and AT (the level of appropriate trust) can be measured based on commonly used metrics in ML for comparative evaluations.

- **A novel explanation method that generates trustworthy, stable, and robust explanations with uncertainty information**
  Paper IV demonstrates that explanations generated by calibrated models show increased performance. The set of feature weights obtained after calibration is associated with a more accurate confidence value, which enhances the trustworthiness of the explanations. In other words, calibrating the underlying model positively influences the quality of the explanations when they are connected with more accurate feature weights. Paper V corroborates the findings of Paper IV, and presents a novel explanation method called *Calibrated Explanations* (CE). This approach produces explanations using calibrated probabilities from the model, and adds uncertainty information to the presentation. Extensive evaluations of CE demonstrate its ability to generate robust and accurate explanations, at a low computational cost. By offering calibrated and interpretable explanations with uncertainty information, CE paves the way for more transparent and trustworthy AI systems, which can enable users to make better-informed decisions based on a deeper understanding of model predictions. CE is highly effective in terms of running time, making it a practical and valuable tool for real-world applications, thereby further contributing to the advancement of explanation methods in the field of ML.

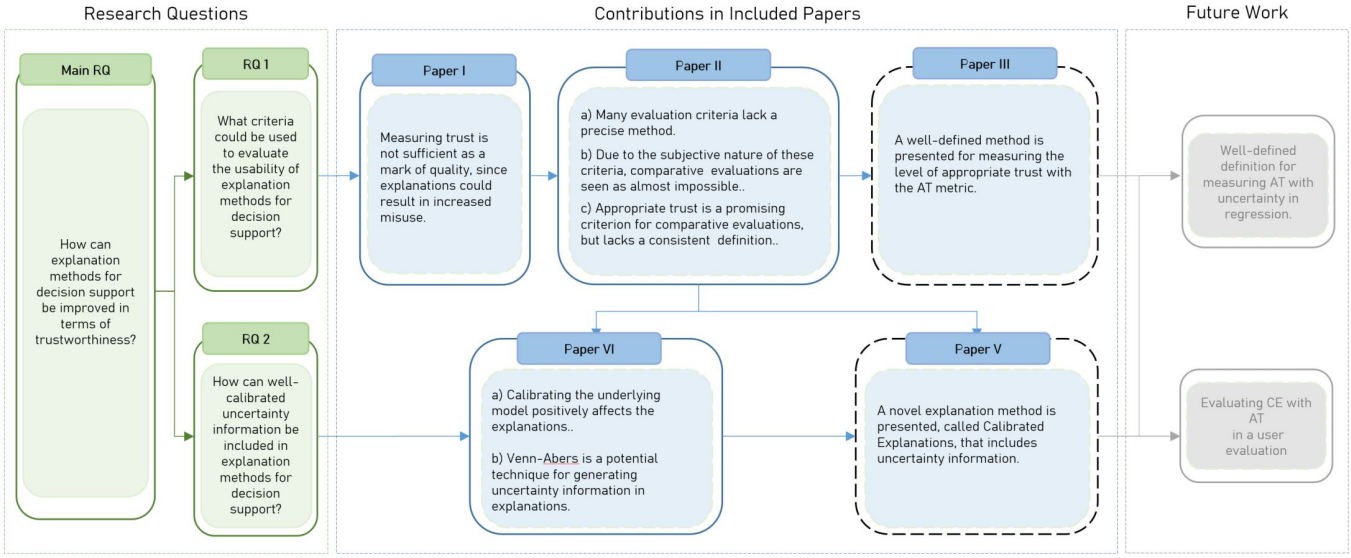

Figure 1.1: Structure of the contributions based on the research questions in the thesis.Primary and secondary contributions to the thesis are highlighted with dotted black frames

1.5 Limitations of the Thesis

Research into ML explanation methods can be divided into two fields: transparency through the use of accurate and inherently interpretable models, and *post hoc* methods for explaining black-box models. Post hoc explanations may be either model-based (i.e., explaining the entire model) or instance-based (i.e., explaining a single instance) (Adadi & Berrada, 2018). The work in this thesis focuses exclusively on post hoc explanation methods. Although the majority of the papers included in this work primarily study instance explanations, the conclusions are not limited to this type. For example, in Paper I, the proposed explanation method is evaluated at both the instance level and the model level. It is worth noting that this thesis explores explanation methods for decision support rather than DSSs *per se*, although explanations may form an integral part of a complex DSS.

Although various calibration methods are considered in the studies, this thesis exclusively addresses applications involving classification with Venn-Abers and the possibilities of this method in terms of producing additional information about prediction uncertainty. Venn-Abers is used for problems of a binary character, and the conclusions drawn in the studies are therefore limited to binary classification problems. However, the novel explanation method presented in Paper V has since been updated with to include functionality for regression, conjunctional rules, and multiclass problems.

The introduction of a novel explanation method commonly involves conducting evaluations with users. Given that the first research question of this thesis focuses on the evaluation of explanations, it would be a logical step to assess the method with users. In Paper II, the evaluation of the explanation quality emphasises the centrality of the user aspect. Paper III defines the AT metric, which could be naturally integrated into a user evaluation. Initially, it was planned to include a user evaluation in the thesis, and the study design was nearly complete; however, unforeseen changes in employment circumstances resulted in a complete restart of the work in this thesis 1.5 years into the project, making it unfeasible to conduct a user evaluation within the limited remaining time.

As a result, the suggested explanation method is primarily verified in terms of its effectiveness in terms of the model and explanations. Nevertheless, the user aspect has still been taken into account. When developing the user interface in Paper V, recommendations from the literature on designing human-machine-friendly interfaces and insights from various well-known explanation methods were considered. Although a complete user evaluation had to be left for future work due to time constraints, the aspect of users and user-friendliness were thoughtfully considered in the design and development of the proposed explanation method and its interface.

1.6 Thesis outline

The structure of the remaining chapters of the thesis is outlined below. The thesis is divided into two parts: Part I presents the cover story, while Part II contains the final versions of the peer-reviewed papers included in the thesis.

PART I: Containing the thesis' cover story:

CHAPTER 2 Theoretical background: This chapter introduces relevant background literature, including several definitions that are used later in the thesis and in the included papers.

CHAPTER 3 Research Design: In this chapter, the research design, data collection process and the analysis carried out in the thesis are described.

CHAPTER 4 Research: This presents a thorough description of how each paper contributes to the thesis. The chapter concludes with a summation of the contributions and implications of the findings

CHAPTER 5 Concluding Remarks: The thesis is concluded with a discussion of the contributions and how they answer the research questions. The chapter ends with suggestions for future work.

PART II The final versions of the peer-reviewed papers included in the thesis.

PAPER I Interpretable Instance-based Text Classification for Social Science Research Projects.

PAPER II A Meta Survey of Quality Evaluation Criteria in Explanation Methods

PAPER III On the Definition of Appropriate Trust: and the Tools that Come with it

PAPER IV Investigating the Impact of Calibration on the Quality of Explanations.

PAPER V Calibrated Explanations with Uncertainty Information and Counterfactuals.

2 Theoretical Background

"*The simplification of anything is always sensational.*"

— G.K. Chesterton

In this chapter, some necessary background information related to the thesis is presented. The reader is introduced to decision theory and decisions under uncertainty, with a short description of predictive models and the fundamentals of model evaluation. An introduction to calibration, Venn-Abers predictors (VAs), and explainable artificial intelligence is also given, which includes explanation methods and their evaluation. The chapter ends with a short summary of earlier work, which provides valuable context for the research.

2.1 Decision Theory

In *Decision Theory* (DT), the decisions of a user or actor are studied. The theory has its roots in the 18^{th} century debates over the value of gambles, with Daniel Bernouilli (Risk & Bernoulli, 1954) giving the earliest precise statement of something akin to the principle of maximising expected utility (Bradley, 2018). DT is interdisciplinary (combining disciplines such as mathematics, psychology, and philosophy) and is closely related to game theory, as described below.

2.1.1 Decisions Under Uncertainty

The focus of DT *under uncertainty* is the study of the logic and the mathematical properties of decision-making under uncertainty, that is, when the consequences of a decision are not entirely predictable, as events in the future may affect the consequences of actions taken now (Bradley, 2018).DT could be seen as a formalisation of common sense (North, 1968), with mathematics providing an unambiguous language in which a decision problem may be represented. This view could refer to two dimensions of a decision: the value, calculated by means of *utility theory*, and the information, calculated by means of *probability theory*. North points out that using this representation, the large and complex problems of systems analysis become conceptually equivalent to simple problems in daily life that we solve by "common sense".

Game theory is a framework for understanding choices in situations involving competing players. It can help players reach optimal decisions when confronted by independent and competing actors in a strategic setting. Furthermore, it can be used in many fields, such as business, finance, and economics, to improve decision-making. A typical form of a "game" that arises in economic and business situations is the prisoner's dilemma (Parmigiani & Inoue, 2009), where an individual

decision-maker always has an incentive to choose in a way that creates a less-than-optimal outcome for the individuals as a group.

A formal theory of decision making should begin with uncertainty (or risk) as its fundamental premise, and should recognise that a precise knowledge of outcomes is an exceptional case, with limitations (Parmigiani & Inoue, 2009).DT serves as a procedural approach that combines all pertinent information to enable the most logical decision possible to be found. Although the aim is to minimise the adverse consequences associated with unfavourable outcomes, it cannot shield the user from every instance of ”bad luck.” Ultimately, the strongest safeguard against unfavourable outcomes lies in making in formed and prudent decisions

2.2 Predictive Modelling

Predictive modelling is a typical application of ML algorithms in DSS and other AI-related tasks. The ML algorithm constructs a model based on patterns of historical data that consist of pairs of two variables, the known input (x) and the output (y) , and this process is known as model *learning* or *training*. To train the model, the data set is divided into two different parts, called the *training set* and the *test set*, as shown in Figure 2.1. The training set is used for training the model, while the test set is used for evaluating the trained model. The error rate on the test set is called the generalisation error (Géron, 2022). If the errors at the testing stage are low, i.e., the model makes few mistakes on the test set, the training is finished and can be used in a production setting. In situations where the number of errors is low at the training stage but high when the model is tested, the model is said to be *overfitted* and be retrained with different algorithm settings or a different set of input features (Géron, 2022).

Careful selection of data is crucial when using ML algorithms, as including irrelevant features can lead to noise and impact the accuracy of the predictions. To ensure the quality of the model, the data set used for training must reflect the situation at hand, and pre-processing the data to fit the modelling requirements is often the most challenging and time-consuming aspect of building a predictive model (Ascarza et al., 2018; Coussement & Van den Poel, 2008; Davoudi et al., 2017; Dechant et al., 2019; Flach, 2012). It is important to note that the quality of the model is limited by the quality of the data used, which highlights the importance of extracting the correct features for the dataset.

For example, when a model is trained to identify diabetes on a dataset that contains a lot of irrelevant and noisy information, the characteristic patterns of diabetes may become obscured and partially concealed. This can significantly reduce the accuracy of the model. In the worst-case scenario, individuals with diabetes might go undetected, while those without the condition could be incorrectly diagnosed as having it; in either situation, the consequences could be life-threatening.

Holdout validation is commonly used in ML (Negnevitsky, 2005) to find the best setup for a model on a specific dataset. This process involves holding out part of the training set to evaluate several candidate models, and then selecting the best one. The new held-out set is called the *validation set*(or sometimes the development set, or ‘dev set’). More specifically, multiple models with various hyperparameters are trained on the reduced training set (i.e., the full training set minus the validation set), and the model that performs best on the validation set is selected. After this holdout validation process, the best model is trained on the full training set (including the validation

set), and this gives the final model. This final model is then evaluated on the test set to get an estimate of the generalisation error. Holdout validation is a widely used method for evaluating ML models, but can lead to imprecise evaluations if the validation set is too small or too large (Géron, 2022). To address this issue, *cross-validation* (Berthold et al., 2020) is employed, which involves using multiple small validation sets (Nti et al., 2021).

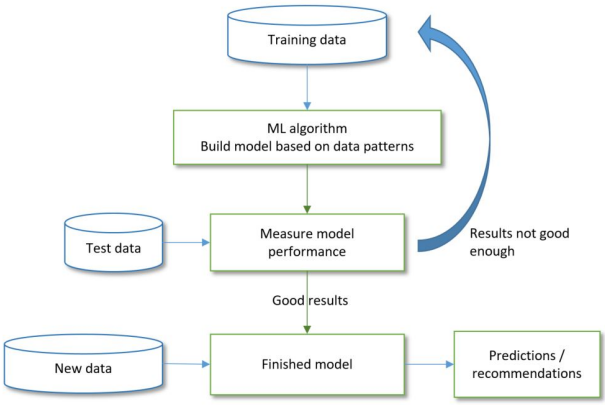

Figure 2.1: Process of building a predictive model.

2.2.1 Classification

Predictive modelling can take various forms, and one of the most commonly used is *classification* (Negnevitsky, 2005). In classification, a set of data, such as documents or images, is divided into subsets that have something in common. These subsets are called *classes* or categories. The *labels* of the classes are symbolic, and are intended to characterise or explain, e.g., the documents in a class. Common examples of the classes used for classification tasks include genre, language, topic, and sentiment (Baeza-Yates, Ribeiro-Neto, et al., 2011; Eklund, 2016; Flach, 2012).

Let us assume that we have a data set $\mathcal{D}$ consisting of a set of n instances $\mathbf{Z} = (z_1, z_2, ..., z_n)$. Each instance is represented by a set of d input features $\mathbf{x} = (x_1, x_2, ..., x_d)$ and a corresponding class y. The task of classification is to learn a function $f(\mathbf{x})$ that maps the input features $\mathbf{x}$ to the corresponding class y (Dunham H, 2003). The function should faithfully predict the class of a new, previously unseen data point (z_{n+1}), based on the relationships between the input features and the classes learned from the training data.

The simplest form of classification is *binary classification*. In this type, there are only two classes, often referred to as the *positive* and *negative* classes. Binary classification is commonly used in applications such as spam e-mail filtering and medical diagnosis, where the classes are spam/non-spam or healthy (non-diabetes)/sick (diabetes), respectively. In these examples, the positive class indicates confirmation of a hypothesis, while the negative class represents a rejection of the hypothesis.

2.2.2 Probabilistic prediction

To conduct a comprehensive analysis, it is important not only to consider the predictions themselves but also to assess the level of confidence in these predictions (Parmigiani & Inoue, 2009). *Probabilistic prediction* is a fundamental aspect of classification, where the goal is not only to predict the class but also to obtain the probability associated with that label, which provides a measure of *confidence* in the prediction. *Probability distributions*serve as the standard representation of uncertainty; they were axiomatically developed by Kolmogorov in the early 1900s (Kolmogorov & Bharucha-Reid, 2018), and can effectively capture various aspects of the uncertainty affecting decision-making processes.

In probability theory, the probability of an event is defined as the proportion of cases in which the event occurs, which represents a scientific measure of chance (Negnevitsky, 2005). Mathematically, this probability is expressed as a numerical index ranging from zero (denoting absolute impossibility) to one (signifying absolute certainty). Most events have a probability index strictly between zero (0) and one (1), indicating the possibility of two outcomes: a favourable outcome or success, and an unfavourable outcome or failure.

The probabilities of the positive and negative classes for a model with a test set Z_T can be determined as follows, where p is the number of times the model predicts the positive class, and n is the number of times the model predicts the negative class:

$$P(positive) = \frac{p}{p+n}$$

$$P(negative) = \frac{n}{p+n}$$

2.2.3 Performance Evaluation of Classification Models

It is essential to consider several aspects when evaluating classification models. One of the most frequently used criteria for classification problems is the accuracy, or conversely, the error rate, in which the fraction of correct or incorrect predictions for the test set is calculated. The accuracy is often given as a number between zero (0) and one (1), and is interpreted as a percentage; for example, an accuracy of 0.65 is interpreted as 65% accuracy. However, for imbalanced datasets, the minority class tends to be less noteworthy for the overall measure (Géron, 2022). In many situations, it is the minority class that is the important class to predict; for instance, customer churn in a company is seldom the majority class in a customer dataset.

There are alternative approaches for evaluating models for binary problems if the dataset is im- balanced. One very common method is the *confusion matrix* (see Figure 2.2). In this matrix, one of the classes is referred to as the positive class, where the choice of this class is problem-specific and depends on the circumstances. From a confusion matrix containing the numbers of correctly predicted positive (TP), correctly predicted negative (TN), incorrectly predicted positive (FP), and incorrectly predicted positive (FP) outcomes, a closer analysis of the model is possible (Géron, 2022). The general idea is to count the number of times the model confuses the predictions, i.e., attaches the wrong class label to an instance. The number of correctly and incorrectly predicted instances are then compared, and several aspects of the model performance can be calculated (Géron, 2022).

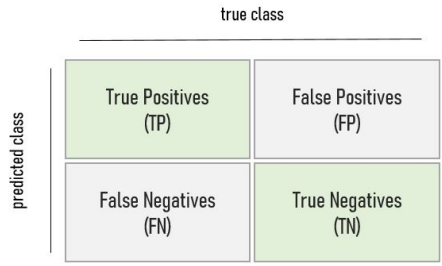

Figure 2.2: Confusion matrix for evaluation of a binary classification model.

Two of the most commonly used metrics based on the confusion matrix are *precision* and *recall*, as described below.

Precision:

This metric represents the proportion of positive predictions that are correct. A high precision means that the model gives a high number of true positives (e.g., where the predictions of patients having diabetes are correct) and a limited number of false positives (e.g., where the predictions of patients having diabetes are incorrect). In other words, the model succeeds at correctly predicting the instances (Géron, 2022). This metric is expressed as the number of true positives divided by the total number of predicted positives:

$$precision = \frac{TP}{TP + FP}$$

Recall:

Recall, also called sensitivity, is typically used together with precision. This metric is used to iden- tify the proportion of positive instances that are correctly identified by the model (e.g., the proportion of patients correctly identified as having diabetes). A high value for recall signals that the model can success- fully identify the positive instances (Géron, 2022):

$$recall = \frac{TP}{TP + FN}$$

F_1 score

Precision and recall can be combined into a single metric called the F_1 score.score. This metric is calculated as the harmonic mean between recall and precision, and gives much more weight to low values, with the result that the classifier only achieves a high F_1 score if both the recall and precision are high (Géron, 2022):

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

Receiver Operating Characteristic Curve and Area Under the Curve:

The *Receiver Operating Characteristic* (ROC) curve is one of the most frequently used tools for evaluating classification models. It is a plot of the *True Positive Rate* (TPR), i.e., the sensitivity, against the *False Positive Rate* (FPR) at various threshold values (Linoff & Berry, 2011). The ROC curve can also be defined as a sensitivity versus 1-specificity plot.

$$FPR = \frac{FP}{TN + FP}$$

The *Area Under the Curve* (AUC)) makes it possible to compare classifiers. A perfect classifier will have an AUC equal to one (1), whereas a purely random classifier will have an AUC equal to 0.5.

2.3 Calibration

In predictive modelling, probability estimates are often *poorly calibrated* (Van Calster et al., 2019) meaning that the predicted probabilities do not match the observed accuracy. This is illustrated in Figure 2.3[1], where the diagonal line indicates a perfectly calibrated model. For a model to be well calibrated, the predicted probabilities must be matched by the observed accuracy. In other words, the predicted probability of an event should correctly correspond to the observed probability for that event; for example, predictions with a probability of 70% should be correct in 70% of cases.Conversely, in a poorly calibrated model, the estimated probabilities will be either be *overconfident* (less than 70% of the predictions are correct) or *underconfident* (more than 70% of the predictions are correct).

Calibration can be expressed as the *probability* P of the class a, given that the predicted probability P_a is, almost surely, indeed P_a (Nouretdinov et al., 2018):

$$\mathbb{P}[Y = a | P_a] \approx P_a$$

In the context of calibration, when dealing with a poorly calibrated model, it is common practice to employ an external calibration method, which utilises a distinct subset of the labelled dataset known as the *calibration set*. One widely used method for calibration is called *Isotonic calibrators* (Zadrozny & Elkan, 2002). In isotonic regression, the predictor variable is linked to the target variable in a monotonically increasing or decreasing form, meaning that as the value of the predictor variable increases, the value of the target variable also consistently increases or decreases, respectively, without any oscillations or irregularities.

Isotonic calibrators are step-wise, non-decreasing regression functions that are typically produced with the pair-adjacent violators algorithm (PAVA) (Tibshirani et al., 2011). This algorithm starts with a set of input probability intervals, where the scores of the calibration instances are used as borders. Adjacent intervals in which the lower interval contains a higher (or equally high) fraction of examples belonging to the positive class are then repeatedly merged. This process continues until no such pair of intervals can be found. Upon termination, the algorithm outputs a function that returns the fraction of positive examples in the calibration set for each interval.

[1] scikit-learn.sourceforge.net/dev/modules/calibration.html

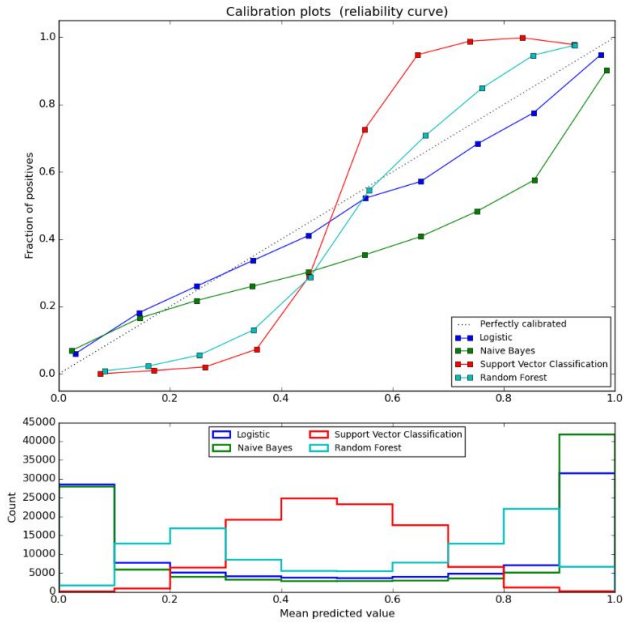

Figure 2.3: Levels of calibration of commonly used models, from Scikit-learn.

2.3.1 Venn predictors

In probabilistic predictions, the validity of a model is determined by how well the predicted probability distributions perform against statistical tests based on observed labels. However, achieving validity in a general sense is not possible (Vovk et al., 2005). *Venn predictors*, on the other hand, are probabilistic predictors that bypass this general impossibility result by restricting statistical tests for validity to calibration, and outputting multiple probability distributions of which one is valid (Vovk et al., 2004). These multi-probabilistic predictions can be converted into probability intervals for each label, where the interval size indicates the confidence in the estimation.

In Venn prediction (Lambrou et al., 2015), a Venn taxonomy is used to divide calibration instances into categories based on the underlying model. Each class receives the relative frequency of calibration instances as the estimated probability for test instances falling into that category. By including the test instance in the calculation, validity is achieved. Since the correct class is unknown for the test instances, every possible class is tried, and the resulting probability distribution is calculated.

2.3.2 Venn-Abers predictors

Choosing an appropriate taxonomy is essential when using Venn predictors. One approach to this is the use of *Venn-Abers predictors* (VAs) (Vovk & Petej, 2012), where the taxonomy is automatically optimised using isotonic regression. VAs are used together with *scoring classifiers*, and since VAs are Venn predictors, they inherit the validity guarantees. The output of a two-class

scoring classifier, when predicting a test instance x_i, is a *prediction score* $s(x_i)$, where a higher value of $s(x_i)$ signals a higher belief in class 1. To obtain the predicted class from a scoring classifier, the score is compared to a fixed threshold value t. The prediction has a value of one (1) if $s(x) > t$, and zero (0) otherwise.

When using VA with scoring classifiers, the threshold is not a fixed value t. Instead, the increasing function g is fitted through isotonic regression (Zadrozny & Elkan, 2001), based on a number of prediction scores for which the true targets are known. The resulting function, $g(s(x))$, can be interpreted as the probability that the class of x is 1, i.e., it is a calibrator. A VA predictor produces a multiprobabilistic prediction for a new test instance x_{n+1} as follows:

1. Define a training set as $Z_{Tr} = \{z_1, \ldots, z_n\}$. Each instance $z_i = (x_i, y_i)$ consists of two parts; an *object* x_i and a *class* y_i.
2. Divide the training set into a proper training set $Z_T = \{z_1, \ldots, z_q\}$ and a calibration set $Z_c = \{z_{q+1}, \ldots, z_l\}$, where $n = q + l$.
3. Use the proper training set Z_T to train a scoring classifier c.
4. Using c, predict $\{x_1, \ldots, x_l, x_{n+1}\}$ to produce the prediction scores s.
5. Let g_0 be a isotonic calibrator for $\{(s(x_1), y_1), \ldots, (s(x_l), y_l), (s(x_{n+1}), 0)\}$, and let g_1 be a isotonic calibrator for $\{(s(x_1), y_1), \ldots, (s(x_l), y_l), (s(x_{n+1}), 1)\}$.
6. Let the probability interval for $y_{n+1} = 1$ be $[g_0(s_0(x_{n+1})), g_1(s_1(x_{n+1}))]$.

2.3.3 Evaluation of Calibration Performance

Expected Calibration Error

When calculating the *Expected Calibration Error* (ECE), the probability estimates for the positive class are divided into M equally sized bins before calculating a weighted average of the absolute differences between the fraction of positives (*fop*) predictions and the mean of the probabilities for the positive class (*mopp*):

$$\mathrm{ECE} = \sum_{i=1}^{M} \frac{\#B_i}{n} |\mathrm{fop}(B_i) - \mathrm{mopp}(B_i)|$$

where n is the size of the data set and $\#B_i$ represents the number of instances in bin i.

Log loss

In ML, the log loss (also known as the logarithmic or cross-entropy loss) is a loss function used to measure the performance of a classification model. It is commonly used when the model's output is a probability score between zero (0) and one (1) for each class. 'Log' is a logarithm, e.g., the binary or natural logarithm, and p is the estimate for the predicted label. In Python, to avoid infinite results, the log loss function clips the probabilities to ensure that they never are exactly zero (0) or one (1) (Géron, 2022). Simply put, log loss penalises overconfident models that make incorrect predictions.

The *log loss* is calculated as follows:

$$\lambda_{log} = \begin{cases} -\log p & \text{if correct} \\ -\log(1-p) & \text{if incorrect} \end{cases}$$

2.4 Explainable Artificial Intelligence

Although predictive modelling algorithms are widely recognised for their high accuracy, many of the most precise ML algorithms produce models that are lacking in transparency and comprehensibility (Gunning & Aha, 2019), rendering them opaque (see Figure 2.4). Although these models excel at uncovering novel patterns in data, their output is limited to predictions accompanied by estimated probability values. The limited nature of the output from these opaque models poses challenges when they are used for decision support, as they cannot provide answers to fundamental questions such as why the model has come to its conclusions. In the example given above of a doctor deciding on a diagnosis for a patient, the output from such a model might be that the patient has diabetes with a confidence of 79% . However, the model cannot explain how it has come to this conclusion.

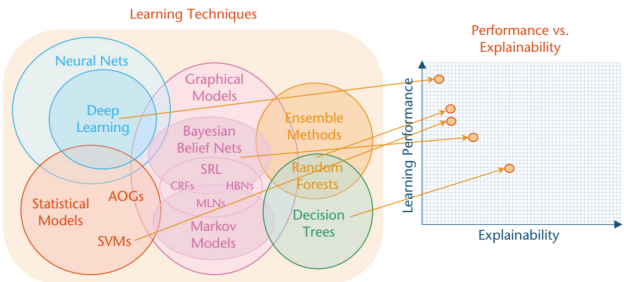

Figure 2.4: Learning Performance Versus Explainability for Several Categories of Learning Techniques, from (Gunning & Aha, 2019).

In response to this limitation, the field of *explainable artificial intelligence* (XAI) was developed, in which the aim is to enhance the transparency, trustworthiness, and user satisfaction of AI models, thus enabling high-quality decision-making (Zhang & Chen, 2018). A key objective of XAI is to empower users to identify and correct erroneous predictions, such as an incorrect medical diagnosis of diabetes, for example (Gunning & Aha, 2019).

A well-known criterion in XAI is *appropriate trust*, i.e., whether the user can detect correct and erroneous predictions when using explanation methods (Yang et al., 2020). When the user has a low level of appropriate trust in the model, two unwanted situations can occur: *misuse* or *disuse* (see Figure 2.5):

Misuse: A user who is over-reliant on the model may fail to identify errors or biases in the model (Adya & Phillips-Wren, 2020; Sheridan et al., 2002). This term can also be defined as a higher trust in the model than is appropriate, meaning that doubts about the model's trustworthiness are neglected. An evaluation of the user's ability to detect correct and erroneous

predictions would show a higher number of trusted predictions, including erroneous predictions. This is also known as *overtrust* (Yang et al., 2020). In our example involving the diagnosis of diabetes in patients, the doctor would accept the prediction from the model in this case, and would suppress any doubts.

Disuse: This is often seen as more problematic, since the user has a lower reliance on the model than is appropriate, given the model's actual performance (Alvarado-Valencia & Barrero, 2014; Buçinca et al., 2020). In the worst-case scenario, disuse could lead to neglect or under-utilisation of the DSS. This is also known as *undertrust* (Yang et al., 2020). In our example involving diagnosing diabetes in patients, the doctor would doubt the model in this case, and would not use it or follow any recommendations, although it is not unusual that the predictions of such a model are more accurate than those of humans.

Misuse and disuse are two situations that can be prevented by explanations, and can be detected when measuring the user's level of appropriate trust in the model.

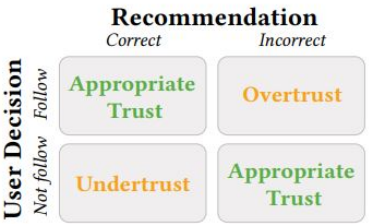

Figure 2.5: Identifying the level of appropriate trust in classification, from (Yang et al., 2020).

2.4.1 Explainability vs. Interpretability

A distinction can be made between models that are interpretable by design and those that can be explained through the use of an external XAI technique. This difference is also regarded as the difference between interpretable models and model interpretability (Arrieta et al., 2020; Gunning & Aha, 2019). One of the main reasons that the evaluation of explanation methods is seen as complicated is the vagueness of the terms *explainability* and *interpretability* (Carvalho et al., 2019; Linardatos et al., 2021; Lipton, 2018). In this thesis, the following definitions are employed from (Arrieta et al., 2020):

- Interpretability: A passive property of the model (a model is said to be interpretable) that is similar to comprehensibility and transparency.
- Explainability: An active property (an explanation method is said to *explain to* the user).

2.4.2 Constructing Understandable Explanations

Two different levels of explanations are considered in predictive modelling: *Model* (or global) explanation, and *instance* (or local) explanation (Martens & Foster, 2014; Nguyen, 2018). The model explanation provides a greater understanding of the entire classification model and its performance. A model explanation provides a greater understanding of the entire classification model and its

performance, whereas an instance explanation provides a greater understanding of the model's predictions of a specific instance, e.g., the prediction of a single patient (Martens & Foster, 2014; Robnik-Šikonja & Kononenko, 2008).The goal of explanation methods is to make the rationale of the model transparent, independently of the accuracy of the prediction. However, it is reasonable to assume that the quality, or at least the *usefulness*, of the explanations increases with higher accuracy (Robnik-Šikonja & Kononenko, 2008).

To generate an explanation for an instance, an *instance-based explanation method* is employed. Various approaches can be utilised when constructing these explanations, and the most straightforward strategy is to use an *interpretable model*such as a decision tree. For an interpretable model, the explanation for each instance is derived directly from the model. In a decision tree, for example, the path to the leaf predicting the instance serves as an explanation of the model's prediction. One weakness of interpretable models is that they are generally less accurate than more complex, opaque models such as ensembles, SVM:s or different types of neural network.

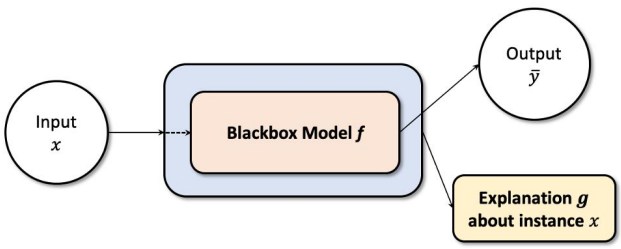

Figure 2.6: Constructing explanations from an opaque model (from Das and Rad).

The construction of explanations for a predictive model is typically done in a separate step after training the underlying predictive model (see Figure 2.6). An explanation method that is based on predictions is called a *post hoc* (Arrieta et al., 2020).These methods use the predictions of the model and extract relationships between feature values and predictions. The relationships are presented via an *explanation interface* as explanations, and may include pixels in a picture, or words in a piece of text. Although there are many names for this type of explanation method, they all emphasise the importance of the features as a basis for the explanation of the prediction, e.g., *feature attribution*, *feature relevance*, or *feature importance* methods (Holzinger et al., 2020).

The importance of a feature is measured as the increase in the prediction error of the model after permutation of the value of the feature, which breaks the relationship between the feature and the true outcome (Molnar, 2022). The concept is straightforward (Breiman, 2001; Fisher et al., 2019): a feature is important if the model error increases, since the prediction is dependent on the feature, whereas a feature is unimportant if the model error does not change, since the model ignores the feature when making the prediction.

Local Interpretable Model-agnostic Explanation (LIME) is a well-known method for post hoc explanations (Ribeiro et al., 2016). LIME is model agnostic in the sense that the method treats every model as a black box, and does not make any assumptions about its inner workings. To explain the prediction of an instance z using LIME an interpretable model is trained based on

perturbed instances in the vicinity of z in the underlying model h. To generate these instances, the method first randomly selects a number of k non-zero features x of z to be used in the explanation. The instances are weighted based on their proximity to z. The number of chosen and perturbed instances n are then fed into the underlying model to obtain the labels, y . These labels are then used to train the interpretable model, which is used to create the explanations e (Rahnama & Boström, 2019).

There are some notable disadvantages to this type of method. Since the meaningfulness of the explanations is dependent on the size of the instance space, a space that is too small makes the explanation hard to generalise. At the same time, an instance space that is too large may decrease the fidelity (if the explanation accurately captures the behaviour of the underlying model) and incur heavy computational costs (van der Waa et al., 2020).

More formally, the local surrogate model with interpretability constraint can be described in the following way (Molnar, 2022):

$$\text{explanation}(z) = \arg\min_{g \in G} L(h, g, \pi_z) + \Omega(g),$$

where g interpretable model (ridge regression is used as the default in the Python implementation) that minimises the loss L (e.g., the mean squared error), representing how close it is to h (the underlying model). G is the family of possible explanations (e.g., the possible ridge regression models), and π_z defines the proximity around instance z to be taken into account. The model complexity, $\Omega(g)$, is not optimised by LIME but is controlled through parameters (e.g., the number of features to include in the explanation).

SHapley-Additive exPlanations (SHAP) (Lundberg & Lee, 2017) is another well-known post-hoc explanation method. This method uses Shapley values from game theory, which have the fundamental property that they always sum to the difference between the game outcome when all players are present and the game outcome when no players are present. Since SHAP is an additive feature relevance method, it attributes an effect ϕ_x to each feature, which sums the effects of all feature weights approximating the output $h(z)$ of the original model h, which is the main difference between the feature weights used in SHAP and LIME. The value assigned to each feature is the average marginal contribution of the feature among all possible coalitions, creating an explanation model that is a linear function of binary variables.

Explanations may have a *factual* or *counterfactual* character. Factual explanations play a central role in XAI, and aim to provide insights into why a prediction is made by offering reasons based on the data and highlighting the features influencing the outcome. These explanations shed light on the relationships between the input data and the resulting prediction, addressing the 'why', 'when', and 'how' aspects of the prediction process. Counterfactual explanations serve a different purpose and focus on answering contrasting questions, providing the reasoning behind why an alternative prediction or outcome was not realised. By exploring the factors that could have led to a different class, counterfactual explanations contribute to a deeper understanding of the decision-making process (Mueller et al., 2019).

2.4.3 Essential Characteristics and Evaluation of Explanations

Creating high-quality explanations in XAI is a multidisciplinary effort in which knowledge is drawn from *Human Computer Interaction* (HCI) and ML. The quality of an explanation method

depends on its goals, which may vary. For example, an assessment of the user's level of appropriate trust differs from evaluating the fidelity of an explanation to the model. Nonetheless, certain specific characteristics are universally desirable for post hoc explanation methods, and are crucial to consider when evaluating explanations, such as:

1. **The interpretability of the explanations**, i.e., whether they provide a qualitative understanding of the relationship between the input variables and the (Ribeiro et al., 2016).The interpretability of an explanation is closely connected to subjective criteria such as *trust*, *curiosity*, or *explanation satisfaction* and is difficult to measure in such a way as to give comparative results.
   The user's mental model is often treated as having a high level of importance in terms of the quality of an explanation. However, it is not a criterion in itself, and could be considered a container of other subjective criteria. The user's mental model could be changed positively or negatively during use of the model, i.e., this could cause the criteria of the user's mental model to change (Kulesza et al., 2013). According to Hoffman et al. could the level of appropriate trust be said to be the output from the mental model.

2. **The model-agnostic nature of the explanation model**, meaning that it treats the model as a black box (Ribeiro et al., 2016; Slack et al., 2021).

3. **The reliability**, which is another critical characteristic of explanation methods (Adadi & Berrada, 2018; Agarwal et al., 2022; Carvalho et al., 2019; Hoffman et al., 2018; Moradi & Samwald, 2020; Mueller et al., 2019; Wang et al., 2019).
   There are various different aspects of reliability, including:

   - *Faithfulness* or *fidelity* to the underlying model, meaning that it accurately captures the behaviour of the underlying model (Carvalho et al., 2019; Gallant, 1993; Ribeiro et al., 2016; Sestito & Dillon, 1991). The quality of an explanation depends not only on the user's reactions but also on how faithfully the explanation can mirror the underlying inner rationale of the model (Doshi-Velez & Kim, 2017; Gilpin et al., 2018; Murdoch et al., 2019).

   - *Stability*, i.e., whether the same instance produces identical explanations across multiple runs. An explanation method is considered stable if it generates identical explanations from numerous runs (Slack et al., 2021).

   - *Robustness* refers to the ability of an explanation method to produce consistent results even when an instance undergoes small perturbations (Dimanov et al., 2020).It is considered robust if it produces consistent explanations when feature values are varied within the boundaries of the explanation's rules (Agarwal et al., 2022; Alvarez-Melis & Jaakkola, 2018).

In summary, the essential characteristics of an explanation method are that it should be *model agnostic*, *faithful* to the underlying model, *stable* over multiple runs, and *robust* against small perturbations. In addition, the underlying model must be trustworthy and interpretable to the user, in order to ensure that the explanations are of high quality and can generate a high level of appropriate user trust (Dimanov et al., 2020; Hoffman et al., 2018).

2.5 Related Work

Explanations play a vital role in the acceptance of DSSs driven by ML models, as they allow users to understand the reasoning behind recommendations and actions and enable them to make informed decisions (Dzindolet et al., 2003; Lacave & Diez, 2002, 2004; Pavlidis et al., 2012). However, explanations can also introduce new errors, negatively influencing the quality of judgements (Hopkins et al., 2016, 2019; Skitka et al., 1999; Weisberg et al., 2015; Weisberg et al., 2008).

In recent years, instance-based explanations, particularly LIME (Ribeiro et al., 2016), SHAP (Lundberg & Lee, 2017), ANCHOR (Guidotti et al., 2019), or MAPLE (Plumb et al., 2018), have gained in popularity, as they provide insights into the processes of machine learning models and make them more transparent and interpretable (Martens & Foster, 2014; Ribeiro et al., 2016).However, evaluating the quality of explanations in machine learning is challenging, due to the vagueness of terms such as 'trust' and 'expectation' (Carvalho et al., 2019; Lipton, 2018). Nevertheless, the focus of the evaluation is often on subjective criteria such as trust and their impact on decision-making (Adadi & Berrada, 2018; Chiou & Wong, 2010; Dzindolet et al., 2003; Lacave & Diez, 2002, 2004; Pavlidis et al., 2012; Ribeiro et al., 2016). This focus on subjective criteria means that comparative evaluations of explanations are seen as almost impossible, creating a lack of rigour in the field, and more research is needed in this area (Slack et al., 2021; Zhou et al., 2021). A high level of appropriate trust is frequently seen within XAI as an accepted criterion for measuring the quality of explanations (Adadi & Berrada, 2018; Arrieta et al., 2020; Carvalho et al., 2019; Chromik & Schuessler, 2020; Das & Rad, 2020; Doshi-Velez & Kim, 2017; Gunning & Aha, 2019; Hoff & Bashir, 2015; Hoffman et al., 2018; Mohseni et al., 2018; Wang et al., 2019; Zhang & Chen, 2018). An explanation method may create a high level of trust but a lower level of appropriate trust, and a good mental model is seen as a requirement for developing appropriate trust in the model (Hoff & Bashir, 2015). Although the meaning of appropriate trust is straightforward (i.e., the extent to which a user can identify correct and incorrect predictions), it lacks a consistent definition of method in the literature. With authors measuring it using likert scales and subjective evaluation of explanation effectiveness (Holzinger et al., 2020), to the alignment between the perceived and actual performance of the system and related to the user's abilities to recognise when the system is correct and when it is incorrect (Jacovi et al., 2021; Yang et al., 2020). Some authors have emphasised subjective aspects of trust; for example, Hoffman et al. focus on the user's evaluation of the experience of the system, whereas others, such as (Yang et al., 2020), are interested in the degree to which the user can distinguish erroneous predictions.

A deterministic environment implies the availability of perfect information, where each decision has a single outcome. However, the real world is uncertain, and precise probabilities are not sufficient; the level and type of uncertainty can affect the capabilities of AI-based systems to make intelligent decisions. Badings et al. point out that there are likelihoods of information loss in the training data, which can lead to sub-optimal outcomes as performance deterioration in predictive models. Uncertainty models remove this assumption by incorporating uncertainty as sets of probabilities, and it has been argued the most informative form of output for a predictive problem (Nouretdinov et al., 2018).

One crucial factor that can impact the quality of explanations is the calibration of the underlying ML model. Poorly calibrated models may mislead explanation methods, as the probability estimates they provide may be incorrect (Adadi & Berrada, 2018). Calibration techniques such as

Venn-Abers, can provide probability intervals that convey a level of certainty and confidence in prediction estimates (Vovk et al., 2005); however, there is limited research on the consequences in terms of explanation quality when using poorly calibrated models and the consequences when calibrating the underlying model (Slack et al., 2021). Trustworthy models rely on correct calibration and transparency, and explanations depend on human understanding and effective communication (Adadi & Berrada, 2018; Bhatt et al., 2021; Chromik & Schuessler, 2020; Mehrtash et al., 2020; Ribeiro et al., 2016). Understanding and effective communication are crucial for successful explanations in these domains (Gilpin et al., 2018).

3 Research Design

"*The definition of insanity is doing the same thing over and over again and expecting a different result.*"

Unknown origin

This chapter describes the method adopted to address the research questions of this thesis. The decisions that were taken to determine the research method, the data collection process, and the analysis in this thesis took into consideration three aspects, as described by Creswell (2014):(i) the researcher's philosophical assumptions towards the study; (ii) the procedures of inquiry; and (iii) the choice of research method. The methods used for data collection methods are then discussed, and the chapter ends with a presentation of how several analysis methods are used in the thesis. For a more detailed description of each paper, see Chapter 4.

3.1 Philosophical assumptions

In this thesis, a positivist research philosophy is applied, which is guided by a realist and objectivist ontology; in other words, this research is based on the belief that an objective reality can be studied through quantitative experiments and logical reasoning (Recker, 2012). The goal is to achieve generalisation, validation, and replicability, which are considered to be the core pillars of science.

However, as highlighted in the introduction, explanations for ML models are produced in a context where the goal is to allow the user to identify when to trust the model's predictions. In RQ1, the focus is on evaluating the explanation methods and trying to understand whether and how the user's reaction indicates the quality of the explanations, among other aspects. Consequently, questions such as how and why a user reacts in a certain way (as considered in Paper I) become crucial. These questions require a study design that allows us to find the meaning of actions in a given context, rather than validation or generalisation (McCusker & Gunaydin, 2015; Queirós et al., 2017).

In summary, although the thesis is founded on a positivist philosophy, the scope of the thesis includes an investigation of the patterns behind the user's reactions.

3.2 Procedures for Inquiry

Research is primarily guided by three strategies: *induction*, *deduction*, or *abduction* (Recker, 2012). These strategies dictate how the researcher employs data to arrive at conclusions, for example by forming a theory based on the data, or by using data to test and validate an existing theory.

These strategies are typically associated with either *qualitative* research methods, which seek to understand complex realities and the meanings of actions in a specific context, or *quantitative* research methods which aim to obtain accurate and reliable measurements for statistical analysis (Queirós et al., 2017) (see Table 3.1).

Table 3.1: Key questions in the methodology of research (adapted from da Silva et al. (2018), and Wilson (2014))

	Qualitative Approach	Quantitative Approach	Pragmatic Approach
Theory and data connection	Induction	Deduction	Abduction (Deduction/ Induction)
Ontology	Subjectivity	Objectivity	Objective or subjective
Inference from the data	Contextual	Generalisation	Contextual/ Generalisation

An inductive strategy is suitable for qualitative methods, while a deductive strategy corresponds to quantitative methods (Oates, 2005). In some cases, a pragmatic methodological approach is adopted, which combines both inductive and deductive approaches to theory and data, as in Paper I. It can be seen from Table 3.1, that this approach includes both both the subjectivity of qualitative studies and the objectivity of quantitative studies.

Qualitative methods are important when trying to comprehend how users perceive explanations, as they capture the nuances of human understanding (McCusker & Gunaydin, 2015). This aspect is crucial in addressing RQ1. For RQ2, which focuses on developing well-calibrated explanations with uncertainty information, experiments and logical reasoning play a key role, thus making quantitative methods more suitable (Queirós et al., 2017; Recker, 2012).

Consequently, this thesis adopts a pragmatic approach that integrates both quantitative and qualitative methods. The combined use of these research methods can offer a comprehensive understanding of the subject under investigation (Wilson, 2014), i.e., how to develop trustworthy explanations.

The advantages and drawbacks associated with employing qualitative and quantitative research methods must be carefully weighed. Challenges related to sample size can arise when using qualitative methods to delve into attitudes and to conduct in-depth examinations of behaviour within smaller user groups, as in the expert user evaluation in Paper I. In these situations, researchers must give thoughtful consideration to their sampling selection and address privacy concerns. Nevertheless, qualitative methods offer valuable benefits, such as flexibility, promoting meaningful discussions, and the potential to unearth novel patterns and attitudes (McCusker & Gunaydin, 2015).

In contrast, quantitative methods employ larger sample sizes and anonymous data, facilitating faster and more straightforward data collection. These methods often yield reliable results, unlike qualitative approaches, which focus on uncovering new patterns. However, in quantitative research, it is impossible to delve deeper into the respondents' answers, and the outcomes are constrained by a predetermined set of questions, as seen in surveys, for instance.

3.3 Method

As mentioned above, both qualitative and quantitative methods were required to answer the research questions of this thesis. In one of the papers, both qualitative and quantitative methods are used. There are few methods that support this type of research, and it is often recommended that the researcher choose either a qualitative or quantitative research method, since the analysis can be time-consuming (McCusker & Gunaydin, 2015; Recker, 2012).As this thesis is based on a compilation of papers, their individual contributions must be considered when choosing a methodology; however, it is also essential to consider the contribution of the thesis as a whole.

Mixed methods research (MMR) is a type of research method that combines qualitative and quantitative methods for data collection and analysis, in either a sequential or concurrent fashion (Ågerfalk, 2013; Recker, 2012). This method provides a pragmatic, pluralistic approach to research that combines the strengths of both qualitative and quantitative methods, in order to answer research questions that could not be answered in other ways (Creamer, 2017).

Table 3.2: Five main rationales of MMR

Purpose	Rationale
Triangulation	Different methods and designs are used to study the same phenomenon, with the intention to identify convergence and corroboration (Ågerfalk, 2013; Greene et al., 1989).
Complementary	Qualitative and quantitative methods are used to measure different facets of a phenomenon, and the researcher seeks to elaborate, enhance, or clarify the results from one method based on results from another method (Ågerfalk, 2013; Greene et al., 1989; Recker, 2012).
Development	The findings from one method are used to inform a research design, involving another method (Ågerfalk, 2013; Recker, 2012).
Initiation	The researcher attempts to discover paradoxes and contradictions that lead to a reframing of the research questions (Ågerfalk, 2013; Greene et al., 1989; Recker, 2012).
Expansion	The depth and breadth of the research are expanded through the use of different methods for different components of inquiry (Ågerfalk, 2013; Recker, 2012).

There are five different rationales for conducting MMR (see Table 3.2). One of these is *development*, where the findings from one method are used to inform a research design involving another method (Ågerfalk, 2013; Recker, 2012). The two sub-questions of the thesis revolve around how to improve explanation methods with well-calibrated explanations conveying uncertainty information: the findings from the first research question (RQ1) are used to provide the second research question (RQ2) with information about how to measure the quality of explanation methods, resulting in a development rationale between the contributions answering the two research questions.

The primary method used to answer RQ1 is based on an qualitative approach, with an *expansion* rationale in which the depth and breadth of knowledge about evaluating the quality of an explanation are expanded. However, within Paper I, a *triangulation* rationale was adopted, as different methods and designs were used to study the users' reactions and the effectiveness of the WITE explanation method.

To summarise: To answer RQ1, an *expansion* rationale was primarily adopted. The *development* rationalewas adopted to guide the design of the evaluations in RQ2 based on the findings from RQ1. Controlled experiments were used to answer RQ2,using a single (quantitative) method approach.

MMR can be used in different contexts and in different research designs. This type of approach can be divided into four categories, which address different dimensions of the research:

- **Timing**: This refers to the time ordering of the qualitative and quantitative approaches (Ågerfalk, 2013; Recker, 2012)
- **Weighing** or **dominance**: The qualitative or quantitative approach may be the dominant one in the research, or they may contribute evenly (Ågerfalk, 2013; Recker, 2012).
- **Mixing**: The degree to which the research is mixed forms a continuum, from a single method to a fully mixed-method (Recker, 2012).
- **Placing**: This refers to the decision on where mixing takes place: in the research questions, in the methods of data collection, in the research methods, during the data analysis, or at the data interpretation stage (Recker, 2012).

As described above, there is a distinguishable *timing* order between the two research questions of this thesis, as qualitative approaches are primarily used to answer RQ1 and quantitative ones for RQ2. Although it appears that the qualitative and quantitative approaches are applied evenly, the contribution of this thesis that is likely to be the most easily applied in practice is the novel explanation method in Paper V, and this can be seen as the primary contribution, although the AT metric proposed in Paper III could offer vital help to researchers evaluating the level of appropriate trust for a user. The quantitative approach could therefore be said to be the *dominant* one.

The research approach to the thesis as a whole is *mixed*; however, research questions RQ1 and RQ2 are answered using mono-methods, meaning that the placing of mixing lies primarily in the research questions. It should also be noted that there is no clear separation between the two research questions (quantitative/qualitative); for example, the user evaluations in Paper I were based on a quantitative approach for the online evaluation, and a controlled experiment was carried out.

It is essential to differentiate between *multi-method research* and MMR. In multi-method research, two or more research methods are incorporated in one study, whereas in MMR, the researcher studies the same phenomenon with different methods. In short, all mixed-method studies use multi-methods, but not all multi-method studies use mixed methods.

The research methodology known as *design science research* (DSR) could have been a viable choice for this study, as it allows for the incorporation of both quantitative and qualitative methods, and is thus aligned with the paradigmatic stance of MMR. DSR is commonly employed in the development of DSSs, and requires the production of an artefact as a significant outcome (Hevner et al., 2004). Notably, almost every contribution in this thesis results in an artefact, with the explanation method being the primary contribution.

DSR emphasises conducting evaluations preferably within real-world company settings. The contributions in this thesis could be essential for companies using DSS and one of the papers was produced in close collaboration with a company. However, the other papers lack such a relationship. Additionally, DSR follows a cyclic methodology, involving iterative development and

evaluation processes (Vom Brocke et al., 2020). As no cyclic behaviour was applied in the thesis, the DSR approach was a less suitable option, leading to the selection of MMR instead.

3.4 Datasets

Throughout the course of the studies in the thesis, a diverse range of datasets were used to address the research questions posed in the papers, reflecting the research methodology of each paper. For a more comprehensive overview of the construction and usage of the datasets, see Chapter 4.

3.4.1 State-of-the-Art-Datasets

In Papers IV and V, a total of 25 SOTA datasets were employed to evaluate the proposed solutions. These datasets are well-known within the research community, and were pre-processed to ensure their quality and relevance to the study.

3.4.2 Data from a Real-World Dataset

The study in Paper I used data collected from a real-world research project. However, before using the dataset in the context of ML, pre-processing was required to make it suitable for analysis. The dataset consisted of a limited subset of 178 instances specifically pertaining to politics or social issues, extracted from a larger corpus of approximately 5,000 manually classified news articles.

3.4.3 Data Collected through User Evaluations

The data from the user evaluations were collected using two distinct approaches: an online user evaluation, and a questionnaire with open questions. Each type of evaluation approached the research questions from a different perspective: the online user evaluations sought to achieve reliability by involving a larger number of non-expert respondents, while the questionnaires with open items aimed to identify patterns, albeit with a smaller pool of expert respondents.

Online User Evaluation: The dataset considered in Paper I was based on closed questions and involved non-expert users. The evaluations were conducted anonymously, ensuring that the researcher could not trace back individual responses.

Questionnaire with Open-Ended Questions, Expert Users: In Paper I, a questionnaire with open questions was administered to six expert users to explore potential patterns and uncover new phenomena.

3.5 Learning Algorithms

All the learning algorithms described in this thesis were implemented in the Python programming language (Python Software Foundation, 2021). In the experiments, the, `RandomForestClassifier` (RF) from *scikit-learn* and `xGBoost` from *xgboost* were used as learning learning algorithms. Both techniques were used with default settings, except that the `objective` parameter in xGBoost was set to `binary:logistic`. The algorithms were chosen to generate models representative of different characteristics: RF generates rather well-calibrated models whereas

XGBoost is known for generating poorly calibrated models. By choosing representative examples for different characteristics, it is possible to study how calibration affects explanations under different circumstances, for example. For a comprehensive description of the learning algorithms considered in each paper, see Chapter 4.

3.6 Analysis of Research

The thesis used a mixed method, and consequently different types of analysis were involved, such as evaluation of controlled experiments, quantitative analysis of online user evaluations, and qualitative analysis of coded texts, as follows:

Paper I contributed to answering RQ1 with an emphasis on the evaluation of explanations in, among others, a qualitative study. The answers from the expert users were coded and organised based on the type of prediction (correct, incorrect, or uncertain). The aim of the analysis was to find new patterns in the answers that could highlight the reasons behind the increase in user trust.

Paper II contributed to answering RQ1. Building on a textual analysis of larger numbers of papers (between 739 and 2,384), the analysis aimed to demonstrate validity and generalisation. The surveys included in the study were read and coded based on the definition and method for each evaluation criteria. The analysis focused on finding existing criteria, their definitions and their relationships.

Paper III contributed to answering RQ1 through a qualitative analysis of definitions of appropriate trust in the literature. A method based on the AT metric (measuring the level of appropriate trust by users) was derived based on these definitions and similarities to methods of evaluating the performance of a binary classifier.

Paper IV contributed to answering RQ2, and emphasised the development of trustworthy explanations through calibration of the underlying model. The analysis of the experiments focused on the accuracy and calibration quality of the models, predictions, and explanations.

Paper V contributed to answering RQ2, and emphasised the development of trustworthy explanations with uncertainty information. The analysis of the experiments did not focus solely on the calibration quality, and different metrics were also used to explore the changes in quality of the explanations. Due to the notable differences between the results for the explanation methods when evaluating the computational costs, no statistical methods were needed to validate the findings.

3.6.1 Evaluation Criteria

When evaluating the performance of the underlying models in the experiments and the generated explanations, commonly known metrics were used, although not all of these metrics were applied in all papers. For a comprehensive description of each criterion, see Chapter 2.

Paper I: The model considered in this paper was trained on a deliberately balanced dataset, to ensure equal representation of the two classes. The performance of the model was measured based on the *Accuracy*. Additionally, *Recall*, *Precision*, and the F_1 score were used to assess the model's performance due to its relatively low level of accuracy (74%). The level of

accuracy was as expected, considering the challenging nature of the task of predicting two closely related classes (politics and social issues) of texts from newspapers.

Paper IV: This paper included two distinct experiments. The first evaluated the performance of the underlying model, with and without calibration, using metrics such as the *accuracy*, *AUC*, *ECE*, and *log loss*, while the second focused on assessing the impact of calibration, including *ECE*, on the explanations provided by the LIME explanation method.

Paper V: This paper consisted of an experiment where the explanations from the presented explanation method were evaluated based on the execution time, stability, and robustness. Standard 10×10-fold stratified cross-validation was used, and all results were averaged over 100 folds.

3.6.2 Qualitative Analysis

Two types of qualitative analysis were conducted in the thesis: coding of the texts of answers to the questionnaire with open questions in Paper I, and an analysis of the articles in the meta survey in Paper II and the additional articles included in Paper III.

Paper I: In this paper, the responses to the open questions were analysed to identify the reasons behind the respondents' trust or lack of trust in the predictions. The results obtained from this process were categorised based on the type of prediction, distinguishing between instances that were correctly or incorrectly predicted.

Paper II: In the final sample, the surveys underwent a thorough analysis to identify evaluation criteria, including their definitions, usage, and quality threshold values. Each evaluation criterion was initially documented individually, and then grouped based on its definition and aspect of explanation quality.

Paper III: In this paper, an analysis of various papers was conducted to extract the definitions and usage related to appropriate trust, misuse, and disuse. From these findings, definitions were formulated. A study was also conducted to examine the similarities between the definitions and evaluation metrics for binary classifiers such as the *precision*, *recall*, and F_1 score. Based on this analysis, metrics for evaluating user performance evaluations were derived.

3.7 Ethics

Ethical considerations are essential, especially when conducting research involving human participants. Careful attention was therefore paid to addressing ethical aspects throughout the research process[1]. Two user evaluations were conducted as described in Paper I.

- In the first evaluation, non-expert respondents used an online questionnaire service that guaranteed complete anonymity.
- Respondents (expert users) provided informed consent before participating in the second user evaluation. Their responses were anonymised to ensure that they could not be traced back to individual participants, although no sensitive information was collected or handled.

[1] https://www.vr.se/english/analysis/reports/our-reports/2017-08-31-good-research-practice.html

Ethical questions also arise when using open datasets, especially in terms of ensuring the fairness of data usage. The majority of the benchmark datasets used in this thesis have been widely employed in research without any known ethical concerns. In one of the datasets, Diabetes, all patients were females, aged at least 21 and of Pima Indian heritage[2],which could affect the level of bias if generalised to other groups. However, the dataset was only used for experimental analyses. Notably, the only dataset created within this work and used in the human evaluation consisted of articles that had already been subject to ethical consideration as part of the research project in which the dataset was originally collected.

[2]https://www.kaggle.com/datasets/mathchi/diabetes-data-set

4 Research

"*Now the trouble about trying to make yourself stupider than you really are is that you very often succeed.*"

- C. S. Lewis

This chapter comprehensively explores the contributions made by each peer-reviewed article included in the thesis, in terms of addressing the research questions. It commences with a structured overview of the organisation of the papers, and subsequently presents a thorough analysis of their individual contributions to answering the research questions.

4.1 Organisation of the Included Papers

This thesis represents the culmination of extensive research that resulted in the publication of five papers (as shown in Figure 1.1).The interconnections and contributions of these papers, which are summarised in Table 4.1, provide an overview of the questions that they answer. The table reveals a con- vergence of various types of knowledge in the thesis(Wildemuth, 2009), stemming from the fact that each paper built upon the findings of its predecessors, except Papers III and Paper IV. Notably, Papers IV and V had an explicit dependency. At the same time, although not explicitly discussed, Paper IV relied on the insights presented in Paper II pertaining to the evaluation of explanation quality.

Although the primary author of the papers was the main contributor to these studies, varying degrees of cooperation with other co-authors was also involved (see Table 4.2). In the first two papers, the role of the co-authors was limited to proofreading and discussion. The studies described in Papers IV and V were extensive, and lasted well over a year; there were numerous discussions with the co-authors during these studies, and the experiments were redesigned several times, resulting in several different versions of the papers. An additional researcher, Maria Riverio, also initially participated at an early stage of the study described in Paper IV. Although the first author initially carried out the coding for an early version of the explanation method in Paper V, this required the cooperation of two people. Rudy Matela also contributed to the launch of the package of *Calibrated Explanations* on PiPy[1]

[1] See https://pypi.org/project/calibrated-explanations/

Table 4.1: Mapping of the contributions to the research questions.

Paper	Contribution	RQ1	RQ2	Type of Knowledge
Paper I	Measuring trust is not sufficient as a mark of quality, since explanations could result in increased misuse.	X		Descriptive
Paper II	a) Many evaluation criteria lack a precise method.	X		Descriptive
Paper II	b) Due to the subjective nature of these criteria, comparative evaluations are seen as almost impossible.	X		Descriptive
Paper II	c) Appropriate trust is a promising criterion for comparative evaluations, but lacks a consistent definition.	X		Exploratory
Paper III	A well-defined method is presented for measuring the level of appropriate trust with the AT metric.	X		Exploratory
Paper IV	a) Calibrating the underlying model positively affects the explanations.		X	Prescriptive
Paper IV	b) Venn-Abers is a potential technique for generating uncertainty information in explanations.		X	Prescriptive
Paper V.	A novel explanation method is presented, called Calibrated Explanations, that includes uncertainty information.		X	Prescriptive

4.2 Papers Related to Research Question I

The first three papers examined the distinguishing characteristics of explanations, and proposed methods of measuring their quality. These contributions address the first sub-question: *"What criteria could be used to evaluate the usability of explanation methods for decision support?"* The research question was approached from various perspectives in these three papers.

4.2.1 Paper I

trust, it not only reinforces the user's decisions but also mitigates the risk of user disengagement, or disuse. The pioneering investigations in Paper I, delved, among other things, into the intricate factors contributing to increased user trust through a qualitative study involving expert users.

Contributions

The findings suggested that the increase in user acceptance of incorrect predictions could be attributed to the persuasive nature of the explanations provided. Users expressed initial doubt about erroneous predictions, but when presented with "logical" explanations, they tended to accept the system's reasoning as potentially correct and subsequently trusted it, which increased misuse (overtrust).

Table 4.2: Contributions of the authors to the papers included in the thesis.

Paper	First author	Co-authors	Author contribution
Paper I	Helena Löfström (HL)	Tuwe Löfström (TL), Ulf Johansson (UJ)	The study was designed and completed by HL. TL contributed with discussions and proof reading, UJ contributed with discussions.
Paper II	Helena Löfström	Karl Hammar (KH), Ulf Johansson	The study was designed and completed by HL. The contributions from KH and UJ were limited to proof reading.
Paper III	Helena Löfström	-	The study was designed and completed by HL as a single author.
Paper IV	Helena Löfström	Tuwe Löfström, Ulf Johansson, Cecilia Sönströd (CS)	The study was designed and completed by HL, with support from the contributing authors in the following order: TL, UJ, and CS. TL contributed to the discussions, coding, experimental setup, and writing. UJ and CS contributed with discussions, writing, and proof reading.
Paper V	Helena Löfström	Tuwe Löfström, Ulf Johansson, Cecilia Sönströd	The study was designed and completed by HL, with support from the contributing authors in the following order: TL, UJ, and CS. TL contributed to the discussions, coding, experimental setup, and writing. UJ contributed with ideation, discussions, and proof reading. CS contributed with discussions.

The explanation methods that are typically used today offer only probability-based estimates of the model's confidence in the predictions, along with single-valued feature weights. As described in regard to the contributions from Paper I, these types of explanations present a seemingly conclusive picture to the user, and often ignore the uncertainty. Hence, the communication of explanations to the user must be more sensitive to the possibility of persuasion.

Explanation methods aim to support decisions by enhancing the transparency and understanding of the underlying model's predictions. However, comparative evaluations are challenging due to the numerous subjective evaluation criteria that are closely tied to the respondent's or user's perspective. In Paper I, trust was measured when users were presented with a traditional explanation method for text classification. The results indicated a seemingly high level of trust among the users, but a concerning pattern emerged when this trust was compared to the accuracy of the trusted predictions. The users' trust increased irrespective of whether the trusted predictions were correct or incorrect, and measurements of the level of trust alone did not reflect the quality of the decisions based on the explanations. This pattern was further reinforced when users explained why they trusted the predictions, with some attributing their trust to the explanations provided. Hence, trust, as a metric of quality, could not be used to evaluate whether the explanation method effectively assisted users in identifying when to trust predictions.

Method

The dataset considered in this study focused on politics and social issues, and was a subset extracted from a larger corpus of manually classified news articles from a research project in Sweden. Human coders underwent thorough training and testing to ensure the highest level of consistency in their document classification.

Initially, the original data corpus consisted of around 5,000 documents, but only about 1,000 were digitally accessible PDF files (see Figure 4.1). These articles covered 15 different topics or classes, each with varying frequency. The two most frequent classes were chosen for this study, namely politics and social issues. After careful sampling, the final dataset comprised 178 documents, evenly distributed between the two classes, with 89 articles per class. Human coders assigned each document to one of these two classes (politics or social issues), and their judgements were used as the ground truth for this study.

It is worth noting that some documents proved challenging to classify due to their complex nature, as they covered multiple topics. In addition, the documents had a wide variation of sizes, ranging from as few as 27 words to as many as 1,446 *words*. All articles were written in Swedish, meaning that all descriptive words were in the Swedish language.

At the pre-processing stage of KNIME (Berthold et al., 2008), underwent a digital transformation into text files. Various filtering steps were applied, including the removal of numbers, punctuation characters, stop words (from a predefined list of Swedish stop words), and one-letter words. The documents were then transformed into indicator features using the bag-of-words approach. Furthermore, indicator variables representing words occurring in less than three documents were filtered out. It is important to note that stemming was not used, since the words needed to be presented in the questionnaire in their original form, and this could have had a minor impact on the predictive performance. For comprehensive information on the data pre-processing step, Appendix A.1.

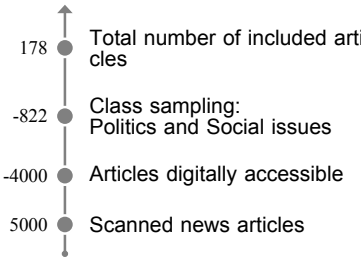

Figure 4.1: Development of the dataset in Paper I (to be read from bottom to top).

Learning Algorithms The leave-one-out evaluation method was employed, in which one underlying model was evaluated for each document and the remaining documents served as the training set. A threshold value of 0.01 was applied to the feature weights, since the explanation method presented in the paper (see Appendix A.1) used all the words in the documents. The top 10 words with the highest weights from each class were presented to the users to explain the prediction for the document. Some predictions had very few words indicating a specific class, due to the low weights associated with certain words, especially for predictions with high probability estimates.

User Evaluations In the first evaluation, non-expert users participated in an online questionnaire to evaluate the overall performance of the words identified as most important by the explanation method. The aim of this evaluation was to capture the main characteristics of the classes. The user group was recruited from a group of adults on Facebook with a diverse range of societal

backgrounds, opinions, and interests. Of the 525 group members, 24 individuals completed the survey. The 14 globally highest weighted words were selected from each class, and participants were asked to determine which class they believed best suited each word. To avoid bias, the words were presented randomly to each participant.

A questionnaire with open questions was used to find patterns among the responses from expert users and to explore new phenomena. Three of the respondents were senior researchers working in media science, who used manual classification in their research projects. They were familiar with the documents, which gave them a unique competence in terms of evaluating the results, as they were familiar with the manual text classification process and the evaluated data. They were chosen because their expert knowledge of the data and the code instructions might have made them more inclined to question erroneous predictions, and be more in line with the manual classification. The other three experts in the group were non-researchers working in a media house in Sweden; they had different positions in the company, but were all familiar with similar texts to those in the study. Furthermore, through their work, this group had a unique knowledge of the challenges associated with categorising articles and experience of using automated content clustering for news articles. They were chosen because they handled newspaper articles on a daily basis and had a deep knowledge of the problems involved in fuzzy newspaper categories.

Based on the classification results, six documents were selected to represent: (i) correctly predicted documents with a probability estimate (PE) above 0.8; (ii) incorrectly predicted documents with a PE above 0.8; and (iii) uncertainly predicted documents with a PE of around 0.5, from each class. The respondents were presented with the article text (in which the important words had been highlighted), the prediction, the probability estimates, and the list of top explanatory words, sorted based on their weights. The respondents were asked to reflect on the prediction concerning the text if the essential words had been selected, and to state whether they agreed with the prediction. The manual classification results, which were used as the ground truth and targets, were not revealed. For further details of the user evaluation, see Appendix A.1.

4.2.2 Paper II

Ensuring that explanation methods play a constructive role in decision support without inadvertently promoting misuse is paramount. In the meta-survey described in this paper, criteria for assessing the quality of explanations were organised into a high-level model. The model was used to systematically classify the identified evaluation criteria across three pivotal aspects of the explanation quality— namely, the *model*, *explanation*, and *user*, as illustrated in Figure 4.2.

Contributions

Comparative evaluations of explanation methods are challenging, and the development of a general computational benchmark for all possible explanation methods is seen as unlikely (e.g., by Zhou et al. (2021) due to the subjective nature of the characteristics. However, it is generally accepted(Hoffman et al., 2018) that the user's mental model affects the level of appropriate trust in and reliance on the system. If the mental model is considered as a container of the subjective criteria, the outcome of the mental model could be measured based on the changes in the criterion of appropriate trust, with and without explanations. Although appropriate trust does not explicitly reflect how the user experiences the explanations, it demonstrates if it fulfils one of the most crucial goals of explanation methods: *allowing the user to detect correct and erroneous predictions*.

By measuring the outcome of the mental model through appropriate trust, an objective metric is derived for these subjective measurements, thereby creating the possibility of comparative evaluations of explanation methods. In other words, the user experience can be verified, as highlighted in Chapter 1.3.

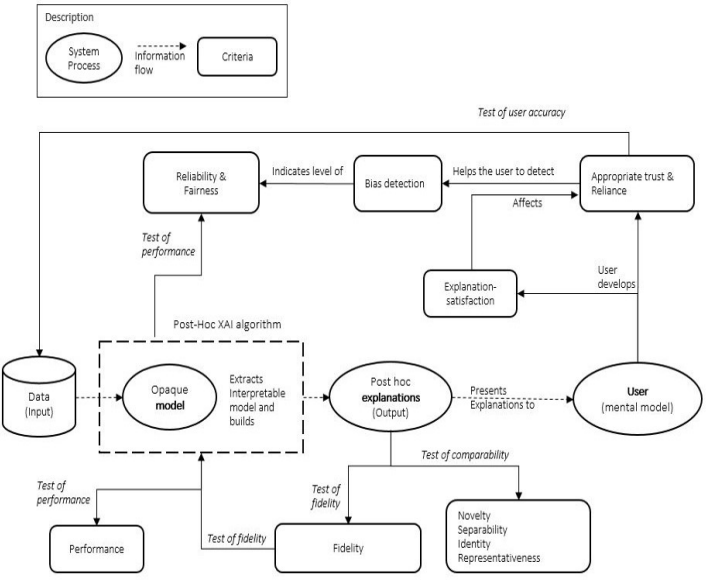

Figure 4.2: model over explanation quality, from Paper II.

It is essential to highlight that subjective measurements are not considered unnecessary or unimportant, but are crucial in terms of quality. When evaluating a single explanation method, it could be vital to follow the changes in subjective criteria. However, the criterion of appropriate trust is recommended if the researcher intends to obtain comparable results from the evaluation.

It has been highlighted in earlier research (see, e.g., (Hoffman et al., 2018))that the criterion of appropriate trust is closely connected to the user's level of appropriate reliance. In Paper II, the criterion of appropriate trust is also directly connected with the user's performance. However, several challenges are associated with this metric: for example, the name is frequently confused with the subjective measurement of trust, which creates confusion in terms of its definition and usage.

Method

The study in Paper II was semi-structured, and the selection of literature was conducted in three phases. In the first phase, a broad literature search was conducted based on the terms *'explanation method'* and *'trust'*. The term 'explanation method' was chosen since it focuses on the evaluation of explanation methods. 'Trust'was also chosen as a search term since it is intimately connected to the measurement of the quality of explanation methods in the literature.

The initial search resulted in 250 articles (138 + 112). This low number of articles may have been a result of the limiting the search to the subject areas of *Computer Science*, *Social Science*, *Business and Marketing*, and *Decision Science*. Only articles written in English were considered. The abstracts were read, and articles with a focus that coincided with the study were chosen. A total of 214 articles were found not to relate to evaluation or some form of measurement, and were excluded, leaving 36 articles (see Figure 4.3). Of these 36, only surveys were chosen, and 27 additional articles were excluded. Phase 1 resulted in nine surveys covering the chosen research areas. The list of references in the results was studied, and three additional surveys were found, resulting in a total of 12 surveys at the end of Phase 2. Articles that cited the surveys were also studied. Since at least one of the surveys was cited over 1,000 times, this was a narrow search. Three more surveys were found, and Phase 3 resulted in 15 unique surveys, covering a total of between 739 and 2,750 articles spanning from maximum to minimum overlap in cited articles, thus indicating the breadth of this study.

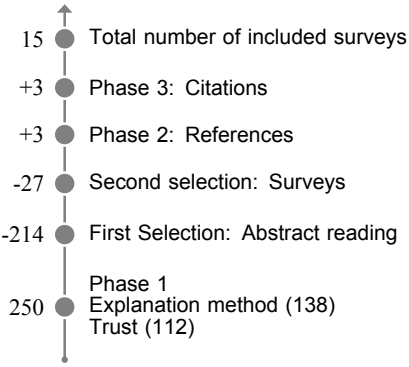

Figure 4.3: Phases in the article selection process (to be read from bottom to top).

4.2.3 Paper III

To effectively support the users' decisions, explanations must strike a balance that avoids the pitfalls of both disuse (undertrust) and misuse (overtrust). In Paper III, the AT metric was developed to measure the level of appropriate trust by drawing upon the parallels between model evaluation metrics for ML binary classification tasks and the established definitions of appropriate trust, misuse, and disuse in the literature. For an in-depth exploration of the derivation of this metric, see Appendix A.3.

Contributions

From the existing definitions in the literature and the combined confusion matrix in Figure 4.4, *misuse* and *disuse* were defined. Misuse was defined as when the user showed a low rate of true trust *(Tt)* in the total amount of trusted predictions (Tt+Ft), and disuse as when the user showed a high rate of false mistrust (Fm) in relation to the total amount of correct predictions (Tt+Fm). From these definitions, and their strong similarities to precision and recall, methods for calculating them were derived. Misuse was defined as the result of a user evaluation indicating low user precision, and disuse as when the user shows low user recall, as follows:

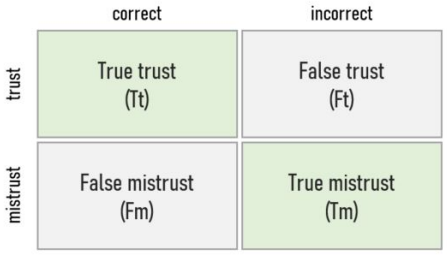

Figure 4.4: Combined confusion matrix for model performance and trust from user evaluations (inspired by (Yang et al., 2020))

User precision (misuse) U_{pr}**:** $U_{pr} = \frac{Tt}{Tt+Ft}$

User recall (disuse) U_{rc}**:** $U_{rc} = \frac{Tt}{Tt+Fm}$

Finally, *appropriate trust* was defined based on the literature as the performance of the user, in terms of the degree to which the user is capable of identifying when to trust or distrust the predictions. The definition given in the literature also states that a high level of appropriate trust is a result of a low level of misuse and disuse. In view of the strong similarities to the measurement of the harmonic mean between precision and recall (i.e., a binary classifier F_1 score), a method for the AT metric (for measuring the level of appropriate trust) could be derived. A high AT score was defined as when the user evaluation results in high user precision, U_{pr}, and high user recall, Upr:

AT (User F_1 **score)** U_{at}**:** $U_{at} = 2 * \frac{U_{pr} * U_{rc}}{U_{pr}+U_{rc}}$

Using commonly accepted methods to evaluate the user's reliance, a researcher or any other individual who wants to evaluate the quality of explanations through, for example, changes in AT, has tools for creating comparable results.

AT metric: Setup

Although this paper primarily introduces a well-defined metric for comparative evaluations, it is also necessary to consider recommendations for group sizes and sampling methods. As discussed in prior work by (Ribeiro et al., 2016), the number of users needed to evaluate ML models and their explanations may be influenced by factors such as the complexity of the model and the diversity of the user population. When considering evaluations in a context where the decision makers are known and well-defined, user sampling is relatively intuitive; however, when considering a more homogeneous and less well-defined user group, user sampling becomes more complex for the researcher, since the diversity of the intended user groups must be considered and may affect the evaluation outcome. It may, for example, be challenging to find representative users for an evaluation.

Ribeiro et al. (2016)proposed that evaluations involving a minimum of 20 users might be necessary to attain statistically significant results. However, it is worth noting that the International Organization for Standardization (ISO), which offers usability testing guidelines, has suggested a different approach based on a minimum of only five users for each user group or persona. Their guidelines are rooted in the observation that most usability issues can be detected through testing with five

users, as tests with additional users above this threshold may primarily uncover minor issues or reconfirm previously identified problems. Similarly, in a study by (Nielsen & Landauer, 1993), it was suggested that testing with five users could reveal 85% of usability problems, whereas expanding the user group to 15 could reveal 95% of these issues. The authors argued that this marginal increase in problem detection may not justify the added cost and time associated with testing more users. In view of this, it is advisable to apply the AT with a user group of at least 15 ndividuals, provided that the resources in terms of cost and time permit a sample of this size.

Since the AT metric reflects the user accuracy, it emphasises the number of tested instances rather than the number of users involved in the evaluation. Consequently, it is essential to assess various scenarios that are applicable to the specific situation and context under investigation. For example, in the context of Paper I, the pool of expert users was relatively small. Nevertheless, the users encountered a variety of critical scenarios, each of which was represented by two distinct instances. When conducting user evaluations with closed questions, the inclusion of a higher number of instances per user is naturally more feasible; conversely, administering questionnaires with open questions and a higher number of queries can be time-consuming, and might discourage users from completing the questionnaire. In accordance with recommendations from the ISO and the guidelines provided by Nielsen and Landauer (1993), it is suggested that a minimum of five instances should be tested, which is manageable for questionnaires with open questions. However, the inclusion of up to 15 instances for evaluations featuring closed questions may be advisable for statistical significance, according to the recommendations of Ribeiro et al. (2016).

It is crucial to emphasise that although the AT metric assesses the level of the users' appropriate trust in an evaluation, measuring the initial level of AT is equally essential. To effectively evaluate whether explanations can enhance appropriate trust, the measurement of AT without access to any explanatory information is imperative. In addition, it is possible to use AT without an initial evaluation, for example to measure the level of the users' disuse.

Method

The approach used in this paper was to map methods for evaluating binary classification models in ML to the definitions of appropriate trust, misuse, and disuse in the literature. This mapping was achieved by first identifying definitions of these terms, primarily from the literature reviewed in the meta-survey in Paper II. Next, a combined confusion matrix for the model performance and the definitions of appropriate trust, misuse, and disuse was developed (see Figure 4.4).). From this confusion matrix and methods for binary classification evaluation, definitions of AT, misuse, and disuse were derived and finally exemplified based on the results from the expert user evaluation in Paper I. For a comprehensive description of this method, see the Appendix A.3.

4.2.4 Summary of Contributions

In these papers the first research question has been answered through three distinct steps:

1. The first paper raised the point that relying solely on subjective criteria such as *trust* is inadequate to represent the quality of explanations.
2. In the second paper, an extensive literature survey was conducted to identify and understand the existing evaluation criteria and their interrelationships in the literature. The concept of encompassing various aspects of explanation quality and existing evaluation criteria for each

aspect was introduced. The criterion of *appropriate trust* was emphasised as the outcome of the subjective criteria which makes it possible to give them objective values (although with a vaguely defined method in the literature), thus making comparative evaluations possible.

3. In the third paper, a method for calculating the *AT* metric, which measures the level of *appropriate trust*, as emphasised in the second paper, was precisely defined. The paper established a valuable toolset that can allow researchers to perform comparative evaluations of explanation methods with users.

4.3 Papers Related to Research Question II

The last two papers in the thesis explore and suggest improvements to explanation methods with the addition of calibration techniques, to address the second sub-question: *"How can well-calibrated uncertainty information be included in explanation methods for decision support?"* These papers specifically examine two aspects of ensuring the trustworthiness of explanations: firstly, they emphasise the importance of accurately reflecting a trustworthy underlying model, and secondly, they highlight the significance of effectively conveying uncertainty information to users in order to minimise any potential persuasive elements, as indicated in Paper I.

4.3.1 Paper IV

In the context of explainability, if probability estimates are poorly calibrated and then used to construct explanation methods, the explanations conveyed to the user may differ significantly from reality. In such circumstances, it will be challenging for the user to develop appropriate trust in the model, since the information provided by both the underlying model and the explanation method will be deficient. This issue is important for explanation methods such as LIME, which utilise probability estimates to quantify the relevance of features when creating their explanation reports. A poorly calibrated probability estimate from the underlying model will likely result in misleading feature weights.

Contributions

The findings of Paper IV showed how calibration affects not only the underlying model but also the explanations. A well-calibrated model represents actual predictions in data and, consequently, reality. The findings showed that the calibration of explanations was affected by the degree of calibration of the underlying model. In a poorly calibrated model, the explanations tend to have a higher log loss than the model, meaning that when the explanations contain notable errors, these tend to be larger than those of the underlying model. The conclusion was that a better calibrated model resulted in explanations that were more aligned with the underlying model. Furthermore, since a higher level of calibration means a more accurate representation of reality, calibrated explanations are a more accurate representation of reality. Venn-Abers was also highlighted as appropriate for explanation methods, since it outputs valid probability intervals, thereby providing opportunities to enrich the explanations with additional insights.

In the last two years, uncertainty (Bhatt et al., 2021; Slack et al., 2021) has been highlighted as a way of communication between the user and the model. In Paper IV, the Venn-Abers calibration technique is shown to produce well-calibrated probability estimates and feature values for explanations, meaning that it has the potential to increase the trustworthiness of explanations. The

paper also highlights how this technique can be used to estimate the uncertainty of the underlying model. In other words, this technique can increase trustworthiness by communicating well-calibrated probability estimates and providing transparency when communicating the uncertainty of the underlying model to the user.

Method

In Papers IV and V, the datasets used to evaluate the suggested solutions were pre- processed. These were SOTA datasets that are well-known within the research community.

Table 4.3: Descriptions of the datasets used in Papers IV and V

Name	#inst.	#attrib.	Name	#inst.	#attrib.
colic	357	60	kc2	369	22
creditA	690	43	kc3	325	40
diabetes	768	9	liver	341	7
german	955	28	pc1req	104	9
haberman	283	4	pc4	1343	38
heartC	302	23	sonar	208	61
heartH	293	21	spect	218	23
Hearts	270	14	spectf	267	45
hepati	155	20	transfusion	502	5
iono	350	34	tictactoe	958	28
je4042	270	9	vote	517	17
je4243	363	9	wbc	463	10
kc1	1192	22			

All 25 datasets contained binary classification data and are publicly available from either the UCI repository (Dua & Graff, 2022) or the PROMISE Software Engineering Repository (Sayyad Shirabad & Menzies, 2005). The characteristics of the datasets are presented in Table 4.3, where *#inst.* is the number of instances, and *#attrib.* is the number of input attributes.

Learning Algorithms: LIME was chosen as the explanation method, since it is well known and it is evident from the rules which feature values are not covered, making evaluation straight- forward. Both `RandomForestClassifier` and `xGBoost` were used with default settings, except that the `objective` parameter in xGBoost was set to `binary:logistic`.

Paper IV contributes to the second sub-question by emphasising the improvement in explanations. The analysis of the experiments in the paper consequently focuses on the accuracy and calibration quality of the predictions and models.

Two experiments were conducted and their results were evaluated: first, the calibration of the underlying models and calibration techniques were evaluated, and the impact of calibration on the explanations from LIME was then evaluated.

Experiment 1. Evaluating the calibration quality of the underlying models: *Accuracy* and *AUC* were used to measure the predictive performance. In order to investigate the quality of the calibration, the values of the *log losses* and the *ECE* were reported. Standard $10x10$-fold stratified cross-validation was used, thus all results were averaged over 100 folds.

Experiment 2. Evaluation of the Impact of Calibration on LIME: When evaluating LIME, the same setup as in Experiment 1 was used, with one small but important difference. In order

to measure the effect of calibration alone, exactly the same underlying model was used in each case (with or without calibration, using Platt scaling or Venn-Abers) when extracting explanations using LIME. This means that the underlying model, used as both an uncalibrated and a calibrated model, was trained using $2/3$ of the training set (and calibrated using $1/3$). Furthermore, 10-fold stratified cross-validation was used, instead of $10x10$-fold cross-validation.

The underlying assumption in Experiment 2 was that the feature weight w_f of feature f iwas indicative of how much (and in what direction) it affected the probability estimate p of the underlying model. Thus, if the substitute probability estimate p'_f for an instance without feature f (or with other feature values) could be estimated, it follows that $p \approx p'_f + w_f$. The impact of calibration could be evaluated by measuring the expected calibration error or log loss on $p'_f + w_f$. LIME provides feature weights for all features, and only p'_f remains to be calculated.

The explanations produced by LIME include a rule in the form of a condition, e.g., "`Feature 4 > 2`" or "`1 < Feature 1` $\leq$ 2". To find the values used in the rules, LIME discretises feature values into a number of bins B_f for each feature f. By using the mean of each bin b_f as a substitute feature value for feature f in the instance, resulting in a new substitute instance x_{b_f}, substitute probability estimates can be calculated by applying the evaluated model h on each substitute instance: $p'_{b_f} = h(x_{b_f})$. In Paper IV, the model h is the uncalibrated or calibrated model. The substitute probability estimate p'_f for feature f is the weighted average over all bins except the bin covered by the rule:

$$p'_f = \sum_{b_f \in B_f \setminus b_r} \frac{p'_{b_f}}{|b_f|},$$

where b_r is the bin covered by the LIME rule and $|b_f|$ is the number of instances falling into bin b_f. This corresponds to how the rules are created in LIME

To calculate the expected calibration error over all features, the following expression was used, where $|F|$ is the number of features:

$$p' + w = \frac{1}{|F|} \sum_{f=1\ldots|F|} p'_f + w_f$$

4.3.2 Paper V

In this paper, the advancements made in Paper IV were extended and applied to develop a novel explanation method in which Venn-Abers intervals were used to quantify prediction uncertainty and feature uncertainty, where a wider interval indicates lower certainty and a narrower interval indicates higher certainty. The method was designed to produce stable rules. Reliability follows from the inherent calibration, meaning that both predictions and explanations become better representations of the actual underlying distribution. As explained in Section 2.1.1, it needs to be recognised that a precise knowledge of the outcomes is an exceptional case, and any decision-making theory should begin with uncertainty as its fundamental premise. The explanations generated in the method presented in Paper V built on this premise and included uncertainty information to support the user in making informed and prudent decisions.

Contributions

A novel explanation method called *Calibrated Explanations* (CE) was presented. This is model-agnostic and can be applied to different models. The explanation method can be downloaded as a Python package and installed from PiPy, or Github.

Description of Calibrated Explanations
The steps described below at a high level were taken to generate a local explanation in CE for a test object x.
After training a scoring classifier on the proper training set Z_q, Venn-Abers was used to calibrate the underlying model to get the probability interval $[p_l, p_h]$ and the calibrated probability estimate p. The Venn-Abers technique was then used to estimate the probability intervals and calibrated probability estimates for slightly perturbed versions of the test object x by systematically changing one feature at a time with the existing values in Z_q.

The uncertainty interval for a specific feature f was generated by first calculating the average for the alternative probability estimate(s) and probability interval(s) achieved on the perturbed instance(s) related to f. The difference between the average values and p was then calculated, and an upper and lower bound on the uncertainty interval were obtained. For a detailed description of the method, see the Appendix A.5.

Since CE is built on Venn-Abers, it generates well-calibrated probability estimates. One of the strengths of this method is straightforward communication, which can provide comprehensive insights into how each feature affects the probability estimate, either with (*Uncertainty plot*) or without (*Regular plot*) uncertainty information, as follows:

- **Regular**: The most basic plot in CE, seen in Figure 4.5, shows how each feature affects the probability estimate depending on whether the feature value is above or below a specific threshold. Binary rules are used in CE to enhance the clarity of communication for the user. Unlike interval-based rules, binary rules make it easier for users to interpret how the prediction changes with feature values. Each feature has a corresponding conditional rule and weight, allowing for an unambiguous understanding of the meaning of the rule and the weight it contributes to the prediction.

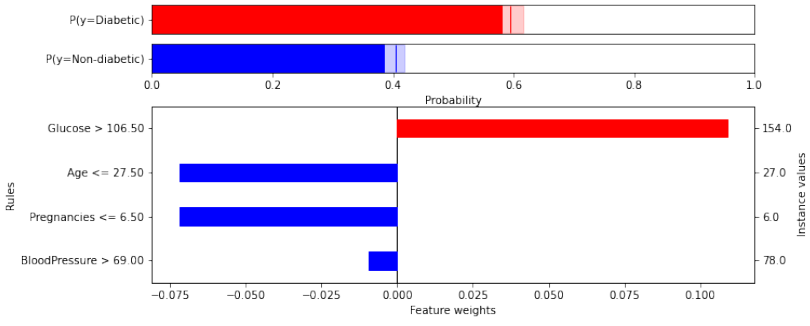

Figure 4.5: Regular CE plot showing information on how each feature affects the probability estimate when it is above or below a specific value.

Regular CE explanations share similarities with LIME explanations in several respects. Firstly, each feature in regular CE has a corresponding conditional rule that specifies the condition under which the weight is defined (written out at the left side of the plot). Secondly, every rule in Regular CE has an associated feature weight, calculated such that subtracting the probability achieved if the rule condition is violated yields the probability of the instances covered by that rule. This calculation ensures that the meaning of the rule and the weight it produces can be clearly and unambiguously understood.

- **Uncertainty**: In the uncertainty CE plot, intervals are added to the regular bar plot to provide information about the uncertainty associated with each feature's contribution to the prediction. Narrow intervals indicate low uncertainty, whereas wider intervals indicate higher uncertainty. The feature interval also shows how the feature affects the prediction estimate. The uncertainty plot in Figure 4.6 presents the identical prediction as in Figure 4.5.

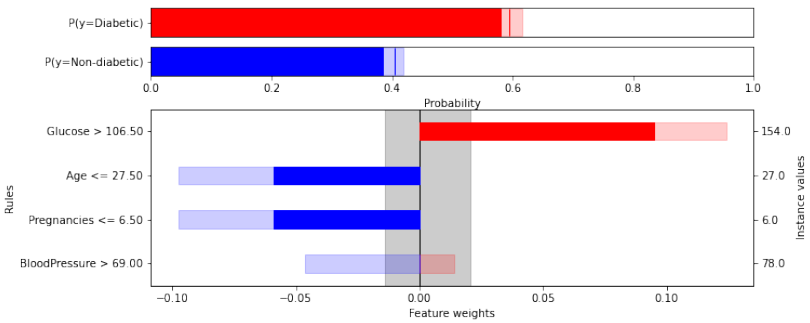

Figure 4.6: Uncertainty CE plot, in which intervals are added to the regular plot to provide information about the uncertainty associated with each feature's contribution to the prediction.

The Venn-Abers interval for the underlying model is indicated in the probability bars at the top, in light blue or light red, and as the light grey area in the central part of the plot. Similarly, the feature weights are solidly coloured, from the black line (which represents the calibrated probability estimate) to the lower bound on the weight interval, and lightly coloured between the lower and upper bounds on the interval. The rules are ordered based on the feature weights of the calibrated prediction. Blue bars indicate that the feature supports Class 0 (*non-diabetic*), whereas red bars indicate support for class 1 (*diabetic*).

- **Counterfactual**: In *Counterfactual CE* (CCE), the plots do not show feature weights; instead, this type of plot focuses on the Venn-Abers probability intervals. Each rule shows the alternative Venn-Abers probability interval resulting from changing the feature value to a value covered by the counterfactual rule condition. Numerical features can result in at most two counterfactual rules (above or below the thresholds surrounding the feature value), whereas one counterfactual rule is created for each alternative categorical value. In the plots, up to the 10 most influential counterfactuals are shown.

  Like the uncertainty plot, the counterfactual plot in Figure 4.7 shows that the two features *Glucose* and *Age* affect the prediction most. However, in the plot, the rules show that even a small reduction in the *Glucose* feature ($\texttt{Glucose} < 152.5$) may change the probability

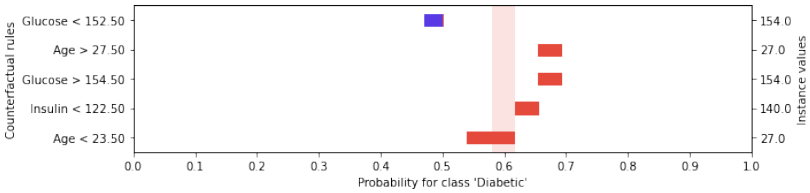

Figure 4.7: CCE plot where each rule shows the alternative Venn-Abers probability interval resulting from changing the feature value to a value covered by the counterfactual rule condition.

estimate enough to change the prediction of the model. Higher *Age* and *Glucose* will increase the likelyhood for the patient being *diabetic*.

Since this paper was published, additional functionality has been introduced to the CE package for multiclass regression and conjunctional rules. As it uses a simple approach and addresses only one feature and its importance at a time, the original CE is fast and easy to understand. However, one of the strengths of SHAP is that it considers all combinations of all features, although this simultaneously increases the computational cost. One additional functionality of the updated CE includes the combined importance of the highest weighted features (conjunctional rules), which strengthens the explanation method without dramatically increasing the computational cost. A paper based on these findings is under review, and these extensions can be used with the Python package.

Method

In this paper, the evaluation consisted of two parts to evaluate the *stability* and the *robustness* of the method. Stability was evaluated through experiments where the same model, calibration set and test set were explained 30 times per dataset. The only source of variation was the random seed. Robustness was evaluated in the second experiment, where training and calibration sets were randomly resampled before a new model was trained and explained. Robustness was measured in this way to avoid inferring perturbed instances which are not from the same distribution as the test instances being explained. The probability estimate of each of the models was computed on the same test set as a comparison to the robustness results. The expectation was that a stable and robust explanation method should result in low variance in the feature importance weights. During the experimentation, the run time was measured and also reported. The test set was 20 stratified instances (making sure both classes were equally represented). Both random forests and xGBoost were used, and factual and counterfactual explanations were evaluated. Counterfactual explanations (CCE) were slightly slower than factual explanations (CE) since more than one counterfactual rule can be created per numerical feature (there is no difference for categorical features). For an extensive description of the method, see the Appendix A.5.

4.3.3 Summary of Contributions

The second research question (RQ2) has been answered in the two final papers with in distinct steps:

1. Paper IV studied the impact of explanation quality when the underlying model was calibrated. Various calibration techniques were considered to find the most effective and suitable for the purpose of increasing the explanation quality. VA was identified as the best alternative for calibration, and the possibilities of obtaining uncertainty intervals was noted as highly interesting for explanation methods.
2. In the final paper, Paper V, the knowledge created in the earlier papers was used to develop a novel explanation method, based on the Venn-Abers technique, that could generate robust, stable, and reliable explanations while also providing valuable uncertainty information with a limited computational cost.

4.4 Contributions Related to Decision Support

This thesis aims to support the user in the decision-making process through the development of explanation methods for predictive models in ML. Uncertainty is a critical factor that often complicates the use of predictive models for decision support. The initial papers making up this thesis demonstrate that even when explanation methods are used, uncertainty may lead to either misuse (from persuasions) or disuse (from not trusting the predictions), and emphasise the critical nature of managing uncertainty in decision support. For example, Phillips-Wren and Adya (2020) highlight user uncertainty as one of four stress factors that affect the quality of a decision.

Recent studies have recognised the necessity of incorporating uncertainty estimation into explanations to enhance the transparency of the underlying models (Antorán et al., 2020; Bhatt et al., 2021; Slack et al., 2021)), although, authors such as Zhou et al. (2021),have pointed out that the major problem for users of predictive models is not primarily the opaqueness or the complexity of the model, but the output uncertainty. In the example of a doctor who is deciding on a diagnosis, revealing the uncertainty of the underlying model could cause the doctor to put the case aside for closer inspection, thereby increasing confidence in the diagnosis. The critical aspect of uncertainty information in explanations aligns with how uncertainty is acknowledged as a foundational element in decision theory (DT) (see Section 2.1.1) (Parmigiani & Inoue, 2009).

Probabilities are typically used to assess the level of confidence in predictions, which often serves as the standard representation of uncertainty and is seen as effectively capturing various aspects of the uncertainty that influences decision-making processes (Kolmogorov & Bharucha-Reid, 2018). How ever, during the study reported in Paper IV, it became apparent that poor calibration of the underlying model not only distorts prediction probabilities but also influences the magnitude of potential errors in explanations. This distortion hampers the user's ability to make informed decisions, as incorrect information creates challenges in terms of arriving at the most logical decision possible.

The explanations in CE, presented in Paper V, do not solve the problem of uncertainty, since uncertainty is an inherent aspect of reality. Nonetheless, they reveal trustworthy information about the model's uncertainty to the user, thus laying a foundation for more prudent and informed decisions, as the only form of protection against poor decisions

4.5 Implications

In this thesis, approaches for explaining complex ML models have been investigated and refined in order to provide suggestions for improving and assessing the quality of explanation methods to support high-quality decisions. This research resulted in an explanation method that can simultaneously calibrate the underlying model and generate robust and reliable (well-calibrated) explanations, while also providing valuable uncertainty information at a limited computational cost. Moreover, the AT criterion (representing the level of appropriate trust) has been defined, which can facilitate comparative user evaluations of explanation methods.

As seen in the literature meta-survey, the area of explanation methods is in need of more rigour. Although there are a plethora of explanation methods, there is no consensus on how to evaluate them for comparative outcomes. This thesis contributes to increasing the rigour in this field through the meta-survey in Paper II, where existing evaluation criteria are collected and organised in a high-level model. However, this study is not limited to theoretical implications. In at least one of the articles that have cited this paper(Galanti et al., 2023)), the results are also used to form suggestions for the design of practices for future work, and the suggestion in Paper II of using criteria such as appropriate trust to get an objective result is highlighted. Appropriate trust is a very promising criterion for comparative results. However, Paper II notes that its definition is sometimes rather vague, and there is almost an absence of method, which could pose challenges for the researchers or practitioners who wish to use it. The outcome of Paper III is a toolset that enables comparative evaluations based on the level of appropriate trust (through AT) with a well-formulated method, thus strengthening the rigour in the field.

A separate explanation method is added to the underlying model when explaining complex ML models with post hoc methods. However, a well-known problem within ML is that predictive models often are poorly calibrated. To manage this problem, the practitioner needs to add a calibration technique to the model. With some of these techniques, it is possible to generate information about their uncertainty, although most do not capture the uncertainty of the underlying model. In this thesis, the proposed explanation method, known as CE, can be used to calibrate the underlying model and generate robust and reliable explanations, revealing the uncertainty from the underlying model at a low computational cost. The proposed method can be installed as a package in Python via Pipy[2] or to follow in Github[3].

There are several implications for practitioners, for example: (i) the underlying model is calibrated, the probability estimates are trustworthy, and the practitioner does not need to add a technique for calibration; (ii) the given explanations are accurate, robust, stable and have a low computational cost, resulting in an effective method; and (iii) the uncertainty information generated as part of the calibration is revealed to the user, meaning that the practitioner does not need to add any other technique than CE to increase the model's trustworthiness.

Summary: The implications of this thesis will primarily (although not exclusively) be of interest to researchers and practitioners. The contributions focus on making it easier for researchers and practitioners to explain complex ML models, and how to assess the quality of an explanation. However, the thesis also includes guidelines on the essential aspects of trustworthy explanation methods and evaluation criteria.

[2]https://pypi.org/project/calibrated-explanations/
[3]https://github.com/Moffran/calibrated_explanations

5 Conclusion and Future Work

"*I am glad you are here with me. Here at the end of all things*"

— J.R.R. Tolkien, The Return of the King

This final chapter presents the conclusions of the thesis, followed by suggestions for future work.

5.1 Conclusions

When we look back at the system controlled by Stanislav Petrov four decades ago, it is remarkable that these critical decisions were based on unstable and imprecise models. Although modern predictive models have improved significantly, they still suffer from errors, particularly in terms of calibration. We find ourselves facing a similar question today: in critical situations, how can we rely on questionably calibrated models that produce skewed explanations, and fail to reveal the uncertainty inherent in the underlying model?

The aim of this thesis has been to assess and improve the quality of explanations for predictive models in ML, in order to lay a foundation for appropriate trust that can support users in the decision-making process. To achieve this aim, and based on identified areas of inquiry (the development and evaluation of explanation methods in ML), a research question was formulated as follows: *"How can explanation methods for decision support be improved in terms of trustworthiness?"*.Two sub-questions were formulated to answer this research question: RQ1) *"What criteria could be used to evaluate the usability of explanation methods for decision support?"* and RQ2) *"How can well-calibrated uncertainty information be included in explanation methods for decision support?"*. The thesis includes five studies in which a mixed-method approach was used to answer the two sub-questions and the main question from various angles.

The first research question was answered through the contributions of the three first papers. Paper I emphasised that relying solely on the subjective criterion of trust is not sufficient for an assessment of the quality of an explanation method, since users may trust both correct and incorrect predictions. In Paper II, a literature survey was conducted to find existing criteria, and it was highlighted that the quality of an explanation comprises three crucial aspects: the model, the explanation, and the user. To overcome the challenges preventing comparative studies, which arise from the use of subjective criteria by the user, it was suggested that the objective criterion of appropriate trust should be used, which comprises the outcome of the subjective criteria. However, appropriate trust was found to have a definition that was often vague, and almost an absence of method. To make it possible to use appropriate trust as a criterion, as suggested in Paper II, Paper III properly defined appropriate

trust based on the literature, and derived a metric for measuring appropriate trust, *AT*, from the strong similarities between the definitions and the evaluation of binary classifiers.

The second research question was answered through the contributions from the two last papers in the thesis. The findings in Paper IV demonstrated that the calibration of the underlying model was crucial to ensure the accuracy and reliability of feature relevance explanation methods, as a well-calibrated model results in better calibrated explanations that accurately represent reality. Different calibration techniques were studied to find the most suitable one to strengthen the explanation quality. The VA technique was identified as the best option for producing well-calibrated probability estimates, and the possibility of obtaining uncertainty intervals was noted as highly interesting for explanation methods. In the final paper, the knowledge gathered from the earlier papers was used to develop a novel explanation method, CE, that generates robust, stable, and reliable explanations while also providing valuable uncertainty information at a low computational cost.

In conclusion, the contributions from the two sub-questions were combined to answer the main research question, firstly through guidelines on the evaluation of the trustworthiness of explanation methods, and secondly by providing an explanation method that generates robust, stable, and reliable explanations with uncertainty information at a low computational cost, thereby enhancing trustworthiness and effectiveness.

The findings of this thesis are primarily aimed at researchers and practitioners, as they facilitate the explanations of complex ML models and the assessment of explanation quality. The guidelines developed here encompass essential aspects of trustworthy explanation methods and evaluation criteria.

5.2 Future Work

There are several areas that have been left for future work, since there was not enough time during the work on the thesis to include them, for example:

Evaluation of CE with expert users . CE is an explanation method with a strong ability to help users gain an appropriate level of trust in the underlying model. Evaluations of explanation methods are often conducted either with online users or through a short, time-limited evaluation with users who are not accustomed to explanation methods, which could skew the results. In future work, evaluations could be conducted with expert users and their approach to the method could be considered over a longer time window, to explore both how their initial appropriate trust changes when they first meet the explanations and how it develops over time.

Defined expression for AT with uncertainty in regression . When developing the formalised methods for appropriate trust in Paper III, a possible approach to evaluating appropriate trust in regression and including user uncertainty was discussed. In future work, a method for appropriate trust in regression could be studied that includes uncertainty information.

Bibliography

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, *6*, 52138–52160.

Adya, M., & Phillips-Wren, G. (2020). Stressed decision makers and use of decision aids: A literature review and conceptual model. *Information Technology & People*, *33*(2), 710–754.

Agarwal, C., Saxena, E., Krishna, S., Pawelczyk, M., Johnson, N., Puri, I., Zitnik, M., & Lakkaraju, H. (2022). Openxai: Towards a transparent evaluation of model explanations. *arXiv preprint arXiv:2206.11104*.

Ågerfalk, P. J. (2013). Embracing diversity through mixed methods research.

Alvarado-Valencia, J. A., & Barrero, L. H. (2014). Reliance, trust and heuristics in judgmental forecasting. *Computers in human behavior*, *36*, 102–113.

Alvarez-Melis, D., & Jaakkola, T. S. (2018). On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049*.

Antorán, J., Bhatt, U., Adel, T., Weller, A., & Hernández-Lobato, J. M. (2020). Getting a clue: A method for explaining uncertainty estimates. *arXiv preprint arXiv:2006.06848*.

Arnott, D., & Pervan, G. (2014). A critical analysis of decision support systems research revisited: The rise of design science. *Journal of Information Technology*, *29*(4), 269–293. https://doi.org/10.1057/jit.2014.16

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, *58*, 82–115.

Ascarza, E., Netzer, O., & Hardie, B. G. (2018). Some customers would rather leave without saying goodbye. *Marketing Science*, *37*(1), 54–77.

Badings, T., Simão, T. D., Suilen, M., & Jansen, N. (2023). Decision-making under uncertainty: Beyond probabilities: Challenges and perspectives. *International Journal on Software Tools for Technology Transfer*, 1–17.

Baeza-Yates, R., Ribeiro-Neto, B. et al. (2011). *Modern information retrieval* (2. ed.). Addison-Wesley.

Berthold, M. R., Borgelt, C., Höppner, F., Klawonn, F., & Silipo, R. (2020). *Guide to intelligent data science*. Springer.

Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., Ohl, P., Sieb, C., Thiel, K., & Wiswedel, B. (2008). Knime: The konstanz information miner. In C. Preisach, H. Burkhardt, L. Schmidt-Thieme, & R. Decker (Eds.), *Data analysis, machine learning and applications* (pp. 319–326). Springer Berlin Heidelberg.

Bhatt, U., Antorán, J., Zhang, Y., Liao, Q. V., Sattigeri, P., Fogliato, R., Melançon, G., Krishnan, R., Stanley, J., Tickoo, O., et al. (2021). Uncertainty as a form of transparency: Measuring,

communicating, and using uncertainty. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 401–413.

Bradley, R. (2018). Decision theory: A formal philosophical introduction. *Introduction to formal philosophy*, 611–655.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5–32.

Buçinca, Z., Lin, P., Gajos, K. Z., & Glassman, E. L. (2020). Proxy tasks and subjective measures can be misleading in evaluating explainable ai systems. *Proceedings of the 25th international conference on intelligent user interfaces*, 454–464.

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, *8*, 832. https://doi.org/10.3390/electronics8080832

Chiou, A., & Wong, K. W. (2010). Auto-explanation system: Player satisfaction in strategy-based board games. *Entertainment Computing Symposium*, 46–54.

Chromik, M., & Schuessler, M. (2020). A taxonomy for human subject evaluation of black-box explanations in xai. *Exss-atec@ iui*, *1*.

Coussement, K., & Van den Poel, D. (2008). Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. *Expert systems with applications*, *34*(1), 313–327.

Creamer, E. G. (2017). *An introduction to fully integrated mixed methods research*. sage publications.

Creswell, J. W. (2014). Qualitative, quantitative and mixed methods approaches.

Das, A., & Rad, P. (2020). Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv preprint arXiv:2006.11371*.

da Silva, H. C. C., de Oliveira Siqueira, A., Araújo, M. A. V., & Dornelas, J. S. (2018). Let's be pragmatic: Research in information systems with relevance and rigor. *International Journal of Business Management & Economic Research*, *9*(4).

Davoudi, H., Zihayat, M., & An, A. (2017). Time-aware subscription prediction model for user acquisition in digital news media. *Proceedings of the 2017 SIAM International Conference on Data Mining*, 135–143.

Dechant, A., Spann, M., & Becker, J. U. (2019). Positive customer churn: An application to online dating. *Journal of Service Research*, *22*(1), 90–100.

Dimanov, B., Bhatt, U., Jamnik, M., & Weller, A. (2020). You shouldn't trust me: Learning models which conceal unfairness from multiple explanation methods. *Frontiers in Artificial Intelligence and Applications: ECAI 2020*.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Dua, D., & Graff, C. (2022, March 1). UCI machine learning repository. http://archive.ics.uci.edu/ml

Dunham H, M. (2003). Data mining introductory and advanced topics. *Pearsons Education Inc*.

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies*, *58*(6), 697–718.

Eklund, J. (2016). *With or without context: Automatic text categorization using semantic kernels* (Doctoral dissertation). University of Borås. Borås, Sweden. https://urn.kb.se/resolve?urn=urn:nbn:se:hb:diva-8949

Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *J. Mach. Learn. Res.*, *20*(177), 1–81.

Flach, P. (2012). *Machine learning: The art and science of algorithms that make sense of data*. Cambridge University Press.

Galanti, R., de Leoni, M., Monaro, M., Navarin, N., Marazzi, A., Di Stasi, B., & Maldera, S. (2023). An explainable decision support system for predictive process analytics. *Engineering Applications of Artificial Intelligence*, *120*, 105904.

Gallant, S. I. (1993). *Neural network learning and expert systems*. MIT press.

Géron, A. (2022). *Hands-on machine learning with scikit-learn, keras, and tensorflow*. " O'Reilly Media, Inc."

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: An approach to evaluating interpretability of machine learning. *arXiv preprint arXiv:1806.00069*, 118.

Greene, J. C., Caracelli, V. J., & Graham, W. F. (1989). Toward a conceptual framework for mixed-method evaluation designs. *Educational evaluation and policy analysis*, *11*(3), 255–274.

Grushka-Cockayne, Y., Jose, V. R. R., & Lichtendahl Jr, K. C. (2017). Ensembles of overfit and overconfident forecasts. *Management Science*, *63*(4), 1110–1130.

Guidotti, R., Monreale, A., Giannotti, F., Pedreschi, D., Ruggieri, S., & Turini, F. (2019). Factual and counterfactual explanations for black box decision making. *IEEE Intelligent Systems*, *34*(6), 14–23.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, *51*(5), 1–42.

Gunning, D., & Aha, D. W. (2019). Darpa's explainable artificial intelligence program. *AI Magazine*, *40*(2), 44–58.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS quarterly*, 75–105.

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors*, *57*(3), 407–434.

Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). Metrics for explainable ai: Challenges and prospects. *arXiv preprint arXiv:1812.04608*.

Holzinger, A., Saranti, A., Molnar, C., Biecek, P., & Samek, W. (2020). Explainable ai methods-a brief overview. *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*, 13–38.

Hopkins, E. J., Weisberg, D. S., & Taylor, J. C. (2016). The seductive allure is a reductive allure: People prefer scientific explanations that contain logically irrelevant reductive information. *Cognition*, *155*, 67–76.

Hopkins, E. J., Weisberg, D. S., & Taylor, J. C. (2019). Does expertise moderate the seductive allure of reductive explanations? *Acta psychologica*, *198*, 102890.

Jacovi, A., Marasović, A., Miller, T., & Goldberg, Y. (2021). Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai. *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 624–635.

Kolmogorov, A. N., & Bharucha-Reid, A. T. (2018). *Foundations of the theory of probability: Second english edition*. Courier Dover Publications.

Kulesza, T., Stumpf, S., Burnett, M., Yang, S., Kwan, I., & Wong, W.-K. (2013). Too much, too little, or just right? ways explanations impact end users' mental models. *2013 IEEE Symposium on Visual Languages and Human Centric Computing*, 3–10.

Lacave, C., & Diez, F. J. (2002). A review of explanation methods for bayesian networks. *The Knowledge Engineering Review*, *17*(2), 107–127.

Lacave, C., & Diez, F. J. (2004). A review of explanation methods for heuristic expert systems. *The Knowledge Engineering Review*, *19*(02), 133–146.

Lambrou, A., Nouretdinov, I., & Papadopoulos, H. (2015). Inductive venn prediction. *Annals of Mathematics and Artificial Intelligence*, *74*(1), 181–201.

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2021). Explainable ai: A review of machine learning interpretability methods. *Entropy*, *23*(1), 18.

Linoff, G. S., & Berry, M. J. (2011). *Data mining techniques: For marketing, sales, and customer relationship management*. John Wiley & Sons.

Lipton, Z. C. (2018). The mythos of model interpretability. *Queue*, *16*(3), 31–57.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Proceedings of the 31st international conference on neural information processing systems*, 4768–4777.

Martens, D., & Foster, P. (2014). Explaining data-driven document classifications. *MIS Quaterly*, *38*(1), 73–100.

McCusker, K., & Gunaydin, S. (2015). Research using qualitative, quantitative or mixed methods and choice based on the research. *Perfusion*, *30*(7), 537–542.

Mehrtash, A., Wells, W. M., Tempany, C. M., Abolmaesumi, P., & Kapur, T. (2020). Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *IEEE transactions on medical imaging*, *39*(12), 3868–3878.

Mohseni, S., Zarei, N., & Ragan, E. D. (2018). A multidisciplinary survey and framework for design and evaluation of explainable ai systems. *arXiv*, arXiv–1811.

Molnar, C. (2022). *Interpretable machine learning: A guide for making black box models explainable* (2nd ed.). https://christophm.github.io/interpretable-ml-book

Moradi, M., & Samwald, M. (2020). Post-hoc explanation of black-box classifiers using confident itemsets. *arXiv preprint arXiv:2005.01992*.

Mueller, S. T., Hoffman, R. R., Clancey, W., Emrey, A., & Klein, G. (2019). Explanation in human-ai systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable ai. *arXiv preprint arXiv:1902.01876*.

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, *116*(44), 22071–22080.

Negnevitsky, M. (2005). *Artificial intelligence: A guide to intelligent systems*. Pearson education.

Nguyen, D. (2018). Comparing automatic and human evaluation of local explanations for text classification. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1069–1078. https://doi.org/10.18653/v1/N18-1097

Nielsen, J., & Landauer, T. K. (1993). A mathematical model of the finding of usability problems. *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*, 206–213.

North, D. W. (1968). A tutorial introduction to decision theory. *IEEE transactions on systems science and cybernetics*, *4*(3), 200–210.

Nouretdinov, I., Volkhonskiy, D., Lim, P., Toccaceli, P., & Gammerman, A. (2018). Inductive venn-abers predictive distribution. *Conformal and Probabilistic Prediction and Applications*, 15–36.

Nti, I. K., yarko-Boateng, O. N., & Aning, J. (2021). Performance of machine learning algorithms with different k values in k-fold crossvalidation. *International Journal of Information Technology and Computer Science*.

Oates, B. J. (2005). *Researching information systems and computing*. Sage.

Parmigiani, G., & Inoue, L. (2009). *Decision theory: Principles and approaches*. John Wiley & Sons.

Pavlidis, M., Mouratidis, H., Islam, S., & Kearney, P. (2012). Dealing with trust and control: A meta-model for trustworthy information systems development. *2012 Sixth International Conference on Research Challenges in Information Science (RCIS)*, 1–9.

Phillips-Wren, G., & Adya, M. (2020). Decision making under stress: The role of information overload, time pressure, complexity, and uncertainty. *Journal of Decision Systems*, *29*(sup1), 213–225.

Plumb, G., Molitor, D., & Talwalkar, A. S. (2018). Model agnostic supervised local explanations. *Advances in neural information processing systems*, *31*.

Python Software Foundation. (2021, August 30). *Python* (Version 3.9.7). https://www.python.org/downloads/release/python-397/

Queirós, A., Faria, D., & Almeida, F. (2017). Strengths and limitations of qualitative and quantitative research methods. *European journal of education studies*.

Rahnama, A. H. A., & Boström, H. (2019). A study of data and label shift in the lime framework. *arXiv preprint arXiv:1910.14421*.

Recker, J. (2012). *Scientific research in information systems: A beginner's guide*. Springer Science & Business Media.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "why should i trust you?": Explaining the predictions of any classifier. *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778

Risk, O., & Bernoulli, D. (1954). Exposition of a new theory on the measurement. *Econometrica*, *22*(1), 23–36.

Robnik-Šikonja, M., & Kononenko, I. (2008). Explaining classifications for individual instances. *IEEE Transactions on Knowledge and Data Engineering*, *20*(5), 589–600.

Sayyad Shirabad, J., & Menzies, T. (2005). PROMISE Repository of Software Engineering Databases.

Sestito, S., & Dillon, T. (1991). Using single-layered neural networks for the extraction of conjunctive rules and hierarchical classifications. *Applied Intelligence*, *1*, 157–173.

Sheridan, T. B., Sheridan, T. B., Maschinenbauingenieur, K., Sheridan, T. B., & Sheridan, T. B. (2002). *Humans and automation: System design and research issues* (Vol. 280). Human Factors; Ergonomics Society Santa Monica, CA.

Skitka, L. J., Mosier, K. L., & Burdick, M. (1999). Does automation bias decision-making? *International Journal of Human-Computer Studies*, *51*(5), 991–1006.

Slack, D., Hilgard, A., Singh, S., & Lakkaraju, H. (2021). Reliable post hoc explanations: Modeling uncertainty in explainability. *Advances in neural information processing systems*, *34*, 9391–9404.

Tibshirani, R. J., Hoefling, H., & Tibshirani, R. (2011). Nearly-isotonic regression. *Technometrics*, *53*(1), 54–61.

Todd, P., & Benbasat, I. (1992). The use of information in decision making: An experimental investigation of the impact of computer-based decision aids. *Mis Quarterly*, 373–393.

Van Calster, B., McLernon, D. J., Van Smeden, M., Wynants, L., & Steyerberg, E. W. (2019). Calibration: The achilles heel of predictive analytics. *BMC medicine*, *17*(1), 1–7.

van der Waa, J., Schoonderwoerd, T., van Diggelen, J., & Neerincx, M. (2020). Interpretable confidence measures for decision support systems. *International Journal of Human-Computer Studies*, *144*, 102493.

Vom Brocke, J., Hevner, A., & Maedche, A. (2020). Introduction to design science research. *Design science research. Cases*, 1–13.

Vovk, V., Gammerman, A., & Shafer, G. (2005). *Algorithmic learning in a random world*. Springer Science & Business Media.

Vovk, V., & Petej, I. (2012). Venn-Abers predictors. *arXiv preprint arXiv:1211.0025*.

Vovk, V., Shafer, G., & Nouretdinov, I. (2004). Self-calibrating probability forecasting. *Advances in Neural Information Processing Systems*, 1133–1140.

Wang, D., Yang, Q., Abdul, A., & Lim, B. Y. (2019). Designing theory-driven user-centric explainable ai. *Proceedings of the 2019 CHI conference on human factors in computing systems*, 1–15.

Weisberg, D. S., Taylor, J. C. V., & Hopkins, E. J. (2015). Deconstructing the seductive allure of neuroscience explanations. *Judgment and Decision Making*, *10*(5), 429–441.

Weisberg, D. S., Keil, F. C., Goodstein, J., Rawson, E., & Gray, J. R. (2008). The seductive allure of neuroscience explanations. *Journal of cognitive neuroscience*, *20*(3), 470–477.

Wildemuth, B. M. (2009). *Application of social methods to questions in information and library science*. Libraries Unlimited.

Wilson, J. (2014). Essentials of business research: A guide to doing your research project. *Essentials of Business Research*, 1–376.

Wynants, L., Van Calster, B., Collins, G. S., Riley, R. D., Heinze, G., Schuit, E., Albu, E., Arshi, B., Bellou, V., Bonten, M. M., et al. (2020). Prediction models for diagnosis and prognosis of covid-19: Systematic review and critical appraisal. *bmj*, *369*.

Yang, F., Huang, Z., Scholtz, J., & Arendt, D. L. (2020). How do visual explanations foster end users' appropriate trust in machine learning? *Proceedings of the 25th International Conference on Intelligent User Interfaces*, 189–201.

Zadrozny, B., & Elkan, C. (2001). Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. *Proc. 18th International Conference on Machine Learning*, 609–616.

Zadrozny, B., & Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 694–699.

Zhang, Y., & Chen, X. (2018). Explainable recommendation: A survey and new perspectives. *arXiv preprint arXiv:1804.11192*.

Zhou, J., Gandomi, A. H., Chen, F., & Holzinger, A. (2021). Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, *10*(5), 593.

Trustworthy Explanations

Improved Decision Support Through Well-Calibrated Uncertainty Quantification

The use of Artificial Intelligence (AI) has transformed fields like disease diagnosis and defence. Utilising sophisticated Machine Learning (ML) models, AI predicts future events based on historical data, introducing complexity that challenges understanding and decision-making. Previous research emphasizes users' difficulty discerning when to trust predictions due to model complexity, underscoring addressing model complexity and providing transparent explanations as pivotal for facilitating high-quality decisions.

Many ML models offer probability estimates for predictions, commonly used in methods providing explanations to guide users on prediction confidence. However, these probabilities often do not accurately reflect the actual distribution in the data, leading to potential user misinterpretation of prediction trustworthiness. Additionally, most explanation methods fail to convey whether the model's probability is linked to any uncertainty, further diminishing the reliability of the explanations.

Evaluating the quality of explanations for decision support is challenging, and although highlighted as essential in research, there are no benchmark criteria for comparative evaluations.

This thesis introduces an innovative explanation method that generates reliable explanations, incorporating uncertainty information supporting users in determining when to trust the model's predictions. The thesis also outlines strategies for evaluating explanation quality and facilitating comparative evaluations. Through empirical evaluations and user studies, the thesis provides practical insights to support decision-making utilising complex ML models.

HELENA LÖFSTRÖM received a Bachelor's degree in Computer Science in 1997 and a Master's in Library and Information Science in 2018, both from University of Borås. Since 2018, she has been a PhD student at Jönköping International Business School, Jönköping University, connected to the INSiDR and MIT research schools. The main fields of her research interests are explanation methods for complex machine learning models.

ISSN 1403-0470
ISBN 978-91-7914-031-1 (Printed version)
ISBN 978-91-7914-032-8 (Online version)